CARL VON OSSIETZKY universität OLDENBURG

VERY LARGE BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Abschlussbericht

der Projektgruppe HAWK

Version zur Veröffentlichung

Verantwortlicher: Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer: Dipl.-Wirt. Inform. Andreas Solsbach, M.Sc. Andreas Messler

Vorgelegt von: Jad Asswad, Björn Berger, Jonas Borrmann, Simon Brosig, Sören Flathmann, Marco Lucht, Catharina Petry, Benjamin Reinecke, Yvonne Tevs, Christoph Walther, Jan-Hendrik Wege, Jan-Patrick Weiß

Abgabetermin: 31. März 2014

Die Mitglieder der Projektgruppe HAWK bedanken sich bei ihren Betreuern Prof. Dr.-Ing. Jorge Marx Gómez, M.Sc. Andreas Messler, Dipl.-Wirt. Inform. Andreas Solsbach sowie Viktor Dmitriyev und Hans-Hermann Redenius für die Koordination und die Unterstützung während der gesamten Projektlaufzeit. Hierbei gilt unser Dank auch für die Zeit der wöchentlichen Projektgruppentreffen und der Teilnahme an öffentlichen Terminen. Über den gesamten Zeitraum waren sie Ansprechpartner für inhaltliche sowie organisatorische Fragen und haben mit interessierten Diskussionen zur Verbesserung unseres Projektergebnisses beigetragen. Darüber hinaus gilt unser Dank ebenso für die Initiierung und Organisation der Ausstellung auf der Messe CeBIT in Hannover, da es uns die Möglichkeit eröffnet hat, Teile unserer Projektarbeit einem interessierten Publikum vorzustellen und weiteres Feedback zu erhalten.

Wir möchten uns außerdem ganz herzlich bei unseren Partnerunternehmen, der Volkswagen AG und der CX4U AG, für die vertrauensvolle und erfolgreiche Zusammenarbeit bedanken. Insbesondere Frau Tanja Führmann, Herrn Thomas Pröttel und Herrn Gerhard Korths von der Volkswagen AG sowie Herrn Okko Nannen und Herrn Frank Locher von der CX4U AG danken wir für die inhaltliche Unterstützung sowie die Zeit für regelmäßige Treffen und Telefonate. Aufgrund der Bereitstellung der technischen Systeme war eine erfolgreiche Projektarbeit erst möglich.

Projektgruppe HAWK im März 2014

Inhaltsverzeichnis

Abkürzungsverzeichnis **IV**

Abbildungsverzeichnis **VIII**

Tabellenverzeichnis **IX**

Listingverzeichnis **X**

1 Einleitung **1**
1.1 Motivation . . . 1
1.2 Problemstellung und Zielsetzung der Projektgruppe HAWK . . . 2
1.3 Aufbau der Arbeit . . . 3

2 Rahmenbedingungen **4**
2.1 Organisatorische Rahmenbedingungen . . . 4
2.1.1 Konzept und Ziel einer Projektgruppe . . . 4
2.1.2 Projektgruppenmitglieder und ihre Rollen . . . 5
2.1.3 Vorbereitung und Seminararbeitsthemen . . . 6
2.1.4 Projektverlauf . . . 8
2.2 Technische Rahmenbedingungen . . . 10
2.2.1 Confluence als interne Kommunikationsplattform . . . 10
2.2.2 Dokumentation mittels LaTeX und Jabref . . . 11
2.2.3 Homepagearbeit . . . 14
2.2.4 Nutzung von Redmine für die Projektgruppenorganisation . . . 14
2.2.5 Datensicherheit durch SVN . . . 14

3 Teilgruppe Big Data im IT-Capacity Management **15**
3.1 Projektdefinition . . . 18
3.1.1 Praxispartner . . . 18
3.1.2 Aufgabenstellung . . . 19
3.1.3 IT-Capacity Management . . . 21
3.2 Projektorganisation . . . 24
3.2.1 Organisationsstruktur . . . 24
3.2.2 Projektplan . . . 25
3.2.3 Projektverlauf . . . 26
3.3 Referenzarchitekturen . . . 28
3.3.1 Einleitung . . . 28
3.3.2 Szenario A . . . 30
3.3.3 Szenario B . . . 32
3.3.4 Szenario C . . . 40
3.3.5 Besondere Herausforderungen . . . 42
3.3.6 Beispielhafte Umsetzung einer Referenzarchitektur . . . 47
3.3.7 Installation eines Hadoop-Clusters mit Hortonworks . . . 51
3.4 Marktüberblick . . . 60
3.4.1 Problemorientierte Einleitung . . . 60
3.4.2 Anbieter und deren Produkte . . . 69
3.4.3 Vorstellung der Ergebnisse . . . 133

3.4.4 Schlussbetrachtung . . . 139
3.5 Erstellung eines Prototyps zur Datenintegration . . . 141
3.5.1 Aufgabenstellung . . . 141
3.5.2 Beschreibung der Werkzeuge . . . 141
3.5.3 Konzeption . . . 142
3.5.4 Umsetzung . . . 146
3.6 Erstellung eines Prototypen zur „Simulation für Capacity Management“ . . . 161
3.6.1 Einleitung . . . 161
3.6.2 Beschreibung des Werkzeugs IBM Cognos TM1 . . . 162
3.6.3 Konzeption . . . 168
3.6.4 Umsetzung . . . 170
3.6.5 Benutzerhandbuch . . . 186
3.7 Schlussbetrachtung . . . 200

4 Teilgruppe Smart Metering 202
4.1 Projektdefinition . . . 205
4.1.1 Praxispartner . . . 205
4.1.2 Aufgabenstellung . . . 205
4.2 Projektorganisation . . . 206
4.2.1 Organisationsstruktur . . . 206
4.2.2 Projektplan . . . 207
4.2.3 Projektverlauf . . . 208
4.3 Anforderungsspezifikation . . . 211
4.3.1 Unternehmensbeschreibung der CX4U AG . . . 211
4.3.2 Zweck der Software . . . 213
4.3.3 Stakeholder . . . 216
4.3.4 Geschäftsanwendungsfälle . . . 216
4.3.5 Systemanwendungsfälle . . . 217
4.3.6 Systemanforderungen . . . 233
4.4 Systementwurf . . . 234
4.4.1 Systemzerlegung . . . 234
4.4.2 Hard- und Softwarekomponenten . . . 234
4.4.3 VPN-Verbindung zwischen den Systemen . . . 235
4.4.4 Dynamisches Modell . . . 235
4.4.5 Fachliches Datenmodell . . . 238
4.4.6 Sicherheitsaspekte und Zugriffsrechte . . . 238
4.5 Machbarkeitsstudie . . . 240
4.5.1 Aktueller Stand . . . 240
4.5.2 Beschreibung der Umsetzungsalternativen . . . 241
4.5.3 Bewertung der Umsetzungsalternativen . . . 256
4.5.4 Prototypen . . . 261
4.5.5 Schlussbetrachtung . . . 274
4.6 Qualitätsmanagement . . . 276
4.6.1 QM-Maßnahmen . . . 276
4.6.2 Funktionale Testfälle . . . 283
4.7 Öffentlichkeitsarbeit . . . 285
4.7.1 Future SOC Day . . . 285
4.7.2 CeBIT . . . 285

4.8 Schlussbetrachtung 287

5 Fazit **289**

F Anhang **i**

Literatur **lvii**

Abschließende Erklärung **lviii**

Abkürzungsverzeichnis

ADBC . . . . . ABAP Database Connectivity
API . . . . . Application Programming Interface
APL . . . . . Application Layer
BI . . . . . Buiness Intelligence
BI-CC . . . . . Business Intelligence Competence Center
BIRT . . . . . Business Intelligence and Reporting Tool
B2B . . . . . Business-to-business
CCL . . . . . Continous Computation Language
CDH . . . . . Cloudera Distribution mit Apache Hadoop
CeBIT . . . . . Centrum für Büroautomation, Informationstechnologie und Telekommunikation
CEP . . . . . Complex Event Processing Engines
CPU . . . . . Central processing unit
CRM . . . . . Customer-Relationship-Management
DBMS . . . . . Datenbankmanagementsystem
DBSL . . . . . Database Shared Library
DMAG . . . . . Deutsche Messe AG
DSMS . . . . . Data Stream Management System
DWH . . . . . Data-Warehouse
EDI . . . . . Electronic Data Interchange
EDM . . . . . Energy Data Management System
EMITS . . . . . Electronic Mail Interface Technical Support
ERP . . . . . Enterprise-Resource-Planning
ETL . . . . . Extract, Transform, Load
FSOC . . . . . Future SOC
GB . . . . . Gigabyte
GPS . . . . . Global Positioning System
GPU . . . . . Graphics processing unit
GUI . . . . . Graphical User Interface
HDFS . . . . . Hadoop Distributed File System
HPCI . . . . . High Performance Compressed Index
HPI . . . . . Hasso-Plattner-Institut
HQL . . . . . Hive Query Language
IMDSO . . . . . In-Memory Data Store Object
ITIL . . . . . IT Infrastructure Library
KPI . . . . . Key Performance Indicators
KMU . . . . . Kleine und mittlere Unternehmen
LDAP . . . . . Lightweight Directory Access Protocol
LZO . . . . . Lempel-Ziv-Oberhumer
MDUS . . . . . Meta Data Unification & Sync System
MDM . . . . . Meta Data Management System
MPP . . . . . Massively Parallel Processing
MWK . . . . . Niedersächsisches Ministerium für Wissenschaft und Kultur
oData . . . . . Open Data Protocol
ODS . . . . . Operational Data Store
OGC . . . . . Office of Government Commerce
OLAP . . . . . Online Analytical Processing

OLTP . . . . . Online-Transaction-Processing
POS . . . . . Point of Sale
PPTP . . . . . Point-to-Point Tunneling Protocoll
RAM . . . . . Random-Access Memory
RDBMS . . . . . relationales Datenbankmanagementsystem
REST . . . . . Representational State Transfer
RFC . . . . . Remote Function Call
RFID . . . . . Radio-Frequency Identification
RHEL . . . . . Red Hat Enterprise Linux
SaaS . . . . . Software-as-a-Service
SAP BW . . . . . SAP Business Warehouse
SAP ERP . . . . . SAP Enterprise Ressource Planing
SAP SRS . . . . . SAP Sybase Replication Server
SLT . . . . . SAP Landscape Transformation
SMP . . . . . Symmetrisches Multiprozessorsystem
SPS . . . . . Support Package Stack
T-KF BIT . . . . . Technologie-Komepetenzfeld Business Information Technologies
TB . . . . . Terabyte
TNS . . . . . Transparent Network Substrate
VEE . . . . . Validation, Estimation, Editing
VM . . . . . Virtuelle Maschine
VPN . . . . . Virtual Private Network
WMI . . . . . Windows Management Instrumentation
WYSIWYG . . . . . What you see is what you get
XS . . . . . SAP HANA Extended Application Services
ZFA . . . . . Zählerfernauslesung

Abbildungsverzeichnis

2.1 Projektgruppenmitglieder . . . 6
2.2 Screenshot Texmaker . . . 12
2.3 Screenshot Jabref . . . 13
3.1 Leistungs-Portfolio des BI-CCs . . . 19
3.2 Prozess des Capacity Managements . . . 23
3.3 Mitglieder der Teilprojektgruppe Big Data im IT-Capacity Management . . . 24
3.4 Projektplan der Teilgruppe Big Data im IT-Capacity Management . . . 25
3.5 Referenzarchitekturen . . . 29
3.6 Szenario A im Detail . . . 31
3.7 Referenzarchitektur - Szenario B . . . 32
3.8 Ansicht des SAP BO Data Services Designer . . . 35
3.9 Ansicht des Informatica Power Center Developer . . . 36
3.10 Referenzarchitektur - Szenario B - Daten nachladen . . . 38
3.11 Referenzarchitektur - Szenario B - Datenstand der APL zu Beginn . . . 39
3.12 Referenzarchitektur - Szenario B - Datenstand nach der zweiten Abfrage . . . 40
3.13 Referenzarchitektur - Szenario C . . . 40
3.14 Klassifikation von Schema-Matching-Methoden . . . 44
3.15 Domänenmodell zur Access-Log-Auswertung . . . 48
3.16 Extraktion und Transformation der Get-Requests . . . 48
3.17 HiveQL Abfrage zur Ermittlung der Anzahl der Zugriffe je Use Case . . . 48
3.18 Anzahl der Zugriffe pro Use Case . . . 49
3.19 Anzahl Zugriffe pro Stunde . . . 50
3.20 Use Case Payload in Byte . . . 50
3.21 BIRT Viewer . . . 51
3.22 Aufbau des Hadoop-Clusters . . . 52
3.23 Übersicht im Amazon Web Service EC2 über die Server - I . . . 53
3.24 Übersicht im Amazon Web Service EC2 über die Server -II . . . 53
3.25 Konfiguration des Dienstes HBase . . . 55
3.26 Das Dashboard . . . 56
3.27 Die Heatmaps . . . 56
3.28 Übersicht über die registrierten Hosts . . . 57
3.29 Details zu einem Hosts . . . 57
3.30 Die Dienste . . . 58
3.31 Die 4 Charakteristika von Big Data . . . 61
3.32 Produktkategorisierung Cloudera . . . 73
3.33 Produktkategorisierung Datameer . . . 76
3.34 Produktkategorisierung Datawatch . . . 78
3.35 Produktkategorisierung EMC . . . 81
3.36 Produktkategorisierung EXASOL . . . 85
3.37 Produktkategorisierung Hortonworks . . . 88
3.38 Produktkategorisierung Hewlett-Packard . . . 90
3.39 Produktkategorisierung IBM . . . 95
3.40 Produktkategorisierung Informatica . . . 98
3.41 Produktkategorisierung Jaspersoft . . . 101
3.42 Produktkategorisierung MicroStrategy . . . 103
3.43 Produktkategorisierung Oracle . . . 107

3.44 Produktkategorisierung ParStream . . . 109
3.45 Produktkategorisierung Pentaho . . . 113
3.46 Produktkategorisierung QlikTech . . . 116
3.47 Produktkategorisierung SAP . . . 121
3.48 Produktkategorisierung SAS . . . 124
3.49 Produktkategorisierung Software AG . . . 127
3.50 Produktkategorisierung Splunk . . . 130
3.51 Produktkategorisierung Teradata . . . 133
3.52 Anteil der Produkte in den aufgestellten Kategorien . . . 134
3.53 Erklärungen zu den Wertungen innerhalb der Kategorieübersichten . . . 135
3.54 Kategorieübersicht Speicherkonzepte . . . 136
3.55 Kategorieübersicht Anwenderwerkzeuge . . . 137
3.56 Kategorieübersicht Datenmanagement . . . 138
3.57 Logisches Datenmodell der ODS-Schicht . . . 144
3.58 Logisches Datenmodell der DWH-Schicht . . . 145
3.59 Quelltabelle zur relationalen Speicherung von Access Logs . . . 148
3.60 Zieltabelle zur Speicherung von Access Logs im ODS . . . 148
3.61 Mapping zur Speicherung von Access Logs im ODS . . . 149
3.62 Konfiguration des Workflows zur Speicherung von Access Logs im ODS . . . 149
3.63 Mapping-Tabellen für die Stamm- und Nutzerkontextdaten . . . 150
3.64 Mapping zur Datenanreicherung der Access Logs . . . 151
3.65 Zieltabelle für die Staging Area . . . 152
3.66 Zieltabellen für die Dimensionen des Zielschemas . . . 154
3.67 Mapping zur Datenbefüllung einer Dimensionstabelle . . . 154
3.68 Faktentabelle zur Auswertung der Access Logs . . . 155
3.69 Mapping zur Datenbefüllung der Faktentabelle . . . 155
3.70 Auswahl der Datenfelder . . . 157
3.71 Report Zugriffe je Stunde und Use Case . . . 158
3.72 Report Zugriffe je Stunde . . . 159
3.73 Report Zugriffe je Use Case . . . 159
3.74 Definieren eines Filters in BIRT . . . 160
3.75 Report Zugriffe je Land . . . 160
3.76 Dimensionen eines Cubes . . . 163
3.77 Konzeptmodell der Simulation . . . 169
3.78 Formeln für die neue Auslastung Prozessor und Speicher . . . 170
3.79 Modell der Dimensionen und Cubes . . . 173
3.80 Modell der Relationen zwischen den Cubes . . . 181
3.81 Prozessschritte der Simulation . . . 182
3.82 Anmeldeseite von TM1 Web . . . 187
3.83 Startseite von TM1 Web . . . 188
3.84 Alle Anwendungen im Navigationsbereich . . . 189
3.85 Seite Simulation-1-Neu des Simulationsassistenten . . . 189
3.86 Seite Simulation-2-Varianten des Simulationsassistenten . . . 190
3.87 Seite Simulation-3-Server des Simulationsassistenten . . . 191
3.88 Dialog zur Auswahl des Serves oder der Servergruppe . . . 191
3.89 Seite Ergebnis-Server des Simulationsassistenten . . . 192
3.90 Seite Ergebnis-Maximum des Simulationsassistenten . . . 193
3.91 Seite Ergebnis-Überschreitungen des Simulationsassistenten . . . 194

3.92 Seite Ergebnis-Ausstattung des Simulationsassistenten . . . 195
3.93 Seite Ergebnis-Szenario des Simulationsassistenten . . . 196
3.94 Seite Ergebnis-Details des Simulationsassistenten . . . 197
3.95 Seite Zugriffe der Analyse der Ist-Daten . . . 198
3.96 Seite Details der Analyse der Ist-Daten . . . 199
3.97 Seite Konfiguration der Analyse der Ist-Daten . . . 199
4.1 Organisationsstruktur der Teilgruppe Smart Metering . . . 206
4.2 Projektplan der Teilgruppe Smart Metering Soll-/ Ist-Vergleich . . . 208
4.3 Unternehmensgliederung . . . 211
4.4 Systemarchitektur MDUS . . . 213
4.5 Systemanwendungsfälle . . . 218
4.6 Validierung . . . 221
4.7 Estimation . . . 225
4.8 Regelauswahl . . . 228
4.9 Versionierung . . . 230
4.10 Aktivitätsdiagramm Editing . . . 232
4.11 Systemzerlegung des HANA-Systems . . . 234
4.12 Dynamisches Modell des HTTP-Requests . . . 236
4.13 Dynamisches Modell des VEE-Frameworks . . . 237
4.14 Darstellung der Daten im HANA-System . . . 238
4.15 Aktuelle Systemumgebung . . . 241
4.16 Betrachtete Technologien . . . 242
4.17 SAP Event Stream Processor . . . 243
4.18 SAP Smart Data Access . . . 244
4.19 Secondary Database Connection . . . 247
4.20 Produktübersicht . . . 251
4.21 SLT mit SAP-Quellsystem . . . 252
4.22 SLT mit SAP-fremden-Quellsystem . . . 252
4.23 Beispieldaten aus der Tabelle „Messdaten“ . . . 262
4.24 VEE-Webinterface . . . 263
4.25 Konfiguration der HTTP-Destination . . . 269
4.26 Anmeldedaten der HTTP-Destination . . . 270
4.27 Ausgabe der Daten der HTTP-Response mit write . . . 273
4.28 Ausgabe der Daten der HTTP-Response in einer GUI . . . 274
4.29 Fehlerkosten 10er-Regel . . . 276
4.30 Konstruktive Qualitätsmanagementmaßnahmen . . . 278
4.31 Wasserfallmodell . . . 279
4.32 Spiralmodell . . . 280
4.33 Prototypen . . . 281
4.34 Funktionale Testfälle . . . 284
F.1 Kompetenzmatrix der TG-CX4U . . . ii
F.2 TG-CX4U Redmine Projektplan . . . iii

Tabellenverzeichnis

2.1 Seminararbeitsthemen . . . 7
3.2 Erklärung der Untersuchungsmerkmale . . . 67
3.3 Die zu untersuchenden Anbieter im Rahmen des Marktüberblicks . . . 69
3.4 Ergebnisse Untersuchungsmerkmale Cloudera Enterprise . . . 72
3.5 Ergebnisse Untersuchungsmerkmale Datameer . . . 75
3.6 Ergebnisse Untersuchungsmerkmale Datawatch Panopticon . . . 78
3.7 Ergebnisse Untersuchungsmerkmale Greenplum Database & HD . . . 81
3.8 Ergebnisse Untersuchungsmerkmale EXASOL EXASolution . . . 85
3.9 Ergebnisse Untersuchungsmerkmale Hortonworks Data Platform . . . 87
3.10 Ergebnisse Untersuchungsmerkmale HP Vertica Analytics Platform . . . 89
3.11 Ergebnisse Untersuchungsmerkmale IBM InfoSphere BigInsights & Streams, Netezza, Cognos BI . . . 94
3.12 Ergebnisse Untersuchungsmerkmale Informatica PowerCenter Big Data Edition 98
3.13 Ergebnisse Untersuchungsmerkmale Jaspersoft Instant Big Data Analytics . . . 101
3.14 Ergebnisse Untersuchungsmerkmale MicroStrategy 9.3 . . . 103
3.15 Ergebnisse Untersuchungsmerkmale Oracle Big Data Appliance . . . 107
3.16 Ergebnisse Untersuchungsmerkmale ParStream . . . 109
3.17 Ergebnisse Untersuchungsmerkmale Pentaho Data Integration & Business Analytics 112
3.18 Ergebnisse Untersuchungsmerkmale QlikTech QlikView . . . 115
3.19 Ergebnisse Untersuchungsmerkmale SAP HANA . . . 119
3.20 Ergebnisse Untersuchungsmerkmale SAP Data Services . . . 120
3.21 Ergebnisse Untersuchungsmerkmale SAS Visual Analytics . . . 123
3.22 Ergebnisse Untersuchungsmerkmale Terracotta BigMemory . . . 127
3.23 Ergebnisse Untersuchungsmerkmale Splunk Enterprise . . . 129
3.24 Ergebnisse Untersuchungsmerkmale Teradata Aster Big Analytics Appliance . . 132
3.25 Aufbau einer Access-Log-Datei . . . 143
3.26 Aufbau der Tabelle für die Nutzerkontextdaten . . . 143
3.27 Dimensionen und Elemente Cube ServerAuslastung . . . 175
3.28 Dimensionen und Elemente Cube ServerAusstattung . . . 175
3.29 Dimensionen und Elemente Cube ServerZugriffe . . . 176
3.30 Dimensionen und Elemente Cube SimulationServerAuslastung . . . 177
3.31 Dimensionen und Elemente Cube SimulationServerAusstattung . . . 177
3.32 Dimensionen und Elemente Cube SimulationsVarianteSzenarioParameter . . . 178
3.33 Dimensionen und Elemente Cube SimulationSzenarioParameter . . . 179
3.34 Dimensionen und Elemente Cube SimulationSzenarioServerAuswahl . . . 179
3.35 Dimensionen und Elemente Cube SimulationVarianteSzenarioResult . . . 180
3.36 Kompatible Web-Browser . . . 187
4.37 Validierung . . . 220
4.38 Estimation . . . 224
4.39 Regelauswahl . . . 227
4.40 Versionierung . . . 230
4.41 Editing . . . 232
4.42 Vergleich der Umsetzungsalternativen 1 . . . 258
4.43 Vergleich der Umsetzungsalternativen 2 . . . 260

Listingverzeichnis

3.1 Verknüpfung zwischen Access Log und Use Case . . . 151
3.2 Verknüpfung zwischen Access Log und aufrufender Anwendung . . . 151
3.3 Verknüpfung zwischen Access Log und Nutzerkontextdaten . . . 151
3.4 SQL-Statement zur Befüllung der Staging Area . . . 152
3.5 View für Zugriffe je Land . . . 156
3.6 View für Zugriffe je Stunde und Use Case . . . 156
3.7 Rules des Cubes ServerAuslastung . . . 175
3.8 Rules des Cubes ServerAusstattung . . . 175
3.9 Rules des Cubes ServerZugriffe . . . 176
3.10 Rules des Cubes SimulationServerAuslastung . . . 177
3.11 Rules des Cubes SimulationServerAusstattung . . . 177
3.12 Rules des Cubes SimulationsVarianteSzenarioParameter . . . 178
3.13 Rules des Cubes SimulationSzenarioParameter . . . 179
3.14 Rules des Cubes SimulationSzenarioServerAuswahl . . . 179
3.15 Rules des Cubes SimulationVarianteSzenarioResult . . . 180
3.16 Prolog des Prozesses p01_simulation_maximum . . . 182
3.17 Epilog des Prozesses p01_simulation_maximum . . . 184
3.18 Die DBR-Funktion . . . 185
4.1 Quellcode einer oData-Schnittstelle . . . 246
4.2 Erhalten der ZählerID von der Benutzeroberfläche . . . 262
4.3 DB Verbindung . . . 263
4.4 Datenvalidierung durch die While-Schleife . . . 263
4.5 Nullwerte Ersatzwertbildung . . . 265
4.6 Fehlende Werte Ersatzwertbildung . . . 265
4.7 Zeitstempel Korrektur . . . 267
4.8 Konvertierung zum UTC-Zeitstempel-Format . . . 267
4.9 Messwert Korrektur . . . 268
4.10 Änderungen Dokumentation . . . 268
4.11 HTTP-Client-Objekt . . . 270
4.12 Deklaration der Variablen des HTTP-Requests . . . 271
4.13 Header-Felder des HTTP-Requests setzen . . . 271
4.14 Versenden und Empfang des HTTP-Requests . . . 271
4.15 Laden der Header-Felder und schließen der Verbindung . . . 272
4.16 Aufbereitung der HTTP-Respone Daten . . . 273
6.1 Quellcode HTTP-Request . . . iv

1 Einleitung

Kapitelübersicht

1.1 Motivation **1**
1.2 Problemstellung und Zielsetzung der Projektgruppe HAWK . **2**
1.3 Aufbau der Arbeit **3**

„Daten sind die Rohstoffe unserer Zeit,
sie müssen nutzbar, smart, und verfügbar gemacht werden."

(Krempl 2013) Hans-Joachim Otto, der parlamentarische Staatssekretär im Wirtschaftsministerium, spricht damit die Notwendigkeit der effizienten Datenwertschöpfung im Kontext von Big Data an (Vgl. Krempl 2013).

Big Data meint den Einsatz großer, unterschiedlich strukturierter Datenmengen aus einer Vielzahl von Quellen in einer hohen Verarbeitungsgeschwindigkeit (Vgl. Bitkom 2013, S.19 ff.). Mit der Thematik Big Data gehen für Unternehmen jeglicher Branche diverse Herausforderungen einher, besonders im Hinblick auf die Datenintegration sowie die Verarbeitung und Auswertung der Datenmengen. Aufgrund der Hochaktualität und der Brisanz hat sich die Projektgruppe HAWK dieser Thematik in Form der zwei Teilprojektthemen „Herausforderungen durch Big Data am Beispiel des IT-Capacity Managements der Volkswagen AG" und „Prozessorientierte Business Intelligence mit SAP HANA" am Beispiel der CX4U AG angenommen.

Fachlich begleitet und betreut wurde die Projektgruppe durch die Fachabteilung Very Large Business Applications (VLBA) der Carl von Ossietzky Universität Oldenburg in dem Zeitraum vom 01. April 2013 - 31. März 2014 unter der Leitung von Prof. Dr. Jorge Marx Gómez und seinen Mitarbeitern M.Sc. Andreas Messler und Dipl.-Wirt.Inform. Andreas Solsbach. Dieser Projektbericht soll den thematischen Hintergrund der Projektgruppe darstellen und die projektbezogenen Rahmenbedingungen beschreiben. Der Hauptbestandteil dieses Berichtes bildet dabei die Vorstellung der erarbeiteten Inhalte der Teilprojektgruppen sowie deren Vorgehensweisen und Ergebnisse.

1.1 Motivation

Unternehmen befinden sich heute in einer verschärften Wettbewerbssituation durch die globalisierten Märkte und zunehmende Transparenz der unternehmensinternen Daten. Dadurch sind Unternehmen gezwungen, vorausschauend zu handeln und flexibel auf Marktveränderungen reagieren zu können. Um dies zu gewährleisten, müssen den Unternehmen entscheidungsrelevante Informationen zur Verfügung stehen. Ein wesentlicher Erfolgsfaktor für Unternehmen ist daher die Aufbereitung dieser Informationen. Eine integrierte und zeitnahe Sicht auf Informationen aus allen Unternehmensbereichen ist wesentlich für das frühzeitige Erkennen und Aufdecken

von Problemen sowie Potenzialen und das dementsprechend notwendige Handeln (Vgl. Engels 2008, S.8). „[Denn] die Verzahnung und Rückkopplung von strategischen und operativen Entscheidungen gewährt eine Durchgängigkeit der Unternehmenssteuerung.“ (Engels 2008, S.8) Business Intelligence (BI) soll mithilfe entsprechender Techniken und Anwendungen durch das Aufbereiten operativer Daten, die durch Geschäftsprozesse im Unternehmen erzeugt werden, den zielgerichteten Zugriff auf entscheidungsrelevante Informationen vereinfachen. Dadurch sollen betriebswirtschaftliche Entscheidungsprozesse verbessert werden. Der Begriff BI setzt sich aus den zwei Themengebieten

1. Business = Grundlegende betriebswirtschaftliche Aufgaben und
2. Intelligence = Methoden zur intelligenten Datenaufbereitung

zusammen und subsumiert daher viele betriebswirtschaftlichen Funktionen. Bestandteile sind hierbei (Vgl. Engels 2008, S.4f.):

- Datenbereitstellung, -haltung und -abfrage und
- Analytische Applikationen.

Demnach trägt effektives BI zu einer erfolgreichen Unternehmenssteuerung bei. Die VLBA-Abteilung der CvO-Universität Oldenburg beschäftigt sich ebenfalls mit dieser Thematik. Unter VLBA versteht man unternehmensweite betriebliche und zwischenbetriebliche Informationssysteme, die die Ausführung der Geschäftsprozesse entlang der Wertschöpfungskette unterstützen, z.B. ERP-Systeme (Vgl. VLBA 2012a). Da sich die Abteilung ebenfalls mit aktuellen Problemstellungen in diesem Zusammenhang auseinandersetzt, sind im Kontext der Thematik Big Data die zwei besagten Projektgruppenthemen entstanden, dessen sich die Projektgruppe HAWK annahm.

1.2 Problemstellung und Zielsetzung der Projektgruppe HAWK

Das Themengebiet Prozessorientierte Business Intelligence ist breit gefächert. Daher ist es nicht verwunderlich, dass aus diesem zwei Themen abgeleitet wurden. Die Projektgruppe HAWK befasst sich deshalb mit den zwei aktuellen Themengebieten:

- Prozessorientierte Business Intelligence mit SAP HANA
- Herausforderungen durch Big Data am Beispiel des IT-Capacity Managements der Volkswagen AG

Bei beiden Themen geht es um die Verarbeitung von großen Datenmengen und die Möglichkeit der Integration dieser in vorhandene IT-Systemlandschaften. Die Ausprägungen der zwei Themen sind entstanden, da die Projektgruppe mit zwei Praxispartnern aus der Wirtschaft zusammengearbeitet hat und sich diesbezüglich verschiedene Aufgabengebiete ergaben. Aus diesem Grund entstanden zwei Teilgruppen mit je sechs Gruppenteilnehmern.

Die Teilgruppe „Big Data im IT-Capacity Management“ arbeitete mit dem Automobilhersteller Volkswagen AG zusammen. Fachlich unterstützt wurde die Gruppe durch das BI-CC des Volkswagen Konzerns. Das Thema der Zusammenarbeit beinhaltete „Herausforderungen durch Big Data am Beispiel des IT-Capacity Managements der Volkswagen AG“. Die hierbei entstandenen Ergebnisse sowie Informationen zur Teilprojektgruppe und dem Projektverlauf lassen sich dem Kapitel 3 dieses Berichts entnehmen.

Die Teilgruppe „Smart Metering“ kooperierte mit dem Dienstleistungsunternehmen CX4U AG, welches seinen Sitz in Oldenburg hat. Hierbei wurde das Themengebiet „Prozessorientierte Business Intelligence mit SAP HANA“ unternehmensbezogen bearbeitet. Die dabei entstandenen Ergebnisse sowie Informationen zur Teilprojektgruppe und dem Projektverlauf lassen sich dem Kapitel 4 dieses Berichts entnehmen.

1.3 Aufbau der Arbeit

In Anschluss an die Einleitung in die Problematik und den Aufbau dieser Arbeit sollen sowohl die organisatorischen als auch technischen Rahmenbedingungen der Projektarbeit dargestellt werden (Kap. 2). Dabei wird das Konzept und Ziel dieser Projektgruppe vorgestellt. Hierbei soll auch verdeutlicht werden, welche Rolle die Projektgruppenarbeit im Masterstudium Wirtschaftsinformatik der CvO-Universität einnimmt. Ebenso werden die einzelnen Projektteilnehmer genannt sowie die ihnen zugeteilten Rollen und Verantwortlichkeiten innerhalb der Projektgruppe aufgezeigt. Des Weiteren wird ein Einblick in die Vorbereitungen für die Bearbeitung der Projektthemen gegeben und in diesem Zusammenhang die dabei behandelten Seminararbeitsthemen aufgeführt. Im Anschluss daran wird zunächst die Teilgruppe „Big Data im IT-Capacity Management“ vorgestellt (Kap. 3). Darauf folgt die Vorstellung der Teilgruppe „Smart Metering“ (Kap. 4). Beide Kapitel werden mit allgemeinen Informationen zu der Teilprojektgruppe eingeleitet. Dabei wird sowohl auf die Teilnehmer als auch auf die organisatorischen Hintergründe und Abläufe der Teilgruppe eingegangen. Weiter werden die während der Projektarbeit zu bearbeitenden Aufgaben und ihre Inhalte sowie Ergebnisse vorgestellt. Abschließen wird der Projektbericht der Projektgruppe HAWK mit einer Schlussbetrachtung (Kap. 5).

2 Rahmenbedingungen

Kapitelübersicht

2.1 Organisatorische Rahmenbedingungen 4
2.1.1 Konzept und Ziel einer Projektgruppe 4
2.1.2 Projektgruppenmitglieder und ihre Rollen 5
2.1.3 Vorbereitung und Seminararbeitsthemen 6
2.1.4 Projektverlauf 8
2.2 Technische Rahmenbedingungen 10
2.2.1 Confluence als interne Kommunikationsplattform 10
2.2.2 Dokumentation mittels LaTeX und Jabref 11
2.2.3 Homepagearbeit 14
2.2.4 Nutzung von Redmine für die Projektgruppenorganisation 14
2.2.5 Datensicherheit durch SVN 14

Jedes erfolgreiche Projekt bedarf einer gründlichen Planung. Diese wird im Verlauf des Projektes meist angepasst, z.B. durch Veränderungen der Gegebenheiten oder sogar des Zieles der Arbeit. Dennoch ist die Planung allein nicht der Schlüssel für eine erfolgreiche Projektarbeit. Wichtig ist, die entsprechenden Rahmenbedingungen für eine erfolgreiche Projektumsetzung zu schaffen. Die Rahmenbedingungen eines Projektes ergeben sich vor allem aus den Kompetenzen und Kenntnissen der beteiligten Gruppenmitglieder sowie der technischen Voraussetzungen. In diesem Kapitel werden daher die organisatorischen und technischen Rahmenbedingungen der Projektgruppe HAWK betrachtet.

2.1 Organisatorische Rahmenbedingungen

In diesem Kapitel soll zunächst ein Einblick in die Inhalte der Projektarbeit innerhalb des Masterstudienganges an der Carl von Ossietzky Universität Oldenburg gegeben werden. Weiter sollen die Mitglieder der Projektgruppe HAWK vorgestellt werden. Ebenso soll dem Leser dargestellt werden, wie sich die Teilnehmer mit den thematischen Hintergründen der einzelnen Aufgaben auseinandergesetzt haben und welche Vorbereitungen für die Projektarbeit getroffen wurden. Abschließend wird der Projektverlauf aus der Sicht der gesamten Projektgruppe dargestellt.

2.1.1 Konzept und Ziel einer Projektgruppe

Der Masterstudiengang der Wirtschaftsinformatik an der Carl von Ossietzky Universität Oldenburg wird i.d.R. innerhalb von 4 Semestern abgeschlossen, davon sind allein 2 Semester (inklusive der vorlesungsfreien Zeit) für die Mitarbeit an einer Projektgruppe vorgesehen, die sich aus 6-12 Masterstudierenden zusammensetzt. Dies zeigt, dass innerhalb des Masterstudiums großer

Wert auf die Anwendung der im Laufe des Studiums angeeigneten theoretischen Kenntnisse in die Praxis gelegt wird. Die Projektarbeit beinhaltet nicht nur die Durchführung eines Projektes anhand einer vorgegebenen Fragestellung von der Problemanalyse, über die Erarbeitung der gegebenen Thematik bis zur endgültigen Realisierung, sondern auch die Förderung von Softskills sowie das Erlernen von berufstypischen Arbeitsweisen. Innerhalb der Projektgruppe soll somit vieles vermittelt werden, was sich theoretisch kaum lehren lässt. Besonderen Wert wird in diesem Kontext u.a. auf folgende Aspekte der Projektgruppe gelegt (Vgl. Boles 2009, S.1; VLBA 2012b):

- intensive Teamarbeit
- Sammeln von Erfahrungen innerhalb der Teamarbeit
- Notwendigkeit zu längerfristigem Projekt- und Zeitmanagement
- umfassende Auseinandersetzung mit einer Thematik bzw. Aufgabenstellung
- starke persönliche Bindung zu den Lehrenden

Trotz eigenständiger Organisation und Arbeit, wird jede Projektgruppe von Lehrenden der CvO-Universität begleitet. Die Projektgruppe HAWK der Fachabteilung Very Large Business Applications (VLBA) wurde unter der Leitung von Prof. Dr. Jorge Marx Gómez von seinen Mitarbeitern M.Sc. Andreas Messler und Dipl.-Wirt.Inform. Andreas Solsbach im Zeitraum 01. April 2013 bis 31. März 2014 betreut.

2.1.2 Projektgruppenmitglieder und ihre Rollen

Die Projektgruppe HAWK besteht aus 12 Masterstudenten der CvO-Universität. Das Akronym HAWK steht dabei für:

High **A**nalytics to **W**in **K**nowledge

Der Name wurde von den Studenten selbst entwickelt und findet sich auch symbolisch im Logo der Projektgruppe wieder. Im Rahmen der Projektgruppe werden Rollen und Verantwortlichkeiten verteilt. Für diese soll das zuständige Teammitglied während der gesamten Projektlaufzeit die Verantwortung tragen. Dabei können sich Rollen und Verantwortlichkeiten auch erst im Laufe der Projektarbeit entwickeln, da am Anfang eines Projektes viele Begebenheiten und Aufgaben noch nicht definiert sind. Ebenso ist die Anzahl der Projektgruppenmitglieder nicht gleich der Rollen bzw. Verantwortlichkeiten. Deshalb übt nicht jedes Teammitglied eine bestimmte Rolle aus. Die nachfolgende Abbildung 2.1 soll die Projektgruppenmitglieder und ihre zugewiesenen Aufgabenbereiche vorstellen.

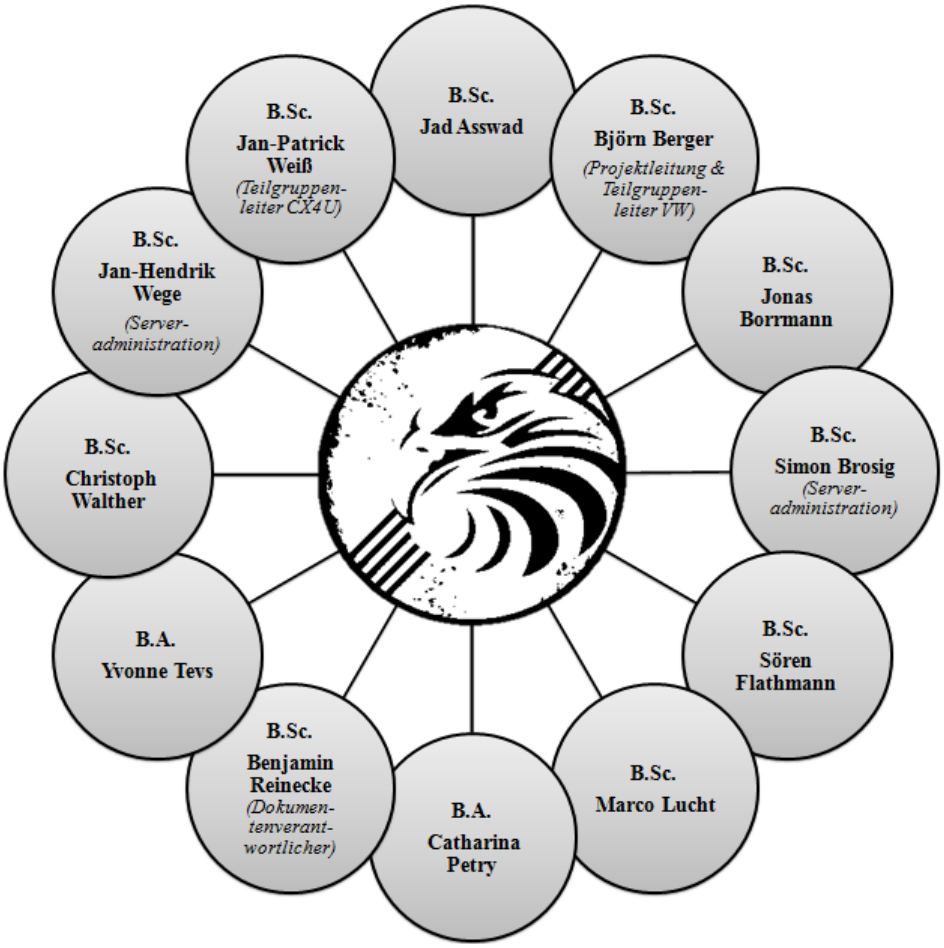


Abbildung 2.1: Projektgruppenmitglieder

Durch die Einteilung in Teilgruppen sind jedoch auch Rollen definiert worden, die lediglich für die Teilgruppe gelten. Die einzelnen Organisationsstrukturen der Teilgruppen werden in den Kapiteln 3.2 und 4.2 dargestellt.

2.1.3 Vorbereitung und Seminararbeitsthemen

Zu Beginn der Projektarbeitsphase wurde die erste Zeit für grundlegende Tätigkeiten verwendet, die die zukünftige Projektarbeit unterstützen sollten. Dazu zählt zum einen die Kontaktaufnahme mit den Praxispartnern. Zum anderen gilt es vor jeder Projektarbeit festzustellen, welche Kompetenzen das Team bereits aufweisen kann und welche Fachkenntnisse es noch zu vermitteln gilt. Während demnach auf der einen Seite der Kontakt zu den Praxispartnern hergestellt und die für die Projektarbeit relevanten Aufgabenstellungen beider Teilgruppen definiert wurden, wurden auf der anderen Seite parallel auch die Projektmitglieder gleichermaßen geschult. In diesem Zusammenhang absolvierte jedes Projektmitglied einen Online-Kurs des openHPI[1], der

[1] Offene und interaktive Bildungsplattform des Hasso-Plattner-Institutes.

sich mit der zum Zeitpunkt des Projektbeginns hochaktuellen Thematik SAP HANA beschäftigt. Dieser Online-Kurs (Introduction to Software Development on SAP HANA) wurde binnen sechs Wochen abgehalten und beinhaltete einen durchschnittlichen Wochen-Workload von ca. 4 Stunden sowie wöchentliche Lernkontrolltests. Dabei wurden folgende Inhalte behandelt:

- Developing Applications for SAP HANA
- Database Tasks, Loading, and Modeling
- SQLScript Basics
- Exposing and Consuming Data with OData
- Exposing and Consuming Data with Server-Side JavaScript
- SAP HANA Advanced Development Options

Abgeschlossen wurde der Kurs mit einer umfassenden Online-Prüfung zum Ende der Kurszeit. Um sich weiter in für die Projektarbeit relevante Themengebiete einzuarbeiten, wurden von den Projektgruppenmitgliedern zusätzlich Seminararbeiten verfasst. Die Bearbeitungszeit betrug hierfür rund einen Monat und endete ebenfalls im Juli 2013. Abgeschlossen wurde diese Arbeitsphase mit einer Abschlusspräsentation. In der nachfolgenden Tabelle sollen die bearbeiteten Themen vorgestellt werden.

Verfasser	**Thema der Seminararbeit**
Jonas Borrmann Christoph Walther	Data Mining Techniken im In-Memory Computing
Yvonne Tevs Catharina Petry	Datenanalyse in Zeiten von Big Data: Darstellung der Chancen und Herausforderungen für Unternehmen
Björn Berger Sören Flathmann	BI in SAP HANA und Apache Hadoop
Simon Brosig Jad Asswad	SAP HANA und Apache Hadoop Cluster
Benjamin Reinecke Jan-Patrick Weiß	Auswertungen und Möglichkeiten des Einsatzes von R und SQLScript
Marco Lucht Jan-Hendrik Wege	Technologievergleich zur Visualisierung mittels HTML5/CSS3/Javascript/Business Objects 4.0/Rest Services

Tabelle 2.1: Seminararbeitsthemen

Die verfassten Seminararbeiten lassen sich entsprechend des Verweises der Spalte „Anhang Nr.“ dem Anhang dieses Projektberichtes entnehmen. Nachdem die Arbeitsphase der Vorbereitung

und Ausarbeitung der Seminararbeitsthemen abgeschlossen wurde, konnte ab August 2013 mit den zu bearbeitenden Inhalten bezüglich der Aufgabenstellungen der einzelnen Teilgruppen begonnen werden.

2.1.4 Projektverlauf

Die Projektlaufzeit begann am 04. April 2013 mit der Begrüßung durch Professor Marx Gómez. Der Projektgruppe stellte er das Thema „Prozessorientierte Business Intelligence mit SAP HANA“ und die herausragende Bedeutung der neuen Datenbank-Appliance SAP HANA für den IT-Markt vor. Die Betreuer Andreas Messler und Andreas Solsbach erklärten die grundsätzliche Organisation der Projektgruppe. Im zweiten Treffen der Projektgruppe in der darauffolgenden Woche am 11. April 2013 machte Professor Marx Gómez die Projektgruppenmitglieder mit einem der zwei Praxispartner, der CX4U AG bekannt. Dazu stellten Frank Locher und Okko Nannen von der CX4U AG das Unternehmen und die Aufgabenstellung für die Projektgruppe vor.

In den folgenden zwei Wochen wurden weitere vorbereitende Aufgaben zur Durchführung des Projektes erledigt. Dazu gehörte die Einrichtung des Projektgruppenraums, die Beauftragung und Konfiguration der Server für die Systeme SVN und Redmine sowie des E-Mail-Verteilers und die Einrichtung des SAP HANA Studios und der Entwicklungsumgebung. Im Anschluss an jedes Projektgruppentreffen, die in der Regel Donnerstags von 14:00 Uhr bis 15:00 Uhr durchgeführt wurden, fand ein internes Treffen der Projektgruppe statt um weitere Punkte zu klären oder die Bearbeitung der Aufgaben weiterzuführen.

Im Rahmen des Projektgruppentreffens vom 16. Mai 2013 informierte Professor Marx Gómez die Projektgruppe über den Fortschritt der Verhandlungen mit dem zweiten Praxispartner der Volkswagen AG. Themenschwerpunkt sollte Big Data aus Log-Dateien aus IT-Systemen sein. Es war erforderlich, die Projektgruppe in zwei Teilgruppen zu teilen, die sich jeweils mit der Aufgabenstellung eines Praxispartners befassen. Dem kam die Projektgruppe in diesem Treffen durch einen Losentscheid nach. Die Aufteilung wurde durch einen Abgleich der zuvor von jedem Projektgruppenmitglied ausgefüllten Kompentenzmatrix gegen die im Projekt notwendigen Fähigkeiten überprüft. Die Betreuung der Teilprojektgruppe mit dem Praxispartner CX4U (kurz TG CX4U, im späteren Verlauf dann „Projektgruppe Smart Metering“) übernahm Andreas Solsbach und die Betreuung der Teilprojektgruppe mit dem Praxispartner Volkswagen AG (kurz TG VW, im späteren Verlauf dann „Projektgruppe Big Data im IT-Capacity Management“) übernahm Andreas Messler.

Mit Beginn der folgenden Woche vom 20. Mai bis zum 14. Juli 2013 erstellten die Projektgruppenmitglieder in 2er-Teams Seminararbeiten zur Vorbereitung auf kommende Inhalte. Parallel dazu startete am 27. Mai ein SAP HANA-Online-Kurs, der die Grundlagen zur Entwicklung von Anwendungen mit SAP HANA lehren sollte.

Erste Informationen zum Projektstart erhielt die TG VW im Projektgruppentreffen am 30. Mai

2013. In einem KickOff-Meeting am 04. Juni 2013 sollen die Mitglieder der TG VW die exakte Aufgabenstellung und Anforderungen erfahren. Zu diesem Zeitpunkt einigte sich die gesamte Projektgruppe zum Einsatz von Confluence des Herstellers Atlassian als Wiki zur internen Kommunikation und zum Wissensaustausch.

Im Projektgruppentreffen am 06. Juni 2013 nach dem Kick-Off in Wolfsburg mit der Volkswagen AG informierte die TG VW die gesamte Projektgruppe und die Betreuer über die neuen Informationen. Es stellte sich heraus, dass das Kernsystem SAP HANA aus der ursprünglichen Aufgabenstellung der Projektgruppe in diesem Projekt für die Volkswagen AG nicht relevant ist. Dies bestätigte sich auch in einem weiteren Telefonat von Professor Marx Gómez mit der Auftraggeberin und Teamleiterin des BI-CC der Volkswagen AG. Aus diesem Grund wurde zwischenzeitlich eine Ersatzaufgabe für den Einsatz von SAP HANA eingeplant, die schlussendlich wieder verworfen wurde, da die Erfüllung der Aufgaben im Projekt der TG VW die gesamte Projektlaufzeit in Anspruch nahm.

Der Startschuss im Projekt mit der CX4U AG fiel für die Teilprojektgruppe CX4U am 13. Juni in einem Meeting in der Universität. Die Auftraggeber der CX4U AG Frank Locher und Okko Nannen stellten die Aufgabenstellung vor und der weitere organisatorische Ablauf wurde geklärt. Weitere Informationen aus der Wirtschaft konnten die Projektgruppenmitglieder Jan-Patrick Weiß und Benjamin Reinecke beim Info Day bei der abat AG in Bremen am 20. Juni 2013 sammeln. Die Erkenntnisse des Tages halfen bei der Erfüllung der Aufgaben.

Das weitere Verfahren in den Teilprojektgruppen lief fortan getrennt voneinander ab. Die Ergebnisse und der Projektverlauf werden in den Kapiteln 3 und 4 beschrieben. Dennoch gab es weitere Berührungspunkte. Die grundsätzliche Organisation der Teilprojektgruppen erfolgte seitens der Betreuer in der Regel für beide gleichzeitig. So gab es stets ein wöchentliches Projektgruppentreffen für beide Teilgruppen. Treffen außerhalb der Projektarbeit zum Grillen und Pokern wurden ebenfalls in der gesamten Projektgruppe durchgeführt. Ein erstes Feedback zur Bewertung der Seminararbeiten und zur generellen Einschätzung der Projektgruppenmitglieder erfolgte ab dem 12. September. Gegenstand waren auch eine Selbsteinschätzung sowie Hinweise zur Verbesserung der Leistungen seitens der Betreuer. Der Projektverlauf wurde von zwei Urlaubsphasen unterbrochen. Die erste war eine Woche Sommerurlaub vom 12. bis 18. August 2013. Die zweite Unterbrechung ergab sich durch zweiwöchige Weihnachtsferien bis zum 05. Januar 2014. Im März 2014 stellten die Mitglieder beider Teilprojektgruppen die Ergebnisse der TG CX4U auf der CeBIT in Hannover auf dem Messestand des Landes Niedersachsen vor. Auch sind in den Projektplänen beider Teilgruppen Phasen zur Erstellung der Dokumentation zu finden, die in diesem gemeinsamen Abschlussbericht münden.

2.2 Technische Rahmenbedingungen

Für eine erfolgreiche und reibungslose Teamarbeit innerhalb der Projektgruppe war es erforderlich, die notwendigen technischen Rahmenbedingungen zu schaffen. Die Komplexität der Arbeit untereinander ergab sich vor allem in der Tatsache, dass die Gruppenmitglieder z.B. unterschiedliche Betriebssysteme verwenden, dass bei einer Anzahl von 12 Mitgliedern die Arbeit an einem Dokument, wie z.B. dem Endbericht, einem hohen Koordinationsaufwand bedarf sowie weitere spezielle Gegebenheiten, die beachtet werden mussten. Hierbei musste vor allem im Bezug auf die Zusammenarbeit mit den Praxispartnern die Datensicherheit gewährleistet sein, um die unternehmensinternen Informationen der Praxispartner vor Dritten zu schützen. Hinzu kommt, dass eine geeignete Plattform für die Projektkommunikation geschaffen werden musste. Dabei ging es nicht nur um die Verteilung von Aufgaben, sondern auch um die Nachvollziehbarkeit des Projektverlaufes. Weiter mussten verschiedene E-Mail-Verteiler aufgrund der unterschiedlichen Interessengruppen von E-Mail-Inhalten erstellt werden.

Daher werden in diesem Kapitel die genutzten technischen Instrumente und Mittel vorgestellt, die einen reibungslosen Verlauf hinsichtlich der Kommunikation sowie des Datenaustausches während der Projektarbeit gewährleisten sollten. Die technischen Rahmenbedingungen setzen sich dabei wie folgt zusammen:

- Kommunikation innerhalb der Gruppe mittels E-Mail-Verteiler
- Kommunikation nach außen durch eine Projektgruppen-Homepage
- Gewährleistung eines sicheren Datenaustausches innerhalb der Gruppe mittels verschlüsselter SVN-Verbindung
- Nachvollziehbarkeit des Projektverlaufs sowie Aufgabenverteilung an die Gruppenmitglieder mit Hilfe von Redmine und Confluence
- Schaffen von einheitlichen LaTeX-Dokumentenvorlagen sowie deren Pflege

Auf die Umsetzung der technischen Rahmenbedingungen soll nachfolgend weiter eingegangen werden. Dazu werden die verwendeten Instrumente sowie ihr Nutzen für die Projektgruppe vorgestellt. Ebenfalls soll hier an den entsprechenden Stellen auf die Besonderheit der Wahrung von Betriebsgeheimnissen der Praxispartner eingegangen werden. Interessant in diesem Zusammenhang ist es, welche Instrumente im Bezug auf die Datensicherheit zu nennen sind und mit welcher Begründung eben diese für die Verwendung innerhalb der Projektgruppe ausgewählt wurden.

2.2.1 Confluence als interne Kommunikationsplattform

Damit die Projektgruppenmitglieder gemeinsam arbeiten, Vorschläge sammeln und Anleitungen notieren können, wurde eine Plattform für eben diese Aufgaben benötigt. Die Entscheidung ist hierbei auf Confluence gefallen, da es sich um ein sehr professionelles, auf Unternehmen

ausgelegtes Produkt handelt, welches u.a. auch regelmäßig mit Updates versorgt wird. Hier lassen sich Bereiche erstellen, in denen dann bestimmte Benutzer oder alle Benutzer arbeiten dürfen. Es wurden z.B. für die jeweiligen Teilgruppen Bereiche erstellt, zu denen nur deren Mitglieder Zugang haben. Es existieren auch Bereiche, um innerhalb der Projektgruppe Organisatorisches zu klären oder Bereiche, in denen den Mitgliedern Anleitungen zur Verfügung stehen. Ein weiteres positiv hervorzuhebendes Feature von Confluence ist der integrierte Kalender, mittels dessen sich Termine auf einfachste Weise verwalten lassen. Die Termine lassen sich hier auch nochmal in die einzelnen Teilgruppen aufschlüsseln. Confluence läuft auf dem PG-Server mittels eines Tomcat-Servers. Confluence unterstützt zudem eine E-Mail-Anbindung, mittels welcher Nutzer z.B. über Änderungen in einem ihrer Beiträge informiert werden können.

2.2.2 Dokumentation mittels LaTeX und Jabref

Zur Anfertigung des Projektgruppenberichts, der Seminararbeiten, der Protokolle sowie der sonstigen Dokumente wurde LaTeX verwendet. Anders als in anderen Textverarbeitungsprogrammen, wie beispielsweise in Microsoft Word, OpenOffice Writer oder dem LibreOffice Writer, ist in LaTeX das Aussehen des Dokuments vom eigentlichen Inhalt getrennt. Sind diese beiden Bereiche verknüpft spricht man im Allgemeinen von „What you see is what you get“ (WYSIWYG). Beim Arbeiten mit LaTeX hat man die grafische Aufbereitung des Textes nicht permanent vor Augen - und diese steht auch nicht im Mittelpunkt des Schaffens. Arbeitet man mit LaTeX, schreibt man reinen Text und ergänzt diesen mit einem so genannten Markup. Hierbei handelt es sich um Befehle die das spätere Layout beeinflussen. Beim Gestalten und Setzen (technisch: kompilieren) des Dokuments werden dann Markup und Text interpretiert und zu einem druckbaren Dokument im PDF-Format zusammengefügt (Vgl. Schlosser 2009, S. 17).

Nachteilig an diesem Vorgehen ist zum einen, dass die Layout-Befehle bekannt sein müssen und zum anderen, dass das Arbeiten gerade mit Grafiken und Tabellen schwierig sein kann, da diese nicht sofort angezeigt werden. Allerdings kompensieren die folgenden Vorteile diese negativen Aspekte, weswegen sich die Projektgruppe für die Arbeit mit LaTeX entschieden hat (Vgl. Schlosser 2009, S. 18 ff.):

- Das Layout einer wissenschaftlichen Arbeit soll in erster Linie den Zweck erfüllen, angenehm lesbar zu sein. Bei WYSIWYG-Textverarbeitungsprogrammen besteht zu jeder Zeit die Gefahr, dass sich der Autor zu sehr auf die Form als auf den eigentlichen Text konzentriert.

- Die meisten Textverarbeitungsprogramme speichern ihre Dokumente in so genannten Binärdateien. Diese können in der Regel ausschließlich von der jeweiligen Software gelesen und gespeichert werden. Zwar kann beispielsweise das freie LibreOffice Dokumente aus Microsoft Word importieren, das Ergebnis lässt jedoch oft zu wünschen übrig. LaTeX-Dokumente werden als einfache Textdateien gespeichert. So können diese unter den

verschiedensten Betriebssystemen und Versionen geöffnet und bearbeitet werden.

- Es ist sehr einfach, LaTeX-Dokumente aufzuteilen und in verschiedenen Dateien und Ordnern zu strukturieren. Dies ermöglicht im Zusammenspiel mit dem eingesetzten SVN die sinnvolle Zusammenarbeit an einem Dokument mit mehreren Personen.

Um mit LaTeX arbeiten zu können, wird eine LaTeX-Distribution und ein Editor benötigt. Während die Distribution dafür zuständig ist, aus Eingaben eine Ausgabe zu generieren, werden die Eingaben in dem Editor getätigt. Die Distribution besteht aus Programmen, die zum Erstellen von PDF-Dateien nötig sind, sowie aus Paketen für den Zeichensatz und das Layout. Im Rahmen der Projektgruppe wurden die Distributionen MiKTeX[2] unter Microsoft Windows und MacTeX[3] unter Apple Macintosh Betriebssystemen eingesetzt. Als Editor kam in beiden Fällen Texmaker[4] zum Einsatz (Vgl. Abb. 2.2). Durch das Verwenden nur eines Editors unter beiden Betriebssystemen konnten Benutzerfragen, die durch eine nicht konsistente Bedienung ausgelöst werden könnten, weitestgehend ausgeschlossen werden. Weitere Gründe für den Einsatz von Texmaker sind das Vorhandensein einer Syntax-Highlight-Funktion sowie der Schnellzugriff auf oft benötigte Layout-Befehle über die Menüleiste.

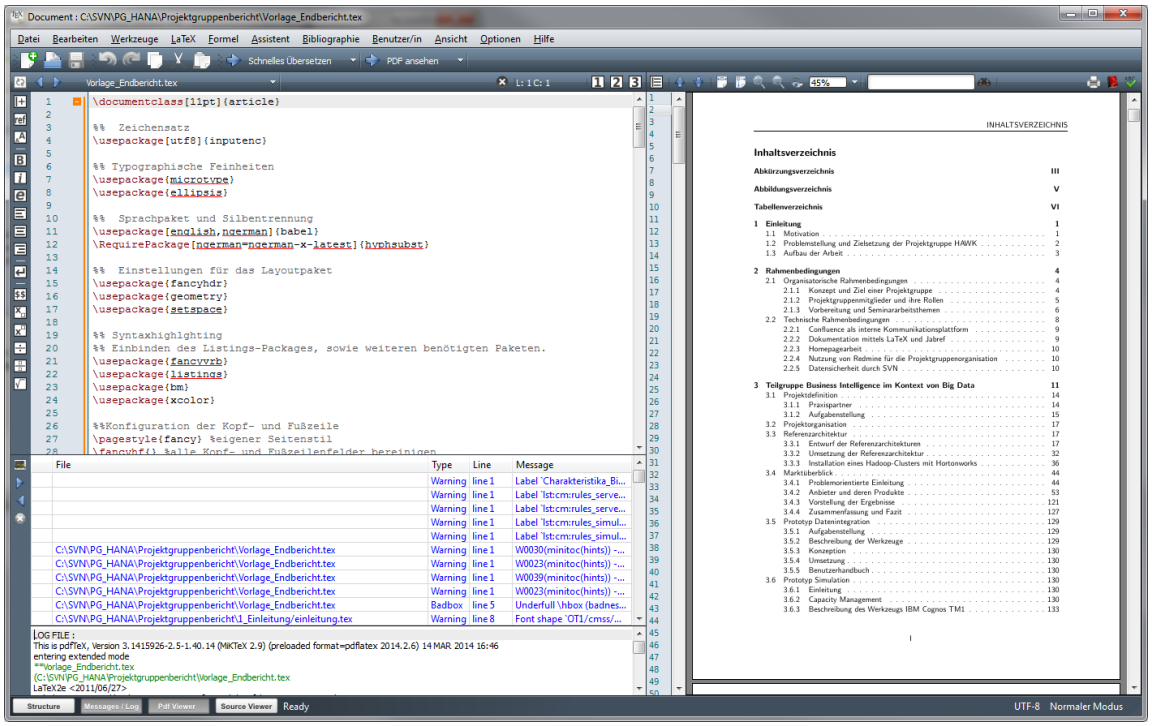


Abbildung 2.2: Screenshot Texmaker

Um in LaTeX Literatur verwalten und einbinden zu können, existiert die BiBTeX-Lösung. Die gesamte Literatur, unabhängig ob es sich hierbei um Bücher, Magazine oder Internetquellen

[2] http://miktex.org/download
[3] http://www.tug.org/mactex/index.html
[4] http://www.xm1math.net/texmaker/download.html

handelt, wird in einer Datei erfasst. In der Arbeit wird nun nur auf diese Datei, bzw. auf die Quelle in der Datei, verwiesen. Die Art und Weise wie das Zitat dann im fertigen Dokument und im Literaturverzeichnis dargestellt wird, wird einmal im Dokument konfiguriert (Vgl. Schlosser 2009, S. 169). Zum Erstellen und Verwalten dieser einen Literaturdatei wurde in der Projektgruppe Jabref[5] eingesetzt. Diese Software basiert auf Java, so dass auch hier eine Verwendung unabhängig des eingesetzten Betriebssystems möglich ist. Die grafische Oberfläche von Jabref erleichtert die Arbeit mit der Bibliotheksdatei, bei der es sich ebenfalls um eine einfache Textdatei handelt (Vgl. Abb. 2.3).

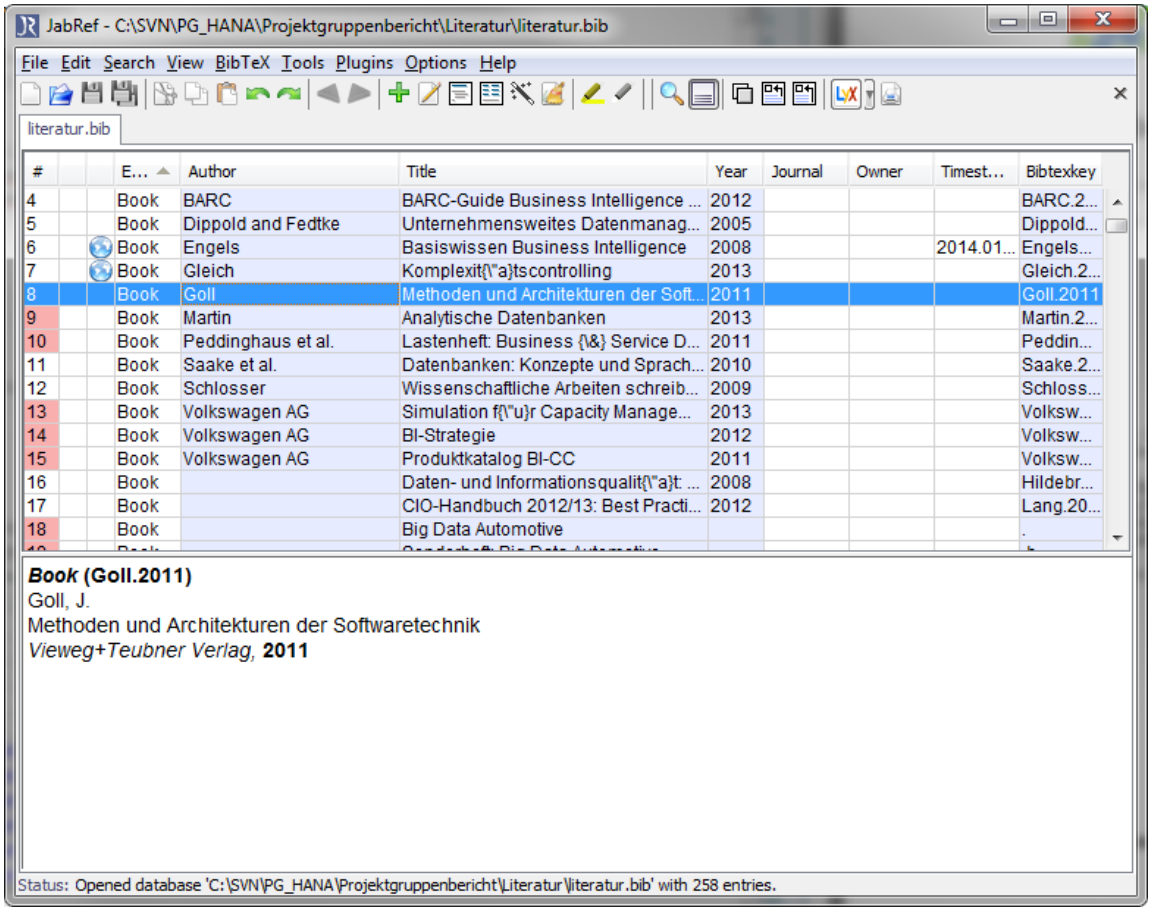


Abbildung 2.3: Screenshot Jabref

Da nur wenige Mitglieder der Projektgruppe zu Beginn der Projektzeit über LaTeX-Kenntnisse verfügten, wurde vom Dokumentenbeauftragten ein LaTeX-Wiki im Confluence gepflegt (Vgl. Anhang F.8). In diesem wurden neben Hinweisen zu der Arbeit mit Jabref und Texmaker auch Code-Konventionen gepflegt und die üblichsten Fehlerursachen dargestellt. Auch Links zu den Downloads der eingesetzten Software sind im Wiki zu finden.

[5] `http://www.xm1math.net/texmaker/download.html`

2.2.3 Homepagearbeit

Um Informationen zu der Projektgruppe HAWK auch außenstehenden Personen zugänglich machen zu können, wurde eine Projektgruppen-Website erstellt. Diese ist unter der Adresse `http://hawk.informatik.uni-oldenburg.de` erreichbar und stellt in erster Linie die Ziele, Vorgehensweisen und die untersuchten Technologien der beiden Teilprojektgruppen vor. Um eine gemeinschaftliche Erstellung, Bearbeitung und Organisation der Inhalte gewährleisten zu können, wurde die Website mithilfe des Content-Management-Systems WordPress entwickelt. Das Hosting erfolgt über einen von der Uni Oldenburg bereitgestellten Server, über den auch die anderen Dienste (SVN, Redmine und Confluence) für die Projektgruppe bereitgestellt werden.

2.2.4 Nutzung von Redmine für die Projektgruppenorganisation

Zur besseren Planung und Steuerung des Projekts, wurde Redmine eingeführt. Hierbei handelt es sich um ein webbasiertes und quelloffenes Projektmanagement-Tool. Nach erfolgreicher Installation und Konfiguration auf dem PG-eigenen Server, wurde ein neues Projekt mit zwei Unterprojekten in Redmine angelegt. Durch diese Struktur konnte sowohl eine Teilprojektgruppen-spezifische als auch Teilprojektgruppen-übergreifende Planung durchgeführt werden. Anschließend wurden die einzelnen Phasen der Projekte (z.B. Anforderungsanalyse, Konzeption, Realisierung) über sog. Tracker abgebildet. Des Weiteren wurde die Konvention vereinbart, dass jede Projekttätigkeit entweder als Aufgabe oder als Arbeitspaket über die entsprechende Option in Redmine zu kategorisieren ist. Nachdem schließlich im letzten Schritt für jedes Projektgruppenmitglied ein Redmine-User mit entsprechenden Rechten angelegt wurde, konnten konkrete Aufgaben durch die Projektleitung definiert und einem Bearbeiter zugeordnet werden. Anschließend war es möglich, die Zeit- und Ressourcenplanung in einem Gantt-Diagramm auszuwerten.

2.2.5 Datensicherheit durch SVN

Damit die anfallenden Daten der PG-Mitglieder mit anderen Mitgliedern ausgetauscht und im gleichen Moment versioniert werden können, wurde auf die Versionierungs-Software SVN (Subversion) zurückgegriffen. Gründe hierfür waren u.a. die Tatsache, dass SVN bereits in früheren Projekten genutzt wurde und sich hier bewährt hat. SVN kann mit großen Datenmengen und vielen Revisionen umgehen. Nach der Installation von SVN auf dem Server musste sichergestellt werden, dass die Daten verschlüsselt übertragen werden. Dies geschah, indem der SVN-Client des Users nicht wie gewöhnlich die Verbindung über den SVN-Port des PG-Servers aufgenommen hat, sondern eine sichere SSH-Verbindung nutzt, über welche die Verbindung zum SVN-Dienst aufgenommen wurde. Es wurden zwei Ordner angelegt, auf welche per SVN zugegriffen werden konnte. In dem Ordner „pg“ finden sich alle Daten der PG-Teilnehmer wieder und im Ordner „cx4u“ die Daten, auf welche die externe Kooperations-Firma CX4U zugreifen muss. Sehr wichtig bei SVN ist, dass eine Bearbeitung einer Datei durch mehrere Nutzer vermieden werden sollte.

3 Teilgruppe Big Data im IT-Capacity Management

Kapitelübersicht

3.1 Projektdefinition . . . 18

3.1.1 Praxispartner . . . 18

3.1.2 Aufgabenstellung . . . 19

3.1.2.1 Auftragsdefinition . . . 20

3.1.2.2 Interpretation und Anpassungen . . . 21

3.1.3 IT-Capacity Management . . . 21

3.2 Projektorganisation . . . 24

3.2.1 Organisationsstruktur . . . 24

3.2.2 Projektplan . . . 25

3.2.3 Projektverlauf . . . 26

3.3 Referenzarchitekturen . . . 28

3.3.1 Einleitung . . . 28

3.3.2 Szenario A . . . 30

3.3.2.1 Architektur . . . 30

3.3.2.2 Anpassungen zur Verarbeitung von Big Data . . . 31

3.3.3 Szenario B . . . 32

3.3.3.1 Architektur . . . 32

3.3.3.2 Datenbankintegration . . . 33

3.3.3.3 Integration via ETL-Tool . . . 34

3.3.3.4 Bereitstellung der Daten . . . 36

3.3.3.5 Procedures . . . 37

3.3.4 Szenario C . . . 40

3.3.4.1 Architektur . . . 40

3.3.4.2 Prozessübersicht . . . 41

3.3.5 Besondere Herausforderungen . . . 42

3.3.5.1 Variety . . . 42

3.3.5.2 Velocity . . . 45

3.3.5.2.1 Datengenerierung und -übertragung . . . 45

3.3.5.2.2 Datenverarbeitung . . . 46

3.3.6 Beispielhafte Umsetzung einer Referenzarchitektur . . . 47

3.3.6.1 Zielsetzung . . . 47

3.3.6.2 Vorgehen . . . 47

3.3.6.3 Reporting . . . 49

3.3.7 Installation eines Hadoop-Clusters mit Hortonworks . . . 51

3.3.7.1 Motivation 51
3.3.7.2 Durchführung 52
3.3.7.2.1 Aufbau 52
3.3.7.2.2 Installation 52
3.3.7.2.3 Konfiguration 54
3.3.7.3 Administration 55
3.3.7.4 Schlussbetrachtung 58
3.4 Marktüberblick 60
3.4.1 Problemorientierte Einleitung 60
3.4.1.1 Motivation und Vorgehen 62
3.4.1.2 Kategorien und Untersuchungsmerkmale 63
3.4.1.3 Auswahl der Big Data Produkte 68
3.4.2 Anbieter und deren Produkte 69
3.4.2.1 Cloudera 70
3.4.2.2 Datameer 73
3.4.2.3 Datawatch 76
3.4.2.4 EMC 79
3.4.2.5 EXASOL 82
3.4.2.6 Hortonworks 86
3.4.2.7 Hewlett-Packard 88
3.4.2.8 IBM 90
3.4.2.9 Informatica 95
3.4.2.10 Jaspersoft 99
3.4.2.11 MicroStrategy 101
3.4.2.12 Oracle 104
3.4.2.13 ParStream 108
3.4.2.14 Pentaho 110
3.4.2.15 QlikTech 113
3.4.2.16 SAP 116
3.4.2.17 SAS 121
3.4.2.18 Software AG 124
3.4.2.19 Splunk 127
3.4.2.20 Teradata 130
3.4.3 Vorstellung der Ergebnisse 133
3.4.4 Schlussbetrachtung 139
3.5 Erstellung eines Prototyps zur Datenintegration 141
3.5.1 Aufgabenstellung 141
3.5.2 Beschreibung der Werkzeuge 141

3.5.3 Konzeption 142
3.5.3.1 Integration der Quelldaten 143
3.5.3.2 ODS-Schicht 144
3.5.3.3 Staging Area 145
3.5.3.4 Data Warehouse 145
3.5.3.5 Reporting 146
3.5.4 Umsetzung 146
3.5.4.1 Installation und Konfiguration 146
3.5.4.2 Beschreibung des Prototyps 147
3.5.4.2.1 Integration der Access Logs 148
3.5.4.2.2 Integration der Stamm- u. Nutzerkontextdaten 150
3.5.4.2.3 Datenextraktion 152
3.5.4.2.4 DWH-Integration 154
3.5.4.2.5 Reporting 155
3.5.4.3 Schlussbetrachtung 159
3.6 Erstellung eines Prototypen zur „Simulation für Capacity Management“ 161
3.6.1 Einleitung 161
3.6.2 Beschreibung des Werkzeugs IBM Cognos TM1 162
3.6.2.1 Aufbau von TM1 162
3.6.2.1.1 Cubes und Dimensionen 163
3.6.2.1.2 Dimensionselemente 163
3.6.2.1.3 Cube-Regeln 163
3.6.2.2 Architektur 164
3.6.2.3 Sicherheitskonzept 166
3.6.2.3.1 Serverauthentifizierung 166
3.6.2.3.2 Objektsicherheit 166
3.6.2.4 Datenimport 167
3.6.2.5 Handbücher 167
3.6.3 Konzeption 168
3.6.4 Umsetzung 170
3.6.4.1 Installation und Konfiguration 171
3.6.4.2 Beschreibung des Prototyps 173
3.6.4.2.1 Die Datenhaltung 173
3.6.4.2.2 Die Simulation 181
3.6.4.2.3 Die Oberfläche 184
3.6.5 Benutzerhandbuch 186
3.6.5.1 Allgemeines 186

3.6.5.2 Anmelden 186
3.6.5.3 Durchführen einer Simulation 189
3.6.5.4 Analysieren der Ist-Daten 198
3.7 Schlussbetrachtung 200

Das folgende Kapitel stellt die Arbeit der Teilprojektgruppe Big Data im IT-Capacity Management zu dem Thema „Herausforderungen durch Big Data am Beispiel des IT-Capacity Managements der Volkswagen AG“ vor. Das Projekt wurde in Zusammenarbeit mit der Volkswagen AG durchgeführt.

Im nachfolgenden Unterkapitel 3.1 wird der Praxispartner vorgestellt. Anschließend erfolgt die Erarbeitung und Vorstellung der diesem Projekt zugrundeliegenden Aufgabenstellung. Das darauf folgende Kapitel 3.2 befasst sich mit der Projektorganisation. Dazu zählt die Zusammensetzung der Teilprojektgruppe aus Mitgliedern der Projektgruppe HAWK und die Rollenverteilung innerhalb der Gruppe. Zudem wird der Projektverlauf erläutert und anhand des Projektplans dargestellt. Anschließend folgt im Kapitel 3.3 die Vorstellung der entwickelten Referenzarchitekturen für die Verarbeitung von Big Data. Darauf aufbauend wird die exemplarische Umsetzung mit Hadoop beschrieben. Der Abschnitt schließt mit der Erläuterung zur Installation eines Hortonworks-Clusters. Um einen Überblick zu erhalten, welche auf dem Markt erhältlichen Softwarelösungen zur Verarbeitung von Big Data vorhanden sind, werden die Ergebnisse der Analyse in Kapitel 3.4 dargestellt. Der Marktüberblick ordnet schließlich die untersuchten Produkte verschiedenen Anwendungsfällen zu. In Zusammenarbeit mit dem Praxispartner wird die im Unternehmen vorhandene Architektur untersucht und für die speziellen Anforderungen ein Prototyp zur Datenintegration entwickelt. Dieses Vorgehen und die konkrete Umsetzung sind in Kapitel 3.5 dargestellt. Die Konzeption und Umsetzung des Prototys zur „Simulation für Capacity Management“ wird im Kapitel 3.6 beschrieben. Im Anschluss an die Vorstellung der erarbeiteten Ergebnisse folgt eine Schlussbetrachtung.

3.1 Projektdefinition

In diesem Kapitel wird zunächst der Praxispartner der Teilgruppe (Kap. 3.1.1) und anschließend die dem Projekt zugrundeliegende Aufgabenstellung (Kap. 3.1.2) vorgestellt. Abschließend führt das Kapitel „IT-Capacity Management“ (Kap. 3.1.3) in das Thema „Herausforderungen durch Big Data am Beispiel des IT-Capacity Managements der Volkswagen AG“ ein.

3.1.1 Praxispartner

Die Teilgruppe „Big Data im IT-Capacity Management“ führte das Projekt mit der Unterstützung und der Betreuung durch die Volkswagen AG in Wolfsburg als Praxispartner durch. Die Volkswagen AG ist ein weltweit agierender Konzern dessen Hauptgeschäft die Produktion von Fahrzeugen ist. Zu ihm gehören im PKW-Segment die Marken Volkswagen PKW, Audi, Seat, Skoda, Bentley, Bugatti, Lamborghini und Porsche, im Nutzfahrzeug-Segment

Volkswagen Nutzfahrzeuge, Scania und MAN sowie als Motorradhersteller die Marke Ducati. Neben der Produktion von Fahrzeugen ist das Unternehmen ebenfalls in der Herstellung von Großdieselmotoren für maritime und stationäre Anwendungen, Turboladern, Turbomaschinen, Kompressoren, chemischen Reaktoren, Spezialgetrieben für Fahrzeuge und Windräder, Gleitlagern, Kupplungen sowie Prüfzentren für den Mobilitätssektor tätig. Zudem werden eine Vielzahl an Finanzdienstleistungen angeboten (Vgl. Volkswagen AG 2014).

Für diese Projektarbeit war das Business Intelligence Competence Center (BI-CC) der Volkswagen AG zuständig, das ein Teil des gesamten Technologie-Kompetenzfeldes Business Information Technologies (T-KF BIT) ist. Das BI-CC wurde im Jahr 2009 gegründet und stellt im Bereich BI projektspezifische Beratungs- und Unterstützungsleistungen, Lösungsbausteine, Leit- und Richtlinien, Standards, Bereitstellungen und Betrieb von Infrastrukturen für ausgewählte BI-Tools und sonstige BI spezifische Dienstleistungen innerhalb des Konzerns bereit (Vgl. Volkswagen AG 2011, S. 10; Volkswagen AG 2012, S.3). Die folgende Abbildung stellt das gesamte Leistungs-Portfolio des BI-CCs dar.

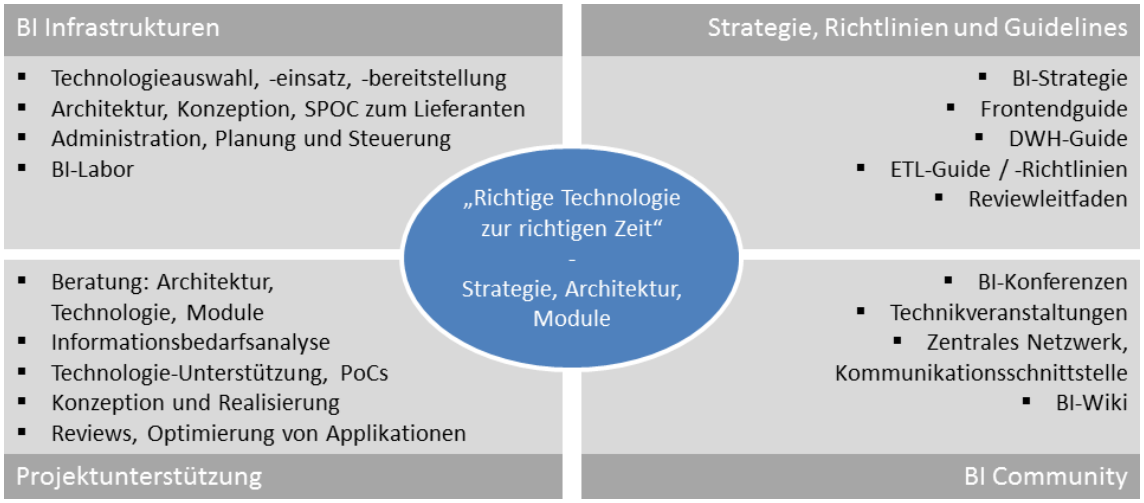


Abbildung 3.1: Leistungs-Portfolio des BI-CCs
Quelle: Volkswagen AG (2011, S. 32)

3.1.2 Aufgabenstellung

Zu Beginn des Projektes wurde von der Teilprojektgruppe eine Auftragsdefinition erstellt. Diese enthält alle Anforderungen an die Teilprojektgruppe. Auftraggeber für das Projekt ist die Abteilung Very Large Business Applications der Universität Oldenburg und das Business Intelligence Competence Center der Volkswagen AG. Die Auftragsdefinition und die Einordnung sind in den folgenden Abschnitten dargestellt.

3.1.2.1 Auftragsdefinition

Die Aufgabenstellung ist insgesamt in das IT-Capacity Management einzuordnen. Im Rahmen der Erweiterung des bestehenden Kapazitätsmanagements der IT-Architektur der Volkswagen AG um mehrere Detaillierungsebenen ist das BI-CC auf Herausforderungen gestoßen, die im Kontext von Big Data diskutiert werden. Bei der Verarbeitung von Access-Log-Dateien zur Ermittlung der notwendigen Kennzahlen wurde erkannt, dass beim Anbinden aller Applikationen die Kapazitäten der BI-Infrastrukturen überschritten würden. Durch die großen Datenmengen, die innerhalb von kurzen Zeitspannen von verschiedenen Servern in unterschiedlicher Struktur erstellt werden und im anschließenden ETL-Prozess verarbeitet werden müssen, ergeben sich neue Herausforderungen. Diese sollen durch die Projektgruppe analysiert werden und es sollen Lösungsvorschläge erarbeitet werden. Darüber hinaus soll gezeigt werden, welchen Nutzen die neuen Kennzahlen bieten. Eine Simulation der Serverauslastungen und -zugriffe im IT-Capacity Management ermöglicht eine bessere Vorhersage und Planung der zukünftigen Anforderungen an die IT-Architektur.

Daraus ergeben sich zwei Aufgabenbereiche für die Teilprojektgruppe „Big Data im IT-Capacity Management“. Zunächst soll nach der Erstellung der Seminararbeiten und der Einarbeitungsphase in projektrelevante Themen eine allgemeine Konzeption für den Aufbau einer Big Data-verarbeitenden IT-Architektur erstellt und exemplarisch umgesetzt werden. In diesem Kontext wird eine Marktanalyse über Big Data-verarbeitende Produkte in Auftrag gegeben. Die Erkenntnisse aus dieser Untersuchung sollen im Anschluss auf die spezifischen Anforderungen der Volkswagen AG angewandt und bewertet werden.

Es soll ein Konzept erstellt werden, in dem dargestellt wird, wie in einem Big Data-Umfeld semi-strukturierte Daten in der bestehenden IT-Architektur der Volkswagen AG verarbeitet werden können. Dabei wird besonderer Wert auf die Integrationsfähigkeit gelegt. Mithilfe einer Marktanalyse sollen mögliche Anbieter und deren Lösungen aufgezeigt und untersucht werden. Im Anschluss daran soll eine Empfehlung für eines der untersuchten Produkte erfolgen. Dabei sollen die Unterschiede zwischen den Produkten deutlich gemacht werden und es soll eine Zuordnung zu Anwendungsfällen erfolgen. Weiter ist es notwendig, dass ein Prototyp für die Verarbeitung von Access-Log-Dateien entwickelt wird.

Der Prototyp soll mit den aufgeführten Tools umgesetzt werden. In der Konzeption können weitere, bisher nicht aufgeführte Werkzeuge vorgeschlagen werden. Es sollen mehrere Datenmodelle erstellt werden. Die Lösung mit den bei der Volkswagen AG nicht eingesetzten Tools soll bei einer Einsatzempfehlung entsprechend begründet sein. Abschließend soll eine Simulations- und Planungsvariante für das IT-Capacity Management erstellt werden. Für diese Simulation kann z.B. Oracle Essbase verwendet werden. Das Ergebnis stellt die berechneten Hardwareanforderungen für die Gegenwart und die simulierte Zukunft dar. Mit der Simulation soll die Entscheidungsfindung unterstützt und ein Beschaffungsplan erstellt werden.

3.1.2.2 Interpretation und Anpassungen

In der zuvor dargestellten Auftragsdefinition wird eine allgemeine Konzeption für den Aufbau einer Big Data-verarbeitenden IT-Architektur gefordert. Die daraufhin entwickelten Referenzarchitekturen werden im Kapitel 3.3 vorgestellt.

Die genauen Anforderungen ergeben sich aus den von der Volkswagen AG bereitgestellten Lastenheften. Für die Erstellung des Prototypen zur Datenintegration wurde das Lastenheft „Business & Service Daten Elsa Pro Access Logs“ (Vgl. Peddinghaus u. a. 2011) zur Verfügung gestellt. Aus diesem Lastenheft leitet sich das Pflichtenheft „Datenintegration“ ab und definiert die Anforderungen (siehe Kapitel F.3 im Anhang). Die Grundlage für die Erstellung des Prototyps zur Simulation im IT-Capacity Management bildet das Lastenheft „Simulation für Capacity Management“ (Vgl. Volkswagen AG 2013) und das abgeleitete Pflichtenheft (Vgl. Kap. F.4 im Anhang).

Der anhand der Pflichtenhefte entwickelte Prototyp für die Datenintegration wird im Kapitel 3.5 beschrieben. Anschließend wird in Kapitel 3.6 der Prototyp für die Simulation vorgestellt. Die Simulation wird dabei entgegen der ersten Auftragsdefinition in Abstimmung mit den Auftraggebern mit IBM Cognos TM1 erstellt. Gründe für diese Entscheidung ergaben sich während des Projektverlaufs durch gravierende Verzögerungen in vorherigen Phasen. Der dadurch entstandene Termindruck zwang die Teilgruppe dazu, das bereits vorhandene Know-how zu nutzen und auf ein in Umfang und Funktionalität ähnliches Tool zu wechseln.

In der Auftragsdefinition ist zudem eine Marktanalyse vorgesehen, in deren Anschluss eine Empfehlung für eines der untersuchten Produkte erfolgen soll. Bei der Analyse der erhältlichen Big Data-Produkte wurde jedoch deutlich, dass die möglichen Anwendungsfälle sehr heterogen sind und die verschiedenen Produkte mit ihren unterschiedlichen Funktionsumfängen und Kombinationsmöglichkeiten mit anderen Produkten nur schwer vergleichbar sind. So ist es nicht möglich aufzuzeigen, für welche Anwendungsfälle bestimmte Produkte zu empfehlen sind. Aufgrund dieser Erkenntnis wird die ursprünglich vorgesehene Empfehlung für eines oder mehrere Produkte verworfen. Stattdessen wird ein allgemein gehaltener Marktüberblick gegeben, um eine weitreichendere Sicht ohne eingrenzende Beschränkungen durch ein konkretes Anwendungsszenario zu erhalten. Das Kapitel 3.4 stellt den Marktüberblick dar.

3.1.3 IT-Capacity Management

Im Rahmen dieses Projektes wird der Prozess des IT-Capacity Managements betrachtet. Dabei geht es um die Balance zwischen Kosten und Kapazität, die Stych und Zeppenfeld mit den folgenden Fragen definieren (Vgl. Stych u. Zeppenfeld 2009, S. 102ff.):

- „Wie viel kostet die Anschaffung von Kapazitäten, um die geschäftlichen Anforderungen zu erfüllen?
- Wird die Kapazität effizient genutzt?

- Ist die Kapazität ausreichend für die künftige Nachfrage des Kunden?
- Wann ist die Performanz ausgereizt und wann sollte zusätzliche Kapazität geschaffen werden?"

Der Begriff des Capacity Managements stammt aus der IT Infrastructure Library (kurz ITIL). Dies ist ein herstellerunabhängiges Framework, das von der britischen Regierung in den 80er Jahren in Auftrag gegeben wurde, um die IT-Ressourcen des öffentlichen Sektors effizienter und kosteneffektiver nutzen zu können. Bis heute wird ITIL vom OGC (Office of Government Commerce) weiterentwickelt und dabei von verschiedenen Unternehmen unterstützt. Die erste Version dieser Sammlung wurde 1995 abgeschlossen, die zweite 1999 und seit 2007 gibt es die dritte Version von ITIL. Zusammenfassend ist ITIL eine „Bibliothek von Büchern, die das bestbewährte Erfahrungswissen aus der Praxis von IT-Organisationen vereinheitlichend zusammenfasst, in dem die Inhalte, Prozesse und Ziele innerhalb eines Unternehmens als ‚was zu tun ist' verdeutlicht werden, ohne eine genaue Vorschrift über das ‚wie es zu tun ist' festzulegen." (Stych u. Zeppenfeld 2009, S.11f.).

ITIL umfasst zahlreiche Prozesse, welche in Service-Support und Service Delivery unterteilt werden. Zu den Prozessen des Service-Supports zählen das Incident Management mit der Funktion des Service-Desks, das Problem-Management, das Configuration Management, das Change Management und das Release-Management. Diese Prozesse dienen der Unterstützung von IT-Services und behandeln Kunden- und Anwenderschnittstellen (Vgl. Stych u. Zeppenfeld 2009, S. 41). Zum Service Delivery zählen das Service-Level-Management, das Financial-Management, das Capacity Management, das Availability Management, das Continuity Management und das Security Management. Diese sind für die Planung und die Service-Bereitstellung, definieren aber auch Voraussetzungen und Maßnahmen dafür (Vgl. Stych u. Zeppenfeld 2009, S. 85f.).

Das Capacity Management besteht aus drei Subprozessen: Business Capacity Management, Service Capacity Management und Resource Capacity Management. Beim Business Capacity Management werden Vorhersagen über zukünftige Bedürfnisse und Anforderungen erstellt. Dabei werden ein Kapazitätsplan erstellt, Schätzungen, Analysen und Simulationen modelliert und das Application-Sizing durchgeführt. Beim Service Capacity Management wird die Nutzung von IT-Dienstleistungen bestimmt und beim Resource-Capacity-Management die Nutzung der IT-Infrastruktur (Vgl. Stych u. Zeppenfeld 2009, S. 104f.). Die folgende Abbildung zeigt diese Unterteilung und welche Schnittstellen das Capacity-Management zu anderen Prozessen hat.
Der Kapazitätsplan, der beim Capacity Management erstellt wird ist für dieses Projekt besonders relevant. Er beinhaltet die Rahmenbedingungen der Planung, Annahmen und Voraussetzungen, den Bewertungsgegenstand und Geschäftsszenarien, ein Mengengerüst und eine Ressourcenübersicht, den Kostenplan, Empfehlungen und eine Nutzen-Risiko-Analyse sowie eine Management-Zusammenfassung und stellt somit das Ergebnis des Capacity-Managements dar (Vgl. Stych u. Zeppenfeld 2009, S. 103).

Abbildung 3.2: Prozess des Capacity Managements
Quelle: Stych u. Zeppenfeld (Vgl. 2009, S. 103)

Für eine erfolgreiche Durchführung des Capacity Management-Prozesses sind im Vorfeld einige Dinge zu beachten. Stych und Zeppenfeld nennen die folgenden kritischen Erfolgsfaktoren für die Implementierung (Vgl. Stych u. Zeppenfeld 2009, S. 106):

- „Genaue Vorhersagen und Prognosen für das Unternehmen
- Kenntnis der IT-Strategie und IT-Planung
- Kenntnis der Entwicklung im Bereich der Technologie
- Zusammenarbeit mit anderen Prozessen“

Für die Überprüfung des Erfolges des Capacity Managements nennen sie die folgenden Key Performance Indicators (KPI) (Vgl. Stych u. Zeppenfeld 2009, S. 106):

- „Abweichung zwischen Business-Plan und Kapazitätsplan
- Anzahl von Überkapazitäten
- Anzahl der Vertragsverletzungen wegen Kapazitätsengpässen
- Verringerung der Anzahl der Panikkäufe
- Reduzierung der Störungen auf Grund von Performanzproblemen“

Zusammenfassend lässt sich mithilfe des Capacity Managements das Risiko der Überlastung von Ressourcen minimieren. Dies wird durch die ständige Überwachung und die Kenntnis, welche Auswirkungen Änderungen haben, ermöglicht. Durch eine frühzeitige Abstimmung von benötigten und verfügbaren Ressourcen lässt sich die Effizienz steigern, was zu einer Reduzierung der Kosten führen kann.

3.2 Projektorganisation

Im folgenden Kapitel wird auf die Projektorganisation der Teilprojektgruppe „Big Data im IT-Capacity Management“ eingegangen, indem die Organisationsstruktur (Kap. 3.2.1), der Projektplan (Kap. 3.2.2) und der Projektverlauf (Kap. 3.2.3) erläutert werden.

3.2.1 Organisationsstruktur

Die Teilprojektgruppe „Big Data im IT-Capacity Management“ besteht aus sechs Mitgliedern, die während der Projektlaufzeit verschiedene Rollen und Aufgaben übernahmen. Die Aufteilung erfolgte in Abhängigkeit des Bedarfs und der vorhandenen Ressourcen. Den Rahmen bildete der Projektplan (Kap. 3.2.2), der ebenfalls die Ressourcen berücksichtigte.

Die Projektleitung übernahm Björn Berger. Zusammen mit Simon Brosig und Sören Flathmann übernahm er vor allem die Arbeitspakete zur Erstellung der Referenzarchitekturen und zur technischen Umsetzung. Die Projektgruppenmitglieder Catharina Petry, Yvonne Tevs und Jonas Borrmann legten ihren Schwerpunkt vornehmlich in die Erstellung des Marktüberblicks. Die Abbildung 3.3 zeigt das Organigramm der Teilprojektgruppe.

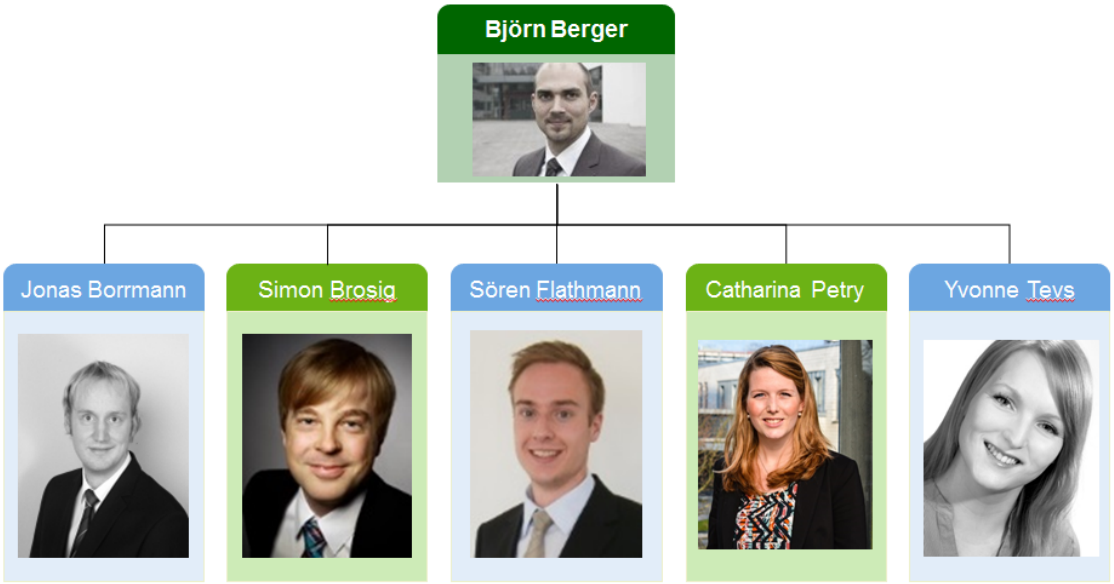
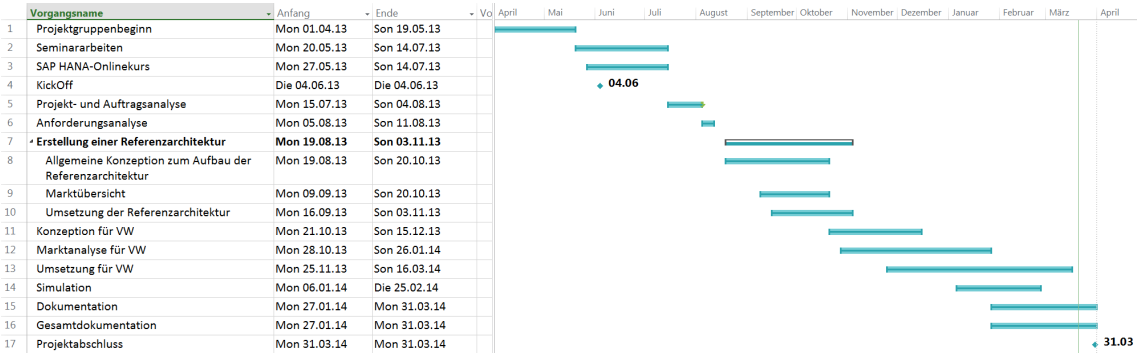


Abbildung 3.3: Mitglieder der Teilprojektgruppe Big Data im IT-Capacity Management

In den unterschiedlichen Phasen des Projekts wurden von den Projektmitgliedern verschiedene Aufgaben durchgeführt. Je nach Arbeitspaket nahm eine Person eine andere Rolle ein. So gehörten zur technischen Umsetzung die Rollen Designer, Entwickler, Spezifizierer, Tester und Dokumentierer. Die Arbeitspakete zur Erstellung des Marktüberblicks erforderten dagegen die Rollen Marktanalyst und Dokumentierer.

3.2.2 Projektplan

Zu Beginn des Projektes wurden anhand der Aufgabenstellung die verschiedenen Tätigkeiten ermittelt. Diese wurden in einen Projektplan integriert, indem die Abhängigkeiten und vorhandenen Ressourcen einbezogen wurden. Die Abbildung 3.4 stellt diesen dar. Der Projektplan besteht aus 17 Vorgängen, wovon 12 als Projektphase deklariert wurden.



	Vorgangsname	Anfang	Ende
1	Projektgruppenbeginn	Mon 01.04.13	Son 19.05.13
2	Seminararbeiten	Mon 20.05.13	Son 14.07.13
3	SAP HANA-Onlinekurs	Mon 27.05.13	Son 14.07.13
4	KickOff	Die 04.06.13	Die 04.06.13
5	Projekt- und Auftragsanalyse	Mon 15.07.13	Son 04.08.13
6	Anforderungsanalyse	Mon 05.08.13	Son 11.08.13
7	**Erstellung einer Referenzarchitektur**	**Mon 19.08.13**	**Son 03.11.13**
8	Allgemeine Konzeption zum Aufbau der Referenzarchitektur	Mon 19.08.13	Son 20.10.13
9	Marktübersicht	Mon 09.09.13	Son 20.10.13
10	Umsetzung der Referenzarchitektur	Mon 16.09.13	Son 03.11.13
11	Konzeption für VW	Mon 21.10.13	Son 15.12.13
12	Marktanalyse für VW	Mon 28.10.13	Son 26.01.14
13	Umsetzung für VW	Mon 25.11.13	Son 16.03.14
14	Simulation	Mon 06.01.14	Die 25.02.14
15	Dokumentation	Mon 27.01.14	Mon 31.03.14
16	Gesamtdokumentation	Mon 27.01.14	Mon 31.03.14
17	Projektabschluss	Mon 31.03.14	Mon 31.03.14

Abbildung 3.4: Projektplan der Teilgruppe Big Data im IT-Capacity Management

Jedes Ende einer Phase stellt einen Meilenstein dar. Durch ein vereinbartes iteratives Vorgehen wurden die Ergebnisse jeder Phase besprochen und in Telefonkonferenzen mit dem Praxispartner abgestimmt. Notwendige Anpassungen am Projektplan wurden anhand neuer Erkenntnisse nach Abschluss einer Phase durchgeführt und in den Konferenzen dargelegt. Die Telefonkonferenzen wurden im Beisein eines Betreuers in monatlichem Turnus ab dem Juli 2013 mit dem Praxispartner durchgeführt. In den Besprechungen wurden die Ergebnisse vorgestellt und diskutiert sowie weitere Aufgaben besprochen.

Die erste Phase vom 01. April 2013 bis zum 19. Mai 2013 hat den Namen Projektbeginn. Sie beinhaltet die Vorbereitungen zur Durchführung des Projekts und die Bereitstellung der technischen Systeme zur Organisation der Projektgruppe. Im anschließenden Zeitraum bis zum 14. Juli 2013 wurden die Seminararbeiten erstellt. Nahezu zeitgleich vom 27. Mai 2013 bis zum 14. Juli 2013 führten die Projektgruppenmitglieder einen SAP HANA-Onlinekurs durch.

Für die Teilgruppe „Big Data im IT-Capacity Management“ von besonderer Bedeutung ist der Meilenstein KickOff vom 04. Juni 2013, der die Vorstellung der Aufgabenstellung durch den Praxispartner Volkswagen AG bedeutet. In der Zeit vom 15. Juli 2013 bis zum 04. August 2013 wurde in der Phase „Projekt- und Auftragsanalyse“ die vorgestellte Aufgabenstellung untersucht. In der Woche vom 05. August 2013 bis zum 11. August 2013 wurden die Anforderungen erstellt und mit dem Praxispartner abgestimmt.

Nach einer Woche Projektgruppenurlaub begann die erste Umsetzungsphase zur „Erstellung einer Referenzarchitektur“ und endete am 03. November 2013. Sie bestand aus den drei Phasen „Allgemeine Konzeption zum Aufbau der Referenzarchitektur“, „Marktüberblick“ und „Umsetzung der Referenzarchitektur“, die sich größtenteils überlappten.

Am 21. Oktober 2013 konnte mit der Phase „Konzeption für VW“ begonnen werden. Diese Phase schloss sich thematisch und zeitlich unmittelbar an die Phase „Allgemeine Konzeption zum Aufbau der Referenzarchitektur“ an. Sie endete am 15. Dezember 2013. Eine Woche nach der Phase „Konzeption für VW“ begann die Phase „Marktanalyse für VW“, die inhaltlich auf den Ergebnissen der Phase Marktüberblick aufbaute und die Phase am 26. Januar 2014 abschloss.

Die Phase „Umsetzung für VW“ war für den Zeitraum vom 25. November 2013 bis zum 16. März 2014 eingeplant. Mit Beginn der ersten Woche im Jahr 2014 wurde die Arbeit in der Phase „Simulation“ aufgenommen und am 25. Februar 2014 abgeschlossen. Die Dokumentation der Teilgruppenergebnisse sowie die Gesamtdokumentation wurde vom 27. Januar 2014 bis zum Ende der Projektgruppenlaufzeit am 31. März 2014 erstellt. Der 31. März 2014 stellte den letzten Meilenstein zum Abschluss des Projekts dar.

3.2.3 Projektverlauf

Der Startschuss für die Teilprojektgruppe „Big Data im IT-Capacity Management“ fiel mit dem KickOff-Meeting bei dem Praxispartner Volkswagen AG. Auf Einladung des BI-CC reisten Prof. Dr. Jorge Marx Gómez, die Betreuer Andreas Messler und Andreas Solsbach und die Teilgruppenmitglieder am 04. Juni 2013 nach Wolfsburg, um zum einen die Auftraggeber der Volkswagen AG kennenzulernen und zum anderen die genaue Aufgabenstellung zu erfahren und den Projektrahmen kennenzulernen. Es konnten bereits erste mögliche Umsetzungsszenarien abgestimmt werden und der Erwartungshorizont wurde definiert.

Mit der Projekt- und Auftragsanalyse konnte fünf Wochen später in den Phasen „Projekt- und Auftragsanalyse“ und „Anforderungsanalyse“ begonnen werden. Die ersten Rückmeldungen erhielt die Teilgruppe für die Auftragsdefinition in der Telefonkonferenz vom 30. Juli 2013. Anhand der übermittelten Lastenhefte wurden Pflichtenhefte erstellt, die im Anschluss abgestimmt wurden. Diese bildeten die Basis für die Umsetzung in den Phasen „Umsetzung für VW“ und „Simulation“.

Nach Abschluss dieser Phase konnte mit den ersten Arbeitspaketen begonnen werden. Zur Annäherung an die grundsätzlichen Herausforderungen wurden zunächst Konzepte für Referenzarchitekturen anhand von Best-Practice-Ansätzen erstellt und zum Test implementiert. Zeitgleich wurde von drei Projektmitgliedern der Marktüberblick über bereits verfügbare Lösungen erstellt. Zunächst interpretierte die Teilgruppe die Aufgabenstellung in der Art, dass eine allgemein gültige Empfehlung für bestimmte Lösungen erarbeitet werden sollte. Es stellte sich nach ersten Recherchen heraus, dass dies aufgrund der ungewissen Rahmenbedingungen nicht möglich ist. In Rücksprache mit dem Praxispartner wurde sich darauf geeinigt, dass die Teilgruppe einen Marktüberblick für Big Data-Lösungen erstellt, der ergebnisneutral die Funktionen und Möglichkeiten der Produkte darstellt. Der Marktüberblick sollte schließlich um eine Klassifizierung und Gruppierung der Produkte erweitert werden. Aufgrund des erhöhten Aufwand in diesen Arbeitspaketen wurden die Phasen um sechs Wochen erweitert.

Die andere Hälfte der Teilgruppe beschäftigte sich zunächst mit der Konzeption der Referenzarchitekturen. Im Anschluss sollte ein Szenario exemplarisch in einer vorgegebenen Softwareumgebung des Praxispartners umgesetzt werden. Durch Verzögerungen in der Bereitstellung der Entwicklungsumgebung wurde die Umsetzungsphase um fünf Wochen nach hinten terminiert. Der tatsächliche Beginn der Entwicklung nach Erstellung der Konzeptionen verschob sich aufgrund von verschiedenen Verbindungs- und Bereitstellungsproblemen zur Erstellung des Prototypen zur Datenintegration bis Anfang Februar 2014. Somit wurde bis zum 16. März 2014 der Prototyp zur Datenintegration von zwei Projektgruppenmitgliedern entwickelt.

Grundsätzlich organisierte die Teilprojektgruppe sich in Abhängigkeit zu den Phasen des Projektplans. Die Zuordnung der Teilgruppenmitglieder zu den Arbeitspaketen erfolgte unter Einschätzung der Verfügbarkeit und der erforderlichen Kenntnisse. Neben dem obligatorischen, wöchentlichen internen Projektgruppentreffen wurde sich in den Kleingruppen einer Projektphase mindestens ein weiteres Mal pro Woche für mehrere Stunden getroffen, um die Ergebnisse zu diskutieren, gemeinsam zu arbeiten oder die weiteren Tätigkeiten zu besprechen. Darüber hinaus wurde die generelle Ausrichtung regelmäßig in der gesamten Teilgruppe diskutiert.

Um die Aufgaben und Ergebnisse mit dem Praxispartner zu besprechen, wurde monatlich eine Telefonkonferenz durchgeführt. Anhand eines Leitfadens wurden in der Konferenz die Fortschritte und weiteren Tätigkeiten besprochen, der von der Teilgruppe vor jedem Treffen erstellt wurde.

Neben der Projektarbeit konnten die Mitglieder der Teilprojektgruppe an der BI-Konferenz des BI-CC der Volkswagen AG am 28. November 2013 in Wolfsburg teilnehmen. Wiederrum lud das BI-CC ein, um verschiedene Vorträge von internen und externen Referenten zum Thema Big Data zu hören. Die Vorträge dieses Tages fügten sich nahtlos in die Grundthematik der Teilgruppe ein und gaben Anstöße zu neuen Überlegungen in den verschiedenen Arbeitspaketen.

3.3 Referenzarchitekturen

Die folgenden Abschnitte stellen die Ergebnisse und die Vorgehensweise zur Entwicklung der Referenzarchitekturen im Rahmen des Projektes dar. Zunächst wird der Bezug der Aufgabenstellung zum Hauptthema der Projektgruppe in einer Einleitung (Kap. 3.3.1) hergestellt. Es folgt die detaillierte Beschreibung von drei Szenarien (Kap. 3.3.2, 3.3.3 und 3.3.4). Im Anschluss daran werden besondere Herausforderungen in diesem Themenkomplex aufgezeigt. Die beispielhafte Umsetzung einer Referenzarchitektur wird in dem Abschnitt 3.3.6 beschrieben. Das Kapitel schließt mit der Beschreibung der Installation eines Hadoop-Clusters mit Hortonworks.

3.3.1 Einleitung

Zur Bereitstellung reportingrelevanter Informationen aus einer Quelle mit Daten in unterschiedlicher Strukturierung sind verschiedene Architekturen denkbar. Jede Architektur beantwortet unterschiedliche Anforderungen an die Datenherkunft, den Umfang, die Häufigkeit, das Datenvolumen und die speziellen Herausforderungen, die sich durch den Umfang des Reportings ergeben. Mit einher gehen immer auch Rahmenbedingungen wie die verfügbare Speicherkapazität und die Verarbeitungsgeschwindigkeit.

Besondere Hürden ergeben sich im Kontext von Big Data durch drei Merkmale: Velocity, Variety und Volume (Vgl. IBM Corporation 2013a; Bitkom 2013).

- **Velocity** (Geschwindigkeit) beschreibt in welcher Geschwindigkeit die Daten generiert bzw. übertragen werden und ob dies beispielsweise Realtime oder in einem bestimmten Rhythmus geschieht. Echtzeitanalysen sind in den Bereichen Wetter, Verkehr, Gesundheit oder Börsenhandel von großer Bedeutung. Dabei stellt sich die Frage, ob und wie die Daten zeitnah extrahiert und verarbeitet werden können.

- **Variety** (Datenvielfalt) definiert, ob es sich bei den Daten um Fremd- oder Firmendaten handelt, diese un-, semi- oder strukturierte sind und aus welchen Quellen diese stammen. Sowohl die unterschiedlichen Strukturen in einer Quelle als auch die große Anzahl an verschiedenen Quellen stellen eine Herausforderung an den Integrationsprozess und das ETL-Tool dar. Bisher eher weniger beachtete Informationen, die sich z.B. im Social Media, in Emails oder in Berichten befinden, bringen die Verarbeitung auf eine neue Ebene.

- **Volume** (Datenmenge) definiert die gesamte Menge an Daten bzw. deren Dateigröße. So wird unter Big Data häufig auch eine große zu verarbeitende Datenmenge verstanden, die an bestehende Systeme neue Herausforderungen stellt.

Die Abbildung 3.5 zeigt die Unterscheidung der entwickelten Zielarchitekturen in drei Szenarien. Es handelt sich bei der Darstellung immer um die vollständige Aufbereitung der operativen Daten von der Quelle bis zum Reporting.

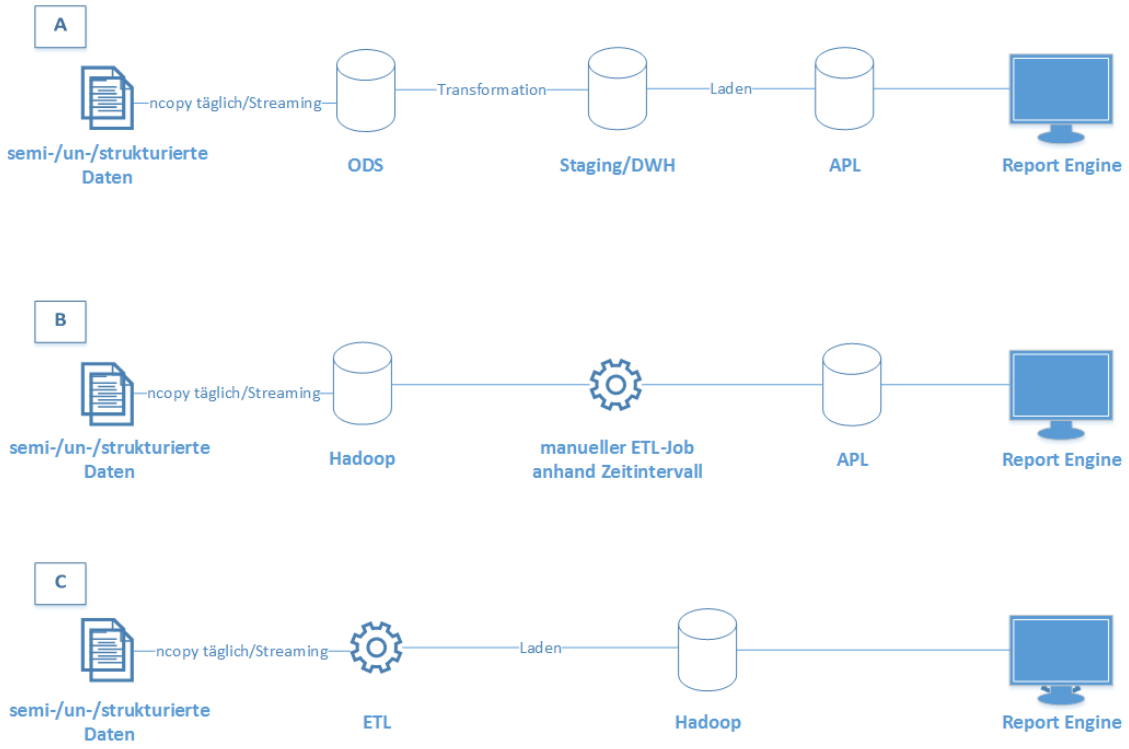


Abbildung 3.5: Referenzarchitekturen

Szenario A ähnelt dem standardisierten ETL-Prozess. In diesem werden die Daten zunächst aus den Quellen extrahiert. Diese können in unterschiedlicher Strukturierung vorliegen. In einem ersten Schritt der Extraktion werden diese im Operational Data Store (ODS) historisiert. Anschließend erfolgt die Aufbereitung der Daten durch die Transformation. Das Ergebnis sind aufbereitete Daten in der Staging Area bzw. in der Datawarehouseing-Schicht. Zur Bereitstellung der Daten zu Reporting- und Analysezwecken, werden diese in das bzw. die Application Layer übertragen, auf die dann die Report Engines zugreifen können.

Szenario B stellt eine Möglichkeit dar, die Techniken integriert, die im Kontext von Big Data diskutiert werden. Ein verteiltes Hadoop-System mit dem Hadoop Dateisystem (HDFS) stellt die Speicherungskomponente dar, die im Vergleich zu den klassischen Speichermöglichkeiten deutlich kostengünstiger sein soll und in diesem Szenario sämtliche Daten historisiert vorhält. Ein manueller ETL-Job kann unter Angabe eines Zeitintervalls Zeitscheiben adhoc in der Auswertungsschicht (APL) bereitstellen. Dieses Szenario eignet sich im Vergleich zu Szenario A eher für große Datenmengen, die sich durch eine große Detailtiefe und einen großen Zeitumfang auszeichnen.

Szenario C gleicht einem klassischen ETL-Prozess. In einem ersten Schritt werden die Daten extrahiert und transformiert. Anschließend werden sie in einen Hadoop-Cluster geladen, der die Daten auf mehrere Nodes verteilt und weiterverarbeitet. Die Report Engine kann über geeignete Konnektoren direkt auf den Hadoop-Cluster zugreifen und Abfragen ausführen.

Dieses Kapitel stellt in den folgenden Abschnitten neben der ausführlichen Diskussion der Szenarien A, B und C die besonderen Herausforderungen durch die Parameter Variety und Velocity dar. Im Anschluss wird die beispielhafte Umsetzung einer Referenzarchitektur dargestellt. Darüber hinaus wird die Installation eines Hadoop-Clusters mit Hortonworks beschrieben.

3.3.2 Szenario A

In diesem Abschnitt wird das Szenario A der Referenzarchitekturen beschrieben. Hierbei wird zunächst auf die Architektur an sich eingegangen. Abgeschlossen wird dieser Abschnitt anschließend mit einer Erläuterung möglicher Anpassungen der Architektur zur Verarbeitung von Big Data.

3.3.2.1 Architektur

In diesem Szenario, welches dem klassischen ETL-Prozess ähnelt, werden die verschiedenen Daten (semi-/un-/strukturiert) aus den unterschiedlichen Quellen extrahiert und zunächst in einen Operational Data Store (ODS) überführt. Dabei wird bewusst offengehalten, ob dies beispielsweise einmal täglich oder über einen permanenten Daten-Stream geschieht. Ein Operational Data Store ist eine spezielle Datenbank, welche dazu dient Daten aus unterschiedlichen Quellen zusammenzuführen und zu vereinheitlichen. Des Weiteren bietet ein Operational Data Store die Möglichkeit operative Daten für Analysen bereitzustellen. Es ermöglicht damit neben der strategischen Sichtweise in klassischen Data Warehouse Systemen auch eine operative Sicht einfließen zu lassen. Das ODS bildet in dieser Architektur jedoch lediglich eine Vorstufe zum eigentlichen Data Warehouse (Vgl. Enzyklopaedie der Wirtschaftsinformatik 2013a). Da die Daten nicht im ODS verbleiben, werden sie anschließend mit Hilfe eines Transformationsprozesses weiter in das Data Warehouse (DWH) transportiert. In der Ebene des Data Warehouse kommen die Daten zunächst temporär in eine Staging Area. Diese dient zur Vorbereitung der Daten für das eigentliche Data Warehouse und ist eine typische Komponente eines ETL-Prozesses bzw. einer Data Warehouse Architektur. In dieser Ebene werden die Daten transformiert, so werden bspw. Datenbereinigungen durchgeführt oder auch Datensätze aggregiert. Die Hauptkomponente des Data Warehouse ist die zentrale Basisdatenbank. Diese dient zur Speicherung der Daten in der feinsten Granularität und muss daher große Datenmengen bewältigen können. Bei diesen Daten handelt es sich je nach Art des Data Warehouses um historische Daten beispielsweise vom Vortag (klassisches Data Warehouse) oder um Echtzeitdaten direkt aus den operativen Systemen (Real Time Data Warehouse) (Vgl. Enzyklopaedie der Wirtschaftsinformatik 2013b). Zur Unterteilung des Data Warehouses können sogenannte Data Marts verwendet werden, welche eine Auswahl von Daten aus der Basisdatenbank darstellen. Bei den Daten handelt es sich jedoch lediglich um eine temporäre Kopie. Mit Hilfe solcher Data Marts können Daten für Analyse- und Reporting-Zwecke zusammengeführt und in speziellen Sichten bereitgestellt werden (Vgl. DWDM 2013). Auf diese Data Marts greifen nun die verschiedenen Report- und Simulations Tools zu. Zur Übersicht über diese Architektur dient die Abbildung 3.6.

