



Joachim Dorschel *Hrsg.*

Praxishandbuch Big Data

Wirtschaft – Recht – Technik



Springer Gabler

Praxishandbuch Big Data

Joachim Dorschel
Herausgeber

Praxishandbuch Big Data

Wirtschaft – Recht – Technik



Springer Gabler

Herausgeber

Joachim Dorschel
Karlsruhe, Deutschland

ISBN 978-3-658-07288-9
DOI 10.1007/978-3-658-07289-6

ISBN 978-3-658-07289-6 (eBook)

Die Deutsche Nationalbibliothek verzeichnetet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Springer Gabler

© Springer Fachmedien Wiesbaden 2015

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Ein-speicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen.

Lektorat: Anna Pietras, Sylvia Meier

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Springer Fachmedien Wiesbaden GmbH ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Inhaltsverzeichnis

1 Einführung	1
Werner Dorschel und Joachim Dorschel	
1.1 Keynote: „Die Zeit ist reif für Big Data“	1
1.2 Einführung und Überblick	5
1.2.1 Definitionen	6
1.2.1.1 Volume	7
1.2.1.2 Velocity	7
1.2.1.3 Variety	8
1.2.1.4 Veracity	8
1.2.2 Perspektiven	8
1.2.2.1 Gesellschafts- und rechtspolitische Sicht	9
1.2.2.2 Ökonomische Sicht	9
1.2.2.3 Technische Sichtweise	10
1.2.3 Gegenstand dieses Handbuchs	11
Literatur	13
2 Wirtschaft	15
Joachim Dorschel, Werner Dorschel, Ulrich Föhl, Wilhelmus van Geenen, Dieter Hertweck, Martin Kinitzki, Philipp Küller, Carsten Lanquillon, Hauke Mallow, Lothar März, Fouad Omri, Sigurd Schacht, Alphonse Stremler und Elke Theobald	
2.1 Datenorientierung statt Bauchentscheidung: Führungs- und Organisationskultur in der datenorientierten Unternehmung	15
2.1.1 Unternehmerische Sinnhaftigkeit von Big Data Entscheidungen	17
2.1.2 Fakten erhöhen die Verantwortlichkeit der einzelnen Akteure	22
2.1.3 Kreativität der Mitarbeiter als Teil einer Big Data freundlichen Unternehmenskultur	23
2.1.4 Informations- und Kommunikationskompetenz und Verände- rungskompetenz als Basis schneller Reaktionszeiten	26

2.1.5	Führung wird komplexer und bedarf der Unternehmensmodellierung, sowie des aktiven Managements der Unternehmensarchitektur	28
2.1.6	Zusammenfassung: Tipps für Entscheider, die es bei der Einführung einer Datengetriebenen Entscheidungskultur zu beachten gibt	30
2.2	Enterprise Architecture Management und Big Data	32
2.2.1	Enterprise Architecture Management und Big Data	32
2.2.1.1	EAM ein kurzer Überblick	33
2.2.1.2	Competitive Advantage durch Big Data	38
2.2.2	EAM als Ausgangspunkt für die Etablierung von Big Data im Unternehmen	41
2.2.2.1	Einführung und Entwicklung einer Unternehmensarchitektur	42
2.2.2.2	Einführung von Big Data unter besonderer Beachtung der Unternehmensarchitektur	51
2.2.3	Fazit	54
2.3	Advanced Analytics mit Big Data	55
2.3.1	Begriffsdefinitionen und Varianten	55
2.3.1.1	Analyse und Analytics	55
2.3.1.2	Analytics-Varianten	56
2.3.1.3	Analytics trifft auf Big Data	63
2.3.2	Analyseaufgaben	63
2.3.2.1	Prädiktive Analyseaufgaben	64
2.3.2.2	Beschreibende Analyseaufgaben	66
2.3.3	CRISP-DM: Ein Prozessmodell für Analyseprozesse	68
2.3.3.1	Business Understanding	70
2.3.3.2	Data Understanding	71
2.3.3.3	Data Preparation	71
2.3.3.4	Modeling	71
2.3.3.5	Evaluation	73
2.3.3.6	Deployment	73
2.3.4	Big Data Analytics: Was ist anders?	74
2.3.4.1	Einfluss der Daten auf den Analyseprozess	74
2.3.4.2	Technologische Aspekte	83
2.3.4.3	Der Mensch im Unternehmen als Einflussfaktor	86
2.3.5	Zusammenfassung und Ausblick	88
2.4	Simulation: Neue Einsatzfelder durch Big Data	89
2.4.1	Einführung	89
2.4.2	Planungsablauf in der Fahrzeugindustrie	90
2.4.3	Herausforderungen an die Planung	93
2.4.3.1	Erhöhung der Planungsgenauigkeit	93