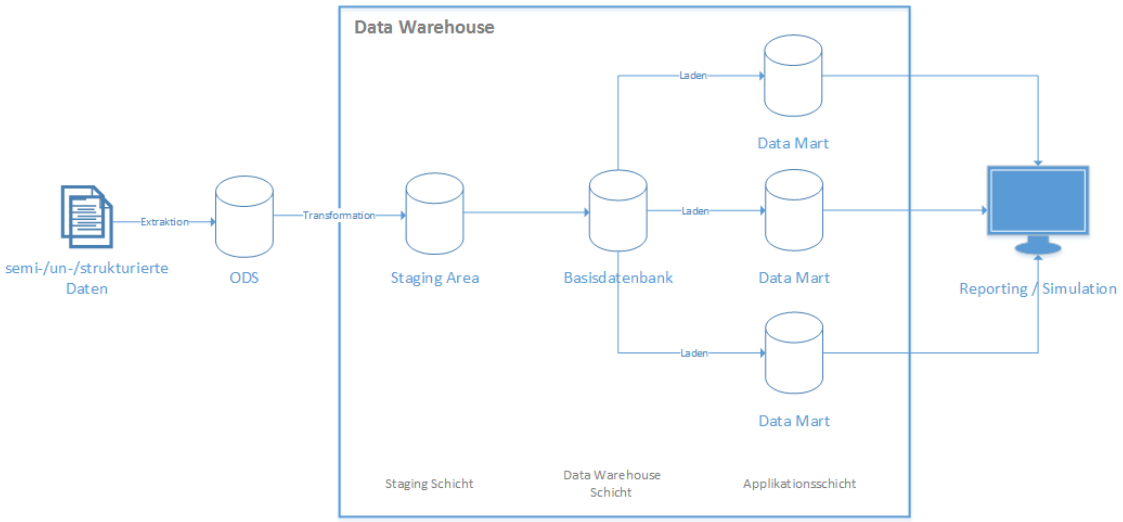


Abbildung 3.6: Szenario A im Detail

3.3.2.2 Anpassungen zur Verarbeitung von Big Data

Damit die Architektur dieses Szenarios, die einem klassischen Data Warehouse ähnelt, auch für den Einsatz in einem Big Data Umfeld geeignet ist, kann eine Reihe von Anpassungen bzw. Ergänzungen vorgenommen werden:

- Die einzelnen Komponenten könnten auf verschiedene Systeme verteilt werden, beispielsweise kann/können die Staging Area(s) auf mehrere Systeme verteilt werden und bieten somit mehr Ressourcen und damit eine bessere Performance.

- Es können Data Marts eingesetzt werden, um bestimmte Datenbereiche zur Analyse auszulagern. Diese Data Marts können auf extra Systemen umgesetzt werden, um die Ressourcen besser zu verteilen und somit schnellere Zugriffs-/Antwortzeiten zu ermöglichen. So wäre es in der Architektur des Praxispartners beispielsweise denkbar, für die verschiedenen Kennzahlenbereiche jeweils eigene Data Marts bereitzustellen, auf diese die jeweiligen Simulations- und Reporting-Anwendungen dann zugreifen.

- Weiter wäre es denkbar die Data Marts mit Hilfe von Technologien wie In-Memory-Datenbanken zu realisieren und hierdurch sehr schnelle Analysen zu ermöglichen. Diese benötigen jedoch weitere, herkömmliche Datenspeicher zur Unterstützung, da es nicht realistisch ist, große Datenmengen komplett im Hautspeicher der In-Memory Datenbank vorzuhalten.

- Häufig geht Big Data mit dem Begriff Echtzeitanalysen einher (Vgl. Ortega 2013). Da klassische Data Warehouses nur begrenzt Echtzeitanalysen ermöglichen, ist dieses auch nur begrenzt für Big Data geeignet. Hierfür kann jedoch ein Realtime Data Warehouse verwendet werden, welches beispielsweise durch einen permanenten Daten-Stream von Daten aus den operativen Quellsystemen in die Basisdatenbank erreicht wird.

Die konkrete Umsetzung einer Big Data verarbeitenden Architektur in einem Unternehmen hängt, besonders wenn sie wie in diesem Szenario auf einem Data Warehouse basiert, von den drei "V" Velocity, Variety und Volume ab, da hierbei die genauen Anforderungen des Unternehmens berücksichtigt werden müssen. So wäre es denkbar, dass ein Unternehmen für sich Big Data eher in Richtung Geschwindigkeit und Datenvielfalt definiert und weniger auf die Datenmenge. Oder umgekehrt, dass viele sehr ähnliche Daten in großer Anzahl in dem Unternehmen als Big Data verstanden und verarbeitet werden sollen.

3.3.3 Szenario B

Dieser Abschnitt beschreibt zunächst die Architektur des Szenarios B. Die folgenden Abschnitte Datenbankintegration, Integration via ETL-Tool, Bereitstellung der Daten und Procedures diskutieren verschiedene Lösungsmöglichkeiten anhand der Architektur.

3.3.3.1 Architektur

Dieses Szenario kann die Probleme beheben, die durch die Integration von Big Data in eine BI-Landschaft entstehen können. Eine Herausforderung, die Big Data mit sich bringt wird unter dem Begriff „Volume", also der Datenmenge zusammengefasst und meint die außerordentlichen Aufgaben, die durch eine sehr große Datenmenge entstehen und gemeistert werden müssen. Eine bestehende BI-Infrastruktur in einem Unternehmen ist dafür in der Regel nicht ausgelegt. Eine größere Datenmenge kann mit den bisherigen Systemen nicht immer in der vorgegebenen Zeit verarbeitet werden und die Speichermöglichkeiten können nicht ausreichend sein.

Eine Möglichkeit, diese Problematik zu umgehen ist auf andere Techniken zurückzugreifen. Unter der Annahme, dass die Report Engine und die Datenbank der Anwendungsschicht (APL) weiter genutzt werden sollen, aber der Speicher der APL begrenzt ist, müssen zudem organisatorische Regeln definiert werden. Der Vorteil dieser Vorgehensweise ist, dass die neuen Big Data-Kennzahlen mit den bestehenden Geschäftsdaten verbunden werden können und gemeinsam ausgewertet werden können. Die Abbildung 3.7 stellt den Aufbau dar.

Abbildung 3.7: Referenzarchitektur - Szenario B

Eine organisatorische Regel kann sein, dass die Anwendungsschicht nicht die übliche Detaillierungstiefe oder die übliche Intervallbreite für die Big Data-Kennzahlen vorhält. Dies bedeutet, dass der Anwender zum einen über diese Tatsache Bescheid weiß und bei der Nutzung diesen

Umstand beachtet. Darüber hinaus kann er Schritte veranlassen, dass für seine Auswertungen die benötigten Daten in der benötigten Detaillierungstiefe vorliegen.

Nachteil dieser Lösung ist, dass die neuen Big Data-Kennzahlen nicht adhoc auswertbar sind und nicht permanent bereit stehen. Fertige Analysen können nicht spontan abgerufen oder aktualisiert werden. Durch das erneute Laden geht viel Zeit verloren.

Seine Vorteile hat diese Lösung allerdings in der ganzheitlichen Betrachtung. Eine bestehende BI-Architektur muss nicht vollständig geändert werden und ermöglicht dennoch die Auswertung feingranularer Daten. Die Report Engine kann ohne Anpassung weiter genutzt werden und auf die neuen Daten zugreifen. Von dem klassischen Vorgehen muss hingegen zur Bereitstellung der Daten abgewichen werden.

Zur dauerhaften Speicherung der Daten, wird ein Hadoop-System in die BI-Architektur integriert. Dieses ermöglicht die verteilte Speicherung und Verarbeitung der Daten im Cluster und wird laufend mit den operativen Daten befüllt. Schon hier können Aggregationsregeln die Datenmenge vermindern und aus den Daten Informationen abgeleitet werden. Dieser Schritt kann aber auch zunächst vernachlässigt werden.

Benötigt der Anwender nun bestimmte Zeitintervalle der Big Data-Kennzahlen, so kann er über organisatorische Maßnahmen die Bereitstellung in der APL verlangen. Nach einer vordefinierten Bereitstellungszeit, kann er die Daten zur Analyse oder Reporting nutzen. Dabei stehen verschiedene Wege der Bereitstellung der Daten zur Auswahl:

- Der Anwender informiert die IT-Abteilung, dass er die Daten in einer bestimmten Detaillierungstiefe und zeitlichen Intervallbreite benötigt. Durch das manuelle Starten des ETL-Jobs, werden die Daten aus dem Hadoop-Cluster in die APL geladen.
- Der Anwender hat selbst die Möglichkeit, durch gezieltes Starten des ETL-Jobs die gewünschten Daten zu laden.
- Bei der Abfrage der Daten schränkt der Anwender seine Quelle geeignet ein. Die Abfrage ermittelt automatisiert den Bedarf der Daten und gleicht diesen ab. Werden weitere Daten benötigt, so werden diese zunächst nachgeladen, bevor die Anfrage ausgeführt und das Ergebnis zurückgegeben werden kann.

3.3.3.2 Datenbankintegration

Beispielhaft werden die Systeme SAP HANA und Oracle Database 11g untersucht, inwieweit sie die direkte Integration von Hadoop unterstützen. SAP HANA ist dazu geeignet als OLAP-Quelle in der BI-Architektur zu fungieren. Um externe Quellen wie ein Hadoop-Cluster zu integrieren steht die Lösung „SAP HANA smart data access“ zur Verfügung. In der aktuellen Version SAP HANA Support Package Stack (SPS) 6 unterstützt die Appliance die direkte Integration der Intel Distribution for Apache Hadoop in der Version 2.3, die Apache Hadoop 1.0.3 und

Apache Hive 0.9.0 beinhaltet. Darüber hinaus können auch weitere Datenbanken der Hersteller Teradata und Sybase eingebunden werden.

Die Integration dieser Datenbanken, insbesondere Apache Hadoop, ermöglicht einen direkten Zugriff auf die Daten mittels SQL-Skripten. Dadurch ist es möglich, auf die Datenbestände in den Hive-Tabellen lesend zuzugreifen und Datenbestände in SAP HANA zu laden.

SAP verspricht sich durch die Anbindung von Hadoop an die HANA-Appliance einen kosteneffizienten Datenspeicher, mit dem große Volumen an strukturierten, semi-struktuierten und unstrukturierten Daten verarbeitet werden können. Besonders werden Vorteile in der Batch-Verarbeitung, bei komplexen Abläufen, eine gute Archivierung der Daten und die Möglichkeit Rohdaten vorzuhalten hervorgehoben (Vgl. Gardella 2013b).

Viele BI-Architekturen basieren zur Datenhaltung auf Produkten von Oracle. Gartner zählt Oracle wie SAP zu den führenden Herstellern im Bereich Business Intelligence und Analytics Plattformen (Vgl. Gartner, Inc. 2013b). Um auf Daten aus Hadoop-Clustern zugreifen zu können bietet Oracle verschiedene Möglichkeiten. Zwei davon sind sind der Oracle Loader for Hadoop und der Oracle SQL Connector for HDFS aus der Suite der Oracle Big Data Connectors. Dabei geht Oracle stets davon aus, dass die nicht-relationale Datenbank Hive die Daten im Hadoop-Cluster vorhält.

Mit dem Oracle Loader for Hadoop können Daten aus einem Hadoop-Cluster sowohl im Online- als auch im Offline-Modus geladen werden. Im Online-Modus werden die Daten mittels JDBC oder OCI in die Datenbank geladen. Dagegen bietet der Offline-Modus eine höhere Performanz durch die Verwendung des Oracle Data Pump-Treibers. Zunächst werden die Daten aus dem Hadoop System extrahiert und im Hadoop-Cluster als CSV-Dateien gespeichert. Dabei können die Daten in beiden Modi transformiert, aggregiert und sortiert werden. Nach dem Transfer auf das Datenbanksystem können die Dateien mit dem SQL-Loader eingelesen werden (Vgl. Oracle Corporation 2013c).

Im Fall von Oracle SQL Connector for HDFS erstellt Oracle sogenannte externe Tabellen zum direkten Datenzugriff via SQL. Es gibt drei verschiedene Möglichkeiten zur Einbindung. Es können zum einen Data Pump-Dateien aus einem HDFS eingebunden werden, zum anderen können auch direkt Hive-Tabellen integriert werden oder CSV-Dateien aus einem HDFS (Vgl. Oracle Corporation 2013a).

3.3.3.3 Integration via ETL-Tool

Neben der direkten Einbindung der Hadoop-Datenquelle in die Datenbank, ist eine Verknüpfung über ein ETL-Tool möglich. Dabei muss das ETL-Tool das Hadoop-Cluster als Datenquelle einbinden und die Daten zielgenau filtern können.
SAP bietet im Rahmen des Big Data-Bundles verschiedene Möglichkeiten zur Integration an. Mithilfe der SAP BusinessObjects BI suite und dem SAP BusinessObjects Data Services Designer kann Hadoop per Drag & Drop in einen Extraktionsprozess eingebunden werden (Vgl.

Abb. 3.8). Abhängig von der Datenauswahl wird die Anfrage in die Hive Query Language (HQL) zur Abfrage von Hive Tabellen oder in ein Pig-Skript zur Abfrage von Dateien im HDFS übersetzt. Wie gewohnt werden diese Abfragen für Hadoop in MapReduce-Jobs konvertiert. Die Ergebnisse werden in Dateien auf dem HDFS gespeichert. SAP Data Services liest diese mit mehreren Prozessen aus. Mit diesen Daten kann nun der nächste definierte Schritt, beispielsweise Datenqualitätsaufgaben oder direkt das Laden der Ergebnisse in eine Datenbank gestartet werden.

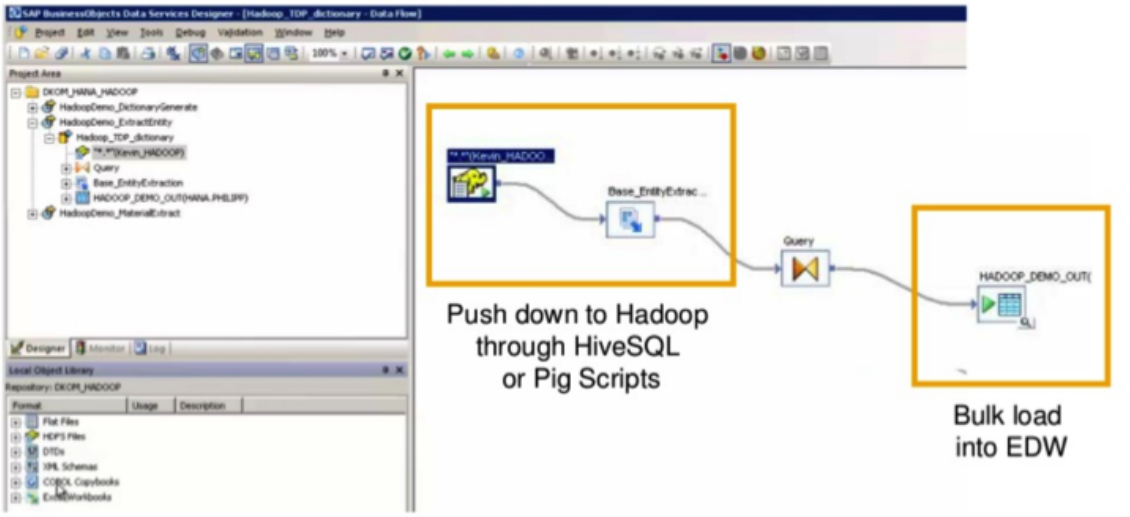


Abbildung 3.8: Ansicht des SAP BO Data Services Designer
Quelle: Gardella (2013a, Folie 19)

Mit dem Informatica Power Center in der Big Data Edition ist es in ähnlicher Weise möglich, Hadoop in einen bestehenden ETL-Prozess zu integrieren. Die Anwendung Informatica Developer bietet dazu ebenfalls eine Oberfläche, die zur Definition des ETL-Jobs dient und die Hadoop-Datenquelle mit einbeziehen kann (Vgl. Abb. 3.9).

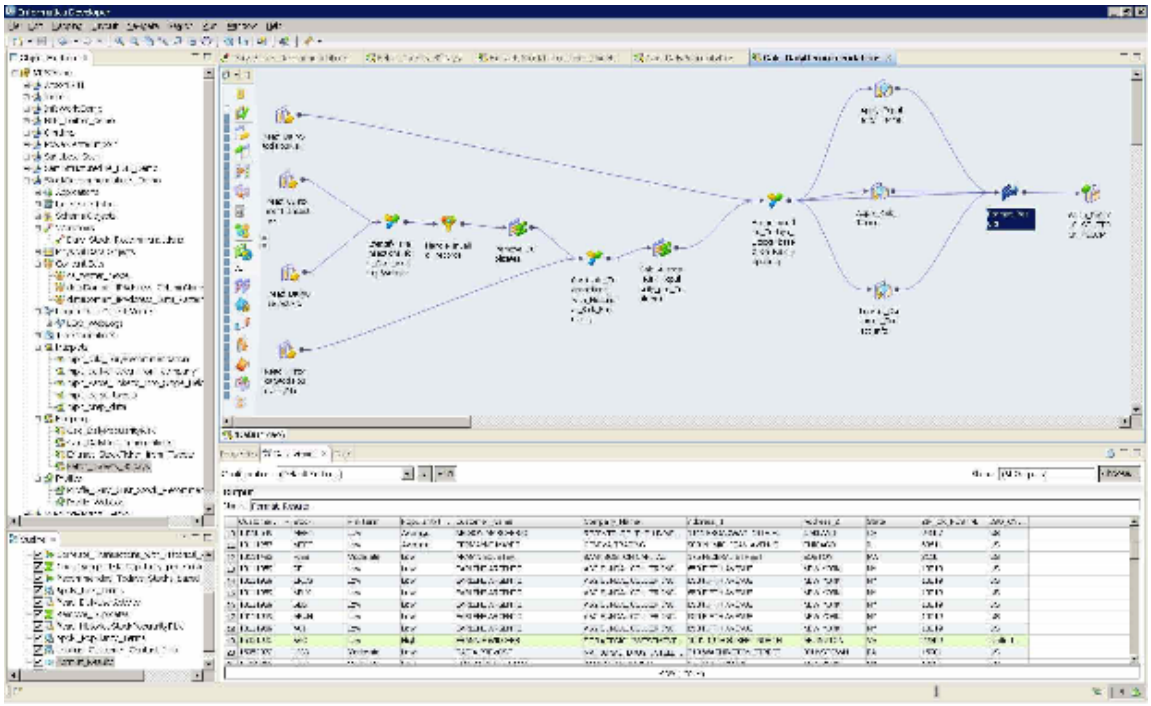


Abbildung 3.9: Ansicht des Informatica Power Center Developer
Quelle: Informatica (2013c)

3.3.3.4 Bereitstellung der Daten

Damit die gewünschten Daten bei Bedarf adhoc in die Anwendungsschicht geladen werden können, müssen sie zunächst aus den Quellsystemen extrahiert und im Hadoop-System bereitgestellt werden. Dieses Szenario ermöglicht eine Vorverarbeitung und Aggregation der Daten in dieser Stufe. Die Daten können im Hadoop-System als Dateien im HDFS hinterlegt werden oder wie logische Entitäten in Tabellen in der Erweiterung Hive, die eine nicht-relationale Datenbank auf der Basis von Hadoop implementiert.

Bis zu dieser Stufe ist eine automatisierte Verarbeitung möglich. Die Daten können via kontinuierlichem Datenfluss integriert werden oder in regelmäßigem Turnus beispielsweise nach Geschäftsschluss per Batch-Verarbeitung. Diese Vorverarbeitung sorgt für die grundlegende Bereitstellung der Daten und kann die Geschäftsdaten um weitere Informationen aus anderen Datenquellen anreichern und transformieren.

Auf Anfrage wird ein Ausschnitt der benötigten Daten in die Anwendungsschicht geladen und steht nach Abschluss des Ladeprozesses für Analysen zur Verfügung. Dadurch ergeben sich verschiedene Anforderungen an den Datenbestand im Hadoop-Cluster:

- Hadoop beinhaltet alle eventuell benötigten zeitlichen Ausschnitte der Daten.
- Die Daten liegen historisiert und redundant vor.
- Das System kann auf spontane Anfragen die gewünschten Daten bereitstellen und

gleichzeitig die Ladeprozesse aus den Quellsystemen durchführen.

Sind die Daten in die Anwendungsschicht geladen worden, muss der Anwender entweder die Daten erhalten oder über die Bereitstellung informiert werden. Abhängig von der Implementierung sind dazu organisatorische Prozessvereinbarungen oder technische Abläufe von Nöten.

Dagegen ist eine manuelle Bereitstellung der Daten relativ einfach möglich. Üblicherweise verfügt der Anwender nicht über Zugriff auf das ETL-Tool in der produktiven Umgebung. Somit sind Prozeduren und Abläufe zu etablieren, die unter Angabe der gewünschten Ausschnitte den ETL-Job starten. Diese Tätigkeit kann zum Beispiel durch die IT-Abteilung oder zentrale Ablaufsteuerung durchgeführt werden.

3.3.3.5 Procedures

Eine mögliche Umsetzung des Szenarios B ist die Integration des neuen Prozesses in eine bestehende BI-Architektur. Die bestehende Architektur basiert auf einer Oracle Datenbank (mindestens in der Version 11g erforderlich) zur Datenhaltung. Es sind die verschiedenen Schichten ODS, Staging, DWH und APL definiert. Die APL-Schicht ist nach der dritten Normalform modelliert und stellt die Daten in einem Star-Schema dar. Die Schlüsselfelder und Hierarchiestufenschlüssel sind mit Bitmap-Indizes versehen. Zentrale Bestandteile des Schemas sind die Faktentabellen und die Zeitdimension, die mit jeder Faktentabelle verknüpft ist. Eine Reporting Engine stellt die Anfragen per SQL über eine Standard ODBC-Schnittstelle an die APL-Schicht und visualisiert die Ergebnisse in beliebiger Tabellen- oder Diagrammform.

Über die aus Szenario A bekannte Lösung hinaus geht nun die Bereitstellung der Daten. Die Annahme, die zu diesem Szenario B führt ist, dass es aus Gründen des Speicherplatzmangels und den zu langen Laufzeiten von Abfragen nicht möglich bzw. zweckmäßig ist, sämtliche reportingrelevante Daten in der APL-Schicht vorzuhalten. Von einer vorgeschalteten Aggregation wird an dieser Stelle abgesehen, da die Daten feingranular für die Auswertung benötigt werden.

Im Geschäftsmodell wird für diese Daten festgehalten, dass im Normalfall lediglich die letzten n Zeitscheiben t_n ausgehend vom aktuellsten Datum y vorgehalten werden und für die direkte Analyse bereitstehen. Dies bedeutet, dass die APL-Schicht die Zeitscheiben $t_{y-n}..t_y$ vorhält. Die Daten, die vor dem Zeitpunkt t_{y-n} liegen, also $t_0..t_{y-n}$, sind im Hadoop-Cluster gespeichert. Das Hadoop-Cluster kann mittels SQL über eine direkte Integration der Datenquelle in die APL-Schicht angesprochen werden und stellt eine langsamere, aber mit deutlich mehr Speicherplatz versehene Speicherlösung dar.

Wenn nun eine Abfrage Daten aus Zeitscheiben benötigt, die sich nicht in dem schnellen Speicherbereich der APL-Schicht befinden, aber über den Hadoop-Connector angesprochen werden können, so können diese zunächst aus dem Hadoop-Cluster in die schnelle APL-Schicht befördert werden. Reicht der aktuelle Speicherplatz nicht aus, so können nicht benötigte Zeitscheiben entfernt werden, da diese weiterhin im Hadoop-Cluster vorliegen. Sind die Zeitscheiben nachgeladen worden, so kann die Abfrage auf der APL-Schicht durchgeführt werden.

Der Schritt des Nachladens kann adhoc mittels Prozeduraufruf durch den Benutzer erfolgen, allerdings kann dies auch automatisiert geschehen. Moderne BI-Suiten wie z.B. IBM Cognos können als Datenquelle auch so genannte Stored Procedures ansprechen (Vgl. Volitich 2008, S. 471). Diese Stored Procedure gibt eine Referenz auf einen erstellten Cursor zurück, der die Abfrage darstellt. Zuvor kann die Prozedur die benötigten Zeitscheiben nachgeladen und bereitgestellt haben. Die Abbildung 3.10 stellt die Vorgehensweise bei einer Anfrage an eine Stored Procedure mit den oben beschriebenen Aufgaben dar.

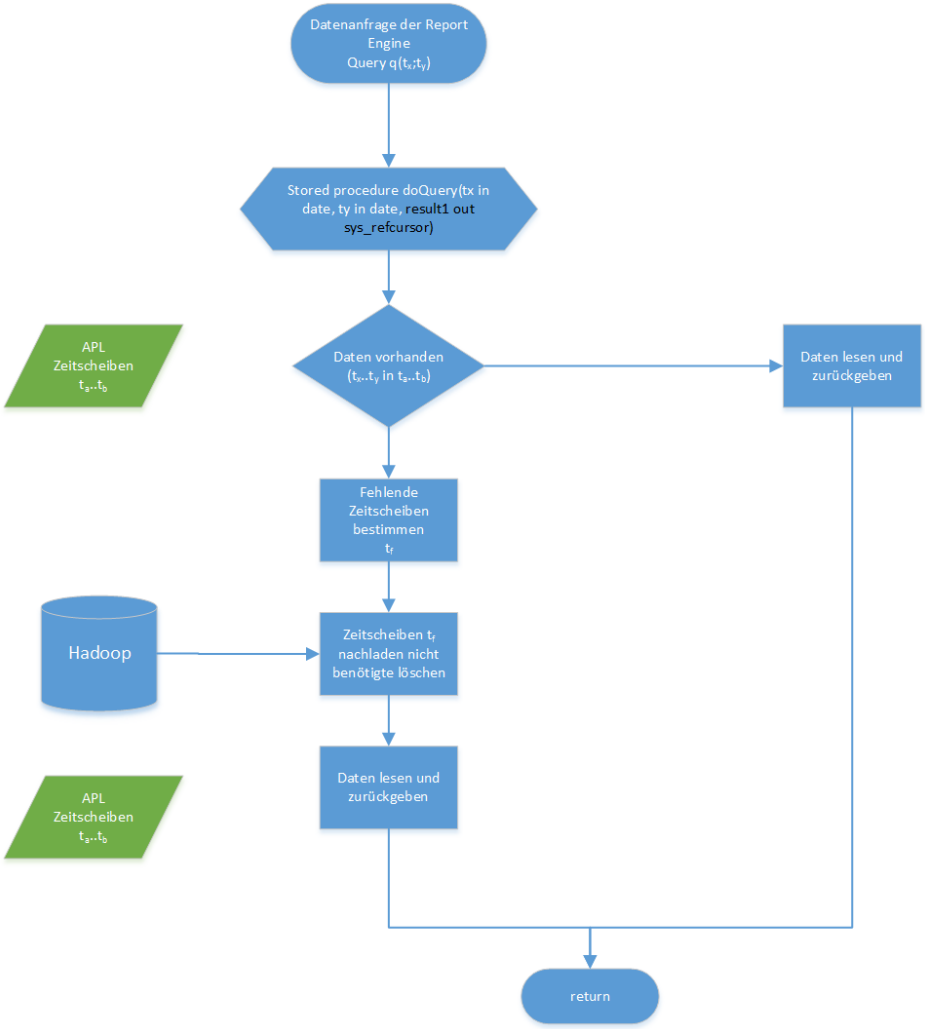


Abbildung 3.10: Referenzarchitektur - Szenario B - Daten nachladen

Deutlicher wird der Datenstand der APL-Schicht durch die nachfolgenden Grafiken. Die blauen Balken v stellen die vorhandenen Zeitscheiben in der APL-Schicht dar. Die hier nicht dargestellten Zeitscheiben zu den weiteren Zeitpunkten $t_0..t_y$ sind in dem Hadoop-Cluster vorhanden und können nachgeladen werden.

Zunächst sind alle Zeitscheiben vorhanden und die für die erste Abfrage q_1 benötigten Zeitscheiben sind in der APL-Schicht vorhanden, sodass diese direkt beantwortet werden kann. Dagegen

erfordert die zweite Abfrage q_2 weitere Zeitscheiben, nämlich $t_5..t_{12}$, wovon $t_{10}..t_{12}$ nicht in die APL-Schicht geladen worden sind. Diese Tatsache ist dem Benutzer allerdings nicht bekannt, da ihm nur eine logische Datenquelle bekannt ist. Abbildung 3.11 zeigt den Datenstand der APL-Schicht zum Zeitpunkt der Abfrage q_1.

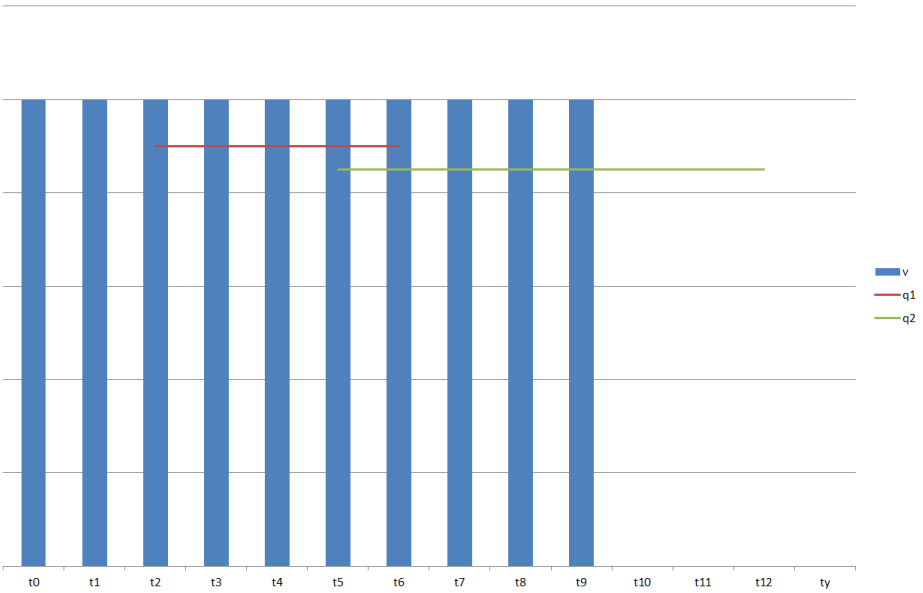


Abbildung 3.11: Referenzarchitektur - Szenario B - Datenstand der APL zu Beginn

Diese logische Datenquelle wird durch die Stored Procedure dargestellt, die nun anhand der Parameter der Abfrage feststellt, dass Zeitscheiben nachzuladen sind und importiert diese aus dem Hadoop-Cluster und entfernt die nicht benötigten Zeitscheiben, nämlich $t_0..t_4$. Anschließend wird die Abfrage durchgeführt und die Stored Procedure gibt das Ergebnis als Referenz auf den Cursor zurück.

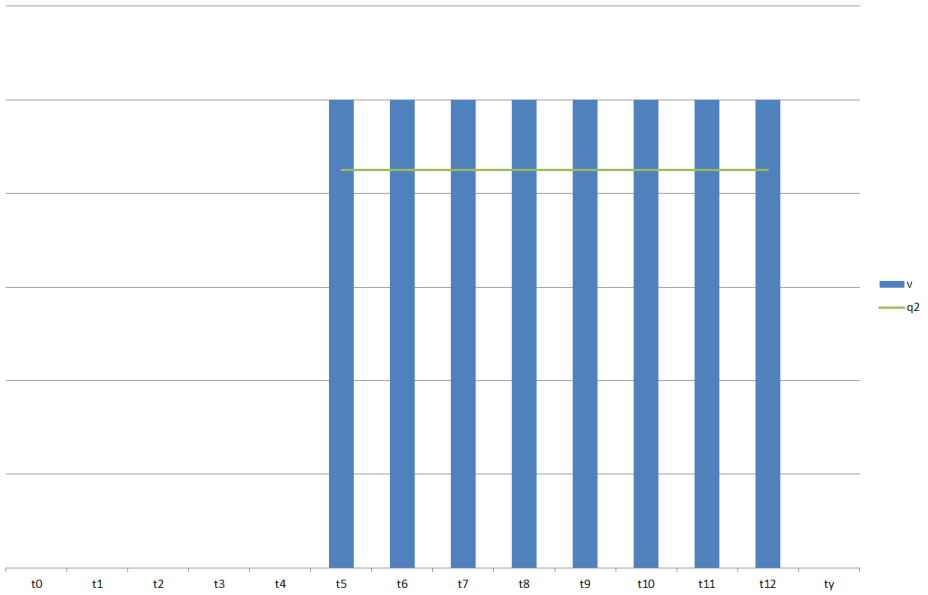


Abbildung 3.12: Referenzarchitektur - Szenario B - Datenstand nach der zweiten Abfrage

Nach der Ausführung beinhaltet die APL-Schicht die Zeitscheiben $t_5..t_{12}$ (Vgl. Abb. 3.12).

3.3.4 Szenario C

Das Ziel des in diesem Abschnitt beschriebenen Szenarios besteht darin, eine Referenzarchitektur zur Verarbeitung von Big Data bereitzustellen. Dabei sollen die in diesem Kontext relevanten Kriterien hinsichtlich Volume (Datenvolumen), Variety (Datenvielfalt) und Velocity (Geschwindigkeit) fokussiert und mithilfe bestehender Softwarelösungen umgesetzt werden.

3.3.4.1 Architektur

Im Unterschied zu den bisher vorgestellten Ansätzen, wird hierbei allerdings keine separate Anwendungsschicht (APL) vorgesehen. Stattdessen können alle Daten, die für das Reporting benötigt werden, über eine Schnittstelle (z.B. ODBC) direkt von einer Datenquelle auf der Basis eines Hadoop-Clusters, der in diesem Szenario das Zielsystem darstellt, abgefragt und analysiert werden (Vgl. Abb. 3.13). Die Daten werden somit entweder direkt im Hadoop-eigenen Filesystem (HDFS) oder in einer darauf aufbauenden Datenbank - wie zum Beispiel Apache HBase - gespeichert.

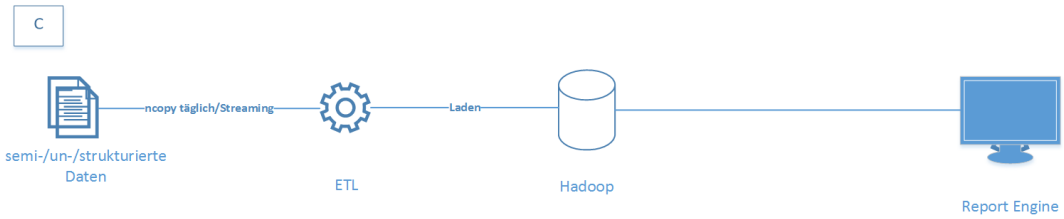


Abbildung 3.13: Referenzarchitektur - Szenario C

Durch diese Zentralisierung der Daten, können Big-Data-Kennzahlen ad-hoc ausgewertet werden - ohne dass ein aufwendiges Nachladen - wie beispielsweise in Szenario B - erforderlich ist. Da die Daten des Weiteren mithilfe von Hadoop gespeichert werden, liegen diese in redundanter Form auf mehreren Nodes vor, sodass stets eine hohe Verfügbarkeit gewährleistet wird. Durch die Elimination der Anwendungsschicht, entsteht außerdem ein Kostenvorteil, da zum einen Investitionskosten für zusätzliche Hardware entfallen und zum anderen der administrative Aufwand geringer ist. Ein Teil der Kostenersparnis wird jedoch in die Anschaffung der zusätzlich benötigten Server-Kapazitäten des Hadoop-Clusters fließen.

3.3.4.2 Prozessübersicht

Ausgangspunkt des Prozesses sind die für das Reporting benötigten Daten, die in semistrukturierter, unstrukturierter oder strukturierter Form und in einem bestimmten Format (z.B. CSV) vorliegen. Nach der Extraktion der Daten aus den Quellsystemen, können diese entweder über einen automatisierten Vorgang in regelmäßigen Zeitintervallen direkt in das HDFS kopiert werden oder über eine Streaming-Lösung - wie zum Beispiel Apache Flume (Vgl. Hortonworks 2013c) - in den Hadoop-Cluster geladen werden.

Bezüglich des Transformationsprozesses lassen sich grundsätzlich zwei Ansätze unterscheiden: Entweder werden die zu verarbeitenden Daten erst auf dem Zielsystem oder bereits vorher - während der Übertragung in das Zielsystem („on-the-fly“) - transformiert. Für beide Ansätze finden sich im Hadoop-Umfeld bereits entsprechende Lösungen.

So bietet der Softwarehersteller Cloudera mit Cloudera Morphlines ein Open Source Framework an, das es erlaubt, Daten jeglicher Art und aus verschieden Quellen zunächst im Hauptspeicher zu transformieren, bevor diese in Hadoop eingelesen werden. Hierbei kann der Anwender auf eine Reihe von Befehlen zurückgreifen, mit denen beispielsweise einzelne Felder aus einem Access Log ausgelesen und weiterverarbeitet werden können. Alle gewünschten Transformationsschritte werden dabei in einer zentralen Konfigurationsdatei (morphline.conf) definiert, die dann über ein Java-Programm und mithilfe der von Cloudera bereitgestellten API (com.cloudera.cdk.morphline.api) ausgeführt wird (Vgl. Cloudera 2013e).

Soll die Transformation dagegen - wie im ersten Ansatz - direkt auf einem Hadoop-Cluster erfolgen, kann dies mit der Erweiterung Apache Pig realisiert werden. Mithilfe von Apache Pig können komplexe MapReduce Jobs u.a. zur Datentransformation über eine Skriptsprache (Pig Latin) definiert und ausgeführt werden. Dazu werden diverse SQL-ähnliche Befehle bereitgestellt, durch die Daten beispielsweise aggregiert und sortiert werden können. Des Weiteren kann Apache Pig aber auch dazu verwendet werden, um Daten u.a. aus einer CSV-Datei in ein relationales Schema zu überführen und in Hadoop zu speichern (Vgl. Hortonworks 2013g). Als Alternative zu Apache Pig sei an dieser Stelle noch Apache Hive erwähnt, mit der sich ein komplettes Data Warehouse auf Basis eines Hadoop-Cluster realisieren lässt (Vgl. Hortonworks 2013e).

Am Ende des in diesem Abschnitt beschriebenen Prozesses, steht die Report Engine. Da

die gängigsten Hadoop-Distributionen (dazu gehören u.a. Cloudera und Hortonworks) jedoch diverse Standard-Schnittstellen wie ODBC und JDBC zur Verfügung stellen und die Anbindung somit trivial ist, wird hier nicht näher darauf eingegangen. Ein später folgendes Kapitel wird dieses Thema jedoch nochmals aufgreifen und beispielhaft darlegen, wie das Zusammenspiel zwischen einer konkreten Report Engine und Apache Hadoop aussehen kann.

3.3.5 Besondere Herausforderungen

Im folgenden Kapitel wird noch einmal speziell auf die Herausforderungen an eine Architektur zur Verarbeitung von Big Data, anhand der zwei Charakteristika Variety sowie Velocity eingegangen.

3.3.5.1 Variety

Um ein fachbereichs- und anwendungsübergreifendes Reporting ermöglichen zu können, müssen Daten nicht nur von einem, sondern gleich von mehreren Quellsystemen in die Zielarchitekturen (siehe oben) überführt werden. Dabei liegen die zu integrierenden Daten in unterschiedlichen Formaten vor und können strukturiert, semistrukturiert oder gar unstrukturiert sein. Diese Begebenheit wird im Kontext von Big Data auch als Variety bezeichnet und stellt eine besondere Herausforderung bei der Integration von Daten dar.

Strukturierte, semistrukturierte und unstrukturierte Daten

Daten, deren Struktur durch ein bestimmtes Schema vorgegeben wird, werden als strukturierte Daten bezeichnet. Ein klassisches Beispiel dafür sind operative Daten, die in einer relationalen Datenbank vorliegen. Da hier u.a. Objekte aus der Realität in Form von Tabellen mit entsprechenden Spaltenbezeichnungen gespeichert werden, lässt sich außerdem eine Aussage über die Bedeutung der Daten treffen. Weist beispielsweise eine Tabelle die Spalten ProduktNr, ProduktBez und ProduktArt auf, so ist es naheliegend, dass es sich hierbei um eine Relation handelt, in der die Produkte eines Unternehmens gespeichert werden. Über die einzelnen Spalten können dann bestimmte Attribute eines Produktes in Erfahrung gebracht werden (Vgl. Leser 2007, S. 17).

Auch semistrukturierte Daten sind nach einem Schema strukturiert. Im Unterschied zu strukturierten Daten kann die zugrunde liegende Struktur jedoch für verschiedene Datensätze variieren. Als Beispiel sind an dieser Stelle XML-Dokumente ohne begleitendes XML-Schema zu nennen: In XML-Dokumenten werden Daten mithilfe sogenannter Tags, die die Attribute eines bestimmten Objektes spezifizieren, strukturiert. Im Gegensatz zu einer relationalen Datenbank müssen aber nicht immer alle Attribute vorhanden sein. Ferner können diese sogar innerhalb eines XML-Dokuments beliebig geschachtelt werden. Wenn also - basierend auf dem oben eingeführten Beispiel - nur zu fünf von 1.000 Produkten die Produktart bekannt ist, so erlaubt es XML, dass nur für diese fünf Datensätze ein entsprechender Tag vorgesehen werden muss. Durch semistrukturierte Daten lässt sich so zwar eine flexiblere Modellierung realisieren, jedoch wird durch das Fehlen einer festen Struktur die Verarbeitung der Daten erschwert. Dies begründet sich darin, dass für ein und dasselbe Attribut unterschiedliche Bezeichnungen definiert werden

können, was beispielsweise zu fehlerhaften Werten bei der Aggregation von Daten führen kann (Vgl. Leser 2007, S. 28).

Als besonders schwierig erweist sich die automatisierte Verarbeitung und Analyse von Texten, die beispielsweise in Form von Briefen, E-Mails, Berichten, Artikeln oder Dokumenten vorliegen. Da diese Texte kein Datenmodell und kein maschinenlesbares Schema aufweisen, wird in diesem Kontext von unstrukturierten Daten gesprochen. Eine erste Annäherung zur Verarbeitung unstrukturierter Daten ist das sogenannte Text Mining. Hierbei kommen Methoden der maschinellen Sprachverarbeitung und des maschinellen Lernens zum Einsatz, die es beispielsweise ermöglichen bestimmte Entitäten und deren Beziehungen zueinander in Textdokumenten zu identifizieren (Vgl. Leser 2007, S. 29).

Um strukturierte und semi-strukturierte Daten aus verschiedenen Quellsystemen zentralisiert verarbeiten zu können, bedarf es bestimmter Methoden und Techniken der Informationsintegration. Diese sind Inhalt des folgenden Abschnitts.

Schemamanagement

Ein wesentliches Problem bei der Integration von strukturierten und semistrukturierten Daten aus mehreren unterschiedlichen Quellsystemen besteht in der sogenannten strukturellen Heterogenität. Diese ist dadurch bedingt, dass jedes Quellsystem seine Daten in einem anderen Schema speichert. Um dem entgegenzuwirken, bedient man sich des Schemamanagements, das diverse Techniken zur Erkennung und Überwindung struktureller Heterogenität auf Schemaebene zusammenfasst.

Zu diesen Techniken gehören u.a. die Schemaintegration und das Schema Mapping. Diese unterscheiden sich folgendermaßen: Während die Schemaintegration die Daten aus den verschiedenen Schemata in ein neues (globales) Schema überführt, integriert das Schema Mapping die Daten über einen Transformationsprozess in ein bestehendes Schema. Als Basis beider Ansätze gilt das Konzept der Korrespondenz. Mithilfe der Korrespondenz werden semantische Zusammenhänge zwischen Elementen verschiedener Schemata, die in ein neues globales bzw. in ein bestehendes Schema integriert werden sollen, beschrieben. Die Ermittlung dieser Zusammenhänge erfolgt entweder manuell oder semiautomatisch durch das sogenannte Schema Matching, das im Folgenden näher betrachtet wird (Vgl. Leser 2007, S. 116).

Schema Matching

Aufgrund der hohen Komplexität sind dem manuellen Auffinden von Schema-Korrespondenzen gewisse Grenzen gesetzt. Auch wenn dabei meist Experten mit detaillierten Kenntnissen der Anwendungsdomäne zum Einsatz kommen, sind falsche und fehlende Korrespondenzen nicht auszuschließen. Vor allem dann, wenn besonders große Schemata mit einer Vielzahl an Relationen und Attributen miteinander in Verbindung gebracht werden müssen, die zudem noch in mehrfacher Beziehung zueinander stehen (Vgl. Leser 2007, S. 143).

Abhilfe schafft hier das sogenannte Schema Matching, das eine Reihe an Verfahren zur automatisierten Erkennung von Korrespondenzen zwischen semantisch äquivalenten Elementen zweier

Schemata erlaubt. Ganz ohne manuelles Zutun kommt allerdings auch das Schema Matching nicht aus, da die generierten Korrespondenzen als Vorschläge zu werten sind, die abschließend von einem Domänenexperten verifiziert werden müssen. Insofern ist der Gesamtprozess des Schema Matching als semiautomatisch anzusehen.

Das grundlegende Prinzip des Schema Matching lässt sich wie folgt erklären: Zunächst werden alle Attribute des Quellschemas mit den Attributen des Zielschemas unter Zuhilfenahme eines Ähnlichkeitsmaßes verglichen. Anhand eines zuvor definierten Schwellwerts, werden dann die Attribute, die eine hinreichende Ähnlichkeit aufweisen, dem Domänenexperten als Korrespondenz vorgeschlagen. Welches Ähnlichkeitsmaß bei der initialen Schemaanalyse zugrunde gelegt wird, ist abhängig von der gewählten Matching-Methode.

Dabei lassen sich zunächst einmal individuelle und kombinierte Ansätze bzw. Methoden unterscheiden. So sind individuelle Methoden dadurch gekennzeichnet, dass lediglich ein einziges Ähnlichkeitsmaß verwendet wird. Wohingegen kombinierte Ansätze den Vergleich zwischen Elementen des Quell- und Zielschemas anhand mehrerer Kriterien zur Ähnlichkeit vornehmen. Hybride und zusammengesetzte Ansätze sind Spezialisierungen und somit konkrete Ausprägungen des kombinierten Ansatzes. Wobei ersterer mehrere Ähnlichkeitsmaße in ein einziges Maß integriert und letzterer Ergebnisse unabhängiger Matching-Methoden miteinander verknüpft.

Im Kontext der individuellen Ansätze lassen sich ferner schemabasierte und instanzbasierte Verfahren voneinander differenzieren. Diese unterscheiden sich insofern, dass bei der schemabasierten Methode die Elemente eines Schemas sowie deren Eigenschaften (Name, Datentyp, Primär- und Fremdschlüsselbeziehungen etc.) betrachtet werden. Während bei dem instanzbasierten Verfahren die gespeicherten Daten analysiert werden. Linguistische Methoden, constraintbasierte Methoden und duplikatbasierte Methoden sind konkrete Ausprägungen der schemabasierten bzw. instanzbasierten Methode. Auf diese wird an dieser Stelle allerdings nicht weiter eingegangen (Vgl. Leser 2007, S. 145f.).

Eine Übersicht aller Schema-Matching-Methoden kann der folgenden Abbildung 3.14 entnommen werden.

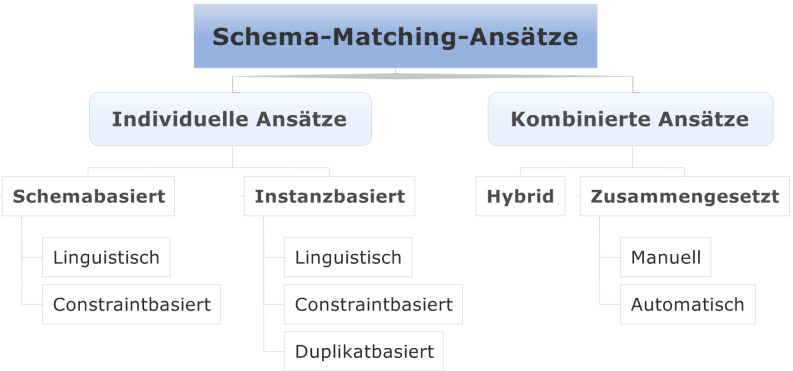


Abbildung 3.14: Klassifikation von Schema-Matching-Methoden
Quelle: Eigene Darstellung in Anlehnung an Leser (2007, S. 146)

3.3.5.2 Velocity

Hinter einem der drei Big Data charakterisierenden V's verbirgt sich der Begriff Velocity. Dieser steht für die Geschwindigkeit der Datenverarbeitung und ist damit eine der Kerneigenschaften von Big Data. Unterscheiden lässt sich hierbei die Geschwindigkeit der Datengenerierung bzw. Datenübertragung sowie der Verarbeitung im Zielsystem.

3.3.5.2.1 Datengenerierung und -übertragung

Zur Beschleunigung der Datengenerierung und anschließenden Übertragung gibt es eine Reihe von Möglichkeiten. Zuerst lassen sich in diesem Kontext jedoch grob zwei Arten von Daten unterscheiden. Bei diesen zwei Arten handelt es sich um Realtime-Daten (Echtzeitdaten) sowie Near-Realtime-Daten (nahe Echtzeitdaten). Wichtig ist hierbei die exakte Definition dieser Begriffe. Diese Definition muss für jedes Szenario vorgenommen werden, da in manchen Anwendungsfällen eine minütliche Übertragung der Daten als Realtime gilt, in anderen Fällen wiederum nur die Übertragung der tatsächlich in der letzten Sekunde generierten Daten.

Eine Option zur Beschleunigung der Datengenerierung und -übertragung ist das sogenannte Streaming. Hierbei werden die Daten direkt aus dem Quellsystem mit Hilfe eines konstanten Datenstroms in das Zielsystem übertragen. Damit sich der Einsatz einer solchen Streaming-Lösung jedoch lohnt, sollten die hierbei übertragenen Datenmengen auch ausreichend groß sein. Bei kleineren Datenmengen wären auch die einzelnen Pakete im Datenstrom dementsprechend klein und der Vorteil große Datenmengen schnell bewältigen zu können wäre obsolet. Ein weiterer Vorteil des Streamings ist das Übermitteln und Verarbeiten von echten Realtime Daten, da immer die grade generierten Daten in den Datenstrom gelangen und in das Zielsystem übertragen werden. Auch für den Fall, dass die generierten Datenmengen für die Speicherung im Zielsystem zu groß sind oder eine solche Speicherung gar nicht gewünscht ist, bietet das Streaming Abhilfe. In diesem Fall können die relevanten und gegebenenfalls aggregierten Daten aus dem Datenstrom gefiltert und im Zielsystem gespeichert werden. Die restlichen Daten werden anschließend nicht benötigt und somit besteht auch kein Bedarf diese zu speichern. Ein Praxisbeispiel für eine Streaming-Anwendung ist das Produkt InfoSphere Streams der Firma IBM. Mithilfe dessen lassen sich relationale Datenbanken mit einem konstanten Datenstrom befüllen (Vgl. Kap. 3.4.2.8). Auch für Hadoop-Systeme gibt es Streaming-Anwendungen. Ein Beispiel hierfür ist die Hadoop Erweiterung Apache Flume, welche es ermöglicht ein Hadoop-System permanent mit Hilfe eines Datenstroms mit Daten zu befüllen.

Eine weitere Möglichkeit zur Übertragung der Daten ist ein ETL-Job. Dieser extrahiert in festgelegten Intervallen die Daten aus dem Quellsystem und überträgt diese anschließend in das Zielsystem. Ein solcher ETL-Prozess bietet jedoch im Gegensatz zum Streaming keine Realtime-, sondern maximal Near-Realtime-Daten, da kein konstanter Datenfluss erzeugt wird. Für gewöhnlich besteht ein solcher Prozess aus den drei Schritten der Extraktion (Extraction), Transformation (Transform) sowie dem Laden (Load). Im Schritt der Extraktion werden die Daten zunächst aus dem Quellsystem extrahiert und in der Regel in einer Staging-Area oder

einem Operational Data Store (ODS) temporär zwischengespeichert. Anschließend erfolgt der Schritt der Transformation, hierbei werden die Daten beispielsweise auf ihre Integrität geprüft und aggregiert. Im letzten Schritt erfolgt dann das Laden in das Zielsystem, etwa in eine einfache Datenbank oder ein Data Warehouse (Vgl. Leser 2007). Der ETL-Prozess eignet sich somit sowohl für kleine als auch große Datenmengen. Ein Nachteil hierbei ist jedoch, dass durch das Fehlen eines konstanten Datenstroms keine echten Realtime Daten, wie beim Streaming, ermöglicht werden.

3.3.5.2.2 Datenverarbeitung

Neben der im vorherigen Abschnitt erläuterten Beschleunigung der Datengenerierung und Datenübertragung kann auch im Bereich der Datenverarbeitung die Geschwindigkeit verbessert werden. Hierbei gibt es unterschiedliche Ansätze, welche auch miteinander kombiniert werden können.

Die erste Möglichkeit ist der Einsatz von In-Memory-Technologien. Ein bekanntes Beispiel auf diesem Gebiet ist die In-Memory-Datenbank HANA der Firma SAP. Im Gegensatz zu herkömmlichen Datenbanksystemen, findet die Datenverarbeitung sowie das Datenmanagement beim In-Memory-Computing nicht auf der Festplatte, sondern direkt im Hauptspeicher, also In-Memory statt. Dies ermöglicht eine Verarbeitung von großen Datenmengen mit hoher Geschwindigkeit. Eine Besonderheit der In-Memory Datenbank SAP Hana ist die spaltenorientierte Arbeitsweise, eine herkömmliche Datenbank hingegen arbeitet zeilenorientiert. In HANA werden diese beiden Arbeitsweisen miteinander kombiniert und somit eine hohe Performance erreicht. Neben dem Wunsch große Datenmengen sehr schnell verarbeiten zu können ist ein weiteres Ziel von HANA das transaktionale OLTP- sowie das analytische OLAP-Konzept in einem System zu vereinen (Vgl. ManageIT 2012). Nachteile einer In-Memory-Datenbank sind die, im Gegensatz zu einer klassischen Datenbank, höheren Hardwarekosten sowie der Bedarf an speziellem Know-How. Des Weiteren gibt es, im Falle von SAP HANA, das gesamte Produkt nur in einer Appliance, also in einem Bundle bestehend aus Hard- und Software direkt vom Hersteller.

Die zweite Möglichkeit ist das sogenannte Clustering. Dahinter versteckt sich das Aufteilen (clustern) von Ressourcen auf mehrere Geräte, um so in der Summe eine höhere Leistung sowie eine bessere Lastverteilung zu erreichen. Ein Beispiel hierfür ist die Hadoop-Technologie der Apache Foundation, welche es ermöglicht die Verarbeitung von großen Datenmengen auf Cluster zu verteilen und somit die Verarbeitung zu beschleunigen. Hadoop basiert hierbei auf dem Google MapReduce-Algorithmus. Es gibt eine Vielzahl von Herstellern, welche kommerzielle Hadoop Varianten anbieten. Bekannte Anbieter sind beispielsweise Cloudera und Hortonworks (Vgl. Kap. 3.4) (Vgl. Fischer 2010).

3.3.6 Beispielhafte Umsetzung einer Referenzarchitektur

In diesem Kapitel wird die beispielhafte Umsetzung des Szenario C der Referenzarchitekturen beschrieben. Diese Beschreibung gliedert sich in die Abschnitte Zielsetzung, der Beschreibung des Vorgehens sowie der Realisierung einiger Reports auf Basis dieser Lösung.

3.3.6.1 Zielsetzung

Um eine Bewertung hinsichtlich der technischen Realisierbarkeit der bisher nur konzeptionell beschriebenen Referenzarchitektur aus dem Szenario C vornehmen zu können, wurde diese gemäß der folgenden Zielsetzung umgesetzt.

Mithilfe einer Hadoop-Distribution sollen konkrete Reports aus dem VW-Lastenheft beispielhaft umgesetzt werden. Dabei soll der zeitliche Aufwand für die Installation und Konfiguration möglichst gering ausfallen ($<$ 1 MT) und alle Auswertungen auf der Basis des von VW bereitgestellten Web-Access-Logs vom 12. Mai 2013 durchgeführt werden. Um auch Auswertungen hinsichtlich unterschiedlicher Use-Cases vornehmen zu können, soll die dafür notwendige Mapping-Tabelle aus dem VW-Lastenheft (Vgl. Peddinghaus u. a. 2011) ebenfalls in das Zielsystem integriert werden. Des Weiteren ist bei der Wahl der Hadoop-Distribution darauf zu achten, dass diese auf der verfügbaren Hardware (Notebook mit Vier-Kern-Prozessor und 8 Gigabyte RAM) lauffähig ist.

3.3.6.2 Vorgehen

Um den aus der Zielsetzung hervorgehenden Anforderungen hinsichtlich der Installation und Konfiguration gerecht zu werden, fiel die Wahl bzgl. der Hadoop-Distribution auf die Hortonworks Data Plattform. Diese ist als sog. Sandboxsystem in Form einer virtuellen Maschine - sowohl für VM-Ware als auch für Oracle VM VirtualBox - frei verfügbar und zeichnet sich durch eine einfache Bedienbarkeit aus, da essentielle Hadoop-Funktionen und -Erweiterungen wie z.B. Apache Pig und Apache Hive über eine Web-Oberfläche bereitgestellt werden.

Nachdem die Wahl des Zielsystems feststand, wurde die Zielsetzung insofern konkretisiert, dass drei verschiedene Reports aus dem VW-Lastenheft zur Umsetzung ausgewählt wurden. Somit sollte die Anzahl der Zugriffe pro Stunde, die Anzahl der Zugriffe je Use-Case und der Workload je Use-Case ausgewertet werden. Bevor aber die Quelldateien (Access Log und Mapping-Tabelle) in das Zielsystem geladen und dort für das Reporting entsprechend aufbereitet wurden, wurde zunächst ein Domänenmodell konzipiert, das die benötigten Entitäten und deren Beziehungen zueinander identifiziert.

So wird beispielsweise ein bestimmter Use-Case indirekt durch eine bzw. durch mehrere Call URLs aus dem Access Log repräsentiert: Eine Call URL referenziert einen Kennzeichner der Entität Qualifier, der seinerseits genau einem Use-Case zugewiesen wird. Ist jedoch für eine bestimmte Call URL kein entsprechender Kennzeichner vorhanden, kann auch keine Zuweisung zum Use-Case erfolgen (siehe Abbildung 3.15). Hinsichtlich der Auswertungen der Anzahl der

Zugriffe pro Stunde und des Workload je Use-Case werden des Weiteren die Entitäten bzw. Felder Timestamp und Payload aus dem Access Log benötigt.

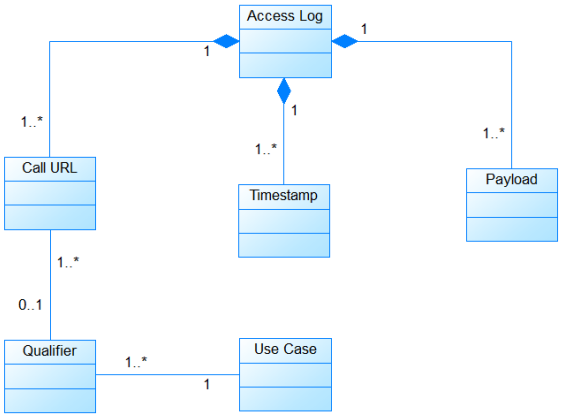


Abbildung 3.15: Domänenmodell zur Access-Log-Auswertung

Die eigentliche Umsetzung in Hortonworks erfolgte dann wie folgt: Zunächst wurde die von VW bereitgestellte Access-Log-Datei, in der die einzelnen Werte durch ein Leerzeichen getrennt sind, eingelesen und über die Import-Funktion des sog. HCatalog in ein relationales Schema überführt. Über HiveQL wurde dann eine Abfrage (Vgl. Abb. 3.16) formuliert, mit der die Call URLs aller Get-Requests inklusive Zeitstempel (Timestamp) und Größe (Payload) aus dem Access Log extrahiert, transformiert und in einer Tabelle gespeichert wurden.

```
1  insert overwrite table t_vw_getrequests_ts_payload
2  select
3  regexp_extract(regexp_extract(callurl,'(.*)HTTP\\/(.*)', 1),'GET(.*)', 1),
4  from_unixtime(unix_timestamp(regexp_extract(entrydate,'\\[(.*)', 1),'dd/MMM/yyyy:HH:mm:ss')),
5  cast(datasize as int)
6  from t_vw_logs
7  where regexp_extract(regexp_extract(callurl,'(.*)HTTP\\/(.*)', 1),'GET(.*)', 1) NOT LIKE '';
```

Abbildung 3.16: Extraktion und Transformation der Get-Requests

Im nächsten Schritt wurden die Mapping-Informationen zwecks Use-Case-Zuordnung aus dem Lastenheft in eine CSV-Datei überführt und anschließend ebenfalls über den HCatalog importiert. Somit war es möglich, die Anzahl der Zugriffe je Use Case mithilfe der folgenden HiveQL-Abfrage zu ermitteln (Vgl. Abb. 3.17).

```
1  insert overwrite table t_vw_anzusecases
2  select b.usecase, count(a.callurl) anzahl
3  from t_vw_getrequests_ts_payload a
4  join t_vw_qualifier_to_uc b ON (a.callurl = b.nqualifier)
5  group BY b.usecase
6  order BY anzahl desc;
```

Abbildung 3.17: HiveQL Abfrage zur Ermittlung der Anzahl der Zugriffe je Use Case

Für die anderen beiden Auswertungen (Anzahl der Zugriffe pro Stunde und Workload je Use-Case) wurden ebenfalls entsprechende HiveQL-Abfragen formuliert.

Nachdem also die Daten aus dem Access Log wie oben beschrieben aufbereitet und in Tabellen gespeichert wurden, können diese mithilfe eines geeigneten Reporting-Tools über die ODBC-Schnittstelle von Hortonworks ausgewertet werden.

3.3.6.3 Reporting

Als Reporting Tool wird in dieser beispielhaften Umsetzung das Open Source Projekt BIRT (Business Intelligence Reporting Tool) der Eclipse Foundation verwendet. Es basiert auf der Eclipse Plattform und bietet Funktionalitäten für das Berichtswesen, Business Intelligence sowie mehrdimensionales OLAP (Online Analytical Processing). Das Tool stellt einerseits einen grafischen Editor bereit mit Hilfe dessen der Nutzer seine Reports erstellen kann. Außerdem enthalten ist eine Laufzeitkomponente mit deren Hilfe die Reports anschließend ausgeführt werden können. Das Programm kann kostenlos als Bundle mit einer Eclipse Distribution von der Homepage der Eclipse Foundation heruntergeladen werden.

Als Datenquelle können unter anderem SQL-Datenbanken (JDBC), Web Services oder XML Dateien dienen. Die Anbindung an die Hortonworks Sandbox wird über einen Apache Hive Treiber realisiert, dieser steht ebenfalls auf der Projekt Webseite zum Download bereit.

Nachdem der Hive Treiber eingebunden und ein neues Projekt angelegt ist, kann eine Verbindung zwischen Hortonworks und BIRT hergestellt werden. Anschließend können aus der angebundenen Datenbank über Abfragen (in BIRT „DataSets“ genannt) die Daten aus dem Hadoop System extrahiert und für die Reports verwendet werden, in diesem Fall handelt es sich bei den Abfragen um SQL-Statements.

Anschließend wurden im Editor die drei Reports aus dem Lastenheft beispielhaft umgesetzt:

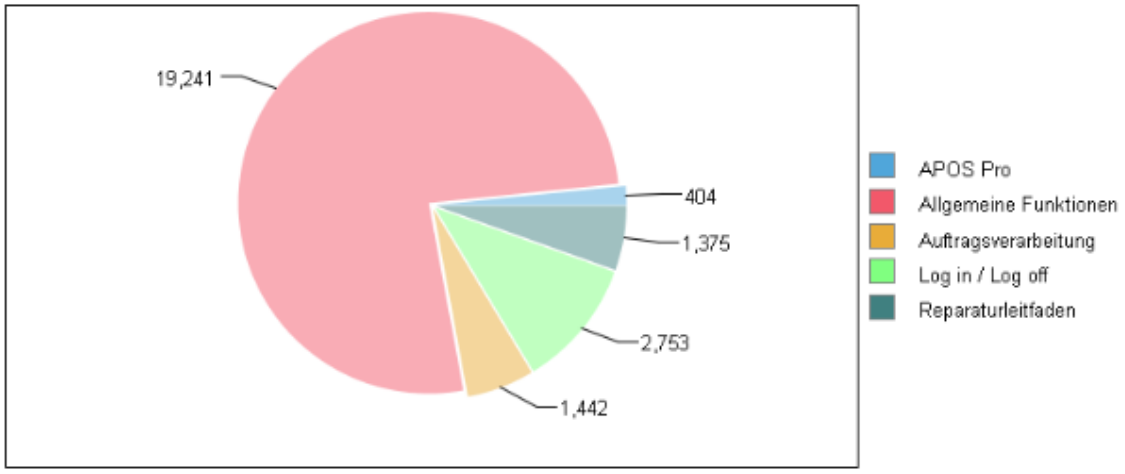


Abbildung 3.18: Anzahl der Zugriffe pro Use Case

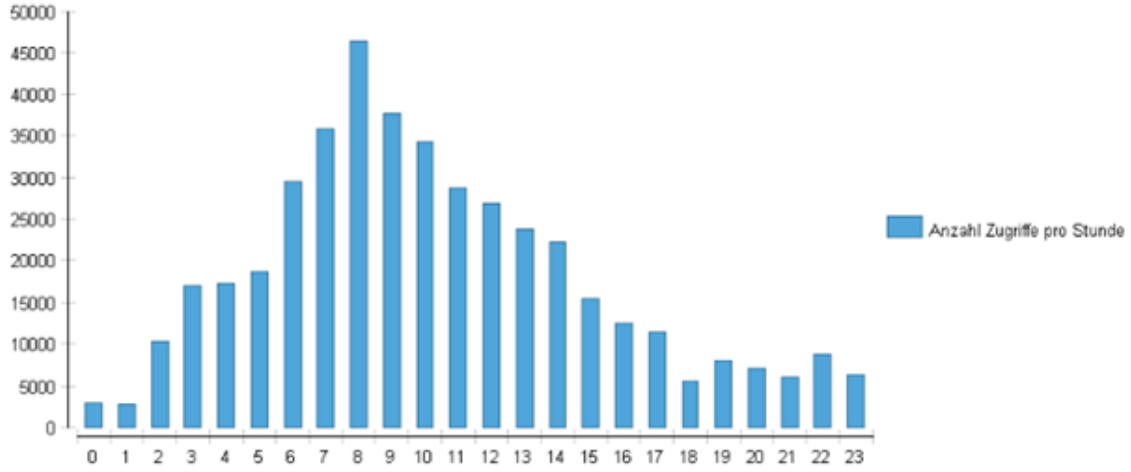


Abbildung 3.19: Anzahl Zugriffe pro Stunde

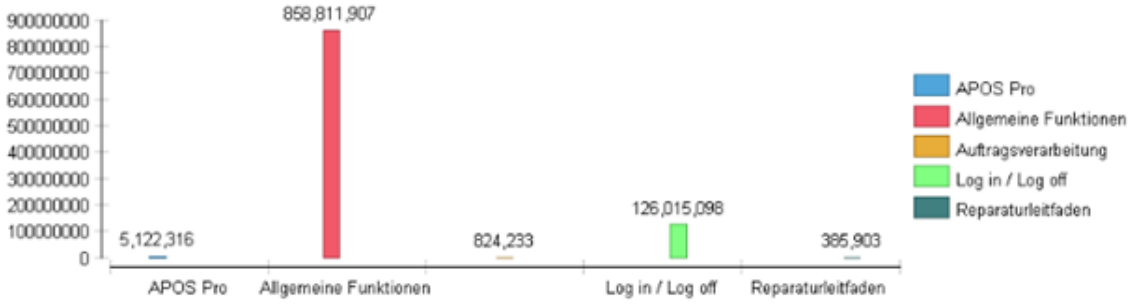


Abbildung 3.20: Use Case Payload in Byte

Neben dem eigentlichen Reporting Tool beeinhaltet das BIRT Projekt auch eine Lösung zur Anzeige der Reports im Browser namens BIRT Viewer. Hiermit wird der Zugriff per Browser auf die Reports über einen (Web-) Server ermöglicht. Die Basis hierfür bildet ein Tomcat Server, welcher durch den BIRT Viewer ergänzt wird. Sowohl der Tomcat Server als auch der BIRT Viewer stehen, wie alle anderen Komponenten in dieser beispielhaften Umsetzung, kostenlos zum Download bereit. Der Hive-Treiber aus dem Reporting Tool muss ebenfalls in den BIRT Viewer eingebunden werden, anschließend kann über den Tomcat Server jeder Report bereitgestellt werden. Der Nutzer kann per Browser auf die gewünschten Reports zugreifen, hierbei wird bei jedem Aufruf der aktuelle Datenstand aus der Hadoop Datenbank abgerufen und damit immer ein aktueller Report angezeigt.

Der folgende Screenshot zeigt die Darstellung eines Reports im BIRT Viewer:

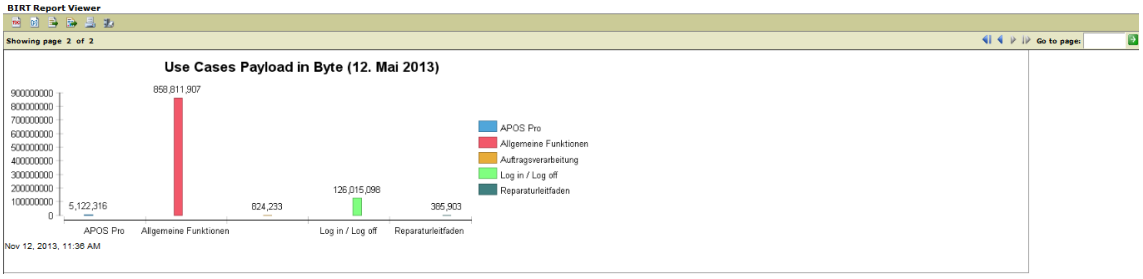


Abbildung 3.21: BIRT Viewer

3.3.7 Installation eines Hadoop-Clusters mit Hortonworks

Der folgende Abschnitt beschreibt nach der Darstellung der Motivation die Durchführung der Installation und die Administration. Abschließend folgt eine Bewertung in einem Fazit.

3.3.7.1 Motivation

Mit der Hortonworks Sandbox bietet das Hortonworks-Team interessierten Entwicklern eine einfache Möglichkeit ein Apache Hadoop System out-of-the-box zu testen. Dabei handelt es sich um das Abbild einer virtuellen Maschine (VM), die lediglich gestartet werden muss.

Die vorkonfigurierte VM stellt eine Installation von Hadoop als Single-Node-Cluster dar. Sämtliche Dienste, die zum Betrieb eines Hadoop Clusters benötigt werden, sind zentral installiert und starten beim Hochfahren der VM automatisch. Zusätzlich ist die Weboberfläche „Hue“ installiert worden, die zur einfachen Nutzung des Clusters dienen soll und Zugang zu gängigen Hadoop-Diensten ermöglicht. Ein integriertes Tutorial mit mehreren Lektionen vervollständigt den einfachen Einstieg in die Thematik.

Auf der Download-Webseite wird diese Form der Installation in der aktuellen Version 2.0 angeboten, um den Umgang mit Hadoop zu erlernen, ein Proof-of-Concept durchzuführen oder neue Funktionalität zu testen. Für einen produktiven Einsatz jedoch eignet sie sich nicht. Dies hat verschiedene Gründe. Zum Einen benötigen diese Formen der VM ein laufendes Betriebssystem und eine Virtualisierungssoftware, die bereits belegte Ressourcen des Betriebssystems dem Virtuellen bereitstellt. Somit sind die Kapazitäten wie Festplatten-, Arbeitsspeicher und Prozessor nicht frei für das Hadoop-System verfügbar und in der Regel stark begrenzt. Zum Anderen können in der Installationsform als Single-Node-Cluster nicht die Vorteile eines verteilten Systems wie Redundanz, Datensicherung und Lastverteilung genutzt werden. Aus diesen Gründen ist es interessant, den Einsatz und den Aufwand für Installation und Konfiguration eines größeren Clusters selbst zu erfahren.

Die folgenden Kapitel stellen die Herausforderungen und Möglichkeiten bei der Installation von Hortonworks in einem Hadoop-Cluster dar. Hortonworks bietet eine umfangreiche Dokumentation auf der eigenen Webseite über APIs, Foren und Wikis an (Vgl. Hortonworks 2013b).

3.3.7.2 Durchführung

Der Abschnitt Durchführung beschreibt zunächst den Aufbau des Systems und im Anschluss die Installation und Konfiguration des Hadoop-Clusters. Danach werden die Administrationsmöglichkeiten aufgezeigt. Der Abschnitt schließt mit einer Schlussbetrachtung.

3.3.7.2.1 Aufbau

Die technische Basis bildet der EC2 Service von Amazon AWS. Dieser Service eignet sich besonders durch die schnelle und kostengünstige Bereitstellung von Servern. Zur Verwaltung des Hadoop-Clusters wird das Open-Source Framework Apache Ambari genutzt, das eine Anleitung zur Installation von Hadoop-Clustern und darüber hinaus Werkzeuge zur Administration und Überwachung des Clusters anbietet. Mit der Ambari-Administrationsoberfläche ist es möglich, die Größe des Clusters zu verändern und einzelne Hadoop-Dienste zu verschieben, zu starten, zu stoppen und zu konfigurieren.

Der Cluster besteht aus sechs Servern, wobei ein Server als zentraler Verwaltungsserver dient. Dem nachgelagert sind die zwei Hadoopmaster, die zentrale Dienste ausführen. Daran angeschlossen sind drei Hadoop-Slaves, die vornehmlich für die Datenhaltung zuständig sind.

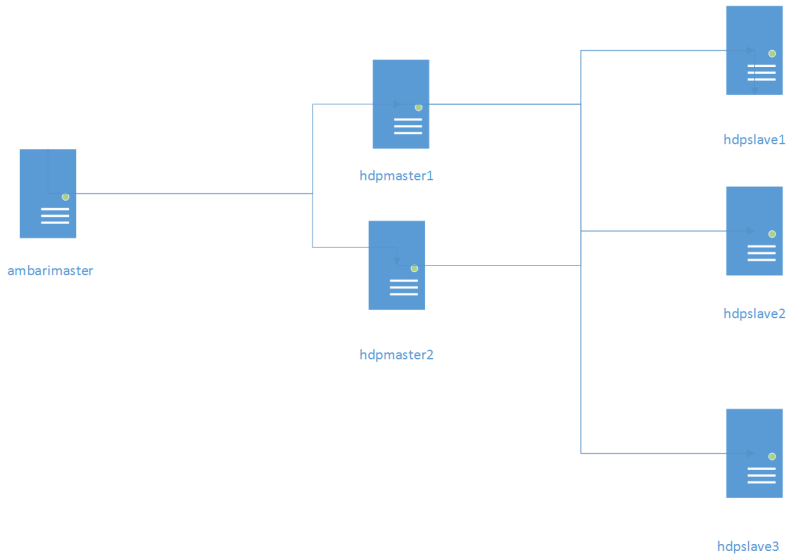


Abbildung 3.22: Aufbau des Hadoop-Clusters

3.3.7.2.2 Installation

Für die Installation werden sechs RHEL[6] 6.3 64bit Instanzen ausgewählt. Das Cluster besteht aus einem zentralen Ambari-Server, zwei Hadoop-Mastern und drei Slaves. Die Hadoop-Master werden in der Konfiguration „m1.large“ aufgesetzt, alle weiteren als „m1.medium“. Die folgende Abbildung 3.23 zeigt die Übersicht über die eingerichteten Server im Amazon Web Server EC2. Alle Server befinden sich im Status „running“ und haben eine Public IP zugewiesen bekommen.

[6] Red Hat Enterprise Linux

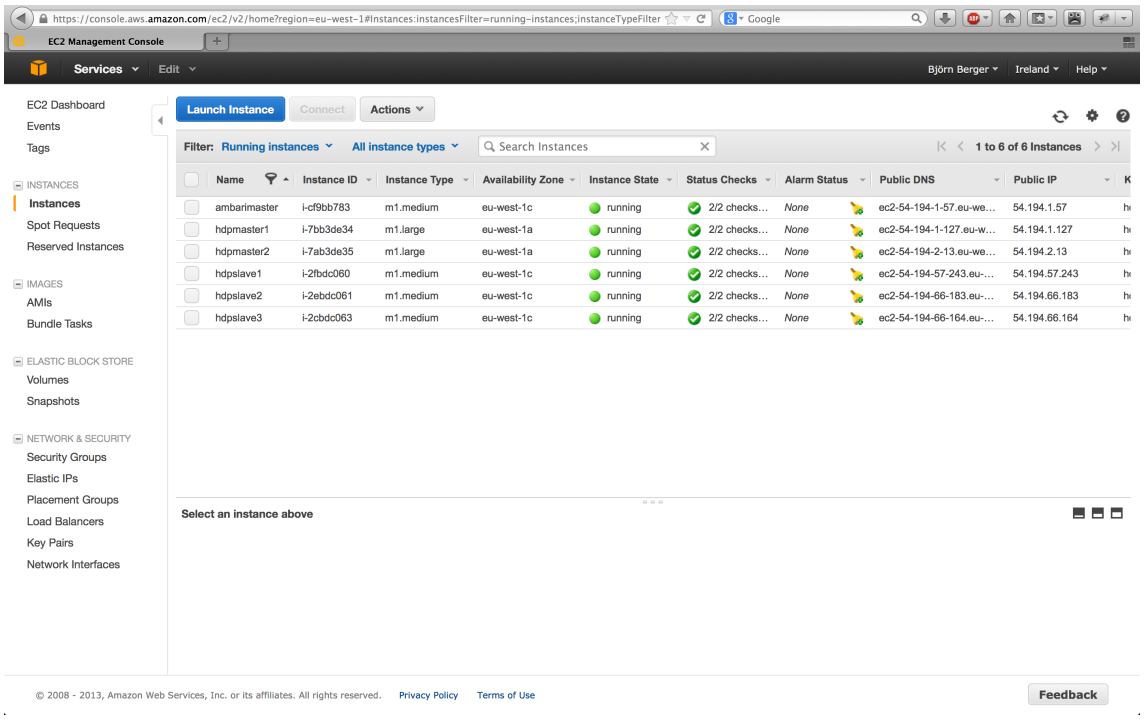


Abbildung 3.23: Übersicht im Amazon Web Service EC2 über die Server - I

Sie unterscheiden sich damit in der unterschiedlichen Bereitstellung virtueller Prozessorkapazität und Arbeitsspeicher. Alle Instanzen erhalten ein virtuelles Festplattenvolumen von 100 GB (Vgl. Abb. 3.24).

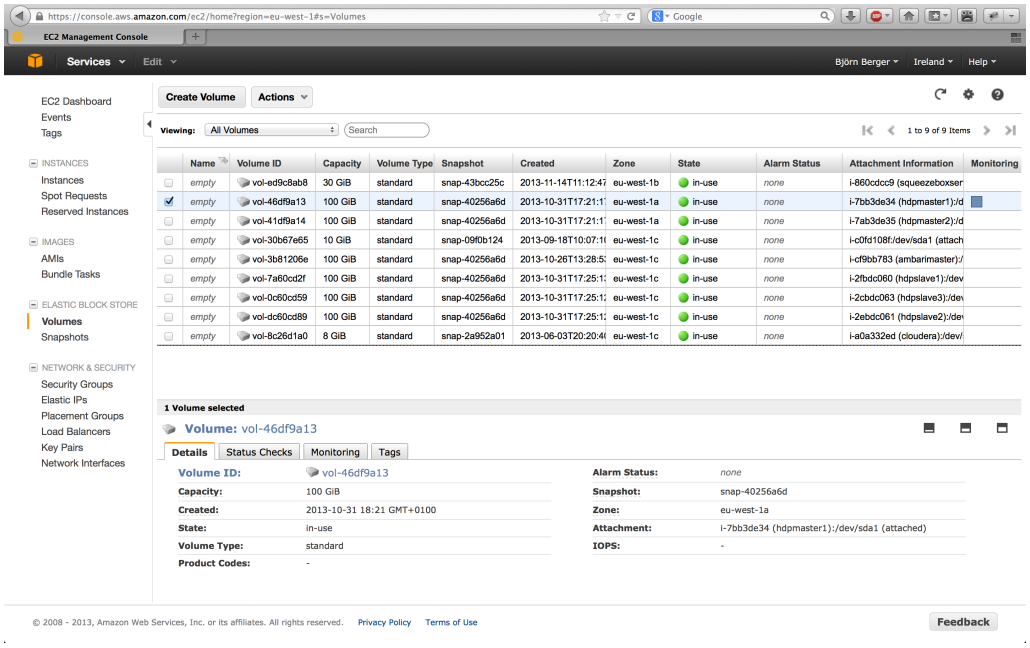


Abbildung 3.24: Übersicht im Amazon Web Service EC2 über die Server -II

Der Zugriff auf die Server geschieht zunächst via SSH über die Konsole mittels während der Installation eingerichteten Security-Token. Über diese Verbindung werden die einzelnen Server untereinander bekannt gemacht, indem die Netzwerkeinstellungen und die -routen eingerichtet werden. Der öffentliche Schlüssel des Servers ambarimaster wird auf den anderen Servern bekannt gegeben, sodass der ambarimaster sich ohne Authentifizierung anmelden kann.

Zur Installation des ambarimaster befindet sich ein yum Repository auf dem Download-Server von Hortonworks, das mittels „wget“ geladen werden kann. Nach der Installation kann der Dienst ambari-server gestartet werden. Von nun an steht eine Weboberfläche zur Verfügung, mit der die anderen Server und die Dienste eingerichtet werden können.

3.3.7.2.3 Konfiguration

Über die öffentliche IP-Adresse des ambarimaster und über den Port 8080 ist die Ambari Weboberfläche zu erreichen, die nach der Anmeldung mit Standard-Benutzer und -Passwort zur weiteren Konfiguration des Systems auffordert. Zunächst sind die weiteren Server anzugeben. Anschließend ist die RSA-ID-Datei hochzuladen, damit der ambarimaster die Installation des Dienstes ambari-client durchführen kann.

Nachdem die Installation des ambari-client erfolgt ist und alle Server sich bei dem ambarimaster registriert haben, folgt die Zuweisung der Dienste. Für eine Testinstallation kann folgende Konfiguration verwendet werden:

- **hdpmaster1:** NameNode, NagiosServer, GangliaCollector, HBaseMaster, ZooKeeper
- **hdpmaster2:** SNameNode, JobTracker, HiveServer2, HiveMetastore, WebHCatServer, OozieServer, ZooKeeper
- **hdpslave1:** ZooKeeper

Der Dienst NameNode nimmt die Anfrage eines Client auf, bearbeitet die Datenanfrage und speichert Metadaten. Er wird bei einem Ausfall von dem Server mit dem SNameNode vertreten. Mit dem NagiosServer-Dienst und dem GangliaCollector ist das Monitoring der Dienste und des Netzwerkes möglich. Der HBaseMaster organisiert die HBase-Datenbank in dem Hadoop-System, das ein verteiltes spaltenorientiertes Datenbanksystem darstellt.

In einem Hadoop-System mit einer HBase-Installation ist der Dienst Zookeeper unabkömmlich, der für die Konfiguration in verteilten Systemen verantwortlich ist. Ebenfalls für die Organisation im verteilten Cluster ist der Dienst JobTracker zuständig, der die MapReduce-Aufgaben verwaltet. Im Idealfall weist der Dienst die verteilten Rechenaufgaben den Knotenpunkten zu, die auch die entsprechenden Daten vorhalten. Dazu kommuniziert der JobTracker-Dienst mit dem NameNode, um die Verteilung der benötigten Daten zu erfahren und weist die Aufgabe einem TaskTracker in der Nähe der Daten zu.

Die Dienste HiveServer2 und HiveMetastore dienen der Bereitstellung der Datawarehouse-Funktionalitäten und erweitern die Datenbank HBase um die Abfragesprache HiveQL und der

Indizierung der Daten. Dieser erfordert eine detailliertere Konfiguration in der Konfigurationsdatei, die über die Weboberfläche zu pflegen ist (Vgl. Abb. 3.25).

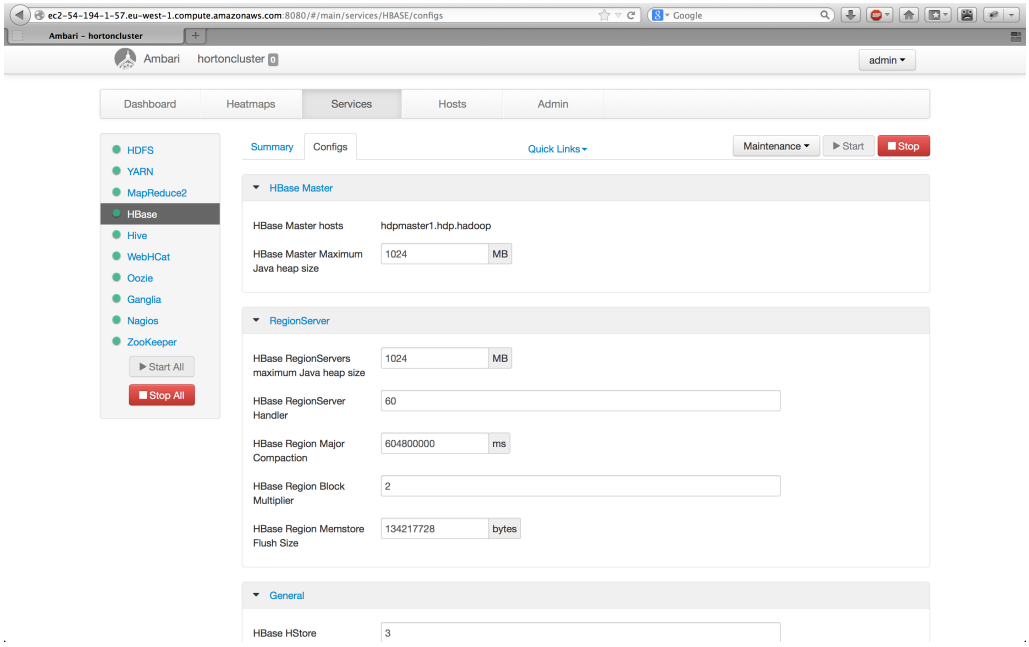


Abbildung 3.25: Konfiguration des Dienstes HBase

Der WebHCatServer-Dienst ist eine Komponente, die eine API für Hadoop und HCatalog bereitstellt. HCatalog ist ein zentraler Bestandteil eines Hadoop-Systems, das Benutzer und Skripte von der Art und dem Ort der Speicherung der Daten trennt und eine Metadaten-Abstraktionsschicht darstellt.

Weiterhin werden alle Server mit Ausnahme des ambarimasters als DataNodes deklariert und stehen somit mit ihrer Festplattenkapazität zur verteilten Speicherung der Daten bereit. In einem produktiven Cluster mit mehreren Servern und einer höheren Speicherkapazität ist sinnvoll, die DataNodes getrennt von den mit der Organisation beschäftigten Servern auszuwählen, damit keine Verarbeitungsengpässe entstehen.

3.3.7.3 Administration

Neben der Durchführung der Installation steht die Ambari-Weboberfläche auch zur Konfiguration und Verwaltung des Hadoop-Clusters zur Verfügung. Ein Dashboard gibt einen ersten Überblick über den Zustand des Clusters. Durch Klick auf ein Kästchen kann die dahinterliegende Grafik vergrößert oder der Dienst direkt bearbeitet werden.

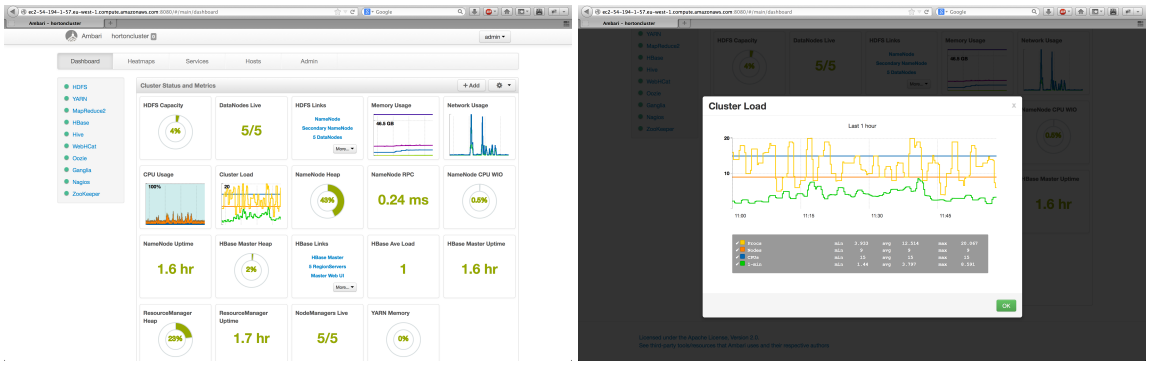


Abbildung 3.26: Das Dashboard

Über die Rubrik Heatmaps können Flaschenhälse in einzelnen Leistungsdisziplinen festgestellt werden. So ist die Host CPU-Wartezeit auf Lese-/Schreiboperationen (HOST CPU Wait I/O %) auf dem ersten von fünf Servern prozentual im bedenklichen Bereich bei 40% bis 60%. Alle anderen Server befinden sich in dieser Disziplin im grünen Bereich. Deutlich schlechter sieht es bei der Überwachung der Hase Memstore Sizes aus, das den Datenbank-Cache bezeichnet. Es fällt sofort auf, dass drei von fünf Servern 80-100MB von bereitstehenden 100MB pro Server bereits ausnutzen und somit die Grenze erreicht haben.

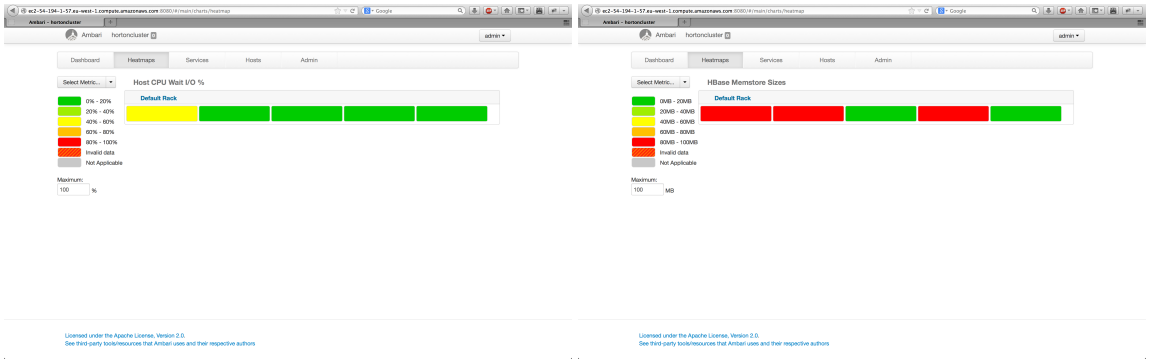


Abbildung 3.27: Die Heatmaps

Die nachfolgende Abbildung 3.28 zeigt die Übersicht über die registrierten Hosts, die Informationen über den Zustand, die IP-Adresse und die Systemauslastung gibt. Darüber hinaus sind die installierten Dienste bei Bedarf einzublenden.

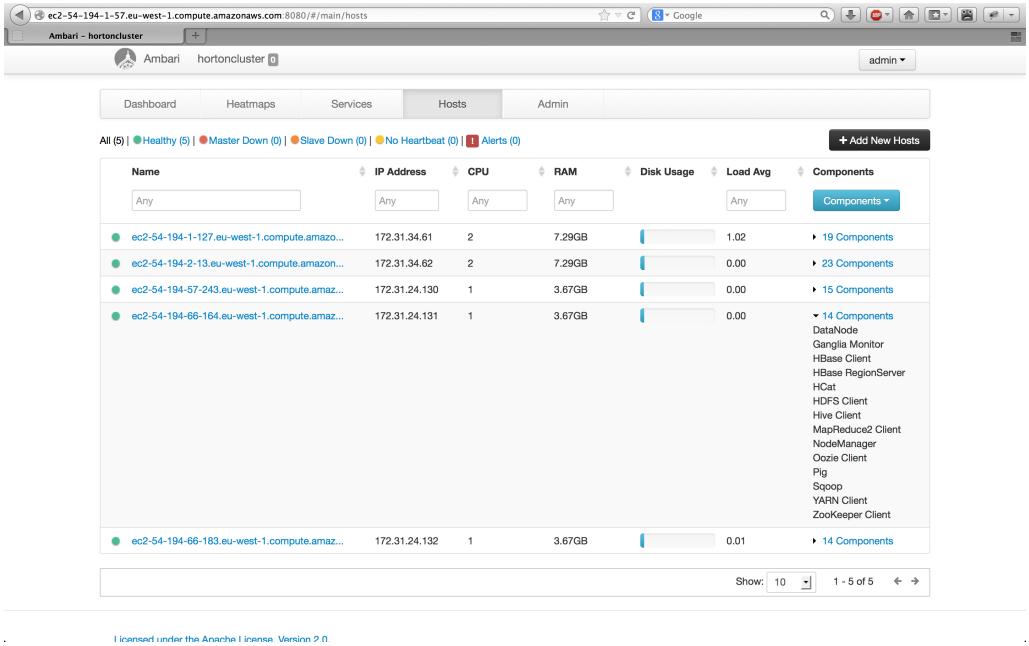


Abbildung 3.28: Übersicht über die registrierten Hosts

Durch Auswahl eines Servers gelangt man zu einer detaillierten Auflistung aller Komponenten des Servers. Diese können über den Action-Button konfiguriert und neu gestartet werden. Das Dashboard gibt Auskunft über die Auslastung des Servers. Mögliche Alarmmeldungen werden in der Kopfzeile dargestellt.

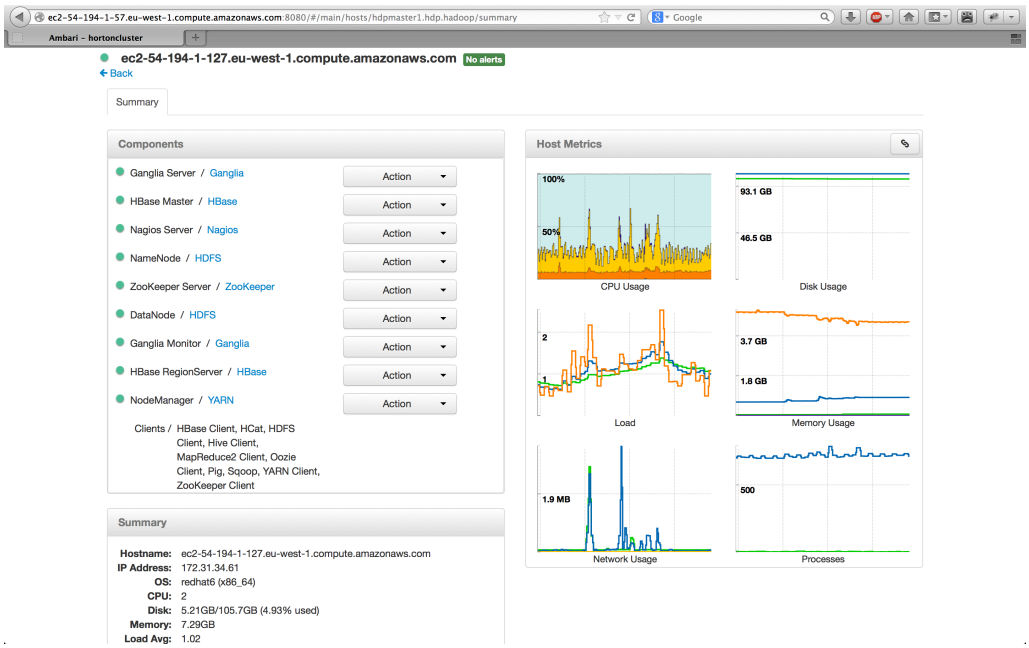


Abbildung 3.29: Details zu einem Hosts

Die Dienste werden über die Rubrik Services überwacht. Im linken Bereich sind alle notwendigen

Dienste aufgeführt und können gestartet und gestoppt werden. Je nach ausgewähltem Dienst stehen unterschiedliche Grafiken bereit, um die wesentlichen Informationen zu visualisieren.

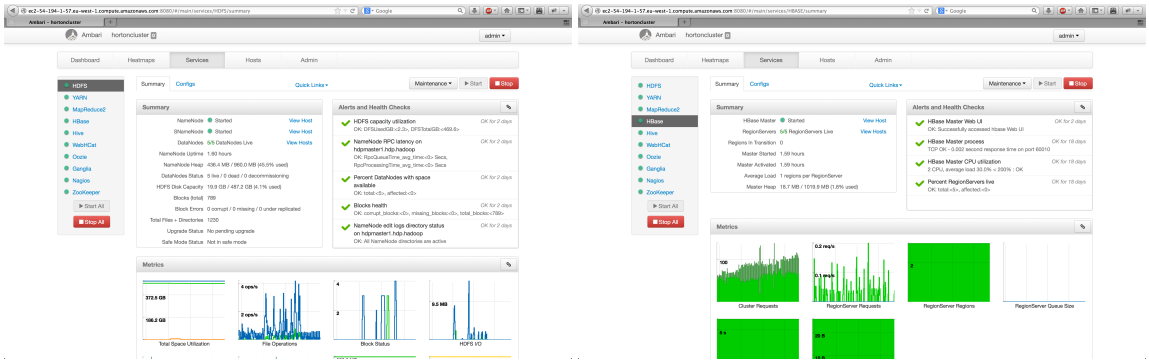


Abbildung 3.30: Die Dienste

3.3.7.4 Schlussbetrachtung

Mit der Zusammenstellung des Software-Pakets zur Installation des Hadoop-Clusters mit der Installations- und Verwaltungsoberfläche Ambari hat Hortonworks den Einrichtungsprozess gegenüber einer manuellen Installation deutlich erleichtert. Die Darstellung des Cluster-Zustands ist übersichtlich und die Konfiguration intuitiv möglich. Die Installation weiterer Dienste wie der Oberfläche Hue ist problemlos möglich. Neue Server können einfach integriert werden. Eine umfangreiche Dokumentation steht im Internet auf den Webseiten von Hortonworks zur Verfügung.

Hortonworks ist eine von vielen Hadoop-Distributionen, die viele interne Konfigurationsaufgaben dem Endnutzer ersparen wollen und ein Gesamtkonzept für eine Hadoop-Architektur bereitstellen. Neben Hortonworks bieten zum Beispiel Cloudera mit Impala, IBM mit der InfoSphere BigInsights Enterprise Edition oder Debian mit einem eigenen Server-Paket verschiedene Distributionen an, die ein Hadoop-Cluster umsetzen oder integrieren. Darüber hinaus beinhalten viele weitere Lösungen anderer Anbieter Hadoop-Funktionalitäten oder können mit diesen umgehen (Vgl. Apache 2013a).

Bei der Installation und Einrichtung eines Hadoop-Clusters stößt man erwartungsgemäß auf Probleme, die einem bei der Einrichtung und Benutzung der Hortonworks-Sandbox verborgen bleiben. Die korrekte Einrichtung des Clusters erfordert Zeit, Erfahrung und eine gute Dokumentation. Die in den Abschnitten zuvor beschriebene Installation eignet sich nicht für einen produktiven Einsatz. Die Verteilung der Dienste auf die zur Verfügung stehenden Server ist nicht ideal, was die Auslastung der einzelnen Dienste und Server zeigt (Vgl. Kap. 3.3.7.3 und Abb. 3.27). Darauf weist allerdings auch Hortonworks selbst hin und verweist auf die vollständige Dokumentation und bietet auf der Webseite Kurse, Tutorials, Enterprise Lösungen, Ausbildungszertifikate, Foren, Partner, Success Stories und Verweise auf Consulting-Unternehmen an.

Es hat sich gezeigt, dass eine Einrichtung eines Hadoop-Clusters möglich ist, ein produktiver Einsatz ist allerdings nur in einer vollständigen Strategie sinnvoll und beinhaltet viele weitere Aufgaben.

3.4 Marktüberblick

Dem Kapitel 3.1.2 lässt sich entnehmen, dass im Rahmen der Projektarbeit ein Marktüberblick über aktuelle Big Data Produkte am Markt erstellt werden soll. Dieses Kapitel nimmt sich diesem Teil der Aufgabenstellung an und stellt nachfolgend die dabei entstandenen Ergebnisse vor.

Dabei wird zunächst auf die Thematik Big Data eingegangen. Hier soll der Leser einen Einblick in die damit einhergehenden Herausforderungen bekommen. Gleichwohl soll daraufhin auf das Vorgehen bei der Erstellung des Marktüberblicks eingegangen sowie die grundlegenden Kriterien und Methoden hierfür vorgestellt werden. Im Anschluss daran wird kurz dargestellt, welche Produkte für die Untersuchung ausgewählt wurden.

Der Hauptteil dieses Kapitels besteht schließlich aus der Vorstellung der untersuchten Produkte sowie den dazugehörigen Anbietern. Im Vordergrund steht hierbei die Ergebnisdarstellung hinsichtlich der von der Projektgruppe aufgestellten Untersuchungsmerkmale. Ausklingen wird dieses Kapitel mit einer Zusammenfassung der Ergebnisse sowie einem abschließenden Fazit.

3.4.1 Problemorientierte Einleitung

Der Begriff Big Data ist seit geraumer Zeit nicht mehr aus dem Fachbereich der Business Intelligence wegzudenken. In den letzten Jahren sind die Datenmengen nicht nur enorm angestiegen, sondern gestalten sich auch vielfältig in ihrer Beschaffenheit. Dies bedeutet, dass die Unternehmen von heute mit sehr großen Datenmengen unterschiedlicher Struktur konfrontiert sind. Hiermit sind vor allem maschinell erzeugte Daten, wie etwa RFID-Funkchip-Erfassungen, Logdaten oder auch Sensordaten sowie Daten aus dem World Wide Web, wie z.B. Social-Media-Daten, gemeint (Vgl. Bange u. Grosser 2012, S. 1).

Das Thema Big Data ist an sich kein neues Thema für Unternehmen. Seit Anfang der 1990er Jahre wird in vielen Bereichen des Einzelhandels bereits das Kundenverhalten analysiert, um Produkte und Werbung zu optimieren. Die Geschmäcker der einzelnen Kunden sind mittlerweile individueller als je zuvor, weshalb die Produkte zunehmend an den Kunden angepasst werden und der Fokus der Unternehmen immer stärker auf die Kunden gerichtet ist. Der Thematik Big Data kommt genau ab diesem Punkt eine besondere Relevanz zu. Die Daten werden zum einen aus der steigenden Anzahl unternehmensinterner Applikationen und zum anderen auch zunehmend aus verschiedensten Quellen erzeugt. Diese Daten müssen verwaltet und ausgewertet werden, was die Komplexität weiter erhöht. Sensoren in Fahrzeugen oder Maschinen, mobile Endgeräte, soziale Netzwerke und viele andere neue Datenquellen können den Unternehmen dabei behilflich sein, Markttrends zu erkennen, indem sie beispielsweise die Stimmungen ihrer Kunden in Foren erfassen, auswerten und so zu einem Mehrwert für das Unternehmen machen. Demnach stehen den Unternehmen durch das World Wide Web beispielsweise gänzlich neue Informationen, z.B. über Kundengruppen, zur Verfügung. Ebenso erweist sich die Auswertung von Sensordaten als

hilfreich für die Optimierung der Produkte und auch der Produktionsprozesse (Vgl. Bretting, S. 5).

Für die Unternehmen stellt sich in diesem Kontext die Frage, wie aus den gesammelten und erzeugten Daten Wettbewerbsvorteile erzielt werden können. Die Masse an Daten muss nicht nur für den Nutzer verfügbar gemacht werden; dies sollte zugleich in einer angemessenen Geschwindigkeit geschehen. An diesem Punkt spielt die Beschaffenheit der IT-Infrastruktur des jeweiligen Unternehmens sowie dessen vorhandenen Transaktionssysteme eine entscheidende Rolle. Denn nicht jedes ERP- oder auch CRM-System kann diese Daten verarbeiten. Damit ist sowohl die unterschiedliche Struktur der Daten als auch die Performance gemeint. Analysetools, die zwar unterschiedlich strukturierte Daten verarbeiten können, aber aufgrund der Menge an Daten ihrer eigentlichen Funktion nicht mehr gerecht werden können, bringen den Unternehmen keinen Mehrwert. Dabei soll Big Data den Unternehmen eben diesen erzeugen (Bange u. Grosser 2012). Big Data meint in diesem Zusammenhang den Einsatz großer unterschiedlich strukturierter Datenmengen aus einer Vielzahl von Quellen und kann anhand von vier wesentlichen Charakteristika (Vgl. Abb. 3.31) beschrieben werden (Vgl. Spies 2012, S. 142).

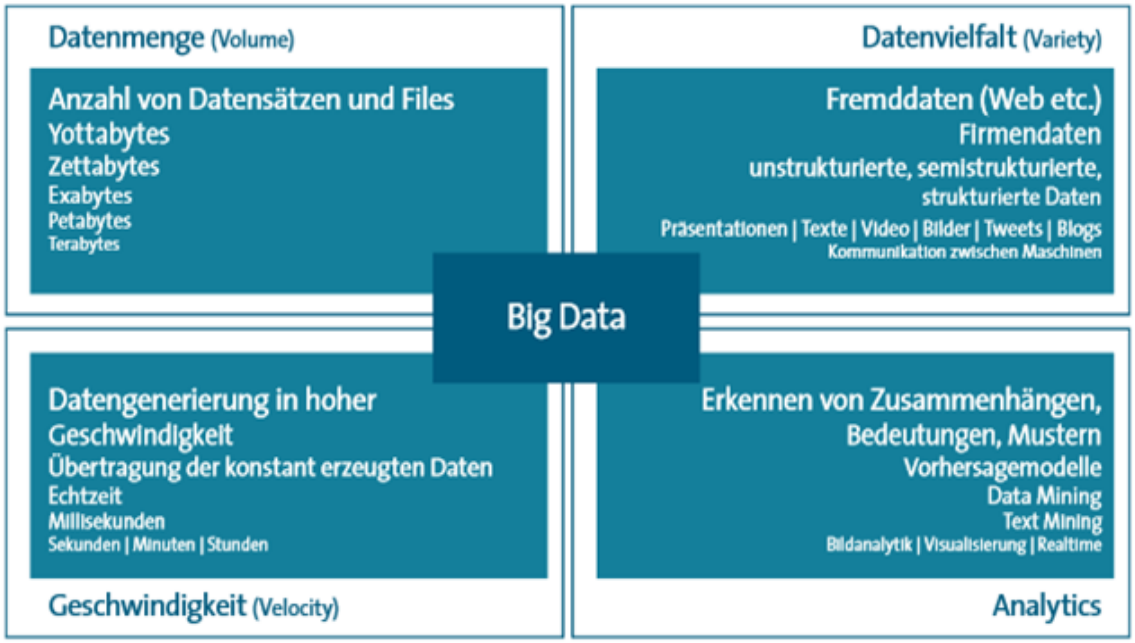


Abbildung 3.31: Die 4 Charakteristika von Big Data
Quelle: BITKOM (2012, S. 19)

Aus der Abbildung 3.31 lässt sich entnehmen, dass Big Data anhand der vier Kriterien Volume, Variety, Velocity und Analytics charakterisiert werden kann. Volume bezieht sich auf die Datenmenge, d.h. der Anzahl von Datensätzen und Files. Damit sind Datenvolumen in Bereichen zwischen Terabytes und Yottabytes gemeint. Variety hingegen steht für die Datenvielfalt, welche sowohl strukturierte als auch semi- und unstrukturierte Daten umfasst. Das dritte Kriterium Velocity zielt auf die Übertragungsgeschwindigkeit der Daten ab. Dies ist vielmehr

eine Anforderung an Big Data als eine Eigenschaft, die durch die Attribute Volume und Variety allerdings für viele Unternehmen und ihre IT-Infrastruktur zur Herausforderung wird ebenso wie das vierte Charakteristikum Analytics. Hiermit ist gemeint, dass die Daten in herkömmlichen Mengen wie bislang alle strukturierten Daten analysierbar sein müssen trotz der zwei wesentlich erschwerenden Eigenschaften Volume und Variety (Vgl. BITKOM 2012, S. 19ff.; Spies 2012, S. 143ff.).

Die Quellen- und Datenvielfalt stellen für die Unternehmen demnach zugleich eine Chance und eine Herausforderung dar. Denn gängige Business-Intelligence-Lösungen können diese großen Datenmengen meist nicht mehr hinreichend erfassen und nicht alle Datenformate verarbeiten. Die Big Data-Problematik ergibt sich demnach aus der Erfassung, Verarbeitung und der Speicherung der neuen Daten sowie der gestiegenen Anzahl von bisher genutzten Daten. Sollte ein Unternehmen jedoch einen Nutzen aus der Big Data-Problematik ziehen wollen, braucht es entsprechende Methoden, Werkzeuge, IT-Infrastrukturen und insbesondere im Bereich des Datenmanagements und der -integration optimierende Lösungen (Vgl. Bretting, S. 5; Spies 2012, S. 142). Mit dieser Thematik befassen sich mittlerweile immer mehr IT-Dienstleister. Der Markt wächst täglich um neue Produkte, die sich der Herausforderung Big Data stellen sollen. Im Rahmen der Allgemeinen Konzeption sowie in Absprache mit dem Partnerunternehmen wurde deshalb ein Marktüberblick für Big Data-Produkte erarbeitet, bei dem auf eine Auswahl von derzeit angebotenen Big Data-Produkten genauer eingegangen wird. Die Vorgehensweise, die Untersuchungsmerkmale und die Ergebnisse werden im Folgenden dargestellt.

3.4.1.1 Motivation und Vorgehen

Der Problematik der Datenerfassung, -verarbeitung und -speicherung im Kontext von Big Data haben sich eine Vielzahl an bestehenden aber auch neuen Anbietern von IT-Produkten und -Dienstleistungen angenommen. Der Markt bietet bereits eine große Bandbreite an Big Data-Produkten. Es erweist sich für Unternehmen jedoch schwierig, eine geeignete Lösung für die eigene IT-Infrastruktur und individuelle Unternehmenssituation zu finden. Aus diesem Grund ist es sinnvoll, diesen Unternehmen einen Überblick über vorhandene Produkte zu geben, die sich der Thematik Big Data annehmen. Im Zuge der Phase der allgemeinen Konzeption sowie der Zusammenarbeit mit dem Partnerunternehmen der Projektgruppe HAWK soll deshalb ein Marktüberblick über aktuell auf dem Markt agierende Anbieter und deren Produkte im Kontext der Big Data Thematik gegeben werden. Mithilfe dieses Marktüberblicks sollen mögliche Anbieter und deren Produkte aufgezeigt und untersucht werden.

Besonders wichtig ist dabei die Einteilung der Produkte in Kategorien, da die unterschiedlichen Produkte nicht für alle Funktionen im Unternehmen einsetzbar sind. Aus diesem Grund werden ausgewählte Produkte nachfolgend im Wesentlichen den drei Kategorien Speicherkonzepte, Datenmanagement und Anwenderwerkzeuge zugeordnet. Innerhalb dieser Kategorien finden sich unterschiedliche Unterkategorien wieder, in welche die Produkte nochmals eingeordnet werden. Dies hat den Grund, dass z. B. der Aufgabenbereich Datenmanagement mehr als

beispielsweise die bloße Datenintegration umfasst, es jedoch Anbieter auf dem Markt gibt, deren Big Data-Produkt nur Teile eines Aufgabenbereichs abdecken. Aus diesem Grund wäre eine reine Zuordnung des Big Data-Produkts zu einer Kategorie nicht hinreichend aussagekräftig für diesen Marktüberblick. Ein weiterer wichtiger Aspekt bezüglich der Auswahl der Kategorien bilden die zuvor genannten vier Charakteristika von Big Data, denn diese machen die Kategorisierung der Produkte unumgänglich und finden sich deshalb sowohl direkt als auch indirekt in diesen wieder. Das Charakteristikum Volume fließt sowohl in die Kategorie Speicherkonzept als auch Datenmanagement (z.B. Datenintegration) ein. Mit der Variety von Big Data müssen entsprechende Anwenderwerkzeuge umgehen können. Ebenso sollte es möglich sein, im Kontext des Charakteristikums Analytics die gewonnenen Daten zu analysieren und das in einer dem Anlass entsprechend sinnvollen Velocity.

Weiter muss unterschieden werden zwischen einem Produkt, welches sich ausschließlich einem oder mehreren der genannten Charakteristika von Big Data widmet und einer Big Data Appliance. Im Zuge dieses Marktüberblicks wird ein Big Data-Produkt definiert als eine reine Software- oder Hardwarelösung für mindestens eine der genannten Kategorien, während mit einer Big Data Appliance ein Komplettsystem gemeint ist, welches Server, Storage, Netzwerk, Datenbanken, Middleware und Applikationen beinhalten kann (Vgl. van Laak 2013).

Das Big Data-Produkt wird im Zuge dieses Marktüberblicks nicht nur einzelnen Kategorien zugeordnet, sondern auch entsprechend eigens aufgestellten Untersuchungsmerkmalen genauer betrachtet. Der Leser soll so nicht nur einen Einblick darüber bekommen, um welche Art von Produkt es sich handelt, sondern auch welche Eigenschaften es aufweist. Diese Untersuchungsmerkmale sollen nachfolgend einschließlich der aufgestellten Kategorien erläutert werden.

3.4.1.2 Kategorien und Untersuchungsmerkmale

Im Rahmen des Marktüberblicks wurde eine Produktübersicht erstellt, in der die untersuchten Produkte der Unternehmen aufgelistet und kategorisiert wurden. Diese Kategorisierung wurde vorgenommen, da der Markt der Big Data-Produkte ein sehr breites Spektrum an unterschiedlichen Lösungen für verschiedenste Anwendungsfälle umfasst. In diesem Kapitel werden die einzelnen Kategorien dargestellt und erläutert. Für die Untersuchung der Produkte wurden Kriterien aufgestellt, um eine Vergleichbarkeit der verschiedenen Produkte innerhalb einer Kategorie gewährleisten zu können. Auf diese Untersuchungsmerkmale wird im Anschluss an die Kategorien eingegangen.

Kategorien

Für die Produktübersicht wurden die untersuchten Produkte in die drei Kategorien Speicherkonzepte, Anwenderwerkzeuge und Datenmanagement eingeordnet. Für eine bessere Übersicht und Vergleichbarkeit der Produkte, wurden diese zusätzlich in verschiedene Unterkategorien eingeteilt. Die Auswahl dieser Kategorien und Unterkategorien erfolgte im Wesentlichen in

Anlehnung an BARC (2012) und BARC (2013). Sie werden in den folgenden Kapiteln genauer spezifiziert.

- **Speicherkonzepte**
  Speicherkonzepte bezeichnen die verschiedenen Möglichkeiten der Datenhaltung im Unternehmen. Bei den untersuchten Speicherkonzepten wird zwischen klassischen relationalen, analytischen und multidimensionalen Datenbanken unterschieden. In Abbildung 3.54 auf Seite 136, sowie in den Kategorisierungen der einzelnen Produkte in Kapitel 3.4.2 wurde mit einem x markiert, welches Speicherkonzept von den jeweiligen Produkten verwendet wird.

  Klassische relationale Datenbanken basieren auf dem 1970 von E. F. Codd entwickelten Relationalen Datenmodell, welches aus Objekten, Operationen und Regeln besteht. Objekte sind dabei nach Goll (Vgl. Goll 2011, S. 589)

  - Relationen bzw. Tabellen,
  - die Tupel oder Datensätze in diesen Tabellen,
  - die Attribute, welche die Spalte einer Tabelle darstellen
  - und der Schlüssel, wodurch der Zugriff auf einzelne Tupel ermöglicht wird.

  Durch Operationen lassen sich Relationen anlegen, ändern, löschen sowie abfragen. Um die Konsistenz der Daten zu gewährleisten, gibt es Regeln, wie die Entity-Integrität[7] und die referenzielle Integrität[8]. Im Rahmen von Big Data stoßen relationale Datenbanken jedoch allein durch die großen Datenmengen an ihre Grenzen.

  Analytische Datenbanken bieten in diesem Bereich durch verschiedene Methoden und Technologien Vorteile, wie „Spaltenorientierung, Kompression, Parallelisierung, In-Memory-Verarbeitung sowie Bündelungen von Hardware und Software“ (Martin 2013, S. 4). Dadurch wird die Skalierbarkeit verbessert und die Performance von analytischen Datenbank-Abfragen gesteigert (Vgl. Martin 2013, S. 6).

  Multidimensionale Datenbanken verwenden ein multidimensionales Datenmodell, welches durch einen Datenwürfel visualisiert werden kann. Dabei werden die einzelnen Daten nach verschiedenen Dimensionen organisiert (Vgl. Saake u. a. 2010, S. 625).

  Hadoop-Distributionen sind von Unternehmen zusammengestellte und weiterentwickelte spezielle Pakete auf Basis des Open Source-Frameworks Hadoop, welches die Verarbeitung von Big Data auf Serverclustern ermöglicht (Vgl. Apache 2013b). Das Framework basiert auf der MapReduce-Technologie, die Such- bzw. Analyseanfragen in mehrere Teilaufgaben aufspaltet, die parallel auf den einzelnen Clustern abgearbeitet werden können (Vgl. Spies 2012, S. 151). Der Kern ist das Hadoop Distributed File System (HDFS), das auch

[7] Jede Entity einer Relation besitzt einen eindeutigen Schlüssel (Vgl. Goll 2011, S. 593).
[8] Jeder Wert eines Fremdschlüssels einer Relation ist ein NULL-Wert oder entspricht dem Wert des Primärschlüssels einer anderen Relation (Vgl. Goll 2011, S. 593).

auf Standardhardware einsetzbar ist und statt den Daten, die Programme zwischen den Clustern verschiebt (Vgl. Gleich 2013, S. 192).

Sonstige Datenbanken sind zum Beispiel Streaming-Datenbanken oder NoSQL-Datenbanken. Bei Streaming-Datenbanken liegt im Gegensatz zu herkömmlichen DBMS ein Data Stream Management System (DSMS) vor, das Anfragen auf heterogene Datenströme verwaltet und die Ergebnisse in Relationen speichert oder als Stream versendet. Hierbei werden Weiterentwicklungen von SQL wie die Sprache Continous Computation Language (CCL) verwendet (Vgl. Rommelspacher 2011, S.118f.). NoSQL-Datenbanken stellen eine mögliche Alternative zu relationalen Datenbanken dar. NoSQL steht dabei für „Not only SQL“ und nicht für das näher liegende „No SQL“. Auf SQL wird also nicht grundsätzlich verzichtet. Häufig werden die Daten dabei nicht in Tabellen, sondern in flexibleren Strukturen gespeichert. Dadurch entstehen Vorteile bei der Skalierbarkeit, da große Datenmengen in einem Cluster gespeichert werden können. Dabei können neben flexibleren Strukturen auch andere Grundsätze wie das CAP-Theorem[9] angewendet werden, wodurch große Cluster möglich sind. Es wird zwischen vier NoSQL-Ausprägungen unterschieden: dokumentenorientierte Datenbanken, Key-Value-Datenbanken, spaltenorientierte Datenbanken (zählen jedoch zu den analytischen Datenbanken) und Graphendatenbanken (Vgl. Walker-Morgan 2010).

- **Anwenderwerkzeuge**
  Unter Anwenderwerkzeugen werden die Produkte verstanden, die der Endanwender direkt bedient und dazu verwendet, Informationen aus den vorhandenen Daten zu generieren. Es gibt unterschiedliche Möglichkeiten, mit denen der Anwender die Daten aus den oben beschriebenen Speicherkonzepten verarbeiten kann. In diesem Marktüberblick wurden die Möglichkeiten für ein Dashboarding, das Berichtswesen, die Analyse und das Data Mining untersucht. In der Tabelle bedeutet ein *, dass diese Funktion von dem untersuchten Produkt abgedeckt wird, zwei ** besagen, dass bei dieser Funktion der Produktschwerpunkt liegt.

  Dashboarding bezeichnet die Bereitstellung von fertigen Reports in einem individuell gestaltbaren Frontend-Tool. Darin werden verschiedene Analyse- oder Berichtsergebnisse in einer Übersicht dargestellt, z.B. als Grafiken oder Tabellen. Dadurch ist es für jeden Benutzer möglich, auch ohne Kenntnisse der Datenbasis aktuelle Analysen zu erstellen (Vgl. Bachmann u. Kemper 2011, S. 109).

  Unter **Berichtswesen** wird die Möglichkeit zur Erstellung von Berichten oder Reports und deren Verteilung verstanden. In der tabellarischen Darstellung zeigen dabei zwei Sterne, dass sowohl formatiertes Druckberichtswesen als auch Ad-hoc-Berichtserstellung unterstützt werden.

[9] Besagt, dass in einem verteilten System maximal zwei der drei Eigenschaften Konsistenz, Verfügbarkeit und Partitionstoleranz gewährleistet werden können (Vgl. solid IT 2014)

Bei der **Analyse** werden verschiedene Analyseverfahren, wie die OLAP-Analyse, die mengenbasierte oder die visuelle Analyse eingesetzt, um aus den vorhandenen Daten neue Informationen zu gewinnen.

Data Mining bezeichnet die „Anwendung von Methoden und Algorithmen zur möglichst automatischen Extraktion empirischer Zusammenhänge zwischen Planungsobjekten, deren Daten in einer hierfür aufgebauten Datenbasis bereitgestellt werden“ (Gabler Wirtschaftslexikon 2013a).