2.4.3.2	Einsatz der Simulation in der Planung	94
2.4.3.3	Simulationsgestützte Planung	96
2.4.3.4	Erhöhte Datenanforderungen	97
2.4.4	Praxisbeispiel Automobilendmontage	98
2.4.4.1	Zielsetzung der Anwendung	98
2.4.4.2	Ablauf einer Anwendung	100
2.4.4.3	Datenanforderungen	101
2.4.5	Fazit und Ausblick	103
2.5	Big Data-Analysen: Anwendungsszenarien und Trends	104
2.5.1	Big Data-Analysen: Anwendungsszenarien	105
2.5.1.1	Marketing und Vertrieb	105
2.5.1.2	Forschung und Entwicklung	106
2.5.1.3	Kundenservice	107
2.5.1.4	Produktion	107
2.5.1.5	Logistik	107
2.5.1.6	IT	109
2.5.1.7	Risikomanagement	110
2.5.2	Big Data-Analysen: Trends	110
2.5.2.1	Trends im Rechtswesen	110
2.5.2.2	Trends im Transportwesen	111
2.5.3	Trends im Sozialen Sektor	111
2.5.4	Trends im Gesundheitswesen	112
2.6	Big Data wird zu Smart Data – Big Data in der Marktforschung	112
2.6.1	Big Data in der Marktforschung – Goldgrube oder Datengrab? .	112
2.6.2	Der Marktforschungsprozess bei Big Data	114
2.6.2.1	Die Forschungsfrage	114
2.6.2.2	Das Forschungsdesign	115
2.6.2.3	Die Erhebungsphase: Die Nadel im Heuhaufen	116
2.6.3	Aktuelle Herausforderungen für den Big Data Einsatz in der Marktforschung	117
2.6.3.1	Datenzugänglichkeit und Repräsentativität	118
2.6.3.2	Herausforderung Text Mining und Social-Media-Analyse	118
2.6.3.3	Pluralität der Meinungen	119
2.6.3.4	Interpretation multimedialer Daten	120
2.6.3.5	Der Kontext macht den Unterschied	120
2.6.3.6	Von Korrelationen und Kausalitäten	121
2.6.3.7	Topaktuell und doch Schnee von gestern	121
2.6.4	Die Zukunft von Big Data in der Marktforschung	122
2.7	Big Data und Electronic Commerce – Neue Erkenntnisse zur Customer Journey	123
2.7.1	Einleitung	123
2.7.2	Aktuelle Themen im E-Commerce	123