- **Datenmanagement**
  Die Aufgabe des Datenmanagements ist, Daten gezielt und möglichst in Echtzeit bereitzustellen, sodass aus diesen neue Informationen und Wissen generiert werden können (Vgl. Dippold u. Fedtke 2005, S. 4). Dabei ist besonders auf die Integrationsmöglichkeiten und die Qualität der Daten zu achten. Ob diese gegeben bzw. gewährleistet werden signalisiert ein x in der tabellarischen Darstellung.

  Datenintegration wird benötigt, wenn Daten an unterschiedlichen Orten gespeichert werden und zur Nutzung zusammengeführt und einheitlich dargestellt werden sollen. Dabei werden die Daten zunächst aus den unterschiedlichen Datenquellen ausgelesen und ggf. vorverarbeitet, dann im sogenannten „Schema Matching“ vereinheitlicht, anschließend auf Dubletten untersucht und abschließend zu einer einzigen Repräsentation zusammengefügt. Auf diese Datenbasis wird im Anschluss von den Anwendungsprogrammen zugegriffen (Vgl. Bleiholder u. Schmid 2008, S. 123ff.).

  Datenqualität ist die „Gesamtheit der Ausprägungen von Qualitätsmerkmalen eines Datenbestandes bezüglich dessen Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen.“ Qualitätsmerkmale sind dabei die Qualität der Datendefinition, die inhaltliche Datenqualität und die Qualität der Datenpräsentation. Die Daten werden auf diese Merkmale überprüft und ggf. bereinigt (Vgl. Gebauer u. Windheuser 2008, S. 89ff.).

Die Einordnung der untersuchten Produkte in die vorgestellten Kategorien erfolgt in Kapitel 3.4.2.

Untersuchungsmerkmale

Neben dieser Produktübersicht wurden die einzelnen Produkte auf verschiedene Merkmale hin untersucht, um herauszufinden, welche Leistungen die ausgewählten Produkte aufweisen können. Diese sollen ermöglichen, dass die Produkte der untersuchten Anbieter in Bezug auf diese Merkmale zumindest grob miteinander verglichen bzw. auf gleiche Merkmale untersucht werden können. In der nachfolgenden Tabelle 3.2 werden diese Merkmale genannt und erläutert:

Untersuchungsmerkmal	Erklärung
Möglichkeit der Integration in bestehende BI-Architekturen	Lässt sich die Lösung in bestehende BI-Architekturen integrieren, um diese funktional zu erweitern?
Skalierbarkeit / Zukunftssicherheit	Wo sind die Grenzen des Produktes? Lässt sich die Lösung auch bei steigendem Datenvolumen weiter verwenden?
Verarbeitung verschiedener Datenquellen (Schnittstellen)	Welche Datenquellen können von dem Produkt verwendet werden? Welche Datenbanken, welche Systeme, welche Anwendungen, welche Technologien können über eine Schnittstelle angesprochen werden?
Verarbeitung verschiedener Datenformate	Welche Dateiformate kann das Produkt verarbeiten?
Umgang mit semi- und unstrukturierten Daten	Können nur strukturierte Daten verarbeitet werden oder auch semi- und unstrukturierte Daten? Strukturierte Daten liegen in den festen Tabellenfeldern relationaler Datenbanken vor. Unstrukturierte Daten liegen in Textform vor (z.B. E-Mails), können aber auch Audio-, Video- oder Bilddaten sein. Semistrukturierte Daten werden durch Tags in aufeinander folgende Datenelemente unterteilt (z.B. XML-Daten). Zusätzlich gibt es noch polystrukturierte Daten, die aus unterschiedlichen Datenstrukturen bestehen, die sich noch verändern können (Vgl. Fujitsu 2013, S. 3).
Benutzeroberfläche / Benutzerfreundlichkeit	Ist eine grafische Benutzeroberfläche vorhanden oder erfolgt die Konfiguration durch Programmiercode?
Softwaredokumentation	Wie umfangreich ist die Dokumentation der Software? Werden alle relevanten Funktionen verständlich erklärt?
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Ist der Anbieter bereits am Markt etabliert? Wie viele und welche Unternehmen benutzen die angebotene Lösung?
Qualität des Supports	Wie ist die Erreichbarkeit des Supports? Gibt es eine Community, die bei der Lösung von Problemen hilft?
Produktkategorie	Hier wird eine Einordnung in die Kategorien Speicherkonzepte, Anwenderwerkzeuge und Datenmanagement vorgenommen, die im vorangegangenen Kapitel 3.4.1.2 erläutert wurden.

Tabelle 3.2: Erklärung der Untersuchungsmerkmale

Die Ergebnisse zu diesen Untersuchungsmerkmalen werden tabellarisch in den Untersuchungen zu den einzelnen Produkten in Kapitel 3.4.2 aufgeführt. Welche Produkte dabei untersucht

wurden, wird im folgenden Kapitel dargestellt.

3.4.1.3 Auswahl der Big Data Produkte

Im August 2013 wurde bereits ein Marktüberblick von der Experton Group AG im Bezug auf Big Data-Produkte erstellt, jedoch nicht veröffentlicht - der Big Data Vendor Benchmark 2013. Ersten öffentlichen Informationen zu dieser Studie lässt sich entnehmen, dass das Unternehmen 68 Anbieter von Big Data-Produkten untersucht hat (Vgl. van Laak 2013). „139 Unternehmen wurden ursprünglich für den Big Data Vendor Benchmark betrachtet. 68 Unternehmen konnten entsprechend der Experton-Group-Definition für Big-Data-Lösungen und -Lösungsbausteine in die Bewertung aufgenommen werden“ (van Laak 2013). Leider stehen der Projektgruppe die Ergebnisse und begrifflichen Erläuterungen dieser Studie nicht zur Verfügung. Dennoch können den ersten öffentlichen Informationen dieser Studie relevante Anbieter aus den Bereichen Analytics und Appliances entnommen werden, weshalb diese der Projektgruppe zunächst als Orientierung bezüglich der Auswahl relevanter Anbieter und Produkte dienten (Vgl. Landrock 2013; Vgl. van Laak 2013). Weiter dienten der Projektgruppe der Barc Guide Business Intelligence 2012/2013 sowie die gleichnamige Ausgabe aus dem Jahr 2013/2014 der B-Eye-Media GmbH ebenso wie der Business Intelligence Lösungs-Guide 2013 der isi Medien GmbH als Basisliteratur für die Auswahl der Anbieter und Produkte für den aufgestellten Marktüberblick. Da im Rahmen der Arbeitsphasen der Allgemeinen Konzeption sowie der Konzeption für den Praxispartner eine Untersuchung sämtlicher Anbieter und Produkte im Kontext der Big Data-Thematik nicht möglich war, musste eine Auswahl getroffen werden. Aus diesem Grund werden nachfolgend 20 Anbieter und deren Big Data-Produkte genauer untersucht.

Da Big Data durch die vier bereits genannten Eigenschaften charakterisiert werden kann, ergeben sich in diesem Kontext auch verschiedene Sichtweisen hinsichtlich des Marktüberblicks. So werden auf dem Markt Big Data-Produkte mit unterschiedlichen Schwerpunkten und Stärken angeboten. Während einige Produkte sich mit der Verarbeitung von großen Datenmassen auseinandersetzen, legen andere Anbieter mit ihren Produkten den Schwerpunkt bspw. auf die Datenvielfalt oder die Analyse von polystrukturierten Daten. Für besonders interessant im Rahmen unserer Arbeit erweist sich der Aspekt der polystrukturierten Daten auch in Bezug auf die Arbeit mit dem Partnerunternehmen. In diesem Kontext wurden u.a. Produkte untersucht, die nicht nur mit der Masse an Daten zurechtkommen, sondern auch polystrukturierte Daten verarbeiten können. Ebenso wurden durch Anregungen seitens des Partnerunternehmens Produkte anvisiert, welche aufgrund der Marktpräsenz der zugehörigen Anbieter interessant erscheinen ebenso wie Anbieter, die ihr Produktportfolio ausschließlich auf Big Data-Produkte spezialisieren.

Schließlich wurden im Rahmen des Marktüberblicks ausgewählte Produkte folgender Anbieter untersucht:

Cloudera	Hortonworks	Oracle	SAS
Datameer	HP	Parstream	SGI
Datawatch	IBM	Pentaho	Software AG
EMC	Informatica	QlikTech	Splunk
EXASOL	Jaspersoft	SAP	Teradata

Tabelle 3.3: Die zu untersuchenden Anbieter im Rahmen des Marktüberblicks

Eine Bewertung der Anbieter und deren Big Data-Produkte ist im Rahmen des Marktüberblicks jedoch nicht möglich. Im Kontext dieser Arbeitsphasen wurde keine unternehmensindividuelle Problematik als Basis für die Recherche genutzt. Ziel war es ebenso nicht, Optimierungsbedarfe oder -möglichkeiten in einem Unternehmen anhand der Vorstellung und des Vergleichs der Big Data-Produkte aufzudecken. Das Ziel ist vielmehr, einen Ausschnitt der aktuellen Big Data-Produkte am Markt zu präsentieren und zu untersuchen. Denn ein Vergleich ohne eine konkrete Problemstellung, wie z.B. das Ausfindigmachen eines geeigneten Analyseproduktes für Big Data für ein bestimmtes Unternehmen und dessen individuelle IT-Infrastruktur, ist kaum bzw. gar nicht möglich. So kann ein Anbieter etwa eine Komplettlösung für die Verarbeitung und Analyse von Big Data anbieten, dieses Produkt muss jedoch nicht zwingend für Unternehmen und Probleme jeglicher Art gleichermaßen empfehlenswert sein. Anbieter, die Teillösungen anbieten, können sich als ebenso effektiv erweisen, wie eine Komplettlösung. Die Entscheidung, welches Produkt in einem Ranking an oberster Stelle stünde, hängt demnach mit den unterschiedlichen unternehmensspezifischen Gegebenheiten zusammen. Um die ausgewählten Anbieter und deren Big Data-Produkte schließlich dennoch genauer zu untersuchen, wurden daher die in 3.4.1.2 genannten Kategorien und Untersuchungsmerkmale aufgestellt. Jene sollen ermöglichen, dass die Produkte der untersuchten Anbieter in Bezug auf diese Merkmale zumindest grob miteinander verglichen werden können.

3.4.2 Anbieter und deren Produkte

Der Hauptgegenstand der Arbeitsphase des Marktüberblicks bildet die Untersuchung der genannten Unternehmen und deren Big Data-Produkte. Dieses Kapitel befasst sich mit eben dieser Thematik. Nachfolgend werden deshalb die genannten Anbieter und deren untersuchten Big Data-Produkte vorgestellt.

Die Vorstellung erfolgt in alphabetischer Reihenfolge und stellt keine hierarchische Rangliste dar. Es werden zunächst grundlegende Informationen zum Anbieter und anschließend zu dessen Produktportfolio gegeben. Daran anknüpfend wird das Big Data-Produkt vorgestellt. In diesem Zusammenhang werden sowohl allgemeine Informationen zu dem Produkt dargestellt als auch eine Kategorisierung vorgenommen sowie die Ergebnisse hinsichtlich der Untersuchungsmerkmale präsentiert. Dabei bedeutet die Abkürzung „k.A.“, dass keine Angabe zu diesem Punkt gemacht werden konnte. Diese Abkürzung wird stets dann aufgeführt, wenn der Hersteller der Öffentlichkeit keine oder unklare Informationen dazu bereitgestellt hat.

3.4.2.1 Cloudera

Cloudera ist ein 2008 gegründetes US-amerikanisches Unternehmen mit Sitz in Palo Alto, Kalifornien, das eine Plattform zur Speicherung, Verarbeitung und Analyse von Big Data anbietet. Diese basiert auf Apache Hadoop und unterliegt den Lizenzbestimmungen der Open Source Initiative (Vgl. Cloudera 2013d). Das Unternehmen bietet mit Cloudera Standard und Cloudera Enterprise zwei Varianten einer Plattform an. Cloudera Standard ist eine eingeschränkte und kostenlose Version wohingegen Cloudera Enterprise den vollen Funktionsumfang bietet. In beiden ist die ebenfalls separat erhältliche Cloudera Distribution mit Apache Hadoop (CDH) enthalten, welche mit Impala eine Massively Parallel Processing (MPP) Query-Engine besitzt (Vgl. Cloudera 2013g). Mit dieser lassen sich über ein SQL-Interface Echtzeitabfragen erstellen, die als Grundlage Quelldaten aus dem Hadoop File System (HDFS) oder HBase verwenden können (Vgl. Ihlenfeld 2012).

Impala ermöglicht dabei (Vgl. Cloudera 2013b)

- sekundenschnelle SQL-Abfragen in der CDH,
- die Nutzung einer MPP Query-Engine,
- die Integration von BI-Tools,
- Zugriff auf HDFS und HBase,
- Zugriff auf verschiedene Dateiformate wie Text, SequenceFiles, Avro, RCFile, LZO und Parquet,
- In-Memory Datentransfer,
- den Einsatz von Metadaten, ODBC-Treibern, SQL-Syntax und Beeswax GUI aus Apache Hive,
- Authentifizierung mit Kerberos und
- rollenbasierte Zugriffskontrolle mit Sentry.

Cloudera Enterprise ermöglicht die Erfassung, Speicherung, Untersuchung, Verarbeitung und Analyse von Daten aus verschiedenen Datenquellen und gibt diese an Data Warehouses, Data Marts, BI-Tools oder spezielle Applikationen weiter (Vgl. Cloudera 2013a).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Cloudera Enterprise in der folgenden Tabelle dargestellt:

Cloudera Enterprise - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• Integration von Data Warehouses, Data Marts und BI-Tools möglich. (Vgl. Cloudera 2013a) • Durch das Cloudera Connect Partner Program, das über 600 Unternehmen umfasst, wird eine Integration in unterschiedliche Unternehmensumgebungen ermöglicht. (Vgl. Cloudera 2013d)
Skalierbarkeit / Zukunftssicherheit	Durch Verwendung von Hadoop ist Cloudera Enterprise skalierbar und flexibel. (Vgl. Cloudera 2013h)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• HDFS, HBase (Vgl. Ihlenfeld 2012) • Strukturierte Daten aus POS, Warenwirtschaftssystemen, etc. • Unstrukturierte Daten wie Anwendungsdaten, Log Files und weitere externe Quellen. (Vgl. Cloudera 2013a)
Verarbeitung verschiedener Datenformate	• Support für eine Vielzahl an Datenformaten wie Text, Sequence-Files, Avro, RCFile, LZO und Parquet. (Vgl. Cloudera 2013b) • Anwendungsdaten, Log Files (Vgl. Cloudera 2013a)
Umgang mit semi- und unstrukturierten Daten	strukturiert und unstrukturiert (Vgl. Cloudera 2013i)
Benutzeroberfläche / Benutzerfreundlichkeit	k.A.

Software-dokumentation	• vollständige Dokumentation die die Installation, Konfiguration und Nutzung beschreibt. • Umfangreiche Wissensdatenbank (Vgl. Cloudera 2013c)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	u.a. ebay, Dell, Samsung (Vgl. Cloudera 2013f)
Qualität des Supports	• Weltweit Support-Mitarbeiter • End-to-end Abdeckung für CDH, Cloudera Manager und Cloudera Navigator • 8x5 oder 24x7 Service Levels • Proaktives Support-Programm • Einfluss auf Open Source Roadmaps (Vgl. Cloudera 2013c)
Produktkategorie	Datenmanagement

Tabelle 3.4: Ergebnisse Untersuchungsmerkmale Cloudera Enterprise

Cloudera Enterprise ist eine Plattform, die eine Hadoop-Distribution beinhaltet und Daten aus verschiedensten Quellen extrahiert und transformiert und anschließend zur Verwendung in Data Warehouses, Data Marts oder diverse Applikationen lädt. Wie in der folgenden Abbildung deutlich wird, ist Cloudera Enterprise der Kategorie Datenmanagement und der Unterkategorie Datenintegration zuzuordnen. Da mit CDH eine eigene Hadoop-Distribution enthalten ist, lässt sich ebenfalls eine Einordnung in die Kategorie Speicherkonzepte vornehmen.

	Unternehmen	Cloudera
	untersuchte Produkte bzw. Lösungen	Cloudera Enterprise
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	x
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	x
	Datenqualität	

Abbildung 3.32: Produktkategorisierung Cloudera
Quelle: Eigene Darstellung in Anlehnung an IS Report (2013, S. 100)

3.4.2.2 Datameer

Das Unternehmen Datameer ist ein US-amerikanisches Unternehmen mit Sitz in San Mateo, Kalifornien, das im Jahr 2009 gegründet wurde. Die Mehrzahl der Programmierer ist in Halle (Saale) angesiedelt, da der Gründer des Unternehmens aus Deutschland kommt. Datameer ist im Bereich Softwareentwicklung bzw. Big Data Analytics tätig und auf die Erstellung von Produkten zur Analyse von Big Data spezialisiert (Vgl. Datameer 2013; Steller 2013; Connecticum 2013).

Das Produkt Datameer des gleichnamigen Unternehmens ist ein Analyse-Tool und beinhaltet eine Technologie, die eine erweiterbare Hadoop-Analytics-Plattform zur Verfügung stellt und so mithilfe des Hadoop-Distributed-File-Systems Daten importiert, exportiert und gespeichert werden können. Aufgrund der Nutzung von MapReduce-Jobs können umfangreiche Analysen durchgeführt werden, die in einem Data-Warehouse wiederverwendet werden können (Vgl. Datameer-Beratung 2013b).

Die wesentlichen Produktmerkmale von Datameer sind:

- eine erweiterbare Hadoop-Analytics-Plattform
- Datameer ist für Hadoop gebaut und nutzt seine „NoSQL-Analysis“-Leistung
- unterstützt Pervasive-BI (Vgl. Datameer-Beratung 2013b)
- Smart Analytics Funktion: Die Smart Analytics Funktion unterstützt beim Verstehen der Daten sowie bei der Erkennung von Mustern und Beziehungen in den Daten. Die vier Techniken sind in der Smart Analytics Funktion enthalten (Vgl. Datameer-Beratung 2013a; IT-Daily 2013b)

- Clustering Funktion: Finden von automatisch verwandten Gruppen und Ähnlichkeiten von Datensätzen innerhalb der Daten
- Entscheidungsbäume: Zum Verstehen der verschiedenen Kombinationen von Daten-Attributen
- Abhängigkeitsmuster: Vergleicht automatisch alle möglichen Daten Attribut-Kombinationen und stuft die Stärken dieser Beziehung visuell ein
- Empfehlungen: Automatisches Prognostizieren der Interessen einer Person

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Datameer in der folgenden Tabelle dargestellt:

Datameer - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Nutzt offene Standards und ist deshalb mit einer bereits bestehenden IT-Infrastruktur kompatibel. (Vgl. Datameer-Beratung 2013b)
Skalierbarkeit / Zukunftssicherheit	offene, auf Hadoop-basierende Plattform (Vgl. Datameer-Beratung 2013b)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• Kann mit über 40 beliebigen Datenquellen verknüpft werden und ist durch die Dateneingabe ohne jegliche Formatierungsforderungen zu bedienen. • Kann mit diversen Daten von Oracle, DB2, MS SQL, MySQL, Teradata, Greenplum, XML, JSON, CSV, HBase und Cassandra verknüpft werden. (Vgl. Datameer-Beratung 2013a)
Verarbeitung verschiedener Datenformate	Strukturierte und unstrukturierte Daten, z.B. von Twitter, Facebook, LinkedIn, Jive wie auch E-Mails, LogFiles, SaaS-CRM, GitHub, Jira und Zendesk. (Vgl. Datameer-Beratung 2013a)
Umgang mit semi- und unstrukturierten Daten	strukturierte, semistrukturierte und unstrukturierte Daten (Vgl. BARC 2013, S. 88)

Benutzeroberfläche / Benutzerfreundlichkeit	• Benutzeroberfläche mit über 250 vorgefertigten Funktionen wie der One-Klick Filter, der Sortier- und der „point-and-click“-Funktionen. (Vgl. Datameer-Beratung 2013a) • HTML5 Webbenutzeroberfläche (Vgl. Datameer-Beratung 2013b)
Software-dokumentation	• eingebautes Benutzerhandbuch (Vgl. Datameer-Beratung 2013a) • Tutorials, Webinars • Setup Guide, Developer's Guide, User's Guide (Vgl. Datameer-Beratung 2013b)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• Kunden: Apple, EMI Music, AT&T • Zusammenarbeit mit IBM und Microsoft (Vgl. Steller 2013)
Qualität des Supports	• Online Support • Webinars • Community
Produktkategorie	Speicherkonzept, Anwenderwerkzeug

Tabelle 3.5: Ergebnisse Untersuchungsmerkmale Datameer

Das Produkt Datameer ist ein Analyse-Tool für große Datenmengen, das eine erweiterbare Hadoop-Analytics-Plattform zur Verfügung stellt und so mithilfe des Hadoop-Distributed-File-Systems Daten importiert, exportiert und gespeichert werden können. Datameer ist deshalb ein Anwenderwerkzeug, das hauptsächlich für Analysen eingesetzt und außerdem für das Dashboarding und das Berichtswesen genutzt werden kann. Da es auf Hadoop basiert, ist es außerdem ein Speicherkonzept. Die entsprechende Einordnung wird in der folgenden Abbildung deutlich.

	Unternehmen	Datameer
	untersuchte Produkte bzw. Lösungen	Datameer
Kategorie	**Unterkategorie**	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	x
	Sonstige	
Anwenderwerkzeuge	Dashboarding	*
	Berichtswesen	*
	Analyse	**
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.33: Produktkategorisierung Datameer
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 172)

3.4.2.3 Datawatch

Datawatch ist ein börsennotiertes US-amerikanisches Unternehmen, das 1985 gegründet wurde und seinen Sitz in Chelmsford, Massachusetts hat. Das Unternehmen ist im Business Intelligence-Umfeld tätig und ist Spezialist für die Real-Time-Visualisierung von Daten (Vgl. Datawatch 2013d; Vgl. BARC 2013, S. 90).

Das schwedische Unternehmen Panopticon ist im September 2013 von Datawatch übernommen worden und seitdem eine hundertprozentige Tochter von Datawatch. Die Produkte von Panopticon stellen somit eine Ergänzung zum bisherigen Produktportfolio von Datawatch dar, indem eine In-Memory-Technologie zur Analyse und Real-Time-Visualisierung von großen Datenmengen bereitgestellt wird. Unter dem Namen Panopticon wird seitdem das Big Data-Produkt von Datwatch vertrieben (Vgl. Datawatch 2013a, S. 1; Vgl. BARC 2013, S. 90).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Panopticon in der folgenden Tabelle dargestellt:

Datawatch Panopticon - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Panopticon besitzt eine Vielzahl von Schnittstellen, z.B. zu Oracle, DB2, MySQL, SQL-Server, SAP Netweaver Gateway und kann somit in bestehende Architekturen integriert werden. (Vgl. Datawatch 2013c)
Skalierbarkeit / Zukunftssicherheit	k.A.

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• PDF-Dateien, EDI (Electronic Data Interchange)-Ströme, Dokumente in Ordnerstrukturen • Maschinendaten, Daten aus Social Media (Vgl. Datawatch 2013e) • Daten aus relationalen Datenbanken (Oracle, DB2, etc.) (Vgl. Datawatch 2013c)
Verarbeitung verschiedener Datenformate	Microsoft Excel, CSV-, SVG- und XML-Dateien (Vgl. Datawatch 2013c)
Umgang mit semi- und unstrukturierten Daten	strukturiert, semi-strukturiert und unstrukturiert (Vgl. Datawatch 2013e)
Benutzeroberfläche / Benutzerfreundlichkeit	• Intuitive Analysen mit einer Vielzahl von Filteroptionen und Benachrichtigungsmöglichkeiten. • Zahlreiche Möglichkeiten der Datenvisualisierung in Form von Treemaps, Heatmaps, Pie Charts, Tabellen etc. (Vgl. Datawatch 2013c; Vgl. BARC 2013, S. 90)
Software-dokumentation	k.A.
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• Credit Suisse, Europäisches Patentamt, Vodafone, XEROX • International agierende Finanzinstitute, Telekommunikationsunternehmen, Energiekonzerne und Pharmaunternehmen nutzen Panopticon. (Vgl. BARC 2013, S. 90)

Qualität des Supports	• Webinars • Video Tutorials • White Papers • Datawatch Customer Support Portal
Produktkategorie	Anwenderwerkzeug

Tabelle 3.6: Ergebnisse Untersuchungsmerkmale Datawatch Panopticon

Die Softwarelösung Panopticon bietet eine Datenanalyse aus fast jeder Datenquelle und ermöglicht Real-Time-Visualisierungen dieser Daten. Es ist dabei auf die Velocity und Variety von Big Data ausgelegt. Außerdem eignet sich die Software zur Erstellung individueller Dashboards und Ad-hoc Analysen. Da Panopticon eine In-Memory-Lösung mit Fokus auf die visuelle Analyse von Daten ist, kann es der Kategorie Anwenderwerkzeuge und dabei den Unterkategorien Dashboarding und Analyse zugeordnet werden. Diese Einordnung wird in der folgenden Tabelle deutlich.

	Unternehmen	**Datawatch**
	untersuchte Produkte bzw. Lösungen	Panopticon
Kategorie	**Unterkategorie**	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	*
	Berichtswesen	
	Analyse	**
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.34: Produktkategorisierung Datawatch
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 172)

3.4.2.4 EMC

Das Unternehmen EMC Corporation wurde im Jahr 1979 gegründet und hat seinen Hauptsitz in Hopkinton, Massachusetts in den Vereinigten Staaten (USA). Es ist ein Hersteller von Hardware und Software, der sich auf die Themen Storage, Cloud Computing und Big Data spezialisiert hat. Das Unternehmen hat Niederlassungen in rund 60 Ländern, darunter auch in Deutschland (Vgl. EMC 2013b).

Die Produktpalette von EMC für Big Data umfasst mehrere Produkte (Vgl. EMC 2013a):

- EMC ISILON: Scale-Out-Speicher für Big Data
- EMC ATMOS: Big Data-Cloud-Speicher
- Greenplum Database: Massiv-Parallel-Processing-Datenbank
- Greenplum HD: Hadoop-Stack für strukturierte und unstrukturierte Daten
- Greenplum Chorus: Analyseplattform

In der Greenplum Unified Analytics Platform (UAP) sind die Greenplum Database, Greenplum HD und Greenplum Chorus vereint. Die Greenplum Database ist eine Massively-Parallel-Processing-Datenbank, die in der Lage ist, Daten mit bis zu 10 TB/Std. zu laden. Um auch unstrukturierte Daten verarbeiten zu können, muss Greenplum HD verwendet werden, da es mit einem Hadoop-Stack arbeitet (Vgl. EMC 2013a, S. 3). Die Greenplum Database und Greenplum HD sind sich ergänzende Technologien, die zusammen für Big Data-Analysen genutzt werden sollten und laut Hersteller, die einzige Big Data-Analyseplattform ihrer Art in der Branche ist. Laut EMC ermöglicht die Kombination beider Produkte folgende Funktionen (Vgl. EMC 2011, S. 2):

- Durchführung von Ad-hoc-Abfragen der im HDFS gespeicherten Daten über die Greenplum Database. Datenbankadministratoren können vertraute SQL-Befehle für die Abfragen nutzen.
- Integration und Verarbeitung von Daten, Pull-Migration sehr großer Datasets auf Basis eines Zeitplans und Verarbeitung der Daten zum Laden in die Greenplum Database.
- Ausführen von Abfragen, die Quelldaten aus der Greenplum Database und aus dem HDFS umfassen.

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Greenplum Database und Greenplum HD in der folgenden Tabelle dargestellt:

Greenplum Database und Greenplum HD - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Anbindung an bereits vorhandene BI- und ETL-Tools möglich (Vgl. Pivotal 2013)
Skalierbarkeit / Zukunftssicherheit	skalierbar, aufgrund der Shared-Nothing bzw. MPP-Architektur von Hadoop (Vgl. Pivotal 2013)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• ODBC, JDBC, OLEDB, MapReduce • unterstützt eine Vielzahl von BI- und ETL-Tools (Vgl. Pivotal 2013)
Verarbeitung verschiedener Datenformate	k.A.
Umgang mit semi- und unstrukturierten Daten	strukturierte, semistrukturierte und unstrukturierte Daten (Vgl. EMC 2011, S. 1)
Benutzeroberfläche / Benutzerfreundlichkeit	k.A.
Softwaredokumentation	Community und Datasheets vorhanden
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• Kunden sind führende Unternehmen aus dem Finanzsektor, dem Gesundheitswesen, der Telekommunikationsbranche sowie Fluggesellschaften. (Vgl. EMC 2013b) • Laut EMC ist die Big Data-Analyseplattform die einzige ihrer Art in der Branche. (Vgl. EMC 2011, S. 2)

Qualität des Supports	• Online-Support, Communities, Livechat • Schulungen • weltweiter End-to-End-Support (Vgl. EMC 2013c; Vgl. EMC 2011, S. 2)
Produktkategorie	Speicherkonzept

Tabelle 3.7: Ergebnisse Untersuchungsmerkmale Greenplum Database & HD

Die Produktkombination von Greenplum Database und Greenplum HD stellt eine Big Data-Analyseplattform bereit. Die Greenplum Database ist dabei eine massiv-parallele Datenbank zur Speicherung und Verarbeitung polystrukturierter Daten und die Greenplum HD beinhaltet einen Hadoop-Stack für strukturierte und unstrukturierte Daten. In der folgenden Abbildung wird dabei deutlich, dass die beiden Datenbanken deshalb in die Kategorie Speicherkonzepte und in die Unterkategorie Analytische Datenbanken eingeordnet werden können.

	Unternehmen	**EMC**	
	untersuchte Produkte bzw. Lösungen	Greenplum Database	Greenplum HD
Kategorie	**Unterkategorie**		
Speicherkonzepte	Relationale Datenbanken		
	Analytische Datenbanken	x	x
	Multidimensionale Datenbanken		
	Hadoop-Distribution		
	Sonstige		
Anwenderwerkzeuge	Dashboarding		
	Berichtswesen		
	Analyse		
	Data Mining		
Datenmanagement	Datenintegration		
	Datenqualität		

Abbildung 3.35: Produktkategorisierung EMC
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 196)

3.4.2.5 EXASOL

Die EXASOL AG wurde im Jahr 2000 gegründet und hat ihren Hauptsitz in Nürnberg. Das Unternehmen hat sich auf Datenbankmanagementsysteme konzentriert und entwickelt in diesem Rahmen das Produkt EXASolution für Analytics und Data Warehousing. Schwerpunktmäßig konzentriert sich EXASOL dabei auf die Märkte Europa und Lateinamerika (Vgl. EXASOL AG 2013f).

Das Unternehmen EXASOL AG bietet erstrangig das Produkt EXASolution an. Diese analytische Datenbank dient den Bereichen Analytics und Data Warehousing und findet besonders im Kontext von Big-Data-Analysen Verwendung, da die branchenunabhängige Lösung selbst große Datenmengen analysieren kann. EXASolution soll in diesem Zusammenhang bei Datenbeständen mit einem Volumen von mehr als 1000 Terabyte zeitkritische komplexe Analysen sowie Planung und Reportings ermöglichen (Vgl. EXASOL AG 2013f). Zum Produktportfolio der EXASOL AG zählen zudem noch die Produkte EXAPowerlytics für In-Database Analytics und Hadoop-Integration zur Analyse unstrukturierter Daten sowie EXAAppliance und EXACloud. Dabei sind die beiden zuletzt genannten Produkte Erweiterungen von EXASolution. So stellt die EXAAppliance die Appliance von EXASolution dar. Diese Data Warehouse Appliance beinhaltet zusätzlich zu der Datenbank EXASolution die passende Hardware (x86-Server von Dell) sowie entsprechende Services und Support. Das Produkt EXACloud hingegen ermöglicht die Nutzung der EXASolution Datenbank als Cloud-Dienst. Dieses Datenbank-as-a-Service-Konzept wird aktuell allerdings primär von KMUs genutzt (Vgl. EXASOL AG 2013a, b, c; Platten 2013).

Das Hauptprodukt EXASolution weist nach Platten 2013 folgende Hauptmerkmale auf (Vgl. Platten 2013, S. 1f.):

- Übernimmt Funktionen eines Betriebssystems, einer relationalen und einer analytischen Datenbank,
- arbeitet mit einem 1:1 Mapping zwischen RAM und Festplatte und einer Speicher-Daten-Ratio von 1:10,
- verarbeitet strukturierte Daten (z.B. aus ERP-Systemen) sowie unstrukturierte Daten (z.B. aus Hadoop),
- bringt ein eigenes Betriebssystem mit, ist jedoch auch zusammen mit anderen Betriebssystemen kompatibel und
- läuft auf einer virtuellen Maschine (VM), allerdings mit Performanceeinbußen von 20-25 %.

EXASOL selbst verspricht dem Nutzer von EXASolution dabei folgende Mehrwerte (Vgl. EXASOL AG 2013d):

- Verknüpfte Analysen von Social-Media- und Unternehmensdaten,

- Echtzeitanalysen zur interaktiven Auswertung und Steuerung von Prozessen,
- Benutzerfreundliche Big Data-Analysen auf polystrukturierten Daten und
- High Performance bei komplexen Datenabfragen.

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von EXASolution in der folgenden Tabelle dargestellt:

EXASOL EXASolution - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• Standardschnittstellen wie JDBC ermöglichen eine einfache Integration von EXASolution in vorhandene IT-Infrastrukturen. (Vgl. EXASOL AG 2012, S. 3ff.; BARC 2013, S. 95)
Skalierbarkeit / Zukunftssicherheit	• Shared-Nothing-Architektur für parallele Datenverarbeitung • Die Plattform ist dynamisch skalierbar und kann aus verschiedenen Quellen polystrukturierte Daten in Echtzeit verarbeiten. • Exasolution ermöglicht durch ein zentrales DWH eine Verarbeitung in Echtzeit. • Der Einsatz von EXASolution wird ab einem Datenvolumen von 10 GB empfohlen. (Vgl. EXASOL AG 2012, S. 7; EXASOL AG 2013e)

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• ODBC-, JDBC-, ADO.NET- und MDX-Schnittstellen • Oracle-, LDAP-, SAP-Schnittstelle und eine generische JDBC-Ladeschnittstelle • Mit dem EXAPowerlytics Hadoop Integration Service ermöglicht EXASolution eine Verknüpfung von Data Warehouse-Daten mit Hadoop-Daten, einer nativen Verzahnung mit SQL und einem leistungsstarken Datentransfer zwischen Hadoop und EXASolution. • R wird bei EXASolution auf jeden laufenden Clusterknoten der Cluster-Architektur von EXASolution integriert.[10] • Die Schnittstelle Xtract EXA ermöglicht die Einbindung von Daten aus SAP-Systemen. (Vgl. EXASOL AG 2013e)
Verarbeitung verschiedener Datenformate	k.A.
Umgang mit semi- und unstrukturierten Daten	Es können polystrukturierte Daten und Geodaten analysiert werden. (Vgl. EXASOL AG 2013d; BARC 2013, S. 95)
Benutzeroberfläche / Benutzerfreundlichkeit	• EXASolution weist kein eigenes Frontend auf, ist jedoch mit den üblichen Frontends basierend auf gängigen Software-Lösungen kompatibel, z.B. IBM, Microsoft oder SAP. • Innerhalb von EXASolution lassen sich Rollen und Prioritäten zuweisen. (Vgl. EXASOL AG 2013e, d)
Software-dokumentation	Das EXASolution Technical White Paper ist für jeden Interessenten und Nutzer verfügbar und gibt Informationen darüber, wie EXASolution funktioniert. Es werden auch beispielhafte SQL-Abfragen demonstriert. (Siehe: EXASOL AG 2012)

[10] R ist eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken.

Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Die EXASOL AG gibt auf der Unternehmens-Homepage Auskunft über bekannte Kunden. Hier ein kleiner Auszug: Datamobile AG, Olympus, XING, Zalando, DataForce, Adidas u.v.m. (Vgl. EXASOL AG 2013d)
Qualität des Supports	• Internationaler Support-Dienst • Support-Zeiten von EXASOL: Montag bis Freitag von 08:00 – 18:00 Uhr. • Auf Wunsch können die Support-Zeiten auf 24/7 ausgeweitet werden. (Vgl. EXASOL AG 2013e)
Produktkategorie	Anwenderwerkzeug, Datenintegration

Tabelle 3.8: Ergebnisse Untersuchungsmerkmale EXASOL EXASolution

Mit der massiv-parallelen, spaltenorientierten In-Memory-Datenbank EXASolution kann der gesamte ETL-Prozess abgedeckt werden. EXASolution kann dabei als Data Mart eingesetzt werden und so eine Ergänzung zu einem bestehenden Data Warehouse darstellen. Es kann aber auch als Core Enterprise Warehouse eingesetzt werden. Die Lösung EXASolution wird deshalb der Kategorie Speicherkonzepte und der Unterkategorie Analytische Datenbanken zugeordnet sowie der Datenintegration, dies wird in der folgenden Abbildung deutlich.

	Unternehmen	**Exasol**
	untersuchte Produkte bzw. Lösungen	EXASolution
Kategorie	**Unterkategorie**	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	x
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	x
	Datenqualität	

Abbildung 3.36: Produktkategorisierung EXASOL
Quelle: Eigene Darstellung in Anlehnung an BARC (2012, S. 176), BARC (2013, S. 196), IS Report (2013, S. 100)

3.4.2.6 Hortonworks

Hortonworks ist ein im Jahre 2011 in Palo Alto gegründetes US-amerikanisches Unternehmen der Softwarebranche und ein Spin-off, also eine Ausgliederung und Verselbstständigung eines Unternehmensteils, von Yahoo (Vgl. Hortonworks 2013a). Es bietet mit dem Produkt Hortonworks Data Platform eine Hadoop Distribution an, welche auf Apache Hadoop basiert und zusammen mit einer Vielzahl von Erweiterungen in einem Package angeboten wird. Zu diesen Erweiterungen zählen u.a. Apache Pig zum Scripting, Apache Hive für SQL ähnliche Abfragen, Apache HBase als NoSQL Datenbank sowie Apache Flume zum Import und Export von Datenflüssen. Ein weiterer Bestandteil der Data Platform ist eine Weboberfläche, über die der Anwender auf diverse Hadoop-Erweiterungen und -Funktionen zugreifen kann. Des Weiteren können Objekte innerhalb des Hadoop-Clusters an einer zentralen Stelle über den HCatalog verwaltet werden. Außerdem gibt es die Möglichkeit über eine Import-Funktion u.a. semistrukturierte Daten in ein relationales Schema zu überführen (Vgl. Hortonworks 2013g).

Die Hortonworks Data Platform wird für Windows sowie Linux angeboten und ist das einzige Softwareprodukt der Firma. Neben diesem bietet Hortonworks in seinem Produktportfolio jedoch noch umfangreiche Support-Dienstleistungen für die Data Platform an (Vgl. Hortonworks 2013e). Eine Besonderheit ist das Angebot der Hortonworks Sandbox mit der jeder Nutzer die Distribution lokal auf seinem eigenen Rechner mit Hilfe einer Virtual Machine installieren und testen kann (Vgl. Hortonworks 2013h).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften der Hortonworks Data Platform in der folgenden Tabelle dargestellt:

Hortonworks Data Platform - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Anbindung an bestehende BI-Tools möglich, u.a. SAP Business Objects, MS Excel, Datameer. (Vgl. Hortonworks 2013d)
Skalierbarkeit / Zukunftssicherheit	Das Hadoop-System ist durch den Ansatz des Clusterings sehr gut skalierbar.
Verarbeitung verschiedener Datenquellen (Schnittstellen)	Anbindung über Schnittstellen an weitere Datenbanken möglich, u.a. Oracle, SAP Hana, Teradata, ODBC. Teilweise werden diese Schnittstellen durch Hadoop-Erweiterungen (z.B. Flume) bereitgestellt. (Vgl. Hortonworks 2013d)
Verarbeitung verschiedener Datenformate	Unterstützt unterschiedliche Datenformate, u.a. Access Logs (Vgl. Hortonworks 2013d)
Umgang mit semi- und unstrukturierten Daten	strukturiert, unstrukturiert (Vgl. Hortonworks 2013d)

Benutzeroberfläche / Benutzerfreundlichkeit	• Weboberfläche, zentraler HCatalog • Import-Funktion für u.a. semistrukturierte Daten (Vgl. Hortonworks 2013g)
Software-dokumentation	• Umfangreiche Dokumentation (Vgl. Hortonworks 2013f) • Eigene und aktive Community auf der Homepage des Herstellers mit Tutorials, Foren und Webinars (Vgl. Hortonworks 2013i)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• u.a. eBay, Samsung, XING, Spotify (Vgl. Hortonworks 2013j) • Laut Forrester Research ein technologischer Marktführer im Bereich Hadoop. (Vgl. Hortonworks 2013j)
Qualität des Supports	• Verschiedene Support-Arten im Produktportfolio (u.a. 24/7) • Support wird direkt von Hortonworks angeboten (Vgl. Hortonworks 2013e)
Produktkategorie	Speicherkonzept

Tabelle 3.9: Ergebnisse Untersuchungsmerkmale Hortonworks Data Platform

Die Hortonworks Data Platform ist eine der führenden Hadoop Distributionen auf dem Markt und das einzige Produkt des Anbieters Hortonworks. Durch die bereits integrierten Hadoop-Erweiterungen bildet es bereits von Haus aus ein breites Spektrum von Funktionen ab und erspart dem Benutzer somit zusätzlichen Integrationsaufwand. Auch weitere Eigenschaften wie eine Weboberfläche oder Wizards für beispielsweise Import-Funktionen führen zu einer hohen Benutzerfreundlichkeit. Da Hortonworks Data Platform eine Hadoop-Datenbank ist, die eine verteilte Bearbeitung riesiger Datenmengen ermöglicht, kann es der Kategorie Speicherkonzepte und der Unterkategorie Hadoop-Distribution zugeordnet werden. Dies ist in der folgenden Abbildung verdeutlicht.

	Unternehmen	Hortonworks
	untersuchte Produkte bzw. Lösungen	Data Platform
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	x
	Multidimensionale Datenbanken	
	Hadoop-Distribution	x
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.37: Produktkategorisierung Hortonworks
Quelle: Eigene Darstellung in Anlehnung an IS Report (2013, S. 100)

3.4.2.7 Hewlett-Packard

Das Unternehmen Hewlett-Packard (HP) wurde 1939 in Palo Alto, Kalifornien, USA, gegründet und ist eines der führenden IT-Unternehmen. Es ist in rund 170 Ländern tätig und beschäftigt rund 300.000 Mitarbeiter (Vgl. XING 2013). Das Portfolio von HP ist in Privatkunden (u.a. Drucker, Laptops, Tablets) und Geschäftskunden (u.a. Server, Software, Drucker) aufgeteilt. In beiden Bereichen bietet HP auch vielseitige Dienstleistungen an (Vgl. HP 2013b).

Ein Produkt, das von HP im Kontext von Big Data hauptsächlich genannt wird, ist die Vertica Analytics Platform. Dabei handelt es sich um ein spaltenorientiertes SQL Datenbankmanagementsystem (DBMS) (Vgl. HP 2013a). Durch die Integration von Vertica Flex Zone ist die Untersuchung und Auswertung von semi-strukturierten Daten möglich.

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Vertica Analytics Platform in der folgenden Tabelle dargestellt:

HP Vertica Analytics Platform - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• Aufgrund einer Vielzahl von Schnittstellen und der Kompatibilität mit Windows, Linux, Solaris, AIX etc. kann Vertica in bestehende Architekturen integriert werden. (Vgl. Vertica 2013b; Vgl. HP)
Skalierbarkeit / Zukunftssicherheit	Leicht skalierbar aufgrund der MPP-Architektur, sodass Server oder Storage-Komponenten zum bestehenden Cluster hinzugefügt werden können. (Vgl. Vertica 2013c)

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• ODBC, JDBC, ADO.Net • Schnittstellen zu Produkten von IBM, Pentaho, Cognos, MicroStrategy, Tableau und dem Informatica PowerCenter Plug-in • Schnittstelle zu Hadoop (Vgl. Vertica 2013b; Vgl. Vertica 2013a, S. 4)
Verarbeitung verschiedener Datenformate	Daten aus Social Media, Logfiles, Maschinendaten (Vgl. HP; Vgl. Vertica 2013a, S. 1)
Umgang mit semi- und unstrukturierten Daten	strukturierte und semi-strukturierte Daten (Vgl. HP)
Benutzeroberfläche / Benutzerfreundlichkeit	• web-basierte Umgebung für das Management des Vertica Clusters ermöglicht einfaches Erstellen und Bearbeiten der Datenbank(en) • „One-Click-Schemata“ zum Erstellen von Hochleistungs-Auswertungen (Vgl. Vertica 2013a, S. 3; Vgl. HP)
Software-dokumentation	k.A.
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	HP ist eines der führenden Technologieunternehmen der IT-Branche
Qualität des Supports	• MyVertica-Community, Supportforen • Schulungen und Zertifizierungen • Treiber und Downloads (Vgl. HP 2013c)
Produktkategorie	Speicherkonzept

Tabelle 3.10: Ergebnisse Untersuchungsmerkmale HP Vertica Analytics Platform

Die Vertica Analytics Platform von HP ermöglicht eine Untersuchung und Auswertung sowohl von strukturierten als auch semi-strukturierten Daten. Das Produkt zeichnet sich durch viele Schnittstellen und laut Hersteller durch die Lauffähigkeit auf allen gängigen Hadoop-Distributionen aus. Es handelt sich hierbei um ein Speicherkonzept.

	Unternehmen	**HP**
	untersuchte Produkte bzw. Lösungen	Vertica Analytics Platform
Kategorie	**Unterkategorie**	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	x
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.38: Produktkategorisierung Hewlett-Packard
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 196), BARC (2012, S. 176)

3.4.2.8 IBM

Das multinationale Unternehmen IBM hat für Kunden eine Big-Data-Plattform entwickelt, die viele Disziplinen im Bereich Big Data beherrscht. Dazu wurden bestehende Technologien für strukturierte Routineaufgaben mit neuen Technologien, die neue Möglichkeiten in puncto Geschwindigkeit und Flexibilität bieten und sich für die Ad-hoc-Datenuntersuchung, -erkennung und die Analyse unstrukturierter Daten eignen, kombiniert. Die neue IBM Plattform umfasst vier zentrale Funktionen: die Hadoop-basierte Analyse, Stream-Computing, Data Warehousing sowie Informationsintegration und –governance (Vgl. IBM 2013a). Nach Herstellerangaben besteht die Plattform aus folgenden zentralen Funktionen (Vgl. IBM 2013a):

- **Hadoop-basierte Analyse:** Verarbeitung und Analyse verschiedener Datentypen in Clustern aus gängigen Servern.
- **Stream-Computing:** Kontinuierliche Analyse großer Mengen von Datenströmen mit Reaktionszeiten von unter einer Millisekunde.
- **Data Warehousing:** Umfassende betriebliche Erkenntnisse durch datenbankinterne Analyse.

- **Informationsintegration und -governance:** Gewinnung verlässlicher Informationen sowie Bereinigung, Konvertierung, Steuerung und Bereitstellung dieser Informationen für Geschäftsprozesse.

Diese Funktionen werden durch folgende Dienste umgesetzt (Vgl. IBM 2013a):

- **Visualisierung und Erkennung:** Unterstützung der Endbenutzer bei der Untersuchung großer und komplexer Datasets.
- **Anwendungsentwicklung:** Optimierung der Entwicklung von Big-Data-Anwendungen.
- **Systemmanagement:** Überwachung und Management von Big-Data-Systemen, um sichere, optimierte Leistung zu erreichen.
- **Beschleuniger:** Schnellere Wertschöpfung mit analyseorientierten und branchenspezifischen Modulen.

IBM InfoSphere BigInsights

IBM InfoSphere BigInsights Enterprise Edition stellt Funktionen für die Big Data-Analyse auf einer Unternehmensplattform bereit. Sie kombiniert das Framework Apache Hadoop mit Unternehmensfunktionalität und Integration, damit eine umfangreiche Analyse mit integrierter Ausfallsicherheit und Fehlertoleranz durchgeführt werden kann. Die Software unterstützt strukturierte, semistrukturierte und unstrukturierte Daten in ihrem nativen Format (Vgl. IBM 2013e).

IBM InfoSphere Streams

IBM InfoSphere Streams ist eine IT-Plattform, die dafür sorgt, dass benutzerspezifisch entwickelte Anwendungen in kurzer Zeit Informationen aus Echtzeitquellen aufnehmen, analysieren und korrelieren können. Die Anwendung unterstützt einen hohen Datendurchsatz von bis zu mehreren Millionen Ereignissen oder Nachrichten pro Sekunde (Vgl. IBM 2013f).

IBM Netezza

Bei der Produktfamilie IBM Netezza handelt es sich um speziell konzipierte Data-Warehouse-Appliances, die Datenbank, Server, Speicher und erweiterte Analysefunktionen vereinen. Das Betriebssystem basiert auf einem Red Hat Linux und verfügt über ein eigenes DBMS (Vgl. IBM 2013d).

IBM Cognos Business Intelligence

IBM Cognos Business Intelligence stellt Berichte, Analysefunktionen, Dashboards und Scoreboards bereit und unterstützt die Darstellung der Geschäftsprozesse und -leistung. Mit dieser Applikation können Informationen untersucht, wichtige Fakten analysiert und zusammengefasst werden. Dashboards geben den Benutzern die Möglichkeit, auf Inhalte zuzugreifen, sie interaktiv zu nutzen und zu personalisieren.

Funktionen für die Analyse ermöglichen den Informationszugriff aus verschiedenen Blickwinkeln

und Perspektiven, um eine Entscheidungsgrundlage darzustellen. Funktionen für die Zusammenarbeit umfassen Kommunikationstools und Social Networking, um den Austausch von Ideen während der Entscheidungsfindung zu fördern. Scorecarding-Funktionen automatisieren die Erfassung, das Management und die Überwachung von geschäftlichen Kennzahlen, damit strategische und operative Ziele verglichen werden können (Vgl. IBM 2013b).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von InfoSphere Streams, InfoSphere BigInsights, Netezza Data Warehouse Appliance und Cognos Business Intelligence in der folgenden Tabelle dargestellt:

InfoSphere Streams, InfoSphere BigInsights, Netezza Data Warehouse Appliance, Cognos Business Intelligence - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• InfoSphere BigInsights verfügt über verschiedene Hadoop-Funktionen, ein zentraler Bestandteil ist das Text-Analyse-Modul zur Analyse von großen unstrukturierten Daten. Es handelt sich um eine ergänzende Hadoop-Plattform zu einer bestehenden BI-Architektur (Vgl. IBM 2014a). • InfoSphere Streams ermöglicht eine kontinuierliche Verarbeitung und Überwachung von Dateneingängen wie Emails, Geschäftsdaten, Social Media, GPS-Tracking etc. in großen Mengen (Proof-of-Concepts mit einer Analyse von über 100 MB pro Sekunde und mehr als 10 Millionen Datensätzen pro Stunde). Löst anhand der Erkennung von Ereignissen neue Tätigkeiten aus bzw. bereitet Daten auf. (Vgl. IBM 2013c)
Skalierbarkeit / Zukunftssicherheit	• InfoSphere Streams: Erweiterbar durch verteiltes System • Netezza: Benutzerdatenkapazität von 1 Terabyte auf 1,5 Petabyte (Vgl. IBM 2013d) • Cognos Business Intelligence: Erweiterbar durch weitere Server, Reportgeschwindigkeit hängt von Datenquelle ab

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• InfoSphere BigInsights: IBM DB2, IBM Netezza, JDBC • InfoSphere Streams: alle streambaren Daten • Netezza: diverse • Cognos Business Intelligence: alle gängigen RDBMS, InfoSphere BigInsights
Verarbeitung verschiedener Datenformate	Netezza, InfoSphere BigInsights und InfoSphere Streams: unstrukturierte und strukturierte Daten.
Umgang mit semi- und unstrukturierten Daten	Netezza, InfoSphere BigInsights und InfoSphere Streams: Volle Unterstützung durch Text-Analyse-Modul
Benutzeroberfläche / Benutzerfreundlichkeit	• Netezza, InfoSphere BigInsights und InfoSphere Streams: Verschiedene, Web-Konsole, Eclipse-IDE für die Text-Analyse • Netezza: Datenbanktools
Software-dokumentation	Online-Dokumentation, White-Paper, Developer Paper
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Viele Kunden in verschiedenen Ländern, z. B. Vestas Wind Systems, Fraunhofer Institut, WIdO (Wissenschaftliches Institut der AOK), Berufsgenossenschaft Rohstoffe und chemische Industrie (BG RCI), Oberbank aus Linz. (Vgl. IBM 2013a)
Qualität des Supports	weltweiter Technischer Support, Customer Support, Premium Support, Passport Advantage, Accelerate Value, PartnerWorld for Software, SoftwareXcel Enterprise Edition, Support Line (Vgl. IBM 2014c)

Produktkategorie	• Netezza, InfoSphere BigInsights und InfoSphere Streams: Speicherkonzept, Anwenderwerkzeug, Datenmanagement • Cognos Business Intelligence: Speicherkonzept, Anwenderwerkzeug

Tabelle 3.11: Ergebnisse Untersuchungsmerkmale IBM InfoSphere BigInsights & Streams, Netezza, Cognos BI

Die Plattform InfoSphere BigInsights kann den Kategorien Speicherkonzepte, Anwenderwerkzeuge und Datenmanagement zugeordnet werden. Zum einen ist es eine Softwareplattform zur Speicherung großer Datenmengen, zum anderen ist es auch eine Plattform zur Analyse und zum Sammeln von Erkenntnissen in großen Datenmengen. Weiterhin ist es auch eine Lösung zur Speicherung großer Datenmengen von polystrukturierten Daten aus verschiedenen Quellen. Es ist außerdem eine Hadoop-Distribution.

InfoSphere Streams besteht aus einer Anwendung mit Benutzeroberfläche zur adhoc-Analyse von Streams und zur Erstellung von Batch-Prozessen und einer Server-Lösung zur Ausführung der Prozesse und ist somit in der Einordnung ein Anwenderwerkzeug als auch ein Tool zum Datenmanagement. Es kann deshalb den Unterkategorien Analyse und Data Mining zugeordnet werden. Da es auch Streaming-Daten integriert und die Qualität der Daten verbessern kann, wird es folglich in die Kategorie Datenmanagement eingeordnet.

Netezza ist eine Data-Warehouse- und Analyse-Appliance, die Datenbank, Server, Speicher und erweiterte Analysefunktionen in einem System vereint, sodass sie den Kategorien Speicherkonzept, Anwenderwerkzeug und Datenmanagement zugeordnet werden kann.

Die Cognos Business Intelligence stellt Berichte, Analysefunktionen, Dashboards und Scorecards bereit und wird deshalb bei Anwenderwerkzeugen eingeordnet. Es beinhaltet außerdem die multidimensionale Datenbank Dynamic Cubes (Vgl. IBM Redbooks 2012). Die detaillierten Einordnungen der einzelnen Produkte in die Unterkategorie können der nachfolgenden Abbildung entnommen werden.

	Unternehmen	IBM			
	untersuchte Produkte bzw. Lösungen	InfoSphere BigInsights	InfoSphere Streams	Netezza	Cognos Business Intelligence
Kategorie	Unterkategorie				
Speicherkonzepte	Relationale Datenbanken				
	Analytische Datenbanken	x		x	
	Multidimensionale Datenbanken				x
	Hadoop-Distribution	x			
	Sonstige		x		
Anwenderwerkzeuge	Dashboarding				••
	Berichtswesen				••
	Analyse	•	•	•	••
	Data Mining	•	•		•
Datenmanagement	Datenintegration	x	x	x	
	Datenqualität	x	x	x	

Abbildung 3.39: Produktkategorisierung IBM
Quelle: Eigene Darstellung in Anlehnung an IS Report (Vgl. 2013, S. 101), BARC (Vgl. 2013, S. 174, 196)

3.4.2.9 Informatica

Informatica ist eine US-amerikanische Aktiengesellschaft, die 1993 gegründet wurde und weltweit agiert (Vgl. Informatica 2013g). Das Unternehmen bietet verschiedene Software und Dienstleistungen in den folgenden Bereichen an (Vgl. Informatica 2013f):

- Lebenszyklusmanagement von Daten in Anwendungen
- B2B-Datenaustausch
- Integration von Cloud-basierten Daten
- Verarbeitung komplexer Ereignisse
- Maskierung von Daten[11]
- Datenqualität, -replikation und -virtualisierung
- Stammdatenverwaltung
- Nachrichtenübermittlung

Das Informatica PowerCenter dient der Integration von Unternehmensdaten und wird in vier verschiedenen Versionen angeboten (Vgl. Informatica 2013e):

- Standard Edition
- Advanced Edition
- Big Data Edition
- Real Time Edition

[11] Schutz von sensiblen oder vertraulichen Daten vor unbefugtem Zugriff

Die PowerCenter Big Data Edition ist im Gegensatz zu den anderen Versionen auf die Integration und Verarbeitung von Big Data spezialisiert. Sie bietet eine visuelle Entwicklungsplattform, unterstützt die Wiederverwendung von Ressourcen und die Zusammenarbeit mit einer gemeinsamen integrierten Entwicklungsumgebung. Es wird der Zugriff auf Big Data aus Transaktionsdaten (z. B. RDBMS, OLTP, OLAP, ERP, CRM, Mainframe, Cloud usw.) und Interaktionsdaten (z. B. Daten aus sozialen Medien, Protokolldateien, maschinelle Messdaten, Websites, Blogs, Dokumente, E-Mails und andere unstrukturierte oder mehrfach strukturierte Daten) ermöglicht. Der Zugriff auf und die automatische Syntaxanalyse (Parsing) von komplexen, mehrfach strukturierten, unstrukturierten Daten (z. B. Blogdaten, JSON, XML, Gerätedaten) wird vereinfacht und vorkonfigurierte Parser für Marktdaten und Industriestandards (FIX, SWIFT, ACORD, HL7, HIPAA, EDI) werden geboten. Es wird die Extraktion und Klassifizierung von Dateneinheiten aus unstrukturierten Daten und die Anreicherung von Stammdaten mit Informationen zu Kundenverhalten oder Produkten erlaubt. Mit der Lösung ist das Abrufen, Laden, Replizieren, Transformieren und Extrahieren von Big Data sowie das Erfassen und Extrahieren von Daten über native APIs mit paralleler Verarbeitung möglich. Es wird die Datenverarbeitung auf SMP-Rechnern, in herkömmlichen Grid-Clustern, auf verteilten Computing-Plattformen wie Hadoop oder Data Warehouse-Appliances unterstützt. Es kann Big Data verarbeitet werden und es wird eine Bibliothek vorgefertigter Transformationsfunktionen für Hadoop geboten. Durch ein Failover (Wechsel bei Ausfall von Hard-/Software), einem flexiblen Wiederanlauf und einer Ausfallsicherheit für Verbindungen kann eine nahezu durchgängige Verfügbarkeit gewährleistet werden (Vgl. Informatica 2013a).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften der Informatica PowerCenter Big Data Edition in der folgenden Tabelle dargestellt:

Informatica PowerCenter Big Data Edition - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• Unterstützt neu aufkommende Technologien wie Hadoop und herkömmliche Datenverwaltungsinfrastrukturen. • Über native APIs können Verbindungen zu Quell- und Zielsystemen hergestellt werden. (Vgl. Informatica 2012)
Skalierbarkeit / Zukunftssicherheit	Höchst skalierbar, da Daten bis zur Größe von Petabytes verarbeitet werden können. (Vgl. Informatica 2013a)

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• Zugriff auf Big Data aus Transaktionsdaten, z. B. RDBMS, OLTP, OLAP, ERP, CRM, Mainframe, Cloud usw. • Zugriff auf Big Data aus Interaktionsdaten, z. B. Daten aus sozialen Medien, Protokolldateien, maschinelle Messdaten, Websites, Blogs, Dokumente, E-Mails und andere unstrukturierte oder mehrfach strukturierte Daten. (Vgl. Informatica 2013a) • Über native APIs können Verbindungen zu Quell- und Zielsystemen hergestellt werden. (Vgl. Informatica 2012)
Verarbeitung verschiedener Datenformate	Daten aus sozialen Medien, Protokolldateien, maschinelle Messdaten, Websites, Blogs, Dokumente, E-Mails und andere unstrukturierte oder mehrfach strukturierte Daten. (Vgl. Informatica 2013a)
Umgang mit semi- und unstrukturierten Daten	Ermöglicht Zugriff auf und das Parsing von komplexen, mehrfach strukturierten, unstrukturierten und branchenüblichen Daten, z. B. Blogdaten, JSON, XML sowie Gerätedaten. (Vgl. Informatica 2013a)
Benutzeroberfläche / Benutzerfreundlichkeit	Keine manuelle Programmierung mehr, sondern arbeitet mit einer codefreien visuellen Entwicklungsumgebung, die vorgefertigte Transformationen bietet und die Wiederverwendung von Ressourcen fördert. (Vgl. Informatica 2013a)
Software-dokumentation	Online-Wissensdatenbank mit mehr als 35.000 Artikeln. (Vgl. Informatica 2013b)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Weltweit nutzen über 4.630 Unternehmen Informatica. (Vgl. Informatica 2013d)

Qualität des Supports	• Informatica Global Customer Support Programs (Standard Support, Enterprise Support, Mission-Critical Support) • Probleme können auf myinformatica.com jederzeit gemeldet werden • Global Support Centers in Nord- und Südamerika, Europa und Asien (Vgl. Informatica 2013b)
Produktkategorie	Datenmanagement

Tabelle 3.12: Ergebnisse Untersuchungsmerkmale Informatica PowerCenter Big Data Edition

Informatica PowerCenter Big Data Edition ist eine Software, die Big Data aus verschiedensten Datenquellen extrahieren und transformieren kann, um sie anschließend für die Weiterverarbeitung bereitzustellen. Sie ist deshalb in die Kategorie Datenmanagement und die Unterkategorie Datenintegration einzuordnen.

	Unternehmen	Informatica
	untersuchte Produkte bzw. Lösungen	PowerCenter Big Data Edition
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	x
	Datenqualität	

Abbildung 3.40: Produktkategorisierung Informatica
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 186), IS Report (2013, S. 114)

3.4.2.10 Jaspersoft

Das Unternehmen Jaspersoft ist ein US-amerikanisches Unternehmen mit Sitz in San Francisco, Kalifornien, das 2001 gegründet wurde und Business Intelligence Software entwickelt sowie vertreibt (Vgl. InsideView 2013).

Die Produktpalette von Jaspersoft besteht aus mehreren einzelnen Tools, die in verschiedenen Kombinationsmöglichkeiten in Suiten zu erwerben sind. Die Produkte decken unterschiedliche Bereiche ab und lassen sich folgendermaßen aufteilen (Vgl. Jaspersoft 2013a):

- Reportingsoftware
- Analyse
- Dashboards
- Datenintegration
- BI-Plattform

Eine Lösung von Jaspersoft für den Umgang mit Big Data ist Instant Big Data Analytics. Das Produkt kann Daten aus verschiedenen RDBMS extrahieren und diese in Datenspeichern wie Hadoop, Cassandra oder MongoDB speichern. Die Daten können dann für Analyse und Reporting genutzt werden.

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Instant Big Data Analytics in der folgenden Tabelle dargestellt:

Jaspersoft Instant Big Data Analytics - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Aufgrund der verschiedenen Schnittstellen lässt sich Instant Big Data Analytics in bereits bestehende Architekturen integrieren. (Vgl. Jaspersoft 2013e)
Skalierbarkeit / Zukunftssicherheit	lässt sich in bereits bestehende Architekturen sowie in der Cloud skalieren (Vgl. Jaspersoft 2013e)

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• Jedes RDBMS, das mit JDBC 2.1 und SQL-92 kompatibel ist (hierzu gehören die meisten relationalen Datenbanken). • Andere Datenquellen: Bean, JNDI, JasperAnalysis (Mondrian), XML/A, kundenspezifische Datenquellen (z. B. Hibernate, XML usw.) (Vgl. Jaspersoft 2013e) • Hadoop, Cassandra, MongoDB (Vgl. Jaspersoft 2013b)
Verarbeitung verschiedener Datenformate	k.A.
Umgang mit semi- und unstrukturierten Daten	Polystrukturierte Daten
Benutzeroberfläche / Benutzerfreundlichkeit	• Grafische Design-Umgebung • Prozess-Design-Tool mit Drag & Drop • Aktivitätsüberwachungs-Dashboard zur Nachverfolgung der Job-Ausführung und der Performance
Software-dokumentation	• verschiedene User Guides • Forum und Community • Wiki (Vgl. Jaspersoft 2013c)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• 14.000 Handelskunden in 100 Ländern • Community mit 225.000 Mitgliedern (Vgl. Jaspersoft 2013c)

Qualität des Supports	• umfangreiche Supportmöglichkeiten (Express, Standard, Premium) (Vgl. Jaspersoft 2013d)
Produktkategorie	Datenmanagement

Tabelle 3.13: Ergebnisse Untersuchungsmerkmale Jaspersoft Instant Big Data Analytics

Mit Instant Big Data Analytics können Big-Data-Quellen extrahiert werden und in Form von Berichten, Dashboards und Analysen visualisiert werden. Dabei können verschiedene Datenspeicher wie MongoDB, Hadoop oder Amazon Web Services genutzt werden. Aufgrund der Analysefunktion wird das Produkt der Kategorie Anwenderwerkzeuge zugeordnet.

	Unternehmen	Jaspersoft
	untersuchte Produkte bzw. Lösungen	Instant Big Data Analytics
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	•
	Berichtswesen	•
	Analyse	•
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.41: Produktkategorisierung Jaspersoft

3.4.2.11 MicroStrategy

MicroStrategy ist ein 1989 gegründetes US-amerikanisches Unternehmen mit Sitz in Tysons Corner, Virginia. Das Unternehmen bietet mit Analytics, Mobile, Identity, Loyalty und Cloud verschiedene Versionen seiner BI-Plattform an.

Die Plattform, die aktuell in der Version 9.3 vertrieben wird, ermöglicht die Erstellung von Scorecards und Dashboards, Enterprise-Reporting, erweiterte und prädiktive Analysen, Closed-Loop-BI und bietet im Bereich Big Data ein Konzept für die Verwaltung komplexer Datenbestände (Vgl. MicroStrategy 2013c).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von MicroStrategy 9.3 in der folgenden Tabelle dargestellt:

MicroStrategy 9.3 - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Bietet zahlreiche Schnittstellen (s.u.) zur Integration in die bestehende Architektur.
Skalierbarkeit / Zukunftssicherheit	Skalierbarkeit durch asymmetrisches Clustering, Shared Caching, In-Memory-Cubes und eine sofortige Synchronisierung von Metadatenobjekten und Sicherheitseinstellungen. (Vgl. MicroStrategy 2013d)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• Oracle, DB2, SQL Server, Sybase, Greenplum, MySQL, PostgreSQL, Aster nCluster, SQL-92, Teradata, Informix, NeoView, NonStop, Netezza, Kognito WX2, Red Brick, Vertica • SAP BW, Microsoft Analysis Services, Hyperion Essbase • Schnittstelle zu Hadoop (Vgl. MicroStrategy 2013b, d, f)
Verarbeitung verschiedener Datenformate	Spreadsheets, Web Services, Text Files und Datenformate der oben angegebenen Quellen (Vgl. MicroStrategy 2013d)
Umgang mit semi- und unstrukturierten Daten	Verarbeitung von strukturierten und unstrukturierten Daten, die in Hadoop gespeichert sind. (Vgl. Finanznachrichten 2012)
Benutzeroberfläche / Benutzerfreundlichkeit	Oberfläche mit Navigation in einer Ordnerstruktur, Menübänder, Akkordeon-Steuerelementen, Mehrfachauswahl durch STRG+Klicken sowie Kontextmenüs und -aktionen. (Vgl. MicroStrategy 2013d)
Software-dokumentation	Dokumente von der Produktarchitektur bis zur Projektimplementierung, Handbücher, Readme-Dateien usw. (Vgl. MicroStrategy 2013e)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	MicroStrategy hat weltweit knapp 4000 Kunden (in Deutschland z. B. Deutsche Bank, Deutsche Telekom etc.) (Vgl. MicroStrategy 2013a)

Qualität des Supports	• Verschiedene Support-Pakete (Standard, Extended, Managed, Dedicated, Elite) • Knowledge Base • Diskussionsforum (Vgl. MicroStrategy 2013e)
Produktkategorie	Anwenderwerkzeug

Tabelle 3.14: Ergebnisse Untersuchungsmerkmale MicroStrategy 9.3

MicroStrategy 9.3 ist eine BI-Plattform und bietet die klassischen BI-Funktionen wie Analysen, Reporting und Dashboards. Sie gehört damit in die Kategorie der Anwenderwerkzeuge. Wie in der nachfolgenden Abbildung zu sehen ist, liegt der Produktschwerpunkt in den Unterkategorien Dashboarding, Berichtswesen und Analyse. Zusätzlich wird die Funktion des Data Mining abgedeckt.

	Unternehmen	MicroStrategy
	untersuchte Produkte bzw. Lösungen	MicroStrategy 9.3
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	••
	Berichtswesen	••
	Analyse	••
	Data Mining	•
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.42: Produktkategorisierung MicroStrategy
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 176)

3.4.2.12 Oracle

Die Oracle Corporation ist ein weltweit agierender Software- und Hardwarehersteller und hat bisweilen ca. 400.000 Kunden in über 45 Ländern. Gegründet wurde Oracle 1977 in Kalifornien, Vereinigte Staaten. Der Hauptsitz des Unternehmens liegt in Redwood Shores (Silicon Valley, Kalifornien). Vorrangig war Oracle bis zum Jahre 2010 ein reiner Softwarehersteller, was sich durch die Übernahme von Sun Microsystems im Januar 2010 änderte und sich Oracles Produktportfolio um Hardware erweiterte (Vgl. o.A. 2012; Vgl. Oracle Corporation 2013e, S. 1). Mittlerweile weist Oracle somit ein breites Spektrum an Hardware und Software auf. Generell bietet Oracle Produkte in folgenden Produktbereichen an (Vgl. Oracle Corporation 2013d):

- Datenbanken
- Anwendungen
- Engineered Systems
- Unternehmensmanagement
- Entwicklungstools
- Betriebssysteme u.v.m.

Im Bereich Big Data gestaltet Oracle sein Produktportfolio in die Bereiche Datenintegration, Datenmanagement und Datenanalyse. Eine Komplettlösung stellt beispielsweise die Big Data Appliance von Oracle dar, die Hard- und Software für die Integration, das Datenmanagement und die Datenanalyse von Big Data kombiniert und ist im Produktbereich der Engineered Systems zu finden. Diese Appliance ist für den Erwerb, die Organisation und das Laden unstrukturierter Daten in Oracle entworfen worden und soll wie die anderen Engineered Systems auch, die Kosten und Komplexität von IT-Infrastrukturen senken sowie gleichwohl die Leistungsfähigkeit jener steigern (Vgl. Oracle Corporation 2013d; Oracle Corporation 2013f). Die mitgelieferten Server der Appliance weisen dabei eine Gesamtkapazität von 648 Terabyte auf. Weiter werden in der Appliance Open Source und kostenpflichtige Software kombiniert. Insgesamt umfasst die Oracle Big Data Appliance folgende Softwarekomponenten (Vgl. Dijcks 2013, S. 8ff.; Vgl. Oracle Corporation 2013b, S. 1f.):

- Oracle Linux 6.4 Betriebssystem
- Oracle Big Data Appliance Plug-In für den Enterprise Manager
- Cloudera Software
- Cloudera Manager
- Oracle NoSQL Database Community Edition
- Open Source Distribution von R

- Oracle Java - JDK7

Optionale Software besteht dabei aus (Vgl. Oracle Corporation 2013b, S. 2)

- Oracle Connectoren für Hadoop (sogenannte Oracle Big Data Connectors): SQL, Loader (OLH), XQuery, R, Data Integrator Application Adapter
- Oracle Audit Vault and Database Firewall for Hadoop Auditing
- Oracle Data Integrator
- Oracle NoSQL Database Enterprise Edition

Spezialisiert ist die Oracle Big Data Appliance auf Big Data Analysen mit einer möglichst hohen Leistung, Verfügbarkeit, Sicherheit und Unterstützung verschiedenster Datentypen. Hadoop und die Datenbank NoSQL bilden dabei die Basis der Appliance. Damit unterscheidet sich diese Appliance von anderen Oracle Komplettsystemen, bei denen üblicherweise eigene (relationale) Oracle Datenbanken zum Einsatz kommen. Diese Big Data Lösung soll gut geeignet sein für schnell wachsende Webdienste mit interaktiven Seiten (Vgl. Manhart 2012). „Die Kombination aus Hadoop und NoSQL könnte es Unternehmen ermöglichen, unstrukturierte Daten aus dem Web zu ziehen und daraus aussagekräftige Berichte zu generieren“ (Manhart 2012).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften der Big Data Appliance von Oracle in der folgenden Tabelle dargestellt:

Oracle Big Data Appliance - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Durch das Komplettsystem wird eine vollständige Architektur geliefert, sodass eine einfache Integration möglich ist. (Vgl. Oracle Corporation 2013b, S. 1)
Skalierbarkeit / Zukunftssicherheit	k.A.
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• diverse Connectoren für Hadoop • Cloudera Connectoren • Oracle Connectoren für Oracle-Produkte wie z.B. Oracle Database • HBase (Vgl. Oracle Corporation 2013b, S. 1ff.)

Verarbeitung verschiedener Datenformate	• JSON • XML • Avro • weitere Formate (Vgl. Oracle Corporation 2013b, S. 1)
Umgang mit semi- und unstrukturierten Daten	strukturierte, unstrukturierte und polystrukturierte Daten (Vgl. Manhart 2012)
Benutzeroberfläche / Benutzerfreundlichkeit	k.A.
Softwaredokumentation	Oracle bietet eine Reihe an Informationen zu den Big Data Produkten und den Komponenten der Big Data Appliance an, u.a. Videos, die in diesem Kontext den Umgang mit Hadoop erklären.[12]
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• Oracle ist einer der führenden Softwarehersteller in der Branche und kann dabei über 400.000 Kunden in über 145 Ländern zu seinem Kundenstamm zählen. Davon beziehen allein 310.000 Kunden Datenbanken. • Oracle beschäftigt über 120.000 Mitarbeiter (Vgl. Oracle Corporation 2013e, S. 1)

[12] siehe z.B. hier: `http://www.oracle.com/de/products/database/big-data-appliance/resources/index.html`

Qualität des Supports	• Oracle Advanced Customer Support Services und Oracle Premier Support bieten einen integrierten, unternehmenskritischen Support für komplexe IT-Umgebungen, um die Leistung zu maximieren, eine höhere Verfügbarkeit zu erreichen und Risiken zu minimieren. • Oracle Cloud-Services für die Cloud-Nutzer: Möglichst kostengünstige Nutzung, Risikominimierung, Leistungsbeschleunigung. (Vgl. Oracle Corporation 2013e, S. 2)
Produktkategorie	Die Big Data Appliance von Oracle deckt alle aufgestellten Kategorien ab: Speicherkonzept, Anwenderwerkzeug und Datenmanagement

Tabelle 3.15: Ergebnisse Untersuchungsmerkmale Oracle Big Data Appliance

Die Big Data Appliance von Oracle umfasst sowohl Hardware als auch Software für den Umgang mit Big Data und stellt sich somit als eine vollständige Lösung für große unterschiedliche strukturierte Datenmengen heraus. Aufgrund der Tatsache, dass es sich hierbei um ein Komplettsystem handelt, werden zugleich mehrere Kategorien abgedeckt, wie sich der nachfolgenden Grafik entnehmen lässt:

	Unternehmen	**Oracle**
	untersuchte Produkte bzw. Lösungen	Big Data Appliance
Kategorie	**Unterkategorie**	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	x
	Sonstige	x
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	**
	Data Mining	
Datenmanagement	Datenintegration	x
	Datenqualität	

Abbildung 3.43: Produktkategorisierung Oracle

Da der Oracle Data Integrator und die Oracle Data Integrator Applikation für Hadoop die Datenintegration ermöglichen, kann hier eine entsprechende Einordnung vorgenommen werden. Die Oracle Big Data Connectors sind allesamt Hadoop-Connectoren, da die Appliance zwecks der Analyse auf Hadoop basiert. Die Konnektoren sind zwar optional, allerdings bleibt es bei der Hadoop-Basis (Vgl. Dijcks 2013, S. 8ff.; Manhart 2012; Oracle Corporation 2013b, S. 2). Durch die Nutzung von NoSQL kann ebenfalls eine Einordnung in sonstige Speicherkonzepte vorgenommen werden. Ebenso ist die Oracle Big Data Appliance aufgrund der bislang aufgeführten Ergebnisse zu den Untersuchungsmerkmalen den Kategorien Anwenderwerkzeuge (Analyse) sowie Datenmanagement (Datenintegration) zuzuordnen. Dass Komponenten der Appliance sich in allen vorgestellten Kategorien wiederfinden zeigt somit, dass die Appliance eine mögliche Komplettlösung für die Big Data Problematik darstellt.

3.4.2.13 ParStream

Das Unternehmen ParStream wurde 2008 gegründet und ist in Köln ansässig (Vgl. ParStream 2013a). Das gleichnamige Produkt ParStream ist eine Massively-Parallel-Processing-Datenbank, die auf einer shared-nothing-Architektur basiert und sich durch ihre Schnelligkeit auszeichnet. Die Schnelligkeit wird vor allem durch den High Performance Compressed Index (HPCI) erreicht, da die Dekompression von Daten entfällt (Vgl. ParStream 2013d; Vgl. ParStream 2013b, S. 1).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von ParStream in der folgenden Tabelle dargestellt:

ParStream - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	ParStream läuft auf Standard-Infrastrukturen, d.h. einzelnen Servern, Server-Clustern, privaten und öffentlichen Clouds. (Vgl. ParStream 2013d)
Skalierbarkeit / Zukunftssicherheit	k.A.
Verarbeitung verschiedener Datenquellen (Schnittstellen)	JDBC, ODBC (Vgl. ParStream 2013b, S. 2)
Verarbeitung verschiedener Datenformate	integer, timestamp, string (Vgl. ParStream 2013c)
Umgang mit semi- und unstrukturierten Daten	strukturierte und semi-strukturierte Daten (Vgl. ParStream 2013c)

Benutzeroberfläche / Benutzerfreundlichkeit	grafisches Frontend (Vgl. ParStream 2013b, S. 2)
Software-dokumentation	k.A.
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	recht junges Start-Up Unternehmen, das u.a. QlikTech und EMC Panopticon seine Partner nennen kann
Qualität des Supports	k.A.
Produktkategorie	Speicherkonzept

Tabelle 3.16: Ergebnisse Untersuchungsmerkmale ParStream

Das Produkt ParStream ist eine Massively-Parallel-Processing-Datenbank, die Echtzeit-Analysen sowie eine Verarbeitung wahlweise auf GPU- oder CPU-Basis ermöglicht und sich insgesamt durch Schnelligkeit auszeichnet. ParStream wird deshalb der Kategorie Speicherkonzepte zugeordnet, dies wird in der nachfolgenden Abbildung verdeutlicht.

	Unternehmen	**ParStream**
	untersuchte Produkte bzw. Lösungen	Parstream
Kategorie	**Unterkategorie**	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	x
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.44: Produktkategorisierung ParStream
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 196)

3.4.2.14 Pentaho

Pentaho ist ein 2004 gegründetes, weltweit agierendes, US-amerikanisches Unternehmen welches sich auf den Bereich Business Intelligence fokussiert. Es vertreibt die zwei Produkte Business Analytics und Data Integration.

Pentaho Data Integration ist eine Datenintegrationsplattform, die dem Endbenutzer analysebereite Daten aus verschiedenen Quellen liefert. Über ein grafisches ETL-Tool lassen sich Daten aus Big Data-Quellen verwenden. Es werden dabei Hadoop-Distributionen von Cloudera, Hortonworks, MapR und Intel unterstützt. Für NoSQL-Datenbanken wie Cassandra, HBase und MongoDB sowie Verbindungen zu Datenspeichern wie Amazon Redshift und Splunk gibt es Plug-ins. Zusätzlich besteht die Möglichkeit analytische Datenbanken wie Vertica, Greenplum und Teradata zu verwenden. Klassische relationale Datenbanken wie Oracle, DB2, MySQL und SQL Server lassen sich ebenfalls integrieren. Durch extrahieren und Data Blending von vorhandenen und diversen Daten wie zum Beispiel Webinhalte, Dokumente, Protokolldateien und Inhalte aus sozialen Medien lassen sich konsistente Daten erstellen, die sofort analysiert werden können. Über die Administration lassen sich Sicherheitsberechtigungen für Benutzer und Rollen verwalten und Berechtigungen festlegen (Vgl. Pentaho 2013b, e).

Pentaho Business Analytics erlaubt verschiedenste Analysen mit interaktiven Visualisierungen wie Geo-Mapping, Heat-Maps und Verteilungs-/Blasendiagramme. Durch In-Memory-Caching lassen sich dabei auch große Datenmengen schnell verarbeiten. Durch das webbasierte Drag & Drop-Dashboard-Design lassen sich benutzerdefinierte Ansichten erstellen und die Geschäftsanalysen in andere Anwendungen integrieren. Im Bereich des Reporting gibt es die Möglichkeit mithilfe eines grafischen Editors, Berichte in den Formaten HTML, Excel, CSV, PDF und RTF zu erstellen. Durch lernfähige Algorithmen wie Klassifikation, Regression, Clustering und Assoziation lassen sich zudem prognostische Analysen erstellen. Als Grundlage für diese Verarbeitung ist eine komplette Datenintegration enthalten. Diese kann auf Daten aus Quellen wie Hadoop, NoSQL, Analyse- sowie relationalen Datenbanken zugreifen, die Daten vorbereiten, zusammenführen und bereitstellen. Dies ist ebenfalls für Big Data möglich. Zur Vermeidung von Kodierung und Komplexität ist ein visueller MapReduce-Designer für Hadoop integriert und die gesamte Datenintegration lässt sich über einfaches Drag & Drop einrichten (Vgl. Pentaho 2013c).

Mithilfe von Pentaho Mobile ist es zudem möglich Datenermittlung von unterwegs zu betreiben und interaktive Analysen und Visualisierungen zu erstellen. Zudem ist der Zugriff auf interaktive Dashboards, Analysen und Berichte möglich (Vgl. Pentaho 2013d).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Pentaho Data Integration und Business Analytics in der folgenden Tabelle dargestellt:

Pentaho Data Integration und Business Analytics - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Die Plattform kann direkt vor Ort oder in der Cloud eingesetzt werden. Zudem kann sie in andere Software-Anwendungen eingebettet werden. (Vgl. Pentaho 2013f)
Skalierbarkeit / Zukunftssicherheit	Skalierbare Plattform mit Integrationsmöglichkeit großer Datenmengen. (Vgl. Pentaho 2013f)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• Standardmäßige relationale Datenbanken: Oracle, DB2, MySQL, SQL Server • Hadoop: Apache Hadoop, Cloudera, HortonWorks, MapR • NoSQL-Datenbanken: MongoDB, Cassandra, HBase • Analytische Datenbanken: Vertica, Greenplum, Teradata • Spezialisierte Daten-Stores: Splunk, Amazon Redshift • Packages, Enterprise Applications, SAP • Cloudbasierte und SaaS-Anwendungen: Salesforce, Amazon Web Services (Vgl. Pentaho 2013e)
Verarbeitung verschiedener Datenformate	Dateien: XML, Excel, Flatfile- und Webservice-APIs (Vgl. Pentaho 2013e)
Umgang mit semi- und unstrukturierten Daten	• Datenvorbereitung, Modellierung und Untersuchung nicht strukturierter Datensätze • Konnektivität für eine Vielfalt an Datenquellen: gängige strukturierte, unstrukturierte und semistrukturierte Datenquellen (Vgl. Pentaho 2013e)

Benutzeroberfläche / Benutzerfreundlichkeit	• SQL-Code schreiben und MapReduce-Java-Funktionen erstellen nicht notwendig. • Grafisches ETL-Tool zum Laden und Verarbeiten von Big Data-Quellen in einer vertrauten Umgebung. Der visuelle Editor von Pentaho Data Integration umfasst: • Intuitiver Drag & Drop-Designer • Umfassende Bibliothek mit vorgefertigten Komponenten. • Dynamische Umwandlungen, um Feldzuordnungen, Validierungs- und Anreicherungsregeln mithilfe von Variablen zu bestimmen. • Integrierter Debugger für Tests und Feineinstellung der Auftragsausführung (Vgl. Pentaho 2013c)
Software-dokumentation	Umfangreiche Dokumentationen für Entwickler und Anwender. (Vgl. Pentaho 2013a)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Pentaho ist „führender unabhängiger Anbieter von Geschäftsanalysen für Tausende von globalen Unternehmen und Software-Herstellern“ (Vgl. Pentaho 2013g)
Qualität des Supports	zwei Support-Pakete, Support Portal[13], Wissensdatenbank, Wiki[14], Forum
Produktkategorie	Datenmanagement und Anwenderwerkzeug

Tabelle 3.17: Ergebnisse Untersuchungsmerkmale Pentaho Data Integration & Business Analytics

Pentaho bietet eine Lösung an, in der die Datenintegration mit der Geschäftsanalyse verbunden wird. Dadurch lässt sich der komplette ETL-Prozess sowie die Analyse und das Reporting der Daten in einer Umgebung durchführen. Aufgrund dessen, dass es sich um eine Open-Source Datenintegrationslösung handelt, kann Data Integration in die Kategorie Datenmanagement eingeordnet werden. Da Business Analytics eine Open-Source-Suite mit Produkten für Reporting, Dashboarding und Data Mining ist, kann es den Anwenderwerkzeugen zugeordnet werden.

[13] https://support.pentaho.com/home
[14] http://wiki.pentaho.com/display/COM/Community+Wiki+Home

	Unternehmen	Pentaho	
	untersuchte Produkte bzw. Lösungen	Data Integration	Business Analytics
Kategorie	Unterkategorie		
Speicherkonzepte	Relationale Datenbanken		
	Analytische Datenbanken		
	Multidimensionale Datenbanken		
	Hadoop-Distribution		
	Sonstige		
Anwenderwerkzeuge	Dashboarding		**
	Berichtswesen		**
	Analyse		**
	Data Mining		**
Datenmanagement	Datenintegration	x	
	Datenqualität		

Abbildung 3.45: Produktkategorisierung Pentaho
Quelle: Eigene Darstellung in Anlehnung an BARC (Vgl. 2013, S. 178)

3.4.2.15 QlikTech

QlikTech ist ein 1993 in Schweden gegründetes Unternehmen, welches mittlerweile seinen Hauptsitz in den USA hat. Seit 2010 ist das Unternehmen börsennotiert und wird an der NASDAQ gehandelt. QlikTech vertreibt mit QlikView eine Business-Discovery-Plattform, die anwendergesteuerte Business Intelligence bietet (Vgl. QlikTech 2013b, h). Mit dieser können Daten aus verschiedenen Quellen wie Big-Data-Systemen, Data Warehouses, Datenbanken und Spreadsheets in einer interaktiven Analyseebene integriert und anschließend analysiert werden. Über verschiedene APIs lässt sich ein unternehmenseigenes, offenes Datenprotokoll als Schnittstelle zu Hadoop und weiteren Anbietern von Big Data-Systemen nutzen. Auf weitere Big Data-Quellen wie Attivio, und DataRoket kann über die Produkte von QlikView-Partnern zugegriffen werden. Dies wird durch das QVX[15]-SDK ermöglicht worüber Drittanbieter passende Konnektoren erstellen können, welche der Extraktion der Daten dienen (Vgl. QlikTech 2013f). So bietet Datawatch mit dem „Unstructured Data Connector for QlikView“ eine Möglichkeit, Quellen wie PDF, XBRL, HTML, Textdateien, Maschinendaten und andere nichtrelationale Datenquellen zu verarbeiten (Vgl. Datawatch 2013b). Solche Anwendungen lassen sich über den QlikMarket[16] herunterladen.

Folgende Bestandteile beinhaltet die Business-Discovery-Plattform (Vgl. QlikTech 2013c):

- Server-Bestandteil, z.B. QlikView Server und QlikView Expressor Server

[15] QlikView data eXchange: Format für den Import aus nicht standardmäßigen Datenquellen (wie Google Big Query) nach QlikView (QlikTech 2013c)
[16] `http://market.qlikview.com/`

- Anwendungsentwicklung, z.B. QlikView Desktop
- Auswertung, z.B. QlikViewBrowser
- Administration, z.B. QlikView Expressor Desktop, QlikView Governance Dashboard
- Konnektoren, z.B. Connector für den Einsatz mit SAP Netweaver und Salesforce

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von QlikView in der folgenden Tabelle dargestellt:

QlikTech QlikView - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Bietet die Möglichkeit, eine direkte Verbindung zu gängigen Geschäftsanwendungen und Datenquellen zu erstellen, wie Salesforce.com, SAP Netweaver und Informatica. (Vgl. QlikTech 2013e)
Skalierbarkeit / Zukunftssicherheit	k.A.
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• Offenes Datenprotokoll als Schnittstelle zu den APIs von Hadoop- und anderen Anbietern von Big-Data-Systemen. • Zugriff auf Big-Data-Quellen wie Attivio und DataRoket erfolgt über die Produkte von QlikView-Partnern. • Spezielles QVX-SDK erhältlich, mit dem Entwickler von Drittanbietern passende Konnektoren für jedes System mit einer offenen API erstellen können. • Direkte Verbindung zu gängigen Geschäftsanwendungen und Datenquellen erstellen – etwa Salesforce.com, SAP NetWeaver und Informatica. • Konsolidierung relevanter Daten aus unterschiedlichen Quellen inklusive Big-Data-Quellen. (Vgl. QlikTech 2013f)
Verarbeitung verschiedener Datenformate	Datawatch Unstructured Data Connector für QlikView: Quellen wie PDF, XBRL, HTML, Textdateien, Maschinendaten und andere nichtrelationale Datenquellen aufbereiten und in Datenermittlungs-Dashboards und -anwendungen aufnehmen (Über den QlikMarket). (Vgl. Datawatch 2013b)

Umgang mit semi- und unstrukturierten Daten	strukturiert, semistrukturiert, unstrukturiert (Vgl. Datawatch 2013b)
Benutzeroberfläche / Benutzerfreundlichkeit	• Zugriff auf und Analyse von Big Data mit mobilen Geräten • Visualisierung von Big Data mit ansprechenden State-of-the-Art-Darstellungen (Vgl. QlikTech 2013d)
Software-dokumentation	k.A.
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	29.000 Kunden in 100 Ländern (Vgl. QlikTech 2013a)
Qualität des Supports	• QlikCommunity • Qoncierge Service bei nicht-technischen Post-Sales-Problemen • Technischer Produkt Support für Lösung technischer Probleme rund um QlikView • Kundenportal • Qonnect Portal for QlikView Partners (Vgl. QlikTech 2013g)
Produktkategorie	Anwenderwerkzeug

Tabelle 3.18: Ergebnisse Untersuchungsmerkmale QlikTech QlikView

QlikView ist eine Business-Discovery-Plattform die neben umfangreichen Analyse- und Reportingfunktionen auch den kompletten ETL-Prozess abbilden kann. Dabei ermöglicht QlikView mengenorientierte Analysen und die Erstellung individueller Dashboard-Applikationen. Das Produkt gehört deshalb zur Kategorie der Anwenderwerkzeuge.

	Unternehmen	QlikTech
	untersuchte Produkte bzw. Lösungen	QlikView
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	**
	Berichtswesen	*
	Analyse	**
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.46: Produktkategorisierung QlikTech
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 178)

3.4.2.16 SAP

Die SAP AG wurde 1972 in Walldorf gegründet und ist einer der führenden Anbieter für Unternehmenssoftware (Vgl. SAP AG 2013d). Das Unternehmen hat ein breites Spektrum an Produkten, die nachfolgend einzelnen Bereichen zuzuordnen sind (Vgl. SAP AG 2013c):

- **Geschäftsprozess-Lösungen**, z.B. Customer Relationship Management, Finanzmanagement, Supply Chain Management
- **Geschäftsanalyse-Lösungen**, z.B. Business Intelligence, Risikomanagement und Compliance
- **Datenbank- und Technologie-Lösungen**, z.B. Data Warehousing, Enterprise Information Management, Zusammenarbeit und Content Management

Im Kontext der Big-Data-Thematik hat die SAP AG im Bereich In-Memory-Computing mit HANA (High Performance Analytic Appliance) aus der Kategorie Datenbank- und Technologie eine Möglichkeit geschaffen, große Datenmengen in kürzester bzw. Echtzeit zu analysieren. HANA ist eine In-Memory-Lösung, die in Kombination mit der Hardware von SAP-Unternehmenspartnern Echtzeitanalysen ermöglicht und so Geschäftsprozesse beschleunigen soll. Unternehmen können HANA außerdem auch als Cloud erwerben (Vgl. SAP AG 2013c; Vgl. Alterauge 2013, S. 1; Vgl. Dill u. Füg 2013; Vgl. Franke 2013).

HANA analysiert jedoch vorwiegend die aus der SAP-Umgebung stammenden Daten abgeschlossener Transaktionen und ersetzt dabei überwiegend vorherige SAP-BW-Prozesse. Die Appliance kann jedoch auch datenintensive Verarbeitungen in SAP-Anwendungen beschleunigen (z.B.

HANA FI-CO) und besteht aus speziell aufeinander abgestimmten Hard- und Softwarekomponenten. Die Software arbeitet in diesem Kontext mit spaltenorientierten Lesezugriffen sowie zeilenorientierten Schreibzugriffen und kann damit als Hybrid aus In-Memory- und relationaler Datenbank verstanden werden (Vgl. Franke 2013, S. 1).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von SAP HANA in der folgenden Tabelle dargestellt:

SAP HANA - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• Die Nutzer von HANA sind an eine bestimmte Hardwarearchitektur gebunden, die sich allerdings auch erweitern lässt. • HANA setzt sich insgesamt aus einer In-Memory Software und passender Hardware zusammen: HP, IBM, Fujitsu, Cisco und Dell. (Vgl. Rachel 2011, S. 17f.)
Skalierbarkeit / Zukunftssicherheit	• Verarbeitung von 1 Petabyte Rohdaten in Echtzeit • Komprimierung von Daten, die den benötigten Speicherplatz signifikant reduziert (z.B. durch dictionary coding). • SAP hat 16 Knoten für Partner und Kunden validiert und zertifiziert. • Hohe Verfügbarkeit und Ausfallsicherheit • durchgehende Performanceverbesserung (Vgl. Osten 2012; Vgl. Wolf 2012)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• HANA verfügt über eigene Modellierungswerkzeuge sowie eine eigene Scriptsprache. • HANA hat diverse Schnittstellen wie z.B. JDBC, ODBC, OLAP-Schnittstelle ODBO/MDX sowie das SAP-proprietäre BICS Interface. • Möglichkeit der Einbindung von Excel-Dateien. (Vgl. Alterauge 2013, S. 1; Vgl. Dill u. Füg 2013)

Verarbeitung verschiedener Datenformate	• Datenzugriffsdienste wie ODATA, ATOM, JSON, XML oder XMLA • Microsoft Excel (Vgl. plus-IT GmbH 2013; Vgl. Osten 2012)
Umgang mit semi- und unstrukturierten Daten	Verarbeitung von strukturierten und unstrukturierten Daten (Vgl. plus-IT GmbH 2013)
Benutzeroberfläche / Benutzerfreundlichkeit	Grafische Design-Umgebung: • SAP Business One Replizierung • SAP HANA Studio • Semantic Layer (Finanz-, Verkaufs-, Lager-Module) • HANA Dashboards • Reports • Interaktive Analyse-Tools (Vgl. Versino AG 2013)
Software-dokumentation	• Vor der Installation von HANA finden sich alle notwendigen Informationen zu Downloads und Einrichtung von SAP HANA und Clients online[17]. • SAP HANA Academy als kostenfreie Online-Ressource: Lehrvideos, Live- und aufgezeichnete Webseminare, Trainingseinheiten zu SAP HANA (Vgl. Dill u. Füg 2013; Vgl. SAP AG 2013b)

[17] unter `http://scn.sap.com/docs/DOC-28294`

Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• SAP als internationaler Konzern hat viele Großkunden, wie etwa Procter & Gamble, AOK u.v.m • Die Unternehmen, die HANA nutzen können dabei keiner spezifischen Branche zugeordnet werden. • Nähere Informationen zu den HANA-Erfahrungen diverser Unternehmen sind online zu finden[18]. (Vgl. SAP AG 2013d)
Qualität des Supports	• SAP hat eine Support-Plattform speziell für HANA[19] • Hier können viele Dokumente zur Software heruntergeladen sowie Fragen beantwortet werden.
Produktkategorie	Speicherkonzept

Tabelle 3.19: Ergebnisse Untersuchungsmerkmale SAP HANA

Ein weiteres Produkt aus dem Portfolio der SAP AG welches für Big Data verarbeitende Architekturen in Frage kommt ist das ETL-Tool Data Services. Dieses dient zur Integration von Daten und kann somit den gesamten Prozess von der Extraktion über die Transformation bis hin zum Laden der Daten in die Datenbank abdecken. Vormals wurde das Produkt als BusinessObjects Data Services vertrieben und ist häufig auch noch unter dieser Bezeichnung zu finden.

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Data Services in der folgenden Tabelle dargestellt:

SAP Data Services - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Die SAP Data Services lassen sich durch eine Vielzahl an Schnittstellen in bestehende BI-Architekturen einbinden.
Skalierbarkeit / Zukunftssicherheit	k.A.

[18] http://www.sap.com/germany/solutions/technology/in-memory-computing-platform/reviews/index.epx

[19] http://help.sap.com/hana_platform

Verarbeitung verschiedener Datenquellen (Schnittstellen)	Unterstützt unterschiedliche Datenbanken, u.a. Oracle, DB2, Sybase, SQL, Informix, Teradata, ODBC, SAP HANA. (Vgl. SAP AG 2012)
Verarbeitung verschiedener Datenformate	XML, Excel, JMS, SOAP (Web Services) (Vgl. SAP AG 2012)
Umgang mit semi- und unstrukturierten Daten	Verarbeitung von strukturierten und unstrukturierten Daten (Vgl. SAP AG 2012)
Benutzeroberfläche / Benutzerfreundlichkeit	Grafische Design-Umgebung: Der SAP Data Services Designer bietet eine grafische Oberfläche zur Modellierung und Verwaltung der Datenflüsse.
Software-dokumentation	Online Hilfe, Handbuch, Tutorials (Vgl. SAP AG 2013a)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• SAP als internationaler Konzern hat viele Großkunden, wie etwa Procter & Gamble, AOK u.v.m (Vgl. SAP AG 2013d)
Qualität des Supports	k.A.
Produktkategorie	Datenmanagement

Tabelle 3.20: Ergebnisse Untersuchungsmerkmale SAP Data Services

Das Produkt SAP HANA ist eine In-Memory-Lösung für analytische Auswertungen, die neben der Zeilenorientierung auch eine Spaltenorientierung unterstützt. Sie kann den Speicherkonzepten zugeordnet werden. Das zweite vorgestellte Produkt SAP Data Services beinhaltet eine integrierte Plattform mit Funktionen zur Datenintegration (Data Integrator) und Datenqualität (Data Quality Management). Es kann somit der Kategorie Datenmanagement zugeordnet werden.

In der nachfolgenden Abbildung ist neben HANA und Data Services auch die Business Objects Business Intelligence Suite aufgeführt, da sie von SAP nicht explizit als Big Data-Produkt genannt wird, Business Objects aber bei der Analyse, auch von großen Datenmengen von Bedeutung ist. Eine umfassende Analyse dieses Tools wurde deshalb nicht vorgenommen, in der Übersicht wird damit jedoch auch deutlich, dass SAP eine breite Produktpalette in Bezug auf Big Data anbietet.

	Unternehmen	SAP		
	untersuchte Produkte bzw. Lösungen	Data Services	Business Objects Business Intelligence Suite	HANA
Kategorie	Unterkategorie			
Speicherkonzepte	Relationale Datenbanken			
	Analytische Datenbanken			x
	Multidimensionale Datenbanken			
	Hadoop-Distribution			
	Sonstige			
Anwenderwerkzeuge	Dashboarding		**	
	Berichtswesen		**	
	Analyse		**	
	Data Mining		*	
Datenmanagement	Datenintegration	x		
	Datenqualität	x		

Abbildung 3.47: Produktkategorisierung SAP
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 180; S. 188; S. 196), IS Report (2013, S. 102))

3.4.2.17 SAS

SAS ist ein 1976 in North Carolina (USA) gegründeter Anbieter von Business-Intelligence-Lösungen und -Services. Das Unternehmen bietet eine Vielzahl an branchen- und fachspezifischen Softwarelösungen an, die auf der BI-Plattform SAS9 basieren, welche alle Prozesse der Generierung, Verteilung und Anwendung von Wissen integriert und automatisiert. Somit werden von Datenmanagement und -analyse über das Reporting bis zur grafischen Aufbereitung der Informationen alle Komponenten der Business-Intelligence-Prozesskette von SAS bereitgestellt (Vgl. SAS 2013i). Im Bereich Big Data bietet SAS folgende Technologien an (Vgl. SAS 2013j):

- Datenmanagement
- High-Performance Analytics
- High-Performance Datenvisualisierung (Visual Analytics)
- Flexible Bereitstellungsoptionen für Big Data (Solutions OnDemand[20])

Visual Analytics ist eine BI-Plattform mit Schwerpunkten auf Visualisierung, Analytics, Reporting und Mobile-BI, die auf der In-Memory-Technologie aufbaut. Es kann auf einzelnen Windows-Servern sowie auf einer Blade-Server-Infrastruktur installiert werden, wodurch die Ausbaufähigkeit gegeben ist. Zudem ist eine Cloud-Bereitstellung für öffentliche Cloud-Anbieter und die SAS Cloud möglich (Vgl. SAS 2013d). Die Plattform umfasst (Vgl. SAS 2012):

- SAS LASR Analytics Server (In-Memory Server)
- The Hub (Oberfläche zum starten der einzelnen Elemente)
- Mobile (Modul, um unterwegs Reports ansehen zu können)

[20] z. B. Software-as-a-Service und Enterprise Hosting (Vgl. SAS 2013c)

- Explorer (Instrument zur Visualisierung, Untersuchung und Analyse von Daten)
- Designer (Tool zur Erstellung von Reports und Dashboards)
- Environment Administration (Für die Verwaltung von Rechten, Sicherheit und Daten)

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Visual Analytics in der folgenden Tabelle dargestellt:

SAS Visual Analytics - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• Lässt sich in bestehende Dateninfrastrukturen integrieren. • Ist kompatibel zu Standardhardware z. B. von HP oder Dell. (Vgl. Accantec 2013; SAS 2013a)
Skalierbarkeit / Zukunftssicherheit	• SAS LASR Analytic Server ist eine Komponente von Visual Analytics und nutzt Hadoop als Speicher. • Durch den möglichen Einsatz einer Blade-Server-Infrastruktur ist eine praktisch unbegrenzte Skalierbarkeit gegeben. (Vgl. SAS 2012, 2013d)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	Über die enthaltene SAS/ACCESS-Schnittstelle: • Data Warehouse Appliances (Aster nCluster, Greenplum, Netezza, Oracle Exadata, ODBC, Sybase IQ, Teradata) • Nichtrelationale Datenquellen (ADABAS, DATACOM/DB, IDMS/R, IMS-DL/I, System 2000) • Verteilte Dateisysteme (Hadoop) • Relationale Datenbanken (DB2, Informix, Microsoft SQL, MySQL, ODBC, OLE DB, Oracle, SAP HANA, Sybase, Vertica) (Vgl. SAS 2013g)
Verarbeitung verschiedener Datenformate	z. B. Microsoft Excel, Microsoft Access, dBase, Lotus, SPSS, JMP und Stata (Vgl. SAS 2013g)
Umgang mit semi- und unstrukturierten Daten	Verarbeitung von strukturierten als auch von unstrukturierten Daten (Vgl. SAS 2012)

Benutzeroberfläche / Benutzerfreundlichkeit	• Prognosen on-the-fly und Szenarioanalysen • Automatische Diagrammerstellung • Hilfe-Fenster • Bedienung per Drag & Drop (Vgl. SAS 2013f)
Software-dokumentation	Für Visual Analytics 6.2: • Getting Started with Exploration and Reporting • User's Guide, Verschiedene Videos • Deployment and administration documentation (Vgl. SAS 2013e)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	• Eins der größten Software-Unternehmen in privater Hand, zählt zu den größten Anbietern von Business-Intelligence-Lösungen in Deutschland. • Bietet Softwarelösungen, die auf die Anforderungen einzelner Branchen zugeschnitten sind. (Vgl. SAS 2013i)
Qualität des Supports	• EMITS-System (Electronic Mail Interface Technical Support) (24/7) • Hotline Support (Montags bis donnerstags: 9.00 Uhr bis 17.00 Uhr, freitags: 9:00 Uhr bis 15.00 Uhr) • Knowledge Base (Vgl. SAS 2013h, b)
Produktkategorie	Anwenderwerkzeug

Tabelle 3.21: Ergebnisse Untersuchungsmerkmale SAS Visual Analytics

Visual Analytics von SAS bietet verschiedene Möglichkeiten BI mit der In-Memory-Technologie zu verbinden. Die Plattform ist dabei stark skalierbar und bietet viele Schnittstellen zu vorhandener Hardware im Unternehmen. Aufgrund seiner Funktionalität ist es in die Kategorie der Anwenderwerkzeuge einzuordnen, erfüllt dort die Unterkategorien Dashboarding und Berichtswesen und legt seinen Schwerpunkt im Bereich Analyse.

	Unternehmen	**SAS**
	untersuchte Produkte bzw. Lösungen	Visual Analytics
Kategorie	**Unterkategorie**	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	*
	Berichtswesen	*
	Analyse	**
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.48: Produktkategorisierung SAS
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 180)

3.4.2.18 Software AG

Die Software AG wurde 1969 in Darmstadt gegründet und ist führend in den Bereichen Geschäftsprozesse, Integration und Big Data (Vgl. Software AG 2013e). Das Unternehmen bietet folgende Produkte an (Vgl. Software AG 2013c):

- ARIS
- webMethods
- Terracotta
- Adabas-Natural
- Software AG Live

Im Produktbereich von Terracotta ist das von der Software AG entwickelte Big Data-Tool BigMemory verankert. BigMemory ist eine In-Memory-Datenplattform, die es ermöglichen soll, auf mehrere hundert Terabyte zugreifen zu können und diese in Echtzeit zu analysieren. Folgende Funktionen und Leistungsmerkmale lassen sich nach IT-Daily definieren (Vgl. IT-Daily 2013a):

- Überwachung und Verwaltung von In-Memory-Daten, z.B. durch ein anpassbares Web-Dashboard,
- hohe Verfügbarkeit und schneller Neustart im Rahmen eines Disaster Recovery,
- Unterstützung von Java 7,
- Kompatibilität mit Hadoop und ein
- schneller Zugriff auf Daten in der Größenordnung von Terabyte und größer.

Das Java-basierte BigMemory baut auf den Funktionen von Open-Source-Ehcache auf und wird in Enterprise Ehcache eingebunden, um große Datenmengen im Hauptspeicher zu speichern. Der Kunde muss bei diesem Produkt keine zusätzliche Hardware beschaffen, um BigMemory im vollsten Umfang nutzen zu können (Vgl. Software AG 2013b). BigMemory ist demnach eine Softwarelösung, die Anwendungen auf Basis der verteilten Server-Knoten innerhalb der Caching-Architektur unter Einsatz von Standardhardware hochskalieren lässt. Dabei lässt sich die Architektur durch das Hinzufügen von Servern erweitern (Vgl. Franke 2013, S. 1).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von BigMemory in der folgenden Tabelle dargestellt:

Terracotta BigMemory - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• BigMemory ist eine Softwarelösung und lässt Anwendungen auf Basis der verteilten Server-Knoten unter Einsatz von Standardhardware hochskalieren. • Die verteilte Caching-Architektur lässt sich durch das Hinzufügen von zusätzlichen Servern erweitern. • BigMemory ist vollständig Java-basiert und wird in Enterprise Ehcache eingebunden. • BigMemory läuft auf gängiger Hardware und allen gängigen JVMs, einschließlich HotSpot, IBM JDK und JRockit. (Vgl. Franke 2013, S. 1; Vgl. Software AG 2013b)
Skalierbarkeit / Zukunftssicherheit	k.A.

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• „Die Enterprise-Lösung verfügt über Schnittstellen für Java, .Net oder Python und untersützt darüber hinaus eine Reihe von Messaging-Protokollen wie WebSocket oder MQTT. Als Web-Client unterstützt die Messaging-Plattform Silverlight, JavaScript und HTML5“ (Schindler 2013) • Hadoop Schnittstelle (Vgl. Software AG 2013a)
Verarbeitung verschiedener Datenformate	k.A.
Umgang mit semi- und unstrukturierten Daten	Verarbeitung von strukturierten und semi-strukturierten Daten. (Vgl. Pressebox 2012)
Benutzeroberfläche / Benutzerfreundlichkeit	• hohe Verfügbarkeit und schneller Neustart im Rahmen eines Disaster Recovery • BigMemory bietet ein individuell anpassbares Web-Dashboard und genaue Kontroll- und Messdaten über die verteilten In-Memory-Kerne. (Vgl. IT-Daily 2013a)
Software-dokumentation	Im Resource Center der Software AG sind diverse Materialien zu den einzelnen Produkten erhältlich[21].
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Software AG hat einen breiten Kundenkreis und kann Großunternehmen wie z.B. CosmosDirekt, Springer-Verlag u.v.m. zum Kundenstamm zählen. (Vgl. Software AG 2013d)

[21] http://www.softwareag.com/corporate/rc/

Qualität des Supports	• globaler 24/7- Support • Support kann an Geschäfts- und Supportanforderungen angepasst werden • mehrere Vertragstypen möglich (Vgl. Software AG 2013e)
Produktkategorie	Speicherkonzept

Tabelle 3.22: Ergebnisse Untersuchungsmerkmale Terracotta Big-Memory

Die Big Data Lösung von Software AG Terracotta BigMemory ist der Kategorie Speicherkonzept zuzuordnen. Es handelt sich hierbei um eine analytische Datenbank, die strukturierte und semistrukturierte Daten verarbeiten kann.

	Unternehmen	Software AG
	untersuchte Produkte bzw. Lösungen	Terracotta Big Memory
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	x
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.49: Produktkategorisierung Software AG

3.4.2.19 Splunk

Splunk ist ein US-amerikanisches Unternehmen welches mit Splunk Enterprise eine Engine anbietet, die Computerdaten von Anwendungen, Servern und Geräten in Echtzeit erfasst, indiziert und anschließend verarbeitet (Vgl. Splunk 2013c). Mit Splunk Free gibt es zusätzlich eine kostenlose Version mit eingeschränktem Funktionsumfang. Splunk Cloud hat denselben Funktionsumfang wie Splunk Enterprise, wird jedoch als Cloud-Service angeboten. Mit Hunk Beta wird zudem eine spezielle Lösung für Hadoop geboten (Vgl. Splunk 2013f).

Splunk Enterprise verarbeitet Computerdaten bzw. IT-Daten in Echtzeit und bietet die Möglichkeit, diese zu beobachten und zu analysieren. Dabei wird MapReduce als Basistechnologie eingesetzt (Vgl. Splunk 2013d).

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Splunk Enterprise in der folgenden Tabelle dargestellt:

Splunk Enterprise - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	Integration möglich durch Splunk Hadoop Connect und Splunk DB Connect (Schnittstellen zu Hadoop und zu relationalen Datenbanken) (Vgl. Splunk 2014a, b)
Skalierbarkeit / Zukunftssicherheit	• Skalierbarkeitsarchitektur basiert auf MapReduce • Bei steigendem Volumen und wachsender Anzahl der Datenquellen können weitere Commodity-Server hinzugefügt werden. • unbegrenztes Indizierungsvolumen (Vgl. Splunk 2013c)
Verarbeitung verschiedener Datenquellen (Schnittstellen)	• Websites, Business Applications, Social Media Plattformen, App Server, Hypervisor, Sensoren, traditionelle Datenbanken und Open Source Data Stores (Vgl. Splunk 2013e) • Splunk Hadoop Connect: Hadoop-Schnittstelle (Vgl. Splunk 2014b) • Splunk DB Connect: Schnittstelle zu relationalen Datenbanken (Vgl. Splunk 2014a)
Verarbeitung verschiedener Datenformate	Computerdaten wie Anwendungslogs, Webzugriffslogs, Webproxylogs, Call Detail Records, ClickStream-Daten, Meldungswarteschlangen, Paketdaten, Konfigurationsdateien, Datenbank-Auditlogs und –tabellen, Dateisystem-Auditlogs, Verwaltungs- und Protokollierungs-APIs, Metriken, Status- und Diagnosebefehle des Betriebssystems, Syslog, WMI (Windows Management Instrumentation) und weitere Datenquellen (Vgl. Splunk 2013i)
Umgang mit semi- und unstrukturierten Daten	Computerdaten, relationale und andere unstrukturierte Daten (Vgl. Splunk 2013e)

Benutzeroberfläche / Benutzerfreundlichkeit	Weboberfläche von Splunk ist intuitiv und unterstützt rasche Ad-hoc-Drilldowns auf Suchergebnisse. (Vgl. Splunk 2013h)
Software-dokumentation	Online-Dokumentation für die Splunk-Plattform und für alle anderen Splunk-Produkte. Es gibt Produkthandbücher als PDF, Begriffe können im Splexicon nachgeschlagen werden, und Feedback kann gegeben werden. Eine zusätzliche Quelle für neue Nutzer ist das Splunk Buch. (Vgl. Splunk 2013a)
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Fast die Hälfte der Fortune 100-Unternehmen und mehr als 6,000 Unternehmen, Dienstanbieter und Behörden in 90 Ländern nutzen Splunk. (Vgl. Splunk 2013b)
Qualität des Supports	• Community Support - Dokumentation, Community Wiki, Splunk Answers and IRC sind für jeden verfügbar um technische Fragen zu beantworten. • Enterprise Support - Direkter Kontakt zum Customer Support Team per Telefon. Möglichkeit die Anfragen online zu verwalten. • Global Support - 24x7x365 Support bei kritischen Problemen (Vgl. Splunk 2013g)
Produktkategorie	Anwenderwerkzeug

Tabelle 3.23: Ergebnisse Untersuchungsmerkmale Splunk Enterprise

Splunk bietet mit Splunk Enterprise eine Engine, die in Echtzeit gestreamte Computerdaten analysieren kann. Sie ermöglicht das Management, Monitoring und die Analyse von Maschinendaten und historischen IT-Daten und kann deshalb in die Kategorie der Anwenderwerkzeuge eingeordnet werden. Wie in der nachfolgenden Abbildung zu sehen ist, legt Splunk dabei den Schwerpunkt auf die Unterkategorien Dashboarding und Analyse, lässt sich jedoch auch dem Berichtswesen und Data Mining zuordnen.

	Unternehmen	Splunk
	untersuchte Produkte bzw. Lösungen	Splunk Enterprise
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	**
	Berichtswesen	*
	Analyse	**
	Data Mining	*
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.50: Produktkategorisierung Splunk
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 180)

3.4.2.20 Teradata

Teradata wurde 1979 in Dayton, Ohio (USA) gegründet und bietet analytische Datenlösungen mit den Schwerpunkten integriertes Data Warehousing, Big Data-Analysen und Geschäftsanwendungen an. Dabei besteht die Produktpalette von Teradata aus Produkten innerhalb folgender Kategorien (Vgl. Teradata 2013a; Vgl. Teradata 2013d):

- Arbeitslastspezifische Plattformen
- Data-Warehouse Software
- Discovery Plattform
- Tools und Hilfsmittel
- Hadoop
- Integrierte Marketing-Anwendungen
- Analytik-Anwendungen
- Diverse Dienstleistungen

Die Aster Big Analytics Appliance von Teradata ist den arbeitsplatzspezifischen Plattformen zuzuordnen und führt die Aster-Datenbank, SQL-MapReduce und Apache Hadoop zusammen. Aster Big Analytics Appliance ist zu 100% Open Source Apache Hadoop und läuft über Hortonworks HDP 1.1. Die Appliance bietet ein integriertes Server-Management & Monitoring über Teradata Viewpoint und Teradata Vital Infrastructure an. Es ist eine eng gekoppelte Hard- und Software-Lösung, die Aster Database und Apache Hadoop enthält, um strukturierte,

unstrukturierte und semi-strukturierte Daten verarbeiten zu können. Laut Hersteller lassen sich folgende wesentliche Merkmale festhalten (Vgl. Teradata.Aster 2013; Vgl. Teradata 2013c):

- Hybride Architektur, die Aster Database, Aster SQL-MapReduce und Apache Hadoop enthält für eine vollständige Analytics-Lösung für Big Data.
- SQL-MapReduce-Funktionen für die Datenerfassung, Datenaufbereitung, Analyse und Visualisierung.
- Verfügt über mehr als 80 vorgefertigte SQL–MapReduce-Funktionen und bietet out-of-the-box Analytik u.a. für Grafiken, Text, Marketing-Analysen.
- Aster Big Analytics Appliance ist eine leistungsfähige, ready-to-Run-Plattform, die vorkonfiguriert und optimal für große Datenspeicherung und Analysen ist.

Mit Bezug auf die festgelegten Untersuchungsmerkmale werden weitere Eigenschaften von Aster Big Analytics Appliance in der folgenden Tabelle dargestellt:

Teradata Aster Big Analytics Appliance - Ergebnisse zu den Untersuchungsmerkmalen:	
Möglichkeit der Integration in bestehende BI-Architekturen	• Die hybride Architektur vereint Aster Database, Aster SQL-MapReduce und Apache Hadoop. • Die Teradata Aster Big Analytics Appliance wird als vollständig integriertes System aus Hard- und Software bereitgestellt und ist eine „Plug and Play“-Lösung, d.h. sie kann innerhalb weniger Stunden den Betrieb aufnehmen. • Teradata Aster Big Analytics Appliance ermöglicht mit Aster-Database- und Hadoop-Rechnerknoten eine unternehmensspezifische Anpassung innerhalb vorherrschender Systeme und lässt sich individuell an den analytischen Workload-Bedarf eines Unternehmens anpassen. (Vgl. Ertel u. Pfleghar 2012, S. 1f.)
Skalierbarkeit / Zukunftssicherheit	Teradata Aster Big Analytics Appliance wurde für extreme Anforderungen an Datenanalysen entwickelt, die besonders hohe Rechnerleistung, viel Arbeitsspeicher und Möglichkeiten für Data Movement verlangen. (Vgl. Ertel u. Pfleghar 2012, S. 2)

Verarbeitung verschiedener Datenquellen (Schnittstellen)	• MapReduce • Hadoop • SQL • SQL-MapReduce-Technologie von Aster für die Parallelisierung von Datenverarbeitung und Anwendungen (Vgl. Ertel u. Pfleghar 2012, S. 5)
Verarbeitung verschiedener Datenformate	k.A.
Umgang mit semi- und unstrukturierten Daten	Verarbeitung von strukturierten und unstrukturierten Daten. (Vgl. Ertel u. Pfleghar 2012, S. 2)
Benutzeroberfläche / Benutzerfreundlichkeit	• weist eine integrierte webbasierte Suite von Anwendungen für System-Management und –Monitoring auf • Aster Database- und Apache Hadoop-Administratoren greifen somit direkt auf zentrale Systeminformationen zu. • diverse Monitoring- und Diagnose-Tools (Vgl. Ertel u. Pfleghar 2012, S. 5)
Software-dokumentation	k.A.
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Teradata hat einen breiten Kundenkreis und kann Großunternehmen wie z.B. DHL, diverse Finanzdienstleister, Cisco, Coca Cola u.v.m. zum Kundenstamm zählen. (Vgl. Teradata 2013b, S. 2)
Qualität des Supports	regionaler und globaler Teradata Customer Care Center mit 24x7-Service (Vgl. Conway 2012, S. 2)
Produktkategorie	Speicherkonzept

Tabelle 3.24: Ergebnisse Untersuchungsmerkmale Teradata Aster Big Analytics Appliance

Als Komplettlösung beinhaltet die Aster Big Analytics Appliance von Teradata eine hybride Architektur, die eine Integration in bestehende BI-Architekturen möglich machen soll. Im Kontext der Kategorisierung ist sie den Speicherkonzepten, speziell den analytischen Datenbanken zuzuordnen.