2.7.3	Daten und Datenstrukturen	125
2.7.4	Umfassende Verhaltensanalyse im Rahmen der Customer Journey	127
2.7.4.1	Bedarfs-/Mangelerkennung	128
2.7.4.2	Suche	130
2.7.4.3	Bewertung	130
2.7.4.4	Kauf und Nachkaufphase	131
2.7.5	Wie aus „Big Data“ „Smart Data“ wird	132
2.8	Big Data in der Kreditwirtschaft	134
2.8.1	IT in der Kreditwirtschaft	134
2.8.1.1	Abgrenzung	134
2.8.1.2	Mainframe, Batch, Dialog und Multichannel	134
2.8.1.3	Legacy-Systeme und Standardisierung	135
2.8.1.4	Core-Banking-Systeme und Fachanwendungen	135
2.8.1.5	Datenverwaltung, IDV und Business Intelligence	135
2.8.1.6	Aktuelle Herausforderungen	136
2.8.2	Big Data bewegt die Bank-IT	137
2.8.2.1	Digitalisierung der Kundenbeziehung	137
2.8.2.2	Transparenzanforderungen durch die Bankenaufsicht	138
2.8.3	Einzelne Geschäftsbereiche	139
2.8.3.1	Zahlungsverkehr	139
2.8.3.2	Handel	141
2.8.3.3	Kreditgeschäft	142
2.8.3.4	Gesamtbanksteuerung	144
2.8.3.5	Vertrieb und Multichannel Services	145
2.8.4	Big Data, Outsourcing und Cloud Computing	146
2.8.4.1	Gefahr der Datendesintegration	146
2.8.4.2	Managed Services für Big Data in der Cloud	147
2.8.5	Fazit	147
2.9	Chancen und Herausforderungen von Big Data in der Industrie	148
2.9.1	Unternehmerische Ziele zur Erhöhung der Wertschöpfung	148
2.9.1.1	Anforderungen in Produktion und Logistik	148
2.9.2	Effizienzsteigerung durch integriertes Realtime-Informations- und Datenmanagement in der integrierten Supply Chain	149
2.9.3	Ein Modell der Produktion	150
2.9.4	Leistungssteuerung in Echtzeit für maximale Reaktivität der Supply Chain	150
2.9.5	Ebenen und Stufen der Planung	152
2.9.6	Daten als Schlüsselfaktor des unternehmerischen Erfolges	154
2.9.6.1	Kundenindividuelle Produkte und Leistungen konfigurieren	154
2.9.6.2	Transparenz schaffen	154
2.9.6.3	Reaktionsfähigkeit erhöhen	155

2.9.6.4 Entscheidungen durch Lösungsvorschläge unterstützen	156
2.9.6.5 Neue Produktionskonfigurationen und Produkt-einführungen durch Szenarien absichern	156
2.9.7 Erfolgsfaktoren zum Ausschöpfen der Potenziale von Big Data	158
2.9.7.1 Umgang mit Daten	158
2.9.7.2 Technologien	158
2.9.7.3 Analysetechniken und Algorithmen	159
2.9.7.4 Datenzugriff	160
2.9.7.5 Organisationale Transformation und Führung	160
2.9.8 Fazit	160
Literatur	161
3 Recht	167
Michael Bartsch, Olaf Botzem, Thorsten Culmsee, Joachim Dorschel, Jenny Hubertus, Carsten Ulbricht und Thorsten Walter	
3.1 Datenschutz	167
3.1.1 Prinzipien des Datenschutzrechts	167
3.1.1.1 Einleitung	167
3.1.1.2 Prinzipien des Datenschutzrechts	168
3.1.1.3 Fazit	173
3.1.2 Gesetzliche Erlaubnistatbestände und Interessenabwägung	174
3.1.2.1 Anwendungsbereiche und Abgrenzungen von TMG, TKG und BDSG	175
3.1.2.2 Der Legitimationstatbestand der Einwilligung	175
3.1.2.3 Weitere Befugnisse zur Datenverarbeitung nach TMG und TKG	176
3.1.2.4 Weitere Befugnisse zur Datenverarbeitung nach dem TKG	177
3.1.2.5 Weitere Befugnisse zur Datenverarbeitung nach dem BDSG	178
3.1.3 Anonymisierung und Pseudonymisierung; Verschlüsselung	185
3.1.3.1 Anonymisierung	186
3.1.3.2 Pseudonymisierung	187
3.1.3.3 Verschlüsselung	188
3.1.4 Technologien zur Umsetzung datenschutzrechtlicher Anforderungen	190
3.1.5 Zulässigkeit einzelner Phasen von Big Data-Analysen	190
3.1.5.1 Erhebung von Big Data	191
3.1.5.2 Speichern von Big Data	191
3.1.5.3 Personenbezogene Auswertung von Big Data	191
3.1.5.4 Auswertung von Big Data	192
3.1.5.5 Veröffentlichen von Big Data	192