	Unternehmen	Teradata
	untersuchte Produkte bzw. Lösungen	Aster Database
Kategorie	Unterkategorie	
Speicherkonzepte	Relationale Datenbanken	
	Analytische Datenbanken	x
	Multidimensionale Datenbanken	
	Hadoop-Distribution	
	Sonstige	
Anwenderwerkzeuge	Dashboarding	
	Berichtswesen	
	Analyse	
	Data Mining	
Datenmanagement	Datenintegration	
	Datenqualität	

Abbildung 3.51: Produktkategorisierung Teradata
Quelle: Eigene Darstellung in Anlehnung an BARC (2013, S. 196)

3.4.3 Vorstellung der Ergebnisse

Die vorangegangene Vorstellung der einzelnen Unternehmen und der angebotenen Big Data-Produkte dient dazu, einen Marktüberblick zu geben, der zwar nicht alle auf dem Markt vorhandenen Produkte aufzeigen kann, jedoch eine Auswahl darstellt. Dabei war die Vorgehensweise stets gleich, indem zunächst Informationen zu dem Unternehmen und den angebotenen Produkten aufgezeigt wurden und anschließend das angebotene Big Data-Produkt im Allgemeinen und in Bezug auf die aufgestellten Merkmale untersucht wurde. Am Ende jedes Kapitels fand schließlich eine Kategorisierung der Produkte statt. Insgesamt wurden von den 20 Anbietern 27 Produkte im Zusammenhang mit Big Data genauer untersucht, da einige Anbieter mehrere Produkte anbieten. Jedes Produkt kann einer Kategorie zugeordnet werden, wobei einige auch in mehrere Kategorien passen. Das liegt z.B. daran, dass manche vorgestellten Produkte eine Appliance darstellen und somit umfangreich aufgestellt sind. Aus diesem Grund finden sich in der folgenden Zuordnung zu den einzelnen Kategorien 39 Nennungen.

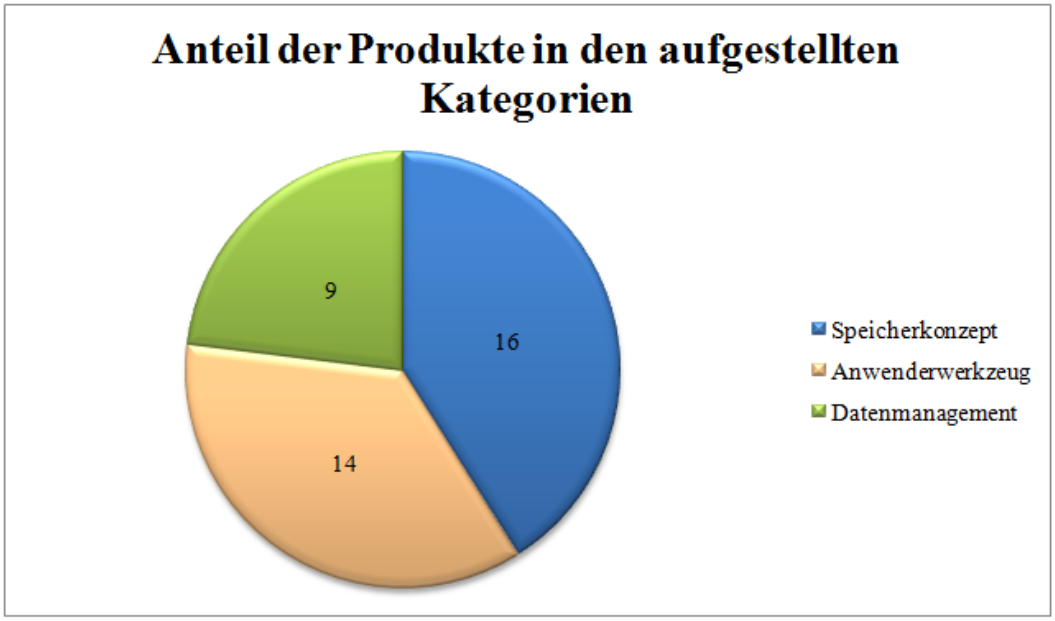


Abbildung 3.52: Anteil der Produkte in den aufgestellten Kategorien

Die Abbildung 3.52 zeigt, dass insgesamt 39 Nennungen innerhalb der Kategorien zu finden sind. Dabei sind in erster Linie Speicherkonzepte (41%) am häufigsten und direkt danach Anwenderwerkzeuge (36%) vertreten. Bei den untersuchten Produkten sind die wenigsten der Kategorie Datenmanagement zuzuordnen (23%). Interessant ist jedoch, welche Produkte letztendlich in welcher Kategorie zu finden sind. Generell lassen sich zwar nicht alle genannten Produkte miteinander vergleichen aus den in Kapitel 3.4.1.1 genannten Gründen. Dennoch kann innerhalb der Kategorien eine allgemeine und unabhängige Wertung vorgenommen werden. So kann z.B. die Qualität des Supports auch unabhängig von einer Unternehmenssituation verglichen werden, ebenso wie die Notwendigkeit der Verarbeitung von verschieden strukturierten Daten im Kontext von Big Data. Nachfolgend wird daher tabellarisch ein Überblick über die in 3.4.1.2 genannten und erläuterten Kategorien und der ihnen zugeordneten Produkte gegeben. Zunächst erfolgt ein Überblick über die Speicherkonzepte, darauf folgt eine Übersicht der Anwenderwerkzeuge und abschließend die Darstellung der Datenmanagement-Produkte. Vorab soll jedoch die für das Verständnis der Übersichten notwendige Erklärung der Wertungen vorgenommen werden (die Erläuterungen hinsichtlich der Untersuchungsmerkmale werden in Kapitel 3.4.1.2 bereits ausführlich dargestellt):

Untersuchungsmerkmal	Kurzbezeichnung	*	**	***
Möglichkeit der Integration in bestehende BI-Architekturen	Integrationsfähigkeit	Die Produktintegration ist an Hardwareprodukte ausgewählter Hersteller gebunden.	Das Produkt ist flexibel anpassbar und nicht an Hardware ausgewählter Hersteller gebunden.	
Skalierbarkeit/ Zukunftssicherheit	Skalierbarkeit	Laut Hersteller ist das Produkt skalierbar, allerdings gibt es zur Grenze keine näheren Informationen.	Das Produkt lässt sich bis zu einer vorgegebenen Grenze skalieren.	Das Produkt lässt sich nach Belieben erweitern ohne vorgegebene Grenzen.
Verarbeitung verschiedener Datenquellen (Schnittstellen)	Schnittstellen	Es sind gängige Schnittstellen vorhanden, z.B. SAP BO, Oracle Database, JDBC u.v.m.	Der Anbieter nennt zusätzlich zu gängigen Schnittstellen auch spezielle Schnittstellen zu bspw. anderen Big Data Produkten, wie z.B. zu Hadoop.	
Verarbeitung verschiedener Datenformate	Datenformate	Verarbeitung gängiger Datenformate, wie z.B. XML.	Verarbeitung von polystrukturierten Daten, wie z.B. Log-Files.	
Umgang mit semi- und unstrukturierten Daten	Datenstruktur	Das Produkt kann strukturierte und semistrukturierte Daten verarbeiten.	Das Produkt kann sowohl strukturierte als polystrukturierte Daten verarbeiten.	
Benutzeroberfläche/ Benutzerfreundlichkeit	Benutzeroberfläche	Das Produkt bringt zwar keine eigene Benutzeroberfläche mit, kann aber an Frontends einiger anderer Produkte angeschlossen bzw. diesen angepasst werden.	Das Produkt bringt eine eigene Benutzeroberfläche mit.	
Softwaredokumentation	Dokumentation	Es werden Dokumente über das Produkt zur Verfügung gestellt.	Zusätzlich zu Dokumenten werden Foren, Tutorials, Webinars u.ä. angeboten.	
Leistungsfähigkeit des Anbieters (Branchenkompetenz)	Branchenkompetenz	Der Anbieter hat einen Kundenkreis von insgesamt < 5.000 Kunden	Der Anbieter hat einen Kundenkreis von insgesamt >5.000 Kunden	Der Anbieter hat einen Kundenkreis von insgesamt >10.000
Qualität des Supports	Support	Es gibt einen Standardsupport für alle Kunden gleichermaßen.	Es stehen mehrere Supportpakete zur Auswahl, die u.a. einen 24/7 Support beinhalten können.	

Abbildung 3.53: Erklärungen zu den Wertungen innerhalb der Kategorieübersichten

In der linken Spalte finden sich die Untersuchungsmerkmale, die als Grundlage für die Untersuchung der Produkte dienten. In der Spalte „Kurzbezeichnung“ wird jedem Untersuchungsmerkmal eine Kurzbezeichnung zugeordnet, um die nachfolgenden Abbildungen übersichtlicher zu gestalten. Die Spalten 3-5 geben Erklärungen wieder, welche Bedeutung die einzelnen Wertungen haben, die in den Abbildungen 3.54, 3.55 und 3.56 vorgenommen wurden. Diese Abbildungen sind schließlich so aufgebaut, dass in den beiden rechten Spalten die untersuchten Produkte innerhalb ihrer Kategorien alphabetisch aufgelistet sind. Danach folgt die Bewertung mit dem in Abbildung 3.53 vorgestellten Bewertungsprinzip. Die Produkte werden nachfolgend zusammenfassend innerhalb ihrer Kategorien ihren Unterkategorien zugeordnet und schließlich bewertet. Die letzte Spalte gibt an, ob einzelne Produkte weiteren Kategorien zuzuordnen sind.

Zusätzlich zu den aus der Abbildung 3.53 gewonnenen Informationen bezüglich der Wertung und Einordnung der Produkte innerhalb ihrer Kategorie, gilt wie auch schon zuvor, dass k.A. dafür steht, dass keine Angaben aus diversen Gründen gemacht werden konnten. Es folgen daher nun die Darstellungen bezüglich der Zuordnungen und Wertungen innerhalb der Kategorien:

Speicherkonzepte		Relationale DB	Analytische DB	Multi-dimensionale DB	Hadoop	Sonstige	Integrations-fähigkeit	Skalierbarkeit	Schnittstellen	Datenformate	Umgang mit Datenstruktur	Benutzer-oberfläche	Dokumentation	Branchen-kompetenz	Support	Weiteren Produktkategorien zuordenbar
Cloudera	Cloudera Enterprise				x		**	*	**	**	**	k.A.	*	k.A.	**	→Datenmanagement
Datameer	Datameer				x		**	*	**	**	**	**	**	*	*	→Anwenderwerkzeug
EMC	Greenplum Database		x				**	*	*	k.A.	**	k.A.	*	k.A.	*	
	Greenplum HD															
EXASOL	EXASolution		x				**	**	**	**	**	*	*	k.A.	**	→Datenmanagement
Hortonworks	Data Platform		x		x		**	*	**	**	**	**	**	k.A.	**	
HP	Vertica Analytics Platform		x				**	*	**	**	*	**	k.A.	k.A.	*	
IBM	InfoSphere BigInsights		x		x		**	k.A.	*	**	**	**	*	k.A.	k.A.	→Anwenderwerkzeug →Datenmanagement
	InfoSphere Streams					x	**	*	*	**	**	**	*	k.A.	k.A.	→Anwenderwerkzeug →Datenmanagement
	Netezza		x				k.A.	**	*	**	**	**	*	k.A.	k.A.	→Anwenderwerkzeug →Datenmanagement
	Cognos Business Intelligence			x			k.A.	***	*	k.A.	k.A.	k.A.	*	k.A.	k.A.	→Anwenderwerkzeug
Oracle	Big Data Appliance				x	x	**	k.A.	**	*	**	k.A.	**	***	**	→Anwenderwerkzeug →Datenmanagement
Parstream	ParStream		x				**	k.A.	*	*	*	**	k.A.	k.A.	k.A.	
SAP	HANA		x				*	**	**	*	**	**	**	***	*	
Software AG	Terracotta Big Memory		x				**	k.A.	*	*	*	**	*	k.A.	**	
Teradata	Aster Big Analytics Appliance		x				**	*	**	**	**	**	k.A.	k.A.	*	

Abbildung 3.54: Kategorieübersicht Speicherkonzepte

Es handelt sich überwiegend um analytische Datenbanken, die wie bereits in Kapitel 3.4.1.2 erläutert wurde, eine bessere Skalierbarkeit als relationale Datenbanken gewährleisten sollen und oftmals im Zusammenhang mit In-Memory-Datenbanken stehen. Auffällig ist ebenfalls, dass es sich bei keiner der untersuchten Produkte um eine relationale Datenbank handelt. Dies könnte daran liegen, dass derartige Datenbanken der Big Data Problematik längst nicht mehr gewachsen sind. Ebenfalls ist festzustellen, dass die Mehrzahl der hier aufgeführten Produkte problemlos und flexibel in bestehende BI-Architekturen integrierbar sein sollen. Bei den Untersuchungsmerkmalen Schnittstellen, Skalierbarkeit und Datenformate herrschen dabei die meisten Unterschiede hinsichtlich der Wertung. Interessant ist, dass IBM und Oracle beispielsweise mit ihren Big Data Lösungen gleich mehrere Produktkategorien abdecken, IBM bei drei von vier Produkten sogar alle drei Kategorien. Zudem ist erkennbar, dass im Zusammenhang mit der Big Data Thematik und der damit einhergehenden Problematik der polystrukturierten Daten, die meisten der hier aufgeführten Produkte der Schwierigkeit der Variety gewachsen sind im Bezug auf die Verarbeitung von semi- und unstrukturierten Daten.

Anwenderwerkzeuge		Dashboarding	Berichtswesen	Analyse	Data Mining	Integrations-fähigkeit	Skalierbarkeit	Schnittstellen	Datenformate	Umgang mit Datenstruktur	Benutzer-oberfläche	Dokumentation	Branchen-kompetenz	Support	Weiteren Produktkategorien zuordenbar
Datameer	Datameer	*	*	**		**	*	**	**	**	**	**	*	*	→Speicherkonzept
Datawatch	Panopticon	*		**		**	k.A.	**	*	**	**	k.A.	***	*	
IBM	InfoSphere BigInsights			*	*	**	k.A.	*	**	**	**	*	k.A.	k.A.	→Speicherkonzept →Datenmanagement
	InfoSphere Streams			*	*	**	*	*	**	**	**	*	k.A.	k.A.	→Speicherkonzept →Datenmanagement
	Netezza			*		k.A.	**	*	**	**	**	*	k.A.	k.A.	→Speicherkonzept →Datenmanagement
	Cognos Business Intelligence			*		k.A.	***	*	k.A.	k.A.	k.A.	*	k.A.	k.A.	→Speicherkonzept
Jaspersoft	Instant Big Data Analytics	*	*	*		**	*	**	k.A.	**	**	**	***	**	
MicroStrategy	MicroStrategy 9.3	**	**	**	*	**	*	**	**	**	**	*	*	**	
Oracle	Big Data Appliance			**		**	k.A.	**	*	**	k.A.	**	***	**	→Speicherkonzept →Datenmanagement
Pentaho	Business Analytics	**	**	**	**	**	*	**	*	**	**	*	k.A.	**	→Datenmanagement
QlikTech	QlikView	**	*	**		**	k.A.	**	**	**	**	k.A.	***	*	
SAP	Business Objects Business Intelligence Suite	**	**	**	*	Keine Untersuchung vorgenommen									
SAS	Visual Analytics	*	*	**		**	***	**	*	**	**	**	***	*	
Splunk	Splunk Enterprise	**	*	**		k.A.	***	**	**	**	**	*	**	***	

Abbildung 3.55: Kategorieübersicht Anwenderwerkzeuge

Bezüglich der Anwenderwerkzeuge verhält es sich ähnlich wie mit den Speicherkonzepten. Denn auch hier werden der Unterkategorie Analyse nicht nur am meisten, sondern sämtliche Produkte der Kategorie Anwenderwerkzeuge zugeordnet. Die meisten Produkte der Unterkategorie Analyse haben sogar ihren Schwerpunkt (gekennzeichnet durch **) auf diese gelegt. Produkte, die der Unterkategorie Data Mining zuzuordnen sind, sind eher weniger vertreten. Allerdings sind Berichtswesen und Dashboarding gleichermaßen vertreten, denn beide Unterkategorien sind im Kontext der Big Data Charakteristik Analytics ebenso wichtig wie die Analyse. Auch bei den Anwenderwerkzeugen zeigt sich, dass die Anbieter Wert auf die Integrationsfähigkeit sowie die Verarbeitung von polystrukturierten Daten legen. Für die Integrationsfähigkeit stellen viele der Anbieter daher verschiedene für die Big Data Thematik interessante und relevante Schnittstellen zur Verfügung auch zu anderen hier untersuchten Produkten, wie z.B. Hadoop oder SAP HANA.

Datenmanagement		Datenintegration	Datenqualität	Integrations-fähigkeit	Skalierbarkeit	Schnittstellen	Datenformate	Umgang mit Datenstruktur	Benutzer-oberfläche	Dokumentation	Branchen-kompetenz	Support	Weiteren Produktkategorien zuordenbar
Cloudera	Cloudera Enterprise	x		**	*	**	**	**	k.A.	*	k.A.	**	→Speicherkonzept
EXASOL	EXASolution	x		**	**	**	**	**	*	*	k.A.	**	→Speicherkonzept
IBM	InfoSphere BigInsights	x	x	**	k.A.	*	**	**	**	*	k.A.	k.A.	→Anwenderwerkzeug →Speicherkonzept
	InfoSphere Streams	x	x	**	*	*	**	**	**	*	k.A.	k.A.	→Anwenderwerkzeug →Speicherkonzept
	Netezza	x	x	k.A.	**	*	**	**	**	*	k.A.	k.A.	→Anwenderwerkzeug →Speicherkonzept
Informatica	PowerCenter Big Data Edition	x		**	***	**	**	**	**	*	*	**	
Oracle	Big Data Appliance	x		**	k.A.	**	*	**	k.A.	**	***	**	→Anwenderwerkzeug →Speicherkonzept
Pentaho	Data Integration	x		**	k.A.	*	*	*	**	k.A.	k.A.	k.A.	→Anwenderwerkzeug
SAP	Data Services	x	**	**	k.A.	**	*	**	**	**	***	k.A.	

Abbildung 3.56: Kategorieübersicht Datenmanagement

In der letzten aufgestellten Kategorie Datenmanagement fällt als erstes auf, dass hier wesentlich weniger Produkte aus der Liste der untersuchten Produkte zu finden sind als bei den zuvor dargestellten Kategorien. Wichtig hierbei ist jedoch, dass der Aspekt Velocity (Übertragungsgeschwindigkeit) im Zusammenhang mit der Big Data Thematik für viele Unternehmen nicht das erstrangige Ziel ist. In vielen Unternehmenssituationen, geht es vielmehr um die Möglichkeit, die großen Datenmengen zu speichern und zu verarbeiten, worauf in den vorherigen Kategorien eingegangen wurde. Das Datenmanagement spielt deshalb allerdings keine untergeordnete Rolle. Schließlich müssen die Daten in die IT-Systemlandschaft der Unternehmen integriert werden. Aus diesem Grund sind alle Produkte in der Kategorie Datenmanagement auch der Unterkategorie Datenintegration zugeordnet. Hier spielt auch das Untersuchungsmerkmal der Verarbeitung verschiedener Datenquellen (hier als Schnittstellen bezeichnet) eine wichtige Rolle. Denn die Datenintegration wird wie in Kapitel 3.4.1.2 bereits erläutert wurde, dafür benötigt, Daten aus unterschiedlichen Quellen auszulesen.

Der vorliegende Marktüberblick bietet keine vollständige Übersicht über die am Markt vorhandenen Big Data-Produkte, verschafft jedoch einen guten Einblick über die Leistungsfähigkeit der Produkte und worin sie sich ähneln oder unterscheiden. Für den Fall, dass ein Unternehmen den Kauf eines Big Data-Produktes in Betracht zieht, ist es ratsam, dass von den Verantwortlichen genau überlegt wird, für welche Zwecke das Produkt eingesetzt werden soll und welche Merkmale es aus diesem Grund haben sollte. In die Entscheidung sollte mit einfließen, ob das Produkt ein Speicherkonzept, ein Anwenderwerkzeug, ein Datenmanagement-Tool oder eine Kombination daraus sein soll. Dadurch können einige Produkte bereits ausgeschlossen werden. Dann können vor allem Schnittstellen und die Möglichkeit der Integration in die bestehende BI-Architektur ausschlaggebend für ein bestimmtes Big Data-Produkt sein. Außerdem muss für die Verantwortlichen klar sein, ob Daten in Echtzeit benötigt werden und ob beispielsweise

Hadoop von dem Produkt unterstützt werden soll. Das Unternehmen kann sich dann noch entscheiden, ob es Einzellösungen möchte oder ob es eine Appliance vorzieht, um aufeinander abgestimmte Software und Hardware aus einer Hand zu haben.

3.4.4 Schlussbetrachtung

Im Zuge dieses Marktüberblicks wurde das Produktportfolio von 20 Anbietern hinsichtlich der Big Data-Produkte untersucht. In Kapitel 3.4.1 wird dem Leser zunächst die Motivation zu diesem Marktüberblick vorgestellt und dabei der Zusammenhang mit der Aufgabenstellung hinsichtlich der Arbeitsphasen der Allgemeinen Konzeption und der Konzeption für das Partnerunternehmen deutlich gemacht. Diesbezüglich wird sowohl auf grundlegende Informationen zu der Big Data-Thematik eingegangen als auch auf die Problematik, die damit einhergeht - die Integration, Speicherung und Analyse von Big Data. Aus diesem Grund wurden Kategorien mit zugehörigen Unterkategorien aufgestellt, die sich mit dieser Problematik befassen und in welche die untersuchten Produkte eingeordnet wurden. Die dazugehörigen Erläuterungen werden in Kapitel 3.4.1.2 dargestellt. Durch die Kapitel 3.4.1.1 und 3.4.1.3 sollen dem Leser schließlich das Vorgehen während der Erstellung des Marktüberblicks nahegebracht werden. Im Kapitel 3.4.2 werden die Anbieter und ihre Big Data-Produkte schließlich auf Basis aufgestellter Untersuchungsmerkmale untersucht, die ebenfalls in Kapitel 3.4.1.2 erklärt werden, und letztendlich den Kategorien sowie Unterkategorien zugeordnet. In Kapitel 3.4.3 werden schließlich die Erkenntnisse, die aus dem Marktüberblick gewonnen werden konnten, dargestellt.

Dabei ist festzustellen, dass bei den drei aufgestellten Kategorien Speicherkonzepte, Anwenderwerkzeuge und Datenmanagement Parallelen zu erkennen sind. Denn es zeigt sich, dass bei den untersuchten Produkten sowohl auf die Integrationsfähigkeit als auch auf die Schnittstellen sowie auf die Verarbeitung von polystrukturierten Daten besonderer Wert gelegt wird. Da diese Untersuchungsmerkmale in einem direkten Zusammenhang mit den in Kapitel 3.4.1 genannten Charakteristika von Big Data stehen, lässt sich festhalten, dass die meisten der untersuchten Produkte der Big Data-Thematik auf ihre individuelle Art und Weise gewachsen sind. Auch ist anzumerken, dass im Kontext der Verarbeitung von verschiedenen Datenquellen viele der Produkte Schnittstellen zu ebenfalls untersuchten Produkten aufweisen, wie z.B. Schnittstellen zu SAP HANA oder Komponenten der Oracle Appliance. Trotz der Tatsache, dass ein Produkt für jedes Unternehmen unterschiedlich attraktiv ist, durch z.B. individuelle IT-Infrastrukturkomponenten oder einer speziellen Datenintegrationsproblematik, konnten Produkte aus einer Kategorie zumindest ansatzweise miteinander verglichen werden. Dadurch konnte deutlich gemacht werden, welche Produkte sich der Charakteristika von Big Data annehmen. Ebenso kann festgestellt werden, dass ein Big Data-Produkt nicht durch den Umfang des Produktportfolios eines Herstellers attraktiver wird. So bietet EXASOL beispielsweise mit EXA-Solution lediglich ein Basisprodukt an, ist jedoch gemäß der Kategorisierung Speicherkonzepte gleichauf hinsichtlich der Untersuchungsmerkmale mit einer Lösung des weltweit bekannten Unternehmens SAP.

Zur Vorgehensweise bei der Erstellung dieses Marktüberblicks ist festzuhalten, dass es sich im Allgemeinen als sehr schwierig erweist, Produkte miteinander zu vergleichen, ohne eine konkrete Unternehmenssituation vor Augen zu haben. So können Empfehlungen für ein Produkt letztendlich nur dann ausgesprochen werden, wenn die IT-Infrastruktur sowie konkrete Anforderungen an ein gewünschtes Produkt vorhanden sind. Dennoch wurde bei der Erstellung des Marktüberblick angestrebt, die Produkte innerhalb ihrer Kategorien zu vergleichen, sodass interessierte Unternehmen einen ersten Einblick über eine Auswahl an aktuellen Big Data-Produkten erhalten. Die Resultate hierzu verdeutlicht das Kapitel 3.4.3.

Die Einordnung in die Kategorien hat sich dabei als weniger schwer erwiesen als die Recherche zu den aufgestellten Untersuchungsmerkmalen. Es ist festzustellen, dass einige Anbieter bei der Herausgabe technischer Informationen bezüglich der angebotenen Produkte zurückhaltend sind bzw. lediglich interessierten Unternehmen Informationen nur auf Anfrage und meist auch kostenpflichtig zur Verfügung stellen. Dies hat zur Folge, dass hinter vielen Produkten ein hoher Rechercheaufwand steckt. Das stellt insbesondere für kleinere Unternehmen eine Hürde dar, da sie zum einen meist nicht genügend finanzielle und zum anderen auch kaum personelle Ressourcen haben, um diese Recherchearbeit zu leisten. Da die Informationsbasis teilweise recht gering ist, wird interessierten Unternehmen ein Vergleich der Produkte und ihrer Merkmale erschwert oder sogar verwehrt. Des Weiteren werden unterschiedliche Fach-Termini verwendet, sodass eine klare Definition einzelner Begriffe im Zuge dieses Marktüberblicks erforderlich war. Diese wurden eingangs in den Kapiteln 3.4.1.2 und 3.4.1.3 festgehalten.

Da das Thema Big Data seit einiger Zeit sehr stark im Fokus der Öffentlichkeit steht, hat sich das Produktportfolio der auf dem Markt agierenden Anbieter entsprechend geändert. Denn hoch aktuelle und brisante Themen bzw. Problemstellungen bieten dem Markt stets die Gelegenheit sich zu erweitern. Es folgen meist nicht nur Produkterweiterungen einzelner bereits auf dem Markt vertretenen Anbieter, sondern auch die Gründung von Unternehmen, die sich auf diese Problematik spezialisieren. Damit geht für die an Big Data-Produkten interessierten Unternehmen die Herausforderung einher, einen Überblick über die neuen sowie aktualisierten Produkte und deren Anbieter zu erhalten, um das für sie bestmögliche Produkt zu finden. Der ausgearbeitete Marktüberblick soll deshalb helfen, eine Übersicht im Bereich der angebotenen Big Data-Produkte zu erhalten. Zusammenfassend lässt sich sagen, dass dieser Marktüberblick versucht die wichtigsten Eigenschaften der insgesamt 27 untersuchten Produkte wiederzugeben sowie die dazugehörigen Unternehmen darstellt. Interessierte Unternehmen erhalten mit diesem Marktüberblick einen Eindruck über die am Markt erhältlichen Big Data-Produkte und Herangehensweisen an das komplexe Thema.

3.5 Erstellung eines Prototyps zur Datenintegration

Der Praxispartner der Teilprojektgruppe möchte ein systemübergreifendes Capacity Management einführen, das es ermöglicht, Kapazitätsauslastungen der IT-Services sowie der IT-Infrastruktur zu analysieren und zu simulieren. Ziel hierbei ist es etwaige Kapazitätsengpässe frühzeitig zu identifizieren. Hierzu soll ein Prototyp entwickelt werden, mit dessen Hilfe Daten aus unterschiedlichen Quellen zum Zwecke der Analyse und der Simulation integriert werden können. Im folgenden Kapitel wird dieser Prototyp beschrieben. Neben der Aufgabenstellung und einer Beschreibung der verwendeten Werkzeuge, wird auch das Konzept sowie die praktische, prototypische Umsetzung erläutert.

3.5.1 Aufgabenstellung

Für den eingangs erwähnten Anwendungsfall soll eine prototypische Umsetzung erfolgen. Das Ziel dieses Prototyps ist die Integration verschiedener Datenquellen in ein gemeinsames Zielschema. Dieses Zielschema dient als Datenbasis für Berichte und Simulationen des Capacity Management. Den Kern dieser zu integrierenden Daten bilden Access-Log-Dateien, welche durch Webanwendungen generiert werden. Weitere zu integrierende Datenquellen sind Nutzerkontextdaten eines Webportals sowie unterschiedliche Arten von Stammdaten, welche zum späteren Verknüpfen, Zuordnen und Aggregieren der Daten dienen. Diese Daten werden vom Praxispartner über ein Server-Verzeichnis bereitgestellt. Im Anschluss an den Datenimport werden die Access Logs mit den oben erwähnten Stammdaten verknüpft, um eine Zuordnung zwischen den einzelnen Einträgen eines Access Logs und fachlichen Anwendungsfällen (Use Cases) zu ermöglichen. Am Ende des Prozesses sollen die Daten nach Land und auf die Minute aggregiert im Zielschema vorliegen. Zur Umsetzung dieses Datenintegrationsprozesses, soll die Software Informatica PowerCenter verwendet werden. Nachdem die Daten im Zielschema vorliegen, sollen beispielhaft einige Berichte auf Basis dieser Daten erstellt werden. Zur weiteren Präzisierung der Aufgabenstellung und Anforderungen des Praxispartners, wurde bereits zu einem früheren Zeitpunkt ein Pflichtenheft zu diesem Prototyp erstellt (Vgl. Anhang F.3).

3.5.2 Beschreibung der Werkzeuge

Für die prototypische Umsetzung des ETL-Prozesses wird vom Praxispartner eine Reihe von Werkzeugen im Rahmen einer Testumgebung bereitgestellt. Im Folgenden werden diese drei Tools genauer erläutert.

Informatica PowerCenter

Das Informatica PowerCenter ist ein Werkzeug zur Datenintegration und wird in diesem Prototyp zur Umsetzung des ETL-Prozesses verwendet. Es bietet die Möglichkeit ETL-Prozesse Schritt für Schritt zu modellieren und anschließend in einem sog. Workflow auszuführen. Des Weiteren wird die Anbindung von Datenquellen und Datenzielen unterschiedlicher Art und Hersteller unterstützt. Die Anwendung unterteilt sich in eine Server- sowie eine Client-Komponente.

Der Server wird auf Seiten des Praxispartners betrieben. Der Zugriff auf den Server wird mit Hilfe eines Repositorys bereitgestellt. Der Client ist wiederum lokal beim Anwender installiert und stellt den Zugriff auf das Repository und die darin enthaltenen Dateien her. Die Client-Installation besteht außerdem aus einer Reihe von verschiedenen Tools. Die für diese Ausarbeitung relevanten Werkzeuge sind jedoch nur der Informatica PowerCenter Designer, der Workflow Manager, der Workflow Monitor sowie der Repository Manager. Alle diese Tools werden jeweils in der Version 9.1 verwendet.

Oracle Datenbank

Die Oracle Datenbank fungiert in diesem Prototyp zur permanenten Speicherung der Daten. Auf Basis dieser Datenbank werden alle Schichten des ETL-Prozesses realisiert (Vgl. Kap. 3.5.3). Die Oracle Datenbank wird, genau wie auch das Repository des Informatica PowerCenters, vom Praxispartner bereitgestellt. Bei der bereitgestellten Datenbank handelt es sich um die Version Oracle Database 11g Enterprise Edition.

Oracle SQL Developer

Der Oracle SQL Developer ist ein Entwicklungs- und Verwaltungstool für Datenbanken. Im Rahmen dieser Umsetzung dient das Tool hauptsächlich zum Zugriff auf die im vorherigen Abschnitt beschriebene Oracle Datenbank und der Anzeige der darin enthaltenen Datenbankobjekte wie beispielsweise Tabellen oder Views. Außerdem können hiermit Verwaltungsaufgaben sowie Testabfragen auf der Datenbank ausgeführt werden. Der Oracle SQL Developer ist als Freeware auf der Homepage des Herstellers Oracle erhältlich. In dieser Umsetzung wurde die zu diesem Zeitpunkt aktuelle Version 4.0 des Oracle SQL Developer verwendet.

Eclipse BIRT

Für die Erstellung der im Rahmen dieser prototypischen Umsetzung vorgesehenen Reports, wurde das Open Source Projekt Eclipse BIRT verwendet. Der Name BIRT ist ein Akronym für Business Intelligence and Reporting Tool. Es handelt sich hierbei um ein Projekt der Eclipse Foundation. BIRT basiert auf der Eclipse Plattform und bietet Funktionalitäten für das Berichtswesen, Business Intelligence sowie mehrdimensionales OLAP (Online Analytical Processing). Das Programm kann über die Homepage der Eclipse Foundation kostenlos heruntergeladen und lokal auf dem Client-PC installiert werden.

3.5.3 Konzeption

Im folgenden Kapitel soll die Konzeption der prototypischen Umsetzung der Datenintegration anhand der einzelnen Schritte betrachtet und beschrieben werden. Zentraler Punkt des Konzeptes ist ein ETL-Prozess, welcher von der Integration der Daten aus den Quelldateien bis hin zum Bereitstellen der Daten für das Reporting alle Funktionen abdeckt.

3.5.3.1 Integration der Quelldaten

Im ersten Schritt der Datenintegration müssen die Quelldateien extrahiert und in die ODS-Schicht überführt werden. Zu diesen Quelldateien zählen die zu analysierenden Access Logs von Webservern, Nutzerkontextdaten aus einem Onlineportal unseres Praxispartners sowie Stammdaten zur Zuordnung der Access Logs zu fachlichen Anwendungsfällen und aufrufenden Anwendungen.

Die erste Art von Daten, die Access Logs, sind die am häufigsten vorkommenden Daten und Kern unserer Datenbank sowie der späteren Auswertungen und Simulationen. Diese Access Logs werden über ein Server-Verzeichnis des Praxispartners bereitgestellt. Sie weisen immer dieselbe Struktur auf, welche in der folgenden Tabelle aufgezeigt wird:

Pos	Bezeichnung	Bedeutung
1	Client IP	IP-Adresse oder DNS-Name des Client
2	Server IP	IP-Adresse oder DNS-Name des Servers
3	IV-User	Name des Users oder -
4	Entry Date	Log Entry Date/Time
5	Call URL	HTTP URL
6	Call URL Parameter	Parameter for HTTP URL
7	HTTP Return Code	HTTP Status
8	Data Size	Größe der Antwort in Byte
9	Referrer URL	Vorgänger-URL
10	Browser	Browser-Typ
11	Responsetime	Antwortzeit in Mikrosekunden

Tabelle 3.25: Aufbau einer Access-Log-Datei

Die Nutzerkontextdaten, welche aus einem Onlineportal des Praxispartners stammen, sind die zweite Art von Daten. Diese sollen im ETL-Prozess zur Aggregation der Daten auf Länderebene dienen. Hierzu wird eine Zuordnung der User aus dem Access Log zu einem Land hergestellt. Die Struktur der Daten sieht wie folgt aus:

Pos	Bezeichnung	Bedeutung
1	Global User ID	Weltweit eindeutige ID
2	User ID	Global User ID oder User ID
3	Land	Land
4	Org ID	Organisationseinheit

Tabelle 3.26: Aufbau der Tabelle für die Nutzerkontextdaten

Die letzte Art von Quelldaten sind die Stammdaten zur Zuordnung der Access Logs zu fachlichen Anwendungsfällen (Use Cases) und aufrufenden Anwendungen. Diese Stammdaten werden, anders als die Access Logs, nicht laufend neu importiert, sondern einmalig in das System

überführt, da sie fortan unverändert bleiben.

3.5.3.2 ODS-Schicht

Der Operational Data Store (ODS) dient in diesem Konzept dazu, die operativen Daten der Quellsysteme in einer relationalen Schicht zusammenzuführen. In diesem Szenario gehören sowohl die Access Logs als auch die Stamm- und Nutzerkontextdaten des Onlineportals zu den operativen Daten. Zur Realisierung der ODS-Schicht, wird eine relationale Datenbank benötigt, in welche die Daten mithilfe eines ETL-Tools importiert werden. Die Tabellen in der Datenbank weisen hierbei die selbe Struktur wie die Quelldateien auf (Vgl. Abb. 3.57), da diese Eins zu Eins in die ODS-Schicht übernommen werden sollen. Die ODS-Schicht wird die Daten jedoch nur temporär und nicht dauerhaft vorhalten.

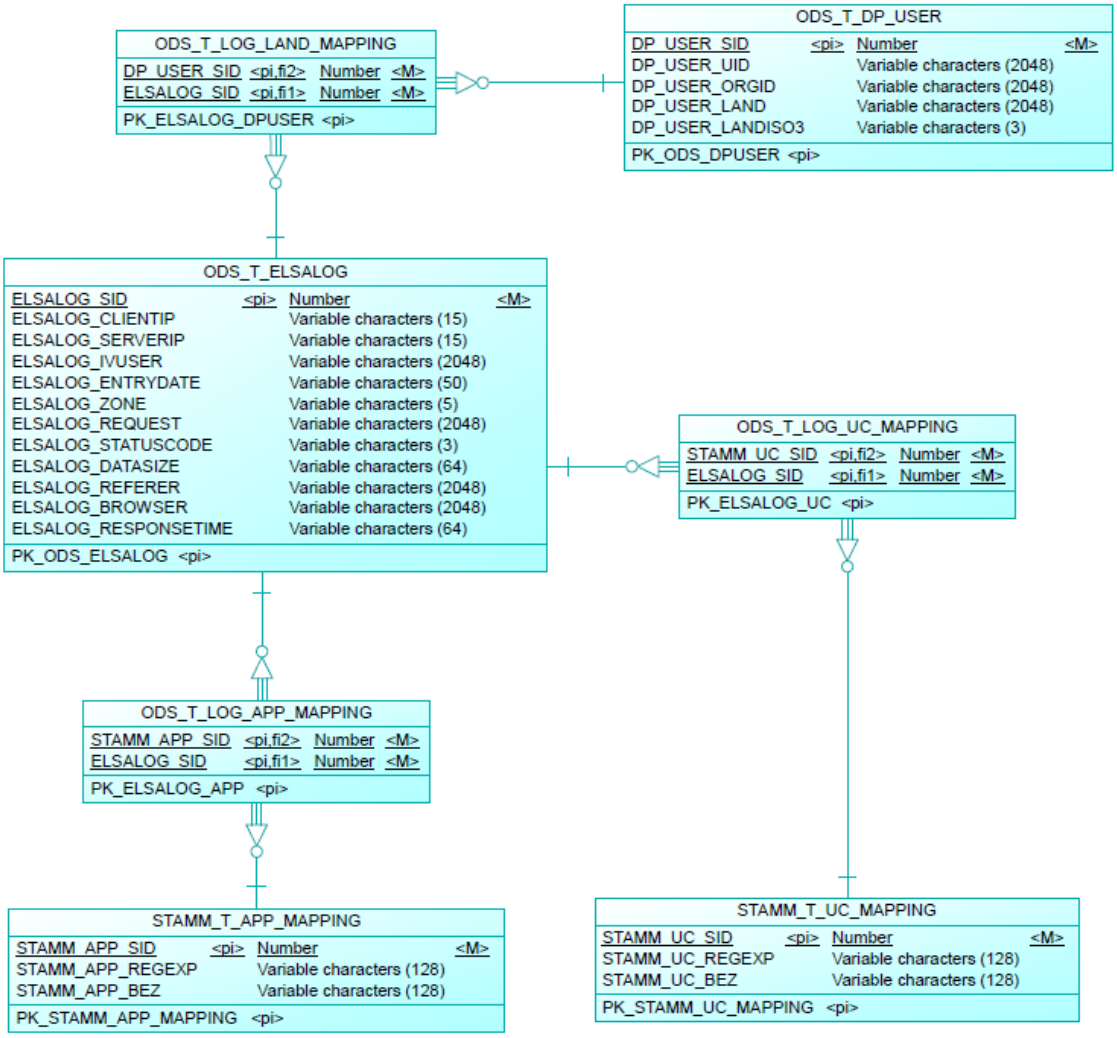


Abbildung 3.57: Logisches Datenmodell der ODS-Schicht

3.5.3.3 Staging Area

Die dritte Ebene der Datenintegration ist die sogenannte Staging Area bzw. der Arbeitsbereich. Dieser dient dazu, die bereits im Operational Data Store zusammengeführten Daten zu vereinheitlichen sowie deren Datenkonsistenz und -qualität zu überprüfen, ehe sie in das eigentliche Data Warehouse geladen werden. Die Staging Area ist lediglich ein temporärer Speicher, welcher die Daten nicht dauerhaft vorhält. In diesem Schritt werden die Daten in eine Staging Tabelle geladen. Hierbei muss zum Beispiel die Zeitangabe, welche im ODS als Charakter vorliegt, in einen Timestamp umgewandelt werden. Des Weiteren müssen in diesem Schritt fehlerhafte Daten, beispielsweise Nullwerte, herausgefiltert werden. Auch bekommen die übrigen Datenfelder, bei denen es notwendig ist, einen anderen Datentyp. Ein Beispiel hierfür ist das Feld „Datasize“, welches für eine spätere Aggregation in einen Integer umgewandelt werden muss. Nach diesen Transformationen geht es weiter in die eigentliche Data Warehouse Schicht.

3.5.3.4 Data Warehouse

In der Ebene des Data Warehouse werden die Daten in Fakten- und Dimensionstabellen unterteilt. So entsteht eine Zeitdimension, welche den Zeitstempel aus einem Access Log auf mehrere Datenfelder für das Jahr, den Monat, den Tag, die Stunde, die Minute sowie die Sekunde aufteilt. Außerdem je eine Dimension für die Use Cases, für die aufrufende Anwendung sowie die Nutzerkontextdaten. Die Verknüpfung dieser Daten erfolgt in der zentralen Faktentabelle. Diese Tabelle beinhaltet einige aggregierte Fakten, welche Kennzahlen abbilden. Daraus ergibt sich das in Abbildung 3.58 gezeigte Zielschema, das die Struktur des Data Warehouse repräsentiert.

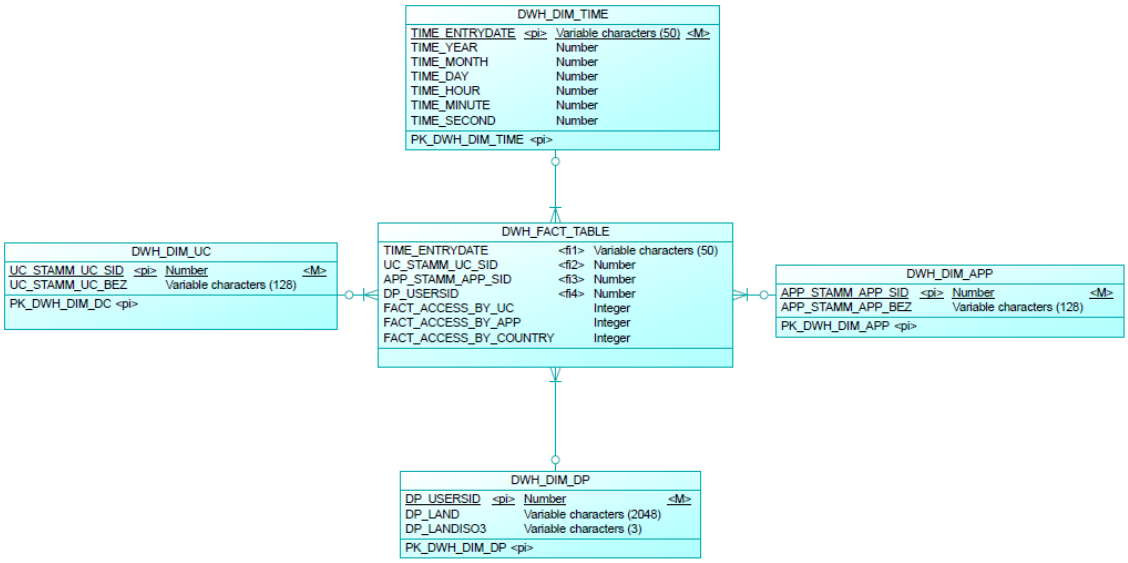


Abbildung 3.58: Logisches Datenmodell der DWH-Schicht

3.5.3.5 Reporting

Zur Bereitstellung der Daten für das Reporting, sollen auf der Datenbank Views erstellt werden. Diese Views dienen anschließend dem Reporting Tool als Datenbasis zur Erstellung der einzelnen Reports. Im Rahmen dieser prototypischen Umsetzung werden nur einige Berichte beispielhaft umgesetzt. So sollen lediglich die Berichte für die Anzahl der Zugriffe je Use Case, die Zugriffe je Stunde, die Zugriffe je Use Case und Stunde sowie Zugriffe je Land umgesetzt werden. Diese Berichte basieren dabei auf unterschiedlichen Arten von Diagrammen.

3.5.4 Umsetzung

In diesem Kapitel wird die konkrete Umsetzung des zu vor dargestellten Prototyps beschrieben. Hierzu wird zunächst auf die Installation und Konfiguration der in Kapitel 3.5.2 aufgeführten Werkzeuge eingegangen. Anschließend folgt die Dokumentation der durchgeführten, prototypischen Entwicklung. Abgeschlossen wird dieser Abschnitt mit einer Schlussbemerkung zur Entwicklung und dabei aufgetretener Probleme.

3.5.4.1 Installation und Konfiguration

Die im Kapitel 3.5.2 vorgestellten Werkzeuge, müssen für die prototypische Umsetzung jeweils auf unterschiedliche Art und Weise konfiguriert und eingerichtet werden. Im folgenden Kapitel wird die Konfiguration der Werkzeuge erläutert.

VPN-Client

Um auf die bereitgestellte Oracle Datenbank sowie das Repository des Informatica PowerCenters zugreifen zu können, muss zunächst eine VPN-Verbindung zum Netzwerk des Praxispartners hergestellt werden. Hierfür muss lokal auf dem Client-PC ein VPN-Client installiert werden, der vom Praxispartner bereitgestellt wurde. Anschließend kann über die zugesandten Logindaten eine Verbindung hergestellt werden.

Datenbankverbindung

Nachdem die Verbindung zum Firmennetzwerk steht, kann mit Hilfe des SQL Developers eine Verbindung zur Oracle Datenbank hergestellt werden. Die Verbindungsdaten hierfür wurden der Projektgruppe ebenfalls vom Praxispartner mitgeteilt. Die Oracle Datenbank besteht aus zwei Instanzen. Mithilfe des Oracle TNS-Protokolls (Transparent Network Substrate) kann auf einfache Art und Weise eine Verbindung zu diesen Instanzen hergestellt werden. Um eine solche Verbindung via TNS nutzen zu können, muss der Datenbank Service in der „tnsnames.ora“-Datei im Installationsordner des Oracle SQL Developers eingetragen werden, anschließend kann hierüber die TNS-Verbindung ausgewählt und eine Verbindung zur Datenbank hergestellt werden.

Repository des Informatica PowerCenter

Neben der Verbindung zur Datenbank, muss auch eine Verbindung zum Repository des Informatica PowerCenter eingerichtet werden. Auch hierfür wurden der Projektgruppe vom Praxispartner die jeweiligen Zugangsdaten mitgeteilt. Die gesamte Einrichtung des Repositorys erfolgt über

den Informatica PowerCenter Repository Manager, welcher Bestandteil der PowerCenter Installation ist. In diesem Repository Manager muss zunächst eine Domain eingetragen werden. Dies geschieht über den Menüpunkt Repository und dort unter „Configure Domains“. Hier kann nun über „Add a new domain“ unter Angabe eines Namens sowie dem Gateway-Host und -Port die Domain angelegt werden. Anschließend wird die eigentliche Verbindung zum Repository konfiguriert. Hierfür gibt es im Menü Repository den Punkt „Add Repository“, wo durch Angabe des Repository- sowie des Benutzernamens das Repository dem PowerCenter hinzugefügt wird. Nun kann der Nutzer unter Angabe seines Users sowie des Passworts auf das Repository und die darin enthaltenen Objekte zugreifen.

Zuätzlich zur Einrichtung des Repositorys, muss im PowerCenter auch noch an zwei Stellen eine Verbindung zur Oracle Datenbank definiert werden. Dies geschieht einmal im PowerCenter Designer in der Ansicht des Source Analyzer über das Menü „Sources“ und „Import from Database“. In dieser Maske kann über einen Klick auf den „...“-Button eine neue Verbindung angelegt werden. Hierzu muss im Reiter „Benutzer-DSN“ zunächst der Oracle Treiber ausgewählt und anschließend die bekannten Verbindungsdaten zur Datenbank angegeben werden. Nun kann im Designer auf die Datenbank zugegriffen werden. Als zweites muss die Verbindung im PowerCenter Workflow Manager hinterlegt werden. Hierzu klickt man in der Menüleiste des Workflow Managers auf „Connections“ und „Relational“. In dem neuen Fenster können über „New“ zunächst der Oracle-Treiber ausgewählt und anschließend wieder die Verbindungsdaten angegeben werden. Nun ist die Einrichtung der Informatica PowerCenter Komponenten abgeschlossen.

BIRT

Zur Erstellung der Reports muss in Eclipse BIRT eine Verbindung zur Oracle Datenbank eingerichtet werden. Hierfür wird zunächst ein Oracle JDBC Datenbanktreiber benötigt. Dieser Treiber kann kostenlos auf der Oracle-Homepage heruntergeladen und anschließend dem Eclipse Pluginordner hinzufgefügt werden. Nun muss in BIRT zunächst ein neues Projekt (File - New - Project) und anschließend in dem Projekt ein leerer Bericht (File - New - Report) angelegt werden. Nun kann der Treiber im Menü zum Erstellen einer neuen Data Source (Ansicht „Data Explorer“, Rechtsklick auf „Data Sources“, „New Data Source“) ausgewählt und die bekannten Verbindungsdaten der Datenbank eingetragen werden. Nun ist ein Zugriff auf die Datenbank und deren Tabellen sowie Views direkt aus BIRT heraus möglich.

3.5.4.2 Beschreibung des Prototyps

Dieser Abschnitt gibt einen Überblick über die einzelnen Implementierungsschritte, die zur Umsetzung des Prototyps nötig sind. Hierzu wird zunächst beschrieben, wie sowohl die Access Logs als auch die Stamm- bzw. Nutzerkontextdaten in den Operational Data Store (ODS) überführt werden. Darauf erfolgt die Implementierung des ETL-Prozesses, durch den die Daten aus dem ODS über eine Staging Area in das Zielschema aus Kapitel 3.5.3.4 geladen werden. Bis auf das Reporting (Vgl. Kap. 3.5.4.2.5) beziehen sich alle im Folgenden dokumentierten

Schritte auf das Informatica PowerCenter.

3.5.4.2.1 Integration der Access Logs

Um die Daten aus einer Access-Log-Datei im Operational Data Store (ODS) speichern zu können, müssen diese zunächst in ein relationales Schema überführt werden. Dazu ist im Source Analyzer des Informatica PowerCenters eine neue Quelltabelle vom Typ „FlatFile“ zu erstellen. Die dazugehörenden Attribute und Datentypen können der Grafik in Abbildung 3.59 entnommen werden.

access_log_elsa_20130512 (Flat File)

K.	Name	Datatype	Length/Precision
	src_clientip	string	15
	src_serverip	string	15
	src_ivuser	string	2048
	FIELD4	string	1
	FIELD5	string	1
	src_entrydate	string	50
	src_zone	string	5
	src_request	string	2048
	src_statuscode	string	3
	src_datasize	string	64
	src_referer	string	2048
	src_browser	string	2048
	src_responsetime	string	64

Abbildung 3.59: Quelltabelle zur relationalen Speicherung von Access Logs

Als nächstes ist im Target Designer eine neue Zieltabelle vom Datenbanktyp Oracle zu definieren. Das dafür benötigte Schema kann der folgenden Abbildung entnommen werden. Aus dieser geht auch hervor, dass das Feld ELSALOG_SID als Primärschlüssel zu definieren ist.

ODS_T_ELSALOG (Oracle)

Key Types	Name	Datatype	Length/Precision
PRIMARY KEY	ELSALOG_SID	number	15
NOT A KEY	ELSALOG_clientip	varchar	15
NOT A KEY	ELSALOG_serverip	varchar	15
NOT A KEY	ELSALOG_ivuser	varchar	2048
NOT A KEY	ELSALOG_FIELD4	varchar	1
NOT A KEY	ELSALOG_FIELD5	varchar	1
NOT A KEY	ELSALOG_entrydate	varchar	50
NOT A KEY	ELSALOG_zone	varchar	5
NOT A KEY	ELSALOG_request	varchar	2048
NOT A KEY	ELSALOG_statuscode	varchar	3
NOT A KEY	ELSALOG_datasize	varchar	64
NOT A KEY	ELSALOG_referer	varchar	2048
NOT A KEY	ELSALOG_browser	varchar	2048
NOT A KEY	ELSALOG_responseti...	varchar	64

Abbildung 3.60: Zieltabelle zur Speicherung von Access Logs im ODS

Danach muss die Zieltabelle über die im Zuge der Installation und Konfiguration angelegte

ODBC Verbindung in der Oracle Datenbank angelegt und über ein Mapping mit Daten befüllt werden. Dieses Mapping ist über den Mapping Designer anzulegen und muss neben den gerade beschriebenen Quell- und Zieltabellen einen „Sequence Generator“ zur Generierung der Primärschlüssel enthalten (Vgl. Abb. 3.61).

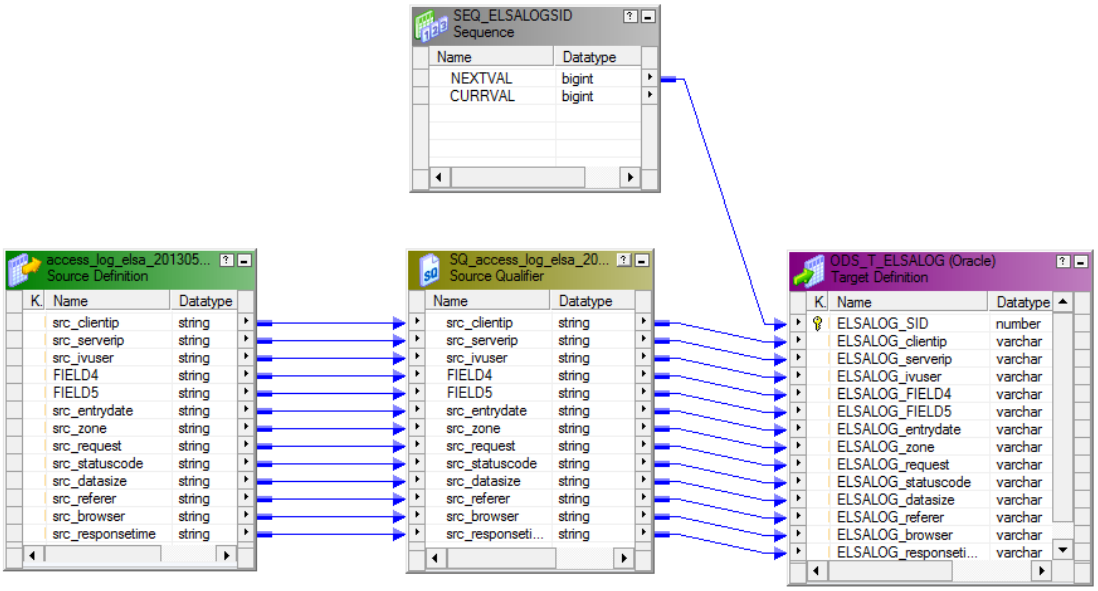


Abbildung 3.61: Mapping zur Speicherung von Access Logs im ODS

Nach der Validierung und Speicherung des Mappings, muss dieses über einen Workflow im Workflow Manager ausgeführt werden. Bei der Erstellung des Workflows muss dann auch das Quellverzeichnis und der Name der Access-Log-Datei angegeben werden, aus dem die Daten extrahiert werden sollen. Die entsprechende Konfiguration des Prototyps ist der Abbildung 3.62 zu entnehmen.

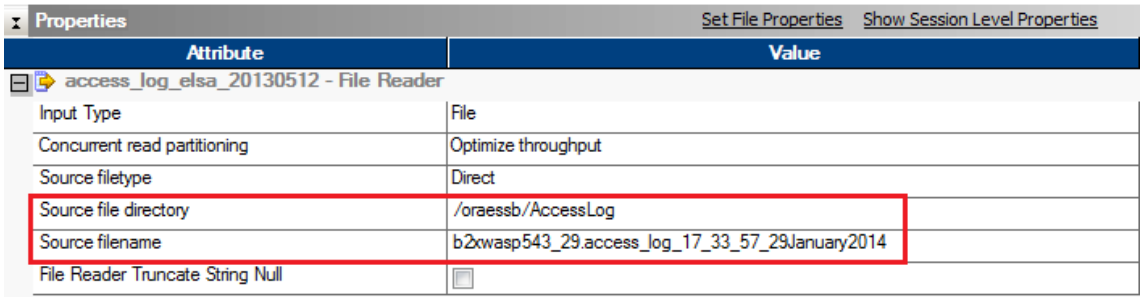


Abbildung 3.62: Konfiguration des Workflows zur Speicherung von Access Logs im ODS

Anschließend ist der Workflow auf der ersten Server Instanz des Informatica PowerCenters (bilab_IS1A) auszuführen. Ist der Workflow erfolgreich durchgelaufen, befinden sich alle Einträge aus der Access-Log-Datei im ODS der Oracle Datenbank.

3.5.4.2.2 Integration der Stamm- u. Nutzerkontextdaten

Nachdem die Access Logs in den Operational Data Store überführt wurden, müssen diese mittels Nutzerkontextdaten und statischen Zuordnungstabellen (Stammdaten) aufbereitet werden. Da die Nutzerkontextdaten - genau wie die Access Logs - als CSV-Datei vorliegen, können diese in analoger Weise zu dem im vorherigen Abschnitt beschriebenen Vorgehen in den ODS überführt werden. Was die Zuordnungstabellen der Stammdaten betrifft, so liegen diese lediglich in textueller Form vor. Aus diesem Grund müssen diese gemäß dem Lastenheft des Praxispartners händisch in die Oracle Datenbank überführt und mit Daten befüllt werden. Wie bereits aus dem Kapitel 3.5.3 hervorgeht, muss dabei sowohl für die Use Cases als auch für die aufrufenden Anwendungen jeweils eine Zuordnungstabelle erstellt werden.

Als nächstes müssen die Tabellen für die Stamm- und Nutzerkontextdaten sowie die Tabelle für die Access Logs als Quelltabellen in das Informatica PowerCenter importiert werden. Um diese Tabellen anschließend gemäß dem in Kapitel 3.5.3.2 vorgestellten Quellschema miteinander verknüpfen zu können, sind drei Zieltabellen im Target Designer anzulegen. Durch eine Referenzierung der entsprechenden Primärschlüsselfelder wird hierbei auch die eigentliche Zuordnung zwischen den Access Logs und den Stamm- bzw. Nutzerkontextdaten ermöglicht (Vgl. Abb. 3.63).

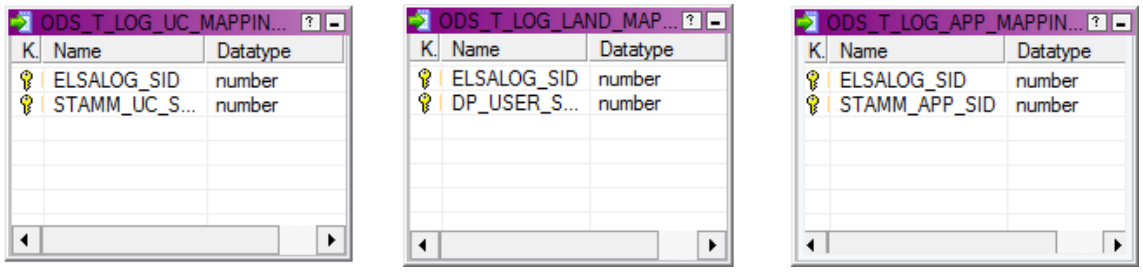


Abbildung 3.63: Mapping-Tabellen für die Stamm- und Nutzerkontextdaten

Nachdem die Zieltabellen erfolgreich definiert und über die ODBC-Schnittstelle im ODS der Oracle Datenbank angelegt wurden, sind diese anschließend mit Daten zu befüllen. Dazu muss für jede Zieltabelle in der oben stehenden Abbildung 3.63 ein Mapping erstellt werden. In jedem Mapping muss dann die Quelltabelle für die Access Logs mit der jeweiligen Quelltabelle für die Stamm- bzw. Nutzerkontextdaten über einen „Source Qualifier“ verknüpft werden. Um dies zu verdeutlichen, wird in Abbildung 3.64 ein beispielhaftes Mapping gezeigt.

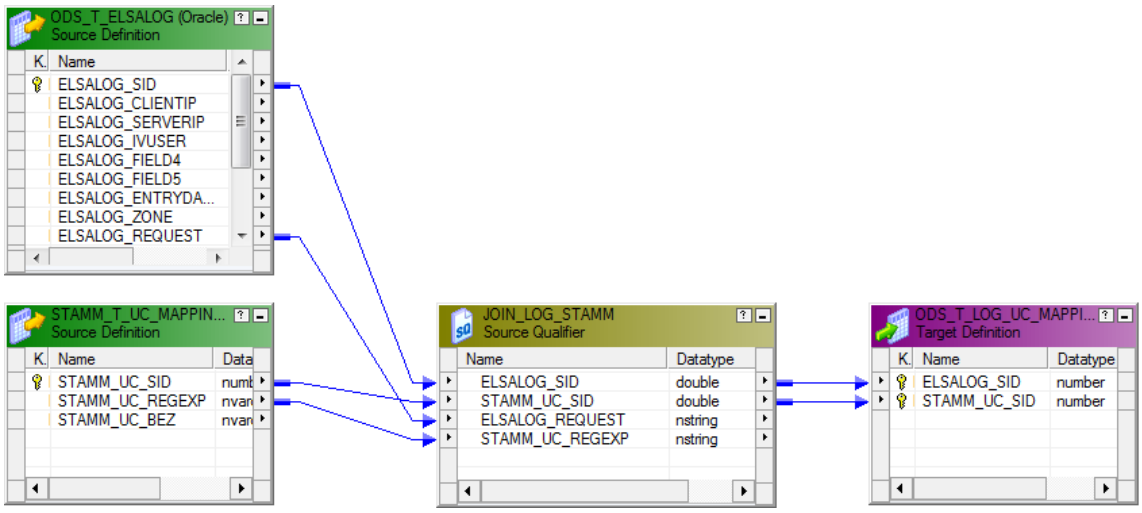


Abbildung 3.64: Mapping zur Datenanreicherung der Access Logs

Damit die Daten aus den Quelltabellen korrekt Verknüpft werden, ist für jedes Mapping im jeweiligen „Source Qualifier“ außerdem eine SQL-Abfrage zur Durchführung des Joins zu formulieren. Für die Zuordnung der Use Cases muss dazu ein direkter Abgleich zwischen dem Access-Log-Feld ELSALOG_REQUEST und dem entsprechenden regulären Ausdruck aus der Zuordnungstabelle vollzogen werden (Vgl. Listing 3.1).

```
1 SELECT ODS_T_ELSALOG.ELSALOG_SID
2 , STAMM_T_UC_MAPPING.STAMM_UC_SID
3 FROM ODS_T_ELSALOG
4 JOIN STAMM_T_UC_MAPPING
5 ON REGEXP_LIKE(ODS_T_ELSALOG.ELSALOG_REQUEST
6 , STAMM_T_UC_MAPPING.STAMM_UC_REGEXP);
```

Listing 3.1: Verknüpfung zwischen Access Log und Use Case

Die Zuordnung zwischen einem Access Log und einer aufrufenden Anwendung erfolgt in analoger Weise (Vgl. Listing 3.2).

```
1 SELECT ODS_T_ELSALOG.ELSALOG_SID
2 , STAMM_T_APP_MAPPING.STAMM_APP_SID
3 FROM ODS_T_ELSALOG
4 LEFT OUTER JOIN STAMM_T_APP_MAPPING
5 ON REGEXP_LIKE(ODS_T_ELSALOG.ELSALOG_REQUEST
6 , STAMM_T_APP_MAPPING.STAMM_APP_REGEXP);
```

Listing 3.2: Verknüpfung zwischen Access Log und aufrufender Anwendung

Um letztlich auch eine Zuordnung zu den Nutzerkontextdaten herstellen zu können, wird folgender SQL-Code benötigt.

```
1 SELECT ODS_T_ELSALOG.ELSALOG_SID
2 , ODS_T_DP_USER.DP_USER_SID
3 FROM ODS_T_DP_USER
```

```
4 INNER JOIN ODS_T_ELSALOG
5 ON (ODS_T_DP_USER.DP_USER_UID = ODS_T_ELSALOG.ELSALOG_IVUSER);
```

Listing 3.3: Verknüpfung zwischen Access Log und Nutzerkontextdaten

Anschließend müssen die Mappings gespeichert und über den Workflow Manager zur Ausführung gebracht werden. Danach kann jeder Access Log einem Use Case, einer aufrufenden Anwendung und einem User aus den Nutzerkontextdaten zugeordnet werden.

3.5.4.2.3 Datenextraktion

Bevor die zuvor integrierten Daten in das Zielschema überführt werden können, müssen diese in einer Staging Area zusammengeführt werden. Zu diesem Zweck müssen alle zum Quellschema gehörenden Tabellen (Vgl. Kap. 3.5.3.2) im Informatica PowerCenter als Quelltabelle vorliegen. Anschließend ist eine neue Zieltabelle anzulegen, die die Staging Area repräsentiert. Das genaue Schema dieser Zieltabelle kann der Abbildung 3.65 entnommen werden.

STAGING_T_MAPPING (Oracle)

K.	Name	Datatype	Length...
	ELSALOG_SID	number	15
	ELSALOG_ENTRYD...	timestamp	26
	ELSALOG_clientip	varchar2	15
	ELSALOG_serverip	varchar2	15
	ELSALOG_ivuser	varchar2	2048
	ELSALOG_zone	varchar2	5
	ELSALOG_request	varchar2	2048
	ELSALOG_statuscode	varchar2	3
	ELSALOG_datasize	number	15
	ELSALOG_referer	varchar2	2048
	ELSALOG_browser	varchar2	2048
	ELSALOG_responset...	varchar2	64
	STAMM_APP_SID	number	15
	STAMM_APP_BEZ	nvarchar2	128
	STAMM_UC_SID	number	15
	STAMM_UC_BEZ	nvarchar2	128
	DP_USER_SID	number	15
	DP_USER_LAND	varchar2	2048
	DP_USER_LANDISO3	varchar2	3

Abbildung 3.65: Zieltabelle für die Staging Area

Auch diese Zieltabelle ist wieder über die ODBC-Schnittstelle in der Oracle Datenbank anzulegen. Anschließend ist ein neues Mapping zu erstellen, das die Staging Area mit Daten befüllt. Dazu müssen die (Quell-) Tabellen aus dem Quellschema über die in Listing 3.4 angegebene SQL-Abfrage in einem „Source Qualifier“ vereint werden.

```
1 SELECT A.ELSALOG_SID
2         ,A.ELSALOG_ENTRYDATE
3         ,A.ELSALOG_CLIENTIP
4         ,A.ELSALOG_SERVERIP
5         ,A.ELSALOG_IVUSER
6         ,A.ELSALOG_ZONE
```

```
7          ,A.ELSALOG_REQUEST
8          ,A.ELSALOG_STATUSCODE
9          ,A.ELSALOG_DATASIZE
10         ,A.ELSALOG_REFERER
11         ,A.ELSALOG_BROWSER
12         ,A.ELSALOG_RESPONSETIME
13         ,STAGING_APP.APP_SID
14         ,STAGING_APP.APP_BEZ
15         ,STAGING_UC.UC_SID
16         ,STAGING_UC.UC_BEZ
17         ,STAGING_DP.DP_SID
18         ,STAGING_DP.DP_LAND
19         ,STAGING_DP.DP_LANDISO3
20 FROM ODS_T_ELSALOG A
21 LEFT OUTER JOIN (
22         SELECT ODS_T_LOG_APP_MAPPING.ELSALOG_SID AS ELSALOG_APP_SID
23                 ,ODS_T_LOG_APP_MAPPING.STAMM_APP_SID AS APP_SID
24                 ,STAMM_T_APP_MAPPING.STAMM_APP_BEZ AS APP_BEZ
25         FROM ODS_T_LOG_APP_MAPPING
26         INNER JOIN STAMM_T_APP_MAPPING ON (STAMM_T_APP_MAPPING.
               STAMM_APP_SID = ODS_T_LOG_APP_MAPPING.STAMM_APP_SID)
27         ) STAGING_APP ON (A.ELSALOG_SID = STAGING_APP.ELSALOG_APP_SID)
28 LEFT OUTER JOIN (
29         SELECT ODS_T_LOG_UC_MAPPING.ELSALOG_SID AS ELSALOG_UC_SID
30                 ,ODS_T_LOG_UC_MAPPING.STAMM_UC_SID AS UC_SID
31                 ,STAMM_T_UC_MAPPING.STAMM_UC_BEZ AS UC_BEZ
32         FROM ODS_T_LOG_UC_MAPPING
33         INNER JOIN STAMM_T_UC_MAPPING ON (STAMM_T_UC_MAPPING.STAMM_UC_SID =
                ODS_T_LOG_UC_MAPPING.STAMM_UC_SID)
34         ) STAGING_UC ON (STAGING_UC.ELSALOG_UC_SID = A.ELSALOG_SID)
35 LEFT OUTER JOIN (
36         SELECT ODS_T_LOG_LAND_MAPPING.ELSALOG_SID ELSALOG_DP_SID
37                 ,ODS_T_LOG_LAND_MAPPING.DP_USER_SID AS DP_SID
38                 ,ODS_T_DP_USER.DP_USER_LAND AS DP_LAND
39                 ,ODS_T_DP_USER.DP_USER_LANDISO3 AS DP_LANDISO3
40         FROM ODS_T_LOG_LAND_MAPPING
41         INNER JOIN ODS_T_DP_USER ON (ODS_T_LOG_LAND_MAPPING.DP_USER_SID =
               ODS_T_DP_USER.DP_USER_SID)
42         ) STAGING_DP ON (STAGING_DP.ELSALOG_DP_SID = A.ELSALOG_SID);
```

Listing 3.4: SQL-Statement zur Befüllung der Staging Area

Nachdem dann die einzelnen Felder des „Source Qualifier“ mit den der Staging Area verbunden wurden, muss das Mapping gespeichert und über einen Workflow ausgeführt werden. Danach können die jeweils zu einem Access-Log-Eintrag zugeordneten Stamm- und Nutzerkontextdaten direkt in der Staging Area eingesehen werden, ohne dass weitere Joins notwendig sind. Auf Basis dieser Daten kann das Zielschema wie im nächsten Abschnitt beschrieben befüllt werden.

3.5.4.2.4 DWH-Integration

Da die im vorherigen Abschnitt erstellte Staging Area als Datenbasis für das im Zuge der Konzeption definierte Zielschema gilt, muss diese zunächst als Quelltabelle aus der Oracle Datenbank in das Informatica PowerCenter importiert werden. Anschließend müssen die im Zielschema enthaltenen Dimensionstabellen als Zieltabellen gemäß der Abbildung 3.66 angelegt werden.

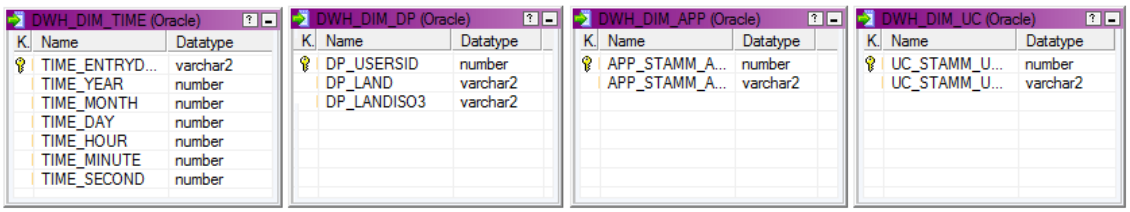


Abbildung 3.66: Zieltabellen für die Dimensionen des Zielschemas

Die für die Dimensionstabellen benötigten Mappings lassen sich jeweils nach dem ungefähr gleichen Prinzip aufbauen. Dabei werden die für die jeweilige Dimension relevanten Felder zunächst über einen Aggregator gruppiert und anschließend in die entsprechende Zieltabelle überführt. Zur Verdeutlichung dieses Vorgehens wird in Abbildung 3.67 ein beispielhaftes Mapping gezeigt.

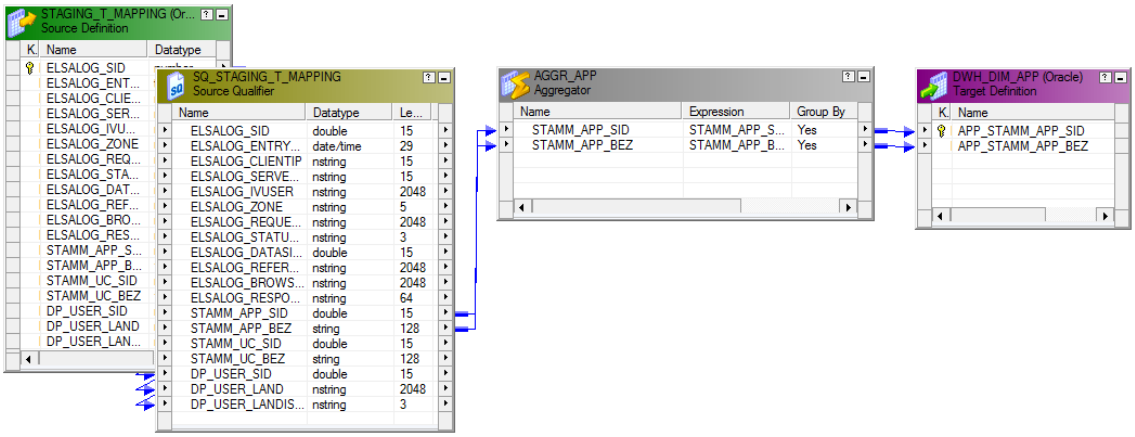


Abbildung 3.67: Mapping zur Datenbefüllung einer Dimensionstabelle

Wie in der oben gezeigten Abbildung 3.67 zu sehen ist, werden für das Mapping der Dimension DWH_DIM_APP die Felder STAMM_APP_SID und STAMM_APP_BEZ aus der Staging Area benötigt. Die für die anderen Dimensionen benötigten Felder sind der folgenden Auflistung zu entnehmen.

- **DWH_DIM_UC**: STAMM_UC_SID, STAMM_UC_BEZ
- **DWH_DIM_DP**: DP_USER_SID, DP_USER_LAND
- **DWH_DIM_TIME**: ELSALOG_ENTRYDATE

Das Mapping für die Dimension DWH_DIM_UC muss außerdem noch um eine Expression erweitertet werden, in der mithilfe der Funktion GET_DATE_PART die einzelnen Aggregationsstufen (Jahr, Monat, Tag etc.) aus dem Feld ELSALOG_ENTRYDATE extrahiert werden. Nachdem die Mappings für die Dimensionen über den Workflow Manager ausgeführt wurden, muss die ebenfalls aus dem Zielschema hervorgehende Faktentabelle als Zieltabelle in das Informatica PowerCenter überführt werden. Die dafür benötigten Datentypen und Feldbezeichnungen können der folgenden Abbildung 3.68 entnommen werden.

DWH_FACTS_ELSALOG (Oracle)

K.	Name	Datatype	Len...
	TIME_ID	varchar2	50
	UC_ID	number	15
	APP_ID	number	15
	DPUSER_ID	number	15
	FACT_ACCESS_BY_UC	number	15
	FACT_ACCESS_BY_APP	number	15
	FACT_ACCESS_BY_COUNTRY	number	15

Abbildung 3.68: Faktentabelle zur Auswertung der Access Logs

Abschließend ist die Faktentabelle über ein Mapping mit Daten zu befüllen. Zur Ermittlung der in der Faktentabelle enthaltenen Kennzahlen, wird dabei auch ein Aggregator benötigt, der die Anzahl der Use Cases, aufrufenden Anwendungen und User aus den Nutzerkontextdaten mittels COUNT ermittelt und diese nach den über die Fremdschlüssel referenzierten Dimensionen gruppiert (Vgl. Abb. 3.69).

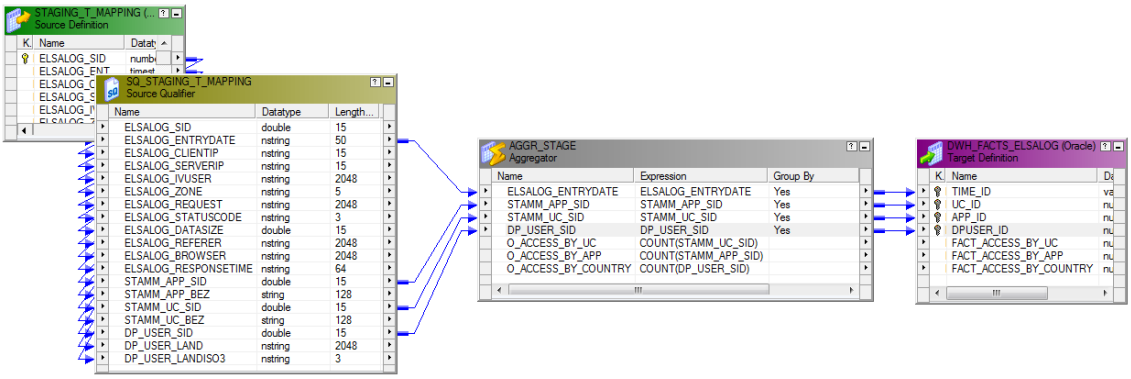


Abbildung 3.69: Mapping zur Datenbefüllung der Faktentabelle

Nachdem letztlich auch das Mapping für die Faktentabelle über einen Workflow ausgeführt wurde, können die Access Logs über das Reporting ausgewertet werden.

3.5.4.2.5 Reporting

In dieser prototypischen Umsetzung sollen vier unterschiedliche Berichte beispielhaft implementiert werden. Diese Berichte sind die Anzahl der Zugriffe je Use Case, die Zugriffe je Stunde,

die Zugriffe je Stunde und Use Case sowie die Zugriffe je Land. Zur Umsetzung dieser Reports wurden zunächst in der Oracle Datenbank zwei Views erstellt. Diese dienen anschließend als Datenbasis für die Berichte. Die Views wurden jeweils mit einem SQL-Befehl angelegt.
Die View für den Report Zugriffe je Land beinhaltet die Felder Land sowie die aggregierten Zugriffe. Der SQL-Code hierfür ist in Listing 3.5 zu finden.

```
1 -- Erstellen der View
2 CREATE VIEW V_Facts_Zugriffe_Land AS
3 -- Auswahl der Attribute
4 SELECT b.dp_land AS Land, SUM(a.fact_access_by_uc) AS Zugriffe
5 FROM dwh_facts_elsalog a
6 -- Join in eine Dimensionstabelle
7 JOIN dwh_dim_dp b ON a.dpuser_id = b.dp_usersid
8 -- Aggregation der Zugriffe
9 GROUP BY b.dp_land;
```

Listing 3.5: View für Zugriffe je Land

Die Berichte Zugriffe je Use Case, Zugriffe je Stunde sowie Zugriffe je Stunde und Use Case basieren auf einer View mit den Attributen Stunde, Use Case und den aggregierten Zugriffen. Der SQL-Code zum Erzeugen dieser View ist in Listing 3.6 zu finden.

```
1 -- Erstellen der View
2 CREATE VIEW v_facts_zugriffe_stunde_uc AS
3 -- Auswahl der Attribute
4 SELECT c.time_hour AS Stunde, b.uc_stamm_uc_bez AS UseCase,
5 SUM(a.fact_access_by_uc) AS Zugriffe
6 FROM dwh_facts_elsalog a
7 -- Zweifacher Join in die Dimensionstabellen
8 JOIN dwh_dim_uc b ON a.uc_id = b.uc_stamm_uc_sid
9 JOIN dwh_dim_time c ON a.time_id = c.time_entrydate
10 -- Aggregation der Zugriffe
11 GROUP BY c.time_hour, b.uc_stamm_uc_bez;
```

Listing 3.6: View für Zugriffe je Stunde und Use Case

Mit diesen beiden Views ist die Arbeit in der Oracle Datenbank erledigt und die Datenbasis bereitgestellt. Dadurch, dass die Reports nicht direkt auf die Daten im Data Warehouse, sondern in den Views zugreifen, ist eine Applikationsschicht entstanden.

Als nächstes muss BIRT aufgerufen werden, um dort die Berichte zu erstellen. Hierfür bietet BIRT eine grafische Oberfläche, welche das Erstellen von Reports sehr angenehm gestaltet. Voraussetzung für diesen Schritt ist die Definition der Verbindung zur Oracle Datenbank (Vgl. Kap. 3.5.4.1). Ist dies erfolgt, kann in dem bereits angelegten, leeren Bericht ein neues Data Set definiert werden (Ansicht „Data Explorer“, Rechtsklick auf „Data Sets“, „New Data Set“). In diesem Fenster wird die angelegte Data Source und als Typ „SQL Select Query“ ausgewählt. Im nächsten Schritt wird ein einfaches SQL „Select all“ Statement auf die jeweilige View als Query Text angegeben. Im Anschluss wird ein weiteres Data Set für die zweite View nach

demselben Prinzip angelegt. Nun kann im grafischen Editor über die Palette ein Chart auf das leere Berichtsblatt gezogen werden, wodurch sich ein neues Fenster öffnet. In diesem Fenster muss zunächst der Chart Type ausgewählt werden. Am Beispiel des Berichts mit den Zugriffen je Use Case und Stunde, wird an dieser Stelle ein Liniendiagramm ausgewählt. Im nächsten Reiter müssen nun die Datenfelder definiert werden (Vgl. Abb. 3.70). Hierzu wählt man im Dropdown-Menü „Use Data from“ das jeweilige, im vorherigen Schritt angelegte Data Set aus und eine Vorschau der Daten wird sichtbar. Nun können per Drag and Drop die gewünschten Spalten auf die X- und Y-Achsen gezogen werden. Wie in der Abbildung 3.70 zu sehen ist, sind im Bericht zu den Zugriffen je Use Case und Stunde auf den Y-Achsen die Zugriffe sowie die Use Cases und auf der X-Achse die Zeitleiste mit den Stunden angeordnet.

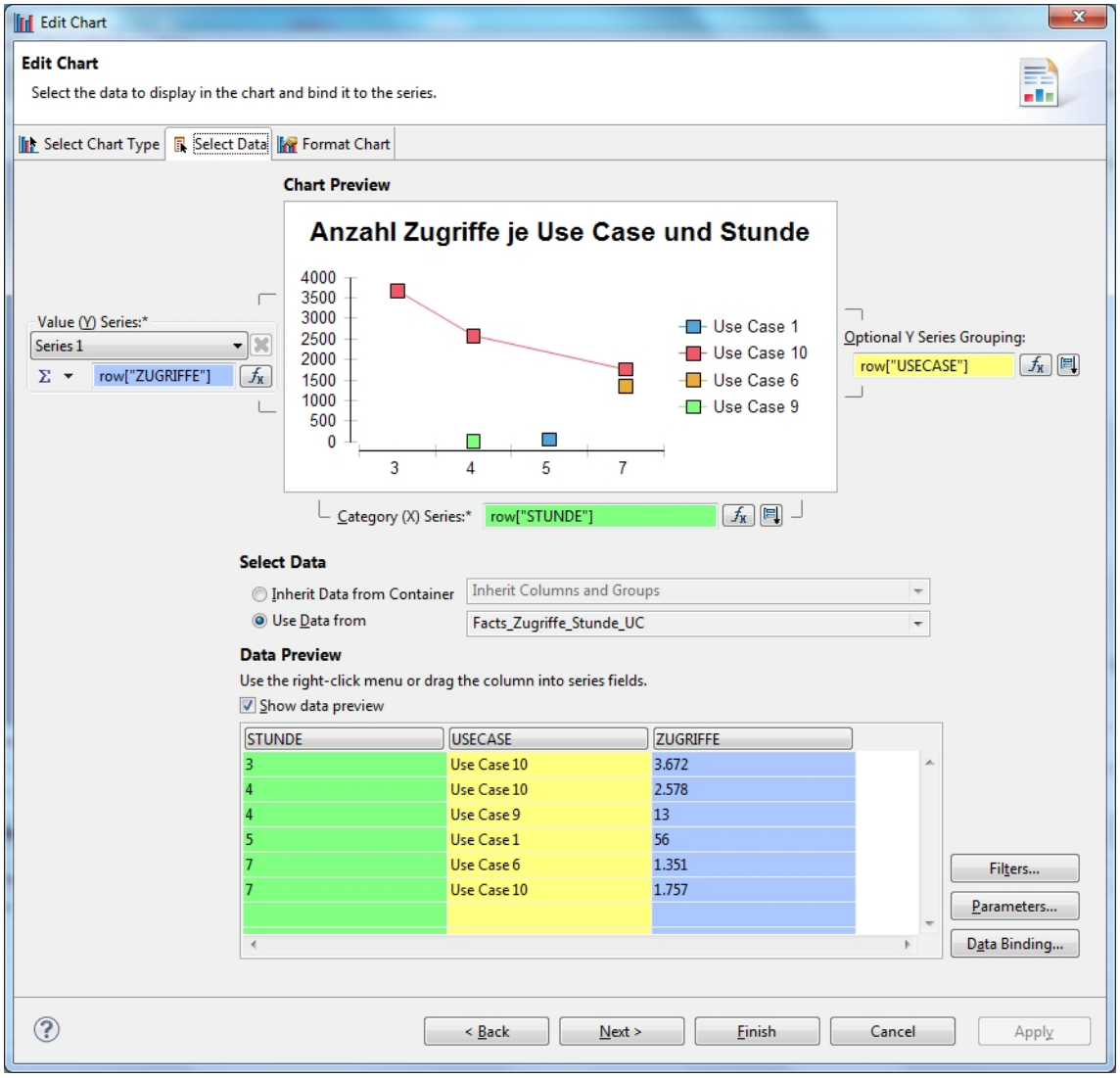


Abbildung 3.70: Auswahl der Datenfelder

Im Reiter „Format Chart“ kann anschließend eine Reihe von Formatierungen zu dem Diagramm

definiert werden. Hier kann beispielsweise eine Beschriftung der X- und Y-Achse sowie ein Titel angegeben werden. Sobald die Formatierung des Diagramms abgeschlossen ist, wird das Fenster über „Finish“ geschlossen. Als Ergebnis dessen wird das Diagramm auf dem Arbeitsblatt des Reports angezeigt. Dieser kann nun über „View Report“, „View Report as HTML“ mit Hilfe des BIRT Report Viewers in einer HTML Vorschau betrachtet werden (Vgl. Abb. 3.71).

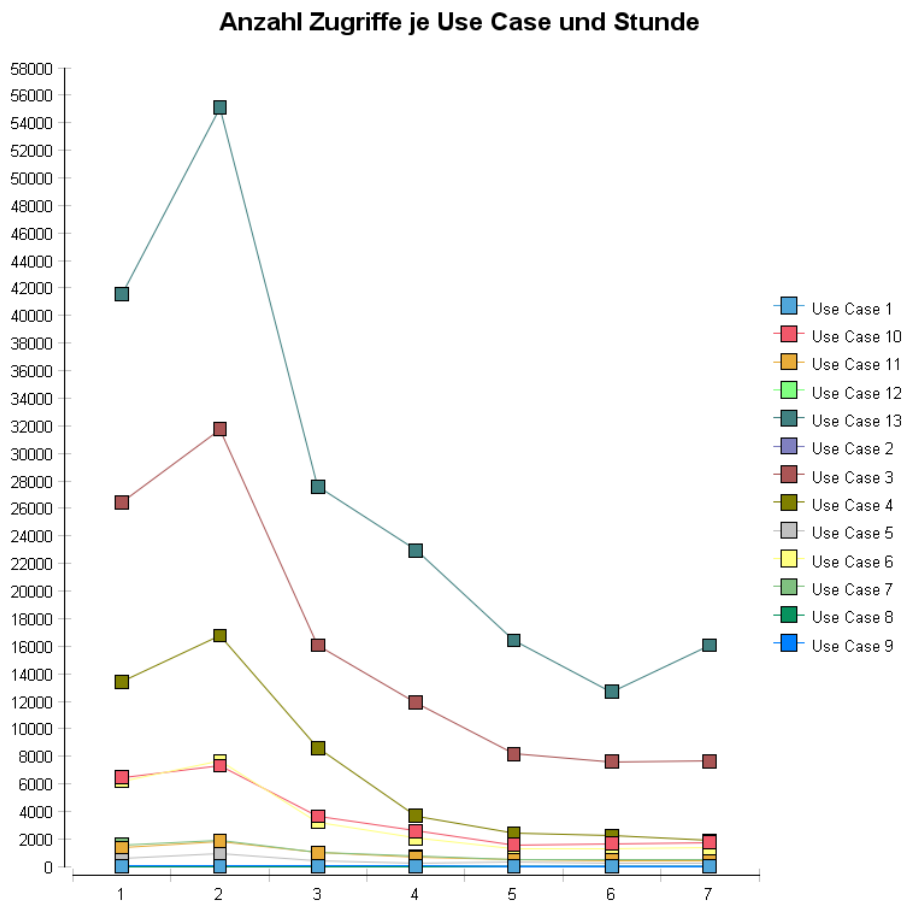


Abbildung 3.71: Report Zugriffe je Stunde und Use Case

Nun ist der erste Bericht fertig. Die Umsetzung der anderen drei Berichte erfolgt auf dieselbe Art und Weise. Lediglich die Diagrammtypen und die Datenfelder unterscheiden sich. Bei den Berichten zu den Zugriffen je Use Case sowie den Zugriffen je Stunde wird jeweils ein Balkendiagramm gewählt, das Data Set bleibt hierbei dasselbe wie im vorherigen Report (Vgl. Abb. 3.72 und 3.73).

Der Bericht zu den Zugriffen je Land basiert im Gegensatz zu den anderen auf einem Kreisdiagramm. Außerdem wird hierbei das zweite Data Set, welches auf die View mit dem Land verweist, verwendet. Des Weiteren wurde, aus Gründen der Übersichtlichkeit, ein Filter eingebaut, welcher nur die Top 15 Länder anzeigt. Also die 15 Länder mit den meisten Zugriffen. Dieser Filter wurde in dem Reiter „Select Data“ über den Button „Filters“ definiert (siehe

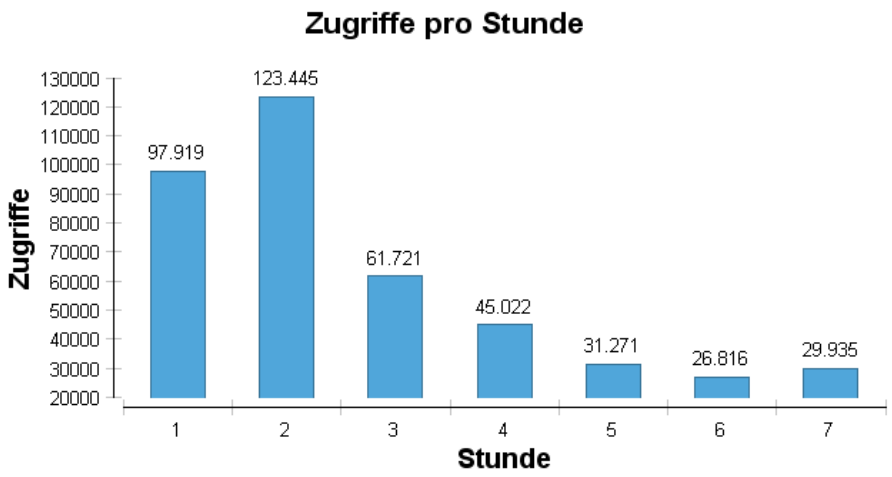


Abbildung 3.72: Report Zugriffe je Stunde

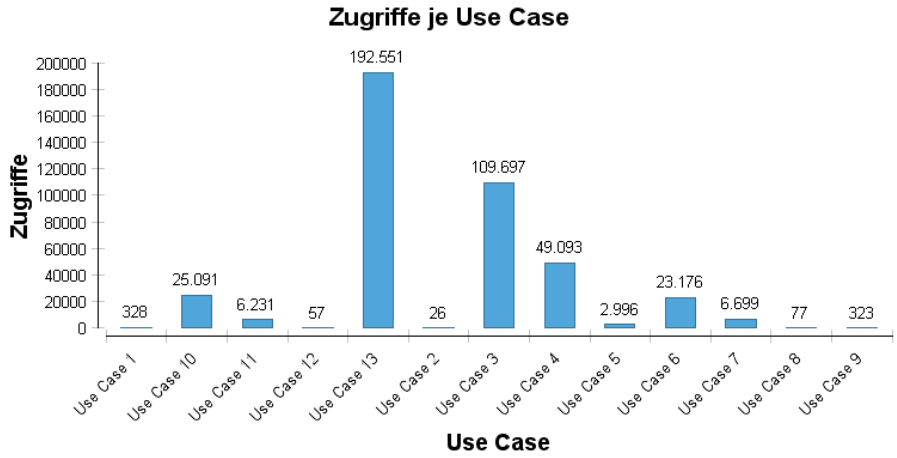


Abbildung 3.73: Report Zugriffe je Use Case

Abbildung 3.74). Der fertige Bericht ist in Abbildung 3.75 zu sehen.

3.5.4.3 Schlussbetrachtung

Der Prototyp zeigt einen möglichen Weg auf, wie mit der bereitgestellten Software der Volkswagen AG Daten aus einem Access Log in ein Data Warehouse überführt und ausgewertet werden können. Es wäre wünschenswert gewesen, einen fortgeschritteneren Entwicklungsstand zu erreichen. Nicht umgesetzt werden konnte schließlich eine Erweiterung des Prototypen um weitere Access-Log-Dateien und damit die Durchführung von Last-Tests. Die größere Ergebnismenge hätte einen breiteren Betrachtungszeitraum für das dargestellte Reporting ermöglicht. Des Weiteren konnte im Rahmen dieser Projektzeit keine Schnittstelle mehr zu der Simulationskomponente des Capacity Managements implementiert werden. Die Umsetzung dieser Erweiterungen hätten eine noch bessere Grundlage zur Bewertung der Machbarkeit

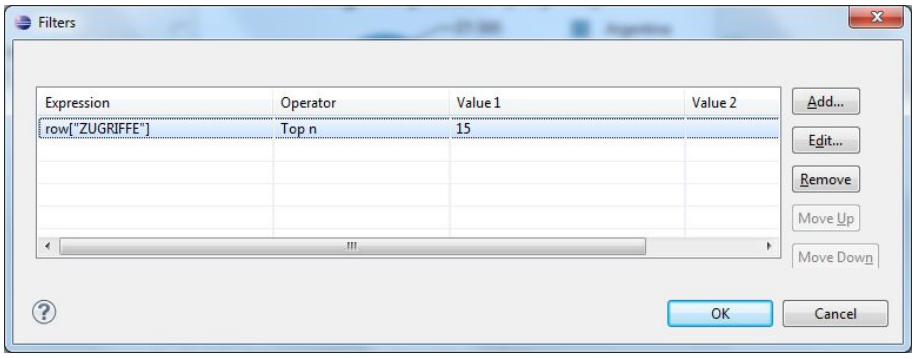


Abbildung 3.74: Definieren eines Filters in BIRT

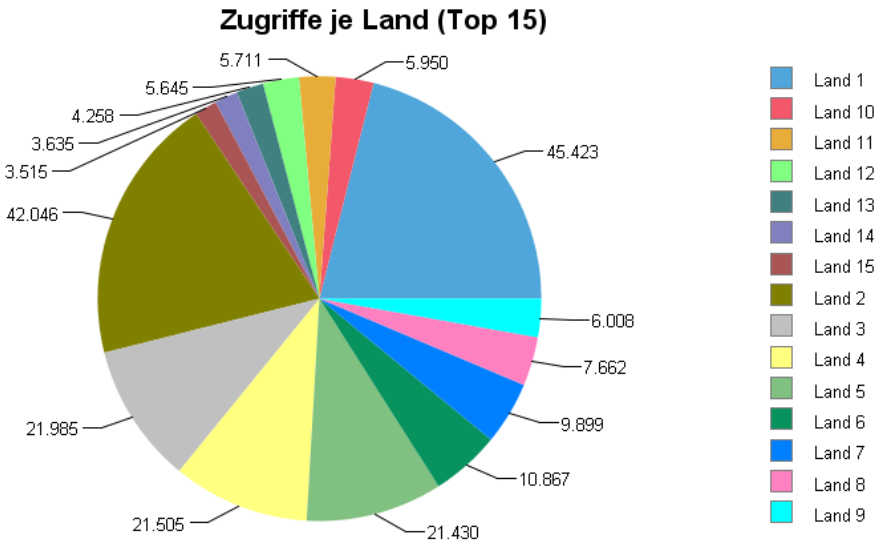


Abbildung 3.75: Report Zugriffe je Land

gegeben. Allerdings zeigt das erreichte Ergebnis, dass auch eine Umsetzung mit konventionellen BI-Systemen die notwendigen Leistungskennzahlen liefern kann.

3.6 Erstellung eines Prototypen zur „Simulation für Capacity Management“

Der folgende Abschnitt beschreibt die Erstellung des Prototypen zur Erfüllung der Anforderungen aus dem Lastenheft „Simulation für Capacity Management“ (Vgl. Volkswagen AG 2013). Nach einleitenden Worten wird zunächst die Grundlagen der verwandten Technologie beschrieben. Es folgt die Darstellung des Konzeptes. Im Anschluss wird die konkrete Umsetzung und das Ergebnis erklärt. Der Abschnitt schließt mit einem Benutzerhandbuch zur Durchführung der Simulation.

3.6.1 Einleitung

Neue geschäftliche Anforderungen führen häufig auch zu einem erhöhten Bedarf an Kapazitäten in der IT-Architektur. Um eine ideale Ressourcennutzung und eine kurze Reaktionszeit auf sich stetig verändernde Anforderungen gewährleisten zu können, ist eine Simulationskomponente im IT-Capacity Management unabdingbar (Vgl. Stych u. Zeppenfeld 2009, S. 106). Der Bedarf einer Planungs- und Vorhersageanwendung für die Ressourcennutzung wird erhöht durch eine komplexe IT-Landschaft und eine Vielzahl von Anwendungen. Durch geschäftsstrategische Anpassungen ergeben sich für einzelne Anwendungen große Änderungen in den Nutzerzahlen. Ebenso führt eine intensive Vernetzung der IT-Dienste zu neuen punktuellen Belastungen in der IT-Architektur.

Die Aufgabenstellung ergibt sich aus dem Lastenheft zur Erstellung einer Simulation für das Capacity Management (Vgl. Volkswagen AG 2013). Ziel ist demnach eine Simulations- und Planungsvariante für das technische Kapazitätsmanagement der IT-Infrastruktur der Volkswagen AG bereitzustellen. Das Ergebnis soll die berechneten Hardwareanforderungen für die Gegenwart und die simulierte Zukunft darstellen. Auf der Basis dieser Simulation sollen sowohl die Entscheidungsfindung als auch die Beschaffung durchgeführt werden. Das endgültige Produkt soll ein Kapazitätskalkulator sein. Diese Anforderungen, die Ergebnisse aus verschiedenen Gesprächen mit den Auftraggebern und die technischen und organisatorischen Rahmenbedingungen sind in ein Pflichtenheft geflossen, das die Auftragsinterpretation darstellt (Vgl. Anhang F.4).

Aus den Anforderungen geht hervor, dass neben der Erfüllung der Funktionalitätsanforderungen Wert auf eine benutzerorientierte und komfortable Lösung gelegt wird. Der Benutzer soll durch die verschiedenen Schritte und Ergebnisse geleitet werden.

Zunächst war die Umsetzung mit der Software Oracle Essbase geplant. In Gesprächen konnte ermittelt werden, dass die Umsetzung der funktionalen und nicht-funktionalen Anforderungen im Fokus steht und keine spezielle Softwareplattform bevorzugt wird. Neben Oracle Essbase ist vor allem IBM Cognos TM1 dazu geeignet, eine Planungsanwendung zu erstellen (Vgl. IBM Corporation 2014). In Absprache mit den Auftraggebern wurde IBM Cognos TM1 als Plattform zur Umsetzung der Simulation gewählt.

Dieses Kapitel beschreibt die Entwicklung des Prototyps zur Durchführung der Simulation im IT-Capacity Management. Dazu wird zunächst das Werkzeug IBM Cognos TM1 beschrieben. Nach der Einführung in die Technologie wird das konzeptionelle Vorgehen erläutert. Die konkrete Umsetzung wird im Anschluss detailliert erläutert. Das Kapitel schließt mit einem Benutzerhandbuch, das den Anwender in die Lage versetzt, ein Simulationsszenario zu erstellen und eine Simulation durchzuführen.

3.6.2 Beschreibung des Werkzeugs IBM Cognos TM1

In diesem Kapitel wird das Werkzeug IBM Cognos TM1 beschrieben, das zur Umsetzung der Simulation des Kapazitätsmanagements verwendet wird. Dabei wird auf den Aufbau, die Architektur, das Sicherheitskonzept und den Datenimport eingegangen. Des Weiteren wird auf die relevanten Handbücher verwiesen.

Die Softwareplattform IBM Cognos TM1 unterstützt Unternehmen bei der Unternehmensplanung, insbesondere bei der Planung, Budgetierung, Prognose und Analyse sowie dem Reporting (Vgl. IBM Corporation 2014). Sie bietet die Möglichkeit zur Anzeige und Bearbeitung großer multidimensionaler Datenmengen sowie Echtzeit-Lese- und Schreibfunktionen. IBM Cognos TM1 arbeitet mit einem 64-Bit In-Memory-OLAP-Server und gilt damit als „Turbo unter den OLAP-Datenbanken“ (Vgl. Braincourt 2014; Selbach Information Systems 2014). Laut IBM hat Cognos TM1 folgende Features (Vgl. IBM Corporation 2014):

- **Leistungsfähige Planung und Analyse:** Erstellung und Analyse von Plänen, Budgets und Prognosen auch bei großen Datasets
- **Integrierte Scorecards und Strategiemanagement:** Modellierung von Kennzahlen, um die Annäherung an die Zielsetzungen zu messen
- **Flexible Modellierung:** Entwicklung und Implementierung von komplexen Planungs- und Analysemodellen
- **Einbindung von mehr Benutzern:** Einbindung von Benutzern in unternehmensweite Planungsvorgänge

Das aktuelle Produkt ist IBM Cognos TM1 10.2, das durch regelmäßige Fix Packs und Erweiterungen stetig aktualisiert wird.

3.6.2.1 Aufbau von TM1

In diesem Abschnitt wird der Aufbau von TM1 erläutert, wobei auf Cubes und Dimensionen, Dimensionselemente sowie die Cube-Regeln eingegangen wird.

3.6.2.1.1 Cubes und Dimensionen

IBM Cognos TM1 speichert Daten in Cubes, wobei jede Zelle eines Cubes eine Kennzahl enthält. In einem Cube sind deshalb Daten für eine oder mehrere Kennzahlen enthalten. Jeder Cube hat mehrere Dimensionen, die die Organisation der Daten darstellen. Beispielsweise sind in einem Cube die Umsatzwerte eines Unternehmens als Kennzahl gespeichert. Durch die Bezeichnung der Dimensionen mit Produkt und Monat kann somit bspw. der Umsatz eines bestimmten Produktes im Monat Januar dargestellt werden. Jeder Cube kann aus zwei bis zu 256 Dimensionen bestehen, wobei sich die Anzahl der Dimensionen meist nach der Dimensionalität der Daten richtet (Vgl. IBM Corporation 2013d, S. 7f.). Abbildung 3.76 veranschaulicht einen Cube mit dazugehörigen Dimensionen.

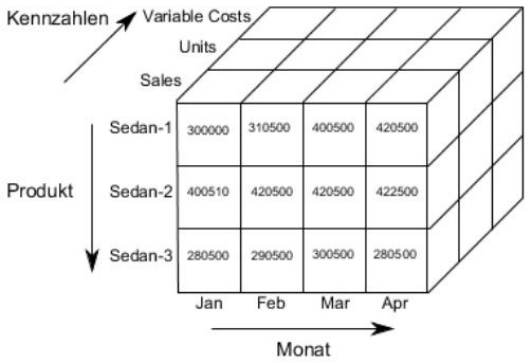


Abbildung 3.76: Dimensionen eines Cubes
Quelle: IBM Corporation (2013d, S.7)

Da bei komplexen Analysen eine Vielzahl von Daten und Kennzahlen auftreten können, besteht die Möglichkeit, die Elemente der Zeilen- und Spaltendimensionen mit sogenannten Subsets einzugrenzen. Die Dimensionensubsets ermöglichen somit beispielsweise eine Eingrenzung in Produktgruppen, Spitzenverkäufer oder Umsatzregionen (Vgl. IBM Corporation 2013b, S. 39).

3.6.2.1.2 Dimensionselemente

Dimensionselemente identifizieren Daten in einem Cube und werden mit Hilfe der Elementattribute beschrieben. Die Software TM1 unterscheidet verschiedene Typen von Elementen: numerische und konsolidierte Elemente sowie Zeichenfolgen. Numerische Elemente identifizieren die niedrigste Detailebene in einer Dimension. Konsolidierte Elemente bilden Aggregationen von niedriger Detailebene. Das Element Zeichenfolge speichert Textzeichenfolgen in Zellen (Vgl. IBM Corporation 2013d, S. 13).

3.6.2.1.3 Cube-Regeln

Regeln bei Cognos TM1 werden auf die individuellen Cubes angewendet und dienen zur Erstellung von Formeln für die Durchführung von Berechnungen. Dabei werden die Berechnungen mithilfe von Konsolidierungshierarchien ausgeführt. Folgende Aufgaben können beispielsweise mit den Cube-Regeln durchgeführt werden (Vgl. IBM Corporation 2013d, S. 45):

- Preis mit Einheiten multiplizieren, um Umsatzbeträge zu berechnen
- Umsatzdaten in einen Cube integrieren, der Gewinn- und Verlustdaten enthält
- die gleichen Werte mehreren Zellen zuweisen

Für die Erstellung von Regeln gibt es verschiedene mathematische Operatoren. Dazu gehören arithmetische Operatoren (z.B. Addition, Multiplikation), Vergleichsoperatoren (z.B. größer als, ungleich) logische Operatoren, um Regelanweisungen zu verbinden (z.B. AND, OR) und viele weitere Regelfunktionen. Außerdem ist die genaue Syntax und das Format zum Erstellen einer Regel festgelegt (Vgl. IBM Corporation 2013d, S. 46 ff.).

3.6.2.2 Architektur

„IBM Cognos TM1 wendet eine verteilte Client-Server-Architektur an, die auf dem IBM Cognos TM1 Server basiert, zu dem eine Kombination unterschiedlicher Clients eine Verbindung herstellen kann“ (IBM Corporation 2013e, S. 25).
Cognos TM1 stellt die folgenden Clients und Benutzerschnittstellen bereit, die im Anschluss näher erläutert werden (Vgl. IBM Corporation 2013e, S. 25; Vgl. IBM Corporation 2013f, S. vii f.):

Endbenutzer-Clients

- Cognos Insight
- Cognos TM1 Application-Portal und -Workflow
- Cognos TM1 Web-Client

Administrations-Clients

- Cognos TM1 Performance Modeler
- Cognos TM1 Architect
- Cognos TM1 Perspectives
- Cognos TM1 Operations Console

IBM Cognos Insight ist sowohl ein Clientanwendung für TM1 Application Web, die lokal oder als Download zur Verfügung steht als auch ein Analysetool, um Daten zu analysieren (Vgl. IBM Corporation 2013f, S. vii).

„IBM Cognos TM1 Application Web ist ein Web-Client ohne lokalen Speicherbedarf, der das Öffnen von und Arbeiten mit Cognos TM1 Applications in einem beliebigen [...] Web-Browser ermöglicht. Ausgehend von der Seite ‚Workflow‘ können in Cognos TM1 Application Web Knoten geöffnet, Eigentumsrechte übernommen und Daten eingegeben werden (...). Cognos TM1 Application Web ist besonders dann von Nutzen, wenn eine unternehmensinterne Richtlinie die Installation eines lokalen Clients untersagt oder wenn ein anderes Betriebssystem als Microsoft Windows verwendet wird [...]“ (IBM Corporation 2013f, S. viii).

Der Web-Client IBM Cognos TM1 Web benötigt keinen lokalen Speicher und unterstützt bei der Analyse und Änderung von TM1-Daten in einem beliebigen Web-Browser. Über Cognos TM1 Web besteht kein Zugriff auf die Seite „Workflow“ von Cognos TM1 Application Web (Vgl. IBM Corporation 2013f, S. viii).

Der IBM Cognos TM1 Performance Modeler ist das neuste Modellierungstool von Cognos TM1 und unterstützt bei der Erstellung von Dimensionen, Cubes, Regeln, Prozessen und anderen Objekten. Das Tool vereinfacht den Modellierungsprozess für den Anwender, indem zum einen Regeln und Feeder automatisch generiert werden und zum anderen der Import von Daten und Metadaten zügig zu erledigen ist. Es ist vorgesehen, dass der Cognos TM1 Performance Modeler als das primäre Verwaltungs- und Entwicklungstool verwendet wird (Vgl. IBM Corporation 2013f, S. viii).

Auch der Cognos TM1 Architect ist ein Modellierungstool, unterstützt jedoch bei der Erstellung und Verwaltung aller TM1-Objekte. Im Gegensatz zum Performance Modeler bietet der Architect keine automatische Generierung von Regeln und geführte Imports. Aus diesem Grund sollte der Performance Modeler anstatt dem Architect als Entwicklungsumgebung für TM1-Modelle genutzt werden (Vgl. IBM Corporation 2013f, S. ix).

Cognos TM1 Perspectives ist das Excel Add-in, das schon seit längerer Zeit verfügbar ist und Microsoft Excel-Funktionen bereitstellt. Es ermöglicht die Entwicklung von TM1-Modellen und kann für die Analyse von Daten über Excel-Funktionen genutzt werden. Das Tool Perspectives unterstützt zwar auch die Erstellung und Verwaltung aller TM1-Objekte, ist jedoch nicht so umfangreich wie der Cognos Performance Modeler. Aus diesem Grund wird auch hier empfohlen, den Performance Modeler als Entwicklungsumgebung zu nutzen (Vgl. IBM Corporation 2013f, S. ix).

Das webbasierte Betriebstool Cognos TM1 Operations Console unterstützt bei der Überwachung, Unterstützung und Verwaltung von Cognos TM1 Servern. Zusätzlich dazu bietet es Einblicke in den Serverbetrieb. Mit Cognos Operations Console können Threadaktivitäten überwacht, sortiert und gefiltert sowie der aktuelle Status jedes Servers abgefragt werden. Für Cognos TM1-Administratoren sollte Cognos Operations Console als primäre Schnittstelle genutzt werden (Vgl. IBM Corporation 2013f, S. ix).

Im vorigen Abschnitt ist deutlich geworden, dass es für die unterschiedlichsten Funktionen von Cognos TM1 verschiedene Tools gibt, die genutzt werden können. Es wird jedoch an mehreren Stellen darauf hingewiesen, den Cognos Performance Modeler als Entwicklungsumgebung zu verwenden anstatt beispielsweise den Cognos Architect oder Cognos Perspectives. Dabei ist jedoch zu berücksichtigen, dass dieser Hinweis hauptsächlich an Mitarbeiter aus Fachabteilungen gerichtet ist, die häufig wenig Informatik-Know-How haben. Ist Know-How vorhanden und es sollen komplexe Dimensionen, Cubes oder auch Prozesse erstellt werden, dann eignet sich der Cognos Architect mehr.

3.6.2.3 Sicherheitskonzept

IBM Cognos TM1 bietet unterschiedliche Sicherheitsmechanismen an, die sich zum einem auf die Serverauthentifizierung und zum anderen auf die Objektsicherheit beziehen. Diese werden im Folgenden vorgestellt.

3.6.2.3.1 Serverauthentifizierung

Die Server-Authentifizierungsmethoden lassen sich nach Installation des Servers einstellen. Die vier folgenden Methoden sind dabei möglich (Vgl. IBM Corporation 2013c, S. 49):

- **TM1-Authentifizierung**
  Eingabe des Benutzernamens und des Kennwortes sind notwendig und werden anhand der Anmeldeinformationen des Sicherheitscubes überprüft.
- **Integrierte Anmeldung**
  Microsoft Windows führt die Cognos TM1-Authentifizierung durch.
- **LDAP[22]-Authentifizierung**
  Eingabe des Benutzernamens und des Kennwortes sind notwendig und werden anhand der Anmeldeinformationen eines externen LDAP-Servers überprüft.
- **IBM Cognos-Sicherheit**
  Benutzer werden anhand von Sicherheitseinrichtungen von IBM Cognos authentifiziert.

3.6.2.3.2 Objektsicherheit

Der Zugriff auf die einzelnen Objekte in Cognos TM1 kann beschränkt werden, unabhängig davon, welche Server-Authentifizierungsmethode ausgewählt wurde. Die Beschränkung kann sich dabei auf Cubes, Elemente, Dimensionen, Prozesse, Jobs, Anwendungen und Referenzen sowie Zellen beziehen. Dabei werden die Beschränkungen meist auf Benutzergruppen angewendet und umfassen folgende Rechte (Vgl. IBM Corporation 2013d, S. 113):

- **Admin**
  Eine Gruppe mit der Rolle Admin hat kompletten Zugriff auf einen Cube, ein Element, eine Dimension oder ein anderes Objekt.
- **Lock**
  Eine Gruppe mit der Rolle Lock kann einen Cube, ein Element, eine Dimension oder ein anderes Objekt sichten und bearbeiten. Diese Objekte können permanent gesperrt werden, um das Aktualisieren durch andere Benutzer zu verhindern.
- **Read**
  Eine Gruppe mit der Rolle Read kann einen Cube, ein Element, eine Dimension, einen Prozess oder einen Job sichten, jedoch das Objekt nicht bearbeiten.

[22] Lightweight Directory Access Protocol

- **Reserve**
  Eine Gruppe mit der Rolle Reserve kann einen Cube, ein Element, eine Dimension oder ein anderes Objekt sichten und bearbeiten. Diese Objekte können vorübergehend gesperrt werden, um das Aktualisieren durch andere Benutzer zu verhindern.

- **Write**
  Eine Gruppe mit der Rolle Write kann einen Cube, ein Element, eine Dimension, einen Prozess oder einen Job sichten und aktualisieren.

- **None**
  Eine Gruppe mit der Rolle None kann einen Cube, ein Element, eine Dimension, einen Prozess oder einen Job nicht sichten und das Objekt nicht bearbeiten.

3.6.2.4 Datenimport

Mit dem Turbo Integrator können Daten in einen IBM Cognos TM1-Cube importiert werden. Dieser ermöglicht das Anlegen eines Prozesses, dessen Ausführung zeitlich festgelegt werden kann. Folgende Datenquellen können dabei mit dem TM1 Turbo Integrator integriert werden (Vgl. IBM Corporation 2013g, S. 1):

- Kommagetrennte Textdateien, einschließlich ASCII-Dateien
- Relationale Datenbanktabellen, die über eine ODBC Datenquelle zugreifbar sind
- Andere Cubes und Ansichten
- Microsoft Analysis Services
- SAP über RFC[23]
- IBM Cognos-Packages

3.6.2.5 Handbücher

Für weitergehende Informationen können die folgenden Handbücher verwendet werden:

- **Cognos TM1 Performance Modeler - Handbuch**[24]
  Beschreibung der Entwicklungs- und Administrationsfunktionen von Cognos TM1 Performance Modeler

- **Cognos TM1 Entwicklerhandbuch**[25]
  Erläuterung der Entwicklungs- und Administrationsfunktionen von Cognos TM1 Architect und Cognos TM1 Perspectives

[23] Remote Function Call
[24] `http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/prfmdl_ug.pdf`
[25] `http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_dg_dvlpr.pdf`

- **Cognos TM1 Operations Console - Handbuch**[26]
  Erläuterung der Überwachung und Verwaltung der Server mithilfe von Cognos TM1 Operations Console
- **Cognos TM1 - Benutzerhandbuch**[27]
  Beschreibung der Analysefunktionen für Endbenutzer von Cognos TM1 Architect und Cognos TM1 Perspectives
- **Cognos TM1 Web - Benutzerhandbuch**[28]
  Beschreibung der Analysefunktionen für Endbenutzer von Cognos TM1 Web
- **Cognos TM1 Applications - Handbuch**[29]
  Beschreibung der Analysefunktionen für Endbenutzer von Cognos TM1 Application Web

3.6.3 Konzeption

Mit dem Werkzeug IBM Cognos TM1 in der Version 10.2 können Datenmodelle sehr dynamisch umgesetzt und optimiert werden. Ziel ist es, die Datenobjekte von TM1 zu nutzen, um die verschiedenen Daten zu speichern. Rules sollen dazu dienen, die Datenobjekte miteinander logisch zu verbinden. Mithilfe von Prozessen sollen die Datenobjekte mit Daten gefüllt werden. Außerdem kann so eine Simulation durchgeführt werden. Dem Endanwender soll über die TM1-Web-Oberfläche ein geführter Dialog angeboten werden, in dem er die Basisdaten einsehen, Simulationsparameter bestimmen und schließlich die Simulation durchführen kann. Das Ergebnis soll ihm ebenfalls über diese Oberfläche bereit stehen. Der Vorteil der TM1-Web-Oberfläche besteht in seiner Interoperabilität. Ein Anwender benötigt lediglich einen Standard-Browser und eine Internetverbindung.

In Absprache mit dem BI-CC wird der Freigabe-Work-Flow nicht umgesetzt, da dafür in der Projektgruppe nicht ausreichend Ressourcen bereit stehen. Als mögliche Erweiterungsstufe des Simulationssystems könnte dieser mit dem IBM Cognos Performance Modeler und einer Anwendung des Typs Genehmigung umgesetzt werden. Dabei dient eine Dimensionshierarchie als Pfad zur Genehmigung von Eingangsdaten.

Dem Lastenheft (Vgl. Volkswagen AG 2013) und dem daraus resultierenden Pflichtenheft (Vgl. Anhang F.4) ist zu entnehmen, dass ein Anwender mehrere Simulationen mit verschiedenen Parametern durchführen kann. Die steuerbaren Parameter sind dabei das Basisjahr, dessen Daten herangezogen werden, die ausgewählten Server, die von der Simulation betroffen sind sowie die drei Varianten best, middle und worst. Für die Varianten wählt der Anwender unterschiedliche

[26] `http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_ops_console.pdf`
[27] `http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_ug.pdf`
[28] `http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_wb.pdf`
[29] `http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_cont_ug.pdf`

prozentuale Veränderungswerte, die als angenommener Aufschlag für die tatsächliche Auslastung dienen.

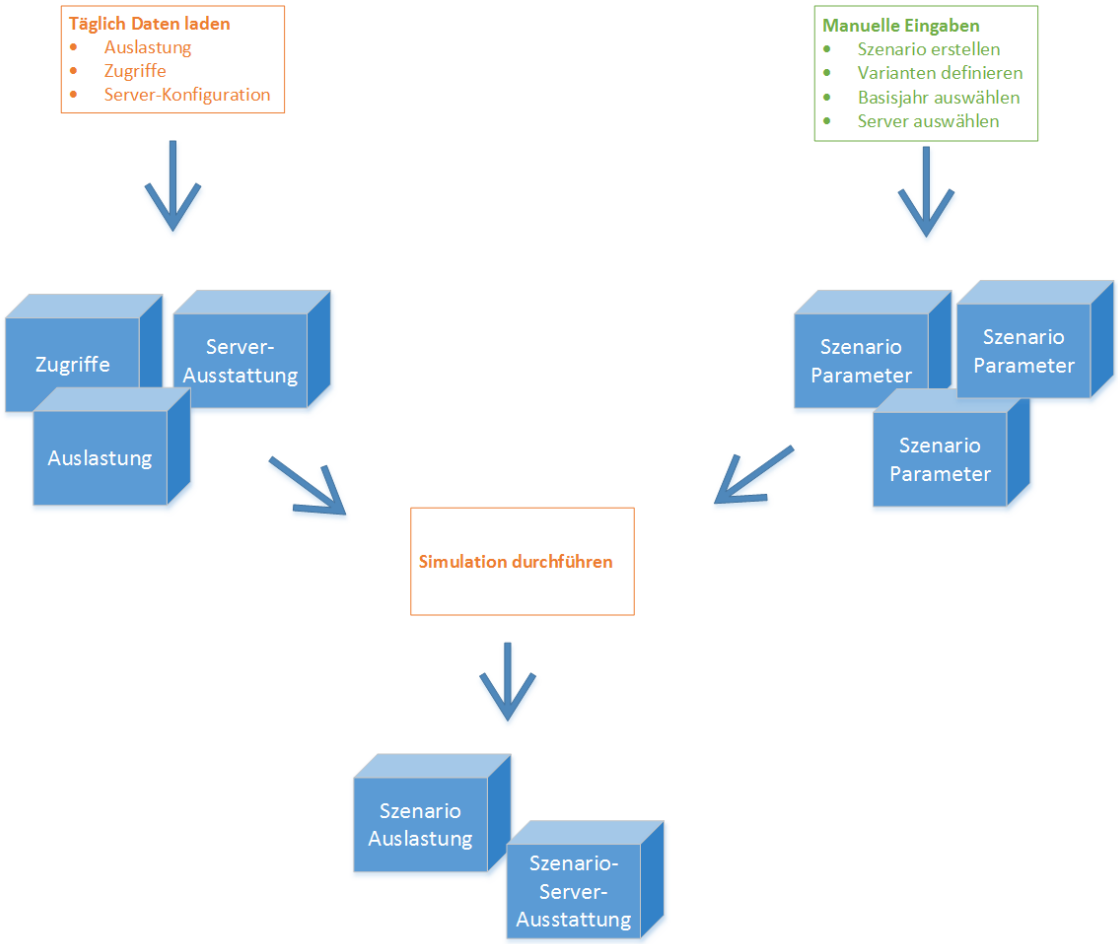


Abbildung 3.77: Konzeptmodell der Simulation

Der Abbildung 3.77 ist der geplante Aufbau des Systems, die benötigten Schnittstellen und Prozesse sowie das Vorgehen zur Durchführung eines Simulationsszenarios zu entnehmen. Das System nutzt verschiedene Cubes, um die Basisdaten über Zugriffe, Auslastung und Server-Ausstattung zu speichern. Tägliche Prozesse laden die Daten des Vortags in die Cubes.