3.1.5.6 Zusammenfassung	193
3.1.6 Betroffenenrechte	193
3.1.6.1 Benachrichtigung des Betroffenen	194
3.1.6.2 Benachrichtigungspflicht bei Web-Crawling und Screen-Scraping?	195
3.1.6.3 Auskunftsanspruch des Betroffenen	196
3.1.6.4 Korrekturrechte	197
3.1.6.5 Das „Recht auf vergessen werden“	198
3.1.7 Internationale Datenverarbeitung	199
3.1.7.1 Anwendbares Recht	199
3.1.7.2 Voraussetzungen für die rechtskonforme Datenverarbeitung in der EU	201
3.1.7.3 Voraussetzungen für die rechtskonforme Datenverarbeitung in Drittstaaten	201
3.1.7.4 Praxisfall Cloud Computing	203
3.1.7.5 Zusammenfassung	204
3.1.8 Big Data in der Personalabteilung	205
3.1.8.1 Einführung	205
3.1.8.2 Daten, Daten und noch mehr Daten	205
3.1.8.3 Problemstellung	206
3.1.8.4 Zusammenfassung	210
3.1.9 Automatisierte Entscheidungen und Scoring	211
3.1.9.1 Automatisierte Einzelentscheidungen	211
3.1.9.2 Scoring	212
3.2 Leistungsschutz	213
3.2.1 Urheberrecht an Daten	213
3.2.1.1 Internationales Urheberrecht	214
3.2.1.2 Urheberrechtliche Schutzfähigkeit von Informationen und Daten	214
3.2.1.3 Urheberrechtlicher Schutz der Einzeldaten	216
3.2.1.4 Urheberrechtlicher Schutz von computergenerierten Werken	216
3.2.1.5 Urheberrechtlicher Schutz von Sammel- oder Datenbankwerken	216
3.2.2 Schutz des Datenbankherstellers	217
3.2.2.1 Der Begriff der Datenbank	218
3.2.2.2 Der Begriff des Datenbankherstellers	218
3.2.2.3 Die Rechte des Datenbankherstellers	219
3.2.2.4 Schranken des Rechts des Datenbankherstellers	220
3.2.3 Unlautere gezielte Mitbewerberbehinderung	222
3.2.4 Sonstige Leistungsschutzrechte	224
3.2.4.1 Schutz des Presseverlegers	224

3.3	Integritätschutz	225
3.3.1	Strafrechtlicher Schutz der Datenintegrität	225
3.3.1.1	Sachbeschädigung (§ 303 StGB)	226
3.3.1.2	§ 303 a Datenveränderung	226
3.3.1.3	Computersabotage (§ 303 b)	228
3.3.1.4	§ 202 a Ausspähen von Daten	228
3.3.1.5	§ 202 b Abfangen von Daten	229
3.3.1.6	§ 202 c Vorbereiten des Ausspähens und Abfangens von Daten	230
3.3.1.7	Ausblick	231
3.3.2	Zivilrechtlicher Schutz: Daten als absolut geschützte Rechtsgüter	231
3.3.2.1	Daten auf eigenen Datenspeichern	232
3.3.2.2	Daten als absolut geschützte Rechtsgüter	233
3.3.2.3	Ansprüche aus Schutzgesetzen	235
3.3.2.4	Rechtsfolgen	235
3.4	Reglementierung der Erhebung von Big Data	237
3.4.1	Rechtliche Bewertung des Screen-Scraping	237
3.4.2	Technische Schutzmaßnahmen	238
3.4.2.1	IP-Sperren	238
3.4.2.2	Captcha	239
3.4.3	Zusammenfassung	239
3.5	Anwendungsszenarien	240
3.5.1	Auswertung des Nutzungsverhaltens im Internet	240
3.5.2	Social Media Analysen	241
3.5.3	Big Data in der Industrie (Industrie 4.0)	242
3.5.4	Zusammenfassung	244
3.6	Verträge über Daten und Datenanalysen	245
3.6.1	Wichtige Vertragstypen	246
3.6.1.1	Kaufverträge über Daten	246
3.6.1.2	Zeitlich begrenzte Datennutzung	246
3.6.1.3	Aufträge zur Datenanalyse	247
3.6.1.4	Datenerhebung im Auftrag	247
3.6.1.5	Datenspeicherung im Auftrag	248
3.6.2	Leistungsstörungen	249
3.6.3	Auftragsdatenverarbeitung	249
Literatur	251
4	Technik	255
Gernot Fels, Carsten Lanquillon, Hauke Mallow, Fritz Schinkel und Christian Schulmeyer		
4.1	Grenzen konventioneller Business-Intelligence-Lösungen	255
4.1.1	Business Intelligence: Ein Überblick	255