Zur Durchführung einer Simulation erstellt der Anwender ein Simulationsszenario und definiert die Parameter. Dabei ist vor allem das Basisjahr der Ist-Daten auszuwählen und die prozentualen Aufschläge der drei Varianten best, middle und worst zu definieren. Die Einstellungen werden in verschiedenen Cubes hinterlegt. Anschließend startet der Anwender die Simulation für das erstellte Szenario.

Dem Lastenheft ist zu entnehmen, wie sich die neue Auslastung inklusive der Aufschläge berechnet. Die Abbildung 3.78 zeigt die Formeln nach denen die lineare Berechnung erfolgt. Der Prozess zur Durchführung der Simulation ermittelt für die ausgewählten Server auf der Basis der Ist-Auslastungsdaten inklusive dem Aufschlag in den drei Varianten best, middle

In dieser zur Veröffentlichung vorgesehenen Version entfernt.

Abbildung 3.78: Formeln für die neue Auslastung Prozessor und Speicher
Quelle: Volkswagen AG (2013, S.4)

und worst die maximalen Auslastungswerte des Basisjahres. Die Ergebnisse werden in einem Ergebnis-Cube gespeichert. Die Überschreitung der maximalen Auslastung ergibt das Verhältnis, um das die neue Ausstattung des Servers erweitert werden muss. Dabei wird jeder Server für sich selbst betrachtet. Übersteigt eine neue Ausstattungsempfehlung einer Variante die maximale Kapazität eines Servers, so muss die Servergruppe um einen neuen Server erweitert werden oder es müssen Server mit höheren Ausbaustufen angeschafft werden.

Für die Entwicklung des Prototyps gelten einige Einschränkungen gegenüber dem im Lastenheft dargestellten Soll-Zustand. So werden zunächst für wenige ausgewählte Server die Ist-Daten von ca. zwei Jahren integriert. Ein täglicher Lade-Prozess mit Datenbankanbindung wird zunächst nicht implementiert. Wie bereits erwähnt wird der Freigabe-Work-Flow nicht abgebildet, da für den Prototyp die Funktionalität der Simulation im Vordergrund steht.

3.6.4 Umsetzung

Das Kapitel Umsetzung unterteilt sich in die Abschnitte „Installation und Konfiguration“ und „Beschreibung des Prototyps“.

3.6.4.1 Installation und Konfiguration

Als Entwicklungsumgebung wurde von der Abteilung Very Large Business Applications der Universität Oldenburg ein Windows Server 2008 bereitgestellt. Darüber hinaus war es möglich über die IBM Academic Initiative die Installationspakete für die IBM Cognos TM1 Software zu bekommen. Außerdem wurde eine Microsoft Excel Installation benötigt. Die Konfiguration des Systems stellt sich wie folgt dar:

- **Server**
  Als Server wird eine virtuelle Instanz über den Open Source Virtualisierer QEMU in der Version 1.1.2 bereitgestellt. Die Instanz verfügt über eine Prozessor mit 2.00 GHz und vier Kernen, 8.00 GB RAM Arbeitsspeicher und 80 GB Festplattenspeicher.

- **Betriebssystem**
  Das Betriebssystem ist eine Installation des Windows Server 2008 R2 Standard in der 64bit-Version. Für die Installation und den Betrieb des Prototyps wird ein Benutzerkonto mit Administrator-Zugriff erstellt.

- **Microsoft Excel**
  Zur Erstellung der TM1-Web-Oberflächen wird Microsoft Excel 2010 in der 32bit-Version benötigt. Zum Einsatz kam die Installation in der Version 14.0.7116.5000 (32bit).

- **IBM Cognos TM1-Server**
  Als TM1-Server-Komponente wird das 64bit-Installationspaket in der Version 10.2.0 gewählt. Dieses installiert den TM1 Server, den TM1 Admin Server und die TM1-Tools. Auf Web Application-Ebene wird mit der Installation das TM1 Application Gateway, der TM1 Application Server, TM1 Web und die TM1 Operations Konsole eingerichtet. Im Anschluss wird über ein weiteres Installationspaket das Fix Pack 1 (Version 10.2.0.1) für die genannten Komponenten eingespielt.

- **IBM Cognos TM1-Clients**
  Da das Excel-AddOn TM1 Perspectives nur in der 32bit-Version bereit steht, werden alle Clients wie auch Microsoft Excel in dieser Version installiert. Dazu steht ein weiteres Installationspaket bereit. Auf der TM1 Client-Ebene wird TM1 Perspectives installiert. Auf der Entwickler-Ebene wird der TM1 Architect und der TM1 Performance Modeler ausgewählt. Auch für die Clients wird im Nachgang das Fix Pack 1 (Version 10.2.5001.145-0) installiert.

Nach erfolgreicher Installation kann das System konfiguriert werden. Um verschiedene TM1-Server-Instanzen auf einem Server logisch und physikalisch voneinander zu trennen, werden je Server-Instanz ein Ordner und ein Windows-Dienst erstellt. Über die Anwendung IBM Cognos Configuration kann die neue TM1 Server-Instanz capacity_management erstellt werden. Nach Angabe des TM1 Server-Konfigurationspfads kann die Server-Instanz registriert und gestartet werden. Ist das Verzeichnis des TM1 Server-Konfigurationspfads leer, so werden die benötigten

Dateien bei erstmaligem Starten mit der Standardkonfiguration erstellt. Eine zentrale Datei stellt die tm1s.cfg dar, die die spezielle Konfiguration für die Server-Instanz beinhaltet. Dazu gehören Einstellungen über die Authentifzierungsart, das Logging, die Parallelisierung, den Servernamen und viele mehr. Diese kann vom Entwickler angepasst werden, ist in diesem Fall aber nicht notwendig. Neben dieser Datei legt TM1 sämtliche Systeminformationen der Instanz ebenfalls in diesem Verzeichnis ab. Diese Dateien sind an der vorangestellten „}“ zu erkennen. So beinhaltet die Datei „}Clients.dim“ die dem System bekannten Benutzer in der Dimension Clients.

Die Installation und Konfiguration der Server-Instanz capacity_management ist nun abgeschlossen. Nun kann mit den Clients TM1 Architect und TM1 Performance Modeler auf die Instanz zugegriffen werden. Standardmäßig ist der Benutzer admin ohne Passwort integriert. Für dieses Szenario wird dieser Benutzer gesperrt und es werden für jeden Benutzer Konten mit eigenem Passwort angelegt. Dazu können in der Anwendung TM1 Architect bei der ausgewählten Server-Instanz über den Menüpunkt „Server - Sicherheit - Benutzer/Gruppen“ die Benutzer erstellt und der DataAdmin-Gruppe zugewiesen werden. Sie haben somit die Berechtigung Daten innerhalb der Datenobjekte zu erstellen, verändern und löschen, dürfen jedoch keine neuen Datenobjekte (Cubes, Dimensionen, etc.) erstellen.

Da nun die Server-Instanz läuft und der Zugriff über die Entwicklungswerkzeuge möglich ist, muss der TM1 Web-Dienst noch konfiguriert und gestartet werden. Es handelt sich dabei um einen vorkonfigurierten Apache Tomcat-Server, der sich im Installationsverzeichnis des TM1 Servers befindet. In dem Unterverzeichnis „tomcat\conf\“ liegt die Datei „server.xml“, in der der Port des Connectors für HTTP-Anfragen definiert ist. Standardmäßig ist dieser mit Port 9510 definiert, zur Vereinfachung wird dieser auf den Standard-HTTP-Port 80 geändert. Der Web-Server kann über den bei der Installation erstellten Windows-Dienst IBM Cognos TM1 Application Server gestartet werden. Nun steht die TM1 Web-Oberfläche über den lokalen Web-Browser unter der Adresse `http://localhost/tm1web` bereit, weitere Clients können diese über die öffentliche IP-Adresse `http://134.106.13.86/tm1web` aufrufen.

Die Konfiguration ist nun vollständig abgeschlossen. Zum jetzigen Zeitpunkt sind alle notwendigen Windows-Server-Dienste gestartet. Diese umfassen den Dienst IBM Congos TM1 Admin Server x64, der den allgemeinen TM1 Server-Dienst darstellt, der IBM Cognos TM1 Application Server, der den Web-Server startet und damit die TM1 Web-Oberfläche bereitstellt sowie den Dienst „IBM Cognos TM1 Server - capacity_management“, der die logische Instanz der Applikation darstellt.

Zur Installation und Konfiguration des TM1-Servers und der TM1-Clients steht das Handbuch „Installation und Konfiguration“ (Vgl. IBM Corporation 2013e) des Herstellers bereit. Die Verwaltung des TM1-Servers ist in dem „Betriebshandbuch“ (Vgl. IBM Corporation 2013c) beschrieben.

3.6.4.2 Beschreibung des Prototyps

Der Abschnitt zur Beschreibung des Prototypen enthält die Unterabschnitte Die Datenhaltung, Die Simulation und Die Oberfläche.

3.6.4.2.1 Die Datenhaltung

Der Prototyp zur Durchführung der Simulation im IT-Capacity Management besteht aus neun verschiedenen Cubes, 14 Dimensionen, wobei neun Dimensionen als Faktendimension fungieren und sechs Prozessen, wobei vier dem einmaligen Import von Daten dienen. Cubes wie Server-Zugriffe, ServerAuslastung und ServerAusstattung haben den Zweck die Ist-Daten zu speichern. Die Cubes SimulationSzenarioParameter, SimulationServerAuswahl und SimulationVarianteSzenarioParameter dienen dazu, die Parameter für die Simulation zu einem Szenario zu speichern. Die Ergebnisse des Simulationsprozesses werden in dem Cube SimulationVarianteSzenarioResult gespeichert. Die Cubes SimulationServerAuslastung und SimulationServerAusstattung sind für die Anzeige der Ergebnisdaten zuständig, beinhalten jedoch keine eigenen Daten, sondern stellen Sichten auf die per Rules definierten Ergebnisse aus der Berechnung der Ist-Daten mit den Simulationsergebnissen dar.

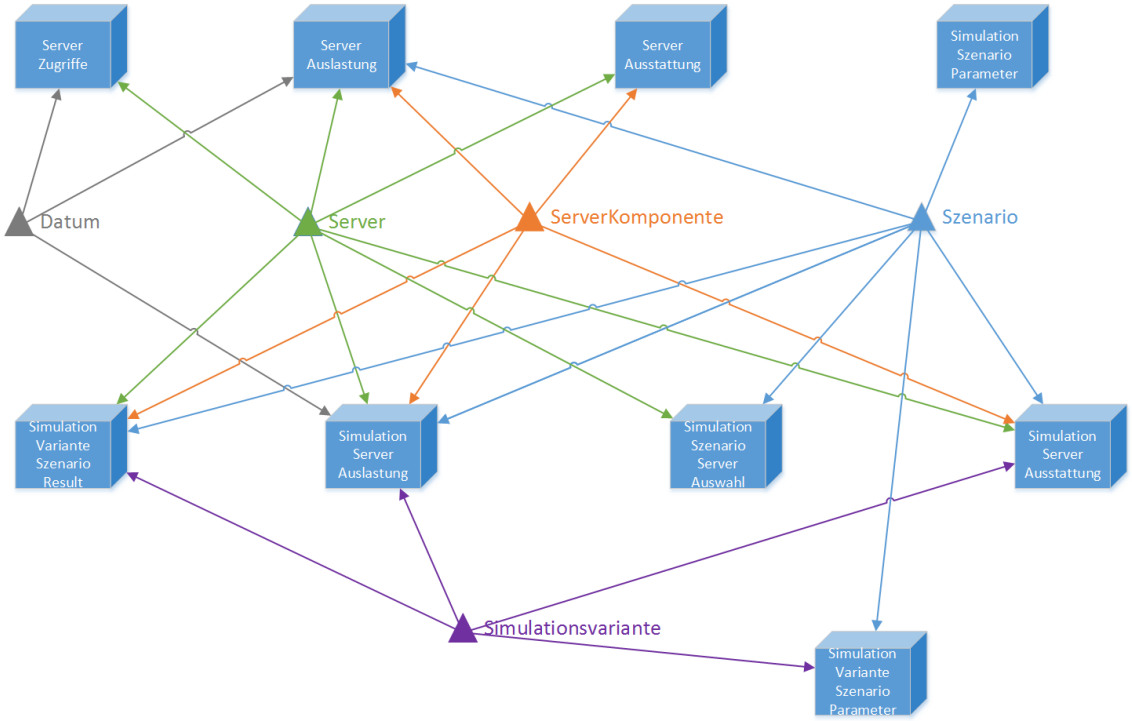


Abbildung 3.79: Modell der Dimensionen und Cubes

Zentrale Bestandteile des Systems sind die fünf Dimensionen Datum, Server, Serverkomponente, Szenario und Simulationsvariante. Diese Dimensionen werden in verschiedenen Cubes wiederverwendet. Die Abbildung 3.79 stellt die Dimensionen der Cubes dar. Der Übersichtlichkeit halber wurden die Faktendimensionen, die je Cube einmalig sind, weggelassen. Folgende Auflistung

stellt die Definition der Dimensionen dar:

- **Dimension Datum**
  Die Dimension Datum stellt eine dreistufige Hierarchie über die Ebenen Jahr und Monat auf den Elementen Datum dar. Von 2011 bis 2016 enthält die Dimension ein Element pro Tag.

- **Dimension Server**
  Die Dimension Server enthält zum Entwicklungszeitpunkt sieben Server als untergeordnete Elemente. Die Konsolidierungsebene über den Servern stellen die Servergruppen dar, die verschiedene Server eines Typs oder einer Anwendung zusammenfassen.

- **Dimension ServerKomponente**
  Als Serverkomponente werden der Arbeitsspeicher (storage) und der Prozessor (Prozessor) in dem System bezeichnet.

- **Dimension SimulationsVariante**
  Die Simulationsvarianten unterscheiden sich in best, middle und worst. Zusätzlich existiert das Element „alt“, um in Simulations-Cubes die Ist-Daten darzustellen.

- **Dimension Szenario**
  Diese Dimension beinhaltet einen Eintrag je Szenario, das der Anwender mit frei wählbarem Namen erstellt. Attribute sind das Erstellungsdatum (creation_date) und der Ersteller (author).

Diese Dimensionen spannen in unterschiedlicher Konstellation mit den Faktendimensionen den Raum zur Speicherung bzw. Darstellung der Werte auf. Nachfolgend werden die Cubes des Prototyps beschrieben. Dabei werden die Informationen über den Zweck des Cubes, die Dimensionen, aus denen der Cube besteht und Beispiele der in den Dimensionen enthaltenen Elemente geliefert.

- **Cube ServerAuslastung**
  Der Cube ServerAuslastung beinhaltet je Server, Serverkomponente und Tagesdatum die Auslastung.

  Dimensionen und Elemente:
  Der Cube ServerAuslastung besteht aus vier Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Datum	2014, 201403, 20140331
Server	Servergruppe1, Server10, Server11
ServerKomponente	prozessor, storage
ServerAuslastungFact	auslastung, auslastung_aktuell_prozent

Tabelle 3.27: Dimensionen und Elemente Cube ServerAuslastung

Rules

```
#Sind Daten vorhanden?
['dummy'] = N: if (['auslastung'] <> 0, 1, 0);
#Durchschnittswert berechnen
[ 'auslastung'] = C: ['hidden_auslastung'] \ ['dummy'];
#Auf unterster Ebene die Auslastung zuweisen
['hidden_auslastung'] = N: ['auslastung'];
#Aktuelle prozentuale Auslastung
['prozessor', 'auslastung_aktuell_prozent'] = ['auslastung'];
#Aktuelle prozentuale Auslastung anhand der Ausbaustufe
['storage', 'auslastung_aktuell_prozent'] = ['auslastung'] \ (DB('
    ServerAusstattung',!Server,'ist','storage')*1000);
```

Listing 3.7: Rules des Cubes ServerAuslastung

- **Cube ServerAusstattung**

  In dem Cube ServerAusstattung wird nur die aktuellste Konfiguration der Server gespeichert. Je Server wird die Ausstattung mit Prozessorkernen und Arbeitsspeicher sowie die maximal mögliche Bestückung festgehalten.

  Dimensionen und Elemente:

  Der Cube ServerAusstattung besteht aus drei Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Server	Servergruppe1, Server10, Server11
ServerKomponente	prozessor, storage
ServerAusstattungFact	ist, max, ratio

Tabelle 3.28: Dimensionen und Elemente Cube ServerAusstattung

Rules

```
['ratio'] = ['ist']\['max'];
```

Listing 3.8: Rules des Cubes ServerAusstattung

- **Cube ServerZugriffe**
  Für jeden Tag werden die Zugriffe je Server in den Cube ServerZugriffe geladen.

  Dimensionen und Elemente:
  Der Cube ServerZugriffe besteht aus drei Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Datum	2014, 201403, 20140331
Server	Servergruppe1, Server10, Server11
ServerZugriffeFact	zugriffe, auslastung_prozessor, auslastung_storage

Tabelle 3.29: Dimensionen und Elemente Cube ServerZugriffe

Rules

```
1 #Die Auslastung aus dem Cube dazu laden
2 ['auslastung_prozessor'] = DB('ServerAuslastung',!Datum,!Server,'
    prozessor','auslastung');
3 ['auslastung_storage'] = DB('ServerAuslastung',!Datum,!Server,'storage
    ','auslastung');
```

Listing 3.9: Rules des Cubes ServerZugriffe

- **Cube SimulationServerAuslastung**
  Der Cube SimulationServerAuslastung stellt den zentralen Cube in der Simulation dar. Anhand der Kennzahl „auslastung“ wird im Simulationsprozess der maximale Wert des Basisjahres ermittelt. Dieser Cube berechnet die Ist-Auslastung inklusive der Aufschläge je Variante, Server, Serverkomponente und Szenario. In den Rules dieses Cubes werden die vorgegebenen Formeln aus dem Lastenheft (Vgl. Abb. 3.78) abgebildet.

  Dimensionen und Elemente:
  Der Cube SimulationServerAuslastung besteht aus sechs Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Datum	2014, 201403, 20140331
Server	Servergruppe1, Server10, Server11
ServerKomponente	prozessor, storage
SimulationsVariante	best, middle, worst
Szenario	User-Szenarien
SimulationServerAuslastungFact	auslastung, zugriffe, auslastung_überschreitung, auslastung_szenario

Tabelle 3.30: Dimensionen und Elemente Cube SimulationServerAuslastung

Rules

```
#Auslastung alt
['alt','auslastung'] = DB('ServerAuslastung',!Datum,!Server,!
    ServerKomponente,'auslastung_aktuell_prozent');
#Zugriffe (Lasttreiber alt)
['alt','zugriffe'] = DB('ServerZugriffe',!Datum,!Server,'zugriffe');
[..]
```

Listing 3.10: Rules des Cubes SimulationServerAuslastung

- **Cube SimulationServerAusstattung**

  Im Cube SimulationServerAusstattung wird die neue empfohlene Ausstattung dargestellt. Anhand der Ist-Ausstattung aus dem Cube ServerAusstattung wird mit den ermittelten maximalen Auslastungswerten aus der Simulation die neue Plan-Ausstattung je Variante, Server und Serverkomponente berechnet.

  Dimensionen und Elemente:

  Der Cube SimulationServerAusstattung besteht aus fünf Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Server	Servergruppe1, Server10, Server11
Szenario	User-Szenarien
ServerKomponente	prozessor, storage
SimulationsVariante	best, middle, worst
SimulationServerAusstattungFact	ist, max, ratio, delta, delta_ratio

Tabelle 3.31: Dimensionen und Elemente Cube SimulationServerAusstattung

Rules

```
#Ist-Ausstattung
```

```
['alt', 'ist'] = DB('ServerAusstattung',!Server,'ist',!
    ServerKomponente);
#Maximale Ausbaustufe
[ 'max'] = DB('ServerAusstattung',!Server,'max',!ServerKomponente);
#Verhältnis
[ 'ratio'] = [ 'ist']\[ 'max'];

[..]

#Delta-Berechnung
[ 'delta'] = N: if ([ 'ist'] = 0, 0, [ 'ist'] - ['alt', 'ist']);
#Veränderungswert
[ 'delta_ratio'] = [ 'delta'] \ ['alt', 'ist'];
```

Listing 3.11: Rules des Cubes SimulationServerAusstattung

- **Cube SimulationsVarianteSzenarioParameter**
  Im Cube SimulationsVarianteSzenarioParameter werden die gewählten Aufschläge je Variante festgehalten, die der Anwender eingangs der Simulation definiert.

  Dimensionen und Elemente:
  Der Cube SimulationsVarianteSzenarioParameter besteht aus drei Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Szenario	User-Szenarien
SimulationsVariante	best, middle, worst
SimulationsVarianteSzenarioParameterFact	aufschlag, chosen_aufschlag

Tabelle 3.32: Dimensionen und Elemente Cube SimulationsVarianteSzenarioParameter

Rules

```
#Wenn 0, dann Default-Werte übernehmen
['chosen_aufschlag'] = N: if (['aufschlag'] = 0, ['default','aufschlag
    '], ['aufschlag']/100);
```

Listing 3.12: Rules des Cubes SimulationsVarianteSzenarioParameter

- **Cube SimulationSzenarioParameter**
  Im Cube SimulationSzenarioParameter werden die Simulationsparameter definiert. Dazu gehören das ausgewählte Basisjahr und ein Kommentar.

  Dimensionen und Elemente:
  Der Cube SimulationSzenarioParameter besteht aus zwei Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Szenario	User-Szenarien
SimulationSzenarioParameterFact	Kommentar, timestamp, abgeschlossen, basisjahr, chosen_basisjahr

Tabelle 3.33: Dimensionen und Elemente Cube SimulationSzenarioParameter

Rules

```
#Wenn 0, dann das Default-Jahr übernehmen
['chosen_basisjahr'] = if (['basisjahr'] = 0, ['default', 'basisjahr
    '], ['basisjahr']);
```

Listing 3.13: Rules des Cubes SimulationSzenarioParameter

- **Cube SimulationSzenarioServerAuswahl**
  Im Cube SimulationSzenarioServerAuswahl wird die Auswahl des Anwenders an Servern im Szenario gespeichert.

  Dimensionen und Elemente:
  Der Cube SimulationSzenarioServerAuswahl besteht aus drei Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Szenario	User-Szenarien
Server	Servergruppe1, Server10, Server11
SimulationSzenarioServerAuswahlFact	auswahl, tmp_auswahl, chosen_auswahl

Tabelle 3.34: Dimensionen und Elemente Cube SimulationSzenarioServerAuswahl

Rules

```
['tmp_auswahl']=  if (DB('SimulationSzenarioServerAuswahl',!Szenario,!
    Server,'auswahl') @= '1'
                  , 1
                  , 0);

```

```
5 #Server auswählen, wenn Servergruppe ausgewählt wurde
6 ['chosen_auswahl']=  if (DB('SimulationSzenarioServerAuswahl',!
      Szenario, ELPAR('Server',!Server, 1), 'tmp_auswahl') = 1
7                  , 1
8                  , if(['tmp_auswahl'] = 1, 1,0));
```

Listing 3.14: Rules des Cubes SimulationSzenarioServerAuswahl

- **Cube SimulationVarianteSzenarioResult**
  Im Cube SimulationVarianteSzenarioResult werden die während der Simulation ermittelten Maximalen Auslassungen je Server, Serverkomponenten, Szenario und Variante gespeichert.

  Dimensionen und Elemente:
  Der Cube SimulationVarianteSzenarioResult besteht aus fünf Dimensionen. Die folgende Tabelle zeigt die Liste der Dimensionen und Beispiele ihrer Elemente. Die Dimensionen sind nach ihrer Anordnung im Cube sortiert.

Dimensionen	Beispielelemente
Szenario	User-Szenarien
SimulationsVariante	best, middle, worst
ServerKomponente	prozessor, storage
Server	Servergruppe1, Server10, Server11
SimulationVarianteSzenarioResultFact	simulation_max, ratio, ist

Tabelle 3.35: Dimensionen und Elemente Cube SimulationVarianteSzenarioResult

Rules

```
1 #Ist-Ausstattung
2 ['ist'] = DB('ServerAusstattung', !Server ,'ist',!ServerKomponente);
3 ['ratio'] =  (['simulation_max']);
4 #Maximum auf C-Ebene darstellen
5 ['hidden_element'] = N: 1;
6 ['hidden_simulation_max'] = N: ['simulation_max'];
7 ['simulation_max'] = C: ['hidden_simulation_max']\ ['hidden_element'];
```

Listing 3.15: Rules des Cubes SimulationVarianteSzenarioResult

Durch die Wiederverwendung derselben Dimensionen in mehreren Cubes ist ein dichtes Netz an Abhängigkeiten entstanden. Jede Änderung an einer Dimension wirkt sich direkt auf alle Cubes aus, die mit dieser Dimension erstellt worden sind. Weitere Abhängigkeiten und Relationen werden durch die Rules definiert. Dadurch ist es möglich, neue Kennzahlen anhand der vorhandenen Kennzahlen im Cube zu erstellen, darüber hinaus können auch die Ergebnisse aus anderen Cubes herangezogen werden. Die Abbildung 3.80 zeigt diese Abhängigkeiten, indem sie darstellt, welche Kennzahlen zur Berechnung in andere Cubes via Rule herangezogen werden (blaue Verbindung zwischen zwei Cubes).

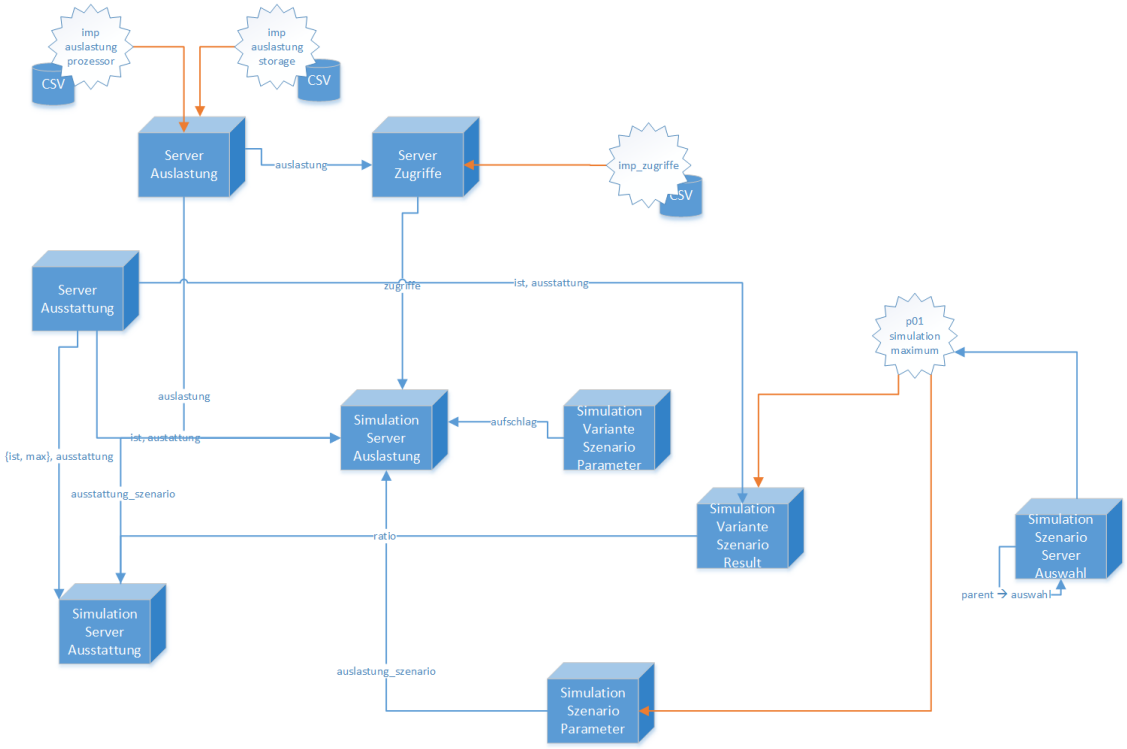


Abbildung 3.80: Modell der Relationen zwischen den Cubes

Des Weiteren stellt diese Abbildung die Prozesse dar (Sternform), die Daten im- oder exportieren. Durch die orangenen Pfeile von den Prozessen zu den Cubes werden Datenbewegungen gekennzeichnet. Die Prozesse imp_auslastung_prozessor, imp_auslastung_storage und imp_zugriffe importieren die Basisdaten und speichern diese in den Cubes ServerAuslastung und ServerZugriffe. Der Prozess p01_simulation_maximum führt anhand der ausgewählten Server und des Basisjahres für die verschiedenen Varianten eines Szenarios die Simulation durch und speichert die Ergebnisse in dem Cube SimulationVarianteSzenarioResult.

3.6.4.2.2 Die Simulation

Die Kernfunktionalität des Systems ist die Simulation. Die Abbildung 3.81 beschreibt die Schritte, die vollzogen werden müssen, um die Simulation starten zu können. Zunächst wählt der Anwender eine beliebige Bezeichnung und erstellt durch Starten des Prozesses p02_szenario_create ein Dimensionselement in der Dimension Szenario.

Im Anschluss kann er für das soeben erstellte Szenario die Parameter definieren. Dazu gehört unter anderem das Basisjahr der Ist-Daten, das als Planungsgrundlage dient. Dieser Parameter wird im Cube SimulationSzenarioParameter für das Szenario hinterlegt. Des Weiteren bestimmt der Anwender die prozentualen Aufschläge je Variante. Diese Einstellungen werden im Cube SimulationVarianteSzenarioParameter gespeichert. Als letzten Parameter wählt der Anwender den Server bzw. die Servergruppe aus, für die die Simulation durchgeführt werden soll. Diese Einstellung wird in dem Cube SimulationSzenarioServerAuswahl gesichert.

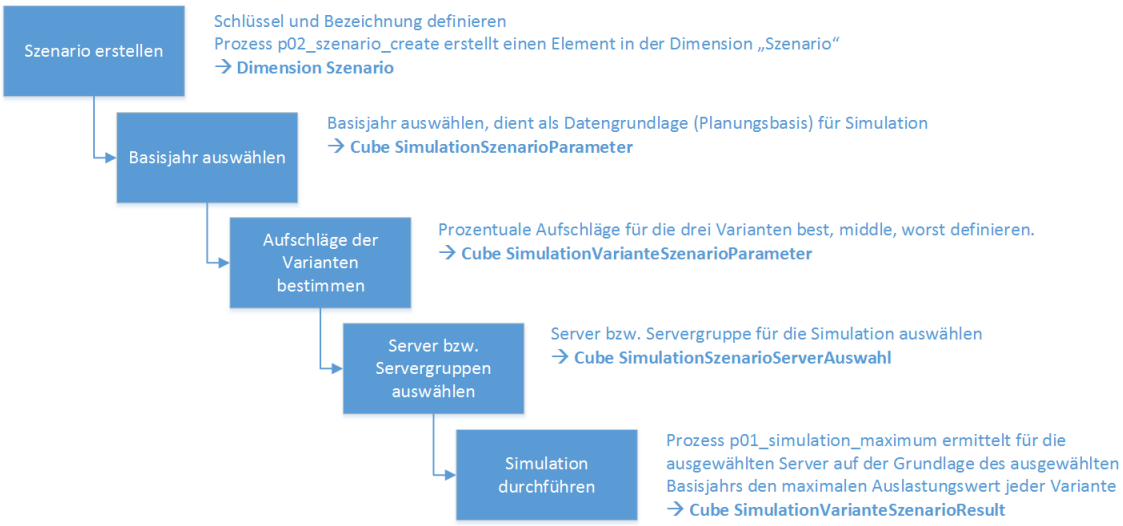


Abbildung 3.81: Prozessschritte der Simulation

Mit diesen Parametern kann die Simulation durchgeführt werden. Dazu startet der Anwender den Prozess p01_simulation_maximum. In dem gewählten Basisjahr wird für die bestimmten Server die maximale Überschreitung der Auslastung inklusive der Aufschläge in den drei Varianten bestimmt und im Cube SimulationVarianteSzenarioResult festgehalten. Das folgende Listing 3.16 stellt den Prolog des Prozesses dar. Als Eingangsparameter erhält der Prozess das Szenario als Zeichenfolge sowie den ausgewählten Server bzw. die Servergruppe. Zunächst wird die Serverauswahl in den Cube SimulationSzenarioServerAuswahl in die Kennzahl „auswahl“ geschrieben (Zeile 2). Somit erhält der Prozess bei einer ausgewählten Servergruppe auch die dazugehörigen Server. Das Basisjahr wird ebenfalls ausgelesen (Zeile 6).

Das Maximum für eine Variante, einen Server, eine Serverkomponente und ein Basisjahr in einem Szenario wird mit der Turbo Integrator-Funktion ConsolidatedMax() ermittelt. Dazu dienen drei verschachtelte Schleifen für alle Server, alle Serverkomponenten und alle Varianten. Die ausgewählten Server werden mit einem dynamischen Subset durch einen MDX-Befehl ermittelt (Zeilen 8-16). Darüber hinaus wird ein Subset für die Darstellung auf der Oberfläche erstellt, das bei einer ausgewählten Servergruppe auch diese neben den Servern beinhaltet (Zeilen 18-26). Der konkrete maximale Wert wird nun durch die Turbo Integrator-Funktion ConsolidatedMax() aus dem Cube SimulationServerAuslastung ermittelt (Zeile 59). Mit der Funktion CellPutN() wird das Ergebnis in den Cube SimulationVarianteSzenarioResult geschrieben (Zeile 60).

```
1 #Serverauswahl in Cube speichern
2 CellPutS('1', 'SimulationSzenarioServerAuswahl', szenario, chosenserver, '
    auswahl');
3 itmpserver = CellGetN('SimulationSzenarioServerAuswahl', szenario,
    chosenserver, 'chosen_auswahl');
4
5 # Das ausgewählte Basisjahr ermitteln
6 jahr = CellGetN('SimulationSzenarioParameter', szenario, 'chosen_basisjahr
```

```
    ');
7
8 #Das Subset anlegen für die ausgewählten Server
9 cDim = 'Server';
10 cSubProcess = 'sub_process_' | szenario;
11 IF (SubsetExists(cDim, cSubProcess) = 1);
12      SubsetDestroy(cDim, cSubProcess);
13 ENDIF;
14
15 SubsetCreatebyMDX(cSubProcess, '{TM1FILTERBYLEVEL({TM1DRILLDOWNMEMBER( {
    FILTER( { HIERARCHIZE( {TM1SUBSETALL( [Server] )} ) }
16 , [SimulationSzenarioServerAuswahl].([Szenario].['| szenario | '], [
    SimulationSzenarioServerAuswahlFact].[chosen_auswahl]) = 1)}, ALL,
    RECURSIVE )}, 0)}');
17
18 #Das Subset anlegen für die ausgewählten Server inkl. Parent für die
    Darstellung in der Oberfläche
19 cSub = 'sub_source_' | szenario;
20 IF (SubsetExists(cDim, cSub) = 1);
21      SubsetDestroy(cDim, cSub);
22 ENDIF;
23
24 SubsetCreatebyMDX(cSub, '{TM1DRILLDOWNMEMBER( {FILTER( { HIERARCHIZE( {
    TM1SUBSETALL( [Server] )} ) }
25 , [SimulationSzenarioServerAuswahl].([Szenario].['| szenario | '], [
    SimulationSzenarioServerAuswahlFact].[chosen_auswahl]) = 1)}
26 , ALL, RECURSIVE )}');
27
28 iServer = 1;
29 iServerCount = SubsetGetSize(cDim, cSubProcess);
30
31 #Schleife für jeden Server
32 while (iServer <= iServerCount);
33
34   server = SubsetGetElementName(cDim, cSubProcess, iServer);
35   i = 0;
36   x = 0;
37
38 #Schleife für die Kompontenten
39 while (x <= 1);
40 if (x = 0);
41   komponente = 'prozessor';
42 else;
43   komponente = 'storage';
44 endif;
45
46 i = 0;
47 #Schleife für jede Variante
48 while (i <= 2);
```

```
49
50 if (i = 0);
51   variante = 'best';
52 elseif (i = 1);
53   variante = 'middle';
54 else;
55   variante = 'worst';
56 endif;
57
58 #Das konsolidierte Maximum mit den ausgewählten Dimensionselementen
      ermitteln und in den Result-Cube schreiben
59 nMax = ConsolidatedMax( 0,  'SimulationServerAuslastung', NumberToString(
      jahr), server, komponente, variante, szenario, 'auslastung' );
60 CellPutN(nMax, 'SimulationVarianteSzenarioResult', szenario, variante,
      komponente, server, 'simulation_max');
61
62 i = i + 1;
63 end;
64 x = x + 1;
65 end;
66 iServer = iServer + 1;
67 end;
```

Listing 3.16: Prolog des Prozesses p01_simulation_maximum

Den Abschluss des Prozesses bildet der Epilog (Vgl. Listing 3.17), der die Simulation für das Szenario auf abgeschlossen setzt und den Timestamp in den Cube SimulationSzenarioParameter schreibt.

```
1 #Simulation abgeschlossen, Timestamp und Flag in den Parametern festhalten.
2 sTimestamp = TIMST(now, '\D.\m.\Y \h:\i:\s', 1);
3 CellPutS(sTimestamp, 'SimulationSzenarioParameter', szenario, 'timestamp');
4 CellPutN(1, 'SimulationSzenarioParameter', szenario, 'abgeschlossen');
```

Listing 3.17: Epilog des Prozesses p01_simulation_maximum

3.6.4.2.3 Die Oberfläche

Zur Darstellung der Inhalte und zur Durchführung der Simulation dient die TM1 Web-Oberfläche. Sie kann mit vielen Standardbrowsern geöffnet werden und bedarf keiner weiteren Installation eines Clients. Über die Adresse `http://134.106.13.86/tm1web` ist die Oberfläche während der Durchführung des Projektes zu erreichen. Nach Auswahl des TM1 Servers capacity_management und unter Angabe eines gültigen Benutzers ist die Anmeldung möglich. Der Anwender erhält im Anschluss an die erfolgreiche Anmeldung eine Übersicht über die für ihn freigeschalteten Cubes und Anwendungen[30]. Unterhalb des Listenpunkts Anwendungen befinden sich die Ordner Analyse und Simulation. Innerhalb des Ordners Analyse befindet sich die Excel-Datei „Analyse

[30] Zum Umgang mit TM1 Web hält der Hersteller ein Web-Benutzerhandbuch bereit (Vgl. IBM Corporation 2013h)

der Zugriffe.xlsm“. Durch Klick auf die Datei öffnet sich im rechten Fensterbereich der erste Reiter der Excel-Datei als Weboberfläche. Die Datei wurde zwar mit Microsoft Excel erstellt, wird in TM1 Web aber als Webseite interpretiert. Durch das Excel-AddOn TM1 Perspectives ist es möglich, Formeln zu erstellen, die einen lesenden oder schreibenden Zugriff auf die Daten in TM1 Cubes zulassen. Diese Funktionalität steht sowohl in Excel zur Entwicklungszeit als auch in TM1 Web nach Veröffentlichung der Datei bereit.

Die Datei „Analyse der Zugriffe.xlsm“ dient zur Analyse der Ist-Daten und zeigt auf den drei Arbeitsblättern Zugriffe, Details und Konfiguration verschiedene Diagramme zu den Auslastungsdaten und Zugriffen im Verlauf der Zeit sowie die aktuelle Konfiguration der dem System bekannten Server.

Im Ordner Simulation befinden sich die zwei Dateien „Simulation Ergebnisse je Tag.xlsx“ und „Simulation.xlsm“. Die erste Datei zeigt die tagesgenauen Ergebnisse der bereits durchgeführten Simulationsszenarien. Hauptbestandteil des Systems ist jedoch die zweite Datei „Simulation.xlsm“, die den Simulationsassistenten darstellt. Sie führt den Benutzer durch alle notwendigen Parametrisierungsschritte, die im vorherigen Kapitel beschrieben wurden. Dabei dienen drei Arbeitsblätter zur Parameterdefinition und zum Ausführen der Simulation und weitere sechs Arbeitsblätter zur Ergebnisdarstellung. Der vollständige Funktionsumfang ist dem Kapitel 3.6.5 Benutzerhandbuch zu entnehmen.

Zentraler Bestandteil der Excel-Datei „Simulation.xlsm“ sind die Funktionen, die den Zugriff auf die Daten in den Cubes des TM1 Servers erlauben. Dazu gehören zum einen Aktionsschaltflächen, die einen TurboIntegrator-Prozess ausführen, zu einem anderen Arbeitsblatt verzweigen, diese Funktionen in Kombination durchführen oder lediglich das Arbeitsblatt neu berechnen bzw. generieren können. Zum Start eines Prozesses können Parameter aus Zellen des Arbeitsblattes ausgelesen und übergeben werden.

Um auf die Elemente einer Dimension zugreifen zu können stehen verschiedene Arbeitsblatt-Funktionen[31] zur Verfügung wie z.B. DIMIX() und DIMNM(). Es empfiehlt sich aber zur kontextabhängigen Darstellung Subsets einer Dimension zu nutzen. Mit der Funktion SUBNM() können einzelne Elemente des Subsets ausgewählt werden. Dem Anwender wird somit der Inhalt des Subsets angezeigt.

Der Inhalt eines Cubes kann mit der Funktion DBR() ausgelesen und in Zellen mit der Funktion DBRW() auch verändert werden. Parameter dieser Funktionen sind der eindeutige Cube-Name inklusive des TM1 Servers sowie die Angabe der Elemente aller Dimensionen des Cubes. Der Planwert für die Serverausstattung der Serverkomponente Storage in der Variante worst für den Server Server1, der im Simulationsszenario „Simulation Admin 03.03.2014“ ermittelt wurde kann demnach mit folgender Formel ausgelesen werden:

[31] Zum Datenzugriff über Arbeitsblätter hält der Hersteller ein Kapitel in einem Benutzerhandbuch bereit (Vgl. IBM Corporation 2013b, S.107ff)

```
1 =DBR("capacity_management:SimulationServerAusstattung; "Server1"; "
    Simulation Admin 03.03.2014"; "Storage"; "worst"; "ist")
```

Listing 3.18: Die DBR-Funktion

Die Daten aus Cubes verhalten sich wie normale Inhalte in der Excel-Datei, sodass Diagramme anhand dieser Daten erstellt werden können. In dem Simulationsassistenten wurden häufig zunächst SUBNM()-Funktionen verwendet, um die Auswahl der Dimensionselemente variabel zu halten. Die Datenansicht verweist dann anstelle einer expliziten Angabe der Dimensionselemente wie in Listing 3.18 zu sehen ist, auf die entsprechenden Arbeitsblatt-Zellen. Dadurch werden die Ergebnisse in Abhängigkeit der Anwenderauswahl dargestellt und es ergibt sich eine dynamische Oberfläche.

Mit den Schaltflächen „Hochladen“ bzw. „Aktualisieren“ im Excel-Menüband des AddOns TM1 kann die Excel-Datei veröffentlicht werden und steht auch in der TM1 Web-Oberfläche zum Aufruf bereit. Wie bei allen Datenobjekten in TM1 können auch für diese Dateien die Zugriffsrechte für Benutzergruppen eingeschränkt werden. Dies ist in diesem Prototyp allerdings nicht notwendig.

3.6.5 Benutzerhandbuch

Das folgende Benutzerhandbuch unterteilt sich in die Abschnitte Allgemeines, Anmelden und Durchführen einer Simulation.

3.6.5.1 Allgemeines

Die prototypische Anwendung zur Durchführung einer Simulation im IT-Capacity Management in der Version 1.0 ermöglicht dem Anwender, eigene Szenarien zu erstellen, die auf realen Daten basieren und die Szenarien mit verschiedenen Parametern zu verändern. Die Ergebnisse einer Simulation werden persistent im System gespeichert und können auch zu späteren Zeitpunkten nachvollzogen werden. Das System ist unter registrierten Anwendern offen und ermöglicht die Einsicht von Szenarien anderer Anwender.

Zur einfachen Durchführung einer eigenen Simulation beinhaltet das System einen Simulationsassistenten. Dieser stellt nach der Durchführung auch die Ergebnisse Schritt für Schritt dar. Verschiedene Diagramme und Ansichten erklären die Daten und stellen die Zusammenhänge und Abhängigkeiten dar. Zu den Ergebnissen gehört auch eine Ausstattungsempfehlung, die Maßnahmen für das Eintreffen der Simulationsannahme vorschlägt.

3.6.5.2 Anmelden

Bevor Sie eine Simulation durchführen können, müssen Sie sich zunächst einmal anmelden. Dazu benötigen Sie einen registrierten Benutzernamen und ein gültiges Passwort. Verfügen Sie über keinen Benutzernamen, so lassen Sie sich von dem Systemadministrator einen Benutzerzugang einrichten. Während der Projektgruppenlaufzeit verfügt jeder Anwender über einen eigenen

Zugang, zum Testen des System steht auch der Benutzername „tester“ mit gleichnamigen Passwort bereit.

Rufen Sie die URL des Servers auf, auf dem die IBM Cognos TM1-Software installiert ist und der TM1-Server capacity_management läuft. Während der Projektgruppenlaufzeit war der Server unter der URL `http://134.106.13.86/tm1web` zu erreichen. Wenn Sie das System gerade selbst installiert und gestartet haben und lokal auf dem Server angemeldet sind, können Sie den TM1 Web-Dienst über die URL `http://localhost/tm1web` aufrufen. Geben Sie die URL in die Adresszeile eines Standard-Browser ein. Die Tabelle 3.36 zeigt die Browser, die mit der TM1 Web-Oberfläche in der verwandten Konfiguration kompatibel sind:

Produktname	**Version**
Google Chrome	28, 31
Microsoft Internet Explorer	8.0, 9.0, 10
Mozilla Firefox	ESR 17.0, 20

Tabelle 3.36: Kompatible Web-Browser
Quelle: IBM (2014b)

Nachdem Sie nun die URL in das Adressfeld Ihres Browser eingegeben haben und auf „Seite aufrufen“ geklickt haben, öffnet sich die Anmeldeseite von TM1 Web (Vgl. Abb. 3.82).

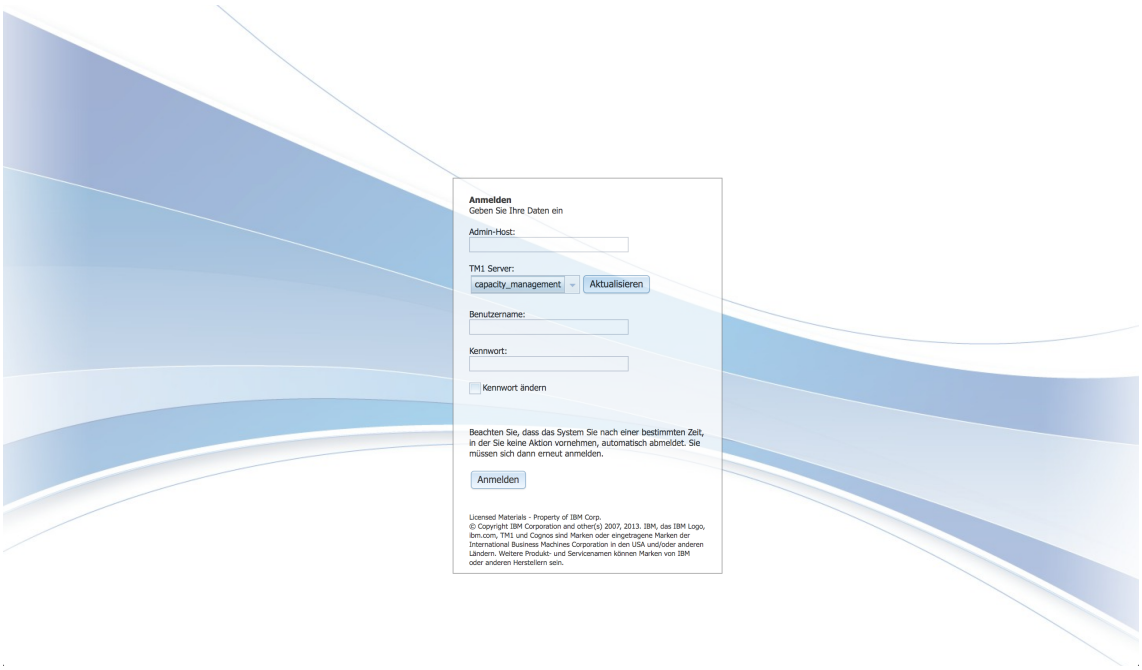


Abbildung 3.82: Anmeldeseite von TM1 Web

Wählen Sie den TM1 Server capacity_management aus, geben Sie Ihren Benutzernamen und das Kennwort ein. Lassen Sie das Feld Admin-Host leer. Klicken Sie auf Anmelden. Wenn Sie

eine falsche Kombination aus Benutzername und Kennwort eingegeben haben, erscheint die Meldung „Anmeldung fehlgeschlagen, versuchen Sie es erneut.“. Geben Sie dann bitte eine gültige Kombination ein oder kontaktieren die Systemadministrator. Haben Sie eine gültige Kombination eingegeben, werden Sie zur Startseite des Systems weitergeleitet (Vgl. Abb. 3.83).

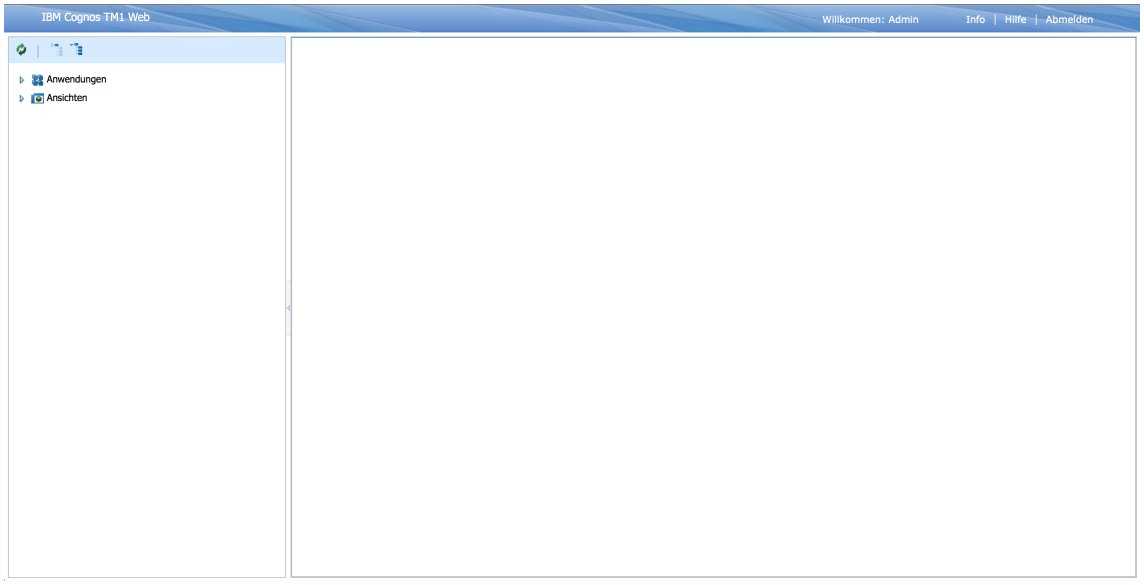


Abbildung 3.83: Startseite von TM1 Web

Die Startseite besteht aus einem linken Navigationsbereich, einer blauen Titelzeile und einem rechten Bereich zur Darstellung der Inhalte. Wenn Sie sich abmelden wollen, klicken Sie bitte oben rechts in der Titelzeile auf Abmelden. Um mehr über den Inhalt des Systems zu erfahren, klicken Sie auf die kleinen Dreiecke, die sich im Navigationsbereich vor den Punkten „Anwendungen“ und „Ansichten“ befinden. Dadurch klappen Sie die Ordnerstruktur auf. Wenn Sie das für alle Ordner des Punkts Anwendungen gemacht haben, erhalten Sie die vollständige Übersicht über die für Sie auswählbaren Anwendungen des Systems (Vgl. Abb. 3.84).

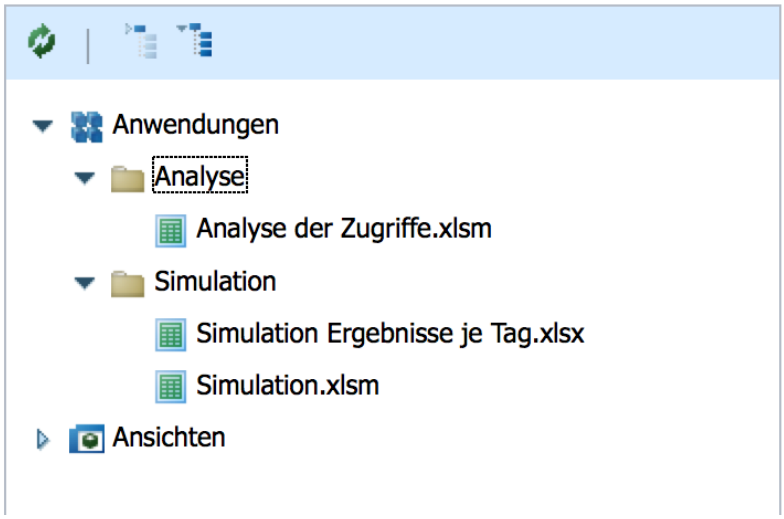


Abbildung 3.84: Alle Anwendungen im Navigationsbereich

Durch Klick auf eine Anwendung öffnen Sie die Anwendung. Sie erkennen diese an den grünen Symbolen vor dem Dateinamen, z.B. „Analyse der Zugriffe.xlsm“.

3.6.5.3 Durchführen einer Simulation

Sie sind nun angemeldet und haben im Navigationsbereich die Anwendungen aufgeklappt, sodass Sie den Eintrag „Simulation.xlsm“ sehen. Klicken Sie auf diesen Eintrag und es öffnet sich die erste Seite des Simulationsassistenten im rechten Inhaltsbereich. Die erste Seite des Simulationsassistenten hat den Namen Simulation-1-Neu (Vgl. Abb. 3.85).

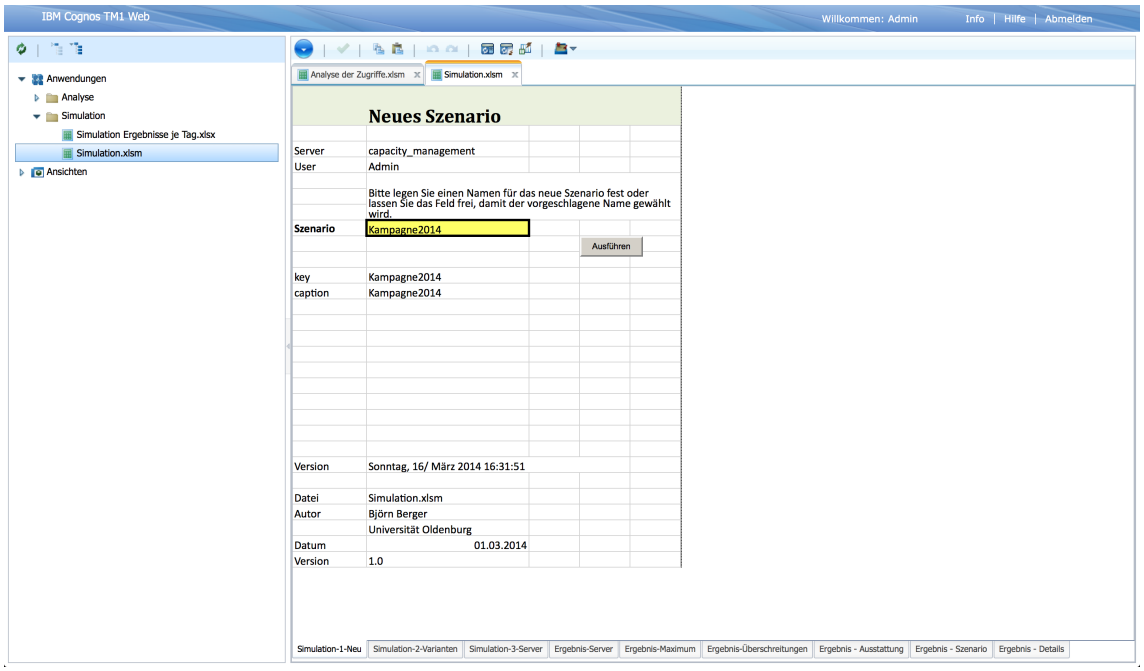


Abbildung 3.85: Seite Simulation-1-Neu des Simulationsassistenten

Sie haben nun die Möglichkeit ein eigenes Szenario anzulegen. Lassen Sie dazu entweder das gelbe Feld „Szenario“ frei um den vorgeschlagenen technischen Namen zu nutzen oder wählen Sie einen eigenen Namen. In diesem Beispiel wurde der Name Kampagne2014 für das Szenario gewählt. Klicken Sie dann auf Ausführen. Es wird ein Szenario mit dem von Ihnen gewählten Namen erstellt und Sie werden zur nächsten Seite im Simulationsassistenten weitergeleitet.

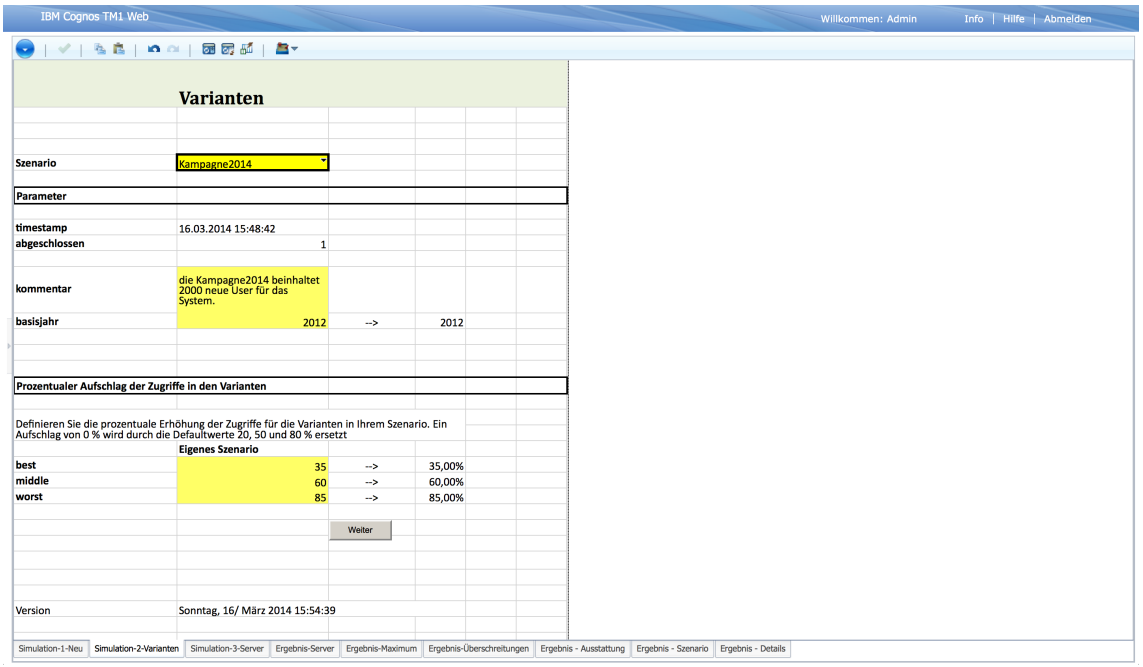


Abbildung 3.86: Seite Simulation-2-Varianten des Simulationsassistenten

Die Seite Simulation-2-Varianten (Vgl. Abb. 3.86) hilft Ihnen beim Ausfüllen der benötigten Parameter. Sie können alle gelben Felder frei lassen. In diesem Fall werden die Standardeinstellungen für die Simulation verwandt. Sie haben aber die Möglichkeit das Basisjahr für die Ist-Daten auszuwählen. Anhand dieses Parameters wird die Daten des ganzen ausgewählten Jahres zur Simulation hinzugezogen. Um weitere beschreibende Informationen zu hinterlegen, können Sie im Feld Kommentar einen Kommentar hinterlegen. Außerdem können Sie die prozentualen Aufschläge in Ihrem Simulationsszenario bestimmen. Ihnen stehen dazu drei verschiedene Varianten mit den vorgegebenen Namen best, middle und worst bereit, um verschiedene Stufen Ihres Szenarios abzubilden.

Beispiel:
Sie gehen davon aus, dass 2000 neue Anwender Ihr System nutzen werden und Sie vermuten, dass dies zu einer prozentualen Erhöhung der Auslastung von 60 % führt. Dann wählen Sie 60 als Wert für die Variante middle. Da es Abweichungen nach oben und unten in Ihrer Vermutung geben kann wählen sie einen kleineren Aufschlag für die Variante best (z.B. 35) und einen höheren Wert für die Variante worst (z.B. 85).

Wenn Sie Ihre prozentualen Aufschläge für Ihr Szenario eingegeben haben, klicken Sie bitte auf

„Weiter“, um zur nächsten Seite im Simulationsassistenten zu gelangen.

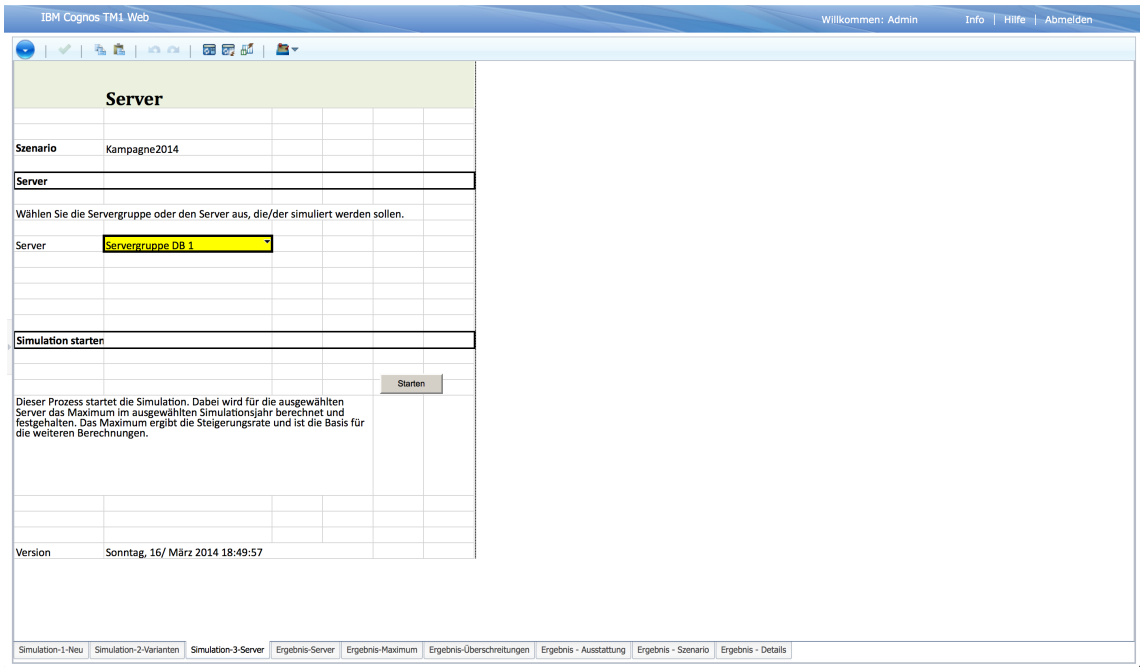


Abbildung 3.87: Seite Simulation-3-Server des Simulationsassistenten

Die nächste Seite Simulation-3-Server des Simulationsassistenten dient zur Auswahl des Servers oder der Servergruppe, für die das Simulationsszenario durchgeführt werden soll (Vgl. Abb. 3.87). Durch Klick auf das gelbe Feld „Server“ erhalten Sie einen Dialog (Vgl. Abb. 3.88) mit sämtlichen dem System bekannten Servern und Servergruppen. Wählen Sie bitte Ihren gewünschten Server aus (in diesem Beispiel „Servergruppe DB 1“).

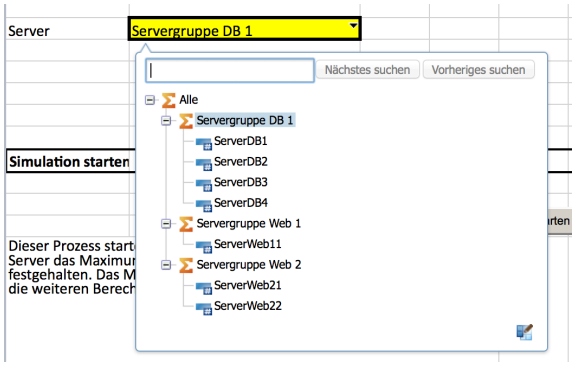


Abbildung 3.88: Dialog zur Auswahl des Serves oder der Servergruppe

Klicken Sie nach erfolgter Auswahl auf den Button „Starten“ um die Simulation durchzuführen. Jetzt wird ein Prozess im Hintergrund gestartet, der anhand Ihrer zuvor bestimmten Parameter die Simulation durchführt. Für jeden ausgewählten Server wird nun in jeder Variante für die Serverkomponenten Prozessor und Storage im ausgewählten Basisjahr die höchste Überschreitung

der Auslastungsgrenze (diese liegt bei 100 %) ermittelt und gespeichert. Nach erfolgreicher Durchführung werden Sie zur ersten Seite der Ergebnisse weitergeleitet.

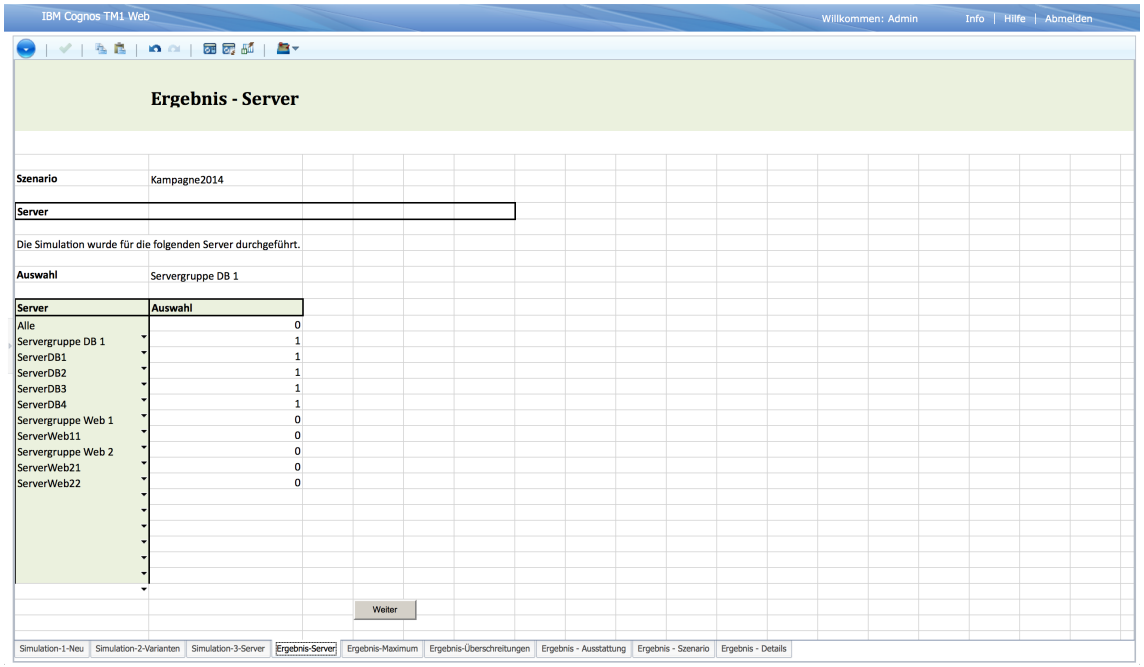


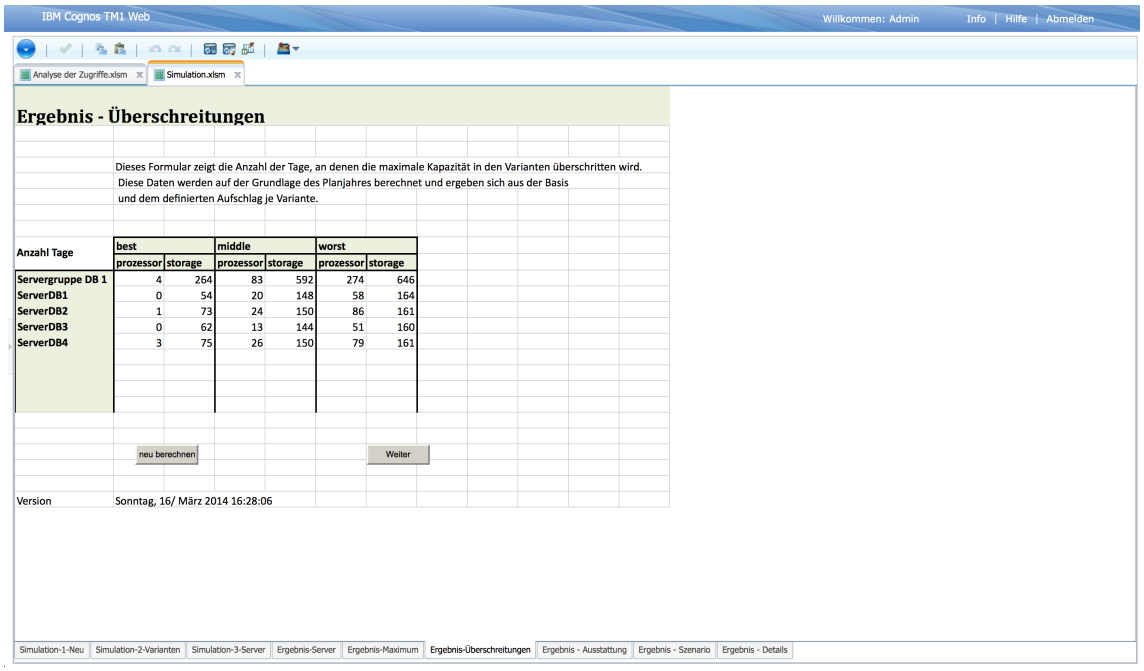
Abbildung 3.89: Seite Ergebnis-Server des Simulationsassistenten

Die Seite Ergebnis-Server (Vgl. Abb. 3.89) stellt Ihnen noch einmal die ausgewählten Server für die Simulation dar. Wenn Sie beispielsweise eine Servergruppe ausgewählt haben, wurde für jeden Server in der Servergruppe die Simulation durchgeführt. Klicken Sie auf „Weiter“, um zur nächsten Seite zu gelangen.

In dieser zur Veröffentlichung vorgesehenen Version entfernt.

Abbildung 3.90: Seite Ergebnis-Maximum des Simulationsassistenten

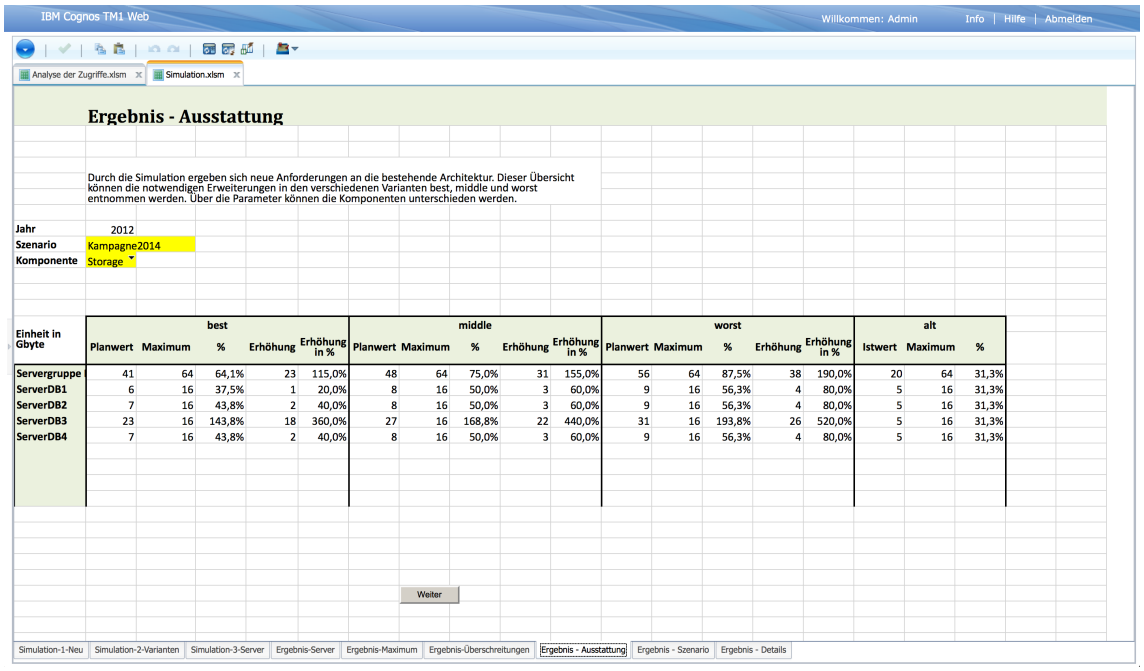
Auf der Seite Ergebnis-Maximum (Vgl. Abb. 3.90) können Sie die unmittelbaren Ergebnisse der Simulation erkennen. Für die ausgewählten Server wurde in jeder Variante für die Serverkomponenten Prozessor und Storage die maximale Auslastung in % ermittelt. Dies stellt die Basis für alle weiteren Ableitungen und Empfehlungen dar. Klicken Sie auf „Weiter“, um zur nächsten Seite der Ergebnisse zu gelangen.



Anzahl Tage	best		middle		worst	
	prozessor	storage	prozessor	storage	prozessor	storage
Servergruppe DB 1	4	264	83	592	274	646
ServerDB1	0	54	20	148	58	164
ServerDB2	1	73	24	150	86	161
ServerDB3	0	62	13	144	51	160
ServerDB4	3	75	26	150	79	161

Abbildung 3.91: Seite Ergebnis-Überschreitungen des Simulationsassistenten

Ein Ziel der Simulation ist die Überprüfung, ob die neuen Anforderungen durch die bestehende Hardware erfüllt werden können. Dies lässt sich dadurch messen, ob es Tage mit einer Überschreitung der maximalen Auslastung gibt. Die Anzahl der Tage mit Überschreitungen auf der Grundlage des ausgewählten Jahres wird in der Tabelle dargestellt. Ist der Wert für eine Komponente und einen Server und eine Variante bei 0, so liegen keine Überschreitungen in dem Jahr und in der Variante vor. Maximal ist ein Wert von 365 für Überschreitungen an jedem Tag möglich. Klicken Sie auf „Weiter“, um zur nächsten Seite zu gelangen.



IBM Cognos TM1 Web — Willkommen: Admin — Info | Hilfe | Abmelden

Analyse der Zugriffe.xlsm | Simulation.xlsm

Ergebnis - Ausstattung

Durch die Simulation ergeben sich neue Anforderungen an die bestehende Architektur. Dieser Übersicht können die notwendigen Erweiterungen in den verschiedenen Varianten best, middle und worst entnommen werden. Über die Parameter können die Komponenten unterschieden werden.

Jahr: 2012
Szenario: Kampagne2014
Komponente: Storage

Einheit in Gbyte	best					middle					worst					alt		
	Planwert	Maximum	%	Erhöhung	Erhöhung in %	Planwert	Maximum	%	Erhöhung	Erhöhung in %	Planwert	Maximum	%	Erhöhung	Erhöhung in %	Istwert	Maximum	%
Servergruppe	41	64	64,1%	23	115,0%	48	64	75,0%	31	155,0%	56	64	87,5%	38	190,0%	20	64	31,3%
ServerDB1	6	16	37,5%	1	20,0%	8	16	50,0%	3	60,0%	9	16	56,3%	4	80,0%	5	16	31,3%
ServerDB2	7	16	43,8%	2	40,0%	8	16	50,0%	3	60,0%	9	16	56,3%	4	80,0%	5	16	31,3%
ServerDB3	23	16	143,8%	18	360,0%	27	16	168,8%	22	440,0%	31	16	193,8%	26	520,0%	5	16	31,3%
ServerDB4	7	16	43,8%	2	40,0%	8	16	50,0%	3	60,0%	9	16	56,3%	4	80,0%	5	16	31,3%

Weiter

Simulation-1-Neu | Simulation-2-Varianten | Simulation-3-Server | Ergebnis-Server | Ergebnis-Maximum | Ergebnis-Überschreitungen | Ergebnis - Ausstattung | Ergebnis - Szenario | Ergebnis - Details

Abbildung 3.92: Seite Ergebnis-Ausstattung des Simulationsassistenten

Anhand dieser Ergebnisse lässt sich eine Empfehlung für die neue Ausstattung der Hardwarekomponenten ermitteln (Vgl. Abb. 3.92). Die Seite Ergebnis-Ausstattung stellt für jede Variante die empfohlene Ausstattung dar. Daneben werden in der Spalte „alt“ die aktuelle Ausstattung dargestellt. Über das gelbe Feld „Komponente“ können Sie zwischen Storage und Prozessor umschalten.

Die Tabelle zeigt für jeden Server und jede Variante den Planwert und die maximal mögliche Ausbaustufe (Maximum) des Servers an. Die Spalte „%“ stellt das Verhältnis der beiden Werte zueinander dar. Die Spalte Erhöhung sagt aus, um wie viele Einheiten der Server ausgebaut werden muss. Dies ergibt eine prozentuale Erweiterung des Systems, die in der Spalte „Erhöhung in %“ dargestellt ist. Klicken Sie auf „Weiter“, um zur nächsten Seite zu gelangen.

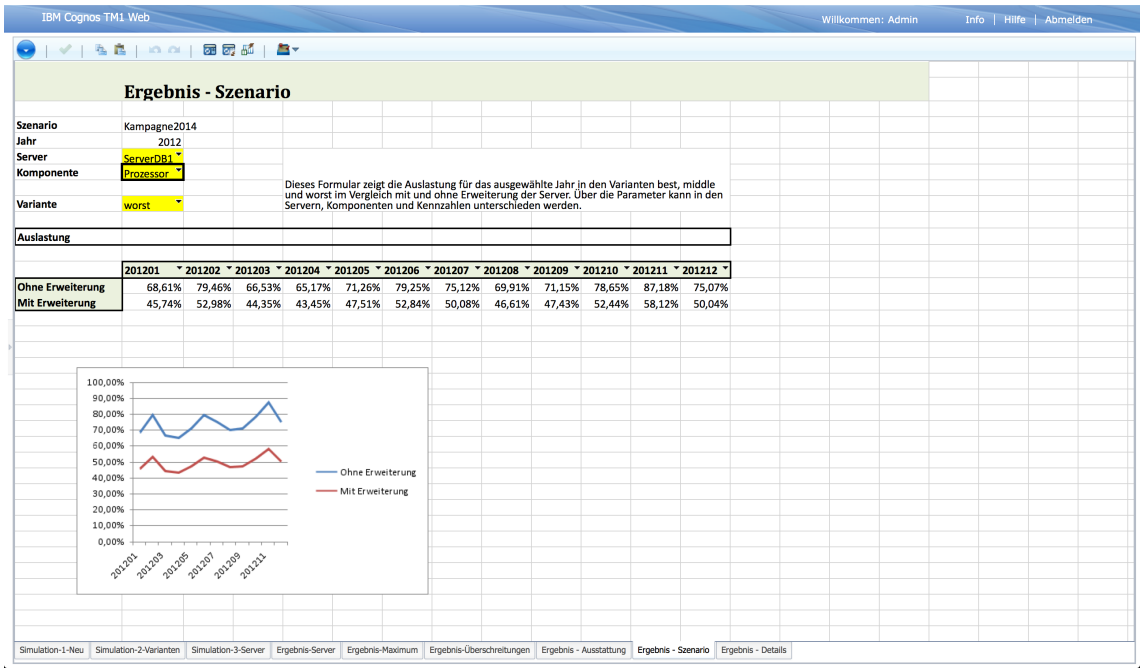


Abbildung 3.93: Seite Ergebnis-Szenario des Simulationsassistenten

Aus der Tabelle und dem Diagramm der Seite Ergebnis-Szenario können Sie entnehmen, wie sich die Auslastung einer Serverkomponente eines Servers in einer Variante mit und ohne Erweiterung des Systems darstellt (Vgl. Abb. 3.93). Über die gelben Felder über der Tabelle können Sie die verschiedenen Elemente auswählen. Klicken Sie auf „Weiter“, um zur nächsten Seite zu gelangen.

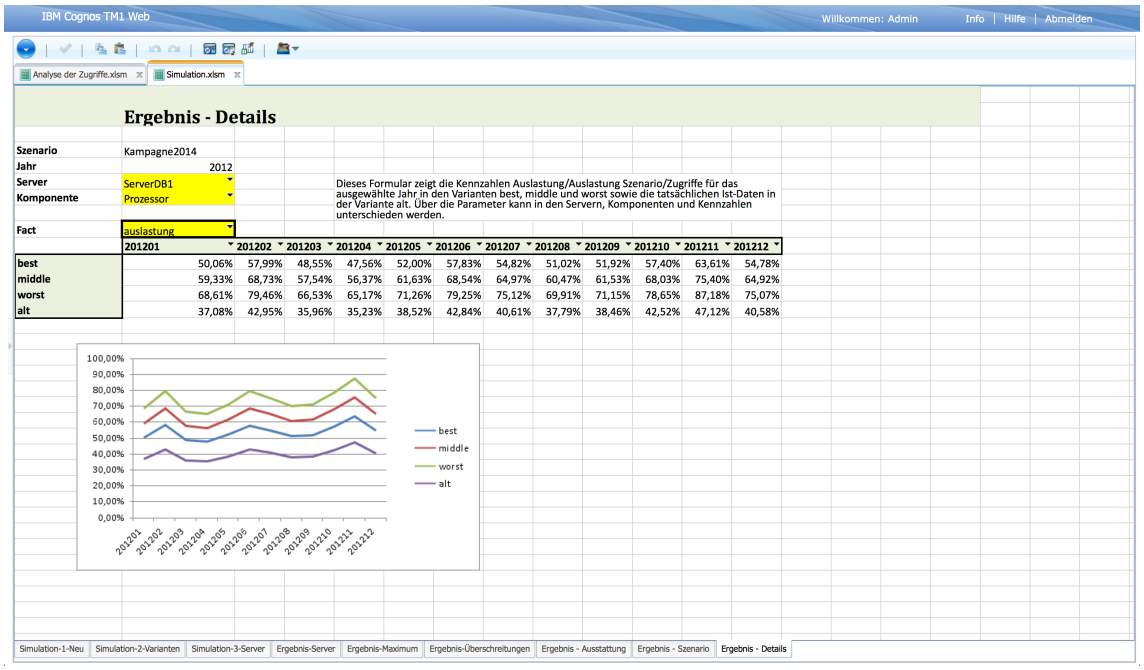


Abbildung 3.94: Seite Ergebnis-Details des Simulationsassistenten

Die letzte Seite zur Darstellung der Ergebnisse stellt die Seite Ergebnis-Details dar (Vgl. Abb. 3.94). Auch hier haben Sie wieder die Möglichkeit, die Ergebnisdarstellung durch Auswahl der Parameter in den gelben Feldern zu verändern. Sie können über die verschiedenen Kennzahlen (Fact) die Auslastung, die Zugriffe, die Auslastung nach Erweiterung des Systems (auslastung_szenario) sowie die Überschreitungen in Tagen und die maximal möglichen Zugriffe des Systems in den drei Varianten vergleichen. Das Diagramm unterhalb stellt diese Informationen grafisch dar. Die Ergebnisse Ihres Simulationsszenarios bleiben im System gespeichert und werden nur durch ein erneutes Ausführen dieses Szenarios überschrieben.

3.6.5.4 Analysieren der Ist-Daten

Zur Analyse der Ist-Daten steht der Eintrag „Analyse der Zugriffe.xlsm“ bereit. Wählen Sie im Navigationsbereich diesen Eintrag aus. Im Inhaltsbereich öffnet sich die erste Seite Zugriffe (Vgl. Abb. 3.95). Sie können über die gelben Schaltflächen oberhalb der Diagramme das Basisjahr und den Server auswählen, für die die Zugriffe pro Monat im Vergleich zur Auslastung des Arbeitsspeichers links und des Prozessors recht dargestellt werden. Unterhalb der Diagramme sehen Sie die aktuelle Ausstattung des ausgewählten Server.

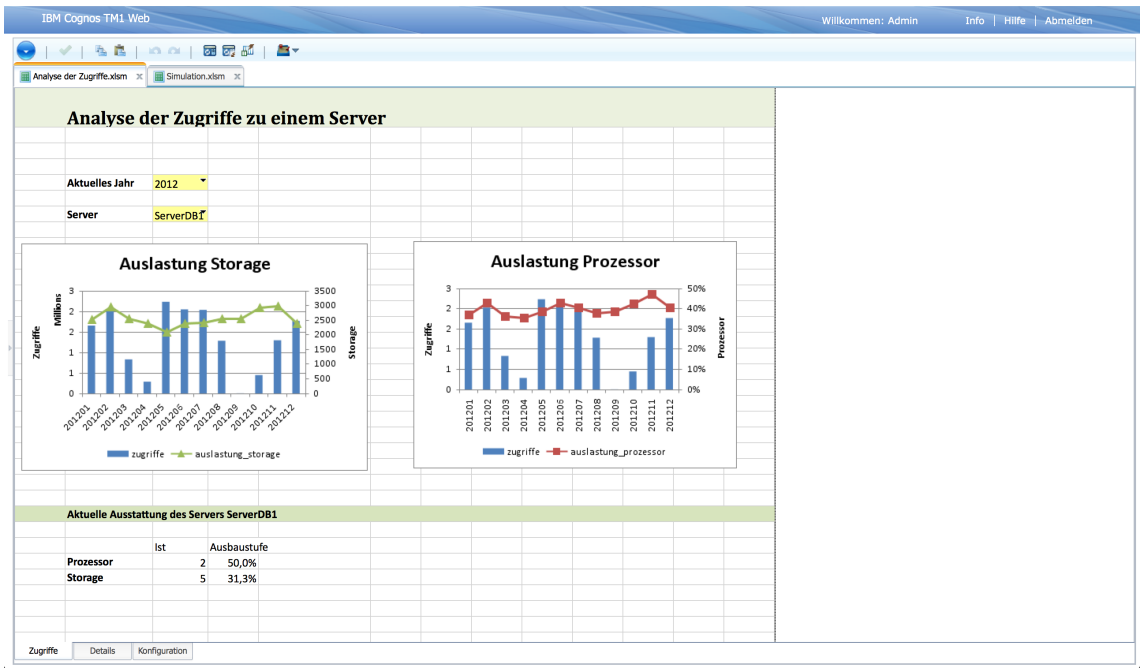


Abbildung 3.95: Seite Zugriffe der Analyse der Ist-Daten

Die Daten, die auf dieser Seite noch pro Monat aggregiert dargestellt worden waren, können Sie nun durch Auswahl der nächsten Seite „Details“ am unteren Bildschirmrand tagesgenau verfolgen (Vgl. Abb. 3.96). Sie haben hier die Möglichkeit, einen Monat und einen Server auszuwählen, für den Sie den Verlauf den Diagrammen entnehmen können.

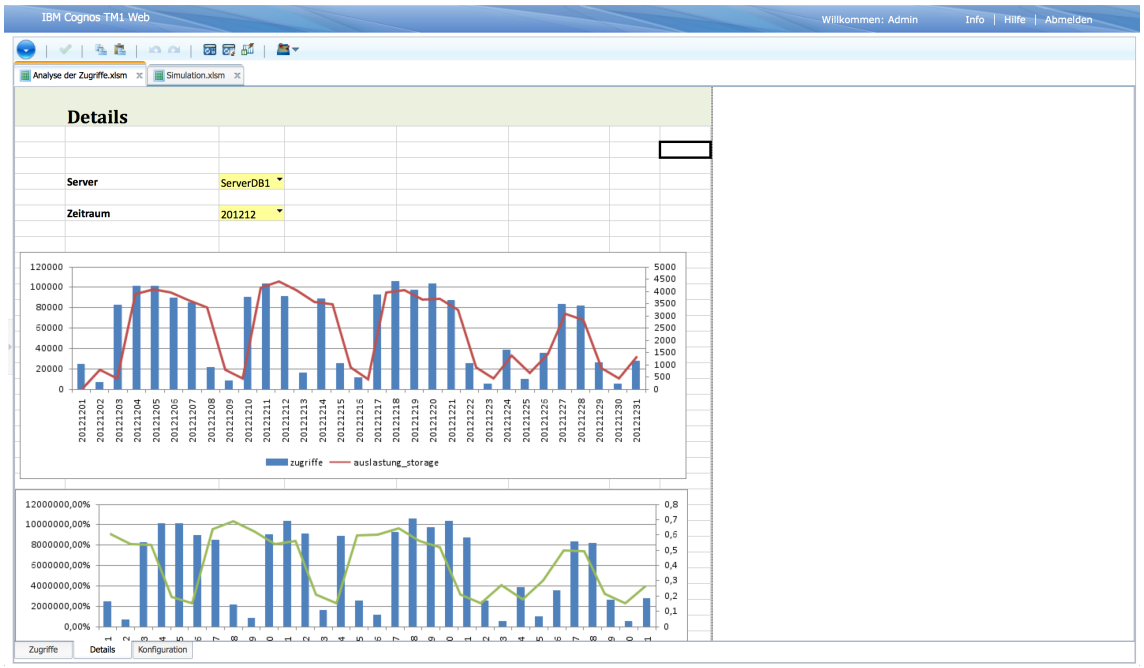


Abbildung 3.96: Seite Details der Analyse der Ist-Daten

Die vollständige aktuelle Konfiguration können Sie der Seite Konfiguration entnehmen (Vgl. Abb. 3.97). Für jeden Server sind in den Serverkomponenten Prozessor und Storage die Ist-Werte im Vergleich zu den maximal möglichen Ausbaustufen dargestellt.

Abbildung 3.97: Seite Konfiguration der Analyse der Ist-Daten

3.7 Schlussbetrachtung

Das Projekt zur Untersuchung der „Herausforderungen durch Big Data am Beispiel des IT-Capacity Managements der Volkswagen AG“ stellte verschiedene Anforderungen an die Teilprojektgruppe aus den Themenbereichen Wirtschaft und Informatik. So mussten verschiedene Disziplinen zum Aufbau eines IT-Capacity Managements durchlaufen werden. Eine besondere Bedeutung kam dem Begriff Big Data zuteil. Durch den Umstand, dass in kurzer Zeit viele Daten in unterschiedlicher Struktur verarbeitet werden mussten, ergaben sich neue Herausforderungen an die eingesetzte Software, an den Umgang mit den Daten und an die IT-Architektur. Diese Probleme und mögliche Lösungswege wurden im Kapitel 3.3 anhand von drei Referenzarchitekturen dargestellt.

Dass es daneben viele Alternativen im Umgang mit Big Data gibt, zeigt das Kapitel 3.4 mit einem Marktüberblick. Dieser umfasst 20 Anbieter mit 27 speziellen Produkten für das Thema Big Data. Die Lösungen einiger Anbieter unterscheiden sich deutlich von den vorgestellten Referenzarchitekturen. Im Rahmen der Erstellung des Marktüberblicks veränderte sich das Verständnis für Produkte und der Umgang mit der Thematik deutlich. Dies spiegelt sich darin wieder, dass entgegen der ursprünglichen Planung anstelle einer Marktanalyse nun ein Überblick erstellt wurde, der die Produkte detailliert untersucht und schließlich kategorisiert. Es stellte sich heraus, dass eine konkrete Bewertung über die Einsetzbarkeit eines Produkts nur dann erfolgen kann, wenn das Szenario und die Rahmenparameter detailliert bekannt sind. In diesem Fall empfahl sich daher ein breit aufgestellter Marktüberblick. Die Schwierigkeiten, die mit der Erstellung verbunden sind, liegen auch in der geringen Erfahrung der Projektgruppenmitglieder. Sowohl der Umgang als auch die Einrichtung von BI-Lösungen konnte bisher nur wenig geprobt werden. Sowohl der zeitliche als auch der finanzielle Rahmen dieses Projekts gaben nicht die Möglichkeit, jedes Produkt selbst zu installieren und auszuprobieren. Daher musste sich die Projektgruppe in vielen Fällen auf das Urteil von Sekundärliteratur, Herstellerangaben oder Expertenmeinungen verlassen, wie sie im BARC-Guide zu finden sind.

Bevor die Daten im IT-Capacity Management analysiert werden konnten, mussten sie zunächst erhoben werden. Beispielhaft wurde ein ETL-Prozess zur Integration von Access Logs in der bestehenden BI-Architektur der Volkswagen AG erstellt. Da es zeitliche Verzögerung in der Bereitstellung der Softwareumgebung gab, konnte in der Zwischenzeit ein Szenario der konzipierten Referenzarchitekturen mit einem im Marktüberblick untersuchten Produkt exemplarisch umgesetzt werden. Die Hadoop-Software von Hortonworks diente als Grundlage, um den ETL-Prozess abzubilden. Anhand dieser Software wurde eine Cluster-Installation ausprobiert. Eine Erkenntnis aus diesem Versuch ist, dass der Installations- und Konfigurationsaufwand für eine Cluster-Lösung deutlich über dem einer einfachen Single-Node-Installation liegt. Schließlich konnte der Prototyp zur Datenintegration aber noch auf der IT-Architektur der Volkswagen AG umgesetzt werden. Die erstellten Reports zeigen, dass die Ergebnisse des Prozesses durchaus als Grundlage für die nächste Aufgabe - die Erstellung des Prototypen zur Simulation - dienen können.

Die Simulation wurde schließlich mit der Software IBM Cognos TM1 auf einem von der Universität bereit gestellten System erstellt. Sie erfüllt die Anforderungen aus dem Lastenheft nahezu vollständig. Anhand dieses Prototypen ist es möglich, eine Weiterentwicklung und den produktiven Einsatz zu evaluieren.

Im Rahmen dieses Projekts konnten die Mitglieder der Teilprojektgruppe viele Erfahrungen über die Bearbeitung der Aufgaben hinaus machen. Die Teilnahme an der BI-Konferenz sowie die verschiedenen Meetings und Telefonkonferenzen mit den Mitarbeitern des BI-CC der Volkswagen AG vermittelten einen guten ersten Eindruck über die Unternehmensphilosophie. Auch die strukturierte Vorgehensweise zur Untersuchung der Chancen und Risiken des Hype-Themas Big Data beeindruckte.

Die Durchführung des Projekts unterlag den Einflüssen von vielen Änderungen der Zielvorstellungen und der Interpretation der Aufgaben. Zudem wurde zu wenig Zeit für die Einrichtung der Entwicklungsumgebungen und für die Installation der benötigten Software eingeplant, sodass es im Projektverlauf zu Verzögerungen kam. Dennoch konnten alle Aufgaben der Anforderungsdefinition vollständig erfüllt oder mindestens teilweise umgesetzt werden.

Ziel der Projektgruppe war zum einen die gestellten Aufgaben zu erfüllen, aber zum anderen deutlich zu machen, dass das Thema Big Data ganz neue Herausforderungen mit sich bringt und alternative Lösungswege bedingt. In diesem Projekt konnten aktuelle Software-Lösungen ausprobiert werden, die moderne Möglichkeiten in der Datenanalyse und Skalierbarkeit bieten. Die Ergebnisse dieser Teilprojektgruppe sind ein weiterer wissenschaftlicher Beitrag zur Aufklärung des Begriffs Big Data. Der Bericht soll der Volkswagen AG helfen das Verständnis für das Thema Big Data zu erhöhen. Das eine nüchterne Einschätzung nicht ganz einfach ist, zeigt auch der Gartner Hype Cycle aus dem Jahr 2013, der das Thema Big Data auf dem Gipfel der überzogenen Erwartungen sieht (Vgl. Gartner, Inc. 2013a).

4 Teilgruppe Smart Metering

Kapitelübersicht

4.1 Projektdefinition . . . 205
4.1.1 Praxispartner . . . 205
4.1.2 Aufgabenstellung . . . 205
4.2 Projektorganisation . . . 206
4.2.1 Organisationsstruktur . . . 206
4.2.2 Projektplan . . . 207
4.2.3 Projektverlauf . . . 208
4.3 Anforderungsspezifikation . . . 211
4.3.1 Unternehmensbeschreibung der CX4U AG . . . 211
4.3.1.1 CX4U Solutions GmbH & Co. KG . . . 212
4.3.1.2 CX4U Consulting GmbH & Co. KG . . . 212
4.3.1.3 CX4U Events GmbH & Co. KG . . . 213
4.3.2 Zweck der Software . . . 213
4.3.2.1 Ausgangssituation . . . 213
4.3.2.2 Zielbestimmung . . . 214
4.3.2.3 Technische Produktumgebung . . . 215
4.3.2.3.1 SAP-System der CX4U-AG . . . 215
4.3.2.3.2 SAP HANA des HPI . . . 215
4.3.3 Stakeholder . . . 216
4.3.4 Geschäftsanwendungsfälle . . . 216
4.3.4.1 Endkundenabrechnung . . . 217
4.3.4.2 Verbrauchsprognose . . . 217
4.3.5 Systemanwendungsfälle . . . 217
4.3.5.1 Akteure . . . 217
4.3.5.2 Validierung . . . 219
4.3.5.3 Estimation . . . 222
4.3.5.4 Regelauswahl . . . 226
4.3.5.5 Versionierung . . . 229
4.3.5.6 Editing . . . 231
4.3.6 Systemanforderungen . . . 233
4.4 Systementwurf . . . 234
4.4.1 Systemzerlegung . . . 234
4.4.2 Hard- und Softwarekomponenten . . . 234
4.4.3 VPN-Verbindung zwischen den Systemen . . . 235

4.4.4 Dynamisches Modell 235
4.4.4.1 HTTP-Request 235
4.4.4.2 VEE-Framework 236
4.4.5 Fachliches Datenmodell 238
4.4.6 Sicherheitsaspekte und Zugriffsrechte 238
4.5 Machbarkeitsstudie 240
4.5.1 Aktueller Stand 240
4.5.2 Beschreibung der Umsetzungsalternativen 241
4.5.2.1 SAP Sybase Event Stream Processor 242
4.5.2.2 SAP HANA Smart Data Access 242
4.5.2.3 SAP HANA Replication Technologies 243
4.5.2.3.1 ETL-basierte Replikation 244
4.5.2.3.2 Extraktor-basierte Replikation 245
4.5.2.3.3 Log-basierte Replikation 245
4.5.2.4 SAP HANA Extended Application Services 245
4.5.2.4.1 Vorteile 245
4.5.2.4.2 Produktübersicht 246
4.5.2.5 Secondary Database Connection 246
4.5.2.5.1 Vorteile 247
4.5.2.5.2 Produktübersicht 247
4.5.2.5.3 Systemvorraussetzung 249
4.5.2.5.4 Umsetzung 249
4.5.2.6 SAP Landscape Transformation Replication Server (SLT) 250
4.5.2.6.1 Vorteile 250
4.5.2.6.2 Produktübersicht 250
4.5.2.6.3 Beschreibung der allgemeinen Architektur . . 251
4.5.2.6.4 Multi System Support 253
4.5.2.6.5 Grundlegende Konzepte der Trigger-basierten Datenreplikation mit SAP LT Replication Server für SAP HANA 253
4.5.2.6.6 Configuration and Monitoring Dashboard (Transaktion LTR) 253
4.5.2.6.7 SAP SLT Replication Server Jobs 255
4.5.2.6.8 Erweiterte Replikationseinstellungen 255
4.5.2.6.9 Zusätzliche Funktionen 256
4.5.3 Bewertung der Umsetzungsalternativen 256
4.5.4 Prototypen 261
4.5.4.1 Prototyp: HANA 261

4.5.4.1.1 Datenbank 261
4.5.4.1.2 Webinterface 262
4.5.4.1.3 OData REST Service 262
4.5.4.1.4 Validation Funktion 262
4.5.4.1.5 Estimation Funktion 264
4.5.4.1.6 Editing Funktion 266
4.5.4.2 HTTP-Request von Netweaver an HANA 268
4.5.4.2.1 Konfiguration der HTTP-Destination 268
4.5.4.2.2 Implementierung des HTTP-Requests 270
4.5.4.2.3 Ausgabe der HTTP-Response Daten 273
4.5.5 Schlussbetrachtung 274
4.6 Qualitätsmanagement 276
4.6.1 QM-Maßnahmen 276
4.6.1.1 Vorgehensmodelle 279
4.6.1.1.1 V-Modell XT 279
4.6.1.1.2 Wasserfallmodell 279
4.6.1.1.3 Spiralmodell 280
4.6.1.1.4 Prototyping 280
4.6.1.1.5 Unified Process 281
4.6.1.2 Testen von Software 282
4.6.1.3 Anwendung 283
4.6.2 Funktionale Testfälle 283
4.7 Öffentlichkeitsarbeit 285
4.7.1 Future SOC Day 285
4.7.2 CeBIT 285
4.8 Schlussbetrachtung 287

4.1 Projektdefinition

Im Rahmen der Projektdefinition werden zunächst die Praxispartner vorgestellt und anschließend die grobe Aufgabenstellung des Themas Prozessorientierte Business Intelligence mit SAP HANA beschrieben. Dieses soll ein grundlegendes Verständnis für die darauffolgenden Kapitel vermitteln.

4.1.1 Praxispartner

Die Teilgruppe Smart Metering[32] arbeitet während des Projektes mit der CX4U AG aus Oldenburg, an dem Thema Prozessorientierte Business Intelligence mit SAP HANA, zusammen. Die CX4U AG ist ein national und international tätiges IT-Beratungsunternehmen mit Konzentration auf den Gas- und Strommarkt. Dabei werden individuelle Softwarelösungen im SAP Umfeld für die Kunden entwickelt. Die CX4U AG liefert der Teilgruppe Smart Metering Fragestellungen aus der Praxis und stellt die zur Beantwortung der Fragestellungen hilfreiche Hardware- und Softwaresysteme bereit. Außerdem arbeitet die Teilgruppe Smart Metering eng mit dem Hasso Plattner Institut (HPI) aus Potsdam zusammen, welches zur Förderung des Austauschs zwischen der Forschungsgemeinschaft und der Industrie ein SAP HANA System bereitstellt. Im folgenden Kapitel soll daher die Aufgabenstellung, unter Betrachtung der verfügbaren Systeme, näher beschrieben werden.

4.1.2 Aufgabenstellung

Durch die Gesetzgebung ändert sich der Energiemarkt bis 2020, sodass in den meisten Haushalten Smart Meter[33] installiert sein werden. Deshalb müssen die IT-Systeme der Energieunternehmen angepasst werden, um mit der wachsenden Menge an Messdaten umgehen zu können. Damit die Echtzeitdatenerfassung der Energielieferanten am besten eingesetzt werden kann, müssen die Daten dauerhaft konsistent, aktuell und fehlerfrei sein. Die Teilgruppe Smart Metering versucht daher, in Zusammenarbeit mit der CX4U AG, die Integrität und Qualität der Daten zu gewährleisten, indem ein VEE-Framework (Validation, Estimation, Editing) implementiert wird. Die Herausforderung der Implementierung besteht darin, dass die Daten innerhalb des VEE-Frameworks zwischen einem SAP Netweaver System und einem SAP HANA System ausgetauscht werden müssen.

In den folgenden Kapiteln werden die Ergebnisse der Teilprojektgruppe Smart Metering dargestellt. Zunächst erfolgt die Beschreibung der Projektorganisation. Eine ausführliche Beschreibung der Aufgaben- und Problemstellung erfolgt innerhalb der Anforderungsspezifikationen.

[32] Smart Metering als Bezug zum Anwendungsszenario und zur Abgrenzung zur Teilgruppe Big Data im IT-Capacity Management

[33] Ein Smart Meter oder auch intelligenter Zähler, ist ein Zähler der den Storm- oder Gasverbrauch misst. Ein Smart Meter zeigt den tatsächlichen Energieverbrauch und die tatsächliche Nutzungszeit an, welche direkt an den Energieversorger übermittelt werden.

4.2 Projektorganisation

Nachdem bereits der grobe Rahmen der Projektgruppe skizziert wurde, soll in den folgenden Kapiteln die Organisationsstruktur der Teilgruppe näher beschrieben werden. Dazu wird zunächst die teilgruppenspezifische Aufteilung erläutert, um anschließend die wesentlichen Phasen des Projektplans zu beschreiben. Abschließend wird der Projektverlauf skizziert, um die wesentlichen Meilensteine des Projekts und deren Einfluss auf das Endergebnis hervorzuheben.

4.2.1 Organisationsstruktur

Innerhalb der Teilgruppe Smart Metering wurde jedem Projektgruppenmitglied eine unterschiedliche Rolle in der Organisationsstruktur zugewiesen. Die Projektrolle gibt an, für welchen Aufgabenbereich das Projektmitglied verantwortlich ist. Das bedeutet nicht, dass dieser Aufgabenreich von dem Projektmitglied alleine gelöst werden soll, sondern dass das Projekt für diesen Teilbereich von dem zuständigen Projektmitglied vorangetrieben und koordiniert wird.

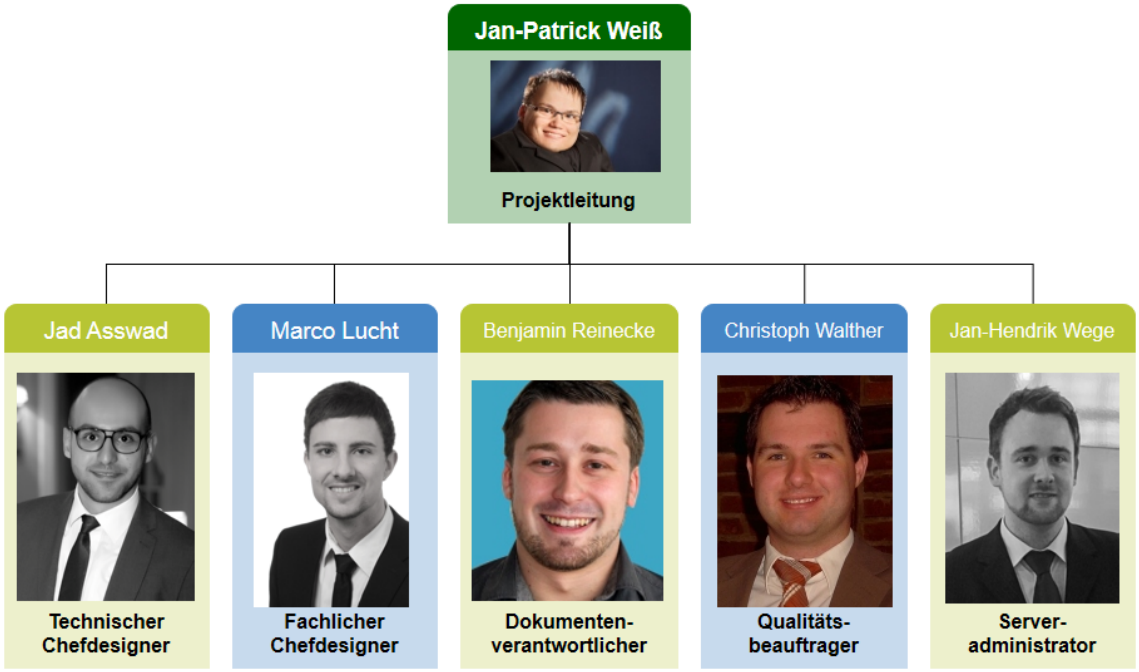


Abbildung 4.1: Organisationsstruktur der Teilgruppe Smart Metering

Die Projektrollen Projektleitung, Dokumentenverantwortlicher und Serveradministration wurden von der Projektgruppe HAWK in die Teilprojektgruppe Smart Metering übernommen, da diese Projektmitglieder mit ihren Funktionen bereits ausgelastet sind und diese ebenfalls Anwendung auf die Teilgruppe Smart Metering finden. Die übrigen Projektrollen sind durch folgende Eigenschaften charakterisiert (Vgl. Schekelmann 2011, S. 9):

Fachlicher Chefdesigner

- ist verantwortlich für die Spezifikation, dazu gehört sowohl die Methodik als auch der Inhalt, zum Aufbau der fachlichen Architektur.

- ist verantwortlich für das Fachkonzept.

Technischer Chefdesigner

- ist verantwortlich für die technische Architektur der Anwendung.
- ist verantwortlich für das DV-Konzept.

Qualitätsbeauftragter

- ist verantwortlich für die Planung und Durchführung von Qualitätssicherungsmaßnahmen.
- ist verantwortlich für den Qualitätsmanagementplan und die Testfälle.

Die Projektrolle des „Entwicklers“, „Spezifizierers“ und „Testers“ wird von jedem Projektmitglied eingenommen; sie sind daher nicht in der Abbildung aufgeführt. Innerhalb dieser Rollen werden die Bestandteile des Softwaresystems erstellt. Zusätzlich wurde eine Kompetenzmatrix erstellt, wodurch für jedes Projektmitglied transparent ist, über welche Fähigkeiten die anderen Projektmitglieder verfügen und wer gegebenenfalls ein Ansprechpartner sein könnte (Vgl. Anhang Abb. F.1).

4.2.2 Projektplan

Die dargestellte Organisationsstruktur wird durch einen Projektplan koordiniert. Dabei wird das Projekt in vier wesentliche Phasen unterteilt. Zu diesen zählen die Seminararbeitsphase, die Konzeptionsphase, die Umsetzungsphase und die Dokumentationsphase. Während der Seminarphase werden die fachlichen Grundlagen erarbeitet. Innerhalb der Konzeptionsphase werden die Anforderungsspezifikation und der Systementwurf erstellt. Basierend auf diesen Dokumenten wird ein Prototyp erstellt, welcher in der Machbarkeitsstudie bewertet wird. Abschließend werden die Ergebnisse in der Dokumentationsphase aufbereitet und festgehalten. Während dieser Projektphasen werden die Ergebnisse regelmäßig mit der CX4U AG besprochen und der Öffentlichkeit im Rahmen des FSOC-Days[34] am HPI und der CeBIT in Hannover vorgestellt. Ein umfangreicher Projektplan liegt als Redmine-Export im Anhang vor (Vgl. Anhang Abb. F.2). Zur besseren Veranschaulichung wird in der folgenden Abbildung der Soll- und Ist-Projektverlauf grafisch aufbereitet.
Die blauen Markierungen repräsentieren den geplanten Projektverlauf und die grünen Markierungen den tatsächlichen Projektverlauf. Die gelben Markierungen stehen für weitere Aufgaben, wie beispielsweise die Vorbereitungen für die CeBIT. Im unteren Teil des Projektplans sind wesentliche Meilensteingespräche mit der CX4U AG und dem HPI aufgeführt. Der Projektverlauf wird im folgenden analysiert, um die zeitlichen Überschreitung der Projektphasen zu skizzieren und damit zu erklären.

[34] Der FSOC-Day ist eine Veranstaltung des HPI, an dem Forschungsergebnisse, aus der Zusammenarbeit zwischen Forschungseinrichtungen und Industriepartnern, präsentiert werden. Näheres dazu im Kap. 4.7 Öffentlichkeitsarbeit.

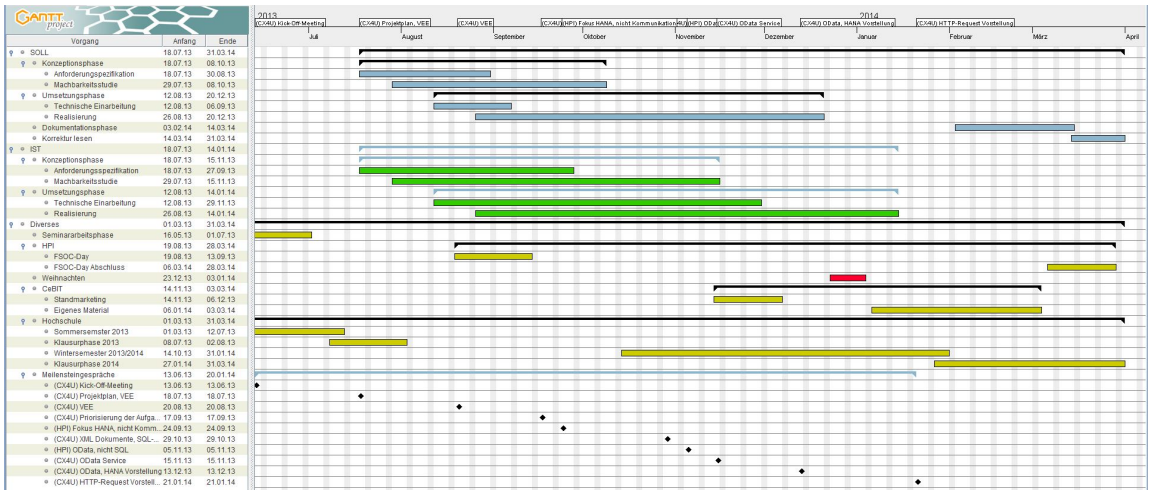


Abbildung 4.2: Projektplan der Teilgruppe Smart Metering Soll-/ Ist-Vergleich

4.2.3 Projektverlauf

Das erste Treffen zwischen der Projektgruppe und dem Unternehmen CX4U AG fand am 11. April 2013 statt. Zunächst stellte sich die CX4U AG vor und skizzierte anschließend grob den Aufgabenbereich. Das initiale Kick-Off-Meeting mit der CX4U AG sollte erfolgen, sobald die Teilnahme des zweiten Praxispartners offiziell bestätigt wird. Erst dann war klar, welche Projektgruppenmitglieder mit der CX4U AG zusammenarbeiten können.

Am 16. Mai 2013 erfolgte die Zusage des zweiten Praxispartners. Die Projektgruppe wurde aufgeteilt und die Teilgruppe Smart Metering hat die Terminfindung mit der CX4U AG für das Kick-Off-Meeting gestartet. Aufgrund von Urlaubszeiten seitens der CX4U AG konnte der Termin erst auf den 13. Juni 2013 gelegt werden.

Am 13. Juni 2013 wurde der Teilgruppe Smart Metering die Aufgabenstellung detaillierter vorgestellt. Da das Konzept VEE sehr komplex ist und nicht während einer Projektbesprechung komplett verstanden werden kann, wird die Teilgruppe Smart Metering seitens der CX4U AG darum gebeten, zur nächsten Projektbesprechung das bisherige Verständnis von VEE zu beschreiben. Solche Ergebnisse sollen auf 1 - 2 Folien als Kurzpräsentation festgehalten werden. Des Weiteren sollte ein Zeitplan mit den wesentlichen Meilensteinen und die Organisationsstruktur aufgestellt werden. Der Arbeitsaufwand für die nächste Projektbesprechung am 18. Juli 2013 wurde bewusst gering gewählt, da sich die Projektgruppenmitglieder zur gleichen Zeit in den Klausurvorbereitungen und dem Abschluss der Seminararbeiten befanden. Die Vorschläge für Seminararbeitsthemen der CX4U AG konnten daher nicht mehr berücksichtigt werden. Am 18. Juli 2013 wurde der Projektplan und das Vorgehensmodell verabschiedet. In dieser Projektbesprechung wurde deutlich, dass VEE ein Theoriekonzept bei einigen Mitarbeitern der CX4U AG und der SAP ist. Die Aufgabe sollte es daher sein, ein VEE-Framework auf Basis von SAP HANA zu entwickeln. Das bedeutet für die Teilprojektgruppe Smart Metering, dass vor der Erstellung des Pflichtenhefts zunächst eine Machbarkeitsstudie durchgeführt werden

musste. Die CX4U AG gab bekannt, dass für das Projekt Zugänge zum SmartGateway bereitgestellt werden. Die Teilgruppe Smart Metering sollte zur nächsten Projektbesprechung den VEE-Prozess skizzieren. Sobald dieser verabschiedet war, konnte mit der Machbarkeitsstudie begonnen werden. Neben der Projektarbeit für die CX4U AG, begann die Phase der Erstellung der Inhalte für den FSOC-Day am HPI.

In der Projektbesprechung am 20. August 2013 wurde das Verständnis der Teilgruppe Smart Metering über die Aufgabenstellung, verabschiedet. Um mit der Machbarkeitsstudie und dem Pflichtenheft beginnen zu können, musste zunächst Zugang zum SmartGateway bestehen, da dort die Beispieldaten vorhanden sind. Währenddessen wurde der Folgeantrag, das Poster und das Paper für den FSOC-Day am HPI vorbereitet. Darüber hinaus wurde der Antrag für die CeBIT erstellt.

In der Projektbesprechung am 17. September 2013 hat die CX4U AG der Projektgruppe Ihre Priorisierung der Aufgabenpakete mitgeteilt. So sollte zunächst innerhalb einer Machbarkeitsstudie die Kommunikation zwischen SAP HANA und dem SmartGateway geprüft werden. Daraufhin sollte das VEE-Konzept in einem ABAP-Modul entwickelt und abschließend das VEE-Konzept mit SAP HANA umgesetzt werden. Am 24. September 2013 ist die Teilgruppe Smart Metering mit diesen Informationen zum FSOC-Day nach Potsdam gefahren. Die Mitarbeiter vom HPI und SAP sehen nicht die Notwendigkeit einer Machbarkeitsstudie für die Kommunikation, da beide Produkte von SAP stammen und daher definitiv miteinander kommunizieren können. Außerdem sollte, aufgrund des wissenschaftlichen Anspruchs, der Schwerpunkt auf die Entwicklung des VEE-Konzept in SAP HANA gelegt werden. Im Nachgespräch mit der CX4U AG wurde festgehalten, dass dennoch eine Machbarkeitsstudie erfolgen soll. Die nötigen Werkzeuge für die Implementierungsarbeiten sollten in den folgenden Wochen bereitgestellt werden. Ein Gespräch mit Prof. Marx Gómez am 10. Oktober 2013 sollte den Vorgang des Zugangs zu den Systemen beschleunigen.

Am 15. Oktober 2013 wurden die Zugänge zu dem SAP-System mit den benötigten Entwicklerschlüsseln übersendet. Nach weiterer Einarbeitung in die Systeme sollten technische Fragen und Fragen bezüglich der beispielhaft zugesandten XML-Dokumente in einer weiteren Projektbesprechung am 29. Oktober 2013 geklärt werden. Der CX4U AG wurde die Verbindung zwischen dem Netweaver und SAP HANA mittels einem nativen SQL Statement gezeigt. Als Ergebnis wurde festgehalten, dass geprüft werden soll, ob die Verbindung mit openSQL innerhalb eines ABAP-Moduls aufgebaut werden kann. Weiterhin wurde festgestellt, dass die Teilgruppe Smart Metering zwar Zugriff auf die Systeme hat, jedoch nicht erkennbar ist, wie auf die relevanten Daten zugegriffen werden kann. Daher sollte von der CX4U AG eine E-Mail folgen, in der die nötigen Funktionsaufrufe der Teilgruppe Smart Metering mitgeteilt werden.

Mit den neuen Informationen aus der Projektbesprechung wurde am 5. November 2013 an einem weiteren HPI Call teilgenommen. Dort wurde der Teilgruppe Smart Metering mitgeteilt, dass es sowohl unüblich ist eine Verbindung mittels nativem SQL aufzubauen als auch mit openSQL innerhalb eines ABAP-Moduls. Um die Vorteile von SAP HANA nutzen zu können, sollte die

XS-Engine von SAP HANA genutzt und somit ein Service bereitgestellt werden, der von dem SmartGateway genutzt werden kann. Darauf wurde von der Teilgruppe Smart Metering eine mögliche Umsetzung als OData-Service geprüft und der CX4U AG in einer Projektbesprechung am 15. November 2013 vorgestellt. Nach ersten Einschätzungen der CX4U AG wird dieses Vorgehen als sinnvoll erachtet und konnte somit umgesetzt werden. Durch die vorangeschrittene Projektzeit sollte zunächst das VEE-Konzept in SAP HANA implementiert und ein Service nach außen bereitgestellt werden. In der nächsten Projektphase sollte der Service in einem ABAP-Modul integriert und die Regelauswahl im SmartGateway umgesetzt werden. In der Projektbesprechung wurden nochmals die noch fehlenden Funktionsaufrufe angesprochen und von Herrn Nannen am folgenden Tag in einem programmierten Report bereitgestellt.