4.1.1.1	Verwendung und Definitionen des Begriffs	255
4.1.1.2	Evolution entscheidungsunterstützender Systeme	256
4.1.1.3	Diskussion um das Analysespektrum	257
4.1.1.4	BI-Referenzarchitektur	258
4.1.2	Grenzen von BI-Lösungen im Kontext von Big Data	260
4.1.2.1	Volume	260
4.1.2.2	Velocity	261
4.1.2.3	Variety	262
4.1.2.4	Veracity	262
4.1.3	Zusammenfassung	263
4.2	Big Data-Lösungen	263
4.2.1	Anforderungen an Big Data-Lösungen	263
4.2.2	Big Data-Referenzarchitekturen	263
4.2.2.1	Funktionale Big Data-Referenzarchitektur	263
4.2.2.2	Erweiterung einer Data-Warehouse-Architektur mit Big Data-Technologien	275
4.2.3	Zusammenfassung und Ausblick	277
4.3	IT-Infrastrukturen für Big Data	278
4.3.1	Herausforderungen an die Infrastruktur	278
4.3.2	Verteilte Parallelverarbeitung großer Datenbestände	279
4.3.2.1	Apache Hadoop	279
4.3.2.2	Reale oder virtuelle Server?	288
4.3.3	NoSQL-Datenbanken	288
4.3.3.1	Key-Value Stores	290
4.3.3.2	Beispiel: Key-Value Store mit Produktinformationen	290
4.3.3.3	Dokument-orientierte Datenbanken (Document Stores)	291
4.3.3.4	Spaltenorientierte Datenbanken (Columnar Stores)	291
4.3.3.5	Graph-Datenbanken (Graph Databases)	292
4.3.4	In-Memory-Technologien	294
4.3.4.1	In-Memory-Datenbanken (IMDB)	295
4.3.4.2	In-Memory Data Grids (IMDG)	296
4.3.5	Verarbeitung großer Ereignisströme	297
4.3.6	Referenzarchitektur für Big Data-Infrastrukturen	300
4.3.7	Lambda-Architektur	301
4.3.7.1	Impala	303
4.3.7.2	Storm	304
4.3.8	Betrieb von Big Data-Infrastrukturen	305
4.3.8.1	IaaS, PaaS, SaaS oder sogar Data Science als Service?	306
4.4	Big Data-Analyse auf Basis technischer Methoden und Systeme	307
4.4.1	Herausforderungen an Big Data-Analyse	307
4.4.1.1	Was sind Big Data aus technischer Sicht?	307
4.4.1.2	Abgrenzung zu BI	307