Am 13. Dezember 2013 wurden die Ergebnisse der Implementierungsarbeiten der Teilgruppe Smart Metering vorgestellt. Innerhalb von SAP HANA wurde eine OData Schnittstelle implementiert, die vom Netweaver für den Datenaustausch genutzt werden soll. Das VEE-Konzept wurde teilweise bereits in SAP HANA umgesetzt und verfügt über fortgeschrittene Logik, wie beispielsweise die Korrektur der Zeitstempel. Daher konnte die CX4U AG bereits Verbesserungstipps geben. Im Netweaver wurden beispielhaft Variablen zur Abfrage der Daten und eine Datenbank zur Speicherung der Regeln angelegt. Außerdem wurde versucht, die OData Schnittstelle zu integrieren. Jedoch konnte dieses nicht erfolgreich getestet werden, da entweder die Systeme der CX4U AG oder die des HPIs nicht erreichbar waren. Das Pflichtenheft wurde von der CX4U AG gelesen und es gibt keine Anmerkungen, sodass die Implementierungen fortgesetzt werden können. Der CX4U AG wurde der Ablauf der CeBIT beschrieben. Anfang 2014 sollten weitere Abstimmungen zwischen den CeBIT Inhalten, welche die Teilgruppe Smart Metering erarbeitet, und der CX4U AG stattfinden

Die Teilgruppe Smart Metering kann den Beweis erbringen, dass der Netweaver einen http-Request auf die OData-Schnittstelle von HANA ausführen kann, ohne dass zusätzliche Software oder Bibliotheken benötigt werden. Der Quellcode wurde der CX4U AG am 21. Januar 2014 vorgestellt. Außerdem wurde an diesem Termin das weitere Vorgehen bezüglich der CeBIT besprochen. Im Januar sollten weitere kleine Implementierungen in HANA und im Netweaver vorgenommen werden. Das Ziel, eine Kommunikation zwischen dem Netweaver und HANA aufzubauen, ohne zusätzliche Produkte einzusetzen, wurde jedoch bereits erfüllt. Der Februar und März sollte zur abschließenden Dokumentation, zur Vorbereitung der CeBIT und des FSOC-Days genutzt werden. Am 7. Februar 2014 findet ein abschließendes Feedback-Gespräch statt, in dem ein Fazit zu den vergangenen Monaten gezogen wurde. In der Woche vom 10. bis 14. März 2014 hat die Projektgruppe ihre Ergebnisse auf dem Niedersachsen-Stand der CeBIT präsentiert. Eine weitere Präsentation der Ergebnisse erfolgt am 9. April 2014 am HPI im Rahmen des FSOC-Days und am 9. Mai 2014 in der Universität Oldenburg im Rahmen einer Abschlusspräsentation.

4.3 Anforderungsspezifikation

Dieses Dokument beschreibt im weiteren Sinne die Hintergründe und Eigenschaften des betrachteten Projektes, sowie deren charakteristischen Merkmale hinsichtlich Konzeption und Umsetzung. Es wird dabei zunächst sowohl der Praxispartner der Teilgruppe, als auch der Sinn und Zweck der für diesen Partner entwickelten Lösung beschrieben. Anschließend werden die mit diesem Projekt adressierten Stakeholder und Geschäftsanwendungsfälle aufgeführt und erläutert, bevor in der Folge einzelne Systemanwendungsfälle, welche im Rahmen eines konzeptionellen VEE-Frameworks erarbeitet wurden, diskutiert werden. Den Abschluss bilden die aus dieser Diskussion entstandenen Anforderungen an das System bzw. an seine Umgebung.

4.3.1 Unternehmensbeschreibung der CX4U AG

Die CX4U AG ist ein national sowie international tätiges IT-Beratungsunternehmen, dessen Hauptaugenmerk auf der Bedienung des jeweiligen Gas-, mit zunehmender Bedeutung jedoch auch des Strommarktes, liegt. Seine Wurzeln hat das Unternehmen in der 2001 gegründeten corporate practice for utilities GbR aus der nach einem Rechtsformwechsel im Jahr 2004 zunächst die corporate practice for utilities GmbH & Co. KG und später die Schwestergesellschaften corporate service for utilities GmbH & Co. KG sowie corporate events for utilities & Co. KG hervorgingen. Einige Zeit später wurden jene drei Gesellschaften der neu gegründeten Holdinggesellschaft CX4U AG zugeordnet und dieser eingegliedert (CX4U (2013a), Vgl. Abb. 4.3).

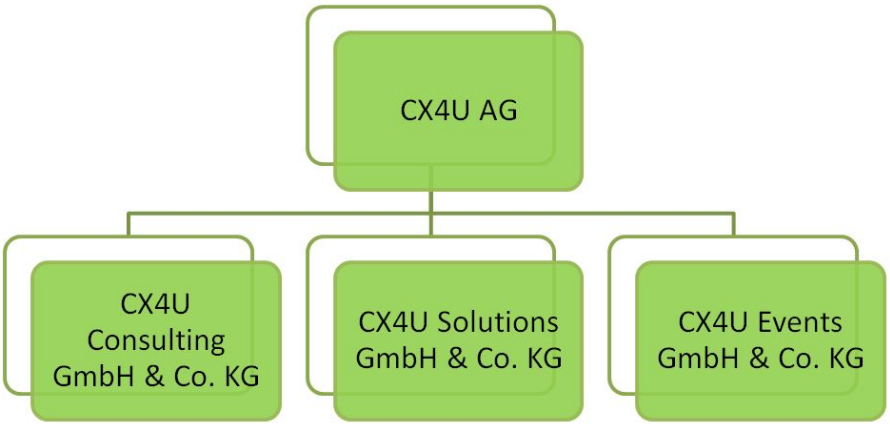


Abbildung 4.3: Unternehmensgliederung

Da die Projektgruppe im Rahmen der Aufgabenbewältigung hauptsächlich mit der CX4U Solutions GmbH & Co. KG zusammen arbeiten wird, soll sich an dieser Stelle auf eben diesen Bereich fokussiert und alle weiteren Unternehmensteile eher gekürzt beschrieben werden.

4.3.1.1 CX4U Solutions GmbH & Co. KG

Die Hauptaufgabe dieser Unternehmenssparte liegt in der Umsetzung projektindividueller Anwendungen und Technologien basierend auf unterschiedlichen Entwicklungsplattformen. Die Arbeit ist dabei seit jeher von einer engen Kooperation mit der SAP AG, genauer mit der IBU Utilities, geprägt und führte auf Grund der Erfahrungen aus verschiedenen Projekten im Laufe der Zeit zu einer sehr hohen Kompetenz in der Fachgebieten SAP IS-U/EDM, SAP BOBJ sowie Smart Metering und Cloud-Technologien. Seit dem Jahr 2007 ist diese Partnerschaft im Bereich Gas durch ein international anerkanntes Zertifikat beschrieben.
Das Portfolio der IT-Tochter Solutions umfasst vier grundlegende Anwendungen, welche sich hinsichtlich des Einsatzgebietes voneinander unterscheiden und zusammengenommen einen Großteil der Anforderungen im Strom- und Gasmanagement abdecken (CX4U (2013b)).

1. **coregas for DSO:** Dieses Produkt unterstützt den Nutzer bei der Bilanzierung von einzelnen Gasübernahmestationen, indem es eine größtenteils automatisierte Abbildung der Standardlastprofile sowie die Darstellung der physikalischen Eigenschaften der Stationen einer integrierten Ansicht ermöglicht. Darüber hinaus werden auch hinreichende Monitoringanforderungen wie Allokationsclearing und Deklarationsclearing unterstützt.

2. **coregas for TSO:** Hierbei handelt es sich um eine Lösung für das Kapazitäts- bzw. Normierungsmanagement eines Netzbetreibers hinsichtlich des Prozesses des Netzzugangs. Dabei werden diverse Kapazitätsverträge in einer integrierten Sicht abgebildet und folglich von der Buchung bis zur Abrechnung sichtbar gemacht. Ebenfalls sind das Auslesen spezifischer Status sowie das Setzen von Schwellenwerten möglich.

3. **coregas for Supplier:** Diese Softwarelösung bietet aus Sicht eines Gaslieferanten eine integrierte Möglichkeit den Prozess des Bilanzkreismanagements auszuführen. Dem Nutzer wird hierzu zum einen eine detaillierte Darstellung der hierarchischen Struktur vom Lieferanten bis zum Stadtwerk angeboten und zum anderen werden die im europäischen Raum üblichen Standardlastprofile automatisiert zur Verfügung gestellt. Diese können etwa im Rahmen der Schattenbilanzierung der einzelnen Bilanzkreise Verwendung finden.

4. **Smart Gateway:** Mit dem Smart Gateway ist es nutzerseitig möglich, unterschiedlichste Daten der intelligenten Zähler auszulesen und im Sinne des Geräte- oder Zeitreihenmanagements sowie des Customer Service per WebServices an die angeschlossene SAP IS-U Schnittstelle weiterzuleiten. Da die Teilgruppe im Rahmen ihrer Arbeit auf dieser Lösung aufbauen wird, wird diese Applikation im Kapitel 4.3.2.1 genauer thematisiert.

4.3.1.2 CX4U Consulting GmbH & Co. KG

Der Fachbereich Consulting beschäftigt sich vorrangig mit der fachlichen und prozessorientierten Beratung der am Gas- bzw, am Strommarkt tätigen Unternehmen. Dabei werden zwei grundlegende Beratungsobjekte unterschieden, nämlich die Unternehmensarchitektur und die fachlichen Fragestellungen. Die Beratungstätigkeit an sich, umfasst dabei neben der Strategieebene auch

die Ebene der Geschäftsarchitektur sowie der IT-Architektur. Je nach betrachtetem Hintergrund werden dabei jeweils unterschiedliche Leistungen angeboten (CX4U (2013b)).

4.3.1.3 CX4U Events GmbH & Co. KG

Der Themenbereich Events befasst sich mit der Planung und Gestaltung der Oldenburger Gastage, welche seit 2010 stattfinden und einen kompetenten Rahmen zur Diskussion aktueller Themen aus den Gebieten Energie- und Wasserwirtschaft mit Vertretern aus Politik und Wirtschaft erlauben (CX4U (2013b)).

4.3.2 Zweck der Software

Dieser Abschnitt dient der Klärung der Hintergründe und Intentionen, welche zur Initialisierung dieses Teilprojektes geführt haben. Es werden daher zunächst die allgemeine Ausgangssituation sowie dessen Herausforderungen und die daraus resultierenden Aufgaben beschrieben. Im Anschluss daran werden die Ziele und die damit einhergehende Vorgehensweise der Teilprojektgruppe beschrieben. Die folgenden Beschreibungen und Ausarbeitungen wurden von der Teilgruppe in Zusammenarbeit mit den Betreuern der CX4U ausgearbeitet.

4.3.2.1 Ausgangssituation

Zum gegenwärtigen Zeitpunkt bietet die CX4U AG mit dem SmartGateway eine integrierte Softwarelösung für das Auslesen und Analysieren von Stromzählerdaten an, welche als Schnittstelle zwischen den intelligenten Zählern (Smart Metern) auf der einen, und den nachgelagerten SAP Abrechnungssystemen der Energieversorger auf der anderen Seite, fungieren.

Aus Sicht der Systemarchitektur bildet das SmartGateway zusammen mit dem von OSI Soft stammenden Meter Data Management System eine von der SAP AG spezifizierte Middleware namens Meter Data Unification & Sync System (MDUS). Der grundlegende Aufbau ist in Abbildung 4.4 zu sehen.

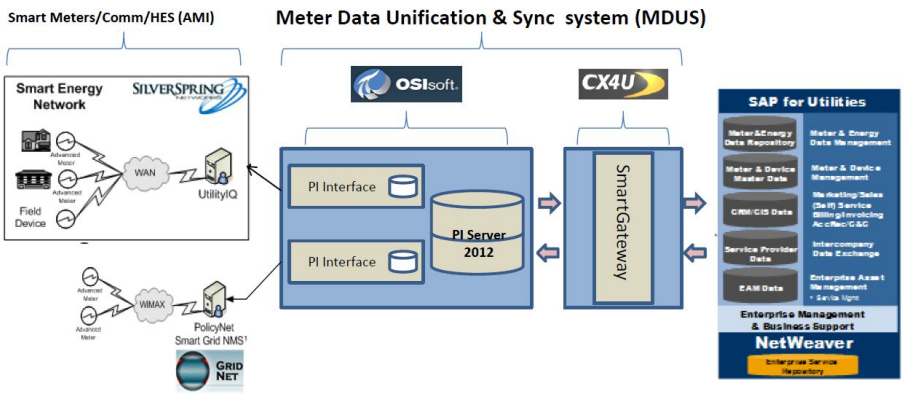


Abbildung 4.4: Systemarchitektur MDUS
Quelle: CX4U AG

Dieses Zähl- und Messdatenverwaltungssystem ist sowohl für das Auslesen der Zähler als auch für die Datenhaltung und Übertragung an die angeschlossenen SAP Systeme verantwortlich. Des Weiteren werden einige Operationen, wie eine Zeitstempelbereinigung, auf den Messwerten ausgeführt, damit diese durch die angeschlossenen Systeme verarbeitet werden können. Der Datenaustausch verläuft dabei beidseitig automatisiert und über MDUS spezifische Webservices.

Im Verlauf der letzten Monate und Jahre haben sich beim Umgang mit Messdaten von intelligenten Zählern einige Problemstellungen herauskristallisiert, welche die CX4U mit ihrem Produktportfolio bzw. dem Funktionsumfang des SmartGateways im speziellen derzeit lediglich hinreichend, aber nicht optimal bedient. Zu diesen Herausforderungen zählt vor allem die enorm gestiegene Menge an Rohdaten die es zu verarbeiten gilt. Ähnlich der Entwicklung in anderen EDV-Bereichen stieg diese auch in der Strom- und Energiebranche deutlich an und bedarf aus diesem Grund einer grundlegenden Überarbeitung.

Grundlegende Funktionen, wie das Prüfen auf Korrektheit oder das Bilden von Ersatzwerten von fehlerhaften Messwerten, werden bislang nicht vom MDUS bzw. SmartGateway angeboten, sondern finden lokal während der Zählerfernauslesung statt. Zusammenfassend betrachtet, fehlt demnach ein System, welches diese Herausforderungen möglichst performant adressiert.

4.3.2.2 Zielbestimmung

Im Rahmen der Teilprojektgruppe soll eine Softwarelösung konzipiert und prototypisch entwickelt werden, die sowohl Methoden zur Überprüfung und gegebenenfalls auch zur Ersatzwertbildung von Messdaten anbietet, als auch eine möglichst performante Verarbeitung großer Datenmengen ermöglicht. Da die SAP AG vor kurzer Zeit ihre In-Memory Appliance namens HANA vorgestellt hat und die CX4U AG ein lizensierter Entwicklungspartner der SAP ist, wurde sich für eine Umsetzung mit eben jener HANA Appliance entschieden. Der große Unterschied im Vegleich zu herkömmlichen Datenbanken liegt in der Datenhaltung bzw. -verarbeitung im Arbeitsspeicher und nicht auf einer regulären Festplatte. Außerdem beinhaltet die Appliance nicht nur die tatsächliche Datenbank, sondern darüber hinaus noch einige Softwarebausteine, deren Logik direkt in der Datenbankschicht läuft. Die dadurch erzielten Performancevorteile sollen es der CX4U somit erlauben, den aktuellen Herausforderungen gerecht zu werden und auch zukünftigen Entwicklungen hinsichtlich eines weiter ansteigenden Datenvolumens erfolgreich zu begegnen.

Die primäre und damit wichtigste Aufgabe bzw. Zielsetzung der Teilprojektgruppe Smart Metering ist zunächst, anhand einer Machbarkeitsstudie potentielle Möglichkeiten zu untersuchen, wie ein SAP HANA System an die vorliegende Systemlandschaft um das Smart Gateway angebunden werden kann.

Um dies zu erreichen, werden unter Berücksichtigung gewisser Rahmenbedingungen, wie möglichst geringe zusätzliche Administrationsaufwände oder das Vermeiden zusätzlicher Kosten, einige in Frage kommende Ansätze untersucht und gegeneinander abgewogen. Gerade vor dem Hintergrund, dass HANA noch immer ein relativ junges Produkt der SAP AG ist und

sich dementsprechend nach wie vor kontinuierlichen Veränderungen unterzogen sieht, ist das Auffinden einer allgemein gültigen und machbaren Lösung keinesfalls als trivial anzusehen.

In der Folge wird sich die Projektgruppe mit der Entwicklung eines HANA-basierten Softwareframeworks beschäftigen, welches eine Datenverarbeitung im Sinne eines von der CX4U vorangetriebenen VEE-Konzeptes erlaubt. Es beschreibt den Prozess der Validierung (Validation), Ersatzwertbildung (Estimation) und Editierung (Editing) von Sensordaten und bietet somit gewissermaßen eine Möglichkeit, die Sensordaten auf ihre Korrektheit hin zu überprüfen, an.

Da der Prozess der Meterdatenverarbeitung an sich sehr komplex und hochautomatisiert abläuft, ist er gleichzeitig auch sehr störanfällig. Falsche oder fehlende Messwerte sind keine Seltenheit und treten meist auf Grund von einzelnen oder einer Kombination mehrerer Faktoren, wie Kommunikations- oder Gerätefehlern auf. Es ist daher unbedingt notwendig, angemessen auf derartige Gegebenheiten reagieren zu können. Der VEE-Prozess zielt nun auf eben diese Anforderungen ab und bietet eine effizienzsteigernde Möglichkeit an, mit fehlerhaften Messdaten umzugehen und somit letztlich die Genauigkeit des anvisierten Anwendungsfalls zu erhöhen.

Durch die Umsetzung des genannten Konzeptes soll vor allem das frühzeitige Erkennen von Problemen bei der Übertragung von Messdaten und damit die Erhöhung der Datenqualität erreicht werden. Weiterhin spielt im Rahmen der Übersichtlichkeit bei der Arbeit mit großen Datenmengen der Umgang mit Änderungen - beispielsweise durch eine Ersatzwertbildung - eine entscheidende Rolle. Aus diesem Grund soll zusätzlich ein Logging-Verfahren implementiert werden, welches jegliche Änderungen an den Rohdaten dokumentiert.

4.3.2.3 Technische Produktumgebung

Während der Teilprojektgruppen-Phase stehen uns jeweils zwei Systeme zur Seite. Einmal ein Netweaver System seitens der CX4U AG und ein SAP Hana System, welches vom HPI gehostet und zur Verfügung gestellt wird.

4.3.2.3.1 SAP-System der CX4U-AG

Das SAP System der CX4U AG liegt in der Version „EHP6 FOR SAP ERP 6.0“ vor und läuft auf einem Windows NT Betriebssystem. Unter dem SAP System liegt eine Oracle Datenbank in der Version „11.2.0.3.0“. Die Module SAP_BASIS, SAP_ABA und SAP_AP sind bspw. in der Version 731 vorhanden. Erreicht wird das System der CX4U AG mittels einer PPTP-VPN Verbindung.

4.3.2.3.2 SAP HANA des HPI

Das SAP HANA System liegt in der Version $1.00.60.379371$ vor und wurde auf einem Server mit „SUSE Linux Enterprise Server 11.1“ - Betriebssystem installiert. Der Linux-Kernel liegt in der Version 2.6.32.59-0.7-default vor und es wurde Hardware des Herstellers Fujitsu verwendet. Bezüglich der Speicherkapazitäten steht ein Data-Volume mit 590 GB Speicherplatz zur Verfügung. Der Datenbank-Arbeitsspeicher wird mit ca. 974 GB bemessen. Um eine Verbindung

zu diesem System aufbauen zu können, muss vorerst eine OpenVPN-Verbindung eingerichtet werden.

4.3.3 Stakeholder

Die auf SAP HANA basierende Anwendung, welche im Rahmen der Projektgruppe konzipiert und prototypisch implementiert wird, adressiert, wie bereits in Kapitel 2 erläutert, vorrangig den Auftraggeber, nämlich die CX4U AG - insbesondere deren Holdingtochter CX4U Solutions GmbH & Co. KG. Gerade vor dem Hintergrund, dass der bereits etablierte Softwarebaustein des Smart Gateways um die Funktionalität einer Ersatzwertbildung erweitert wird, können die Ergebnisse dieser Kooperation als Anreicherung bzw. Optimierung des bestehenden Produktportfolios gesehen werden. Aus diesem Grund ist die CX4U AG sicherlich als erster Stakeholder anzusehen. Darüber hinaus können die entwickelten Resultate zukünftig von der CX4U AG am Markt entsprechend platziert und somit von deren Kunden bezogen und genutzt werden. Dies beschreibt ein - wenn auch indirektes - Interesse der Gas- bzw. Stromanbieter und -versorger und führt dementsprechend dazu, dass jene Industrie ebenfalls als Stakeholder zu bezeichnen ist.

Als dritter großer Interessenvertreter ist letztlich die Carl von Ossietzky Universität Oldenburg zu nennen, welche gleichzeitig die Ausrichtungs- und Organisationsaufgabe innehat. Von den Ergebnissen der Teilprojektgruppe kann insbesondere die ausrichtende Abteilung Very Large Business Applications bei einer längerfristigen Zusammenarbeit mit der CX4U AG profitieren, da das erarbeitete Wissen im Verlauf der Projektarbeit vor allem im Bereich einer zukunftsweisenden Technologie wie SAP HANA, in Form von Dokumenten und Berichten weiterhin zur Verfügung stehen wird und demnach nicht verloren geht.

Schlussendlich hegen alle Studenten, welche an der Projektgruppe teilnehmen, ebenfalls ein besonderes Interesse, da alle Erfahrungen und Fertigkeiten die während der Projektarbeit aufgebaut und vertieft werden, im Zuge einer persönlichen Weiterentwicklung von außerordentlicher Relevanz geprägt sind. Erfahrungen in neuen und aussichtsreichen Technologien wie beispielsweise SAP HANA sind zukünftig sicherlich von großer Bedeutung. Gerade vor dem Hintergrund einer erfolgreichen beruflichen Perspektive sind jene erworbenen Eigenschaften definitiv als hilfreich zu bezeichnen und lassen somit den Begriff des Stakeholders auch für die Gruppe der Studierenden zu.

4.3.4 Geschäftsanwendungsfälle

Die von der Projektgruppe erarbeiteten Lösungen sollen auf langfristige Sicht gesehen, zwei grundlegende Geschäftsanwendungsfälle bei den Kunden der CX4U abdecken, nämlich den der Endkundenabrechnung sowie den der Verbrauchsprognose. In der Folge werden beide Anwendungsfälle kurz beschrieben.

4.3.4.1 Endkundenabrechnung

Die Endkundenabrechnung umfasst prinzipiell alle Tätigkeiten, die auf Seiten des Kunden zur Rechnungsstellung notwendig sind. Dabei werden nicht nur verschiedene Abrechnungszyklen, sondern auch Unterscheidungen nach bestimmten Kriterien, wie etwa der Branche, berücksichtigt. Sie erfolgt kundenseitig über das SAP IS-U System, d.h. sie wird von dort aus angestoßen und liefert auch die jeweiligen Ergebnisse dorthin zurück.

Die Beschaffung der notwendigen Daten aus den intelligenten Zählern erfolgt dabei über das SmartGateway bzw. das MDUS-System, welches unter anderem eine Zeitstempelbereinigung durchführt um die Daten lesbar für das IS-U zu machen. Die VEE-Funktionalitäten werden in diesem Fall nicht von der HANA-Anwendung durchgeführt, sondern von der ABAP-Komponente im SmartGateway, d.h. HANA wird hier nicht eingesetzt. Im Rahmen der Projektarbeit wird dieser Anwendungsfall daher nur theoretisch betrachtet und nicht praktisch umgesetzt.

4.3.4.2 Verbrauchsprognose

Mit Hilfe der Verbrauchsprognose versuchen die Energieversorger den voraussichtlichen Energieverbrauch für einen bestimmten Zeitraum zu prognostizieren. Da sich exakte Angaben nur schwer aufstellen lassen, sind für diesen Anwendungsfall große Mengen an Erfahrungswerten bzw. historischen Daten erforderlich. Diese werden verschiedenen mathematischen Verfahren unterzogen,wodurch sich Aussagen über zukünftige Verbrauche prognostizieren lassen.

Ähnlich der Endkundenabrechnung wird auch dieser Fall über das SAP IS-U angestoßen und dargestellt. Ein wichtiges Unterscheidungsmerkmal liegt jedoch in der Verwendung der HANA-Anwendung als datenverarbeitenden Teil. Im Gegensatz zur Abrechnung übernimmt diese nämlich sämtliche VEE-Schritte. Das SmartGateway dient in diesem Fall vorrangig als Daten- und Informationsquelle sowie als Adressat der Logdateien.

4.3.5 Systemanwendungsfälle

Die in diesem Abschnitt angeführten und dokumentierten Systemanwendungsfälle beschreiben das für außenstehende Akteure wahrnehmbare Verhalten des zu entwickelnden Systems (Vgl. Abb. 4.5). Sie werden innerhalb dieses Kapitels jeweils sowohl textuell, d.h. in diesem Fall in tabellarischer Form beschrieben, als auch in Form eines Aktivitätsdiagramms erläutert. Als Akteur können dabei sowohl Personen, welche mit dem System interagieren, als auch andere Systeme, die in irgendeiner Weise das Verhalten des Systems beeinflussen oder von diesem beeinflusst werden, fungieren. An dieser Stelle soll daher kurz aufgelistet werden, welche Akteure im Rahmen dieser Konzeption im Einzelnen zur Gruppe der Akteure gezählt werden.

4.3.5.1 Akteure

Bevor in der Folge einzelne Systemanwendungsfälle im Detail betrachtet werden, sollen an dieser Stelle zunächst die am Prozess beteiligten Akteure aufgeführt werden.

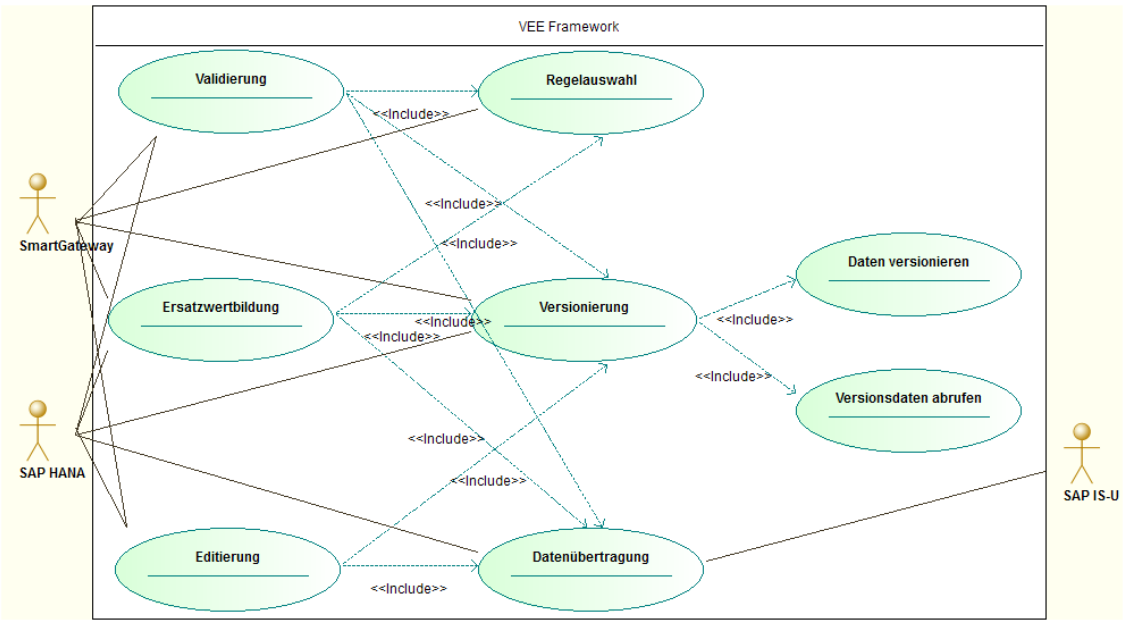


Abbildung 4.5: Systemanwendungsfälle

1. **Smart Gateway:** Das SmartGateway repräsentiert, wie bereits in Kapitel 4.3.2.1 beschrieben, die vorliegende Schnittstelle zwischen der Datenhaltung und den angeschlossenen SAP-Abrechnungssystemen. Es kontrolliert somit die Anfrage und Übermittlung aller in HANA bzw. dem ABAP-Modul notwendigen Informations- und Datenflüsse.

2. **SAP IS-U:** Das SAP IS-U stellt das Endanwendersystem dar, so wie es beim jeweiligen Kunden vorliegt. Es dient gleichzeitig einerseits als Anstoßpunkt für den gesamten Datenverarbeitungsprozess im Sinne des VEE-Konzeptes, andererseits wird das Ergebnis jenes Prozesses auch über dieses System dem Endanwender dargestellt.

3. **SAP HANA:** Die SAP HANA Anwendung führt bei gewissen Anwendungsfällen ebenfalls eigenständig eine Datenverarbeitung nach VEE durch. Die dafür benötigten Daten und Informationen werden über das SmartGateway kommuniziert.

Es sei an dieser Stelle explizit zu erwähnen, dass die aufgeführten Erläuterungen auf konzeptioneller Basis erstellt wurden und nicht den tatsächlichen Funktionsumfang der prototypischen Implementierung widerspiegeln. Die für die Anfertigung der Diagramme verwendete Software heißt Modelio und ist eine Open Source Anwendung zur Erstellung von Modellen jeglicher Art. Erläuterungen zum Import/Export oder zum Bearbeiten von Dateien sind auf der Homepage des Entwicklers zu finden[35].

[35] `http://forge.modelio.org/projects/modelio3-usermanual-english-310/wiki`

4.3.5.2 Validierung

Name	Validierung
Akteur(e)	Smart Gateway, SAP HANA
Kurzbeschreibung	• Auf Basis definierter Regeln werden Messdaten validiert.
Anfangsbedingung	• Die kundenindividuellen Regeln wurden ausgewählt und liegen vor. Des Weiteren sind die Messwerte des betrachteten Zeitraums für die Validierung verfügbar.
Abschlussbedingung - Erfolgsfall	• Auf Basis der ausgewählten Regeln wurden Testroutinen auf die Messwerte angewendet - diese sind nunmehr als gültig bzw. nicht gültig gekennzeichnet. Außerdem wurden jegliche Änderungen versioniert.
Abschlussbedingung - Misserfolgsfall	• Testroutinen konnten nicht angewendet werden und das Ergebnis wird protokolliert.
Abnahmekriterium	• Messwerte wurden validiert und erhalten entweder einen gültigen oder einen nicht gültigen Status. Alle Änderungen wurden protokolliert.

Ablauf - Normalszenario	1. Das SAP HANA erfragt die Validierungsregeln für den Kunden beim SmartGateway. 2. Das SmartGateway überträgt die Validierungsregeln an das SAP HANA. 3. Das SAP HANA wendet die Validierungsregeln an. 4. Das SAP HANA setzt einen Status, um kenntlich zu machen, welcher Wert ersatzwertgebildet werden muss. 5. Das SAP HANA löst die Versionierung aus.
Ablauf - Alternativ	1. zu 3.: 1. Alternative: Es wurden keine Fehlerhaften Werte erkannt, daher ist eine Ersatzwertbildung nicht nötig. 2. zu 3.: 2. Alternative: Es wurden Fehlerhaften Werte erkannt, daher ist eine Ersatzwertbildung nötig.
Ablauf - Fehler	1. zu 1.: Es liegen keine Regeln für den betrachteten Kunden vor. 2. zu 3.: Das SAP HANA kann die Validierungsregeln nicht anwenden.

Tabelle 4.37: Validierung

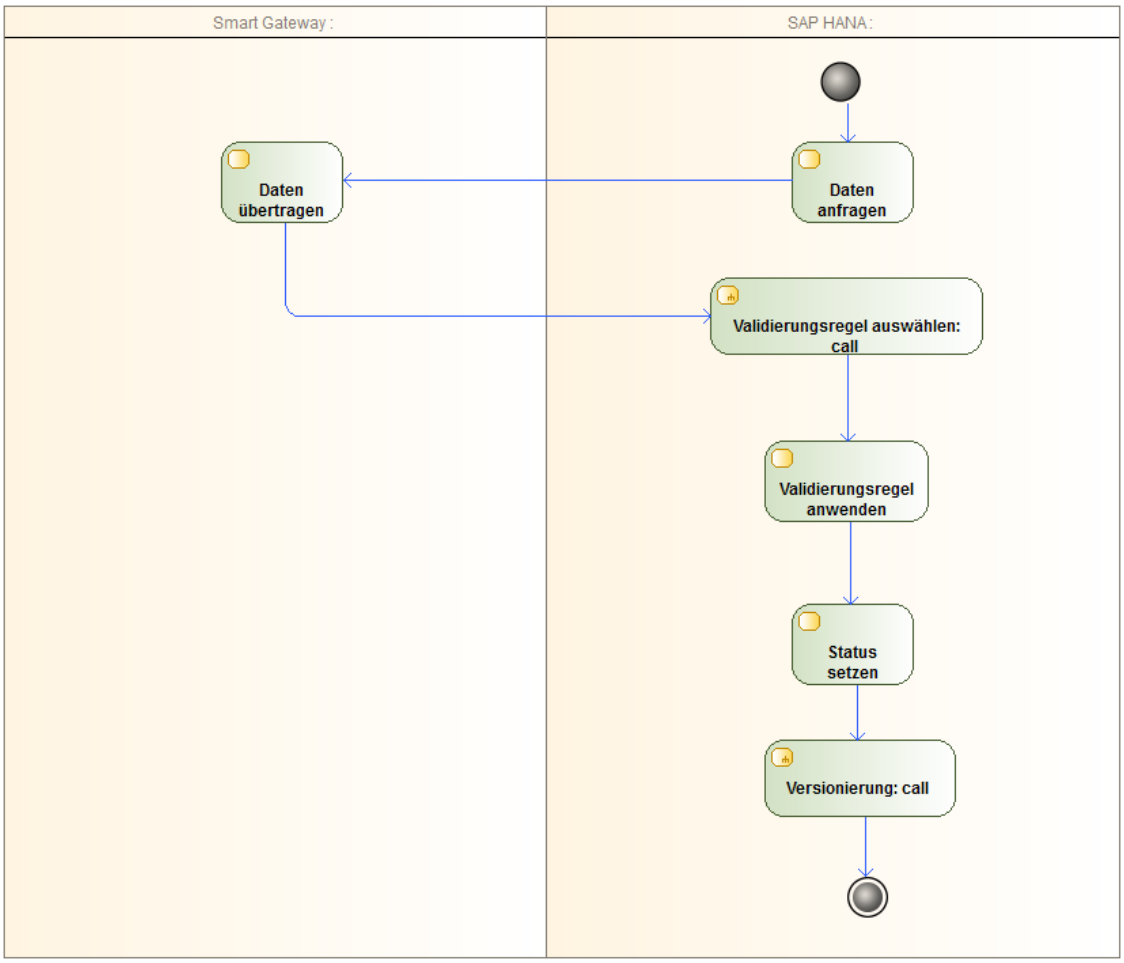


Abbildung 4.6: Validierung

4.3.5.3 Estimation

Name	Estimation
Akteur(e)	SmartGateway, SAP HANA
Kurzbeschreibung	• Auf Basis definierter Regeln werden Ersatzwerte gebildet.
Anfangsbedingung	• Daten, die vorher validiert wurden, liegen vor.
Abschlussbedingung - Erfolgsfall	• Die unterschiedlichen Ergebnisse der Ersatzwertbildung werden versioniert und protokolliert.
Abschlussbedingung - Misserfolgsfall	• Die Methoden zur Ersatzwertbildung können nicht angewendet werden und das Ereignis wird protokolliert.
Abnahmekriterium	• Die Methoden zur Ersatzwertbildung werden durchgeführt und die unterschiedlichen Ergebnisse werden versioniert und protokolliert.

Ablauf - Normalszenario	1. Das SAP HANA ruft den Status des Messwertes ab. 2. Das SAP HANA erfragt die Kundenregel beim SmartGateway. 3. Das SmartGateway überträgt die Kundenregel an das SAP HANA. 4. Das SAP HANA wendet die Kundenregel an. 5. Das SAP HANA erfragt die Estimationregel beim SmartGateway. 6. Das SmartGateway übertragt die Estimationregel an das SAP HANA. 7. Das SAP HANA wendet die Estimationregel an. 8. Das SAP HANA setzt einen Status, um kenntlich zu machen, welcher Wert bei der Ersatzwertbildung errechnet wurde. 9. Das SAP HANA löst die Versionierung aus.
Ablauf - Alternativ	1. zu 4.: 1. Alternative: Eine Ersatzwertbildung ist nach der Kundenregel notwendig. 2. zu 4.: 2. Alternative: Eine Ersatzwertbildung ist nach der Kundenregel nicht notwendig.

Ablauf - Fehler	1. zu 4.: Das SAP HANA kann die Kundenregel nicht anwenden. 2. zu 7.: Das SAP HANA kann die Estimationsregel nicht anwenden.

Tabelle 4.38: Estimation

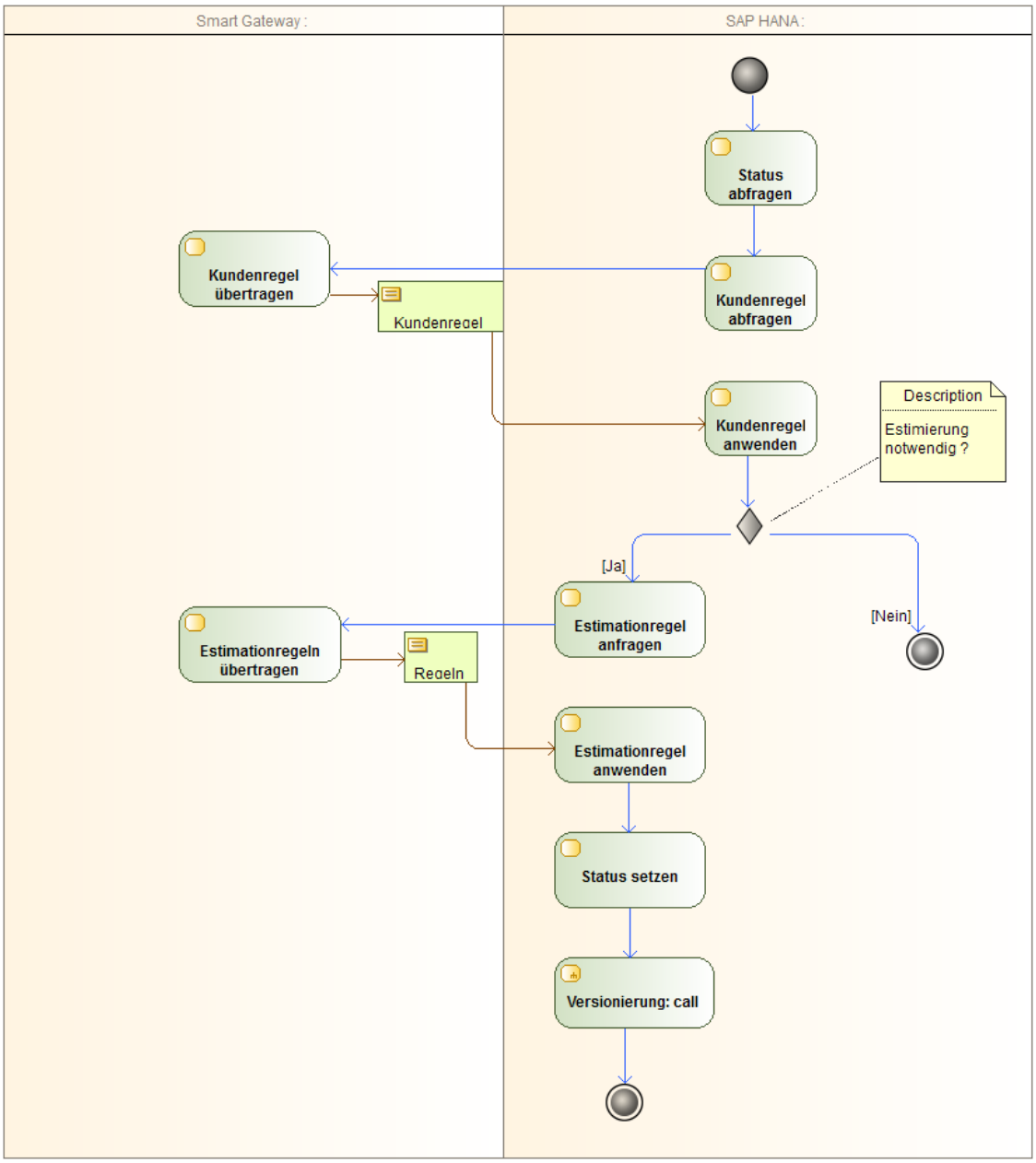


Abbildung 4.7: Estimation

4.3.5.4 Regelauswahl

Name	Regelauswahl
Akteur(e)	Smart Gateway, SAP HANA
Kurzbeschreibung	• Auf Basis des Kundenidentifikationsmerkmals werden die anzuwendenden Regeln sowohl für die Validierung als auch für die Ersatzwertbildung ermittelt.
Anfangsbedingung	• Das Kundenidentifikationsmerkmal liegt vor.
Abschlussbedingung - Erfolgsfall	• Alle anzuwendenden Regeln liegen im Konfigurationsset vor.
Abschlussbedingung - Misserfolgsfall	• Regeln konnten nicht ausgewählt werden und liegen nicht vor.
Abnahmekriterium	• Das Konfigurationsset enthält sowohl Validierungs- als auch Ersatzwertregeln.
Ablauf - Normalszenario	1. Das SAP HANA erfragt den Regelsatz beim SmartGateway. 2. Das SmartGateway identifiziert den Regelsatz anhand der Kunden ID. 3. Das Smart Gateway überträgt den Regelsatz an das SAP HANA. 4. Das SAP HANA empfängt den Regelsatz und wendet ihn an.
Ablauf - Alternativ	-

Ablauf - Fehler	1. zu 1.: Die Verbindung zwischen HANA und SmartGateway kann nicht aufgebaut werden. 2. zu 2.: Das SmartGateway kann keinen Regelsatz zu der Kunden ID finden. 3. zu 3.: Die Verbindung zwischen SmartGateway und HANA kann nicht aufgebaut werden.

Tabelle 4.39: Regelauswahl

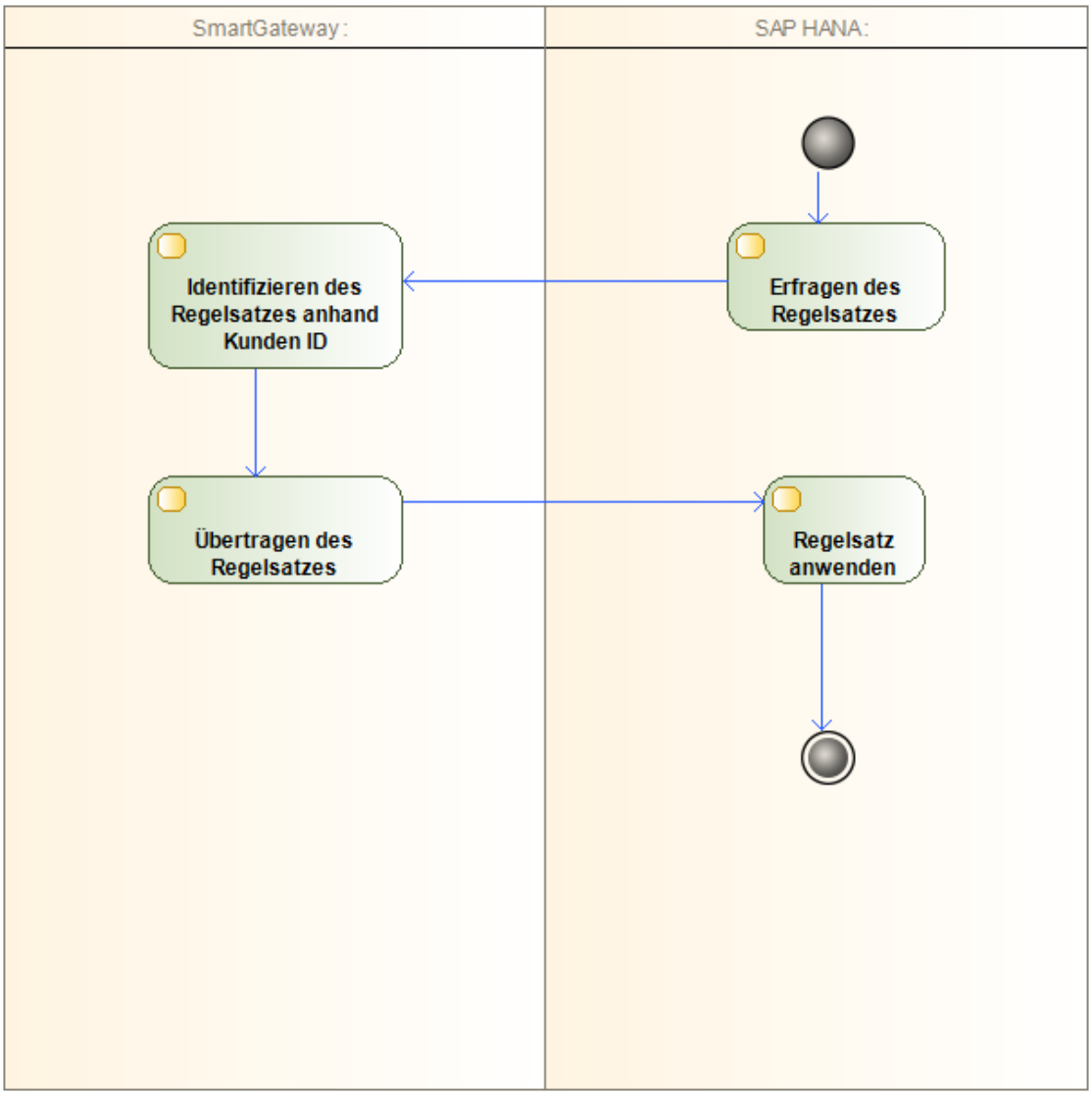


Abbildung 4.8: Regelauswahl

4.3.5.5 Versionierung

Name	Versionierung
Akteur(e)	Smart Gateway, SAP IS-U
Kurzbeschreibung	• Das Vorgehen bei der Bildung von Ersatzwerten wird revisionssicher dokumentiert.
Anfangsbedingung	• Daten, bei denen eine oder mehrere Validierungsmethoden gescheitert sind, liegen vor.
Abschlussbedingung - Erfolgsfall	• Ersatzwerte wurden gebildet und das Vorgehen der Ersatzwertbildung und der Grund der Auswahl dieser Methode wurden dokumentiert.
Abschlussbedingung - Misserfolgsfall	• Es wurden Ersatzwerte gebildet aber das Vorgehen wurde nicht dokumentiert.
Abnahmekriterium	• Die Methoden zur Ersatzwertbildung werden durchgeführt und die unterschiedlichen Ergebnisse werden (versioniert und) protokolliert.
Ablauf - Normalszenario	1. Die nicht validierten Daten werden als solche markiert. 2. Methoden zur Ersatzwertbildung werden ausgewählt und durchgeführt. 3. Ergebnisse der Ersatzwertbildung und die benutzte Methode werden versioniert und protokolliert.

Ablauf - Alternativ	-
Ablauf - Fehler	-

Tabelle 4.40: Versionierung

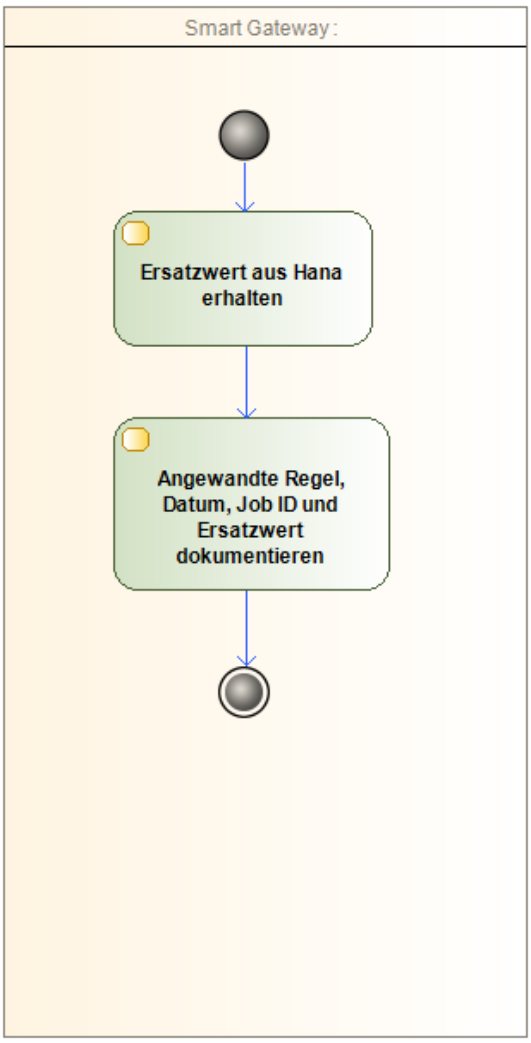


Abbildung 4.9: Versionierung

4.3.5.6 Editing

Name	Editing
Akteur(e)	SmartGateway, SAP HANA, SAP IS-U
Kurzbschreibung	• Im Prozess Editing wird – sofern notwendig – eine Veränderung (Manipulation) der Messwerte vorgenommen, damit später fehlerfreie Daten(sätze) vorliegen.
Anfangsbedingung	• Die Daten aus dem IS-U müssen vorliegen.
Abschlussbedingung - Erfolg	• Daten werden korrigiert an Schnittstelle übergeben.
Abschlussbedingung – Misserfolg	• Daten sind nicht an Schnittstelle übergeben.
Abnahmekriterium	• Die Regeln zum Verändern der Daten sind ermittelt worden und können angewendet werden. Die korrigierten Daten können zurückgegeben und die Änderungen protokolliert werden.
Ablauf – Normalszenario	1. Auslösung des Prozesses durch SAP IS-U. 2. Berichtigung (Manipulation) eines Datensatzes anhand vorher festgelegter Regeln. 3. Nach Aktualisierung der Daten erfolgt die Rückgabe an SAP IS-U. 4. Protokollieren der Änderungen und Versionierung.

Ablauf - Alternativ	-

Tabelle 4.41: Editing

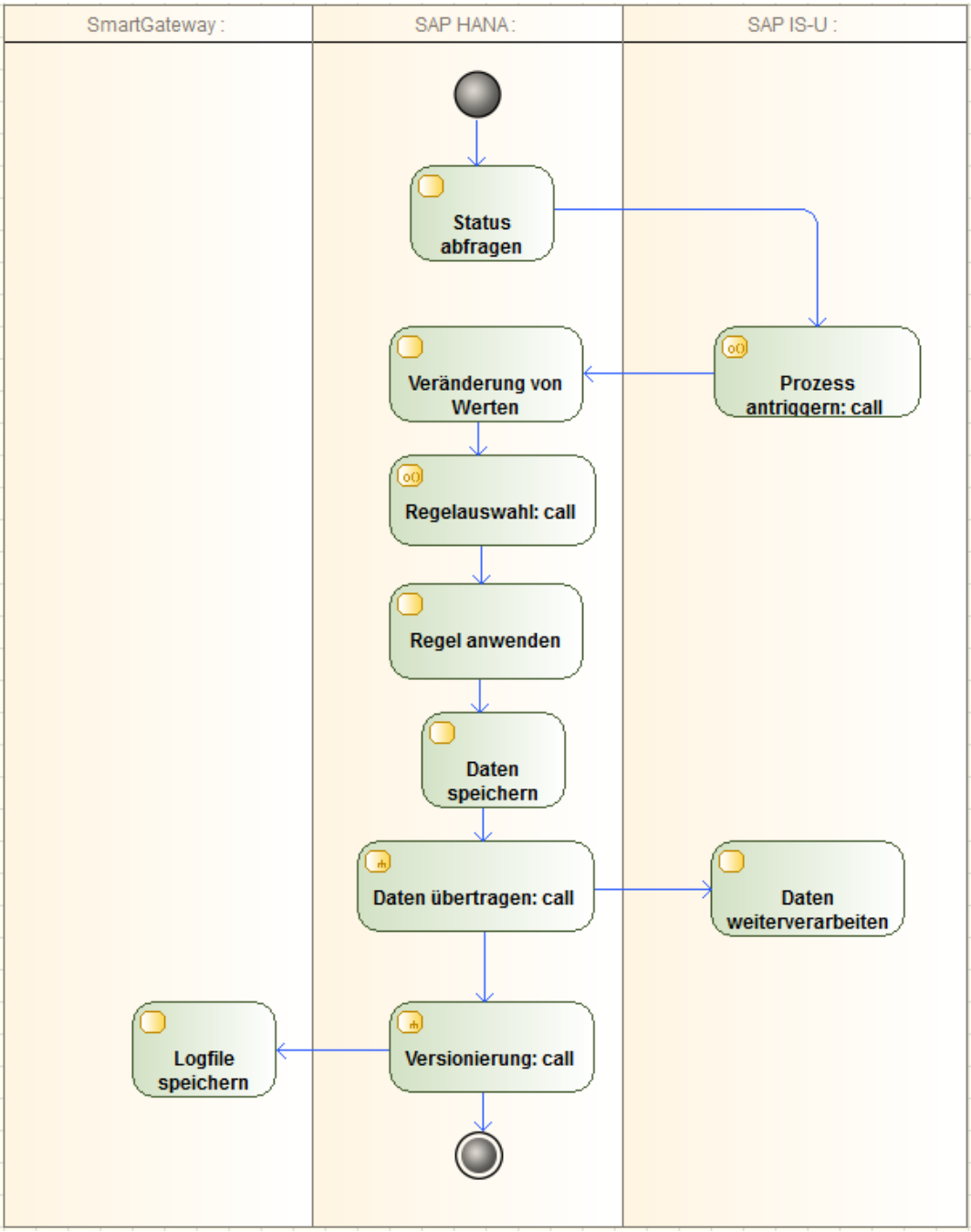


Abbildung 4.10: Aktivitätsdiagramm Editing

4.3.6 Systemanforderungen

In diesem Abschnitt soll beschrieben werden, was das System am Ende der Bearbeitungszeit leisten bzw. welche Dienste es anbieten soll. Es ist dabei zwischen der theoretischen Beschreibung und Ausarbeitung des bereits genannten VEE-Konzeptes und der prototypischen Umsetzung innerhalb des SAP HANA Systems zu unterscheiden.

Die konzeptionellen Ausführungen werden innerhalb der Projektgruppenarbeit hauptsächlich durch die Erstellung des Pflichtenheftes und der Machbarkeitsstudie sowie den darin enthaltenen Beschreibungen abgedeckt. So wird das VEE-Konzept, sowie die darin enthaltenen Besonderheiten und Eigenschaften bereits detailliert im Abschnitt Systemanwendungsfälle betrachtet. Die Inhalte der einzelnen Punkte werden dabei in Absprache mit der CX4U AG entwickelt und entsprechend dokumentiert.

Neben der theoretischen Beleuchtung und Definition des VEE-Konzeptes soll dieses von der Teilprojektgruppe mittels eines prototypischen Frameworks umgesetzt werden, sodass dessen Machbarkeit und Einsatzfähigkeit für die Praxis bewiesen wird. Da die Anbindung an die bestehende Systemlandschaft gleichermaßen sowohl Grundvoraussetzung für weitere Implementierungen, als auch primär fokussiertes Endziel der Projektgruppe ist, wird dieser Aufgabe die größte Relevanz zugeschrieben. Um diese letztlich umsetzen zu können, müssen daher einige Gegebenheiten berücksichtigt werden. So ist es beispielsweise notwendig, das Smart Gateway dahingehend zu erweitern, dass es technisch in der Lage ist mit einem etwaigen SAP HANA System zu kommunizieren. Es muss also eine, vom Smart Gateway angestoßene Verbindung zwischen beiden Systemen hergestellt werden.

Dies impliziert selbstverständlich ebenso, dass das HANA System diese Verbindung akzeptieren und verarbeiten kann. Eine für diesen Punkt geeignete Technologie bzw. ein in Frage kommender Ansatz wird innerhalb der Machbarkeitsstudie diskutiert und ausgewählt werden. Darüber hinaus soll die Datenverarbeitung ebenfalls auf HANA-Seite stattfinden, was die Existenz entsprechender Methoden voraussetzt. Dabei soll ein beliebiger Datensatz im Sinne des VEE-Konzeptes validiert, erstatzwertgebildet und editiert werden können. Es bleibt zu erwähnen, dass das Ziel an dieser Stelle nicht die Vollständigkeit ist, sondern die Schaffung einer Basis, auf welcher zu einem späteren Zeitpunkt aufgebaut werden kann. Nichtsdestotrotz soll auch das prototypische Framework die grundlegende Funktionalität anbieten und beispielhafte Berechnungen auf den Daten ausführen.

4.4 Systementwurf

Für den Entwurf eines möglichen Systems muss zunächst die vorhandene Architektur aller Projektbeteiligten betrachtet werden. Aus diesem Grund werden die eingesetzten Hard- und Softwarekomponenten sowohl von der CX4U AG in Oldenburg als auch vom Hasso-Plattner-Institut der Universität in Potsdam betrachtet. Auch die Verbindung zwischen diesen Systemen wird beschrieben. Mit Hilfe eines dynamischen Modells wird außerdem gezeigt, wie eine prototypische Lösung des Problems aussehen kann. Da unterschiedliche Systeme an verschiedenen Standorten eine zentrale Rolle in diesem Projekt einnehmen, müssen neben generellen Sicherheitsaspekten auch die Vergabe von Zugriffsrechten betrachtet werden.

4.4.1 Systemzerlegung

In der Systemzerlegung der Teilprojektgruppe CX4U müssen das HANA System und das Smart Gateway betrachtet werden. Die Daten in der HANA Datenbank kommen von dem Smart Gateway der Firma CX4U. Diese speist die Daten über die OData-Schnittstelle ein. Auf den über die OData-Schnittstelle importierten Daten arbeiten nun die Scripte für „Validation“, „Estimation“ und „Editing“. Diese Scripte müssen nicht einzeln aufgerufen werden, sondern können über ein einziges Webinterface aufgerufen werden (Vgl. Abb. 4.11).

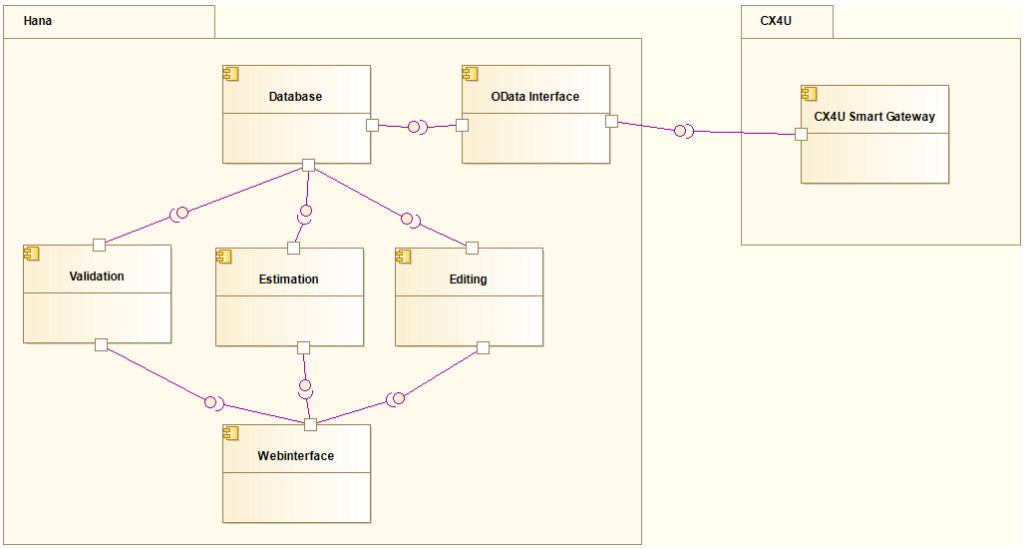


Abbildung 4.11: Systemzerlegung des HANA-Systems

4.4.2 Hard- und Softwarekomponenten

Bezüglich Hard- und Softwarekomponenten sind zwei Systeme zu betrachten. Einmal die Netweaver-Installation bei der CX4U AG und das HANA System beim HPI in Potsdam.

Das von der CX4U AG bereitgestellte System hat einen AMD-Opteron 6136 Prozessor und 4 GB Arbeitsspeicher. Softwareseitig läuft auf dem Server das Betriebssystem Windows Server

2008 R2 und auf diesem der SAP Netweaver mit u.a. dem SAP Basis-Modul in der Version 731. Die Datenbasis des SAP Systems liegt in einer Oracle-Datenbank der Version 11.2.

Bei dem HANA System gestaltet sich dies schwieriger, da das System von mehreren Universitäten genutzt wird und die Ressourcen daher aufgeteilt werden. Auch ist die genaue Hardware-Konstellation nicht zu erkennen. Jedoch kann man sagen, dass HANA zum korrekten Betrieb sehr viel Ressourcen benötigt, vor allem Arbeitsspeicher, da es sich schließlich um eine In-Memory-Lösung handelt. An Arbeitsspeicher steht ca. 1 TB zur Verfügung. SAP HANA läuft auf einem „SUSE Linux Enterprise Server 11.1“ und die beim Verfassen dieses Bericht eingespielte HANA Version war 1.00.60.379371.

4.4.3 VPN-Verbindung zwischen den Systemen

Um auf das HANA System des HPI zugreifen zu können, wird eine OpenVPN-Verbindung zu eben diesem benötigt. Dazu wurden den Studenten die erforderlichen Kennwörter und Zertifikats-Dateien zur Verfügung gestellt.

Damit nun auch über das Netweaver System auf das HANA System zugegriffen werden konnte, musste ebenfalls eine VPN-Verbindung vom Netweaver Server der CX4U AG zum HANA System beim HPI hergestellt werden. Dazu wurde einem Nutzer der Teilgruppe Administratoren-Rechte auf dem besagten Server der CX4U AG eingeräumt. Mittels dieser konnte dort OpenVPN installiert werden. Dies ist sehr einfach und nach erfolgreicher Installation kann die Verbindung gestartet werden, indem man einen Rechtsklick auf die vom HPI bereitgestellte „FSOC-Lab.ovpn“-Datei macht und die Funktion „Start OpenVPN on this config file“ wählt. Hier ist noch hinzuzufügen, das nach einem unbestimmten Zeitraum die Verbindung zum HPI gekappt wurde und diese erneut hergestellt werden musste. Die Anleitung des HPI ist hier zu finden: F.7.

4.4.4 Dynamisches Modell

In den folgenden zwei Kapiteln werden die dynamischen Modelle zum HTTP-Request und zum VEE-Framework erläutert. Das Dynamische Modell wird zunächst durch einen Text erklärt und anschließend in Form eines Sequenzdiagramms konkretisiert.

4.4.4.1 HTTP-Request

Durch das dynamische Modell in Abbildung 4.12 soll verdeutlicht werden, wie der Methodenaufruf in einer prototypischen Umsetzung aussehen kann. Zunächst muss die Verbindung zum Zielsystem hergestellt werden. Die RFC-Verbindung soll an zentraler Stelle mithilfe der Transaktion SM59 gepflegt werden. Diese erzeugt ein Client-Objekt, über welches auf alle benötigten Daten (Response, Header-Felder etc.) zugegriffen werden kann. Bevor der HTTP-Request ausgeführt wird, müssen im Client-Objekt die Methode auf „GET“ und die HTTP-Version auf „HTTP“ gesetzt werden. Anschließend wird der Request an den OData-Service gesendet. Die

Header-Felder und Daten der Antwort müssen mit zwei weiteren GET-Methoden ausgelesen werden. Abschließend werden die Daten ausgegeben und die Verbindung geschlossen.

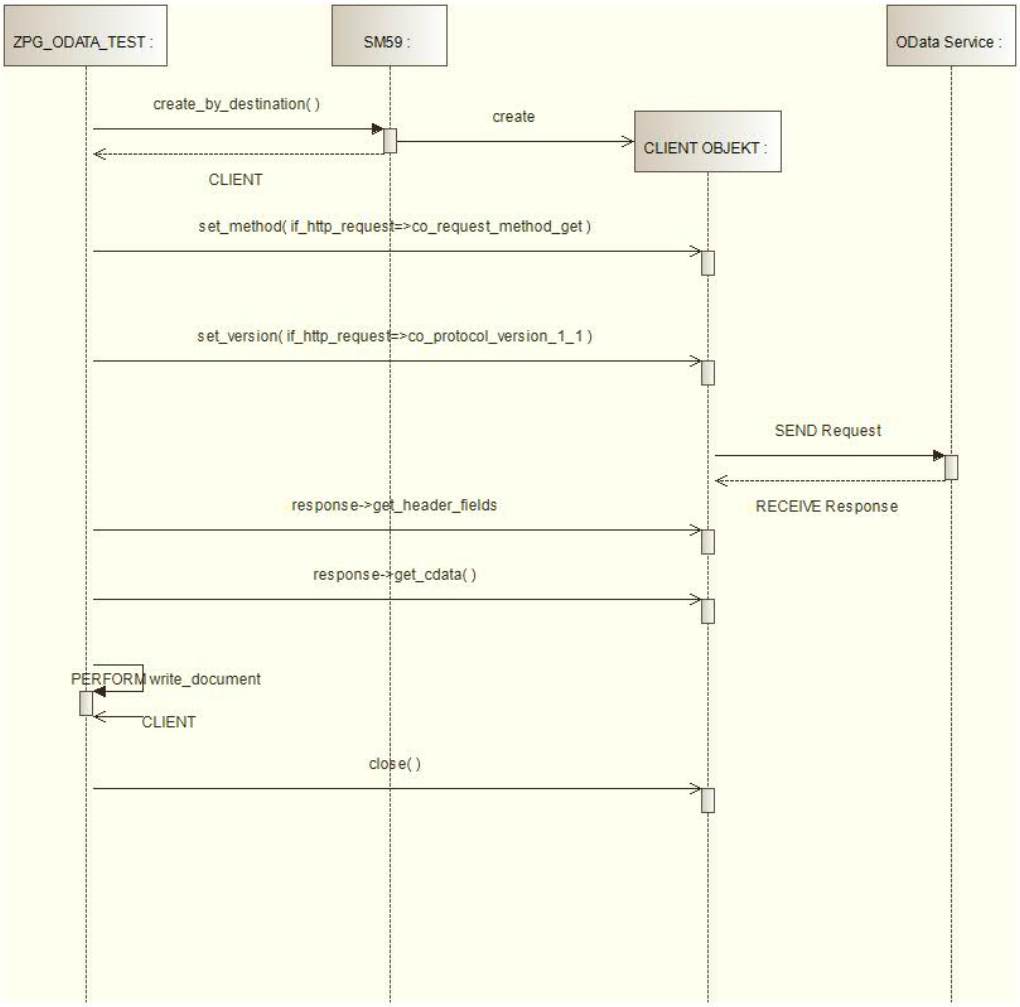


Abbildung 4.12: Dynamisches Modell des HTTP-Requests

4.4.4.2 VEE-Framework

Das Sequenzdiagramm in Abbildung 4.13 zeigt die Funktionsweise des VEE-Frameworks innerhalb des SAP HANA Extended Application Services. Zunächst muss der Nutzer die Zähler-ID eingeben und eine der verschiedenen Funktionen auswählen. Der Nutzer ruft über die Benutzeroberfläche die Funktionen des VEE-Frameworks auf. Durch die Validation Methode werden die Daten des ausgewählten Zählers aus der Datenbank geladen. Dann werden die Daten geprüft und die fehlerhaften Daten werden gekennzeichnet und in der Datenbank aktualisiert. Danach werden die Daten mittels der Estimation-Methode aufgerufen, um die Ersatzwertbildung auszuführen. Die angepassten und korrigierten Werte werden dann in der Datenbank aktualisiert. Anschließend werden die Daten für Editing gesammelt und auf fehlende Zeitstempel geprüft. Die korrigierten Zeitstempel und Intervallmessungen werden daraufhin in der Datenbank aktualisiert.

Nach jeder Funktion des VEE-Frameworks wird die Logfile-Tabelle in der Datenbank aktualisiert und die Ergebnisse werden auf der Benutzeroberfläche angezeigt.

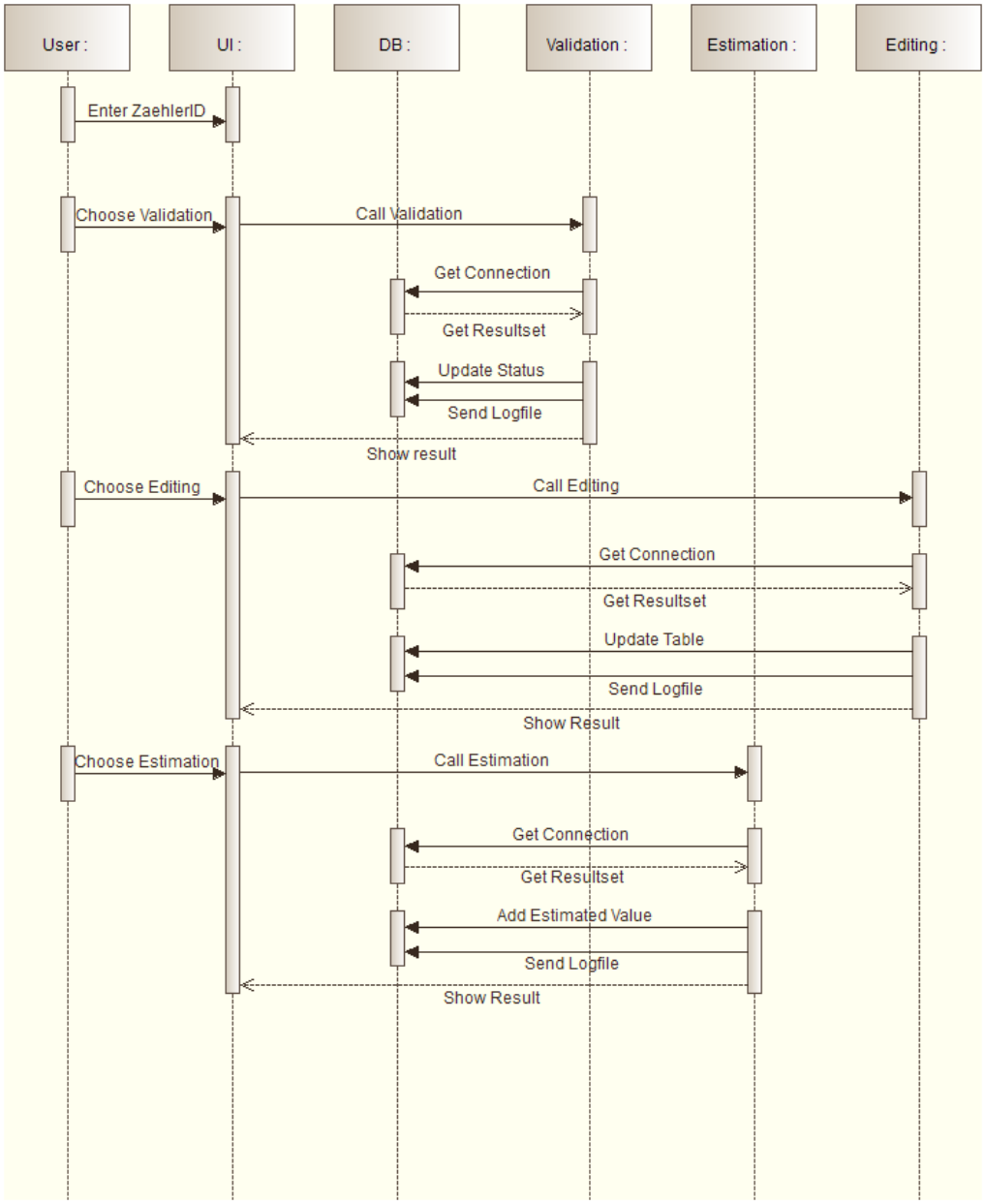


Abbildung 4.13: Dynamisches Modell des VEE-Frameworks

4.4.5 Fachliches Datenmodell

Das fachliche Datenmodell in Abbildung 4.14 verdeutlicht die in der HANA Datenbank vorliegenden Daten. Die Tabelle Zähler enthält Informationen über den jeweiligen Zähler. Dazu gehören die eindeutige ZaehlerID und die drei Regeln für Validation, Estimation und Editing. Der Datensatz eines Zählers besitzt somit drei Regeln. Regeln sind definiert durch eine eindeutige ID, eine Beschreibung und den Typ der Regel. Die Tabelle Messdaten enthält die gemessenen Zählerwerte, die ZahlerID des gemessenen Zählers, den Zeitstempel der Messung und den Status dieses Werts. Primary Key sind hierbei ZaehlerID und Zeitstempel. In der Tabelle „Logging“ werden lediglich gemachte Änderungen dokumentiert und mit einer eindeutigen ID versehen.

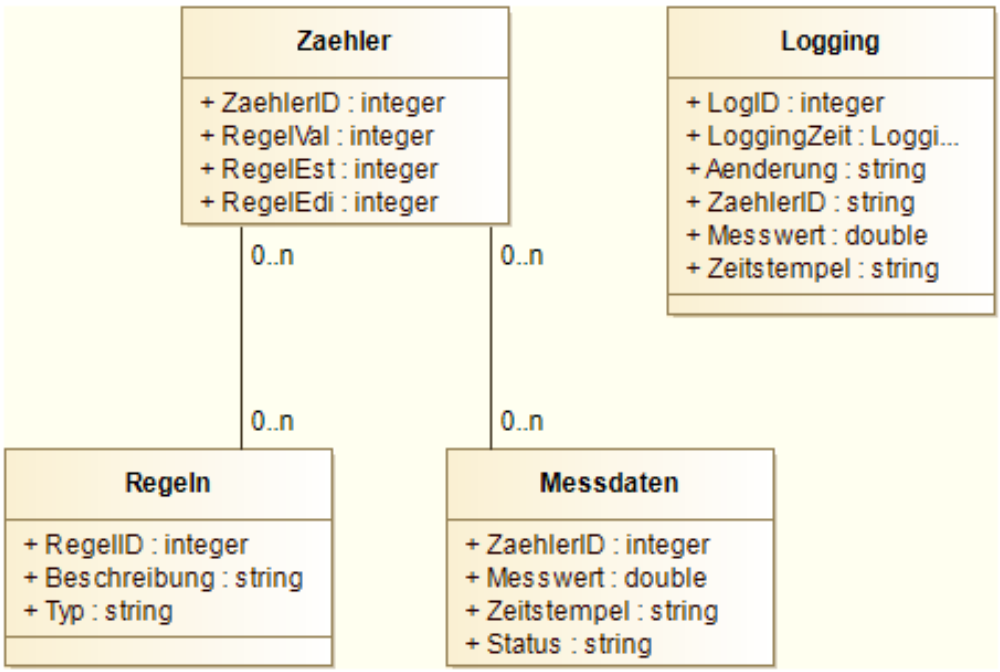


Abbildung 4.14: Darstellung der Daten im HANA-System

4.4.6 Sicherheitsaspekte und Zugriffsrechte

Die zentrale Problemstellung dieses Teilprojekts war es, die Kommunikation zwischen dem SAP NetWeaver der CX4U AG und dem SAP HANA des HPI herzustellen. Da sich diese beiden Systeme an unterschiedlichen Standorten - Oldenburg und Potsdam - befinden, ist eine sichere Kommunikation unbedingt notwendig. Diese wurde mittels mehrerer Virtual Private Networks (VPN) sichergestellt. Sowohl für die Anmeldung am SAP HANA des HPI als auch für die Anmeldung am SAP NetWeaver der CX4U AG wird eine VPN-Verbindung zu den jeweiligen Servern benötigt. Dies bedeutet, dass dies ebenso für die Kommunikation der beiden Systeme untereinander der Fall ist. Aus diesem Grund wurde auf Seiten des NetWeavers ein VPN-Tunnel zum HANA System eingerichtet.

Die Zugriffsrechte auf das Netweaver System der CX4U AG werden selbst vom Unternehmen verwaltet. Die Authentifizierung findet für jeden Benutzer mit einem eigenen Account statt. Es werden ein Benutzername, zusammengesetzt aus dem ersten Buchstaben des Vornamens und dem Nachnamen des Projektgruppenmitglieds, und ein Passwort verwendet. Das ursprünglich von der CX4U vergebene initiale Passwort musste hier von allen Benutzern nach dem ersten Anmelden geändert werden. Im NetWeaver-System selbst standen der Teilprojektgruppe im wesentlichen Lesezugriffe zur Verfügung. Nur für die jeweiligen Transaktionen der ABAP-Workbench sowie den Verbindungseinstellungen zu verbundenen Systemen wurden Schreibrechte eingeräumt. Die Benutzerrechte im SAP HANA System des HPI wurden mittels eines Rollenkonzeptes vergeben. Der Benutzer, namentlich zusammengesetzt aus Vor- und Nachname des Projektgruppenmitglieds, wurde vom HPI angelegt. Alle User gehörten der gleichen Gruppe an, besaßen also dieselben Rechte.

4.5 Machbarkeitsstudie

Dieses Dokument beschreibt einige Alternativen, welche vor dem Hintergrund einer kommunikativen Verbindung der beiden vorliegenden Systeme, also dem SAP Netweaver basierten Smart Gateway auf der einen Seite und dem SAP HANA System auf der anderen Seite, in Frage kommen. Dabei werden zunächst jene Ansätze aufgeführt, welche von der SAP für die Bereitstellung von Daten in HANA propagiert und empfohlen werden. Ausgewählte Konzepte sowie deren Systemvoraussetzungen und Vorteile werden dabei im Detail erläutert.
Im Anschluss werden die betrachteten Technologien an Hand einiger ausgewählter Kriterien miteinander verglichen und auf Basis dessen, der optimalste Ansatz für eine weiterführende Umsetzung im Sinne einer prototypischen Lösung ausgewählt. Jener Prototyp, sowie das für die Umsetzung gewählte Vorgehen und die vorhandenen Funktionen werden in der Folge detailliert beschrieben und an Hand von Grafiken und Listings veranschaulicht. Den Abschluss bildet eine zusammenfassende Schlussbetrachtung.

4.5.1 Aktueller Stand

Dieser Abschnitt widmet sich der Ausführung der aktuell vorliegenden Systemumgebung, sowie der betrachteten Herausforderungen und Problemstellungen. Die Hintergründe bzw. Informationen der dargestellten Beschreibungen stammen von den Betreuern der CX4U. Die CX4U als wirtschaftendes Unternehmen bietet eine SAP-gestützte Lösung zum Auslesen von elektronischen Strom- und Gaszählern sowie der weiteren Verarbeitung dieser Werte an. Es gibt aktuell eine sogenannte MDUS-Lösung, bei der es sich um eine gemeinsame Entwicklung von OSIsoft und der CX4U AG handelt. Dabei werden die eingehenden Daten der Smart Meter über das PI-Interface in den PI-Server geladen. Dieses System dient als Meter Data Management System (MDM) und verwaltet alle Verbindungen von und zu den Zählern. Dieser Bereich kann OSIsoft zugeordnet werden (Vgl. Abb. 4.15). Durch die Verwendung von WebServices gelangen die Daten in das SmartGateway. Von dort aus können sie in unterschiedlichen SAP-Systemen weiter verarbeitet werden. Mithilfe des SmartGateway ist es möglich, „sämtliche WebServices aus den AMI@SAP Enhancement Packages (EhP) 4, 5 und 6 aus den Bereichen Gerätemanagement, Zeitreihenmanagement und auch Customer Service in der Kommunikation innerhalb eines Meter Data Management Systems (MDM) abzubilden“ (Vgl. CX4U 2014). Das SmartGateway wurde durch die CX4U AG vollständig in ABAP als integrierte Lösung zu SAP entwickelt.
Unter Umständen ist es erforderlich, empfangene Messwerte zu korrigieren. Das ist beispielsweise bei unvollständigen Daten (fehlende Zeiten) oder fehlerhaften Daten der Fall. Dieser Vorgang wird als Ersatzwertbildung bezeichnet. In diesem Zusammenhang ist auch der Begriff VEE zu nennen, der sich aus den Anfangsbuchstaben seiner Phasen zusammensetzt.

Validation Es erfolgt eine Validierung der Daten, d.h. es werden unterschiedliche Prüfroutinen auf die eingehenden Daten angewendet. Dabei werden die Daten u.a. auf Konsistenz und Integrität geprüft.

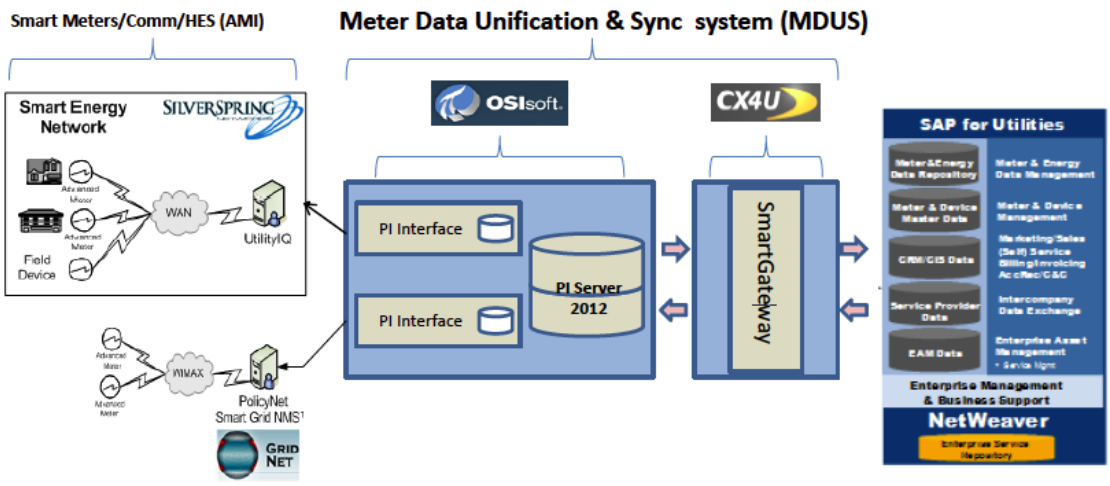


Abbildung 4.15: Aktuelle Systemumgebung
Quelle: Nannen (2013)

Estimation Hier findet die eigentliche Bildung der Ersatzwerte statt. Ungültige Daten erhalten dabei auf Grundlage vorher festgelegter Regeln die Ersatzwerte.

Editing In diesem Schritt werden nicht plausibel erscheinende Daten korrigiert. Dies geschieht ebenfalls anhand von definierten Regeln und kann entweder automatisiert oder manuell erfolgen.

Diese Schritte umfassen die einzelnen Phasen der Ersatzwertbildung. Sie müssen verschiedene Fälle und Regeln unterstützen. Die Regeln müssen frei konfigurierbar sein, damit eine individuelle Anpassung an Kundenwünsche und -vorgaben möglich ist. Bislang handelt es sich bei den betrachteten Daten nur um eine reaktiv kleine Datenmenge. Die Bildung der Ersatzwerte findet mittels einer Zählerfernauslesung (ZFA) statt, in denen die dafür zu verwendenden Regeln fest hinterlegt sind. Die Zeitreihen erhalten – sofern notwendig – ihre Ersatzwerte vor einer weiteren Übertragung. Rohdaten liegen im EDM nicht vor. Weiterhin kann die Ersatzwertbildung auch beim Import der Daten in das SAP EDM durchgeführt werden. Dafür sind die entsprechenden Regeln im EDM verfügbar. Eine individuelle Konfiguration dieser Regeln nach Kundenvorgaben ist möglich. Zukünftig soll es auch die Verarbeitung von großen Datenmengen möglich sein. Zusätzlich dazu sollen Verfahren zur Be-/Verarbeitung der Messdaten in Echtzeit entwickelt werden.

4.5.2 Beschreibung der Umsetzungsalternativen

In diesem Kapitel werden die einzelnen Technologien, welche seitens SAP für die Replizierung und Bereitstellung von Daten angeboten werden, betrachtet (Vgl. Abb. 4.16). Zudem werden die Optionen „SDBC“ und „SAP HANA Extended Application Services“ betrachtet.

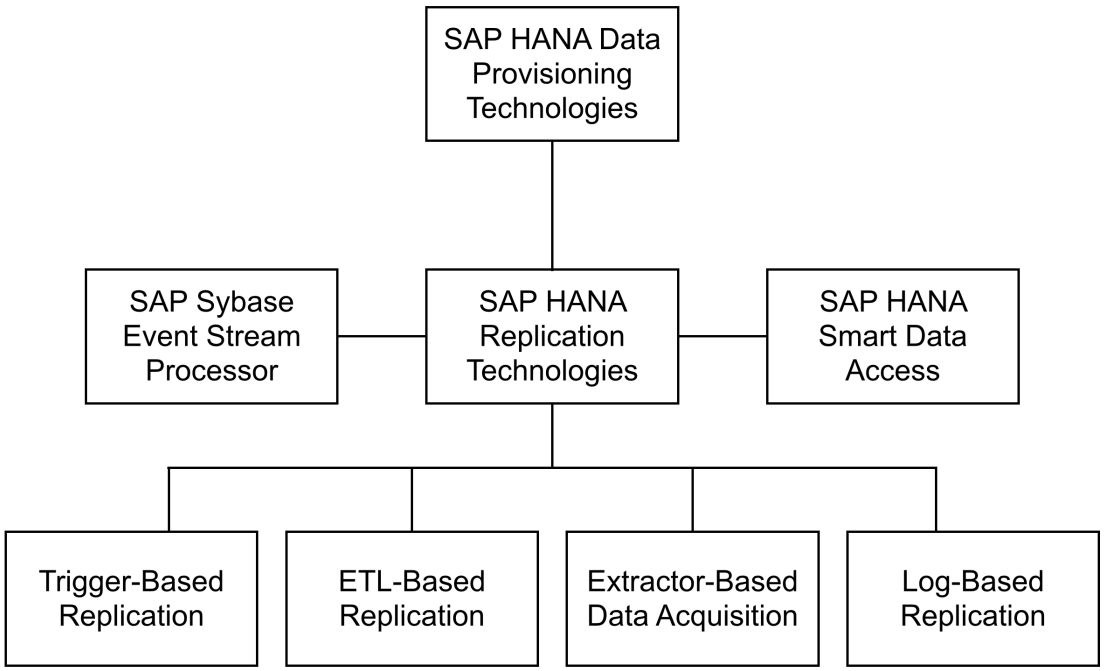


Abbildung 4.16: Betrachtete Technologien

4.5.2.1 SAP Sybase Event Stream Processor

Das U.S. amerikanische Unternehmen Sybase ist ein international tätiges Softwareunternehmen und Anbieter verschiedener Analyse- und Informationsmanagementsoftware. Es ist darüber hinaus eine Tochtergesellschaft der SAP AG. Neben weiteren Lösungen für andere Unternehmensbereiche bietet das Unternehmen seit einiger Zeit den Event Stream Processor an, welcher - wie der Name bereits vermuten lässt - hauptsächlich für die Analyse von Informationsflüssen eingesetzt wird. Dabei ist diese unabhängig von der tatsächlichen Anzahl an Input-Streams (Vgl. Abb. 4.17).

Je nachdem welches Ereignis beim Prozessor eingeht, werden definiert Routinen abgearbeitet und das Ergebnis an die dafür vorgesehenen Endanwendungen weitergeleitet. Diese können beispielsweise Dashboards oder Reporting-Anwendungen, aber eben auch eine Datenbank wie etwa SAP HANA. Der größte Vorteil bei diesem Ansatz ist sicherlich der Echtzeitansatz bei der Datenverarbeitung, da die Daten aus den angeschlossenen Quellen ereignisgesteuert und ohne zeitlichen Aufschub verarbeitet und weitergegeben werden (Vgl. SAP 2011a, S. 3).

4.5.2.2 SAP HANA Smart Data Access

Mit dem Smart Data Access bietet SAP eine Möglichkeit an, auf bestimmte Tabellen verschiedener Datenquellen, per Remotezugriff zuzugreifen. Die Originaldaten werden dabei explizit nicht in HANA kopiert, sondern lediglich für einen Fernzugriff zur Verfügung gestellt. In HANA werden jene Tabellen durch virtuelle Tabellen repräsentiert, welche mittels SQL-Befehlen mit

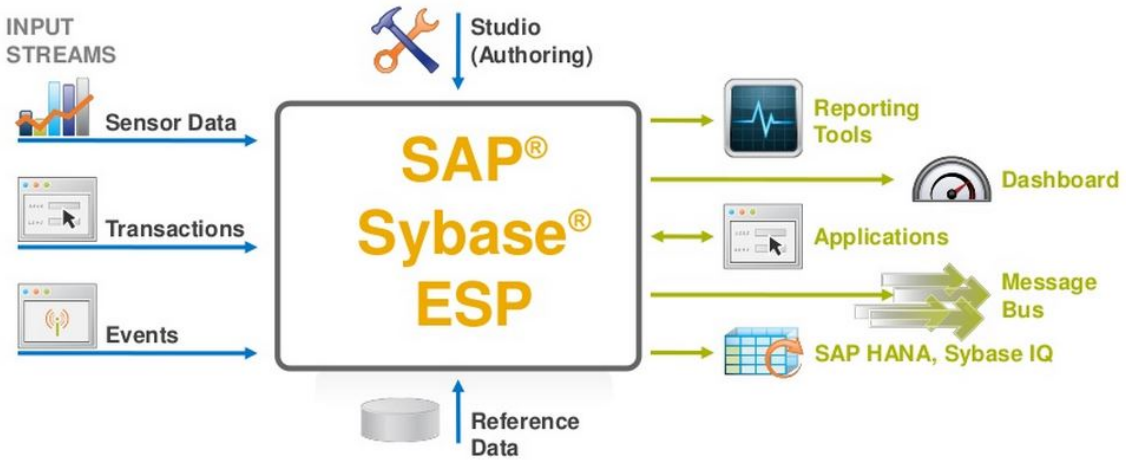


Abbildung 4.17: SAP Event Stream Processor
Quelle: SAP (2011a)

Daten gefüllt werden. Dieses Konzept ermöglicht zum einen die Anbindung verschiedener Datenquellen, sowie deren orts- und zeitunabhängigen Zugriff, zum anderen eine Performance die sich im Bereich der Echtzeitverarbeitung bewegt. Letztere liegt in dem Umgehen des unter Umständen zeitaufwändigen Kopiervorganges der zu bearbeitenden Daten begründet (Vgl. SAP 2013). Die Abbildung 4.18 zeigt einen möglichen schematischen Systemaufbau. SAP Smart Data Access wird seit SAP HANA SP6 standardmäßig mitgeliefert und kann bislang mit folgenden Datenquellen kooperieren

1. Teradata Database: version 13.0SAP
2. Sybase IQ: version 15.4 ESD3 and 16.0
3. SAP Sybase Adaptive Service Enterprise: version 15.7 ESD4
4. Intel Distribution For Apache Hadoop: version 2.3

4.5.2.3 SAP HANA Replication Technologies

In diesem Kapitel werden die Möglichkeiten dargestellt, die es erlauben Daten von einem Quellsystem nach SAP HANA zu laden. SAP bezeichnet diese Konzepte als SAP HANA Replication Technologies. Dabei werden zusätzliche Produkte genutzt, die teilweise bereits mit SAP HANA ausgeliefert werden oder als ein eigenständiges Produkt beschafft werden müssen. Dieses Kapitel soll einen Überblick über weitere Technologien geben, die neben der Trigger-basierten Replikation mit dem SAP Landscape Transformation Server, existieren. Die Technologie SLT wird aufgrund ihres Umfangs gesondert in Kapitel 4.5.2.6 behandelt.

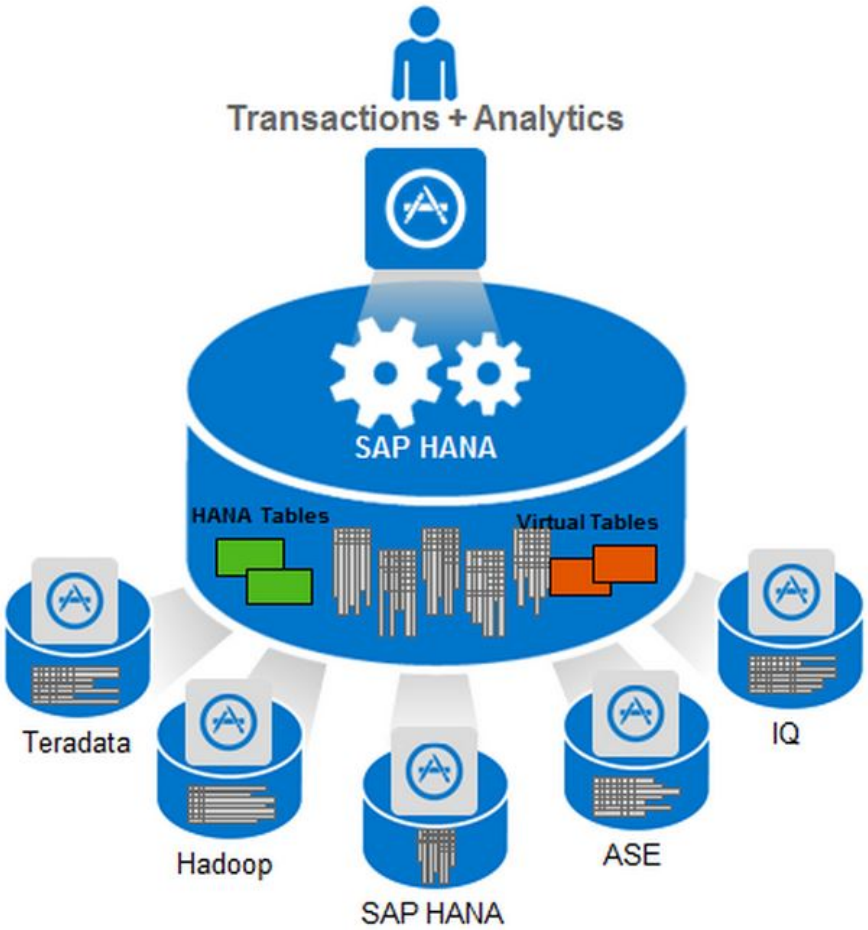


Abbildung 4.18: SAP Smart Data Access
Quelle: SAP (2013)

4.5.2.3.1 ETL-basierte Replikation

Bei der ETL-basierten Replikation (Extraktion-Transformation-Load) werden über SAP Data Services die Daten aus dem Quellsystem SAP ERP in die SAP HANA Datenbank geladen. Im SAP Data Service werden die Parameter eingestellt die nötig sind, um die Daten von der Datenstruktur im SAP ERP in die Datenstruktur der SAP HANA Datenbank zu überführen. Die ETL-basierte Replikation nutzt Batchprozesse wodurch ein Zeitplan im SAP Data Service festgelegt werden muss. Die Parameter für die Datenstruktur werden mit dem Data Service Designer festgelegt und der Zeitplan wird im Job Server festgelegt. Beides sind Komponenten des SAP Data Services. Die Daten werden jedoch nicht aus der Datenbank des SAP ERP-Systems gelesen sondern aus den Businessdaten der Anwendungsschicht. Das SAP Data Service muss zusätzlich zur SAP HANA Datenbank beschafft werden (Vgl. SAPAG 2013b, S. 38 f.).

4.5.2.3.2 Extraktor-basierte Replikation

Die Extraktor-basierte Replikation wird durch die SAP HANA Direct Extractor Connection (DXC) ermöglicht. Dadurch können die Entitäten der SAP Business Suite, über einen ETL-Prozess, in das Datenmodel von SAP HANA überführt werden. Um die Daten in das SAP HANA lesbare In-Memory DataStore Object (IMDSO) zu formatieren, müssen sie vorher für die SAP Business Suite modelliert werden. Andere Quellsysteme werden also nicht unterstützt und die Formatierung erfolgt durch User Exits die zunächst definiert werden müssen. Außerdem ist keine Echtzeitverarbeitung, sondern lediglich eine Stapelverarbeitung möglich. Ein Vorteil ist, dass der DXC zusammen mit dem SAP NetWeaver 7.0 und höher ausgeliefert wird (Vgl. SAPAG 2013b, S. 35 f.).

4.5.2.3.3 Log-basierte Replikation

Die Log-basierte Replikation nutzt den SAP Sybase Replication Server (SRS) zum übertragen und synchronisieren von Daten. Voraussetzung ist eine primäre Datenbank auf Basis von SAP Sybase Adaptive Server, Oracle, DB2 UDB, oder Microsoft SQL Server. Zur Replikation muss zusätzlich ein SAP Sybase Replication Server erworben werden (Vgl. SAPAG 2013b, S. 40 f.).

4.5.2.4 SAP HANA Extended Application Services

Seit SAP HANA SP5 beinhaltet die Distribution einen, im Rahmen der Aufgabenstellung, äußerst interessanten Dienst zur Verarbeitung von Massendaten direkt in HANA. Dieser Dienst wird als SAP HANA Extended Application Services bezeichnet und beschreibt einen Ansatz, der neben einem voll funktionsfähigem Anwendungs- und Webserver auch eine komplette Entwicklungsumgebung in der SAP HANA Distribution vereint. Der große Vorteil bei diesem Ansatz liegt in der vollständigen Integration und dem damit einhergehenden Performancegewinn im Vergleich zu anderen Anwendungen. Um die Extended Application Services Engine nutzen zu können, muss ein SAP HANA System in der Version SP5 oder höher vorliegen (Vgl. Jung 2012c).

4.5.2.4.1 Vorteile

Wie bereits angedeutet, liegt der größte Vorteil der Extended Application Services in der hohen Integrationsstufe innerhalb der HANA Distribution. Dies schlägt sich im Einzelnen dadurch nieder, dass so viel Anwendungslogik wie möglich von der Anwendungsschicht hin zur Datenbankschicht verlagert wird. Auf Grund dieses Vorgehens profitiert die entwickelte Anwendung direkt von den Vorteilen der HANA Datenbank wie etwa der In-Memory-Datenverarbeitung (Vgl. Jung 2012c).

4.5.2.4.2 Produktübersicht

Im Folgenden werden die beiden Produkte Data REST Services und XSJS vorgestellt.

OData REST Services

Eine OData-Schnittstelle ermöglicht es dem Nutzer, Daten aus dem HANA System heraus verfügbar zu machen. Der Vorteil einer OData-Schnittstelle ist, dass diese ein fest definiertes Format hat und somit mit externen Applikationen kommunizieren kann, welche das oData-Protokoll beherrschen. Angesprochen wird die Schnittstelle via REST (Representational State Transfer)-Verben. Am bekanntesten ist wohl das Verb „GET“, mit dessen Hilfe Daten abgerufen werden. Dies war bis SP5 die einzige Methode, welche via oData-Schnittstellen aufgerufen werden konnte. Seit SP6 sind nun auch „INSERT“, „UPDATE“ und „DELETE“ Operationen möglich, welche mittels der REST-Verben „POST“, „PUT“, und „DELETE“ realisiert wurden (Vgl. SAP (2011b)). Daher kann die OData-Schnittstellen nicht nur zum Lesen von Daten in der Datenbank genutzt werden, es können auch Datensätze geschrieben, geändert oder gelöscht werden. Das folgende Code-Beispiel zeigt die Erstellung einer OData-Schnittstelle auf der Datenbank CX4U, in der Tabelle Messdaten. Die Tabelle wird hierbei als „mdaten“ angezeigt.

```
1 service namespace "sap.cx4u.TestProjekt.Haupt" {
2         "CX4U"."MESSDATEN" as "mdaten";
3 }
```

Listing 4.1: Quellcode einer oData-Schnittstelle

Wie zu erkennen ist, handelt es sich hierbei um lediglich drei, sehr einfache Zeilen Code. Und schon dadurch wird eine OData-Schnittstelle definiert. Wenn diese nun mittels einer REST-Operation angesprochen werden soll, muss sich der Benutzer jedoch mit einem Benutzer des Hana-Systems authentifizieren, da die Daten sonst missbräuchlich abgefragt werden könnten.

XSJS

XSJS ist eine in SAP HANA verfügbare Programmiersprache, welche bspw. direkt auf die Datenbasis in HANA zugreifen kann und somit enormes Potenzial bietet. Auch dadurch, dass die XSJS-Engine sehr eng mit der HANA Datenbank verknüpft ist, lassen sich enorme Mengen an Daten in nahezu Echtzeit bearbeiten oder analysieren.

4.5.2.5 Secondary Database Connection

Ist eine Echtzeitverarbeitung von Daten kein notwendiges Kriterium, bietet SAP neben dem bereits erläuterten SLT-Ansatz noch eine weitere Möglichkeit an, ABAP basierte Systeme mit SAP HANA zu verbinden. Dieses, als SAP Secondary Database Connection bezeichnete Konzept, nutzt das Verfahren der Replikation, um bestimmte Datenbestände aus der vorliegenden Produktivdatenbank in die HANA Datenbank zu kopieren. Dementsprechend fungiert HANA nicht als Ersatz der bestehenden Datenbank, sondern vielmehr als eine Art Erweiterung dessen

Hauptaugenmerk auf der Beschleunigung des Lesevorganges der angeschlossenen ABAP-Systeme liegt (Vgl. Jung 2012a).

Ist eine Echtzeitverarbeitung von Daten kein notwendiges Kriterium, bietet SAP neben dem bereits erläuterten SLT-Ansatz noch eine weitere Möglichkeit an, ABAP basierte Systeme mit SAP HANA zu verbinden. Dieses, als SAP Secondary Database Connection bezeichnete Konzept, nutzt das Verfahren der Replikation, um bestimmte Datenbestände aus der vorliegenden Produktivdatenbank in die HANA Datenbank zu kopieren, wie Abbildung 4.19 verdeutlicht. Dementsprechend fungiert HANA nicht als Ersatz der bestehenden Datenbank, sondern vielmehr als eine Art Erweiterung dessen Hauptaugenmerk auf der Beschleunigung des Lesevorganges der angeschlossenen ABAP-Systeme liegt.

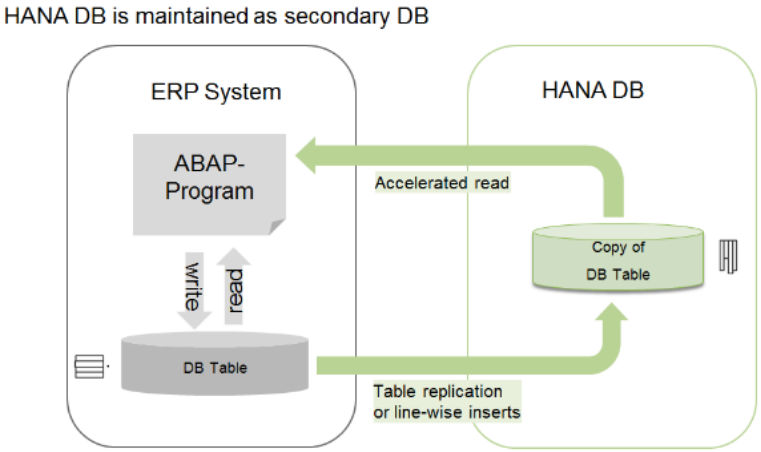


Abbildung 4.19: Secondary Database Connection
Quelle: Jung (2012a)

4.5.2.5.1 Vorteile

Den größten Vorteil, den der Einsatz einer Secondary Database Connection bietet, ist der enorm beschleunigte Datenzugriff der ABAP-basierten Anwendung. Da dieser nicht mehr auf traditionellem Wege, also mittels einer direkten Verbindung der Anwendung zur externen Datenbank, sondern über den optimierten Zugriff auf die HANA Datenbank realisiert wird, ergeben sich im direkten Vergleich Performancevorteile (Vgl. Jung 2012a).

4.5.2.5.2 Produktübersicht

Es stehen drei Möglichkeiten zu Verfügung, die im Anschluss kurz vorgestellt werden (Vgl. Jung 2012b).

Secondary Database Connection über Open SQL

Die einfachste Möglichkeit für die Nutzung von SQL-Operationen von ABAP zur Secondary Database Connection ist die Nutzung der gleichen Open SQL-Anweisungen, die durch die

ABAP-Entwicklungsumgebung bereits bekannt sind. Durch Bereitstellung einer zusätzlichen Syntax (dbcon) kann die Ausführung der Open SQL-Anweisungen gegen die alternative Datenbankverbindung erzwungen werden.

Der Vorteil dieses Ansatzes liegt in seiner Einfachheit. Mit einem kleinen Zusatz zu bereits bestehenden SQL-Anweisungen können die Transaktionen statt der gewöhnlichen Ausführung an HANA weitergeleitet werden. Nachteilig dabei ist jedoch, dass die Tabellen oder Sichten, auf die zugegriffen werden soll, bereits im ABAP Data Dictionary vorhanden sein müssen. Dies stellt jedoch kein großes Problem dar. Alle vorhandenen Daten in lokalen ABAP DBMS werden dazu in HANA repliziert. So sind immer lokale Kopien der Tabellen im ABAP Data Dictionary vorhanden. Dadurch hat man aber keinen Zugriff auf HANA-spezifische Eigenschaften, wie beispielsweise analytische Views oder Prozeduren. Ebenso kann nicht auf Tabellen zugegriffen werden, die HANA als primäre Persistenz nutzen.

Secondary Database Connection über Native SQL

ABAP bietet auch die Möglichkeit, Native SQL zu nutzen. Dabei werden die datenbankspezifischen SQL-Anweisungen genutzt. Dies ermöglicht Zugriff auf Tabellen und andere spezifische Produkte der zugrunde liegenden Datenbank. Auch in Native SQL ist eine Syntax vorhanden, die es erlaubt, besondere Datenbankprozeduren aufzurufen. Dabei ist jedoch etwas mehr Code erforderlich als bei Open SQL-Statements und es ist weniger elegant, da mit einem Datenbank-Cursor gearbeitet wird. Vorteil ist die Nutzung von Funktionen, die sonst nicht verfügbar wären. So können Daten in eine HANA-Tabelle eingefügt und HANA-Abläufe genutzt werden.

Ein Nachteil der Nutzung von Native SQL über EXEC SQL ist, dass es wenig bis gar keine Überprüfung der Syntax der SQL-Anweisungen gibt. Fehler werden erst zur Laufzeit abgefangen und können so zu einem Absturz führen. Dies macht eine manuelle Überprüfung unbedingt erforderlich.

Secondary Database Connection über Native SQL - ADBC

Die dritte Option der Anbindung bietet den Vorteil der nativen SQL Verbindung und verbessert dabei zusätzlich einige Beschränkungen. Dabei wird das Konzept von ADBC (ABAP Database Connectivity) umgesetzt. Es handelt sich hierbei um eine Reihe von Klassen sowie eine Vereinfachung und Zusammenfassung von EXEC SQL-Blöcken.

Dazu wurde das schrittweise Verarbeiten der Datenbank-Cursor durch das Lesen von ganzen Datenpakten ersetzt. Standardmäßig gibt die anfängliche Paketgröße alle resultierenden Datensätze zurück, es lassen sich jedoch auch individuelle Paketgrößen festlegen. Besonders wichtig dabei für HANA ist jedoch, dass ADBC auch den Zugriff auf nicht Data-Dictionary Produkte ermöglicht.

4.5.2.5.3 Systemvorraussetzung

Um die SAP Secondary Database Connection nutzen zu können, sind eine Reihe an Vorbedingungen seitens der Systemadministration zu erfüllen:

- Der SAP HANA Client muss auf jedem ABAP Application Server installiert sein und vom dortigen Betriebssystem unterstützt werden.
- Die SAP HANA Database Shared Library (DBSL) muss installiert sein. Diese ist lediglich ab dem ABAP Kernel 7.20 verfügbar.
- Das ABAP-System muss entweder im Unicode oder im Single Code Page 1100 vorliegen.

4.5.2.5.4 Umsetzung

Um ein SAP-Netweaver System mit einer SAP HANA Datenbank zu verknüpfen, muss wie folgt vorgegangen werden:

- Vorausgesetzt wird eine HANA-Datenbank und eine SAP-Netweaver Server. Für den SAP-Netweaver Server wurde in unserem Fall testweise ein Windows Server 2008 SP2 gewählt.
  Des Weiteren muss auf dem Netweaver-Server eine Version des Netweaver-Systems installiert werden, welche bereits eine HANA Datenbank unterstützt (in unserem Falle Relase 731). Zusätzlich wird der HDB-Client auf dem Netweaver-Server benötigt, um sich zu einer Datenbank zu verbinden.
  Aus dem Installationsverzeichnis des
  HDB-Clients muss ebenfalls die DLL „libSQLDBCHDB.dll“ in das exe-Verzeichnis des Netweavers kopiert werden.
  In unserem Fall war dies C:/usr/sap/NSP/DVEBMGS00/exe.
- Als nächstes werden in der Transaktion „DBCO“ im Netweaver-System die benötigten Informationen zum HANA-Server hinterlegt. Dieses sind Hostname inklusive Port, Benutzername und Passwort.
- Danach muss mittels OpenVPN auf dem Netweaver-System ein VPN-Tunnel zum HPI in Potsdam aufgebaut werden, damit mit dem dortigen Hana-Server kommuniziert werden kann.
- Abschließend muss in der Hana-Datenbank ein Schema angelegt werden, auf welches der in der Transaktion DBCO hinterlegte Nutzer Zugriff hat.
- mit einem einfachen ABAP-Script lässt sich nun auf diesem Schema arbeiten um bspw. Tabellen anzulegen. Die SQL-Verbindung wird z.B. festgelegt mit:
  EXEC SQL. connect to 'HANADB' as 'HANADB' ENDEXEC.
  Hierbei muss HANADB durch den Wert ersetzt werden, welcher in der Transaktion „DBCO“ unter „DB Connection“ eingetragen wurde.

4.5.2.6 SAP Landscape Transformation Replication Server (SLT)

Für Geschäftsanwendungsfälle, in denen die Datenverarbeitung in Echtzeit stattfinden soll, bietet SAP den SAP Landscape Transformation Replication Server (SLT) an. Dieser ermöglicht den Austausch von Daten zwischen SAP-Systemen und nicht-SAP-Systemen.[36] Dadurch können sowohl bestehende Systeme weiterhin genutzt als auch der Geschwindigkeitszuwachs von SAP HANA zum Tragen kommen (Vgl. SAP 2012b, Folie 4).

4.5.2.6.1 Vorteile

Beim Einsatz des SLT Replication Servers werden lediglich die Tabellen bearbeitet, welche für die Datenbereitstellung relevant sind. Dadurch bleibt der Datenaustausch performant. Über das Dashboard kann eingestellt werden ob der Datenaustausch in Echtzeit oder zu geplanten Zeitpunkten stattfinden soll. Darüber hinaus werden seitens der SAP weitere Vorteile wie folgt genannt (Vgl. Koebler 2013, S. 8):

- Datenmigration nach SAP HANA in Echtzeit
- Versionsabdeckung ab SAP R/3 4.6C mit Datenquellen von auf ABAP basierenden SAP Anwendungen
- Bearbeitung von Cluster- und Pooltabellen
- Automatische Konvertierung von nicht-Unicode nach Unicode beim laden/replizieren der Daten
- Tabelleneinstellungen und Transformationskapazitäten bieten die Funktionalität, dass beispielsweise nur bestimmte Werte repliziert werden dürfen
- Komplett integriert im SAP HANA Studio
- Erweiterte Monitoring Möglichkeiten über den SAP Solution Manager 7.1 SP5 und die mobile App SAP Replication Manager

4.5.2.6.2 Produktübersicht

Der SLT Replication Server kopiert die Daten von einem Quellsystem nach SAP HANA. Dabei wird zum Quellsystem eine RFC-Verbindung aufgebaut und das SAP HANA wird über eine DB-Verbindung verbunden. Hier wird von einem Trigger-Basiertem-Konzept gesprochen, da die Tabellen permanent überwacht werden und sobald Änderungen stattfinden, diese über den SLT Replication Server ausgetauscht werden (Vgl. Abb. 4.20).

[36] Insofern die Systeme von SAP unterstützt werden.

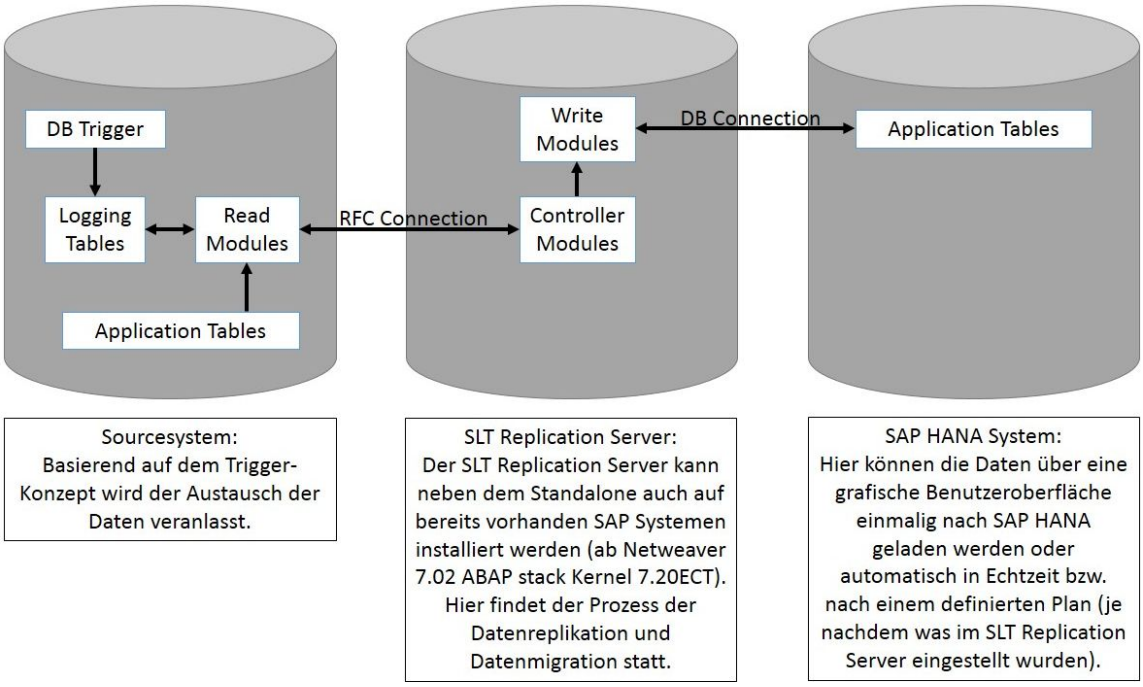


Abbildung 4.20: Produktübersicht
Quelle: Eigene Darstellung in Anlehnung an Khan (2013)

4.5.2.6.3 Beschreibung der allgemeinen Architektur

Im Folgenden soll die grundsätzliche Architektur eines SLT-Systems zum Herstellen einer Datenreplikation zwischen SAP HANA und einem anderen System beschrieben werden. Das SAP LT Replikations Konzept verwendet hierzu einen auf Triggern basierenden Ansatz. Es muss jedoch unterschieden werden, ob es sich bei dem Quell-System um ein SAP- oder ein nicht-SAP-System handelt (Vgl. SAP 2012a, Folie 10).

Betrachtet wird zunächst der Fall, dass ein SAP-System verwendet wird. Im SAP-Quellsystem beginnt die Initialisierung der Datenreplikation. Mittels einer RFC-Verbindung werden die Daten mit dem SLT, der mindestens einem Netweaver mit Version 7.02 entspricht, ausgetauscht. Das SLT-System bietet in diesem Fall eine skalierbare und laut Herstellerangaben zuverlässige Replikation, einschließlicher umfangreicher Funktionen zur Datentransformation, die bereits zur Laufzeit ausgeführt werden. Über eine DB Connection werden die Daten dann an das HANA weitergegeben (Vgl. SAP 2012a, Folie 10).
In dem Fall, dass ein nicht-SAP-System als Quellsystem verwendet wird, wird auch bereits zur Verbindung mit dem SLT eine DB Connection verwendet (Vgl. Abb. 4.22). In einem ersten Schritt überträgt der SAP LT Replication Server alle Table Definitions als Metadaten aus dem nicht-SAP-Quellsystem in das SAP HANA System. Aus dem HANA Studio heraus funktioniert die nicht-SAP Datenreplikation nun äquivalent wie mit SAP Quellen. Sobald eine Tabellen Replikation gestartet wurde, erstellt SLT Logging-Tabellen im Quellsystem. Als Unterschied zum zuvor beschriebenen Vorgehen, werden die Read Modules im SLT erstellt (Vgl. SAP

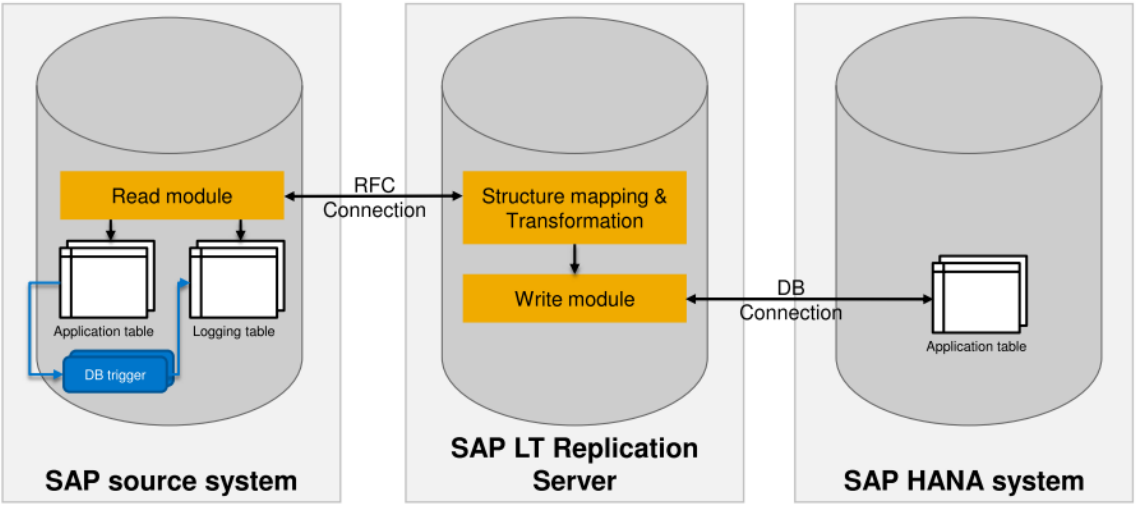


Abbildung 4.21: SLT mit SAP-Quellsystem
Quelle: SAP (2012a)

2012a, Folie 11).

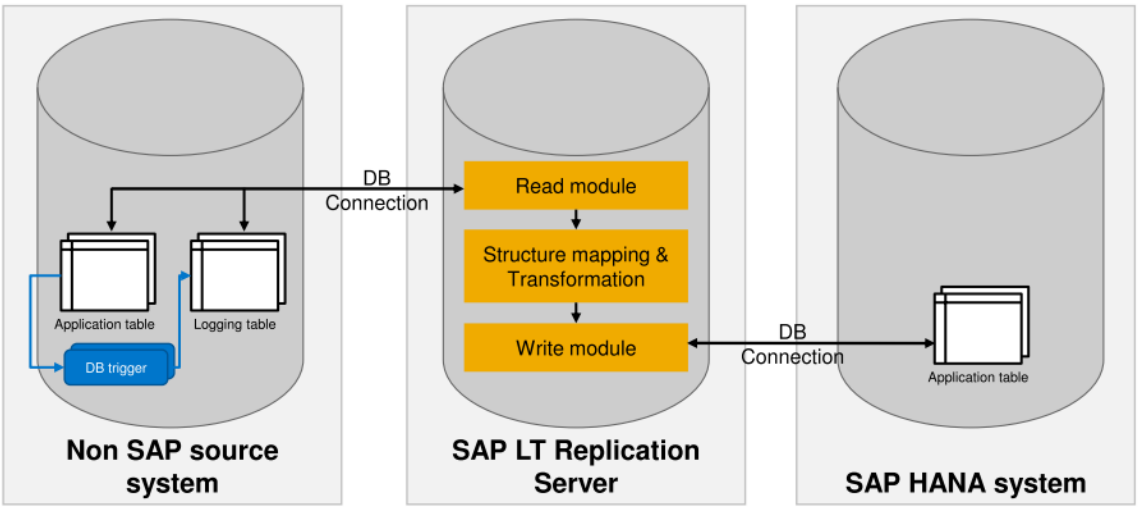


Abbildung 4.22: SLT mit SAP-fremden-Quellsystem
Quelle: SAP (2012a)

4.5.2.6.4 Multi System Support

SLT bietet einen so genannten Multi System Support. Hieraus ergeben sich vier verschiedene Varianten zur Zusammenstellung von Architekturen (Vgl. SAP 2012a, Folie 12 f.):

- Quellsysteme sind mit verschiedenen HANA-Daten-Schemata auf dem gleichen HANA-System verbunden.
- Quellsysteme sind mit verschiedenen HANA-Systemen verbunden. Der Schema Name kann hierbei identisch oder unterschiedlich sein.
- Quellsysteme sind mit dem selben HANA-System und Schema verbunden.
- Ein SAP Quellsystem ist mit verschiedenen HANA-Systemen oder einem HANA-System mit unterschiedlichen Schemata-Namen verbunden.

4.5.2.6.5 Grundlegende Konzepte der Trigger-basierten Datenreplikation mit SAP LT Replication Server für SAP HANA

Der SAP Landscape Transformation (SLT) Replication Server ist die SAP-Technologie, die das Laden und Replizieren von Daten in Echtzeit aus SAP-Quellsystemen und Nicht-SAP-Quellsystemen an eine SAP-HANA-Umgebung ermöglicht. Der SAP SLT Replication Server verwendet einen Trigger-basierten Replikations Ansatz, um Daten aus dem Quellsystem an das Zielsystem zu übertragen. Der Trigger-basierte Datenreplikations-Ansatz ist ein tabellenbasiertes Konzept, welches Daten aus den Anwendungstabellen im Quellsystem abruft. Das SAP HANA Studio wird verwendet, um die Tabellen für die Replikation auszuwählen. Wenn nicht alle Datensätze einer Tabelle übertragen werden sollen, können Transformationsregeln erstellt werden, um selektiv die Daten zu filtern (selektive Datenreplikation) oder um andere Transformationen während der Datenreplikationsprozesse zu ermöglichen.

Der SAP SLT Replication Server kann entweder als separates SAP-System installiert werden, oder auf einem SAP-Quellsystem, wenn die technischen Voraussetzungen gegeben sind. Um Daten zu replizieren, müssen die Parameter, die der SAP SLT Replication Server verwendet, definiert werden. Die Informationen werden in einer Konfiguration definiert. Diese Konfiguration enthält Informationen über das Quellsystem, das Zielsystem und die entsprechenden Verbindungen und sie wird im SAP SLT Replication Server definiert. Die Datenreplikation erfolgt zwischen dem Quellsystem und dem Zieldatenbankschema eines HANA Systems in drei möglichen Beziehungen: (1:1), (N:1), (4:N).

4.5.2.6.6 Configuration and Monitoring Dashboard (Transaktion LTR)

Das Configuration and Monitoring Dashboard wird verwendet, um die Konfigurationen zu erstellen und die Statusinformationen der Replikation zu überwachen.
Das erstellen der Konfiguration erfolgt durch die folgenden Schritte:

1. Angeben der allgemeinen Daten.
In diesem Schritt wird der Konfigurationsname angegeben. Das Schema, das automatisch auf das Zielsystem erstellt wird, hat den gleichen Namen. Zusätzlich wäre es in diesem Schritt möglich, eine Beschreibung und die Autorisierungsgruppe anzugeben.

2. Angaben zum Quellsystem.
Wie bereits erwähnt, kann das Quellsystem entweder ein SAP- oder nicht-SAP-Quellsystem sein. In diesem Schritt wird das Quellsystem und die entsprechenden Kommunikationseinstellungen angegeben.

3. Angaben zum Zielsystem.

- Verbindung zum SAP HANA-System. Hier werden die entsprechenden In-Memory-Datenbank Verbindungsinformationen angegeben.
- Verbindung zum SAP-System über RFC-Connection. Die RFC-Verbindung zu dem SAP-Zielsystem wird definiert.

4. Angeben der Übertragungseinstellungen.
Um die Daten erfolgreich zu übertragen, sollten einige Informationen definiert werden.

- Initial Load Mode: Es gibt verschiedene Verfahren um auf die Daten im Quellsystem zuzugreifen. Diese Reading types werden benutzt, um das Laden von Daten zu beschleunigen.
- Dataclass of Tablespace: Ein Tablespace wird für die logging Tabellen im Quellsystem definiert.
- No. of Data Transfer Jobs: Die Anzahl der Datenübertragung-Jobs wird angegeben.
- No. of Initial Load Jobs: Die Anzahl der Initial-Load -Jobs wird angegeben.
- No. of Calculation Jobs Die Anzahl der Kalkulation-Jobs wird angegeben.

Zusätzlich zu den vorherigen Informationen, soll die Replikationsoption angegeben werden. Es gibt drei verschiedene Optionen:

- Echtzeit: Die Trigger-basierte Datenreplikation Methode, die kontinuierlich alle Datenänderungen an das Zielsystem repliziert.
- Geplant nach Intervall: Zeit und Frequenzbereich für die Replikation werden angegeben. Zum Beispiel alle 30 Minuten oder alle 12 Stunden.
- Geplant nach Zeit: Eine bestimmte Zeit für die Replikation wird angegeben, z. B. 23:00 Uhr.

5. Überprüfen und Erstellen.

4.5.2.6.7 SAP SLT Replication Server Jobs

Im SLT Replication Server finden sich für die verschiedenen Aufgaben, wie Überwachung und Laden von Daten, so genannte Jobs. Diese werden im Folgenden beschrieben.

Master Job (Überwachung Job):
Der Master-Job verarbeitet alle aktiven Konfigurationen.

Master Controller Job:
Dieser Job wird bei Bedarf geplant und ist verantwortlich für:

- Erstellen des Datenbank-Triggers und der Logging-Tabelle im Quellsystem.
- Erstellen von Synonymen.
- Schreiben neuer Einträge in Admin-Tabellen im SLT Server, wenn eine neue Tabelle geladen / repliziert wird.

Data Load Job:
Dieser Job sollte immer aktiv sein und er ist verantwortlich für:

- Datenladen (Laden)
- Replikation von Daten (Replikation)
- Ändern Status-Flag für Einträge in Steuertabellen im SAP SLT Replication Server

Migration Object Definition Job:
Dieser Job definiert das Migrations-Objekt einer bestimmten Tabelle. Die Migrations Objekt-Definition sollte in der Regel recht schnell für alle Tabellen sein.

Access Plan Calculation Job:
Dieser Job berechnet den Zugriffsplan einer bestimmten Tabelle. Für eine normale Größe sollte die Tabelle die Zugangs-Planrechnung ganz schnell beenden (weniger als 1 Minute), während große Tabellen bis zu mehreren Stunden dauern können.

4.5.2.6.8 Erweiterte Replikationseinstellungen

Dieses Programm läuft auf dem SAP SLT Replication Server und es wird verwendet um die erweiterten Replikationseinstellungen anzugeben. Zum Beispiel:

- Modifizierung der Strukturen der Zieltabelle
  - Selektive Datenreplikation / Filterung
  - Änderung der Struktur der Tabelle (addieren, löschen oder mischen von Datenfeldern)
  - Datentransformation
  - Spezielles UI um die Einstellungen der Tabelle anzugeben

- Angeben der Einstellungen der Performanceoptimierung: Durch die erweiterten Replikationseinstellungen, ist es möglich die Leistung des Initial Load der einzelnen Tabellen zu Optimierung.
- Definieren der Transformationsregeln: Die Transformationsregeln transformieren die Daten während der Replikation. Die transformierten Daten werden dann in der Zieltabelle dargestellt. Zum Beispiel ist es möglich Regeln zu definieren für:
  - Konvertieren von Felder
  - Füllen leerer Felder
  - Datensätze zu überspringen

4.5.2.6.9 Zusätzliche Funktionen

Neben der bereits beschriebenen Funktionalität bietet SAP SLT folgende zusätzliche Funktionen:

- Laden von SAP Archiv.
  - Architektur und Key Building Blocks
  - Definieren Ladeobjekt von SLT-System
- Überwachung mit SAP Solution Manager
- LT Replication Manager - Mobile Application
- Technische Enabler für Multiple Data Provisioning Use Cases
- SLT und HANA in der Cloud

4.5.3 Bewertung der Umsetzungsalternativen

Die verschiedenen Umsetzungsalternativen haben ihre Vorteile und Nachteile. Um den Zweck der Machbarkeitsstudie zu erfüllen, werden die Alternativen durch mehrere Kriterien in den Tabellen 4.42 und 4.43 bewertet. Die Kriterien untersuchen die drei Hauptkomponenten, die in allen Replikationsszenarien beteiligt sind: Das Quellsystem, das Zielsystem und die Funktionen der Datenreplikation. Es ist sehr wichtig, die jeweilige Methode bezüglich der Unterstützung nicht-SAP-basierter Quellsysteme, der Unterstützung verschiedener Datenbankquellen und der nötigen Systemvoraussetzungen zu untersuchen. Auf der anderen Seite werden die Transformationsfunktion und die Fähigkeit, die Daten in Echtzeit bzw. durch vordefinierte Intervalle oder geplant zu replizieren, bewertet. Die Kosten spielen eine große Rolle, deshalb wird bei jeder Methode betrachtet, ob sie Anschaffungs- bzw. Betriebskosten hat.

Kriterien	**Trigger-Based Replication (SLT Replication)**	**ETL-Based Replication (Data Services 4.0 – SP2)**	**Extractor-based Data Acquisition (DXC)**	**SAP Sybase Replication Server (SRS)**
Echtzeit, Intervall/Geplant	Echtzeit und Geplant (auf Tabellenebene)	Geplant	Geplant oder Ereignisgesteuert	Echtzeit
Datenquellen	Alle SAP unterstützten Datenbanken, inkl. ASE	Alle SAP unterstützten Datenbanken, inkl. ASE, und andere. (keine MaxDB Unterstützung)	Alle SAP unterstützten Datenbanken	SAP/nicht-SAP Datenquellen: Adaptive Server, Oracle, MS SQL Server, IBM DB2 UDB, SAP ASE
Systemvoraussetzung	SAP R / 3 4.6C - SAP ERP 6.0 (EHP06) Alle anderen ABAP-basierten SAP-Anwendungen (Basis 4.6C-NW7.02)	ERP 4.6c - SAP ERP 6.0	SAP Business Suite Systeme auf Basis von NetWeaver ABAP 4.6C oder höher	Linux - Windows
Unterstützung nicht-SAP-basierter Systeme	Ja (nur für SAP unterstützten Datenbanken)	Ja	Nein	Ja

Transformations-funktionen	Funktionen zur Filterung und Daten-transformation, sowie Daten-Scrambling. Datenfilterung kann entweder über selektive Trigger oder über Replikati-onskonfigurati-onseinstellungen vorgenommen werden	Vollständige ETL-Engine von einfachen Funk-tionen bis zu sehr komplexen Transformatio-nen	Limitiert für die Extraktion, über User-Exits	Erfasst Tabelle von Low-Level-Datenbank-Logfiles und wandelt sie in SQL-Statements, die wiederum auf der SAP HANA Datenbank aus-geführt werden. Aber limitiert für Unicode-Quellendatenbanken
Kosten	Ja (Kommt mit HANA Enterpri-se oder Enterpri-se Extended Edi-tions)	Ja (allein ab 50000 EUR oder kommt mit HA-NA Enterprise oder Enterpri-se Extended Editions)	Ja (SAP Busi-ness Suite ist er-forderlich)	Ja (allein 65000 EUR oder kommt mit HANA Ent-erprise Extended Editions)

Tabelle 4.42: Vergleich der Umsetzungsalternativen 1

Kriterien	**SAP Sybase Event Stream Processor (ESP)**	**SAP HANA Smart Data Access**	**SAP HANA Extended Application Services (XS)**	**Secondary Database Connection**
Echtzeit, Intervall/Geplant	Echtzeit	Echtzeit	Echtzeit	-
Datenquellen	Eingabeströme: Sensordaten, Transaktionen, Ereignisse Streams	Teradata database, SAP Sybase ASE, SAP Sybase IQ, Intel Distribution for Apache Hadoop, SAP HANA	Alle von SAP unterstützten Datenbanken	SAP HANA DB
Systemvoraussetzung	Windows – Linux – Solaris. SAP HANA Plugin	In SAP HANA integriert	SAP HANA System in der Version SP5 oder höher	Installieren des SAP HANA Clients auf jedem ABAP Application Server. Installieren des SAP HANA DBSL. Unicode oder Single Code Page 1100 ABAP System
Unterstützung nicht-SAP-basierter Systeme	Ja	Ja	Ja	Nein

Transformations-funktionen	Funktionen zur Filterung, Aggregation, Zusammenfassung, Join, Merge, und Patterns Finden	automatische Datentypkonvertierung ohne die Daten in die SAP-HANA-Datenbank zu replizieren	Durch OData-REST-Services ist es möglich, die HANA Tabellen oder Views zu entlarven und sie aus dem Web-Browser oder über eine Benutzeroberfläche zu verbrauchen. OData-Services unterstützen CREATE-, READ-, UPDATE- und DELETE-Operationen mit Hilfe der HTTP-Verben	-
Kosten	Ja (Kaufoptionen erfordern Kontaktieren des Verkaufsteams)	-	-	-

Tabelle 4.43: Vergleich der Umsetzungsalternativen 2

Nach der Bewertung der verschiedenen Umsetzungsalternativen ist deutlich geworden, dass die meisten Verfahren zusätzliche Aufwendungen und spezielle Produkte benötigen. Unter den restlichen kostenlosen Alternativen passt SAP HANA Extended Application Services (XS) am besten für die Umsetzung des Projekts. XS erfüllt den Zweck der Umsetzung und es erfordert lediglich einen HTTP-Aufruf von der Netweaver-Seite her auf eine OData-Schnittstelle auf der HANA-Seite.

4.5.4 Prototypen

In diesem Kapitel wird gezeigt, wie und mit welchen Technologien das Projekt umgesetzt wurde. Hierbei haben wir uns auf HANA-Seite für den SAP HANA Extended Application Service entschieden. Auf dem NetWeaver wurde der Abruf mittels einer Programmierung realisiert.

4.5.4.1 Prototyp: HANA

Auf der HANA-Seite haben wir uns für den SAP HANA Extended Application Service entschieden. Gründe für diese Entscheidung waren unter anderem:

- Ist bereits im System enthalten, daher kostenlos.
- Da sich die XS-Engine direkt auf der Datenbank befindet, sind die Zugriffszeiten besser als wenn man bspw. eine Fremdsoftware gewählt hätte.
- Im Einführungskurs wurden bereits Beispiele mit dieser Technologie gezeigt.
- Die OData-Schnittstelle ist ein offenes Format, welches die Datenabfrage per GET-Methode ermöglicht.
- Das Programmieren mit der XS-Engine ist relativ simpel, da die Programmiersprache an JavaScript angelehnt ist.

Im Folgenden soll sowohl die Struktur der Datenbank als auch die Programmierung im Einzelnen gezeigt werden.

4.5.4.1.1 Datenbank

Wie bereits im Kapitel 4.4.5 dargestellt wurde, wurden für den Prototypen in HANA vier Tabellen angelegt.
Die wohl wichtigste dieser Tabellen ist die Tabelle „Messdaten“, da hier die Messungen aller Zähler zusammenlaufen und da auf diesen Daten letztlich auch die VEE-Komponente ausgeführt wird.
Wie man in Abbildung 4.23 erkennen kann, existieren in der Tabelle „Messdaten“ vier Spalten. In der Spalte „ZAEHLERID“ wird die eindeutige ID des Zählers gespeichert. In der Spalte „MESSWERT“ wird der gemessene Stromverbrauch gespeichert und in der Spalte „ZEITSTEMPEL“ wird der Zeitstempel der Messung gespeichert. Der Primärschlüssel setzt sich aus der ZählerID und dem Zeitstempeln zusammen, da auf diese Art und Weise jeder Eintrag eindeutig definiert ist.
Abschließend besitzt die Tabelle noch eine Spalte „STATUS“, in welcher der Status des Datensatzes definiert wird. Hier wird bspw. eingetragen, ob der Wert als fehlerhaft erkannt wurde, ob er bereits korrigiert wurde oder ob der Zeitstempel evtl. fehlerhaft ist.

ZAEHLERID	MESSWERT	ZEITSTEMPEL	STATUS
170.589	0,00876	26.02.2014 18:00:00.000	raw
170.589	0,00543	26.02.2014 18:15:00.000	raw
170.589	0,00854	26.02.2014 18:45:00.000	raw
170.589	0,00343	26.02.2014 20:30:00.000	raw
170.589	0,00123	26.02.2014 22:45:00.000	raw
170.589	0,0061312499999...	26.02.2014 18:30:00.000	raw
170.589	0,00854	26.02.2014 19:00:00.000	raw
170.589	0,00854	26.02.2014 19:15:00.000	raw
170.589	0,0028245915735...	26.02.2014 20:00:00.000	raw
170.589	0,0076302665520...	26.02.2014 20:15:00.000	raw
170.589	0,0044590909090...	26.02.2014 20:45:00.000	raw

Abbildung 4.23: Beispieldaten aus der Tabelle „Messdaten“

4.5.4.1.2 Webinterface

Das Webinterface setzt sich eigentlich aus drei verschiedenen Seiten zusammen, welche jeweils eine der Funktionen der VEE-Komponente übernehmen. Diese einzelnen Funktionen könnten dann in Zukunft automatisiert vom Smart Gateway aufgerufen werden. Übergeben wird diesen Seiten beim Aufruf der Parameter „zaehler“, welcher die ZählerID enthält.
Um den Aufruf zu vereinfachen wurde ein Webinterface in HTML und JavaScript geschrieben, welches asynchron die gewünschte Funktion mit der jeweiligen ZählerID aufruft. Wie in Abbildung 4.24 zu erkennen ist, muss lediglich der auszuwertende Zähler angegeben werden und via Klick auf die jeweilige Funktion wird diese asynchron im Hintergrund (mittels der JavaScript Bibliothek "JQuery“) aufgerufen und die Ausgabe angezeigt. Dies sorgt dafür, dass man alle Funktionen von einer Seite aus aufrufen kann ohne diese Seite zu verlassen. Ausgegeben werden Informationen wie bspw. gefundene Fehler, ersetzte Werte etc.

4.5.4.1.3 OData REST Service

Damit Daten vom Netweaver aus der HANA-Datenbank abgefragt oder in die HANA-Datenbank eingefügt werden können, wird eine OData-Schnittstelle benötigt. Diese erlaubt den Zugriff auf die Tabelle „Messdaten“. Das Vorgehen hierzu wurde in Kapitel 4.5.2.4.2 beschrieben.

4.5.4.1.4 Validation Funktion

Für den Zweck des Prototyps wird die Validation-Funktion wie auch die anderen VEE-Funktionen von der Benutzeroberfläche für einen bestimmten Zähler aufgerufen (Vgl. Listing 4.2).

```
1 var zaehler = $.request.parameters.get('zaehler');
```

Listing 4.2: Erhalten der ZählerID von der Benutzeroberfläche

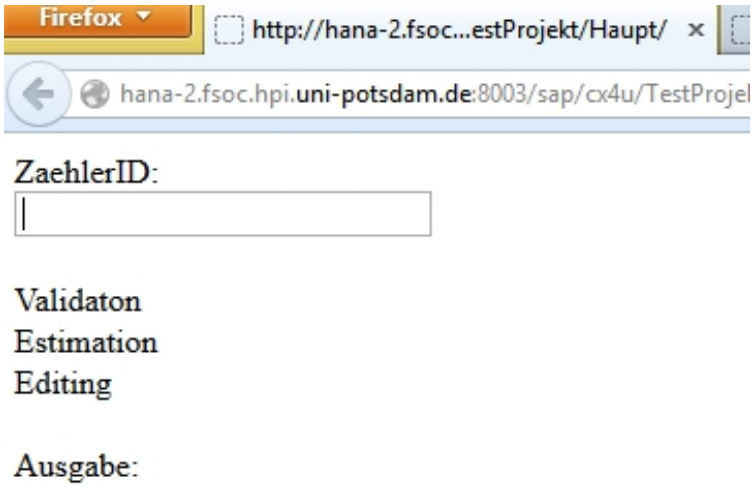


Abbildung 4.24: VEE-Webinterface

Anschließend wird die Verbindung mit der SAP HANA Datenbank hergestellt. Dann werden die Daten für den gewünschten Zähler durch einen SQL-SELECT-Befehl angefordert und in einem ResultSet gespeichert (Vgl. Listing 4.3).

```
1 var conn = $.db.getConnection("sap.cx4u.TestProjekt.Haupt::Verbindung");
2 var pstmt = conn.prepareStatement("SELECT * FROM \"CX4U\".\"MESSDATEN2\"
    WHERE \"ZAEHLERID\" LIKE " +zaehler+ " ORDER BY \"ZEITSTEMPEL\" ASC");
3 var rs = pstmt.executeQuery();
```

Listing 4.3: DB Verbindung

Im nächsten Schritt wird die vordefinierte Validation-Regel des Zählers ausgewählt. Optimalerweise liegen alle Regeln in der Datenbank und sind bestimmten Zählern zugeordnet, im Kontext des Prototypen aber, wird nur eine Regel definiert und direkt im Quellcode eingebunden. Die Messdaten innerhalb des ResultSets werden durch eine While-Schleife nach einander geprüft, um Inkonsistenzen zu bestimmen. Durch die While-Schleife wird jede einzelne Messung auf das Auftreten von Nullwerten, überschrittene Werte oder Unregelmäßigkeiten des Zeitstempels untersucht. Wenn einer dieser Fehler gefunden wurde, wird die Messung als fehlerhaft markiert und der Status wird in der Datenbank für die weitere Verarbeitung geändert. Am Ende der Schleife wird der Status der Messung überprüft. Wenn der Status nicht „raw“ ist, also ein Fehler aufgetreten ist, wird die Änderung in der Logging-Tabelle dokumentiert (Vgl. Listing 4.4).

```
1 while (rs.next()) {
2   var mw = rs.getString(2);
3   var zs = rs.getTimestamp(3);
4   var status = rs.getString(4);
5
6     if(nullposs === 0 && mw === '0')
7     {
8       log = log + "<p><b> NULLWERT: Zaehler "+rs.getString(1)+"/
```

```
          Zeitstempel "+zs+"/ WERT "+mw+" </b></p>";
9         status = "null value";
10      }
11
12      if(mw >= maxval) {
13        log = log + "<p><b> WERTUEBERSCHREITUNG: Zaehler "+rs.getString(1)+"/
            Zeitstempel "+zs+"/ WERT "+mw+" <b></p>";
14        status = "exceeded value";
15      }
16
17      if(prevstamp===="") {
18        prevstamp=zs;
19      } else {
20        jetztiger=zs;
21
22        var diff = Math.abs(jetztiger - prevstamp);
23        diff = diff /60000;
24        var jetztMin = jetztiger.getMinutes();
25
26        if((diff !== 15) && !((jetztMin === 0) || (jetztMin === 15) || (
            jetztMin === 30) || (jetztMin === 45))){
27          log = log + "<p><b> UNREGELMAESSIGKEIT vor: Zaehler "+rs.getString
              (1)+"/ Zeitstempel "+zs+"/ WERT "+mw+" </b></p>";
28          status = "irregularity";
29        }
30
31        if(status !== "raw"){
32          var pstmt2 = conn.prepareStatement("UPDATE \"CX4U\".\"MESSDATEN2\"
              SET \"STATUS\" = '" + status + "' WHERE \"ZAEHLERID\" = "+ rs.
              getString(1)+" AND \"ZEITSTEMPEL\" = '"+ rs.getString(3) +"'");
33          pstmt2.executeUpdate();
34
35          var pstmt4 = conn.prepareStatement("INSERT INTO \"CX4U\".\"LOGGING
              \" VALUES (\"CX4U\".logseq.NEXTVAL, CURRENT_UTCTIMESTAMP, '" +
              status + "', " + rs.getString(1) + ", " + rs.getString(2) + ",
              '" +  rs.getString(3) + "')");
36          pstmt4.executeUpdate();
37        }
38        prevstamp=zs;
39        prevmw=mw;
40      }
41  }
```

Listing 4.4: Datenvalidierung durch die While-Schleife

4.5.4.1.5 Estimation Funktion

Die Daten des eingegebenen Zählers werden wieder von der HANA Datenbank abgerufen und in einem ResultSet gespeichert, um die Ersatzwertbildung auszuführen. Die Ersatzwertbildung

erfolgt durch verschiedene vordefinierte Regeln für jeden Zähler. In diesem Prototyp werden zwei Fehlerfälle behandelt und ihre Werte werden ersetzt. Diese sind: das Auftreten von Nullwerten oder von fehlenden Intervallmessungen. Der fehlerhafte oder fehlende Wert wird durch den letzten gültigen Wert ersetzt. Die beiden Ersetzungs-Funktionen werden innerhalb einer While-Schleife durchgeführt. Wenn keine Nullwerte in den Messungen erlaubt sind und dennoch ein Nullwert auftritt und die vorherige Intervallmessung nicht null ist, dann wird eine Ersetzung durchgeführt und der Nullwert wird mit der vorherigen Messung ersetzt (Vgl. Listing 4.5).

```
1     if(nullposs === 0 && mw === '0' && prevmw !== "")
2     {
3       var status = "null value replaced";
4       mw = prevmw;
5     var pstmt2 = conn.prepareStatement("UPDATE \"CX4U\".\"MESSDATEN2\" SET
        \"MESSWERT\" = " + mw + ", \"STATUS\" = '" + status + "' WHERE \"
        ZAEHLERID\" = "+ rs.getString(1)+" AND \"ZEITSTEMPEL\" = '"+ rs.
        getString(3) +"'");
6     pstmt2.executeUpdate();
7
8         var pstmt3 = conn.prepareStatement("INSERT INTO \"CX4U\".\"LOGGING
            \" VALUES (\"CX4U\".logseq.NEXTVAL, CURRENT_UTCTIMESTAMP, '" +
            status + "', " + rs.getString(1) + ", " + rs.getString(2) + ",
            '" +  rs.getString(3) + "')");
9         pstmt3.executeUpdate();
10
11        log = log + "<p><b> The generated replacement value for the smart
            meter: "+rs.getString(1)+"at the timestamp: "+zs+" is: "+mw+"
            </b></p>";
12    }
```

Listing 4.5: Nullwerte Ersatzwertbildung

Außerdem wird durch den Vergleich zwischen dem aktuellen Zeitstempel und dem vorherigen Zeitstempel die Existenz einer fehlenden Messung überprüft. Dann wird eine neue Messung mit dem Messwert der letzten gültigen Messung hinzugefügt. Durch Addieren des Intervallwerts, in unserem Fall 15 Minuten, zum vorherigen Zeitstempel, wird der Zeitstempel des Ersatzwerts definiert (Vgl. Listing 4.6).

```
1     if(prevstamp==="") {
2         prevstamp=zs;
3     } else {
4       jetztiger=zs;
5
6         var diff = Math.abs(jetztiger - prevstamp);
7         diff = diff /60000;
8         var jetztMin = jetztiger.getMinutes();
9
10        if(diff > 15){
11          addzs = new Date(prevstamp.getTime() + 15*60000);
12          var year = addzs.getFullYear();
```

```
13          var month = addzs.getMonth() < 9 ? '0' + (addzs.getMonth() + 1) :
                 addzs.getMonth() + 1;
14          var day = addzs.getDate() < 10 ? '0' + addzs.getDate() : addzs.
                getDate();
15
16          var TSdate = year + "/" + month + "/" + day;
17          var TStime = addzs.toLocaleTimeString();
18          addzs = TSdate + " " + TStime;
19
20          var pstmt4 = conn.prepareStatement("INSERT INTO \"CX4U\".\"
                MESSDATEN2\" VALUES (" + rs.getString(1) + ", " + prevmw + "
                ,'" + addzs + "','added measurement')");
21          pstmt4.executeUpdate();
22
23          var pstmt5 = conn.prepareStatement("INSERT INTO \"CX4U\".\"
                LOGGING\" VALUES (\"CX4U\".logseq.NEXTVAL,
                CURRENT_UTCTIMESTAMP, 'added measurement', " + rs.getString
                (1) + ", " + rs.getString(2) + ", '" +  rs.getString(3) + "')
                ");
24          pstmt5.executeUpdate();
25
26          log = log + "<p><b> This measurement should be added: ZaehlerId "
                 + rs.getString(1)+ "  Messwert " + prevmw + "  Zeitstempel "
                 + addzs + "  Status 'added measurement' </b></p>";
27        }
28
29        prevstamp=zs;
30        prevmw=mw;
31    }
```

Listing 4.6: Fehlende Werte Ersatzwertbildung

Alle Änderungen werden in der Datenbank aktualisiert und in die Logging-Tabelle eingefügt.

4.5.4.1.6 Editing Funktion

Wie bei der Validation Funktion werden im ersten Schritt die Daten des eingegebenen Zählers von der HANA Datenbank abgerufen und in einem ResultSet gespeichert. Anschließend werden die Zeitstempel der Messdaten in einer While-Schleife geprüft und gegebenenfalls werden die fehlerhaften Zeitstempel und ihre entsprechende Messwerte korrigiert. Die Zeitstempel werden nach einer vordefinierten Intervallmessung geprüft. In unserem Fall wurde ein Intervall von 15 Minuten betrachtet. Die Minuten und die Sekunden werden von dem Zeitstempel extrahiert und auf Basis des 15-Minuten-Intervalls wird der Zeitstempel durch eine IF-Bedingung untersucht. Wenn die Sekunden nicht null sind oder wenn die Minuten nicht null, „15“, „30“, oder „45“ sind, muss der Zeitstempel korrigiert werden. Danach wird die Differenz zwischen dem fehlerhaften und dem korrekten Zeitstempel gespeichert um den Messwert entsprechend zu korrigieren. Anschließend wird der falsche Minutenwert mit dem neuen korrigierten Minutenwert im Zeitstempel ersetzt und der Sekundenwert wird auf null eingestellt (Vgl. Listing 4.7).

```
var mw = rs.getString(2);
var zs = rs.getTimestamp(3);
var min = zs.getMinutes();
var sec = zs.getSeconds();

var newzs = zs;
var corMw = mw;

if (sec !== 0 ||  !((min === 0) || (min === 15) || (min === 30) || (min
    === 45))){
  var minDiff = "";
  var corMin = "";

  if (min > "15" && min < "30"){
    minDiff = min;
    corMin = min - 15;
  }
  else if (min > "30" && min < "45"){
    minDiff = min - 15;
    corMin = min - 30;
  }
  else if (min > "45" && min < "60"){
    minDiff = min - 30;
    corMin = min - 45;
  }
  else if (min > "0" && min < "15"){
    minDiff = min + 15;
    corMin = min;
  }

  newzs = new Date(zs.getTime() - corMin*60000);
  newzs.setSeconds(0);
```

Listing 4.7: Zeitstempel Korrektur

Der korrigierte Zeitstempel wird danach ins UTC-Zeitstempel-Format konvertiert, um ihn wieder in die Datenbank einfügen zu können (Vgl. Listing 4.8).

```
  var year = newzs.getFullYear();
  var month = newzs.getMonth() < 9 ? '0' + (newzs.getMonth() + 1) : newzs
      .getMonth() + 1;
  var day = newzs.getDate() < 10 ? '0' + newzs.getDate() : newzs.getDate
      ();

  var TSdate = year + "/" + month + "/" + day;
  var TStime = newzs.toLocaleTimeString();
  newzs = TSdate + " " + TStime;
```

Listing 4.8: Konvertierung zum UTC-Zeitstempel-Format

Im nächsten Schritt wird der Zeitunterschied zwischen dem falschen und dem korrigierten

Zeitstempel kalkuliert und durch den Unterschied und dem Intervall wird der Messwert korrigiert (Vgl. Listing 4.9).

```
1    var secDiff = "";
2    var timeDiff = "";
3
4    secDiff = sec / 60;
5    timeDiff = minDiff + secDiff;
6    corMw = mw / timeDiff * 15;
```

Listing 4.9: Messwert Korrektur

Abschließend werden die korrigierten Daten in der Datenbank aktualisiert und alle Änderungen werden dokumentiert und in die Logging-Tabelle geschrieben (Vgl. Listing 4.10).

```
1    var pstmt2 = conn.prepareStatement("DELETE FROM \"CX4U\".\"MESSDATEN2\"
         WHERE \"ZAEHLERID\" = " + rs.getString(1) + " AND \"ZEITSTEMPEL\"
        = '" + rs.getString(3) + "'");
2    pstmt2.executeUpdate();
3
4    var pstmt3 = conn.prepareStatement("INSERT INTO \"CX4U\".\"MESSDATEN2\"
         VALUES (" + rs.getString(1) + ", " + corMw + ",'" + newzs + "','
        corrected Timestamp')");
5    pstmt3.executeUpdate();
6
7    var pstmt4 = conn.prepareStatement("INSERT INTO \"CX4U\".\"LOGGING\"
        VALUES (\"CX4U\".logseq.NEXTVAL, CURRENT_UTCTIMESTAMP, 'corrected
        Timestamp', " + rs.getString(1) + ", " + mw + ", '" +  rs.getString
        (3) + "')");
8    pstmt4.executeUpdate();
```

Listing 4.10: Änderungen Dokumentation

4.5.4.2 HTTP-Request von Netweaver an HANA

In diesem Kapitel wird beschrieben, welche Implementierungen vorgenommen werden müssen, um eine Kommunikation zwischen dem Netweaver und HANA aufzubauen. Dazu kommt keine Zusatzsoftware zum Einsatz, sondern lediglich Mechanismen, die bereits in HANA bzw. dem Netweaver integriert sind. Zunächst wird die Einstellung der nötigen Konfigurationsparameter für die HTTP-Verbindung beschrieben. Darauf erfolgt die Implementierung des HTTP-Requests in der Programmiersprache ABAP. Abschließend wird durch die Ausgabe der HTTP-Response-Daten der Beweis erbracht, dass eine Kommunikation aufgebaut werden kann.

4.5.4.2.1 Konfiguration der HTTP-Destination

Bevor der HTTP-Request implementiert werden kann, muss in der Transaktion „SM59“ eine sogenannte HTTP-Destination gepflegt werden. Dort werden alle Parameter eingetragen, die dazu benötigt werden, eine Verbindung zu der von HANA bereitgestellten OData-Schnittstelle aufzubauen. Das hat den Vorteil, dass zum Verbindungsaufbau die HTTP-Destination aufgerufen

werden kann und nicht immer wieder eine eigene „create“ programmiert werden muss. Wie in der folgenden Abbildung ersichtlich, muss der Verbindung ein Name gegeben werden (hier: PG_HANA). Als Verbindungstyp muss „G“ gewählt werden, da eine HTTP-Verbindung zu einem externen Server aufgebaut werden soll. Die Zielmaschine, Portnummer und der Pfadpräfix müssen entsprechend des jeweilig eingesetztem Zielsystems angegeben werden. Der Pfadpräfix ist die OData-Schnittstelle, die innerhalb HANA entwickelt wird (Vgl. SAPAG 2013a).

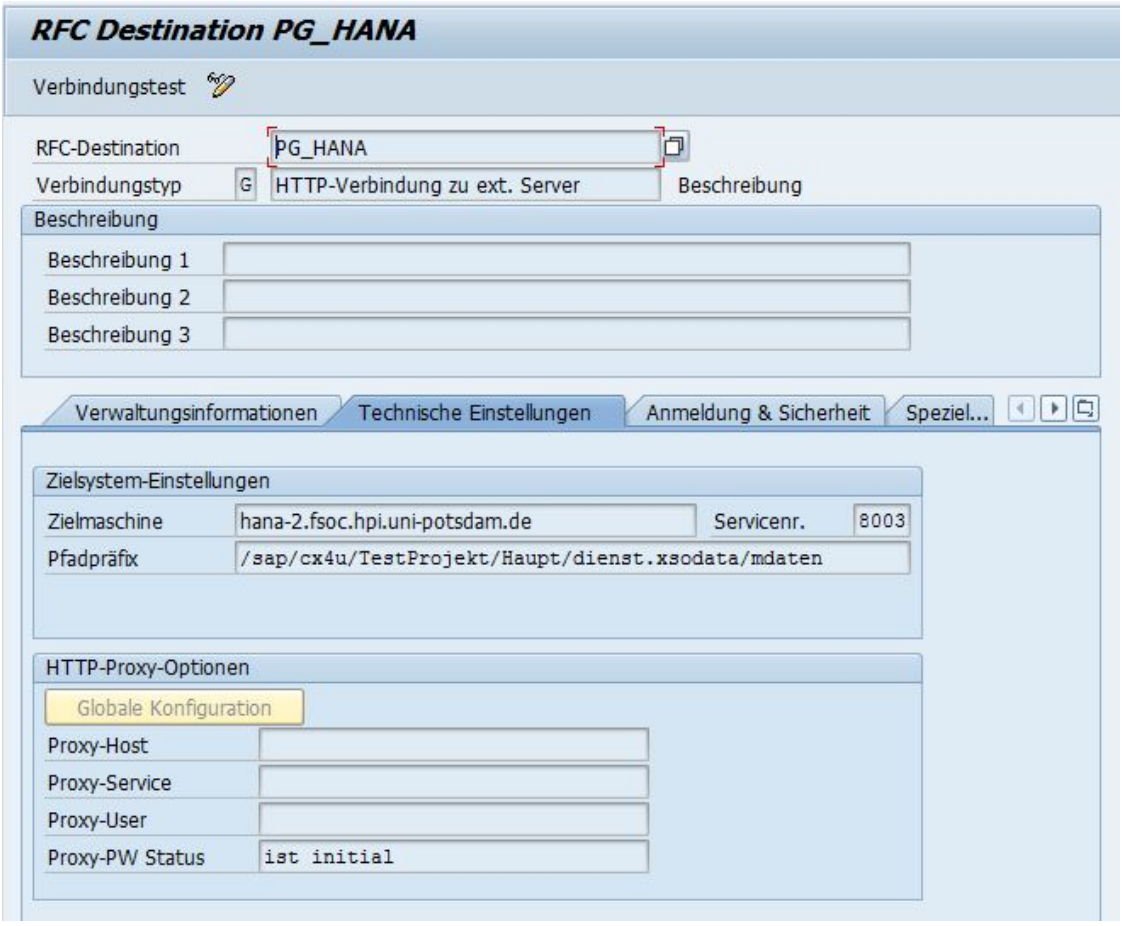


Abbildung 4.25: Konfiguration der HTTP-Destination

Anschließend muss im Reiter „Anmeldung & Sicherheit“ der Benutzername und das Passwort hinterlegt werden, mit dem die OData-Schnittstelle aufgerufen werden soll. Die Anmeldedaten werden dann automatisch in einen String umgewandelt, den HANA verarbeiten kann.
Abschließend kann die HTTP-Destination über den „Verbindungstest“ geprüft werden. Der Verbindungstest zeigt sowohl den Header als auch den Body an, welche die OData-Schnittstelle bereitstellt.

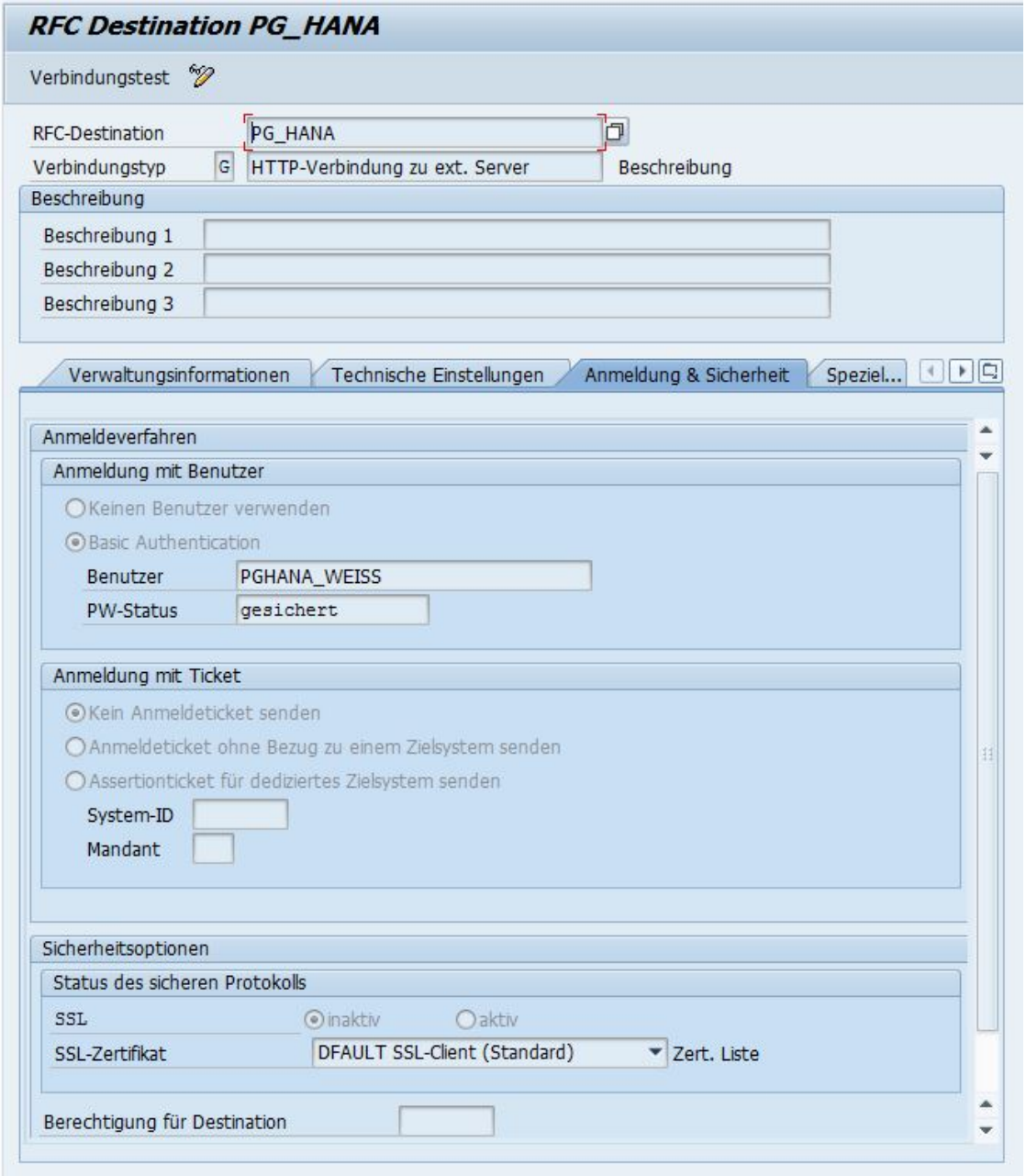


Abbildung 4.26: Anmeldedaten der HTTP-Destination

4.5.4.2.2 Implementierung des HTTP-Requests

Nachdem die Verbindungsdaten über die Transaktion SM59 gepflegt wurden, kann der HTTP-Request implementiert werden (Vgl. SAPAG 2013a). Die angelegte HTTP-Destination wird dazu genutzt um ein Objekt vom Typ „cl_http_client“ zu erzeugen[37] (Vgl. Listing 4.11).

```
1 cl_http_client⇒create_by_destination( EXPORTING
```

[37] Der gesamte Quellcode befindet sich im Anhang (Vgl. Anhang 6.1). In den folgenden Abschnitte werden die wichtigsten Codeabschnitte dargestellt, die für die Erbringung des Beweises nötig sind.

```
2 destination = http_dest IMPORTING CLIENT = CLIENT ).
```

Listing 4.11: HTTP-Client-Objekt

Neben der Destination und dem Client-Objekt, müssen ebenfalls der Request und die Variablen für die Fehlercodes deklariert werden (Vgl. Listing 4.12):

```
1 DATA: http_dest TYPE rfcdest VALUE 'PG_HANA'.
2 DATA: CLIENT TYPE REF TO if_http_client.
3 DATA: request TYPE REF TO if_http_request,
4       subrc          TYPE sysubrc,
5       errortext      TYPE string,
6       dummy          TYPE string,
7       timeout        TYPE I.
```

Listing 4.12: Deklaration der Variablen des HTTP-Requests

Das Objekt Client ermöglicht es, auf die Daten wie beispielsweise den Request, die Response oder die Header-Felder zu zugreifen. Für diesen Prototypen wird die HTTP-Request-Methode auf „GET“ gesetzt, da damit der Beweis erbracht wird, dass der Netweaver und HANA miteinander kommunizieren können. Außerdem wird die HTTP-Version auf „HTTP 1.1“ gesetzt, da diese ebenfalls von der OData-Schnittstelle verwendet wird (Vgl. Listing 4.13).

```
1 CALL METHOD CLIENT->request->set_method(
2 if_http_request=>co_request_method_get ).
3
4 CLIENT->request->set_version(
5 if_http_request=>co_protocol_version_1_1 ).
```

Listing 4.13: Header-Felder des HTTP-Requests setzen

Anschließend wird der Request versendet und die Responsdaten der Antwort in das Client-Objekt geladen. Mit der Methode „get_last_error“ wird sowohl beim „Send“ als auch beim „Receive“ der letzte Fehler abgefragt und gibt somit Informationen darüber, falls der HTTP-Request nicht ausgeführt werden kann (Vgl. Listing 4.14).

```
1 CALL METHOD CLIENT->send
2 EXPORTING  timeout = timeout
3 EXCEPTIONS http_communication_failure  = 1
4   http_invalid_state          = 2
5   http_processing_failed      = 3
6   OTHERS                      = 4.
7
8 IF sy-subrc <> 0.
9   CALL METHOD CLIENT->get_last_error
10  IMPORTING CODE    = subrc
11    MESSAGE = errortext.
12  WRITE: / 'communication_error( send )',
13  / 'code: ', subrc, 'message: ', dummy.
14  EXIT.
15 ENDIF.
```

```
16
17 *receive
18 CALL METHOD CLIENT→receive
19 EXCEPTIONS http_communication_failure  = 1
20   http_invalid_state          = 2
21   http_processing_failed      = 3
22   OTHERS                      = 4.
23
24 IF sy-subrc <> 0.
25   CALL METHOD CLIENT→get_last_error
26   IMPORTING CODE    = subrc
27     MESSAGE = errortext.
28   WRITE: / 'communication_error( receive )',
29   / 'code: ', subrc, 'message: ', dummy.
30   WRITE: errortext.
31
32   EXIT.
33 ELSE.
34 ENDIF.
```

Listing 4.14: Versenden und Empfang des HTTP-Requests

Abschließend werden die Header-Felder der Response-Daten des HTTP-Requests abgefragt, ein Unterprogramm zur Ausgabe der Daten aufgerufen und die Verbindung geschlossen (Vgl. Listing 4.15). Insofern beim Ausführen des Programms keine Fehler angezeigt werden, sollte der HTTP-Request erfolgreich ausgeführt werden können. Um jedoch eindeutig zu testen, ob die von der OData-Schnittstelle bereitgestellten Daten korrekt übermittelt werden, wurde darüber hinaus das eben erwähnte Unterprogramm zur Ausgabe der Daten implementiert.

```
1  DATA: FIELDS TYPE tihttpnvp,
2        fields_wa TYPE ihttpnvp.
3
4  CALL METHOD CLIENT→response→get_header_fields
5  CHANGING FIELDS = FIELDS.
6
7  PERFORM write_document.
8
9  CALL METHOD CLIENT→close
10 EXCEPTIONS
11   http_invalid_state = 1
12   OTHERS             = 2.
```

Listing 4.15: Laden der Header-Felder und schließen der Verbindung

4.5.4.2.3 Ausgabe der HTTP-Response Daten

Werden die Daten über den „write“-Befehl ausgegeben, wird eine Zeichenkette angezeigt, die nicht die von der OData-Schnittstelle bereitgestellten Daten darstellt.

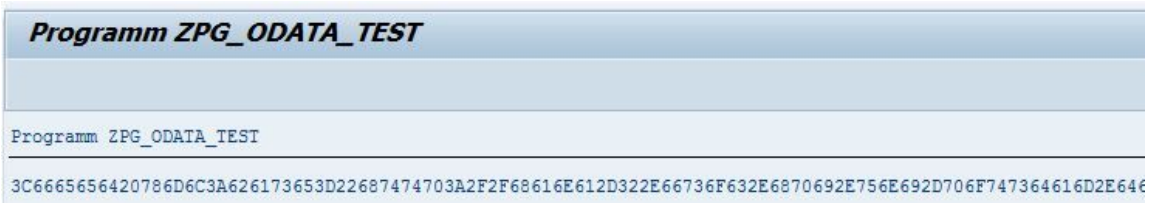


Abbildung 4.27: Ausgabe der Daten der HTTP-Response mit write

Daher muss ein Unterprogramm geschrieben werden, um zu validieren, dass es sich um die korrekten Daten handelt. Die Daten wurden dazu in einem Body-Feld ausgegeben[38]. Für die Machbarkeitsstudie ist es wichtig, wie die Daten aus der Response abgegriffen werden können. Dazu genügt die Methode “get_cdata( )“. In dem folgenden Listing 4.16 ist dargestellt, wie die Daten abgerufen werden. Dazu wird die Länge der Zeichenkette gemessen und mithilfe einer While-Schleife an die Breite des Bodys angepasst.

```
1   DATA = CLIENT→response→get_cdata(  ).
2   len = STRLEN( DATA ).
3
4   WRITE: / sy-ULINE(l_width) NO-gap.
5   FORMAT COLOR col_heading INTENSIFIED off.
6   WRITE: / sy-vline, AT (l_width_b) 'body' .
7   WRITE AT l_width sy-vline.
8   FORMAT COLOR COL_NORMAL INTENSIFIED off.
9   WHILE len > 0.
10    LINE  = DATA+llen(*).
11    WRITE: / sy-vline NO-gap.
12    WRITE:  LINE.
13    WRITE AT l_width sy-vline.
14    len = len - l_width_b.
15    llen = llen + l_width_b.
16  ENDWHILE.
```

Listing 4.16: Aufbereitung der HTTP-Respone Daten

Das Ergebnis ist die korrekte Ausgabe innerhalb eines Bodys der von der OData-Schnittstelle bereitgestellten Daten. Die Implementierungen des Headers sind GUI-Anpassungen, in denen, ebenfalls mit einer While-Schleife, die Daten aus der Methode „get_header_fields“ eingetragen werden.

[38] Nach umfangreicher Recherche konnte kein Parser gefunden werden, der die Daten für das SAP System formatiert. Daher wurde in Anlehnung an folgenden Foreneintrag eine Lösung entwickelt, um den Beweis der Kommunikation zu erbringen: `http://www.abapforum.com/forum/viewtopic.php?f=1&t=16875`

Programm ZPG_ODATA_TEST

Programm ZPG_ODATA_TEST

header fields

field name	field value
~response_line	HTTP/1.1 200 OK
~server_protocol	HTTP/1.1
~status_code	200
~status_reason	OK
set-cookie	xsSessionId=77E139C44184C0418AA903D10E85C18F; path=/; HttpOnly
set-cookie	sapxslb=B76EF569861E1841BEAE45BD085E7A6E; path=/; HttpOnly
content-type	application/atom+xml;charset=utf-8
content-length	6309
cache-control	no-cache
expires	Thu, 01 Jan 1970 00:00:00 GMT
~remote_addr	
~uri_scheme_expanded	NONE

body


```
<feed xml:base="http://hana-2.fsoc.hpi.uni-potsdam.de:8003/sap/cx4u/TestProjekt/Haupt/dienst.xsodata/" xmlns:d="http://schemas.microsoft.com/a
do/2007/08/dataservices" xmlns:m="http://schemas.microsoft.com/ado/2007/08/dataservices/metadata" xmlns="http://www.w3.org/2005/Atom"><title t
ype="text">mdaten</title><id>http://hana-2.fsoc.hpi.uni-potsdam.de:8003/sap/cx4u/TestProjekt/Haupt/dienst.xsodata/mdaten</id><author><name />
/author><link rel="self" title="mdaten" href="mdaten" /><entry><id>http://hana-2.fsoc.hpi.uni-potsdam.de:8003/sap/cx4u/TestProjekt/Haupt/diens
t.xsodata/mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624183000L)</id><title type="text"></title><author><name /></author><link rel="edit" title=
"mdaten" href="mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624183000L)"/><category term="sap.cx4u.TestProjekt.Haupt.mdatenType" scheme="http://sc
hemas.microsoft.com/ado/2007/08/dataservices/scheme" /><content type="application/xml"><m:properties><d:ZAEHLERID m:type="Edm.Int32">170589</d
:ZAEHLERID><d:MESSWERT m:type="Edm.Double">0.0085600000000000005</d:MESSWERT><d:ZEITSTEMPEL m:type="Edm.Int64">20130624183000</d:ZEITSTEMPEL><
d:STATUS m:type="Edm.String">raw</d:STATUS></m:properties></content></entry><entry><id>http://hana-2.fsoc.hpi.uni-potsdam.de:8003/sap/cx4u/Tes
tProjekt/Haupt/dienst.xsodata/mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624184500L)</id><title type="text">/></title><author><name />/></author><li
nk rel="edit" title="mdaten" href="mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624184500L)"/><category term="sap.cx4u.TestProjekt.Haupt.mdatenTyp
e" scheme="http://schemas.microsoft.com/ado/2007/08/dataservices/scheme" /><content type="application/xml"><m:properties><d:ZAEHLERID m:type="
Edm.Int32">170589</d:ZAEHLERID><d:MESSWERT m:type="Edm.Double">0.0075599999999999996</d:MESSWERT><d:ZEITSTEMPEL m:type="Edm.Int64">20130624184
500</d:ZEITSTEMPEL><d:STATUS m:type="Edm.String">raw</d:STATUS></m:properties></content></entry><entry><id>http://hana-2.fsoc.hpi.uni-potsdam.
de:8003/sap/cx4u/TestProjekt/Haupt/dienst.xsodata/mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624190000L)</id><title type="text">/></title><author>
<name />/></author><link rel="edit" title="mdaten" href="mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624190000L)"/><category term="sap.cx4u.TestPro
jekt.Haupt.mdatenType" scheme="http://schemas.microsoft.com/ado/2007/08/dataservices/scheme" /><content type="application/xml"><m:properties><
d:ZAEHLERID m:type="Edm.Int32">170589</d:ZAEHLERID><d:MESSWERT m:type="Edm.Double">0.00554</d:MESSWERT><d:ZEITSTEMPEL m:type="Edm.Int64">20130
624190000</d:ZEITSTEMPEL><d:STATUS m:type="Edm.String">raw</d:STATUS></m:properties></content></entry><entry><id>http://hana-2.fsoc.hpi.uni-po
tsdam.de:8003/sap/cx4u/TestProjekt/Haupt/dienst.xsodata/mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624191500L)</id><title type="text">/></title><a
uthor><name />/></author><link rel="edit" title="mdaten" href="mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624191500L)"/><category term="sap.cx4u.T
estProjekt.Haupt.mdatenType" scheme="http://schemas.microsoft.com/ado/2007/08/dataservices/scheme" /><content type="application/xml"><m:proper
ties><d:ZAEHLERID m:type="Edm.Int32">170589</d:ZAEHLERID><d:MESSWERT m:type="Edm.Double">0.0058399999999999999</d:MESSWERT><d:ZEITSTEMPEL m:ty
pe="Edm.Int64">20130624191500</d:ZEITSTEMPEL><d:STATUS m:type="Edm.String">raw</d:STATUS></m:properties></content></entry><entry><id>http://ha
na-2.fsoc.hpi.uni-potsdam.de:8003/sap/cx4u/TestProjekt/Haupt/dienst.xsodata/mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624193000L)</id><title ty
pe="text">/></title><author><name />/></author><link rel="edit" title="mdaten" href="mdaten(ZAEHLERID=170589,ZEITSTEMPEL=20130624193000L)"/><categ
ory term="sap.cx4u.TestProjekt.Haupt.mdatenType" scheme="http://schemas.microsoft.com/ado/2007/08/dataservices/scheme" /><content type="applic
ation/xml"><m:properties><d:ZAEHLERID m:type="Edm.Int32">170589</d:ZAEHLERID><d:MESSWERT m:type="Edm.Double">2.0058400000000001</d:MESSWERT><d
:ZEITSTEMPEL m:type="Edm.Int64">20130624193000</d:ZEITSTEMPEL><d:STATUS m:type="Edm.String">raw</d:STATUS></m:properties></content></entry><en
try><id>http://hana-2.fsoc.hpi.uni-potsdam.de:8003/sap/cx4u/TestProjekt/Haupt/dienst.xsodata/mdaten(ZAEHLERID=170589,ZEITSTEMPEL=2013062419450
```


Abbildung 4.28: Ausgabe der Daten der HTTP-Response in einer GUI

4.5.5 Schlussbetrachtung

Im Rahmen der Machbarkeitsstudie wurden einige denkbare Möglichkeiten betrachtet, die zusätzliche Produkte benötigen oder Kosten verursacht hätten, um eine Kommunikation zwischen SAP NetWeaver und SAP HANA zu ermöglichen. Aufgabe war es jedoch, sich hierbei auf die vorhandenen Systeme und kostenfreie Erweiterungen zu beschränken. Durch diese Vorgabe der Kostenneutralität konnten einige Alternativen schnell ausgeschlossen werden. Da eine Integration der zukünftigen Lösung in die bestehende Systemlandschaft zwingend notwendig war, wurde dementsprechend viel Zeit in das Erstellen einer möglichst umfassenden Machbarkeitsstudie gelegt. Mit den SAP HANA Extended Application Services konnte eine Lösung gefunden werden, die sich für die Umsetzung des Projektes eignet. Allerdings hat es sich als schwierig herausgestellt, diese Lösung zu finden. Einschlägige Literatur, Beschreibungen in den Foren der SAP AG und auch konkrete Nachfragen beim HPI verwiesen auf die kommerziellen Produkte. Mit der erarbeiteten Lösung konnte jedoch gezeigt werden, dass auch mit verfügbaren,

bzw. freien Technologien ein Erfolg erzielt werden kann - also, dass der Netweaver einen http-Request auf die OData-Schnittstelle von HANA ausführen kann, ohne dass zusätzliche Software oder Bibliotheken benötigt werden.

4.6 Qualitätsmanagement

Dieses Kapitel befasst sich mit der Qualitätssicherung im Projektmanagement, daher auch als Qualitätsmanagement bezeichnet. Es werden im ersten Abschnitt zunächst einige grundlegende Maßnahmen des Qualitätsmanagements vorgestellt. Abschließend werden ausgewählte und festgelegte Testfälle vorgestellt.

4.6.1 QM-Maßnahmen

Die folgenden Abschnitte geben einen kurzen Überblick über Maßnahmen zum Qualitätsmanagement und wie bei der Teilgruppe CX4U versucht wurde, ein solches Management abzubilden. Ein Qualitätsmanagement (QM) umfasst alle Maßnahmen und Tätigkeiten, die der Verbesserung von Produkten, Prozessen oder (Dienst-) Leistungen dienen. Es wird bei der Planung, Steuerung und Prüfung der Qualität von Software sowohl bei fertigen Produkten als auch bei Entwicklungsprozessen eingesetzt. Bei der Erstellung von Software gilt es, bestimmte Standards einzuhalten. Die Qualität von Software bezeichnet alle Merkmale, die ein Softwareprodukt aufweist (Vgl. Gabler Wirtschaftslexikon 2013b). Fehlende Qualität in Softwaresystemen macht sich in den meisten Fällen erst bemerkbar, wenn es zu spät ist und bestimmte Bereiche oder ganze Systeme der Software nicht ordnungsgemäß laufen. Fehler verursachen Kosten und müssen daher auf ein Minimum beschränkt werden. Dies kann durch Qualitätssicherung beim Prozess der Softwareerstellung (Prozessqualität) und durch Prüfung der Software (Produktqualität) erreicht werden. Mit fortschreitender Entwicklung steigen die Kosten für die Behebung von Fehlern. Man spricht hierbei auch von der sogenannten 10er-Regel (Vgl. Abb. 4.29).

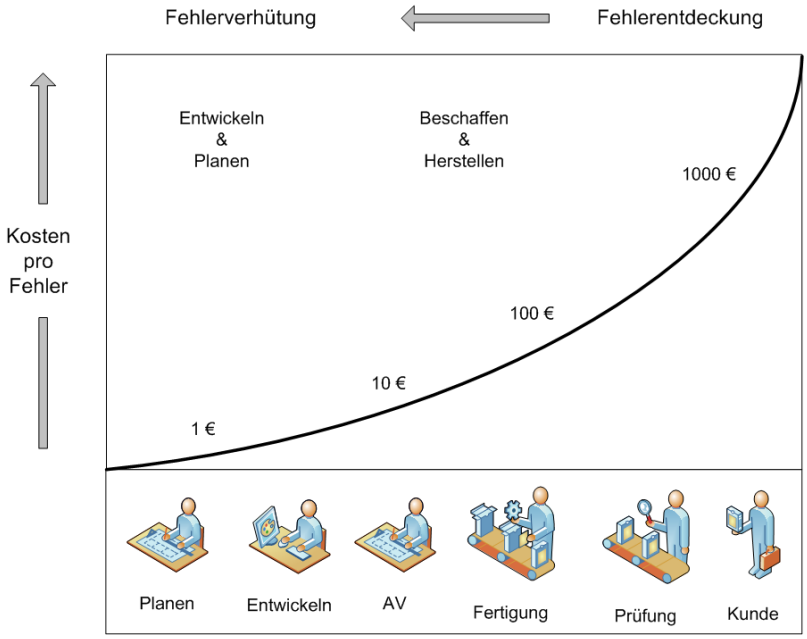


Abbildung 4.29: Fehlerkosten 10er-Regel
Quelle: Schnurr (2014)

Diese besagt, dass je weiter ein Fehler sich unentdeckt in die späten Phasen des Werdeganges eines Produktes oder Prozesses bewegt – oder gar bis zum Kunden –, umso höher werden die Kosten zur Behebung dieses Fehlers (Vgl. Mellis 2008, S. 198). Aus diesem Grund ist die Gewährleistung von Qualität bei der Softwareentwicklung unabdingbar. Das Qualitätsmanagement sollte dabei eine phasenübergreifende Tätigkeit sein und in jeder der Projektphasen angewendet werden. Diese Qualitätsmerkmale von Softwaresystemen sind in der ISO 25010 (vorher ISO 9126) festgehalten und definieren bei der Entwicklung bestimmbare Eigenschaften der Software. Sie können dabei ihrer Wichtigkeit entsprechend hierarchisch geordnet werden. Es ist möglich, die Merkmale weiter zu unterteilen. Dabei entstehen Qualitätsindikatoren. Qualitätsmerkmale/ -kriterien von Softwaresystemen (ISO 25010):

- Funktionalität (Angemessenheit, Richtigkeit, Interoperabilität, Ordnungsgemäßheit, Sicherheit)
- Zuverlässigkeit (Reife, Fehlertoleranz, Wiederherstellbarkeit)
- Benutzbarkeit (Verständlichkeit, Erlernbarkeit, Bedienung)
- Effizienz (Verbrauchsverhalten, Zeitverhalten)
- Änderbarkeit (Analysierbarkeit, Modifizierbarkeit, Stabilität, Prüfbarkeit)
- Übertragbarkeit (Anpassbarkeit, Installierbarkeit, Austauschbarkeit)

Es wird zwischen konstruktiven und analytischen Qualitätsmaßnahmen unterschieden, die in den folgenden Abschnitten kurz vorgestellt werden (Vgl. Balzert 1998, S.278 ff.).

Konstruktive Qualitätsmaßnahmen

Diese Art der Maßnahmen sind die Grundlage für alle weiteren Maßnahmen. Sie sorgen u.a. dafür, dass beim Entwicklungsprozess von Beginn an bestimmte Eigenschaften vorhanden sind. Dazu zählen u.a. Methoden, Sprachen, Werkzeuge, Richtlinien, Checklisten, Standards etc.

Jede Aktivität, die sich positiv auf die Qualität des Produktes oder Prozesses auswirkt, kann als eine konstruktive Qualitätsmaßnahme angesehen werden. Maßnahmen werden unterschieden zwischen ihrer Auswirkung, auf das Produkt (vorgegebene Gliederung für Last- und Pflichtenheft, Entwurfsmuster für die Softwarearchitektur) oder den Entwicklungsprozess (Verwendung eines Vorgehensmodells, Projektmanagement, Definition von Verantwortlichkeiten). Weiterhin kann man zwischen technischen und organisatorischen Maßnahmen unterscheiden. Weitere Aufteilungen dieser beiden Maßnahmen beschreibt Abb. 4.30.

Analytische Qualitätsmaßnahmen

Analytische Maßnahmen bauen auf den konstruktiven Maßnahmen auf und werden verwendet, um die Qualität bzw. das Niveau dieser zu messen und zu bewerten. Es erfolgt eine Unterscheidung in analysierende (sammeln Informationen über das zu prüfende Objekt) und testende

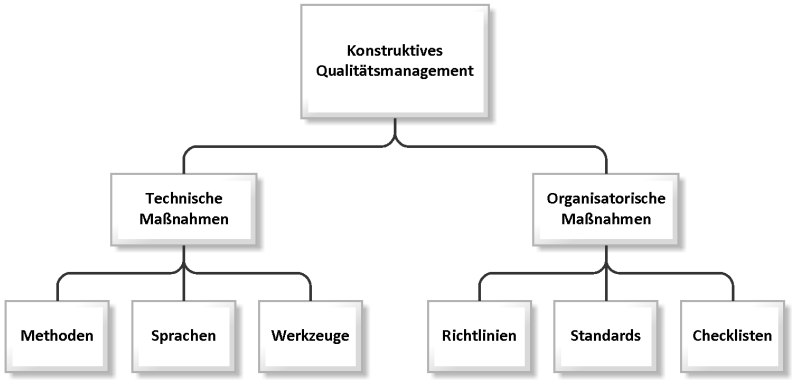


Abbildung 4.30: Konstruktive Qualitätsmanagementmaßnahmen
Quelle: Eigene Darstellung in Anlehnung an (Balzert 1998, S.270)

(testen das Objekt) Verfahren. Weiterhin wird zwischen statischen (nur Prüfung des Quellcodes durch Inaugenscheinnahme) und dynamischen (Ausführung des Quellcodes mit Werten und Parametern und Überprüfung des Ergebnisses) Verfahren unterschieden. Gängige statische Verfahren sind beispielsweise (Vgl. Balzert 1998, S.281):

- Code-Inspektionen: Stark formalisiertes Verfahren bei dem Dokument nach genau festgelegter Vorgehensweise durch Prüfer untersucht und protokolliert wird. Dient lediglich der Fehlerentdeckung, nicht der Beseitigung.
- Review: Manuelle Prüfmethode, die der Inspektion von Ergebnissen dient. Kann als Schreibtischtest erfolgen. Es Erfolgt keine sofortige Korrektur der Fehler.
- Walkthroughs: Dient ausschließlich dem Auffinden von Fehlern, Problemen und Unklarheiten, nicht jedoch der Korrektur. Durchspielen von Szenarien und Testfällen durch Entwickler und Prüfer.

Bei den dynamischen Verfahren lassen sich folgende Alternativen unterscheiden (Vgl. Balzert 1998, S.282):

- Strukturtest (White-Box-Test): Die interne Struktur der Komponente (= Operation, Klasse, Paket oder Gesamtsystem) wird zur Testplanung und -Überwachung herangezogen. Quellcode und Implementierung steht zur Verfügung.
- Funktionstest (Black-Box-Test): Das Ein- und Ausgabeverhalten wird mit der Verwendung von unterschiedlichen Werten getestet. Quellcode ist nicht bekannt.
- Diversifikationstest: Verhalten einer Komponentenversion wird mit dem Verhalten anderer Komponentenversionen verglichen.

Beide Maßnahmen können in allen Phasen der Entwicklung des Softwareprodukts zum Einsatz kommen. Ergänzend dazu gibt es produktorientiertes und prozessorientiertes Qualitätsmanagement.

Bei der produktorientierten Betrachtungsweise setzen die Maßnahmen setzen direkt an der zu erstellenden Software (inkl. Zwischenprodukte) an. Es wird überprüft, ob vorher festgelegte Qualitätsmerkmale eingehalten wurden. Das prozessorientierte Qualitätsmanagement bezieht sich hingegen auf den Entwicklungsprozess und beinhaltet ebenso die Überprüfung auf Einhaltung von vorher festgelegten Qualitätsmerkmalen – hierbei wird jedoch (nur) der (Entwicklungs-) Prozess betrachtet und überprüft.

4.6.1.1 Vorgehensmodelle

Grundsätzlich sollte bei der Entwicklung von Software ein Vorgehensmodell zugrunde liegen. Dort werden alle benötigten Aufgabenbereiche wie beispielsweise das Projektmanagement oder die Qualitätssicherung beschrieben. Es legt weiterhin fest, welche Aktivitäten, in welcher Reihenfolge, von welchen Rollen erledigt werden, welche Ergebnisse dabei entstehen und wie die Qualitätssicherung erfolgt (Vgl. Balzert 1998, S. 98 ff.). Es dient dabei als Muster für die Projektplanung und –steuerung.

4.6.1.1.1 V-Modell XT

Dieses Modell deckt hauptsächlich die Bereiche Projektmanagement, Qualitätssicherung, Ausschreibung und Vergabe sowie die Systementwicklung ab. Entwickelt unter Federführung des Innenministeriums der Bundesrepublik Deutschland und der für die IT zuständigen Stelle der Bundeswehr, ist es bei der Vergabe von Aufträgen durch öffentliche Stellen (Behörden etc.) verpflichtend im Rahmen der notwendigen Ausschreibung einzusetzen und somit auch für Unternehmen aus der Wirtschaft interessant, insbesondere wenn man eine solche Ausschreibung gewinnen möchte. Es dient als Schnittstelle zwischen Auftraggeber und Auftragnehmer.

4.6.1.1.2 Wasserfallmodell

Das Modell hat seinen Namen durch das Prinzip eines Wasserfalls erhalten, da Ergebnisse einer Phase immer in die folgende mit einfließen. Zwischen den jeweils aneinander liegenden Phasen sind Rückkoppelungen möglich, was Abbildung 4.31 verdeutlicht. Grundlegende Idee

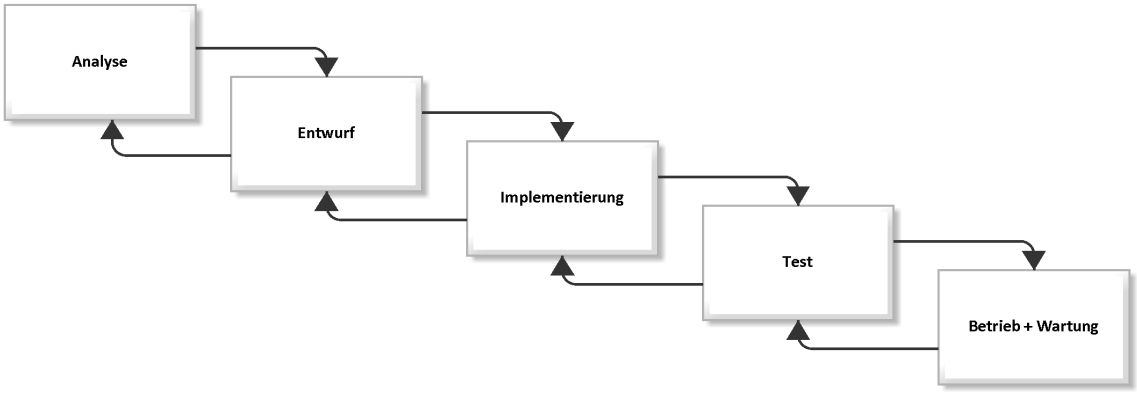


Abbildung 4.31: Wasserfallmodell

dieses Modells ist die Aufteilung des Projekts in Phasen anhand der Entwicklungstätigkeiten in den einzelnen Phasen Analyse, Entwurf, Implementierung, Test sowie Betrieb und Wartung. Jede Phase schließt mit einem zu erreichenden Meilenstein ab. Bevor mit einer nächsten Phase begonnen werden kann, müssen die Aktivitäten der vorangehenden Phase vollständig abgeschlossen sein.

4.6.1.1.3 Spiralmodell

Das Spiralmodell ist ein iteratives Vorgehensmodell für die (Software-) Entwicklung (Vgl. Kurbel u. a. 2014). Primäre Ziele sind die Kontrolle und Minimierung von Projektrisiken. Grundprinzip ist das wiederholte Durchlaufen von einzelnen Phasen wie beispielsweise Analyse, Evaluierung und Review. In Abbildung 4.32 ist das Funktionsprinzip dieses Modells dargestellt.

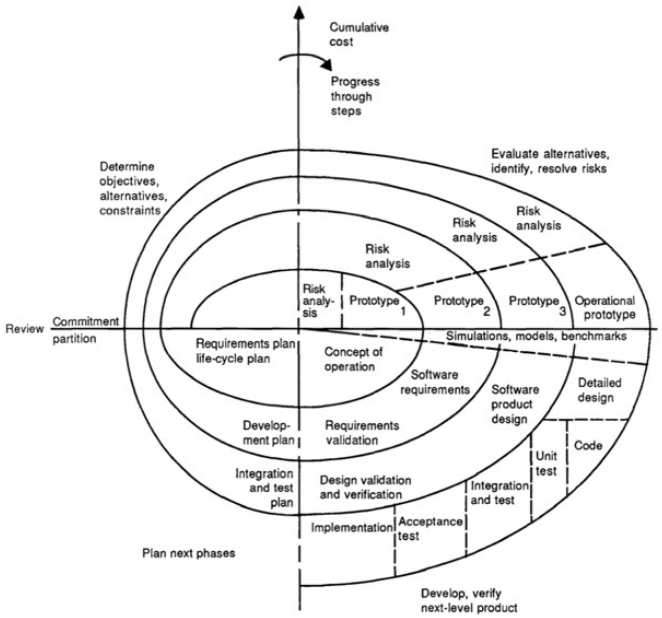


Abbildung 4.32: Spiralmodell
Quelle: Kurbel u. a. (2014)

4.6.1.1.4 Prototyping

Prototypen kommen immer dann zum Einsatz, wenn zu Beginn eines Projekts noch nicht genau definiert ist, was genau entwickelt werden soll bzw. es keine klaren Anforderungen gibt. Mithilfe von kontinuierlich verbesserten Prototypen wird versucht, sich dem Ziel (-system) schrittweise anzunähern. Dieses Verfahren kann u.a. auch eingesetzt werden, um bestimmte Aspekte (z. B. Anforderungen, Vorstellungen der Anwender etc.) schon im Vorfeld besser beleuchten, untersuchen und bewerten zu können. Folgende Arten von Prototypen kommen am häufigsten zum Einsatz:

- Demonstratoren werden in den frühen Phasen eines Projekts eingesetzt und skizzieren grob verschiedene Richtungen sich das (Software-) Projekt entwickeln kann.

- Labormuster werden verwendet, um verschiedene grundlegende Fragestellungen zu untersuchen, wie beispielsweise die Tragfähigkeit einer Softwarearchitektur.
- Pilotsysteme kommen zumeist in späteren Phasen des Projekts zum Einsatz und beinhalten bereits einen Großteil der finalen Implementierung. Ziel ist das Testen des Systems durch den künftigen Anwender, um noch Verbesserungsvorschläge mit einfließen lassen zu können.

Prototypen werden in der Regel nach zwei verschiedenen Mustern konstruiert. Horizontale Prototypen bilden nur einen ausgewählten Bereich einer Architektur ab. Vertikale Prototypen werden hingegen verwendet, um einen gesamten Anwendungsbereich über alle Schichten des Systems abzubilden (Vgl. Abb. 4.33).

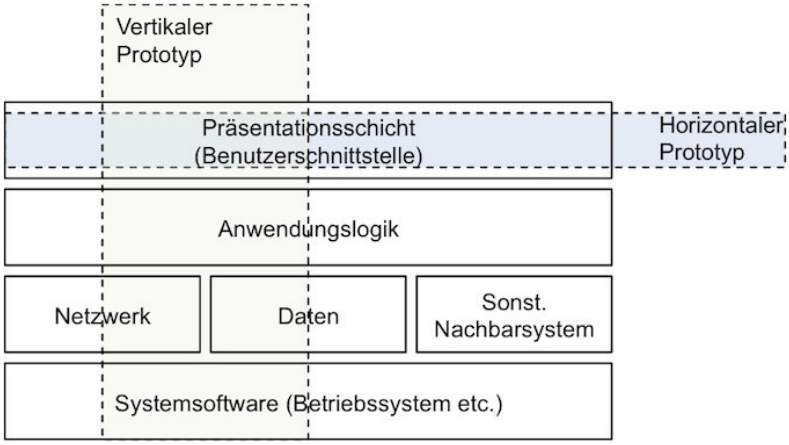


Abbildung 4.33: Prototypen
Quelle: Kurbel u. a. (2014)

4.6.1.1.5 Unified Process

Dieses Modell wird seit 1999 von IBM entwickelt und kommerziell vertrieben. Wesentliches Merkmal ist das Vorantreiben durch Aktivitäten. Es wird insbesondere bei der Softwareentwicklung mit UML eingesetzt. Das Projekt wird in einzelne Phasen (z. B. Teile des Lebenszyklus) unterteilt. Unified Process ist ein sogenannter aktivitätsgetriebener Entwicklungsprozess, d.h. Rollen sind für die Durchführung bestimmter Aktivitäten zuständig. Aktivitäten werden in einer vorher festgelegten Reihenfolge abgearbeitet und sind bestimmten Phasen im Prozess zugeordnet. Das Erreichen bestimmter Phasen und Abschnitte kann als Meilenstein definiert werden. Abschnitte im Lebenszyklus eines Projekts werden im Unified Process in folgende vier Phasen unterteilt:

- Konzeptionsphase
- Entwurfsphase
- Konstruktionsphase

- Übergabephase

In jeder der vier Phasen werden Aktivitäten einzelner Disziplinen gebündelt. Weiterhin sind die einzelnen Phasen in Iterationen unterteilt. Jede Iteration schließt mit einem Meilenstein ab:

- Lifecycle Objectives
- Lifecycle Architecture
- Initial Operational Capability (Entwurfsmodelle und Beta-Release der Software)
- Product Release

Ziel des ersten Meilensteins ist die grobe Definition eines Ziel und eines ersten Anwendungsfallmodells, das die wesentliche Funktionalität des Zielsystems beschreibt. Am Ende der zweiten Phase soll ein Architekturprototyp entstanden sein. Darüber hinaus soll das Anwendungsfallmodell aus der ersten Phase weiter spezifiziert werden. Der dritte Meilenstein hat detaillierte Entwurfsmodelle und eine Beta-Version der Software zum Ergebnis, während im darauffolgenden vierten Meilenstein bereits die fertig entwickelte Software vorliegen muss.

4.6.1.2 Testen von Software

Einen besonders wichtigen Aspekt im Rahmen eines Qualitätsmanagements bei der Softwareentwicklung stellen Tests dar. Sie dienen der Erkennung von nicht geplantem und nicht gewünschtem Verhalten der Software. Es handelt sich dabei oftmals um Stichproben, die jedoch nicht für einen vollständigen Ausschluss von Fehlern garantieren können. Software lässt sich auf mehrere unterschiedliche Arten hin testen. Eine kurze Auswahl nach Balzert (Vgl. Balzert 1998, S. 391 ff.) davon wird an dieser Stelle kurz vorgestellt:

- Tests während der Implementierung dienen der Entdeckung von Syntaxfehler und logischen Fehlern (z.B. Endlosschleifen).
- Komponententests prüfen unter Verwendung von Testdaten und Testprogrammen, ob eine Komponente den Anforderungen genügt.
- Integrationstests überprüfen das Zusammenspiel der (voneinander abhängigen) Komponenten. Der Schwerpunkt liegt auf der Überprüfung von Schnittstellen..
- Systemtests prüfen das Gesamtsystem auf Basis realer Daten.
- Belastungs- und Stresstests dienen der Überprüfung des Verhaltens eines Systems bei (voller) Auslastung.

Die Teilgruppe Smart Metering hat während ihrer Implementierung die ersten drei der eben vorgestellten Testverfahren verwendet. Darüber hinaus wurden funktionale Testfälle entworfen. Weiter Ausführungen dazu sind in Abschnitt 4.6.2 zu finden.

4.6.1.3 Anwendung

Im vorliegenden Projekt der Teilgruppe CX4U wurde durch den Praxispartner zwar kein explizites Modell bzw. eine Vorgehensweise vorgegeben, dennoch erschien es für die Mitglieder der Teilgruppe sinnvoll, dem Projekt ein Vorgehensmodell zu Grunde zu legen. Die Auswahl eines solchen gestaltete sich zunächst schwierig, da zu Beginn des Projekts noch nicht genau fest stand, was und wie es entwickelt werden sollte. Zudem wurden die Spezifikationen sowie der Funktionsumfang während der Einführungsphase noch verändert bzw. im weiteren Verlauf des Projekts noch um weitere Funktionen erweitert. Dies führte dazu, dass schlussendlich das Prototyping zum Einsatz kam. Es wurde nach den Anforderungen entwickelt, anschließend kam ein Feedback vom Auftraggeber mit Verbesserungs- und Änderungswünschen. Diese wurden umgesetzt und erneut zur Prüfung vorgelegt, um dann wieder ein Feedback zu bekommen. So konnte das Programm sukzessive an die Anforderungen und Wünsche des Auftraggebers angepasst und verbessert werden.

4.6.2 Funktionale Testfälle

Um überprüfen zu können, ob die Software die gewünschte Funktionalität bietet, bzw. auf Eingaben wie erwartet reagiert, können funktionale Testfälle genutzt werden. Diese Anwendertests bieten eine strukturierte Möglichkeit, sich dem System zu nähern. Für das NetWeaver System der CX4U AG sind folgende Testfälle (Abb. 4.34) exemplarisch:

ID	Name	Vorbedingung	Testschritte	Erwartetes Ergebnis
Die ID muss den Testfall eindeutig identifizieren	Der Name sollte kurz beschreiben, worum es in dem Testfall geht	Welche Bedingungen müssen erfüllt sein, damit dieser Testfall ausgeführt werden kann? Bsp. Welche Systeme müssen verfügbar sein?	1. Wie muss der Benutzer vorgehen, damit er zu dem erwarteten Ergebnis kommen kann? Hier können Kenntnisse über die Systeme vorausgesetzt werden – es ist also keine Klickanleitung nötig	1. Hier wird beschrieben, wie sich das System verhält, wenn der Benutzer wie beschrieben vorgegangen ist.
NW-F1	NetWeaver Abruf von Daten in HANA	HANA des HPI erreichbar NetWeaver CX4U erreichbar	1. Aufbauen der VPN Verbindung zur CX4U 2. Anmelden am SAP-System 3. Aufrufen der Transaktion SE80 4. Öffnen des Programms ZPG_ODATA_TEST 5. Prüfen des Programms 6. Aktivieren des Programms 7. Ausführen des Programms 8. Prüfen der Ausgabe	1. VPN Verbindung hergestellt 2. Anmeldung erfolgreich 3. Transaktion wird aufgerufen 4. Programm wird geöffnet 5. Es werden keine Syntaxfehler gefunden 6. Aktives Objekt wird generiert 7. Programm wird ohne Fehlermeldung ausgeführt 8. Die Ausgabe entspricht der Erwartung
NW-F2	Fehlerfall: Keine Verbindung zum HANA	NetWeaver CX4U erreichbar	1. Aufbauen der VPN Verbindung zur CX4U 2. Anmelden am SAP-System 3. Aufrufen der Transaktion SE80 4. Öffnen des Programms ZPG_ODATA_TEST 5. Prüfen des Programms 6. Aktivieren des Programms 7. Ausführen des Programms	1. VPN Verbindung hergestellt 2. Anmeldung erfolgreich 3. Transaktion wird aufgerufen 4. Programm wird geöffnet 5. Es werden keine Syntaxfehler gefunden 6. Aktives Objekt wird generiert 7. Es wird eine Fehlermeldung angezeigt
HANA-F1	Test der ODATA Schnittstelle	HANA des HPI erreichbar Google Chrome installiert Erweiterung Advanced Rest Client installiert	1. Aufbauen der VPN Verbindung zum HPI 2. Öffnen von Google Chrome 3. Öffnen des Advanced Rest Clients 4. Eingeben der URL http://hana-2.fsoc.hpi.uni-potsdam.de:8003/sap/cx4u/TestProjekt/Haupt/dienst.xsodata/mdaten?$format=json 5. Auswählen der Methode GET und Klick auf Send 6. Anmelden am System 7. Prüfen der Ausgabe	1. VPN Verbindung hergestellt 2. Chrome wird geöffnet 3. Der Advanced Rest Service wird geöffnet 4. Die URL kann eingegeben werden 5. GET kann ausgewählt und Send angeklickt werden. Das System fragt nach Benutzername und Passwort 6. Anmeldung erfolgreich 7. Die Ausgabe entspricht der Erwartung

Abbildung 4.34: Funktionale Testfälle

4.7 Öffentlichkeitsarbeit

Die Teilprojektgruppe Prozessorientierte Business Intelligence mit SAP HANA stellte ihre Projektarbeit im Rahmen mehrerer Veranstaltungen öffentlichkeitswirksam vor. Hierbei handelt es sich um den Future SOC Day in Potsdam und die Computer- und Technologien-Messe CeBIT in Hannover. Diese Termine werden im Weiteren näher dargestellt.

4.7.1 Future SOC Day

Das Future SOC (FSOC) Lab am Hasso-Plattner-Institut ist eine Kooperation des Instituts mit verschiedenen Partnern aus der Industrie. Ziel des Labs ist die Förderung des Austauschs zwischen Forschungsinstitutionen und der Unternehmen aus der Wirtschaft[39]. Auch die Teilgruppe Prozessorientierte Business Intelligence mit SAP HANA nutzt für ihr Projekt die Ressourcen des Labs. Mit dem FSOC Day findet mehrmals im Jahr eine Informationsveranstaltung statt. Am Mittwoch, den 27. September 2013 beteiligte sich die Teilprojektgruppe am FSOC Day. Dieser findet mehrmals im Jahr am SAP Innovation Center in Potsdam statt. Zur Vorbereitung auf diesen Termin wurden zwei Poster im DIN A1 (Vgl. Anhang F.5.3, F.5.4) und ein mehrseitiges Paper (Vgl. Anhang F.5.1) in der Größe DIN A4 mit der Vorstellung der Projektarbeit im Bereich Smart Metering erstellt. Es wurden insgesamt neun Vorträge von Vertretern verschiedener Hochschulen gehalten, in denen unterschiedliche wissenschaftliche Anwendungsszenarios zur Nutzung von SAP HANA vorgestellt wurden. Außerdem wurden die Poster aller Projekte ausgestellt. Die Teilgruppe Smart Metering hat im Rahmen der Veranstaltung ihr VEE-Konzept vorgestellt. Ein zweiter FSOC Day findet nach Abschluss des Projektzeitraums am 09. April 2014 statt. Auch für diesen hat die Projektgruppe ein Poster (Vgl. Anhang F.5.5) und ein Paper (Vgl. Anhang F.5.2) vorbereitet.

4.7.2 CeBIT

Das Centrum für Büroautomation, Informationstechnologie und Telekommunikation (CeBIT) ist eine Messe im Bereich Informationstechnik. Sie findet seit 1986 auf dem Messegelände in Hannover statt und wird von der Deutschen Messe AG (DMAG) veranstaltet (Vgl. DMAG 2009). Das Land Niedersachsen ist unter anderem mit dem Gemeinschaftsstand „Innovationsland Niedersachsen“ vertreten, wodurch es der Projektgruppe möglich ist, die Projektergebnisse dort auszustellen. Koordiniert wird der Messestand von dem Niedersächsischen Ministerium für Wissenschaft und Kultur (MWK). Insgesamt stellen 18 Hochschulen gemeinsam mit Unternehmen und Gründern ihre Exponate auf dem Gemeinschaftsstand aus (Vgl. MWK 2014). Für den Messestand muss die Projektteilgruppe Smart Metering, dem Corporate Identity des Landes Niedersachsen entsprechend, die Projektergebnisse aufbereiten. Die Präsentation auf dem Messestand besteht im wesentlichen aus den vier Artefakten Standbroschüre, Grafiktafel, Exponat und Flyer. In der Standbroschüre (Vgl. Anhang F.6.1) sind die einzelnen Projekte

[39] `http://www.hpi.uni-potsdam.de/forschung/future_soc_lab.html`

der Aussteller des Gemeinschaftsstandes beschrieben. Sie soll dem Besucher als Leitfaden für den Standbesuch dienen. Sowohl auf deutsch als auch auf englisch wird die Projektidee in der Standbroschüre beschrieben. Außerdem werden zwei Bilder zur besseren Illustration eingefügt. Des Weiteren verfügt jeder Aussteller über eine 2,50 * 1,00 m große Grafiktafel (Vgl. Anhang F.6.2). Diese dient als Blickfang mit aussagekräftigen Kurzinformationen zur Ansprache der Fachbesucher. Der Inhalt wird ebenfalls in beiden Sprachen von der Projektteilgruppe Smart Metering bereitgestellt. Aufgrund der Zeichenbegrenzung muss dies möglichst prägnant erfolgen. Bedingt durch die Größe der Grafiktafel müssen die zwei bereitgestellten Bilder entsprechend skaliert werden.
Der Flyer (Vgl. Anhang F.6.3) soll dazu dienen, dem Besucher weitere Informationen über die Projektgruppe mitzugeben. Neben Kontaktinformationen zu dem Praxispartner CX4U AG und der Universität Oldenburg wird die Projektgruppe kurz vorgestellt. Außerdem wird grundlegend die Projektidee skizziert, das VEE-Framework erklärt und ein Architekturüberblick gegeben. Dies geschieht ebenfalls sowohl auf deutsch als auch auf englisch.
Aufgrund schlechter Erfahrungen mit der Internetverbindung auf der CeBIT in den vergangenen Jahren, werden die Projektergebnisse als Exponat in Form einer Präsentation ausgestellt. Eine Präsentation soll automatisch die ganze Zeit laufen. In dieser wird bereits ein grober Architekturüberblick gezeigt und der VEE-Prozess erklärt. Eine weitere Präsentation dient dazu, interessierten Besuchern die Projektergebnisse detaillierter im direkten Gespräch zu erklären.

4.8 Schlussbetrachtung

Dieser Abschnitt dient der abschließenden und zusammenfassenden Betrachtung der Arbeiten und Ergebnisse der Teilgruppe Smart Metering, so wie sie in den vorangegangenen Kapiteln dargestellt und diskutiert wurden.

In einer ersten Phase wird versucht, die Problemstellung bzw. den Hintergrund der von der Uni sowie der CX4U konzipierten Aufgabenstellung zu analysieren und zu verinnerlichen. Gerade vor dem Hintergrund, dass die betrachtete Branche ebenso wie die verwendeten Technologien für alle Teilnehmer absolutes Neuland waren, ist es nachvollziehbar, dass dieser Vorgang einiges an Zeit in Anspruch nahm. Des Weiteren war die Ausarbeitung und Beschreibung des VEE-Konzeptes von vornherein nicht als starre Vorgabe, sondern vielmehr als kontinuierliche Entwicklung, in welcher sowohl die Betreuer auf Seiten der CX4U, als auch die Mitglieder der Teilgruppe ihre Ansichten und Ideen einbrachten, angedacht. Entsprechend entwickelte sich ein vollständiges und von allen Seiten akzeptiertes Verständnis des Konzeptes und der damit verbundenen Aufgabenstellung nach und nach während dieser ersten Phase.
Das konzeptionelle Ergebnis dieses Zeitraums ist die in Kapitel 4.3 aufgeführte Anforderungsspezifikation, welche sowohl den Hintergrund dieser Projektaufgabe, als auch die grundlegende Aufgabenstellung und daraus abgeleitete Teilergebnisse, sowie die zugehörigen Anforderungen an die Ergebnisse enthält. Sie stellt gewissermaßen die Basis für alle anschließenden Tätigkeiten der Teilgruppe dar.

Auf Grund der von der CX4U angestrebten kurz- sowie langfristigen Weiterentwicklung ihres Produktportfolios und dem daraus resultierenden verbesserten Umgang mit zukünftigen Herausforderungen, wurde schnell klar, dass das Hauptaugenmerk der Projektarbeit auf der Ausarbeitung und Durchführung einer Machbarkeitsstudie lag. Dies war in sofern notwendig, als dass jegliche weiterführende Entwicklungsarbeiten sinnlos geworden wären, falls die bestehende Systemlandschaft nicht mit dem angestrebten Zielsystem kompatibel wäre. Entsprechend viel Zeit wurde in die Untersuchung und Ausarbeitung dieser elementaren Aufgabe investiert. So wurden, wie in Kapitel 4.5 dargestellt, zunächst mehrere denkbare Alternativen betrachtet und unter Berücksichtigung gewisser Parameter wie Kostenneutralität gegeneinander abgewogen. Die am geeignetsten eingestufte Methode wurde dann testweise implementiert um die Machbarkeit in der Praxis zu beweisen.

Die dritte große Bearbeitungsphase stellte die Implementierung im Rahmen einer grundlegenden Umsetzung der VEE-Funktionalität dar. Auf Grund der geringen Erfahrung der einzelnen Mitglieder im Umgang mit Technologien wie ABAP oder XSJS, beinhaltete sie eine notwendige Einarbeitungszeit. Außerdem war es auf Grund der geringen Verbreitung und des innovativen Charakters äußerst schwierig, zielführende Hilfestellungen zu bestimmten Fragestellungen zu bekommen. So stellten beispielsweise prinzipiell simple Probleme enorm große Hürden bei der Umsetzung dar, so dass es insgesamt sehr lange gedauert hat, brauchbare Ergebnisse zu erzielen. Nichtsdestotrotz wurde die grundlegende Funktionalität des VEE-Konzeptes in der

SAP HANA Umgebung umgesetzt. Die Arbeiten sind im Kapitel 4.4, welches somit das dritte große Dokument bildet, dokumentiert.

Neben diesen drei im Allgemeinen aufeinander aufbauenden Projektphasen, fanden während der Projektarbeit noch einige begleitende Aktivitäten Anwendung. So wurde, wie in Kapitel 4.6.1.1.5 erläutert, darauf geachtet, ein durchgängiges Qualitätsmanagement durchzuführen, um die laufenden Arbeiten und die daraus resultierenden Ergebnisse zu prüfen und zu überwachen. Außerdem fand während der Projektzeit zum einen der in Kapitel 4.7 beschriebene FSOC Day beim HPI und zum anderen im März 2014 die CeBIT in Hannover statt, wofür jeweils einiges an Vorarbeit geleistet werden musste.

Zusammenfassend lässt sich sagen, dass die Arbeit der Teilgruppe von einigen Herausforderungen und teilweise sogar von Rückschlägen geprägt war, die vorrangige Zielsetzung, nämlich die Erbringung eines Beweises für die praktische Kommunikationsfähigkeit im Rahmen einer Machbarkeitsstudie, dennoch voll erreicht wurde. Darüber hinaus wurde mit der Implementierung des VEE-Konzeptes in HANA die Basis für weitere Entwicklungen geschaffen, sodass diese, sofern die CX4U sich für eine Migration nach HANA entscheidet, sicherlich als Ausgangspunkt genutzt werden kann.

5 Fazit

Big Data stellt für Daten verarbeitende Unternehmen eine große Chance dar. Der Umgang mit diesem Thema geht jedoch mit diversen Herausforderungen einher. Dass die Praxispartner die Projektgruppe HAWK mit der wissenschaftlichen Untersuchung von Aufgaben im Kontext von Big Data beauftragten, ist ein weiterer Beleg für die Aktualität und die großen Erfolgsaussichten. Die Projektgruppe nahm sich deshalb dieser Thematik an und erzielte innerhalb von zwölf Monaten Projektarbeit die vorgestellten Ergebnisse.

Die Teilprojektgruppe „Big Data im IT-Capacity Management“ behandelte die Aufgabenstellung zum Thema „Herausforderungen durch Big Data am Beispiel des IT-Capacity Managements der Volkswagen AG“, die von der Volkswagen AG beauftragt wurde. Im Rahmen eines BI-Projekts zur Erweiterung des IT-Capacity Managements ergaben sich diese verschiedenen Fragestellungen, da die zu verarbeitende Datenmenge eine neue Dimension erreichte. Daher ergab sich unter anderem die Aufgabe zur Erstellung eines Marktüberblicks, der eine Auswahl an aktuellen Big Data Produkten analysiert (Vgl. Kap. 3.4). Dabei wurden verschiedene Alternativen für den Umgang mit Big Data aufgezeigt. Die Aktualität der Thematik zeigte sich in den deutlichen Unterschieden zwischen den Big Data-Lösungen der Anbieter. Während der Untersuchung wurden einige Produkte durch die Hersteller mehrfach verändert und weiterentwickelt, was häufige Anpassungen für Teilprojektgruppe bedeutete.

Neben der Untersuchung vorhandener Produkte wurden verschiedene Referenzarchitekturen für die Integration von Big Data in bestehende BI-Architekturen entwickelt (Vgl. Kap. 3.3). Eines dieser Szenarien wurde exemplarisch umgesetzt. Möglichkeiten zum Einsatz der Big Data-Technologie Hadoop wurden theoretisch und praktisch untersucht. Um für das IT-Capacity Management eine Datengrundlage zur Herkunft der Kapazitätsanforderungen zu erhalten, leitet das BI-CC der Volkswagen AG Kennzahlen aus den Access Logs der Server ab und integriert sie in die bestehende BI-Architektur. Diesen ETL-Prozess hat die Teilprojektgruppe nachvollzogen und die Ergebnisse dargestellt (Vgl. Kap. 3.5). Anhand dieser Daten wurde ein Prototyp zur Simulation mit der Software IBM Cognos TM1 erstellt (Vgl. Kap. 3.6), der Möglichkeiten zur Analyse, Planung und Steuerung der Kapazitätsauslastung bereitstellt.

Die Zusammenarbeit mit der Volkswagen AG ermöglichte der Teilprojektgruppe aktuelle Software-Lösungen im Kontext von Big Data auszuprobieren. Die zeitliche Begrenzung der Projektgruppe verhinderte, dass weitere Themen, die im Zuge der Projektarbeit aufgekommen sind, nicht weiter vertieft werden konnten. So bietet es sich an, den Prototyp zur Datenintegration mit weiteren Arten von Datenbanken zu testen und die Resultate miteinander zu vergleichen. Durch die Verwendung einer Hadoop- oder In-Memory-Datenbank sind neue Erkenntnisse zu erwarten und ein Performanzvergleich zur gewählten relationalen Datenbank macht Sinn. Da bereits durch das Einlesen der Access Logs eine sehr große Menge an Datensätzen und Transformationen entsteht, konnte in dieser Projektarbeit nur ein geringer Ausschnitt untersucht werden. Weitere Erkenntnisse sind zu erwarten, wenn die vorgestellten Systeme des Marktüberblicks und

die von der Projektgruppe entwickelten Prototypen in einem produktiven Umfeld eingesetzt werden.

Im Zuge der Projektarbeit der Teilprojektgruppe „Smart Metering“ konnte zusätzlich Öffentlichkeitsarbeit in Form von einer Teilnahme an der CeBIT 2014 sowie an dem FSOC Day beim HPI betrieben werden. Die Mitglieder erhielten die Möglichkeit, ihre Ergebnisse, die in Zusammenarbeit mit dem Praxispartner CX4U AG entstanden sind, der Öffentlichkeit vorzustellen und Feedback von Fachleuten zu erhalten. Die Teilprojektgruppe „Smart Metering“ behandelte das Thema „Prozessorientierte Business Intelligence mit SAP HANA“. Die Aufgabe wurde durch die CX4U AG gestellt. Hierbei lag das Hauptaugenmerk auf der Ausarbeitung und Durchführung einer Machbarkeitsstudie (Vgl. Kap. 4.5), die Aussage darüber treffen sollte, ob die bestehende Systemlandschaft mit dem angestrebten Zielsystem kompatibel ist. Im Zuge der Arbeit wurden mehrere Alternativen miteinander verglichen und schließlich eine als geeignet eingestufte Methode implementiert. Ebenso beinhaltete die Aufgabenstellung die Implementierung von VEE-Funktionalität (Vgl. Kap. 4.4). Ein durchgängiges Qualitätsmanagement begleitete die Umsetzung (Vgl. Kap. 4.6.1.1.5).

Die Teilprojektgruppe „Smart Metering“ hat mit ihrer Arbeit das Produktportfolio der CX4U AG anreichern und optimieren können. Der CX4U AG wird somit die Möglichkeit gegeben, die während der Projektarbeit entstandenen Resultate zukünftig weiter auszubauen und entsprechend am Markt zu platzieren. Denn durch die konzeptionellen und praktischen Arbeiten im Rahmen der Machbarkeitsstudie sowie der Implementierung grundlegender VEE-Funktionalitäten in SAP HANA wurde die Basis für weiterführende Arbeiten und zielführende Fortschritte geschaffen. Zu diesen können vorrangig die Entwicklung zusätzlicher Funktionen im Sinne einer umfassenden VEE-Umsetzung, insbesondere der Definition und Anbindung konkreter Kundenregeln, sowie, je nach Anwendungsbedarf, die Anbindung nachgelagerter Endanwendungen, wie z.B. dem SAP IS-U, gezählt werden. Aus Sicht der CX4U AG bleibt darüber hinaus sicherlich die Frage, ob und in welchem Umfang Produktivdaten nach HANA migriert werden sollen, offen. Vor allem da die Anschaffung bzw. der Einsatz der HANA Applicance mit enormen Kosten verbunden ist.

Rückblickend auf die Zusammenarbeit mit beiden Praxispartnern lässt sich feststellen, dass beide die Herausforderungen mit Big Data für ihr Unternehmen angemessen eingeschätzt haben. Bei beiden Praxispartnern musste ein Weg gefunden werden, unterschiedlich strukturierte Daten in großen Mengen in angemessener und in einer auf das Unternehmen abgestimmten Weise in die vorhandene IT-Infrastruktur zu integrieren. Hierfür wurden der Projektgruppe die entsprechenden Testlabore sowie gewünschten Softwarekomponenten zur Verfügung gestellt. Beiden Teilprojektgruppen ist es gelungen, die Anforderungen ihres Praxispartners zu erfüllen und für sich selbst einen Mehrwert durch die Auseinandersetzung mit dem Thema Big Data in einem Praxisbezug zu generieren.

Zusammenfassend kann man sagen, dass die im Marktüberblick vorgestellten Produkte den neuen Herausforderungen durch Big Data gerecht werden können. Allerdings liegt es an den

Unternehmen zu entscheiden, ob die Anschaffung dieser speziellen Produkte als notwendig erachtet wird. So wurde in der Teilprojektgruppe „Smart Metering“ auf Wunsch des Praxispartners mit SAP HANA gearbeitet, das eines der untersuchten Produkte ist. Dagegen wurde bei der prototypischen Umsetzung der Teilprojektgruppe „Big Data im IT-Capacity Management“ keines der vorgestellten Big Data Produkte verwendet. Vielmehr wurde mit der bei der Volkswagen AG bereits vorhandenen Software gearbeitet, dessen Hersteller sogar explizit eigene Big Data Produkte führt. Hier kann die Vermutung aufkommen, dass es nicht immer zwingend notwendig ist, die sogenannten Big Data Produkte zu erwerben. Es ergibt sich eher die Notwendigkeit einer eingehenden Prüfung der vorhandenen IT-Infrastruktur und der neuen Aufgabe, um unnötige Investitionen zu vermeiden. An dieser Stelle würde eine ausführliche Marktanalyse, die unter der Betrachtung verschiedener Unternehmensszenarien Big Data- und BI-Produkte untersucht, den Unternehmen als Entscheidungsunterstützung dienen.

Im Hinblick auf das Konzept „Projektgruppe“ der CvO-Universität lässt sich sagen, dass sowohl die Universität als auch die Studierenden sowie die Praxispartner von dieser Arbeit profitieren. Anhand der Kooperation mit den Praxispartnern wird deutlich, wie wichtig die Zusammenarbeit der Universität mit Unternehmen für beide Seiten ist. Denn hier bedeutet der Praxisbezug eindeutig, dass die gelernte Theorie in der Universität in der Praxis nicht nur ihre Anwendung findet, sondern dass Studierende durch die Anwendung auch einen eigenen wissenschaftlichen Beitrag zu aktuellen Problemstellungen leisten können. Im Fall dieser Projektgruppe konnte sich so mit einer aktuell sehr bedeutenden Thematik auseinandergesetzt werden. Auf der anderen Seite ist die Zusammenarbeit auch für Unternehmen interessant, da so eigene Problemstellungen wissenschaftlich untersucht werden können. Abseits der im Unternehmen etablierten Ansichten und Lösungen können so neue Vorschläge unterbreitet werden.

Die Projektgruppe HAWK teilte sich nach zwei Monaten Projektlaufzeit in zwei Teilprojektgruppen zur Bearbeitung der Aufgabenstellungen der beiden Praxispartner. Begonnen wurde das Thema Big Data gemeinsam mit der Bearbeitung der Seminararbeiten. So ist es sinnvoll, dass sich alle erarbeiteten Inhalte in diesem Abschlussbericht wiederfinden und sich Parallelen ziehen lassen. Es konnten nur zwei Aufgabenstellungen untersucht werden, doch schon allein diese erforderten viele verschiedene Sichtweisen und Disziplinen. Insgesamt ist das Thema aber noch viel breiter gefächert. Big Data gibt ausreichend Grundlage für weitere wissenschaftliche Untersuchungen und hat sehr gute Chancen, in Unternehmen produktiv umgesetzt zu werden.

F Anhang

Kapitelübersicht

F.1 Abbildungen . . . . . **ii**
F.2 Listings . . . . . **iv**
F.3 Pflichtenheft Datenintegration . . . . . **ix**
F.4 Pflichtenheft Simulation . . . . . **x**
F.5 FSOC-Day . . . . . **xi**
F.5.1 FSOC-Paper 2013 xii
F.5.2 FSOC-Paper 2014 xv
F.5.3 FSOC-Poster-1 2013 xviii
F.5.4 FSOC-Poster-2 2013 xix
F.5.5 FSOC-Poster 2014 xx
F.6 CeBit . . . . . **xxi**
F.6.1 CeBIT Standbroschüre xxii
F.6.2 CeBIT Grafiktafel xxiii
F.6.3 CeBIT Flyer xxiv
F.7 OpenVPN-Anleitung vom HPI . . . . . **xxvi**
F.8 Latex-Wiki . . . . . **xxx**

F.1 Abbildungen

Projektgruppe SAP HANA TG-CX4U	Marco	Benjamin	Jad	Christoph	Jan-Hendrik	Jan-Patrick
Teilprojekt	CX4U	CX4U	CX4U	CX4U	CX4U	CX4U
Programmierung						
ABAP	0: kein Vorwissen	0: kein Vorwissen	0: kein Vorwissen	1: einfaches Vorwissen	0: kein Vorwissen	0: kein Vorwissen
Java	1: einfaches Vorwissen	1: einfaches Vorwissen	1: einfaches Vorwissen	1: einfaches Vorwissen	2: gutes Vorwissen	1: einfaches Vorwissen
R	0: kein Vorwissen	1: einfaches Vorwissen	0: kein Vorwissen	0: kein Vorwissen	0: kein Vorwissen	1: einfaches Vorwissen
SQL	0: kein Vorwissen	1: einfaches Vorwissen	2: gutes Vorwissen	1: einfaches Vorwissen	3: praktische Erfahrung	1: einfaches Vorwissen
C	0: kein Vorwissen	1: einfaches Vorwissen	2: gutes Vorwissen	1: einfaches Vorwissen	0: kein Vorwissen	1: einfaches Vorwissen
Anforderungserhebung						
Interviews führen	2: gutes Vorwissen	3: praktische Erfahrung	1: einfaches Vorwissen	1: einfaches Vorwissen	0: kein Vorwissen	3: praktische Erfahrung
Methoden zur Dokumentation (Bsp.: Kontextdiagramm, Strukturierung von Anforderungen, Ereignislisten)	3: praktische Erfahrung	3: praktische Erfahrung	1: einfaches Vorwissen	3: praktische Erfahrung	0: kein Vorwissen	3: praktische Erfahrung
SPSS	0: kein Vorwissen	1: einfaches Vorwissen	0: kein Vorwissen	1: einfaches Vorwissen	0: kein Vorwissen	0: kein Vorwissen
Analyse und Design						
UML	2: gutes Vorwissen	2: gutes Vorwissen	2: gutes Vorwissen	2: gutes Vorwissen	1: einfaches Vorwissen	1: einfaches Vorwissen
fachlicher Überblick (Bsp.: Anwendungsfälle, Aktivitätsdiagramm, fachliches Datenmodell, Dialoge)	2: gutes Vorwissen	2: gutes Vorwissen	3: praktische Erfahrung	2: gutes Vorwissen	1: einfaches Vorwissen	2: gutes Vorwissen
technischer Überblick (Bsp.: Schichten- und Verteilungsmodell (technische Architektur), fachliche Arichtektur (Datenmodell mit Methoden)	2: gutes Vorwissen	1: einfaches Vorwissen	1: einfaches Vorwissen	2: gutes Vorwissen	0: kein Vorwissen	1: einfaches Vorwissen
Dokumentation						
LaTex	0: kein Vorwissen	3: praktische Erfahrung	0: kein Vorwissen	2: gutes Vorwissen	3: praktische Erfahrung	1: einfaches Vorwissen
Bachelorarbeit (Schlüsselbegriffe)	Kennzahlensysteme, Anforderungserhebung, Prozessoptimierung, Prozessbeschreibung, Prozessmodellierung	Anforderungserhebung, Anforderungsanalyse, Prozessmodellierung, Produktvergleich	ERP-System Entwicklung(C#, SQL Server), Prozessanalyse, Prozessmodellierung, Prozessoptimierung, Prozessautomatisierung	Antragstellung, Hochschulförderung, Förderprogramme, Anforderungsanalyse, Entwurf von und Vergleich von Konzepten	Configuration und Asset-Management SOAP PHP MySQL	Lastenheft, Anforderungserhebung, Prozessoptimierung

Abbildung F.1: Kompetenzmatrix der TG-CX4U

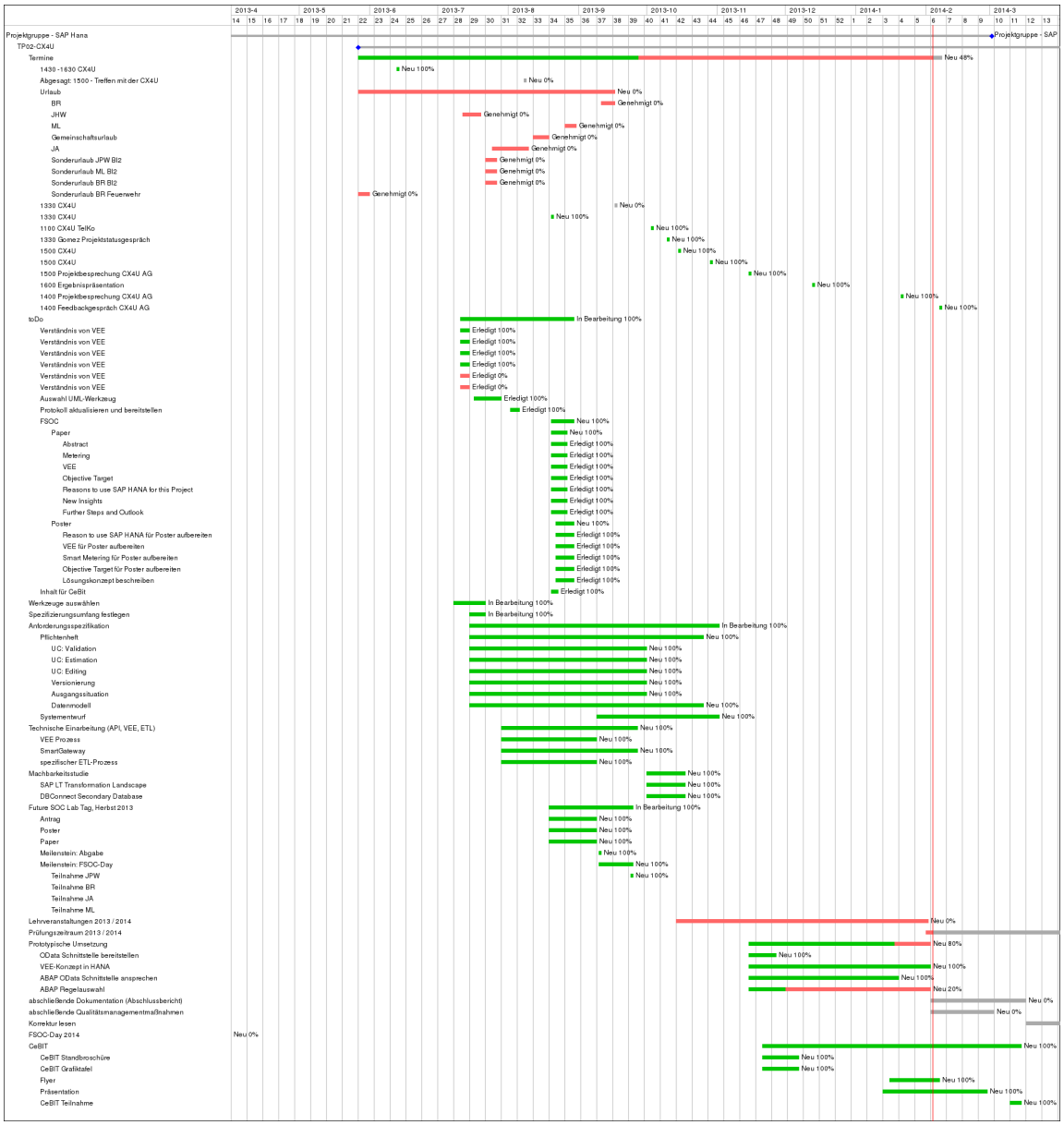


Abbildung F.2: TG-CX4U Redmine Projektplan

F.2 Listings

```
1
2  *&---------------------------------------------------------------------*
3  *& Report  ZPG_ODATA_TEST
4  *&
5  *&---------------------------------------------------------------------*
6  *&
7  *&
8  *&---------------------------------------------------------------------*
9
10 REPORT  ZPG_ODATA_TEST.
11
12
13 DATA: http_dest TYPE rfcdest VALUE 'PG_HANA'.
14 DATA: CLIENT TYPE REF TO if_http_client.
15 DATA: request TYPE REF TO if_http_request,
16       subrc          TYPE sysubrc,
17       errortext      TYPE string,
18       dummy          TYPE string,
19       timeout        TYPE I.
20
21
22 START-OF-SELECTION.
23
24 *Verwendung der mit SM59 gepflegten HTTP-Verbindung
25 cl_http_client=>create_by_destination( EXPORTING
26 destination = http_dest IMPORTING CLIENT = CLIENT ).
27
28 *Methode des Client Objekts auf get setzen
29 CALL METHOD CLIENT->request->set_method(
30 if_http_request=>co_request_method_get ).
31
32 *http Protokoll auswaehlen. Muss dem der Zielmaschine entsprechen
33 CLIENT->request->set_version(
34 if_http_request=>co_protocol_version_1_1 ).
35
36
37 *Send
38 CALL METHOD CLIENT->send
39 EXPORTING  timeout = timeout
40 EXCEPTIONS http_communication_failure  = 1
41   http_invalid_state          = 2
42   http_processing_failed      = 3
43   OTHERS                      = 4.
44
45 *Fehler abfangen
46 IF sy-subrc <> 0.
47   CALL METHOD CLIENT->get_last_error
```

```
48   IMPORTING CODE    = subrc
49     MESSAGE = errortext.
50   WRITE: / 'communication_error( send )',
51   / 'code: ', subrc, 'message: ', dummy.
52   EXIT.
53 ENDIF.
54
55
56 *receive
57 CALL METHOD CLIENT→receive
58 EXCEPTIONS http_communication_failure  = 1
59   http_invalid_state          = 2
60   http_processing_failed      = 3
61   OTHERS                      = 4.
62
63 *Fehler abfangen
64 IF sy-subrc <> 0.
65   CALL METHOD CLIENT→get_last_error
66   IMPORTING CODE    = subrc
67     MESSAGE = errortext.
68   WRITE: / 'communication_error( receive )',
69   / 'code: ', subrc, 'message: ', dummy.
70   WRITE: errortext.
71
72   EXIT.
73 ELSE.
74 ENDIF.
75
76 FORMAT COLOR col_group INTENSIFIED off.
77
78 DATA: FIELDS TYPE tihttpnvp,
79       fields_wa TYPE ihttpnvp.
80
81 *Header-Felder auslesen
82 CALL METHOD CLIENT→response→get_header_fields
83 CHANGING FIELDS = FIELDS.
84
85 *Unter Programm zur Ausgabe der Daten ausfuehren
86 PERFORM write_document.
87
88 *Verbindung schliessen
89 CALL METHOD CLIENT→close
90 EXCEPTIONS
91   http_invalid_state = 1
92   OTHERS             = 2.
93
94
95 *&amp
     ;-----------------------------------------------------------------*
```

```
*&amp;      Form  write_document
*&amp
    ;-----------------------------------------------------------------------*
*       text
*-----------------------------------------------------------------------*
*  --> p1        text
*  <--  p2        text
*-----------------------------------------------------------------------*
FORM write_document.
  DATA: DATA TYPE string.
  DATA: l_width TYPE I,   "length of line
        l_width_f TYPE I, "length of field name output
        l_width_v TYPE I, "length of field value output
        l_width_b TYPE I, "length of body line output
        l_width_h TYPE I. "length of field value for multi-line output
  DATA: LINE(142) TYPE C, "must be of length = l_width_b
        len TYPE I,
        llen TYPE I.
  DATA: l_lvalue TYPE flag. "flag for multi-line output.

*Ab hier wird die GUI zur Ausgabe aufgebaut. Zunaechst fuer den
*Header dann fuer den Body
* init values for output length
  l_width = 145.
  l_width_f = 40.
  l_width_v = 100.
  l_width_b = 142.
  l_width_h = l_width_f + l_width_v.

  WRITE: / sy-ULINE(l_width) NO-gap.
* title
  FORMAT COLOR COL_NORMAL INTENSIFIED ON.
  WRITE: / sy-vline NO-gap,
  AT (l_width_b) 'header fields ' CENTERED.
  WRITE: sy-vline NO-gap.
  WRITE: / sy-ULINE(l_width) NO-gap .

* check length of field values
  LOOP AT FIELDS INTO fields_wa.
    len = STRLEN( fields_wa-VALUE ).
    IF len GT l_width_v.
      l_lvalue = 'X'.
      EXIT.
    ENDIF.
  ENDLOOP.

* table header
  IF l_lvalue IS INITIAL.
    FORMAT COLOR col_heading INTENSIFIED ON.
```

```
    WRITE: / sy-vline NO-gap,
    AT (l_width_f) 'field name' , sy-vline NO-gap,
    AT (l_width_v) 'field value' ,
    sy-vline NO-gap.
    WRITE: / sy-ULINE(l_width) NO-gap.

    FORMAT COLOR COL_NORMAL INTENSIFIED off.
    LOOP AT FIELDS INTO fields_wa.
      WRITE: / sy-vline NO-gap, AT (l_width_f) fields_wa-name ,
      sy-vline NO-gap, AT (l_width_v) fields_wa-VALUE ,
      sy-vline .
    ENDLOOP.
    WRITE: /  sy-ULINE(l_width) NO-gap.
  ELSE.
* multi-line output for header field values
* table header
    FORMAT COLOR col_heading INTENSIFIED ON.
    WRITE: / sy-vline NO-gap,
    AT (l_width_b) 'field name' , sy-vline NO-gap.
    WRITE: / sy-vline NO-gap,
    AT 4(l_width_h) 'field value' ,
    sy-vline NO-gap.
    WRITE: / sy-ULINE(l_width) NO-gap.

    DATA: l_int_flag TYPE C.
    FORMAT COLOR COL_NORMAL.
    LOOP AT FIELDS INTO fields_wa.
      IF l_int_flag IS INITIAL.
        FORMAT INTENSIFIED off.
        l_int_flag = 'X'.
      ELSE.
        FORMAT INTENSIFIED ON.
        CLEAR l_int_flag.
      ENDIF.
      WRITE: / sy-vline NO-gap, AT (l_width_b) fields_wa-name ,
      sy-vline NO-gap.
      len = STRLEN( fields_wa-VALUE ).
      WHILE len > l_width_h.
        WRITE: / sy-vline NO-gap, AT 4(l_width_h) fields_wa-VALUE ,
        sy-vline NO-gap.
        fields_wa-VALUE = fields_wa-VALUE+l_width_h.
        len = STRLEN( fields_wa-VALUE ).
      ENDWHILE.
      WRITE: / sy-vline NO-gap, AT 4(l_width_h) fields_wa-VALUE ,
      sy-vline NO-gap.
    ENDLOOP.
    WRITE: /  sy-ULINE(l_width) NO-gap.

  ENDIF.
```

```
193 *output data
194
195
196 *Hier werden die Daten aus dem zurueckgelieferten Wert ausgelesen
197   DATA = CLIENT→response→get_cdata(  ).
198   len = STRLEN( DATA ).
199
200 *Mit einer Schleife werden die Daten in das Body-Feld geschrieben
201   WRITE: / sy-ULINE(l_width) NO-gap.
202   FORMAT COLOR col_heading INTENSIFIED off.
203   WRITE: / sy-vline, AT (l_width_b) 'body' .
204   WRITE AT l_width sy-vline.
205   FORMAT COLOR COL_NORMAL INTENSIFIED off.
206   WHILE len > 0.
207     LINE  = DATA+llen(*).
208     WRITE: / sy-vline NO-gap.
209     WRITE:  LINE.
210     WRITE AT l_width sy-vline.
211     len = len - l_width_b.
212     llen = llen + l_width_b.
213   ENDWHILE.
214   IF llen < len.
215     WRITE: / sy-vline.
216     LINE  = DATA+llen(*).
217     WRITE AT l_width sy-vline.
218     WRITE:  LINE.
219   ENDIF.
220
221   WRITE: / sy-ULINE(l_width) NO-gap.
222
223 ENDFORM.                    " write_document
```

Listing 6.1: Quellcode HTTP-Request

F.3 Pflichtenheft Datenintegration

In dieser zur Veröffentlichung vorgesehenen Version entfernt.

F.4 Pflichtenheft Simulation

In dieser zur Veröffentlichung vorgesehenen Version entfernt.

F.5 FSOC-Day

Für die Teilnahmen der Teilgruppe Smart Metering am FSOC Day wurden Paper und Poster angefertigt, die im Folgenden zu finden sind.

F.5.1 FSOC-Paper 2013

Integration of a VEE-Framework within a Smart Gateway into SAP HANA

Jan-Patrick Weiß
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
jan-patrick.weiss@uni-oldenburg.de

Benjamin Reinecke
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
benjamin.reinecke@uni-oldenburg.de

Marco Lucht
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
marco.lucht@uni-oldenburg.de

Jad Asswad
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
jad.asswad@uni-oldenburg.de

Jan Hendrik Wege
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
jan-hendrik.wege@uni-oldenburg.de

Christoph Walther
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
christoph.walther@uni-oldenburg.de

Abstract

The energy market is changing by legislation so by 2020 smart meters will be installed in most households. Therefore the IT systems of energy companies have to adjust to deal with the growing amount of measured data. With smart meters it is possible to measure the consumption every 3 minutes and shorter instead of measuring only one annual value. This near real time data consumption allows companies to create new business use cases. They can for example predict the usage of their grids and close short time contracts with suppliers to meet the demands. Furthermore there are new possibilities for analysis and assurance of data quality. In order to assure a defined integrity, possible missing or false values of the available measuring data have to be corrected. The project deals with these problems of data quality and handling of these massive data. To ensure integrity and the best quality of data, a VEE-Framework (Validation, Estimation, Editing) will take a place. The VEE-Framework shall be implemented within a Smart Gateway and integrated into SAP HANA to benefit from the efficiency of in-memory-database technology.

1. Introduction

This document shows the main idea of the project "Integration of a VEE-Framework within a Smart Gateway into SAP HANA". At first the business context of smart metering in modern households is described. After that an explanation of the VEE-Framework will demonstrate how smart meter data can be used and how data is being precessed. In the part "objective target" the goals of the project are described. Furthermore the reasons to use SAP HANA for this project are explained. At the end new insights and further steps and outlook are shown.

2. Smart Metering

Smart Metering describes a method to count, document and analyze the consumption (electricity, gas, water etc.) by using electrical meters. The measurement data can be recorded in individual time inter-

vals (minutes, hours, days etc.) and is automatically submitted to the measurement data responsible, which will submitted to the supplier, which can use this information e.g. for monitoring or billing processes. Smart meters enable two-way communication between the meter and the suppliers' system. Furthermore they can be used to gather information for remote reporting. When using additional technical functions, the consumer is able to monitor his consumption and to make constant analyzes. Energy hogs can be identified and the use of energy-intensive devices can be switched to time frames, where energy is less expensive. That may lead to decrease energy costs. Furthermore smart meters can be part of a smart grid.

3. VEE-Framework - Validation, Estimation, Editing

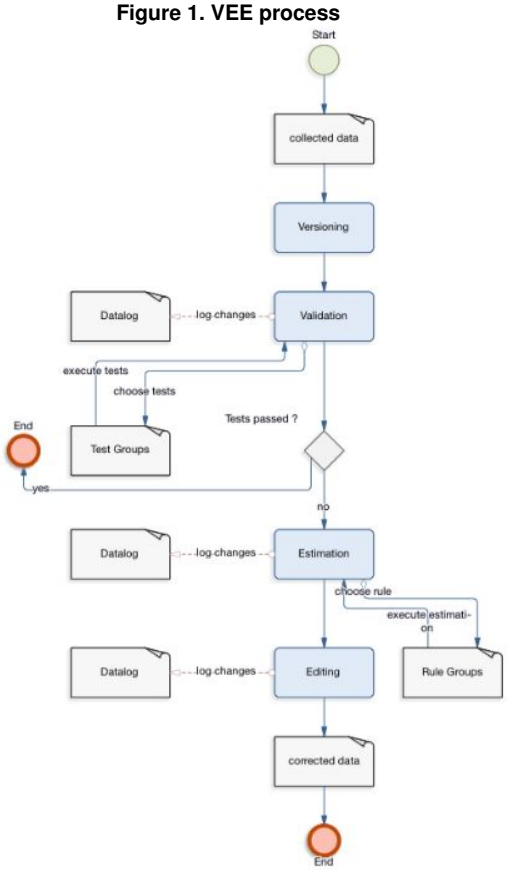
Metering data from different sources is a very extensive process which needs to provide a large amount of functionality and robustness, not only hardware but software based. However, potential errors can never be barred and have to be treated respectively.Therefore a process, called **V**alidation, **E**stimation and **E**diting (VEE) is introduced.The main objective is to offer a set of methods which allow a partly automated approach of specific error handling based on predefines rules. This allows business users to identify erroneous and potentially problematic data before it is used by downstream applications. It is important to state that all changes that are made to the considered data have to be logged in order to reconstruct them in the future.

In a first step, the source dataset needs to be validated in order to determine inconsistencies, e.g. missing values, duplicates or spikes. Therefore the necessary meter data has to be identified, collected and versionized. Afterwards, several tests like Time Tollerance-, Sum- and Spike Checks are peformed and evaluated on the specific values. The various test types are identified and selected automatically via configurable parameters within the meter read. Depending on the test results, the individual values need to be marked as failed and subsequently further processed.

During estimation phase, all previously marked values shall be adapted and corrected in order to correspond to their specific requirements. Therefore several appropriate replacement values are generated by choosing a specific set of customer based methods and applying them to the read meter data. These rules can be assigned to individual customers as well as to groups of various customers and are furthermore exchangeable in any order. It is important that those values which have been estimated or treated in any way are marked accordingly in order to identify them for later use.

The third step of VEE process implies the actual editing of the available meter data. This includes writing the individual replacement value into the data set as well as setting the status parameter to its designated value (see figure 1).

Figure 1. VEE process



Due to the customizable sets of rules, the automatic data validation and estimation and the consecutive logging functionalities, the VEE process guarantees an accurate and quick verification of read meter data which is a necessary assumption for various business use cases.

4. Objective Target

The project group is aiming to develop an SAP-HANA appliance which fits into the existing systems used for measuring customers' energy consumption and creating the customers' invoices. Therefore the following steps shall be taken:

1. When meter-values for a certain customer are requested by SAP e.g. for creating an invoice, SAP HANA loads the data (from the PI-Server or the CX4U Smart-Gateway).
2. The loaded consumption values from the gas meters are checked for faulty or even missing values (either within the Smart Gateway or in SAP HANA).
3. If such a value is found, it is being replaced by a calculated value which was calculated by a certain pattern fitted to the customer.
4. The generated value is logged to a file, so that the invoice stays transparent.
5. Both, calculated and measured values are taken to SAP so that the invoice can be created.

As the calculation method for faulty and missing values depends on the customer and can be changed, the calculation in SAP HANA is just temporary and the values are only passed to the system which requested them and will be deleted afterwards.

5. Reasons to use SAP HANA for this project

SAP HANA is used to optain high performance in processing large amounts of heterogeneous data. By connecting households to smart meters, the massive data volumes increase, since there is not as yet, an annual settlement, but a possibility to measure every 15 minutes, or at even shorter intervals, energy consumption data. Every Smart Meter generates 768 bytes of data per day - with a base of 15,000 meters in the used demonstration system, we are talking about approximately 330 gigabytes of data in an accounting period (normally a month)[1].
In order to forecast electricity consumption for a day, we have to process nearly 11 gigabytes of data. With the available data; analysis, forecasting and quality of results can be increased. In order to generate settlements it is necessary to have correct and complete values. SAP HANA supports this process with statistical functions to find and eliminate missing and non-valid data. This functionality is also called VEE (Validation, Estimation, Editing). When using SAP HANA for processes such as analysis or prediction calculations, it is necessary to integrate the functionalities of SAP HANA with the VEE-Framework within the Smart Gateway.

6. New Insights

Implementing the VEE Framework within SAP HANA and connecting SAP HANA with Smart Gateway leads us to new insights through benefiting from real time processing of the in-memory computing technology. SAP HANA with its simplicity and easy to handle interfaces and functions presents a full database platform in addition to its analytical and transactional capabilities within the attribute, the analytic and the calculation views and other artifacts.

However, despite the capability of HANA to interconnect with other SAP or non-SAP systems or databases, there is always latency in loading and exporting the metering data into and from SAP HANA. These two operations are not automated and time consuming, which is very crucial to our process. In order to handle the data transfer the project group has to provide a solution that transfers the metering data fast, efficiently and accurately.

7. Further Steps and Outlook

Currently the project group cooperates with CX4U Consulting and analyses the data which is provided by the product *SmartGateway*. It is important to understand how the interfaces of the *SmartGateway* work to get the necessary data into SAP HANA. The next steps are to show - within a proof of concept - in which way the VEE-Framework can be implemented into SAP HANA, and in which way SAP HANA can be connected or intgrated with Smart Gateway. There should be several alternatives and the best one should be shown in a prototype.

References

[1] C. Kuhlmann. *E-Mail, 06.08.2013*. CX4U AG, Oldenburg.

F.5.2 FSOC-Paper 2014

Integration of a VEE-Framework into SAP HANA

Jan-Patrick Weiß
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
jan-patrick.weiss@uni-oldenburg.de

Benjamin Reinecke
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
benjamin.reinecke@uni-oldenburg.de

Marco Lucht
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
marco.lucht@uni-oldenburg.de

Jad Asswad
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
jad.asswad@uni-oldenburg.de

Jan Hendrik Wege
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
jan-hendrik.wege@uni-oldenburg.de

Christoph Walther
University of Oldenburg
Department of Computer Science
Uhlhornsweg 84
D-26129 Oldenburg
christoph.walther@uni-oldenburg.de

Abstract

The energy market is changing by legislation so by 2020 smart meters will be installed in most households. Therefore the IT systems of energy companies have to adjust to deal with the growing amount of measured data. With smart metering it is possible to measure the consumption near real time. This allows companies to create new business use cases. They can for example predict the usage of their grids and close short time contracts with suppliers to meet the demands [1][2]. Furthermore there are new possibilities for analysis and assurance of data quality. In order to assure a defined integrity, possible missing or false values of the available measuring data have to be corrected. The project deals with these problems of data quality and handling of these massive data. To ensure integrity and the best quality of data, a VEE-Framework (Validation, Estimation, Editing) has taken place. The VEE-Framework has been implemented into SAP HANA to benefit from the efficiency of in-memory-database technology.

1. Definition of the project

The research project "Integration of a VEE-Framework into SAP HANA" is done in cooperation with CX4U AG. The main idea of the project is to access the Smart Gateway (provided by CX4U) with SAP HANA and to process analysis and predictions of the data within the in-memory database. In the used demonstration system, data of 15,000 smart meters are simulated. Each smart meter generates 768 bytes of data every day which leads to 330 gigabytes of data in an accounting period. It is very important to analyze these data very fast and easily change the rules of analysis depending on the origin of the data. Often there are missing or incorrect values and through the analysis these data get corrected. Each customer and use case has a different rule which defines the methods to correct the data. This function is part of a VEE-Framework in which the data get validated, estimated and edited. The VEE-Framework has been implemented into SAP HANA and within a proof of concept it has been proven how the Smart Gateway and SAP HANA can communicate with each other.

2. Solution alternatives

In the course of the proof of concept, the research led to a variety of different alternatives. The most considered alternatives are provided from SAP itself and listed as SAP Data Provisioning Technologies (Figure 1). Those Technologies are represented by SAP Sybase Event Stream Processor, SAP HANA Smart Data Access and SAP HANA Replication Technologies. The replication Technologies are also divided into four other technologies according to their replication functionality and they are represented by SAP Data Services as ETL-Based Replication, SAP HANA Direct Extractor Connection (DXC) as Extractor-Based Data Acquisition, SAP Sybase Replication Server (SRS) as Log-Based Replication and finally SAP Landscape Transformation Replication Server (SLT) as Trigger-Based Replication [3].

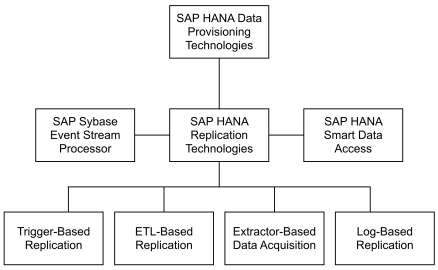


Figure 1. Data Provisioning Technologies

Source: Own figure

SLT, as one of the best provisioning technologies provided from SAP, allows the exchange of data between SAP systems and other SAP or non-SAP systems. With SLT it is possible to continue to use existing system and benefit from the increased speed of SAP HANA. By using the SLT Replication Server, only the tables which are relevant for the data delivery are processed, so that the data exchange remains efficient.

Besides SAP solutions, two other alternatives was considered. One of them is Secondary Database Connection (SDBC). SDBC is used in case real time data processing is not required. This method uses replications of specific data of a database to load it into SAP HANA database. HANA does not act as an alternative of the existing database, furthermore it is an add-on whose main focus is to speed up the reading of data from existing ABAP-Systems. This increased speed is also the main advance of this technology. There are currently three existing alternatives: SDBC with open SQL, native SQL and native SQL - ADBC (ABAP database connectivity). SDBC with native SQL - ADBC provides the ABAP database connectivity concept. By using HANA it is important, that ADBC also provides access to non-data-dictionary-products.

The second alternative is SAP HANA Extended Application Services (XS). The XS-Engine has a high integration level within the SAP HANA itself. With only a simple OData Interface on the HANA side and a HTTP-Request from the NetWeaver side the XS-Engine can fulfill the purpose of the project and assure the connectivity between the two systems in addition to processing the data locally through its server side JavaScript capabilities.

3. Selection criteria

In order to select the suitable solution among the different alternatives, a set of criteria is set to evaluate them. The criteria assess the alternatives according to their efficiency and effectiveness concerning the replication or the communication between SAP HANA and the source system and the transformation capabilities between both of them.

One of the most important aspects is the ability to handle the data at real-time or according to scheduled intervals in a manner that preserve the benefit of using the high speed of SAP HANA in processing metering data. Furthermore, the supporting capabilities of the method are examined to determine the supported databases and the ability to support non-SAP-based systems. On the other hand, the prerequisites of each alternative are listed in order to define the system requirements required to implement the method. The transformation functionalities during the replication or the communication process play a significant role in the evaluation beside the pricing information that help to choose the best alternative within an affordable cost and even better with no cost or extra expenses like the alternative implemented in the project.

The selection criteria are listed as follow:

- Real Time, Interval or Scheduled
- Data Sources
- System Requirements
- Supporting non-SAP-based Systems
- Transformation Functionalities
- Costs

4. Implementation

After evaluating different implementation alternatives, the SAP HANA Extended Application Services (XS-Engine) have been chosen as the suitable solution for this particular project. Among all other alternatives, the XS-Engine fulfils the needed functionalities and provides a real-time performance without any extra expenses or products.

The first step was to create an OData service on the HANA side, so that the Netweaver side could insert data into and get data from the HANA database. At the same time, a functionality on the Netweaver is developed in order to read the OData interface. Afterwards the data had to be processed within the HANA system and therefore several scripts are implemented, which offered the needed VEE-functionality. In addition to the scripts which execute the VEE functions, a web interface had to be developed for managing and calling those implemented functions. The web interface is programmed using HTML and Javascript. Finally it was possible to request data from the OData interface into the Netweaver using a HTTP request. Moreover, the VEE algorithms were able to be used to process the data in the HANA database.

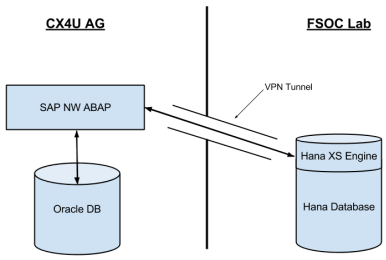


Figure 2. Systemarchitecture
Source: Own figure

5. Results

As a summarization it can be said that the main objectives of this project were two points. First, proving the communication ability between the two separated systems withing a proof of concept. This has been done on a conceptual basis by comparing different implementation methods with each other as well as selecting the most appropriate one and implementing it into a prototypical communication sample.

Second, implementing the VEE-framework which has also been described and evaluated in a conceptual manner, inside the given SAP HANA architecture.
By achieving these goals the results of this project now provide the basic functionality of the VEE-vision and can therefore be seen as a fundamental contribution for further steps of the CX4U in order to handle the existing and future challenges inside the energy and gas market.
Also there has been a huge interest in this particular topic from representatives from different industries as the project group presented its results at the CeBIT 2014 in Hannover, Germany. The feedback which has been gathered at this exhibition shows that the subject of dealing with big data in near real-time or in real-time scenarios is and will become more and more important in the future.

6. Further Steps and Outlook

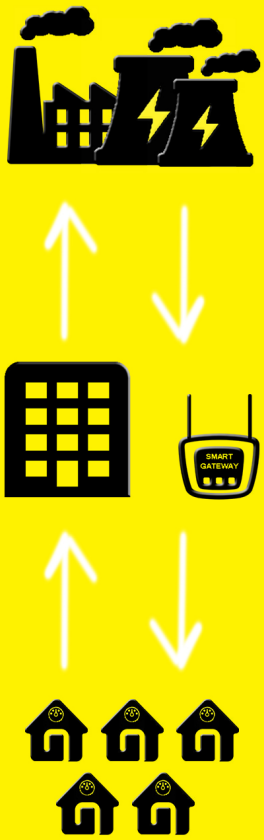
The project has shown that it is possible to handle metering data inside SAP HANA that comes from a specific source system. As this has been a prototypical solution for the desired context, there have to be additional implementations regarding robustness and range of functions. Also specific standards like encrypting the data transfer have to be examined, especially since they come from statutory requirements. Also the set of methods used within the separate VEE-process steps has to be extended in order to achieve customer specific test scenarios.
Considering the increasing amount of assessed data in the near future it seems obvious to deal with software solutions that offer an accelerated performance in handling huge amounts of data. Now that the communication ability has been proven for the productive system at CX4U and SAP HANA, the local management needs to decide if and when they will migrate to SAP HANA.

References

[1] C. Aichele. Innovativ smart meter für die energiemärkte der zukunft nutzen, 2013.
[2] Verbraucherzentrale_Bundesverband. Energiemarkt im wandel. tagung von der verbraucherzentrale bundesverband und der bundnetzagentur, 2013.
[3] SAP AG. SAP HANA Technical Operations Manual, 2013.

F.5.3 FSOC-Poster-1 2013

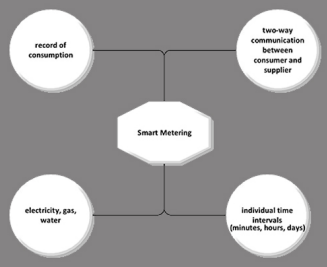
CARL VON OSSIETZKY universität OLDENBURG



Integration of a VEE-Framework within a Smart Gateway into SAP HANA

Smart Metering

Smart Metering is used to count the consumption (electricity, gas, water etc.) by using electrical meters.
Each measurement data can be recorded in indivdual time intervals (seconds, minutes, hours, days etc.).
All data is automatically submitted to the measurement data responsible to use this information for monitoring and billing processes.



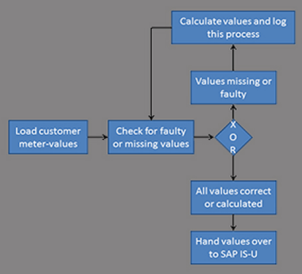
VEE-Framework – Validation, Estimation, Editing

As the process of metering data, whether it is electricity, gas or water energy data, is very complex and highly automated - erroneous or missing data is likely to appear. This fact however may be caused either by an individual factor or a combination of different factors like data communication failures or equipment errors. Therefore it is necessary to treat those circumstances appropriately. The process of Validation, Estimation and Editing introduces a possibility to deal with erroneous data as it not only improves the accuracy of metered data but also consequently raises their significance for the desired purpose. Moreover, all changes that have been applied to the source dataset are logged consecutively in order to be able to reproduce and reconstruct the modifications at a certain point of time. The VEE process is partitioned as follows:

1. **Validation:** In order to determine inconsistencies like missing values, duplicates or spikes, the source dataset is validated with a specific set of tests that are executed on the individual values.
2. **Estimation:** All values that have been marked as erroneous are corrected with a customer based set of methods that are applied on the dataset. All methods inside a certain set are exchangeable customizable in any way.
3. **Editing:** The replacement values of the corrected data are written to the source dataset.

Objective Target

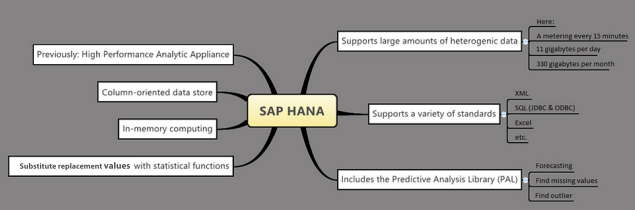
1. Load consumption data of a customer
2. Check the values for faulty or missing ones
3. Calculate values for missing or faulty values after a predened pattern
4. Log the process of calculating values for transparency
5. Send calculated and original values to SAP IS-U for processing



Reasons to use SAP HANA

SAP HANA is used to obtain high performance in processing large amounts of heterogeneous data.

- By connecting households to smart meters, the massive data volumes increase
- Approximately 330 gigabytes of data in an accounting period (normally a month)
- With the available data, analysis and forecasting is possible and the quality of results can be increased
- HANA supports this process with statistical functions to find and eliminate missing and non-valid data

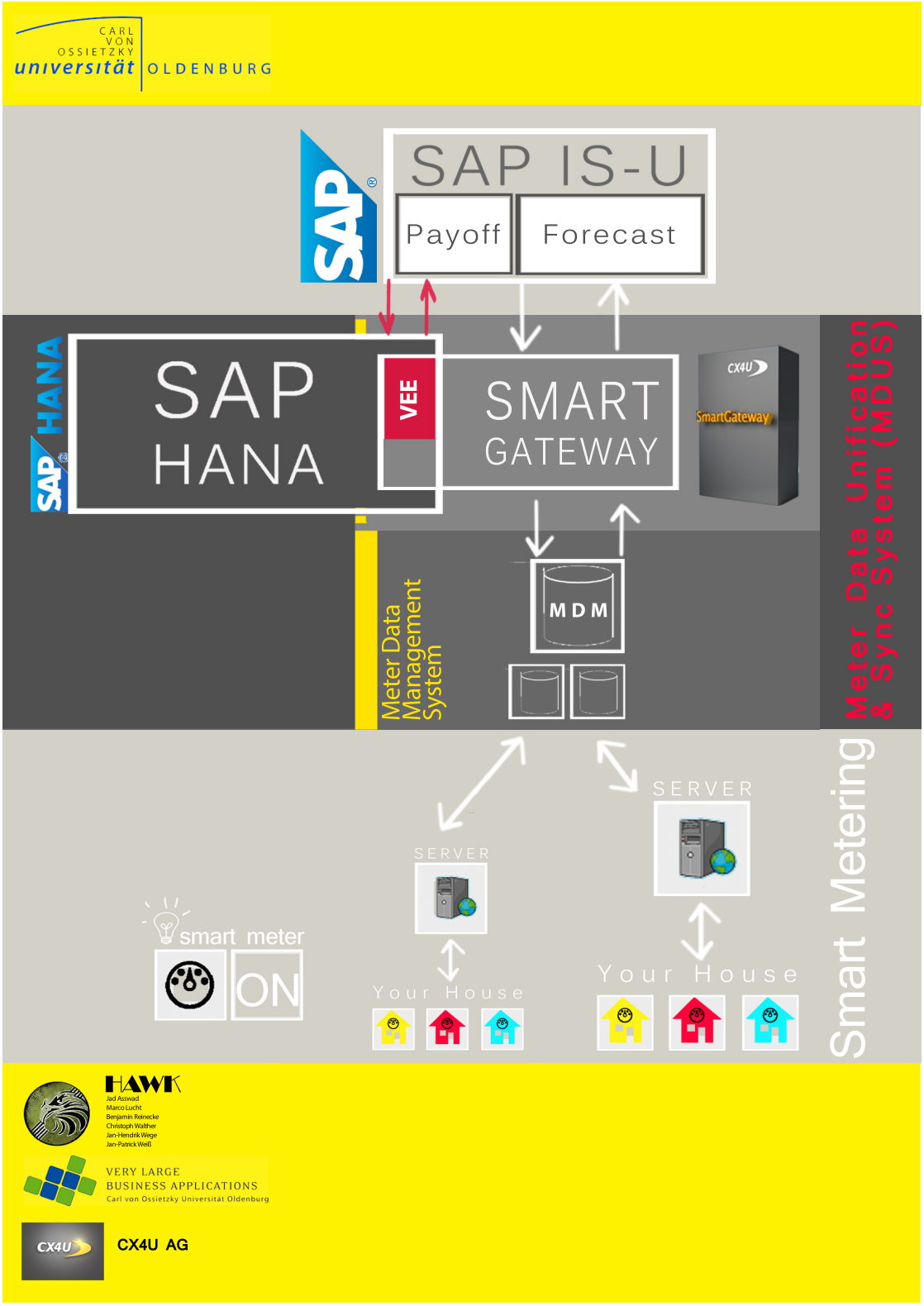


HAWK
Jad Asswad
Marco Lucht
Benjamin Reinecke
Christoph Walther
Jan-Hendrik Wege
Jan-Patrick Weiß

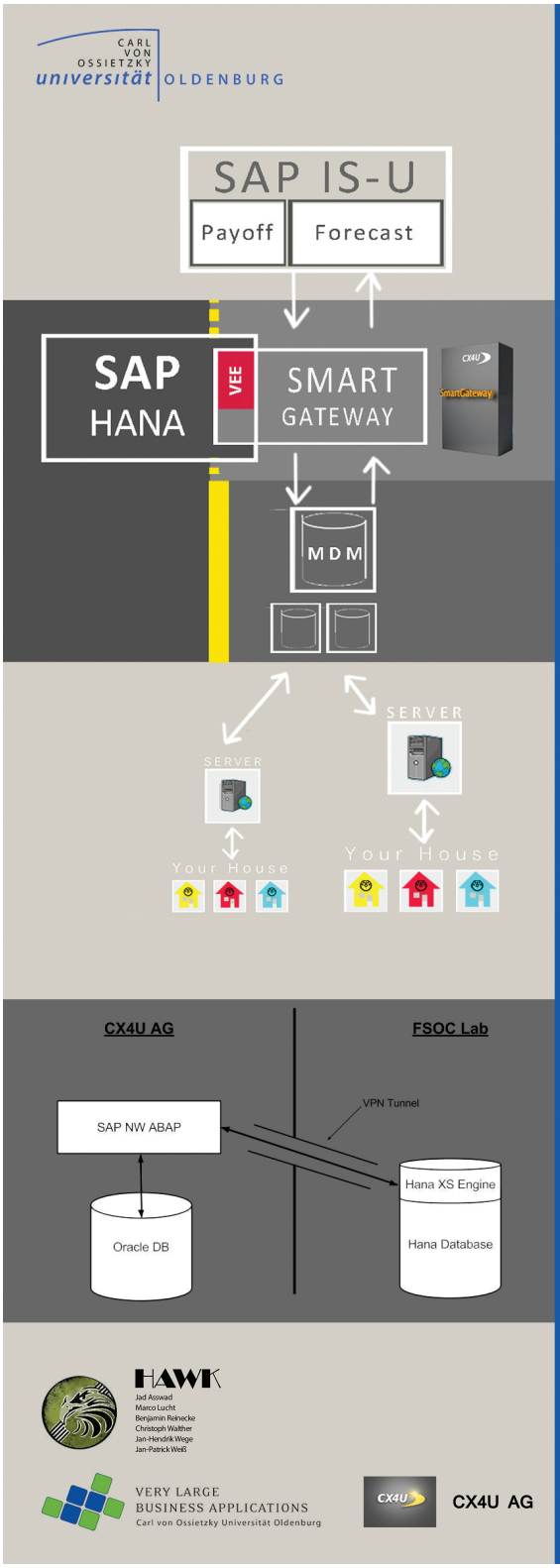
VERY LARGE BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

CX4U AG

F.5.4 FSOC-Poster-2 2013



F.5.5 FSOC-Poster 2014



Integration of a VEE-Framework into SAP HANA

Introduction

The energy market is changing by legislation so by 2020 smart meters will be installed in most households. Therefore the IT systems of energy companies have to adjust to deal with the growing amount of measured data. In order to achieve the best use of the near real time data consumption by energy providers, the data should be permanently consistent, up-to-date and free of faulty values. The project group in cooperation with CX4U Consulting was able to ensure the integrity and the best quality of data through implementing a VEE-Framework (Validation, Estimation, Editing) into SAP HANA to benefit from the efficiency of in-memory-database technology.

VEE (Validation, Estimation, Editing)

- **Validation:** Examination of data in order to determine inconsistencies, e.g. missing values, duplicates or spikes. Depending on the test results, the individual values need to be marked as failed and subsequently further processed.
- **Estimation:** Adapting and correcting all previously marked values with several appropriate replacement values, which are generated by choosing a specific set of predefined energy provider-based methods, in order to correspond to their specific reuirements.
- **Editing:** Examination of the correctness of timestamps and correcting them according to the predefined interval. Moreover adjusting the reading value according to the corrected timestamp.

Implementation

After evaluating different implementation alternatives, the SAP HANA Extended Application Services (XS-Engine) have been chosen as the suitable solution for this particular project. Among all other alternatives, the XS-Engine fulfils the needed functionalities and provides a real-time performance without any extra expenses or products.

The first step was to create an OData service on the HANA side, so that the Netweaver side could insert data into and get data from the HANA database. At the same time, a functionality on the Netweaver is developed in order to read the OData interface. Afterwards the data had to be processed within the HANA system and therefore several scripts are implemented, which offered the needed VEE-functionality. In addition to the scripts which execute the VEE functions, a web interface had to be developed for managing and calling those implemented functions. The web interface is programmed using HTML and Javascript. Finally it was possible to request data from the OData interface into the Netweaver using a HTTP request. Moreover, the VEE algorithms were able to be used to process the data in the HANA database.

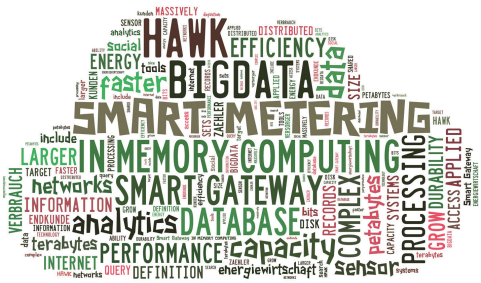
F.6 CeBit

Für die CeBit wurden einen Standbroschüre, eine Grafiktafel, sowie ein Flyer angefertigt, die im Folgenden zu finden sind.

F.6.1 CeBIT Standbroschüre

CARL VON OSSIETZKY universität OLDENBURG

Department für Informatik,
Abteilung Wirtschaftsinformatik /VLBA



Reichweite von Big Data

Smart Metering - Chancen für den Alltag

Big Data – transparente Stromverbräuche durch effiziente Datenverarbeitung

Die Energiewende ist eine große Herausforderung. Technische Entwicklungen wie intelligente Stromzähler können sowohl Energieversorger als auch -verbraucher dabei unterstützen. Diese Smart Meter messen alle drei Sekunden einen Wert und übermitteln ihn an die Energieversorger. Der Trend geht in Richtung Echtzeitverarbeitung. Die Smart Meter ermöglichen eine ressourcensparende Energieverwaltung und transparente Stromabrechnungen. Anstatt nur einen Jahreswert für die Verbrauchsberechnung zugrunde zu legen, lassen sich die tatsächlichen Kosten transparent verfolgen. Der Energiebedarf kann genauer vorhergesagt werden. Energienutzung und -speicherung lassen sich besser planen und umfassendere Analyseverfahren werden ermöglicht.

Um die damit verbundenen steigenden Datenmengen verarbeiten zu können, müssen jedoch zunächst die IT-Systeme der Energieversorger angepasst werden. Das Department für Informatik der Universität Oldenburg hat ein VEE-Framework (Validation, Estimation und Editing) aufgestellt, das die anfallenden großen Datenmengen verarbeitet und die Integrität sicherstellt. Das Framework bildet Ersatzwerte, um fehlerhafte Messwerte zu korrigieren. Die Analysen erfolgen in einer In-Memory Datenbank wie beispielsweise SAP HANA, die schneller rechnet als herkömmliche Datenbanken.

Big Data – Transparent Power Consumption with SAP HANA

The energy market is changing by legislation, so by 2020 smart meters will be installed in most households. Therefore, the IT systems of energy companies have to adjust in order to deal with the growing amount of measured data. Almost real time data consumption allows companies to create new business use cases. For example they can predict the usage of their grids and thus close short time contracts with suppliers to meet the demands. To ensure the reliability and quality of these data, the Department of Computer Science at the Oldenburg University applies a VEE-Framework (Validation, Estimation and Editing). The VEE-Framework is implemented within an in-memory-database to benefit from their efficiency.

Carl von Ossietzky Universität Oldenburg
Department für Informatik
Abteilung Wirtschaftsinformatik /
Very Large Business Applications
Ansprechpartner: Prof. Dr.-Ing. Jorge Marx Gómez
Ammerländer Heerstraße 114-118, D-26129 Oldenburg
Tel. +49 (0)441 798 4470, Fax +49 (0)441 798 4472
jorge.marx.gomez@uni-oldenburg.de
http://vlba.wi-ol.de/

11

F.6.2 CeBIT Grafiktafel

CARL VON OSSIETZKY universität OLDENBURG

Department für Informatik, Abteilung Wirtschaftsinformatik I / VLBA

Smart Meter – Big Data

- Smart Meter übermitteln zeitnah den Energieverbrauch aus den Haushalten an Energieversorger.
- Planungen der Energieversorgung durch den steigenden Anteil der regenerativen Energien benötigen eine aktuelle Datenbasis.
- Transparente Strompreise in Echtzeit erfordern passende Softwareunterstützung.

Smart Meter – Big Data

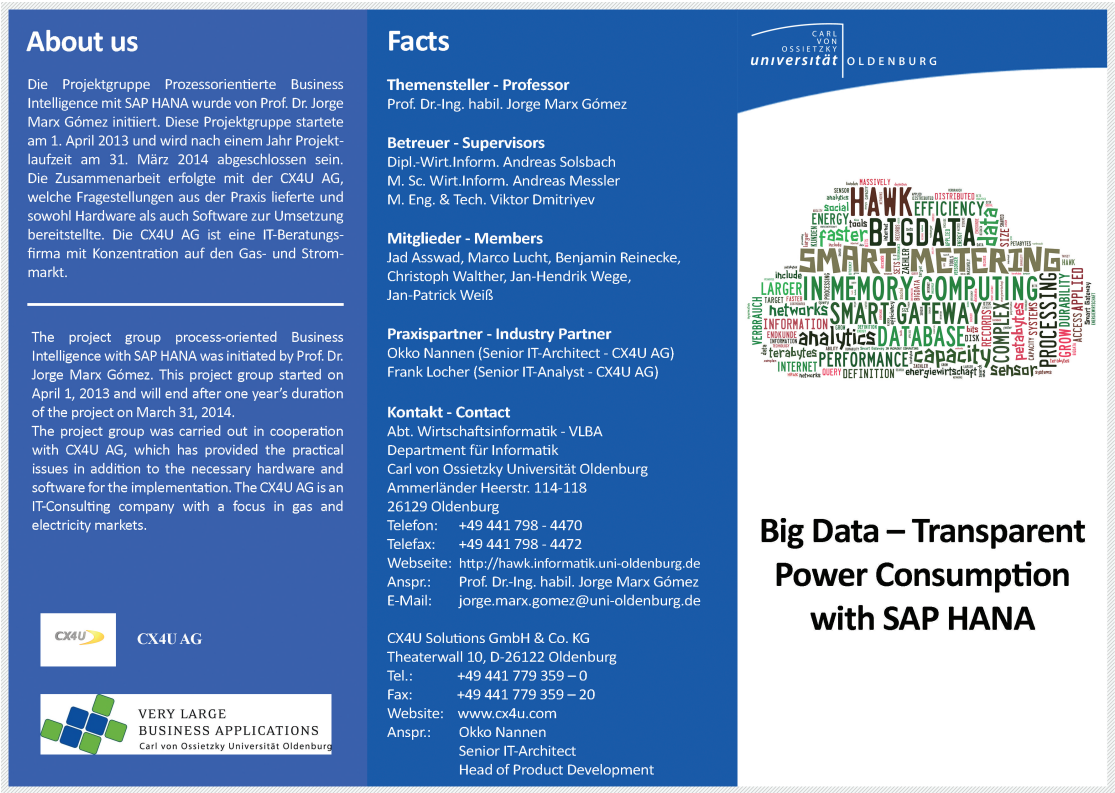
- Smart meters transmit energy consumption of households to the energy provider.
- Due to the increasing share of renewable energies, planning of energy supply requires up-to-date databases.
- Transparent electricity prices provided in real-time require suitable software support.

Niedersachsen

Sie kennen unsere Pferde. Erleben Sie unsere Stärken.

11

F.6.3 CeBIT Flyer



About us

Die Projektgruppe Prozessorientierte Business Intelligence mit SAP HANA wurde von Prof. Dr. Jorge Marx Gómez initiiert. Diese Projektgruppe startete am 1. April 2013 und wird nach einem Jahr Projektlaufzeit am 31. März 2014 abgeschlossen sein. Die Zusammenarbeit erfolgte mit der CX4U AG, welche Fragestellungen aus der Praxis lieferte und sowohl Hardware als auch Software zur Umsetzung bereitstellte. Die CX4U AG ist eine IT-Beratungsfirma mit Konzentration auf den Gas- und Strommarkt.

The project group process-oriented Business Intelligence with SAP HANA was initiated by Prof. Dr. Jorge Marx Gómez. This project group started on April 1, 2013 and will end after one year's duration of the project on March 31, 2014.
The project group was carried out in cooperation with CX4U AG, which has provided the practical issues in addition to the necessary hardware and software for the implementation. The CX4U AG is an IT-Consulting company with a focus in gas and electricity markets.

CX4U AG

VERY LARGE BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Facts

Themensteller - Professor
Prof. Dr.-Ing. habil. Jorge Marx Gómez

Betreuer - Supervisors
Dipl.-Wirt.Inform. Andreas Solsbach
M. Sc. Wirt.Inform. Andreas Messler
M. Eng. & Tech. Viktor Dmitriyev

Mitglieder - Members
Jad Asswad, Marco Lucht, Benjamin Reinecke, Christoph Walther, Jan-Hendrik Wege, Jan-Patrick Weiß

Praxispartner - Industry Partner
Okko Nannen (Senior IT-Architect - CX4U AG)
Frank Locher (Senior IT-Analyst - CX4U AG)

Kontakt - Contact
Abt. Wirtschaftsinformatik - VLBA
Department für Informatik
Carl von Ossietzky Universität Oldenburg
Ammerländer Heerstr. 114-118
26129 Oldenburg
Telefon: +49 441 798 - 4470
Telefax: +49 441 798 - 4472
Webseite: http://hawk.informatik.uni-oldenburg.de
Anspr.: Prof. Dr.-Ing. habil. Jorge Marx Gómez
E-Mail: jorge.marx.gomez@uni-oldenburg.de

CX4U Solutions GmbH & Co. KG
Theaterwall 10, D-26122 Oldenburg
Tel.: +49 441 779 359 – 0
Fax: +49 441 779 359 – 20
Website: www.cx4u.com
Anspr.: Okko Nannen
Senior IT-Architect
Head of Product Development

CARL VON OSSIETZKY universität OLDENBURG

Big Data – Transparent Power Consumption with SAP HANA

The Project

Integration eines VEE-Frameworks innerhalb eines Smart Gateways in SAP HANA

Durch die Gesetzgebung ändert sich der Energiemarkt bis 2020, sodass in den meisten Haushalten Smart Meter installiert sein werden. Deshalb müssen die IT-Systeme der Energieunternehmen angepasst werden, um mit der wachsenden Menge an Messdaten umgehen zu können. Damit die Echtzeitdatenerfassung der Energielieferanten am besten eingesetzt werden kann, müssen die Daten dauerhaft konsistent, aktuell und fehlerfrei sein.

Die Projektgruppe hat es, in Zusammenarbeit mit der CX4U AG, geschafft, die Integrität und Qualität der Daten zu gewährleisten, indem ein VEE-Framework (Validation, Estimation, Editing) in SAP HANA implementiert und somit von den In-Memory Datenbanktechnologie profitiert wurde.

Integration of a VEE-Framework within a Smart Gateway into SAP HANA

The energy market is changing by legislation so by 2020 smart meters will be installed in most households. Therefore the IT systems of energy companies have to adjust to deal with the growing amount of measured data. In order to achieve the best use of the near real time data consumption by energy providers, the data should be permanently consistent, up-to-date and free of faulty values. The project group in cooperation with CX4U Consulting was able to ensure the integrity and the best quality of data through implementing a VEE-Framework (Validation, Estimation, Editing) into SAP HANA to benefit from the efficiency of in-memory-database technology.

VEE-Framework

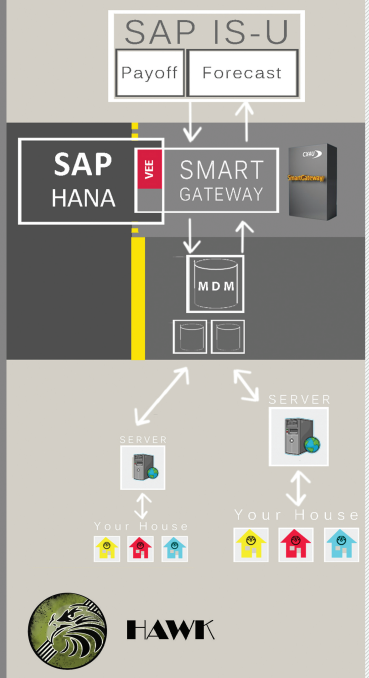
VEE (Validation, Estimation, Editing)

- **Validation:** Die Prüfung der Daten, um Inkonsistenzen zu bestimmen, z. B. fehlende Werte, Duplikate oder falsche Lastspitzen. In Abhängigkeit von den Testergebnissen, müssen die einzelnen Werte als fehlerhaft gekennzeichnet und anschließend weiter verarbeitet werden.
- **Estimation:** Anpassung und Korrektur aller zuvor gekennzeichneten Werte durch entsprechende Ersatzwerte, welche mit Hilfe von vordefinierten Methoden des Energieversorgers errechnet werden.
- **Editing:** Prüfung und gegebenenfalls Korrektur der Zeitstempel und der Intervallmessung.

VEE (Validation, Estimation, Editing)

- **Validation:** Examination of data in order to determine inconsistencies, e.g. missing values, duplicates or spikes. Depending on the test results, the individual values need to be marked as failed and subsequently further processed.
- **Estimation:** Adapting and correcting all previously marked values with several appropriate replacement values, which are generated by choosing a specific set of predefined energy provider-based methods, in order to correspond to their specific reuirements.
- **Editing:** Examination of the correctness of time stamps and correcting them according to the predefined interval. Moreover adjusting the reading value according to the corrected time stamp.

Architecture



F.7 OpenVPN-Anleitung vom HPI

Future SOC Lab Network

Access to the Future SOC Lab is provided through a Virtual Private Network (VPN) connection.

Technical Prerequisites

- Internet Connection
- Capability to send and receive **UDP** (User Datagram Protocol) Packets
- Administrative rights in the Operating System

Software

The VPN connection is provided by the OpenVPN Server software. To establish an OpenVPN connection to the Future SOC Lab an OpenVPN client for your Operating System is required.

Supported operating systems are Windows 2000/XP and higher, OpenBSD, FreeBSD, NetBSD, Linux 2.4 and higher, Mac OS X, and Solaris.

Windows Clients

- OpenVPN GUI (http://openvpn.net/index.php/open-source/downloads.html)
- OpenVPN Portable (http://sourceforge.net/projects/ovpnp/)

MacOS Clients

- Tunnelblick (http://code.google.com/p/tunnelblick/)
- Viscosity (http://www.thesparklabs.com/viscosity/)

Other OSes

- Source-Code available (http://openvpn.net/index.php/open-source/downloads.html)

OpenVPN Configuration

The Future SOC Lab Team provides you a certificate file, a shared secret and an OpenVPN client configuration file.

- `ca.crt` *Root Certificate Authority Certificate file (Root CA)*
- `fsoc-ta.key` *Shared Secret file*
- `FSOC-Lab.ovpn` *OpenVPN Connection Configuration file*
- `fsoc-lab-account.txt` *Future SOC Lab username and initial password*

The configuration file has to be placed on different locations depending on OpenVPN client and operating system.

Certificates

The configuration file `FSOC-Lab.ovpn` contains locations to the certificate files.

Usage (Windows)

Start the OpenVPN GUI application. Windows Vista and Windows 7 may require to execute the OpenVPN client with administrative rights (see Figure 1). In Windows operating systems you have to right click the OpenVPN tray icon and select **Connect** (for the FSOC-Lab entry, if you have multiple configurations).

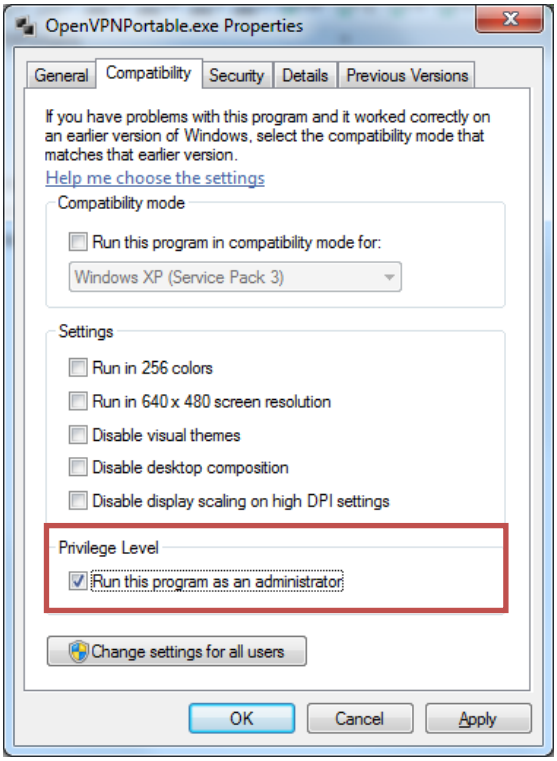


Figure 1: Configure Administrative rights for the OpenVPN client.

A successful connection is indicated by a green colored OpenVPN tray icon. To test the routing table entries please open a command prompt (Windows+R: type cmd.exe) and execute

```
> ping 192.168.42.1
```

You should get similar result as shown in Figure 2.

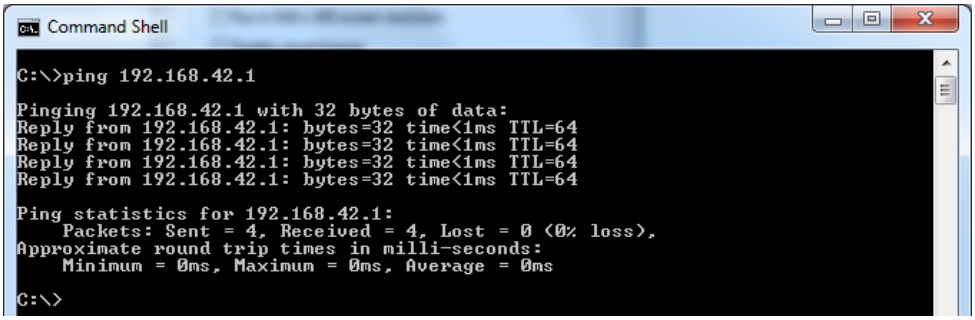


Figure 2: Ping output of successfully connected network

Usage (MacOS X)

For the Apple MacOS X operating system series, we recommend the Tunnelblick OpenVPN client. After installing the software you have to copy the provided certificates and OpenVPN configuration file in to

the folder: <**your user profile>/Library/Application Support/Tunnelblick/Configurations** (see Figure 3)

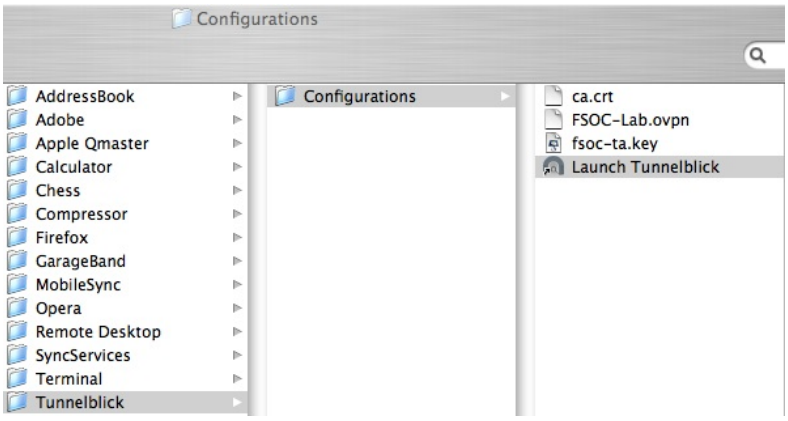


Figure 3: Tunnelblick configuration folder

Make sure you have started the Tunnelblick client. A tunnel icon

is displayed in the upper right of the menu bar (see Figure 5).

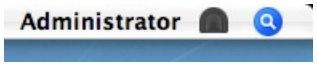


Figure 4: Tunnelblick Icon disconnected

After hitting the Connect entry in the Tunnelblick icon menu, the icon should change to the following:

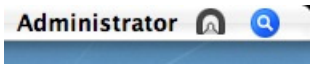


Figure 5: Tunnelblick icon connected

To prove the routing table entries, please open a Terminal and execute the following command:

```
> ping 192.168.42.1
```

The output should be similar to the output in Figure 2.

Usage (Linux, BSD, ...)

Install the OpenVPN package from your Linux distribution repository or build the source package. Copy the configuration file and the certificates to your home directory and execute the openvpn program to establish a connection:

```
> openvpn --config FSOC.ovpn
```

There may be graphical frontends for the openvpn command line program depending on your distribution. You may have to convert the line endings of the FSOC-Lab.ovpn to line feed only, to be accepted by the openvpn program.

F.8 Latex-Wiki

LaTeX Wiki Home

In diesem Bereich findet ihr Arbeitshinweise zu LaTeX und Links zu der benötigten Software. Die Inahlte werden mit den Fragen und Problemen der PG während der Bearbeitung der Berichte anwachsen.

Die LaTeX-Vorlage für den Endbericht ist derart gestaltet, dass Inhalte und Formatierungen strickt getrennt sind. So ist sichergestellt, dass auch bei nachträglichen Änderungen am Header (der Stelle, wo die Dokumenteneigenschaften definiert werden) die Inhalte unangetastet bleiben. Mittels den LaTeX-Funktionen zum Aufteilen des Dokuments für das parallele Arbeiten mehrer Benuzter ist sichergestellt, dass keine oder möglichst wenig Konflikte auftreten.

Es bietet sich an, dass niemals mehr als ein Benutzer in einer .tex-Datei arbeitet. Somit sollte für einzelne Kapitel/Unterkapitel eine eigene Datei erstellt werden. Um eine gewisse Übersichtlichkeit zu gewährleisten, können diese Dateien in Ordnern, benannt nach den Gruppenmitgliedern, abgespeichert werden.

Im Folgenden finden sich Informationen zu den Bereichen

Arbeiten mit JabRef

Arbeiten mit Texmaker

Code-Konventionen

Dos and Don'ts

Downloads

Recent space activity

- **Benjamin Reinecke**
  - Code-Konventionen Mrz 17, 2014 aktualisiert • Änderung anzeigen
  - Dos and Don'ts Feb 14, 2014 aktualisiert • Änderung anzeigen
  - Arbeiten mit Texmaker Feb 14, 2014 aktualisiert • Änderung anzeigen
  - Arbeiten mit JabRef Jan 30, 2014 aktualisiert • Änderung anzeigen
  - LaTeX Wiki Home Okt 17, 2013 aktualisiert • Änderung anzeigen

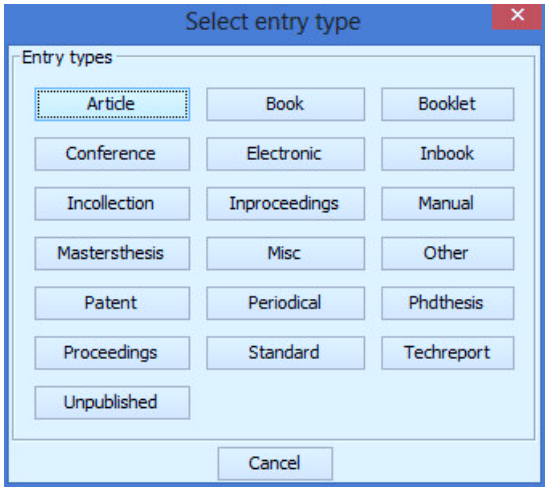
Space contributors

- Benjamin Reinecke (vor 9 Tagen)

Arbeiten mit JabRef

Die BibTeX-Dateien werden in JabRef als Database bezeichnet. Über das Menü ist es sowohl möglich eine neue Datenbank zu erstellen, oder eine vorhandene .bib Datei zu öffnen.

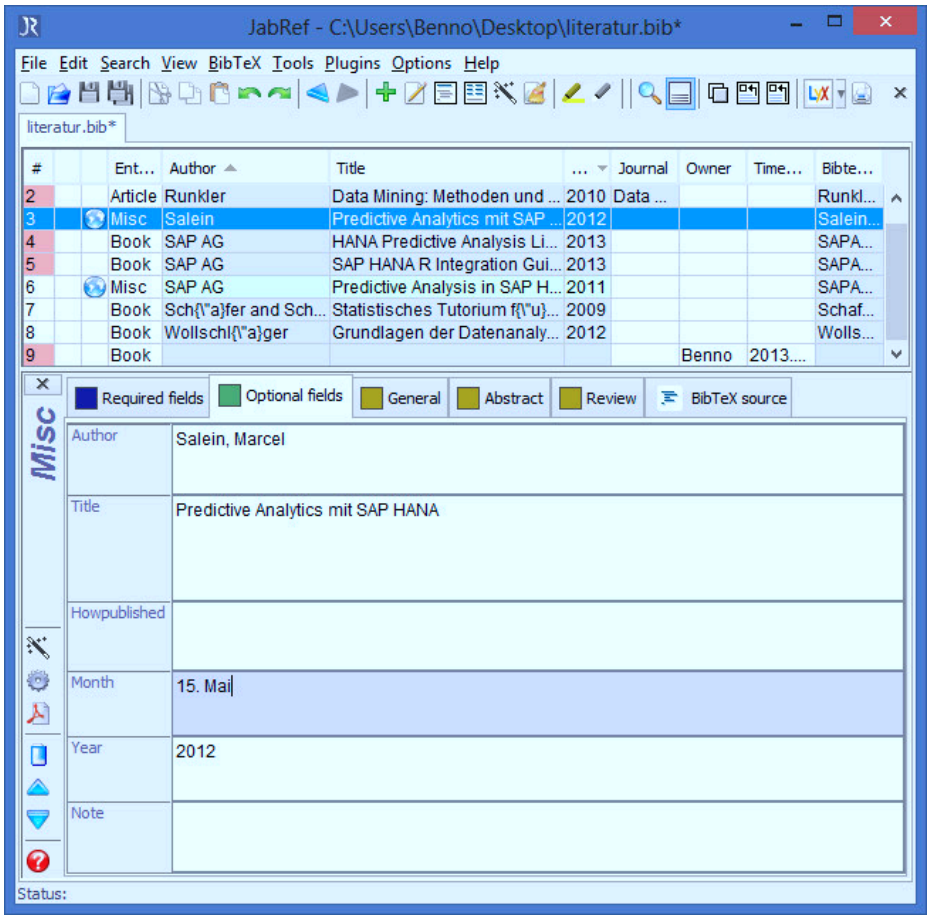
Über das "+"-Symbol werden Einträge zur Datenbank hinzugefügt. Das Programm fragt zunächst, um was für eine Quelle es sich handelt.



Oft wird hier Book oder Article gewählt werden. Eine Internetquelle wird am einfachsten über den Typ Misc angegeben.

Als nächstes müssen die Pflichtfelder gefüllt werden. Bei einem Buch wären dies mindestens Titel, Publisher, Jahr und Autor. Der Bibtexkey muss in jedem Fall und bei jedem Typ angegeben werden: hierüber wird die Quelle im Dokument referenziert. Als Form bietet sich hier *Nach nameDesAutors.Erscheinungsjahr*, also beispielsweise *Runkler.2012* an. Der Bibtexkey muss eindeutig sein und darf nur einmal vergeben werden!

Bei einer Internetquelle ist nur der Bibtexkey das Pflichtfeld, in den optionalen Feldern wird Autor, Titel sowie Monat Jahr des Dokuments angegeben. Im Feld Monat bitte auch den Tag angeben (siehe Abbildung)! Unter dem Reiter General gibt es ein Feld für die URL.



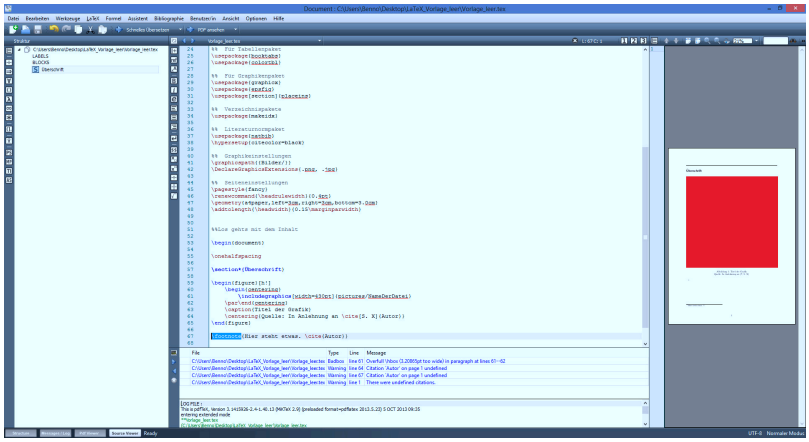
Bei Bibtex-Dateien handelt es sich um Text-Files, die mit einem Editor wie beispielsweise Notepad++ bearbeitet werden könnnen. Dieses manuelle Bearbeiten ist die einfachste Möglichkeit, mehrere .bib-Dateien zu einer zusammenzufügen.

Die Bibtex-Dateien sind ausschließlich mit Jabref oder einem vergleichbaren Editor zu erstellen und zu bearbeiten! So wird verhindert, dass es aufgrund einer Mehrzahl von Datenbanken zu vermeidbaren Fehlern kommt. Im SVN findet sich eine Beispiel-Bibtex-Datei, die zeigt welche Informationen in welcher Form bei den unterschiedlichen Literaturarten anzugeben sind.

Arbeiten mit Texmaker

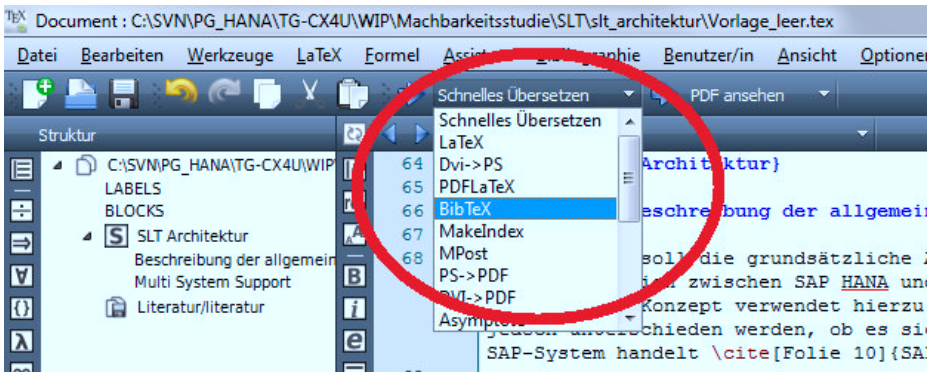
Die Arbeit mit Texmaker gestaltet sich recht intuitiv. Auf der linken Seite sieht man einer Übersicht aller Dateien, die mittels des /include-Befehls mit dem aktuellen Dokument verbunden sind. Im mittleren Bereich befindet sich der Arbeitsbereich. LaTeX-Befehle werden in rot (Formatierungen), bzw. blau (Befehle den Inhalt betreffend) dargestellt. Die rechte Seite bietet nach dem Übersetzen/Klick auf PDF ansehen eine Vorschau auf das Dokument.

In der Menüleiste am oberen Rand befindet sich der Menüpunkt "LaTeX". Über diesen können häufig verwendete Befehle mit zwei bis drei Klicks in das Dokument eingefügt werden.



Ein häufig vorkommendes Problem bei der Arbeit mit Texmaker ist, dass Quellen nicht im Dokument angezeigt werden. An der Stelle, an der mit \cite eine Quellenangabe gemacht werden soll, erscheint nur ein Fragezeichen. Ein Grund hierfür ist in der Regel das Hinzufügen einer neuen Quelle in der BibTex-Datei oder das ändern einer bereits eingetragenen Quelle.

In einem solchen Fall ist es notwendig vor dem "Schnellen Übersetzen" im gleichen Dropdown-Menü "BibTex" auszuwählen, und das Dokument mit dieser Einstellung einmal zu kompilieren. Im Anschluss muss dann ein zweimaliges "Schnelles Übersetzen" durchgeführt werden. Sollten nun weiterhin Anzeigefehler oder fehlende Quellen im Dokument zu finden sein, muss Texmaker einmal neu gestartet werden und im Anschluss ein erneutes "Schnelles Übersetzen" durchgeführt werden. Auch das Löschen allen Dateien **außer der .tex-Files** hilft oft bei solchen Fehlern.



Code-Konventionen

- Label und Referenzen
- Aufzählungen
- Grafiken
- Tabellen
- \hline \hline \endhead
- Zitate und Fußnoten
- Silbentrennung
- Paragraph und Subparagraph

Label und Referenzen

Um Kapitel, Grafiken oder Tabellen referenzieren zu können, müssen die LaTeX-Funktionen **label{name}** (zum Setzen einer Sprungmarke) und **ref{name}** (zum Verweis) verwendet werden.

Zur Bennennung soll das folgende Schema verwendet werden:

- Grafiken: **label{fig_name}**
- Tabellen: **label{tab_name}**
- Kapitel: **label{kap_name}**

Die folgenden Label für Kapitel wurden von mir bereits angelegt:

1 kap_einleitung
1.1 kap_motivation
1.2 kap_problem
1.3 kap_aufbau

2 kap_rahmen
2.1 kap_orga
2.1.1 kap_konzept
2.1.2 kap_rollen
2.1.3 kap_vorbereitung
2.2 kap_technik
2.2.1 kap_confluence
2.2.2 kap_latex
2.2.3 kap_homepage
2.2.4 kap_redmine
2.2.5 kap_svn

3 kap_bi
3.0 kap_bi_def
3.1 kap_bi_orga
3.2 kap_bi_architektur
3.3 kap_bi_markt
3.3.1 kap_bi_probeinleitung
3.3.2 kap_bi_anbieter
3.3.3 kap_bi_marktergebnis
3.3.4 kap_bi_marktfazit

4 kap_cx4u
4.1 kap_cx4u_def
kap_cx4u_partner
kap_cx4u_aufgabe
kap_cx4u_orga
kap_cx4u_struktur
kap_cx4u_plan
kap_cx4u_verlauf
4.2 kap_cx4u_analyse
4.3 kap_cx4u_machbarkeit
4.4 kap_cx4u_oeffentlichkeit
4.5 kap_cx4u_schluss

5 kap_ergebnisse

6 kap_anhang

Aufzählungen

Bei Aufzählungen bietet es sich an, diese mit

\begin{itemize}\itemsep0pt

einzuleiten. Hiermit wird der Abstand zwischen den einzelnen Punkten der Aufzählung verringert.

Grafiken

Um eine Abbildung in ein LaTeX-Dokument einzubinden, empfiehlt sich die Verwendung des folgendes Codes. Die Grafik wird aufgrund des "h!" an exakt der Stelle eingefügt, an der auch der Code im Dokument steht. Sowohl Name als auch Quelle der Grafik stehen unter der Abbildung. Es ist zu empfehlen den Dateinamen der Grafik ohne Leerzeichen zu wählen, da dies ansonsten zu unerwünschten Begleiterscheinungen (doppeltes Anzeigen des Dateinamens über der Grafik o.ä.) führen kann. Die Angabe der Dateiendung ist bei jpg und png nicht nötig.

```
\begin{figure}[h!]
 \begin{centering}
 \includegraphics[width=430pt]{pictures/NameDerDatei}
 \par\end{centering}
 \caption{Titel der Grafik}
 \centering{Quelle: In Anlehnung an \citet[S. X]{NAME.JAHR}}
\end{figure}
```

Tabellen

Für Tabellen gilt für Position, Bezeichnung und Quelle die gleiche Konvention, wie für die Grafiken. Es ergibt sich:

```
\begin {table}[ h! ]
\centering
\begin{tabular}{|l|l|l|}\hline
Author & Title & Year \\\hline\hline
Philip K. Dick & Minority Report & 1956 \\\hline
Philip K. Dick & Do Androids Dream of Electric Sheep? & 1968 \\\hline
Philip K. Dick & A Scanner Darkly & 1977 \\\hline
Neal Stephenson & Snow Crash & 1992 \\\hline
Neal Stephenson & The Diamond Age & 1995 \\\hline
Neal Stephenson & Cryptonomicon & 1999 \\\hline
\end{tabular}
\vspace*{.2cm}
\caption {Titel der Tabelle}
\centering{Quelle: Eigene Darstellung}
\end{table}
```

Neben dem Verwenden von "table" ist auch "longtable" für Tabellen über mehrere Seiten möglich. Hier ein Beispiel (für die Verwendung muss \usepackage{longtable} im LaTeX-Vorspann definiert sein):

```
\begin {longtable}{|l|r|} %%in diesem Fall: 1. Spalte linksbündig und 2. Spalte rechtsbündig

\hline

\multicolumn{2}{|c|}{Dies erscheint zentriert auf jeder Seite}  \\

\hline \hline \endhead

Inhalt 1. Spalte & Inhalt 2. Spalte  \\

Inhalt 1. Spalte &Inhalt 2. Spalte  \\

Inhalt 1. Spalte &Inhalt 2. Spalte  \\

\hline

\caption {Titel der langen Tabelle}

\end {longtable}

\vspace*{-20pt}{\par\centering Quelle: Eigene Darstellung\par}
```

Zitate und Fußnoten

Zitate sind direkt oder indirekt übernommene Gedanken und werden nach der Harvard-Methode im Text kenntlich gemacht. Hierzu den folgenden Code verweden:

```
\cite[S. X]{NAME.JAHR}
```

Im Gegensatz hierzu kennzeichnet eine Fußnote weitergehende Informationen oder Erläuterungen:

```
\footnote{Was hier steht beschreibt beispielsweise einen Sachverhalt der im Text erwähnt, aber nicht weiter verfolgt wird. Auch in Fußnoten können mit dem cite-Befehl Quellen angegben werden.}
```

Zu beachten ist, dass ein Dokument nach dem erstmaligen Zitieren eines Autors in der Regel zwei mal kompiliert werden muss, damit nicht nur ein Fragezeichen an der betreffenden Stelle angezeigt wird.

Anführungszeichen sind mittels **"`** und **"'** zu erzeugen. Die Variante mit **\glqq** und **\grqq** kann zu Fehlern, wie fehlenden Leerzeichen führen.

Achtung: Bei Grafiken und Tabellen werden Quellen mit **\citet** anstett mit \cite eingeleitet! Dies hat zur Folge, dass die Setzung der Klammern an diesen Stellen korrekt ist (der Autorenname steht nicht mit in der Klammer).

Silbentrennung

Für eine optimierte Silbentrennung habe ich das Paket hyphsubst eingefügt, dieses soll besonders bei zusammengesetzten Wörtern gut arbeiten und zumindest das über die Zeile hinaus schreiben beenden.

In manchen Fällen kann eine manuelle Silbentrennung mittels **\-** (ein Wort NUR an dieser Stelle trennen) oder auch **"-** (eine zusätzliche Trennstelle) aber weiterhin notwendig sein.

Paragraph und Subparagraph

Um einen Zeilenumbruch in einem Paragraph und Subparagraph hinzubekommen, ist es am einfachsten einen Workaround anzuwenden: nach setzen von \paragraph{} bzw. \subparagraph{} einfach ein **$\;$** einfügen. Nach dem Befehl muss im Quelltext der tex-Datei eine Leerzeile folgen.

Dos and Don'ts

Dos	Don'ts
Wenn nicht erklärbare Phänomene auftreten, lösche, bis auf die .tex Dateien und eventuelle Grafiken, alle Dateien eines Ordners und kompiliere das Dokument erneut.	Arbeite nicht mit mehr als einer Person in einer .tex Datei, sondern teile das Dokument auf und füge es über den \include bzw. \input Befehl zusammen
Kombiniere das Dokument nach dem Hinzufügen von Quellen zwei Mal, bzw. einmal zusätzlich mit der BibTeX-Option (vgl. Arbeiten mit Texmaker)	

Downloads

Texmaker

Bei Texmaker handelt es sich um einen LaTeX-Editor, der neben Windows auch für Mac und Linux verfügbar ist. Neben einer Syntaxhervorhebung bietet er über eine Menüleiste Zugriff auf oft verwendete Funktionen.

Den Download findet ihr unter http://www.xm1math.net/texmaker/download.html

MiKTeX

Zur Verwendung von Texmaker wird eine so genannte LaTeX-Distribution benötigt. Diese ist das Basis-LaTexsystem, auf dem Texmaker als grafische Oberfläche aussetzt. Ein typischer Vertreter für Windows-Betriebssysteme ist MiKTeX. Bei dieser Distribution ist auch der schon bekannte Editor Texworks enthalten.

Den Download gibt es unter http://miktex.org/download

MacTeX

Da MiKTeX unter MacOS nicht funktioniert, wird hier eine alternative Distribution, wie das weit verbreitete MacTex, benötigt.

Download unter http://www.tug.org/mactex/index.html

JabRef

Unter LaTeX wird für die Literaturveraltung das BibTeX-Format verwendet. Bei JabRef handelt es sich um eine einfache Software für alle gängigen Betriebssysteme (Java-Basis) zur grafischen Bearbeitung der Bibliotheks-Dateien.

Den Download findet ihr unter http://jabref.sourceforge.net

Literatur

[Accantec 2013] ACCANTEC: *SAS Business Analytics*. http://www.accantec.de/software/sas. Version: 2013. – Abgerufen am 15.12.2013

[Alterauge 2013] ALTERAUGE, Markus: *Einsatzszenarien für SAP HANA*. http://www.mayato.com/images/stories/downloads/mayato_whitepaper_einsatzszenarien_sap_hana_final.pdf. Version: 2013. – Abgerufen am 19.10.2013

[Apache 2013a] APACHE: *Distributions and Commercial Support*. http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support. Version: 2013. – Abgerufen am 09.12.2013

[Apache 2013b] APACHE: *Welcome to Apache Hadoop!* http://hadoop.apache.org/. Version: 2013. – Abgerufen am 11.12.2013

[Bachmann u. Kemper 2011] BACHMANN, Ronald ; KEMPER, Guido: *Raus aus der BI-Falle: Wie Business Intelligence zum Erfolg wird*. 2. Aufl. Heidelberg and München and Landsberg and Frechen and Hamburg : mitp, 2011

[Balzert 1998] BALZERT, Helmut: *Lehrbuch der Softwaretechnik: Software-Management, Software-Qualitätssicherung, Unternehmensmodellierung*. Heidelberg, Berlin : Spektrum Akademischer Verlag, 1998

[Bange u. Grosser 2012] BANGE, Carsten ; GROSSER, Timm: *BIG DATA – Business Intelligence der nächsten Generation*. http://www.computerwoche.de/a/big-data-bi-der-naechsten-generation,2505617. Version: 2012. – Abgerufen am 09.12.2013

[BARC 2012] BARC: *BARC-Guide Business Intelligence 2012/13: Der Branchenkatalog von den Marktexperten*. Hagen : B-Eye-Media GmbH, 2012

[BARC 2013] BARC: *BARC-Guide Business Intelligence 2013/14: Der Branchenkatalog der Marktexperten*. Hagen : B-Eye-Media GmbH, 2013

[BITKOM 2012] BITKOM ; BITKOM (Hrsg.): *Big Data im Praxiseinsatz: Szenarien, Beispiele, Effekte*. http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online%281%29.pdf. Version: 2012. – Abgerufen am 29.05.2013

[Bitkom 2013] Abgerufen am 21.10.2013

[Bleiholder u. Schmid 2008] BLEIHOLDER, Jens ; SCHMID, Joachim: Datenintegration und Deduplizierung. In: HILDEBRAND, Knut (Hrsg.): *Daten- und Informationsqualität*. Wiesbaden : Vieweg + Teubner, 2008 (Praxis)

[Boles 2009] BOLES, Dietrich: *Leitfaden zur Durchfuehrung von Projektgruppen*. http://www-is.informatik.uni-oldenburg.de/~dibo/pg_fb10/Leitfaden-PG.pdf. Version: Mai 2009. – Abgerufen am 08.12.2014

[Braincourt 2014] BRAINCOURT: *IBM Cognos TM1*. http://www.braincourt.com/IBM-Cognos-TM1.126.0.html. Version: 2014. – Abgerufen am 11.03.2014

[Bretting] BRETTING, Ralf: Big Data Automotive: Game Changer. http://www.automotiveit.eu/sonderheft-big-data-automotive. In: *Sonderheft: Big Data Automotive*, 4–5. – Abgerufen am 13.12.2013

[Cloudera 2013a] CLOUDERA: *Cloudera Enterprise Datasheet*. http://www.cloudera.com/content/dam/cloudera/Resources/PDF/Cloudera_Datasheet_Cloudera_Enterprise.pdf. Version: 2013. – Abgerufen am 10.12.2013

[Cloudera 2013b] CLOUDERA: *Cloudera Impala: Open Source, Real-time Query for Hadoop*. http://www.cloudera.com/content/cloudera/en/products/cdh/impala.html. Version: 2013. – Abgerufen am 21.10.2013

[Cloudera 2013c] CLOUDERA: *Cloudera Support*. http://www.cloudera.com/content/cloudera/en/products/cloudera-support.html. Version: 2013. – Abgerufen am 21.10.2013

[Cloudera 2013d] CLOUDERA: *Frequently Asked Questions (FAQs)*. http://www.cloudera.com/content/cloudera/en/why-cloudera/faqs.html. Version: 2013. – Abgerufen am 21.10.2013

[Cloudera 2013e] CLOUDERA: *Introducing Morphlines: The Easy Way to Build and Integrate ETL Apps for Hadoop*. http://blog.cloudera.com/blog/2013/07/morphlines-the-easy-way-to-build-and-integrate-etl-apps-for-apache-hadoop/. Version: July 2013. – Abgerufen am 26.10.2013

[Cloudera 2013f] CLOUDERA: *Our Customers*. http://www.cloudera.com/content/cloudera/en/why-cloudera/our-customers.html. Version: 2013. – Abgerufen am 21.10.2013

[Cloudera 2013g] CLOUDERA: *The Platform for Big Data*. http://www.cloudera.com/content/cloudera/en/products.html. Version: 2013. – Abgerufen am 21.10.2013

[Cloudera 2013h] CLOUDERA: *The Platform for Big Data: A Blueprint for Next-Generation Data Management*. http://www.cloudera.com/content/dam/cloudera/Resources/PDF/the-platform-for-big-data.pdf. Version: 2013. – Abgerufen am 10.12.2013

[Cloudera 2013i] CLOUDERA: *Why Cloudera*. http://www.cloudera.com/content/cloudera/en/why-cloudera.html. Version: 2013. – Abgerufen am 21.10.2013

[Connecticum 2013] CONNECTICUM: *Arbeitgeber: Datameer*. http://www.connecticum.de/Arbeitgeber/Datameer-174487.html. Version: 2013. – Abgerufen am 16.10.2013

[Conway 2012] Conway, Dan: *Teradata Big Analytics Appliance Enables New Business Insights on All Enterprise*. http://www.teradata.de/News-Releases/2012/Teradata-Big-Analytics-Appliance-Enables-New-Business-Insights-on--All-Enterprise-Data/?LangType=1031. Version: 2012. – Abgerufen am 19.10.2013

[CX4U 2013a] CX4U: *Historie*. http://www.cx4u.com/cx4u/historie/. Version: 2013. – Abgerufen am 21.03.2014

[CX4U 2013b] CX4U: *Über Uns*. http://www.cx4u.com/cx4u/ueber-uns/. Version: 2013. – Abgerufen am 21.03.2014

[CX4U 2014] CX4U: *Solutions - SmartGateway*. www.cx4u.com/solutions/smartgateway/. Version: 2014. – Abgerufen am 25.03.2014

[Datameer 2013] Datameer: *About Datameer*. http://www.datameer.com/company/index.html. Version: 2013. – Abgerufen am 16.10.2013

[Datameer-Beratung 2013a] Datameer-Beratung: *Datameer Produkte - Datenanalyse für Jedermann*. http://www.datameer-beratung.de/datameer-produkte/. Version: 2013. – Abgerufen am 16.10.2013

[Datameer-Beratung 2013b] Datameer-Beratung: *Datameer Technologie*. http://www.datameer-beratung.de/datameer-technologie/. Version: 2013. – Abgerufen am 16.10.2013

[Datawatch 2013a] Datawatch: *Datawatch macht die Übernahme von Panopticon perfekt*. http://www.datawatch.de/images/PressReleases/MI_Datawatch_Panopticon_2.pdf. Version: 2013. – Abgerufen am 16.10.2013

[Datawatch 2013b] Datawatch ; www.presseportal.de (Hrsg.): *Datawatch Unstructured Data Connector für QlikView jetzt auf QlikMarket, der Börse für Business-Discovery-Lösungen verfügbar*. http://www.presseportal.de/pm/108500/2440618/datawatch-unstructured-data-connector-fuer-qlikview-jetzt-auf-qlikmarket-der-boerse-fuer-business. Version: 2013. – Abgerufen am 21.10.2013

[Datawatch 2013c] Datawatch: *Real-Time Data Visualization for Real-Time Businesses*. http://www.datawatch.com/images/Datasheets/Datawatch_Visualization_RealTimeBusiness_brochure.pdf. Version: 2013. – Abgerufen am 12.12.2013

[Datawatch 2013d] Datawatch: *Über Datawatch*. http://www.datawatch.de/ueber-uns/item/264-ueber-uns. Version: 2013. – Abgerufen am 16.10.2013

[Datawatch 2013e] Datawatch: *Visual Data Discovery*. http://www.datawatch.com/visual-data-discovery. Version: 2013. – Abgerufen am 12.12.2013

[Dijcks 2013] DIJCKS, Jean-Pierre: *Oracle: Big Data for the Enterprise: An Oracle White Paper*. http://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/1453236.pdf. Version: 2013. – Abgerufen am 16.12.2013

[Dill u. Füg 2013] DILL, Marcus ; FÜG, Thorsten: *HANA - Tipps zum Ausprobieren*. http://www.computerwoche.de/a/hana-tipps-zum-ausprobieren,2533668. Version: 2013. – Abgerufen am 19.10.2013

[Dippold u. Fedtke 2005] DIPPOLD, Rolf ; FEDTKE, Stephen: *Unternehmensweites Datenmanagement: Von der Datenbankadministration bis zum Informationsmanagement ; mit 19 Tabellen*. 4., überarb. und erw. Aufl. Braunschweig and Wiesbaden : Vieweg, 2005 (Zielorientiertes Business computing)

[DMAG 2009] DMAG: *Geschichte Entwicklung der CeBIT*. http://www.iris-media.info/IMG/pdf/3/3/6/89.pdf. Version: 2009. – Abgerufen am 24.03.2014

[DWDM 2013] Abgerufen am 21.10.2013

[EMC 2011] EMC: *EMC Greenplum HD CE*. http://germany.emc.com/collateral/software/data-sheet/h8894-greenplum-hd-community-ds.pdf. Version: 2011. – Abgerufen am 16.12.2013

[EMC 2013a] EMC: *Die EMC Big Data-Lösung*. http://germany.emc.com/collateral/brochure/h10769-big-data-cta-brochure.pdf. Version: 2013. – Abgerufen am 16.12.2013

[EMC 2013b] EMC: *EMC im Überblick*. http://germany.emc.com/corporate/emc-at-glance/corporate-profile/index.htm. Version: 2013. – Abgerufen am 16.12.2013

[EMC 2013c] EMC: *Support und Schulung*. http://germany.emc.com/support-training/index.htm. Version: 2013. – Abgerufen am 16.12.2013

[Engels 2008] ENGELS, Christoph: *Basiswissen Business Intelligence*. W3L GmbH, 2008 http://amazon.com/o/ASIN/3937137378/

[Enzyklopaedie der Wirtschaftsinformatik 2013a] Abgerufen am 21.10.2013

[Enzyklopaedie der Wirtschaftsinformatik 2013b] Abgerufen am 21.10.2013

[Ertel u. Pfleghar 2012] ERTEL, Sandra ; PFLEGHAR, Frieder: *Presseinformation: Markteinführung: Teradata Aster Big Analytics Appliance erschließt alle Unternehmensdaten*. http://www.teradata.com/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=20582&libID=20566. Version: 2012. – Abgerufen am 19.10.2013

[EXASOL AG 2012] EXASOL AG: *EXASolution: Technical White Paper*. http://www.exasol.com/fileadmin/content-de/pdf/Whitepaper/EXASolution_TWP_082012.pdf. Version: 2012. – Abgerufen am 19.10.2013

[EXASOL AG 2013a] EXASOL AG: *Data Warehouse Appliance*. http://www.exasol.com/produkte/exaappliance/. Version: 2013. – Abgerufen am 19.10.2013

[EXASOL AG 2013b] EXASOL AG: *EXACloud*. http://www.exasol.com/produkte/exacloud/. Version: 2013. – Abgerufen am 19.10.2013

[EXASOL AG 2013c] EXASOL AG: *EXAPowerlytics*. http://www.exasol.com/produkte/exapowerlytics/. Version: 2013. – Abgerufen am 19.10.2013

[EXASOL AG 2013d] EXASOL AG: *EXASolution - Schnell, smart und simpel*. http://www.exasol.com/produkte/exasolution/. Version: 2013. – Abgerufen am 19.10.2013

[EXASOL AG 2013e] EXASOL AG: *FAQ: Antworten auf die häufigsten Fragen zu EXASolution*. http://www.exasol.com/faq/. Version: 2013. – Abgerufen am 19.10.2013

[EXASOL AG 2013f] EXASOL AG: *Über EXASOL*. http://www.exasol.com/das-unternehmen/. Version: 2013. – Abgerufen am 19.10.2013

[Finanznachrichten 2012] FINANZNACHRICHTEN: *MicroStrategy Enhances Innovative Big Data Capabilities with Certified Integration to Amazon Redshift Analytical Data Warehouse Service*. http://www.finanznachrichten.de/nachrichten-2012-11/25311140-microstrategy-enhances-innovative-big-data-capabilities-with-certified-integration-to-amazon-redshift-analytical-data-warehouse-service-008.htm. Version: 2012. – Abgerufen am 17.12.2013

[Fischer 2010] FISCHER, Oliver: Verarbeiten großer verteilter Datenmengen mit Hadoop / Heise. http://www.heise.de/developer/artikel/Verarbeiten-grosser-verteilter-Datenmengen-mit-Hadoop-964755.html, 2010. – Abgerufen am 16.12.2013

[Franke 2013] FRANKE, Susanne: *In-Memory-Konzepte im Vergleich*. http://www.computerwoche.de/a/in-memory-konzepte-im-vergleich,2523531. Version: 2013. – Abgerufen am 19.10.2013

[Fujitsu 2013] FUJITSU: *White Paper: Lösungsansätze für Big Data*. http://globalsp.ts.fujitsu.com/dmsp/Publications/public/wp-bigdata-solution-approaches-de.pdf. Version: 2013. – Abgerufen am 02.01.2014

[Gabler Wirtschaftslexikon 2013a] GABLER WIRTSCHAFTSLEXIKON ; SPRINGER GABLER VERLAG (Hrsg.): *Stichwort: Data Mining*. http://wirtschaftslexikon.gabler.de/Archiv/57691/data-mining-v7.html. Version: 2013. – Abgerufen am 05.12.2013

[Gabler Wirtschaftslexikon 2013b] GABLER WIRTSCHAFTSLEXIKON ; SPRINGER GABLER VERLAG (Hrsg.): *Stichwort: Softwarequalität*. http://wirtschaftslexikon.gabler.de/Definition/softwarequalitaet.html. Version: 2013. – Abgerufen am 17.01.2014

[Gardella 2013a] Gardella, William: *Big Data Hana Hadoop*. http://de.slideshare.net/willgardella/sf-scalable-systems-meetup. Version: 21. Jan 2013. – Abgerufen am 23.03.2014

[Gardella 2013b] Gardella, William: *HANA and Hadoop for Big Data Management*. http://de.slideshare.net/willgardella/sf-scalable-systems-meetup. Version: 01.10. 2013. – Abgerufen am 22.03.2014

[Gartner, Inc. 2013a] Gartner, Inc.: *Gartner's 2013 Hype Cycle for Emerging Technologies*. http://www.gartner.com/newsroom/id/2575515. Version: 2013. – Abgerufen am 24.03.2014

[Gartner, Inc. 2013b] Gartner, Inc.: *Oracle Gartner*. http://www.gartner.com/technology/reprints.do?id=1-1E1KL1X&ct=130211&st=sb. Version: 2013. – Abgerufen am 23.03.2014

[Gebauer u. Windheuser 2008] Gebauer, Marcus ; Windheuser, Ulricht: Strukturierte Datenanalyse, Profiling und Geschäftsregeln. In: Hildebrand, Knut (Hrsg.): *Daten- und Informationsqualität*. Wiesbaden : Vieweg + Teubner, 2008 (Praxis), S. 88–101

[Gleich 2013] Gleich, Ronald: *Komplexitätscontrolling*. 1. Aufl. s.l : Haufe Verlag, 2013 http://ebooks.ciando.com/book/index.cfm/bok_id/1019210

[Goll 2011] Goll, Joachim: *Methoden und Architekturen der Softwaretechnik*. Wiesbaden : Vieweg+Teubner Verlag, 2011

[Hortonworks 2013a] Hortonworks: *About us*. http://hortonworks.com/about-us/. Version: 2013. – Abgerufen am 12.11.2013

[Hortonworks 2013b] Hortonworks: *Ambari on EC2*. http://hortonworks.com/kb/ambari-on-ec2/. Version: 09.12. 2013. – Abgerufen am 09.12.2013

[Hortonworks 2013c] Hortonworks: *Apache Flume*. http://hortonworks.com/hadoop/flume/. Version: 2013. – Abgerufen am: 26.10.2013

[Hortonworks 2013d] Hortonworks: *Build a Modern Data Architecture*. http://hortonworks.com/hadoop-modern-data-architecture/. Version: 2013. – Abgerufen am 11.12.2013

[Hortonworks 2013e] Hortonworks: *Hadoop Support Subsciption*. http://hortonworks.com/products/hadoop-support/. Version: 2013. – Abgerufen am 12.11.2013

[Hortonworks 2013f] Hortonworks: *HDP for Windows*. http://hortonworks.com/products/hdp-windows/#documentation. Version: 2013. – Abgerufen am 11.12.2013

[Hortonworks 2013g] Hortonworks: *Hortonworks Data Platform*. http://hortonworks.com/products/hdp/. Version: 2013. – Abgerufen am 12.11.2013

[Hortonworks 2013h] HORTONWORKS: *Hortonworks Sandbox*. http://hortonworks.com/products/hortonworks-sandbox/. Version: 2013. – Abgerufen am 11.12.2013

[Hortonworks 2013i] HORTONWORKS: *Join the Community*. http://hortonworks.com/community/. Version: 2013. – Abgerufen am 11.12.2013

[Hortonworks 2013j] HORTONWORKS: *They Do Hadoop*. http://hortonworks.com/customers/. Version: 2013. – Abgerufen am 11.12.2013

[HP] HP: *HP Vertica Crane wirft Licht auf "dunkle Daten"*. http://www8.hp.com/de/de/hp-news/press-release.html?id=1533587#.UrCrUNLuKvM. – Abgerufen am 17.12.2013

[HP 2013a] HP: *Big Data and Analytics*. http://www8.hp.com/de/de/software-solutions/software.html?compURI=1214060#.UrBi39LuKvM. Version: 2013. – Abgerufen am 17.12.2013

[HP 2013b] HP: *HP Startseite*. http://www8.hp.com/de/de/home.html. Version: 2013. – Abgerufen am 17.12.2013

[HP 2013c] HP: *HP Support*. http://www8.hp.com/de/de/support.html. Version: 2013. – Abgerufen am 17.12.2013

[IBM 2013a] IBM: *Big Data - IBM Produktportfolio für Big Data*. http://www-03.ibm.com/software/products/de/de/category/SWP10. Version: 2013. – Abgerufen am 12.12.2013

[IBM 2013b] IBM: *IBM Cognos 10 Intelligence Unleashed*. http://www-01.ibm.com/software/de/analytics/cognos/cognos10/. Version: 2013. – Abgerufen am 12.12.2013

[IBM 2013c] IBM: *IBM InfoSphere Streams: Redefining real-time analytics processing*. http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=SWGE_IM_IM_USEN&htmlfid=IMW14704USEN&attachment=IMW14704USEN.PDF. Version: 2013. – Abgerufen am 07.02.2014

[IBM 2013d] IBM: *IBM Netezza 1000*. http://www-03.ibm.com/software/products/de/ibmnete1000/. Version: 2013. – Abgerufen am 12.12.2013

[IBM 2013e] IBM: *InfoSphere BigInsights Enterprise Edition*. http://www-03.ibm.com/software/products/de/infobigienteedit/. Version: 2013. – Abgerufen am 12.12.2013

[IBM 2013f] IBM: *InfoSphere Streams*. http://www-03.ibm.com/software/products/de/infosphere-streams. Version: 2013. – Abgerufen am 12.12.2013

[IBM 2014a] Abgerufen am 24.03.2014

[IBM 2014b] Abgerufen am 16.03.2014

[IBM 2014c] IBM: *Support Offerings*. http://www-01.ibm.com/software/support/offerings.html. Version: 2014. – Abgerufen am 07.02.2014

[IBM Corporation 2013a] IBM Corporation: *IBM Big Data*. https://www.ibm.com/developerworks/community/blogs/ibm-big-data/entry/october_18_2011_2_41_am1?lang=en. Version: 01.10. 2013. – Abgerufen am 22.10.2013

[IBM Corporation 2013b] IBM Corporation: *IBM Cognos TM1 Version 10.2.0 Benutzerhandbuch*. http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_ug.pdf. Version: 2013. – Abgerufen am 13.02.2014

[IBM Corporation 2013c] IBM Corporation: *IBM Cognos TM1 Version 10.2.0 Betriebshandbuch*. http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_op.pdf. Version: 2013. – Abgerufen am 05.03.2014

[IBM Corporation 2013d] IBM Corporation: *IBM Cognos TM1 Version 10.2.0 Entwicklerhandbuch*. http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_dg_dvlpr.pdf. Version: 2013. – Abgerufen am 28.01.2014

[IBM Corporation 2013e] IBM Corporation: *IBM Cognos TM1 Version 10.2.0 Installation und Konfiguration*. http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_inst.pdf. Version: 2013. – Abgerufen am 13.02.2014

[IBM Corporation 2013f] IBM Corporation: *IBM Cognos TM1 Version 10.2.0 Operations Console Handbuch*. http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_ops_console.pdf. Version: 2013. – Abgerufen am 19.02.2014

[IBM Corporation 2013g] IBM Corporation: *IBM Cognos TM1 Version 10.2.0 Turbo Integrator Handbuch*. http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_turb.pdf. Version: 2013. – Abgerufen am 05.03.2014

[IBM Corporation 2013h] IBM Corporation: *IBM Cognos TM1 Version 10.2.0 Web-Benutzerhandbuch*. http://public.dhe.ibm.com/software/data/cognos/documentation/docs/de/10.2.0/tm1_wb.pdf. Version: 2013. – Abgerufen am 14.03.2014

[IBM Corporation 2014] IBM Corporation: *Cognos TM1*. http://www-03.ibm.com/software/products/de/cognostm1/. Version: 2014. – Abgerufen am 21.01.2014

[IBM Redbooks 2012] IBM Redbooks: *IBM Cognos Dynamic Cubes*. Vervante, 2012 www.redbooks.ibm.com/redbooks/pdfs/sg248064.pdf

[Ihlenfeld 2012] Ihlenfeld, Jens: *Impala: Echtzeitabfragen mit Hadoop*. http://www.golem.de/news/impala-echtzeitabfragen-mit-hadoop-1210-95289.html. Version: 2012. – Abgerufen am 21.10.2013

[Informatica 2012] Informatica: *Informatica PowerCenter Big Data Edition*. http://www.informatica.com/de/Images/02194_powercenter-big-data-edition_ds_de-DE.pdf. Version: 2012. – Abgerufen am 11.12.2013

[Informatica 2013a] Informatica: *Big Data-Integration*. http://www.informatica.com/de/products/enterprise-data-integration/powercenter/powercenter-big-data-edition/. Version: 2013. – Abgerufen am 21.10.2013

[Informatica 2013b] Informatica: *Customer Support Programs*. http://www.informatica.com/us/services-and-training/support-services/support-programs/. Version: 2013. – Abgerufen am 21.10.2013

[Informatica 2013c] Informatica: *Informatica PowerCenter Big Data Edition*. http://www.informatica.com/Images/02194_powercenter-big-data-edition_ds_en-US.pdf. Version: 01.10. 2013. – Abgerufen am 22.10.2013

[Informatica 2013d] Informatica: *Kunden*. http://www.informatica.com/de/customers/. Version: 2013. – Abgerufen am 21.10.2013

[Informatica 2013e] Informatica: *PowerCenter*. http://www.informatica.com/de/products/enterprise-data-integration/powercenter/. Version: 2013. – Abgerufen am 21.10.2013

[Informatica 2013f] Informatica: *Produkte*. http://www.informatica.com/de/products/. Version: 2013. – Abgerufen am 21.10.2013

[Informatica 2013g] Informatica: *Unternehmen*. http://www.informatica.com/de/company/. Version: 2013. – Abgerufen am 21.10.2013

[InsideView 2013] InsideView: *Jaspersoft Corporation*. http://www.insideview.com/directory/jaspersoft-corporation. Version: 2013. – Abgerufen am 16.10.2013

[IS Report 2013] IS Report: *Business Intelligence Lösungs-Guide 2013*. http://www.isreport.de/epaper/Files%20BI/. Version: 2013. – Abgerufen am 11.12.2013

[IT-Daily 2013a] IT-Daily: *BigMemory 4.0 bietet bis zu 60 Prozent schnelleren Zugriff auf Big Data*. http://it-daily.net/it-technologie/aktuelles/7288-mit-bigmemory-4-0-bis-zu-60-prozent-schnellerer-zugriff-auf-big-data. Version: 2013. – Abgerufen am 19.10.2013

[IT-Daily 2013b] IT-Daily: *Datameer 3.0 stellt Smart Analytics vor*. http://www.it-daily.net/it-technologie/aktuelles/7659-datameer-3-0-stellt-smart-analytics-vor. Version: 2013. – Abgerufen am 11.12.2013

[Jaspersoft 2013a] Jaspersoft: *Gute Gründe für Jaspersoft*. http://www.jaspersoft.com/de/gute-gr%C3%BCnde-f%C3%BCr-jaspersoft. Version: 2013. – Abgerufen am 16.10.2013

[Jaspersoft 2013b] Jaspersoft: *Instant Big Data Analytics*. http://www.jaspersoft.com/de/node/54854. Version: 2013. – Abgerufen am 16.10.2013

[Jaspersoft 2013c] Jaspersoft: *Jaspersoft Community*. http://community.jaspersoft.com/. Version: 2013. – Abgerufen am 16.10.2013

[Jaspersoft 2013d] Jaspersoft: *Support Options*. http://www.jaspersoft.com/support-options-table. Version: 2013. – Abgerufen am 16.10.2013

[Jaspersoft 2013e] Jaspersoft: *Unterstützte Plattformen*. http://www.jaspersoft.com/de/supported-platforms. Version: 2013. – Abgerufen am 16.10.2013

[Jung 2012a] Jung, Thomas: *ABAP-HANA*. http://liualex.com/wp-content/uploads/2012/06/ABAPHANA.pdf. Version: 2012. – Abgerufen am 09.03.2014

[Jung 2012b] Jung, Thomas: *Developer's Journal: ABAP/HANA Connectivity via Secondary Database Connection*. http://scn.sap.com/community/hana-in-memory/blog/2012/04/11/test. Version: 04 2012. – Abgerufen am 26.03.2014

[Jung 2012c] Jung, Thomas: *SAP HANA Extended Application Services*. http://scn.sap.com/community/developer-center/hana/blog/2012/11/29/sap-hana-extended-application-services. Version: 2012. – Abgerufen am 26.03.2014

[Khan 2013] Khan, Imran: *Whitepaper LTReplicationServer4SAPHANA Final*. http://www.scribd.com/doc/159675011/Whitepaper-LTReplicationServer4SAPHANA-Final. Version: 2013. – Abgerufen am 09.10.2013

[Koebler 2013] Koebler, Tobias: *Starter Blog for SAP LT Replication Server*. http://scn.sap.com/community/replication-server/blog/2013/06/10/starter-blog-for-sap-lt-replication-server. Version: 10. Juni 2013. – Abgerufen am 07.10.2013

[Krempl 2013] Krempl, Stephan: *Big Data Days: Wirtschaft und Politik im Datenrausch*. http://www.heise.de/newsticker/meldung/Big-Data-Days-Wirtschaft-und-Politik-im-Datenrausch-2044287.html. Version: 11 2013. – Abgerufen am 26.12.2013

[Kurbel u. a. 2014] KURBEL, Karl ; BECKER, Jörg ; GRONAU, Norbert: *Enzyklopädie der Wirtschaftsinformatik*. http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie. Version: 2014. – Abgerufen am 22.03.2014

[van Laak 2013] LAAK, Kirsten van ; EXPERTON GROUP AG (Hrsg.): *Pressemeldung: Experton Group Big Data Vendor Benchmark 2013*. http://www.experton-group.de/press/releases/pressrelease/article/experton-group-big-data-vendor-benchmark-2013.html. Version: 2013. – Abgerufen am 13.12.2013

[Landrock 2013] LANDROCK, Holm: *Experton Group analysiert Big Data-Angebote*. http://www.isreport.de/news-events/news/archiv/2013/08/26/article/experton-grop-analysiert-big-data-angebote.html. Version: 2013. – Abgerufen am 19.10.2013

[Leser 2007] LESER, Naumann: *Informationsintegration*. Heidelberg : dpunkt.Verlag, 2007

[ManageIT 2012] MANAGEIT: SAP HANA: Aufbruch in ein neues Zeitalter? / AP-Verlag. http://ap-verlag.de/Online-Artikel/20120910/m%20In-Memory-Technologie%20SAP%20HANA.htm, 2012. – Forschungsbericht. – Letzter Aufruf: 16.12.2013

[Manhart 2012] MANHART, Klaus: *Big Data Appliance: Große Datenmengen einfach erschließen*. http://www.computerwoche.de/a/grosse-datenmengen-einfach-erschliessen,2502956. Version: 2012. – Abgerufen am 16.12.2013

[Martin 2013] MARTIN, Wolfgang: *Analytische Datenbanken*. 2013

[Mellis 2008] MELLIS, Werner: *Projektmanagement in der Softwareentwicklung*. Friedr. Vieweg & Sohn / GWV Fachverlage, 2008

[MicroStrategy 2013a] MICROSTRATEGY: *About Us*. http://www.microstrategy.com/about-us/overview. Version: 2013. – Abgerufen am 17.12.2013

[MicroStrategy 2013b] MICROSTRATEGY: *Big Data*. http://www.microstrategy.de/software/business-intelligence/big-data/. Version: 2013. – Abgerufen am 17.12.2013

[MicroStrategy 2013c] MICROSTRATEGY: *Business Intelligence für Unternehmen*. http://www.microstrategy.de/software/business-intelligence/. Version: 2013. – Abgerufen am 17.12.2013

[MicroStrategy 2013d] MICROSTRATEGY: *MicroStrategy Business Intelligence Platform: Vereinheitlichte Business Intelligence-Architektur*. http://www.microstrategy.de/software/business-intelligence/platform/. Version: 2013. – Abgerufen am 17.12.2013

[MicroStrategy 2013e] MICROSTRATEGY: *Support*. http://www.microstrategy.de/services-support/support/. Version: 2013. – Abgerufen am 17.12.2013

[MicroStrategy 2013f] MicroStrategy: *Wir präsentieren MicroStrategy 9.3*. http://www.microstrategy.de/software/business-intelligence/9-3/. Version: 2013. – Abgerufen am 17.12.2013

[MWK 2014] MWK: *Innovationsland Niedersachsen*. http://www.cebit.de/aussteller/nds.-ministerium-f.-wissenschaft/M435467. Version: 2014. – Abgerufen am 24.03.2014

[Nannen 2013] Nannen, Okko: *Systemlandschaft CX4U*. 2013

[o.A. 2012] o.A.: *Oracle*. http://de.wikipedia.org/wiki/Oracle. Version: 2012. – Abgerufen am 16.12.2013

[Oracle Corporation 2013a] Oracle Corporation: *HDFS Oracle*. http://docs.oracle.com/cd/E37231_01/doc.20/e36961/sqlch.htm. Version: 2013. – Abgerufen am 22.10.2013

[Oracle Corporation 2013b] Oracle Corporation: *Oracle Big Data Appliance X4-2*. http://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/1883358.pdf. Version: 2013. – Abgerufen am 16.12.2013

[Oracle Corporation 2013c] Oracle Corporation: *Oracle Docs*. http://docs.oracle.com/cd/E27101_01/doc.10/e27365/olh.htm. Version: 2013. – Abgerufen am 22.10.2013

[Oracle Corporation 2013d] Oracle Corporation: *Oracle Engineered Systems: Extreme Performance. Fastest Deployment. Lowest Cost.* http://www.oracle.com/de/products/engineered-systems/index.html. Version: 2013. – Abgerufen am 16.12.2013

[Oracle Corporation 2013e] Oracle Corporation: *Oracle Fact Sheet: Unleash Innovation by Simplifying IT*. http://www.oracle.com/us/corporate/oracle-fact-sheet-079219.pdf?ssSourceSiteId=ocomde. Version: 2013. – Abgerufen am 16.12.2013

[Oracle Corporation 2013f] Oracle Corporation: *Oracle und Big Data*. http://www.oracle.com/de/technologies/big-data/index.html. Version: 2013. – Abgerufen am 16.12.2013

[Ortega 2013] Abgerufen am 21.10.2013

[Osten 2012] Osten, Cathrin v.: *SAP erhöht Reichweite der Plattform von SAP HANA*. http://de.news-sap.com/2012/10/17/sap-erhoht-reichweite-der-plattform-von-sap-hana/. Version: 2012. – Abgerufen am 19.10.2013

[ParStream 2013a] ParStream: *Company*. https://www.parstream.com/about-us/company/. Version: 2013. – Abgerufen am 16.12.2013

[ParStream 2013b] ParStream: *Factsheet ParStream*. https://www.parstream.com/wp-content/uploads/2013/12/factsheet_ParStream.pdf. Version: 2013. – Abgerufen am 16.12.2013

[ParStream 2013c] ParStream: *Features*. https://www.parstream.com/product/features. Version: 2013. – Abgerufen am 16.12.2013

[ParStream 2013d] ParStream: *Technology*. https://www.parstream.com/product/technology/. Version: 2013. – Abgerufen am 16.12.2013

[Peddinghaus u. a. 2011] Peddinghaus, Oliver ; Ringleben, Anselm ; Garapati, Sindhura: *Lastenheft: Business & Service Daten Elsa Pro Access Logs*. 2011

[Pentaho 2013a] Pentaho: *Community Wiki Home*. http://wiki.pentaho.com/display/COM/Community+Wiki+Home. Version: 2013. – Abgerufen am 19.10.2013

[Pentaho 2013b] Pentaho: *Pentaho Business Analytics*. http://www.pentaho.de/explore/pentaho-data-integration/. Version: 2013. – Abgerufen am 21.10.2013

[Pentaho 2013c] Pentaho: *Pentaho Business Analytics*. http://www.pentaho.de/explore/pentaho-business-analytics/. Version: 2013. – Abgerufen am 21.10.2013

[Pentaho 2013d] Pentaho: *Pentaho Business Analytics: Datenintegration ¿ Auswertung ¿ Prognose*. http://www.pentaho.de/assets/pdf_161.html. Version: 2013. – Abgerufen am 21.10.2013

[Pentaho 2013e] Pentaho: *Pentaho Data Integration*. http://www.pentaho.de/assets/pdf_164.html. Version: 2013. – Abgerufen am 19.10.2013

[Pentaho 2013f] Pentaho: *Pentaho Produkte*. http://www.pentaho.de/explore/products/. Version: 2013. – Abgerufen am 19.10.2013

[Pentaho 2013g] Pentaho: *Über uns*. http://www.pentaho.de/about/story/. Version: 2013. – Abgerufen am 19.10.2013

[Pivotal 2013] Pivotal: *Greenplum Database Features*. http://gopivotal.com/products/pivotal-greenplum-database. Version: 2013. – Abgerufen am 16.12.2013

[Platten 2013] Platten, Wilfried: *Data Warehousing und Business Analytics: Wer hat's erfunden? In-Memory-Datenbank von Exasol*. 2013

[plus-IT GmbH 2013] plus-IT GmbH: *SAP HANA (High performance Analytical Appliance)*. http://www.plus-it.de/sap-hana-high-performance-analytical-applience.html. Version: 2013. – Abgerufen am 19.10.2013

[Pressebox 2012] PRESSEBOX: *Software AG ist Finalist beim Big Data Award*. http://www.pressebox.de/pressemitteilung/software-ag/Software-AG-ist-Finalist-beim-Big-Data-Award/boxid/541963. Version: 2012. – Abgerufen am 19.10.2013

[QlikTech 2013a] QLIKTECH: *Customers*. http://www.qlikview.com/de/explore/customers. Version: 2013. – Abgerufen am 21.10.2013

[QlikTech 2013b] QLIKTECH: *QlikTech: Das Unternehmen hinter QlikView*. http://www.qlikview.com/de/company. Version: 2013. – Abgerufen am 21.10.2013

[QlikTech 2013c] QLIKTECH: *Die QlikView Business-Discovery-Plattform*. 2013. – Abgerufen am 12.10.2013

[QlikTech 2013d] QLIKTECH: *QlikView für mobile Endgeräte*. http://www.qlikview.com/de/explore/products/qv-for-mobile. Version: 2013. – Abgerufen am 21.10.2013

[QlikTech 2013e] QLIKTECH: *QlikView im Überblick: Die QlikView Business-Discovery-Plattform*. http://www.qlikview.com/de/explore/products/overview. Version: 2013. – Abgerufen am 21.10.2013

[QlikTech 2013f] QLIKTECH: *QlikView und Big Data*. 2013

[QlikTech 2013g] QLIKTECH: *Support*. http://www.qlikview.com/de/services/support. Version: 2013. – Abgerufen am 21.10.2013

[QlikTech 2013h] QLIKTECH: *Unternehmenshistorie*. http://www.qlikview.com/de/company/history. Version: 2013. – Abgerufen am 21.10.2013

[Rachel 2011] RACHEL, Frank: *SAP In-Memory Appliance: (SAP HANA)*. http://www.dataport.de/ueber-uns/veranstaltungen/Documents/2011/Hausmesse-Vortr%C3%A4ge-SAP/DATAPORT_HAUSMESSE_SAP_HANA_FR.pdf. Version: 2011. – Abgerufen am 19.10.2013

[Rommelspacher 2011] ROMMELSPACHER, Jonas: Automatisierung von Führungsentscheidungen: Framework, Modellierung und Prototyp. In: *Automatisierung von Führungsentscheidungen* (2011)

[Saake u. a. 2010] SAAKE, Gunter ; SATTLER, Kai-Uwe ; HEUER, Andreas: *Datenbanken: Konzepte und Sprachen*. 4. Aufl. Heidelberg and München and Landsberg and Frechen and Hamburg : mitp, 2010 (Biber-Buch)

[SAP 2011a] SAP: *High Performance CEP: The SAP Sysbase Event Stream Processor*. http://www.slideshare.net/SybaseTurkey/esp-5-prodtech-overview-sap-sybase. Version: 2011. – Abgerufen am 28.03.2014

[SAP 2011b] SAP: *OData in a Nutshell*. http://wiki.scn.sap.com/wiki/download/attachments/233739132/02_OData_In_A_Nutshell.pdf?original_fqdn=wiki.sdn.sap.com&version=1. Version: 2011. – Abgerufen am 28.03.2014

[SAP 2012a] SAP: *SAP LT Replication Server*. 2012. – Abgerufen am 30.10.2013

[SAP 2012b] SAP: *SAP LT Replication Server for SAP HANA SPS 05 Overview*. 2012. – Abgerufen am 11.11.2013

[SAP 2013] SAP: *Smart Data Access (Data Virtualization) with SAP HANA*. http://www.saphana.com/community/blogs/blog/2013/07/22/smart-data-access-data-virtualization-with-sap-hana. Version: 2013. – Abgerufen am 26.13.2014

[SAP AG 2012] SAP AG: *CIO Online - SAP Data Services*. http://www.sap-cio.de/neue-beitrage/losungen/database-technologie/sap-bobj-data-services/. Version: 2012. – Abgerufen am 15.12.2013

[SAP AG 2013a] SAP AG: *Help Portal - SAP Data Services 4.2*. https://help.sap.com/bods. Version: 2013. – Abgerufen am 15.12.2013

[SAP AG 2013b] SAP AG: *SAP HANA Academy*. http://www.saphana.com/community/hana-academy. Version: 2013. – Abgerufen am 19.10.2013

[SAP AG 2013c] SAP AG: *SAP-HANA-Lösungen: Überblick*. http://www.sap.com/germany/solutions/technology/in-memory-computing-platform/hana/overview/index.epx. Version: 2013. – Abgerufen am 19.10.2013

[SAP AG 2013d] SAP AG: *Über SAP AG*. http://www.sap.com/germany/about-sap/about-sap.epx. Version: 2013. – Abgerufen am 19.10.2013

[SAPAG 2013a] SAPAG: *Beispielprogramm Ausführung eines HTTP Requests*. http://help.sap.com/saphelp_nw70ehp2/helpdata/de/1f/93163f9959a808e10000000a114084/content.htm. Version: 2013. – Abgerufen am 28.01.2014

[SAPAG 2013b] SAPAG: *SAP HANA Technical Operations Manual*. 2013. – Abgerufen am 21.10.2013

[SAS 2012] SAS: *SAS Visual Analytics: Big Data intuitiv analysieren*. http://www.sas.com/offices/europe/switzerland/news/pressrelease/pi10_SAS_Visual_Analytics.htmlnote={Abgerufenam15.12.2013},urldate={15.12.2013}. Version: 2012

[SAS 2013a] SAS: *Analytics auf den ersten Blick*. http://sas-thinkforward.de/content/beitraege/inhalt_1___2013/sas_visual_analytics/. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013b] SAS: *Customer Support*. http://www.sas.com/offices/europe/germany/services/cts/index.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013c] SAS: *SAS Solutions OnDemand*. http://www.sas.com/solutions/ondemand/index.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013d] SAS: *SAS Visual Analytics: Die Technologie von SAS Visual Analytics*. http://www.sas.com/offices/europe/germany/software/visual-analytics/technology.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013e] SAS: *SAS Visual Analytics: Product Documentation*. http://support.sas.com/documentation/onlinedoc/va/index.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013f] SAS: *SAS Visual Analytics: Warum SAS?* http://www.sas.com/offices/europe/germany/software/visual-analytics/overview.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013g] SAS: *SAS/ACCESS Software*. http://www.sas.com/software/data-management/access/index.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013h] SAS: *Services und Support*. http://www.sas.com/offices/europe/germany/services/index.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013i] SAS: *Über SAS*. http://www.sas.com/offices/europe/germany/sas/sas.html. Version: 2013. – Abgerufen am 15.12.2013

[SAS 2013j] SAS: *Was ist Big Data?* http://www.sas.com/offices/europe/germany/software/datenmanagement/big_data.html. Version: 2013

[Schekelmann 2011] Schekelmann, Andre: *Vorlesung Softwareprojekt Folie Softwareprojekt Unterlagen*. 2011

[Schindler 2013] Schindler, Martin: *Software AG verknüpft Messaging mit Big Data*. http://www.silicon.de/41585642/software-ag-verknupft-messaging-mit-big-data/. Version: 2013. – Abgerufen am 16.12.2013

[Schlosser 2009] Schlosser, Joachim: *Wissenschaftliche Arbeiten schreiben mit LaTeX*. 3., überarbeitete Auflage. Heidelberg and München and Landsberg and Frechen and Hamburg : mitp, 2009

[Schnurr 2014] Schnurr, Roland: *Six Sigma Black Belt*. http://sixsigmablackbelt.de/wp-content/uploads/2013/05/Fehlerkosten-10-er-Regel.png. Version: 2014. – Abgerufen am 22.03.2014

[Selbach Information Systems 2014] Selbach Information Systems: *IBM Cognos TM1*. http://www.selbach-gmbh.de/produkte/ibm-cognos-tm1-multidimensionales-datenbanksystem.html. Version: 2014

[Software AG 2013a] SOFTWARE AG: *Big Data in der Praxis*. http://www.softwareag.com/de/images/Vortrag_BigData_tcm17-108645.pdf. Version: 2013. – Abgerufen am 19.10.2013

[Software AG 2013b] SOFTWARE AG: *Big Memory: Funktionen*. 2013

[Software AG 2013c] SOFTWARE AG: *Produkte A-Z: Online Produktkatalog*. http://www.softwareag.com/de/products/az/aris/default.asp. Version: 2013. – Abgerufen am 19.10.2013

[Software AG 2013d] SOFTWARE AG: *Referenzkunden der Software AG*. http://www.softwareag.com/de/customers/references/default.asp. Version: 2013. – Abgerufen am 19.10.2013

[Software AG 2013e] SOFTWARE AG: *Übersicht: Software AG auf einen Blick*. http://www.softwareag.com/de/company/companyinfo/overview/default.asp. Version: 2013. – Abgerufen am 19.10.2013

[solid IT 2014] SOLID IT: *CAP Theorem*. http://db-engines.com/de/article/CAP+Theorem. Version: 2014. – Abgerufen am 07.02.2014

[Spies 2012] SPIES, Rüdiger: Big Data als Katalysator für neue Business-Analytics-Projekte. In: LANG, Michael (Hrsg.): *CIO-Handbuch 2012/13*. Düsseldorf : Symposion Publishing, 2012, S. 141–158

[Splunk 2013a] SPLUNK: *Documentation*. http://docs.splunk.com/Documentation. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013b] SPLUNK: *Operational Intelligence Solutions*. http://www.splunk.com/solutions. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013c] SPLUNK: *Produktübersicht*. http://de.splunk.com/product. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013d] SPLUNK: *Splunk Enterprise*. http://www.splunk.com/view/splunk/SP-CAAAG57. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013e] SPLUNK: *Splunk for Big Data*. http://www.splunk.com/web_assets/pdfs/secure/Splunk_for_BigData.pdf. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013f] SPLUNK: *Splunk Products*. http://www.splunk.com/product. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013g] SPLUNK: *Support Programs*. http://www.splunk.com/support. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013h] SPLUNK: *Vorteile von Splunk*. http://de.splunk.com/view/benefits/SP-CAAACCS. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2013i] SPLUNK: *Was sind IT-Daten?* http://de.splunk.com/view/SP-CAAACDC. Version: 2013. – Abgerufen am 21.10.2013

[Splunk 2014a] SPLUNK: *Splunk DB Connect*. http://www.splunk.com/view/db-connect/SP-CAAAHR6. Version: 2014. – Abgerufen am 15.01.2014

[Splunk 2014b] SPLUNK: *Splunk Hadoop Connect*. http://www.splunk.com/view/hadoop-connect/SP-CAAAHA3. Version: 2014. – Abgerufen am 15.01.2014

[Steller 2013] STELLER, P.: *Leider wird Innovation auch missbraucht*. http://www.gruenderszene.de/allgemein/datameer. Version: 2013. – Abgerufen am 16.10.2013

[Stych u. Zeppenfeld 2009] STYCH, Christof ; ZEPPENFELD, Klaus: *ITIL (Informatik im Fokus) (German Edition)*. 1. Aufl. 2008. Korr. Nachdruck 2009. Springer, 2009 http://amazon.com/o/ASIN/3540731180/

[Teradata 2013a] TERADATA: *About us*. http://www.teradata.de/about-us/. Version: 2013. – Abgerufen am 19.10.2013

[Teradata 2013b] TERADATA: *Browse our customers*. http://www.teradata.de/customers-list/browse/. Version: 2013. – Abgerufen am 19.10.2013

[Teradata 2013c] TERADATA: *Teradata Aster Big Analytic Appliance: An Industry First*. http://www.asterdata.com/resources/downloads/datasheets/Teradata-Aster-Big-Analytics-Appliance-Datasheet.pdf. Version: 2013. – Abgerufen am 19.10.2013

[Teradata 2013d] TERADATA: *Über uns*. http://www.aprimo.com/About_Us/. Version: 2013. – Abgerufen am 19.10.2013

[Teradata.Aster 2013] TERADATA.ASTER: *Teradata Aster Big Analytics Appliance*. http://www.asterdata.com/product/big-analytics-appliance.php. Version: 2013. – Abgerufen am 19.10.2013

[Versino AG 2013] VERSINO AG: *SAP Business One und SAP HANA*. http://www.versino.de/de/sap-loesungen/sap-business-one/sap-hana-inmemory-computing/sap-hana-abbildungen.htm. Version: 2013. – Abgerufen am 19.10.2013

[Vertica 2013a] VERTICA: *HP Vertica Analytics Platform 7*. http://www.vertica.com/wp-content/uploads/2013/11/HP_7_Whats_New_Data_Sheet.pdf. Version: 2013. – Abgerufen am 17.12.2013

[Vertica 2013b] VERTICA: *Native BI, ETL & Hadoop/MapReduce Integration*. http://www.vertica.com/the-analytics-platform/native-bi-etl-and-hadoop-mapreduce-integration/. Version: 2013. – Abgerufen am 17.12.2013

[Vertica 2013c] VERTICA: *Scale-Out MPP Architecture*. http://www.vertica.com/the-analytics-platform/scale-out-mpp-architecture/. Version: 2013. – Abgerufen am 17.12.2013

[VLBA 2012a] VLBA: *Forschung - Very Large Business Applications*. http://vlba.wi-ol.de/22203.html. Version: Februar 2012. – Abgerufen am 31.01.2014

[VLBA 2012b] VLBA: *Projektgruppen*. http://vlba.wi-ol.de/20261.html. Version: April 2012. – Abgerufen am 31.01.2014

[Volitich 2008] VOLITICH, Dan: *IBM Cognos 8 Business Intelligence: The Official Guide*. Bd. 1. Mcgraw-Hill Professional, 2008

[Volkswagen AG 2011] VOLKSWAGEN AG: *Produktkatalog BI-CC*. 2011

[Volkswagen AG 2012] VOLKSWAGEN AG: *BI-Strategie*. Version 4.1. 2012

[Volkswagen AG 2013] VOLKSWAGEN AG: *Simulation für Capacity Management*. 2013

[Volkswagen AG 2014] VOLKSWAGEN AG: *Der Konzern*. http://www.volkswagenag.com/content/vwcorp/content/de/the_group.html. Version: 2014. – Abgerufen am 25.02.2014

[Walker-Morgan 2010] WALKER-MORGAN, D. J. ; HEISE (Hrsg.): *NoSQL im Überblick*. http://www.heise.de/open/artikel/NoSQL-im-Ueberblick-1012483.html. Version: 2010. – Abgerufen am 02.01.2013

[Wolf 2012] WOLF, Cathrin: *SAP HANA unterstützt Kunden beim Scale-Out*. http://de.news-sap.com/2012/05/16/sap-hana-unterstutzt-kunden-beim-scale-out/. Version: 2012. – Abgerufen am 19.10.2013

[XING 2013] XING: *Unternehmensprofil Hewlett-Packard*. https://www.xing.com/companies/hewlett-packard. Version: 2013. – Abgerufen am 17.12.2013

Abschließende Erklärung

Wir versichern hiermit, dass wir unseren Abschlussbericht selbständig und ohne fremde Hilfe angefertigt haben, und dass wir alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlegenden Ausführungen unserer Arbeit besonders gekennzeichnet und die Quellen zitiert haben.

Oldenburg, den 19. Mai 2014	Jad Asswad
Oldenburg, den 19. Mai 2014	Björn Berger
Oldenburg, den 19. Mai 2014	Jonas Borrmann
Oldenburg, den 19. Mai 2014	Simon Brosig
Oldenburg, den 19. Mai 2014	Sören Flathmann
Oldenburg, den 19. Mai 2014	Marco Lucht
Oldenburg, den 19. Mai 2014	Catharina Petry
Oldenburg, den 19. Mai 2014	Benjamin Reinecke
Oldenburg, den 19. Mai 2014	Yvonne Tevs
Oldenburg, den 19. Mai 2014	Christoph Walther
Oldenburg, den 19. Mai 2014	Jan-Hendrik Wege
Oldenburg, den 19. Mai 2014	Jan Patrick Weiß