4.4.1.3 Datenmengen	308
4.4.1.4 Heterogenität der Datenquellen und der Datenformate sowie fehlende Beschreibung	309
4.4.2 Daten	309
4.4.2.1 Unstrukturierte und semistrukturierte Daten	309
4.4.2.2 Text und nicht-Text-Formate (Audio, Video, Grafik, Bilder)	310
4.4.2.3 Multilinguale Daten	311
4.4.2.4 Datenzugriff	311
4.4.3 Systemische Grundlagen	312
4.4.3.1 Indexerstellung	312
4.4.3.2 In Memory Computing	312
4.4.3.3 MapReduce	313
4.4.3.4 Skalierbarkeit	313
4.4.4 Methoden	314
4.4.4.1 Suche ist nicht gleich Suche	314
4.4.4.2 Keywordbasierte Suche	315
4.4.4.3 Linguistik und Semantik	316
4.4.4.4 Wissensmodelle, Taxonomien und Ontologien	317
4.4.4.5 Assoziative Methoden der Suche	319
4.4.4.6 Case Based Reasoning (CBR)	320
4.4.4.7 Mischformen/Kombinationen	321
4.4.5 Zeitlicher Aspekt	322
4.4.5.1 Retrospektive Analysen	322
4.4.5.2 Echtzeitanalysen	322
4.4.6 Erkenntnisziele der Big Data-Analyse	323
4.4.6.1 Datengold	323
4.4.6.2 Vorhersagen	324
4.4.6.3 Schwache Signale	325
4.4.6.4 Neue Erkenntnisse (knowing the unknown unknown) . .	327
4.4.6.5 Relationen/Verknüpfung von Daten	328
4.4.7 Zusammenfassung	328
Literatur	329
Sachverzeichnis	331

Mitarbeiterverzeichnis

Die Autoren

Prof. Dr. Michael Bartsch ist Rechtsanwalt und Hochschullehrer.

Er studierte Jura und Literaturwissenschaften und wurde 1976 zur Rechtsanwaltschaft in Karlsruhe zugelassen. Er unterrichtet seit 1984 IT-Recht an der Universität Karlsruhe und seit 1995 Urheber- und Medienrecht an der Staatlichen Hochschule für Gestaltung in Karlsruhe.

Bartsch gehört zu den Begründern des IT-Rechts in Deutschland und war auf diesem Gebiet in zahlreichen Funktionen tätig, insbesondere bei der Deutschen Gesellschaft für Recht und Informatik (DGRI) und der Fachzeitschrift „Computer und Recht“.

Mit mehr als 60 Veröffentlichungen zum IT-Recht gehört er zu den bedeutendsten Autoren auf diesem Gebiet.

Die Kanzlei Bartsch Rechtsanwälte berät mittelständische Unternehmen und Unternehmer im Wirtschaftsrecht. Die Kanzlei verbindet exzellente juristische Expertise mit technischem Verständnis. Zu den Mandanten der Kanzlei zählen insbesondere technologieorientierte Unternehmen und deren Vertragspartner.

Die Kanzlei verfügt über ein kleines, gut aufeinander eingespieltes Team. In ihren Fachgebieten arbeiten die Anwälte der Kanzlei hoch spezialisiert und bietet bei fachgebietsübergreifenden Fragestellungen Rechtsberatung aus einer Hand. In Gebieten, die die Kanzlei nicht abdeckt, arbeitet sie mit einem starken Netzwerk aus Spezialisten zusammen. Gleichermaßen gilt bei Fragestellungen, die ausländische Rechtsordnungen berühren.

Bartsch Rechtsanwälte arbeiten mit wissenschaftlichem Anspruch und stellen dies durch Lehraufträge an Universitäten und Hochschulen und durch zahlreiche Veröffentlichungen unter Beweis.

Olaf Botzem ist Rechtsanwalt in der Kanzlei Bartsch Rechtsanwälte.¹ Im Rahmen seiner Tätigkeit berät Olaf Botzem nationale und internationale Mandanten in allen Fragen des IT-Rechts und des Gesellschaftsrechts. Im Bereich des IT-Rechts liegt sein Schwerpunkt auf der Gestaltung der Verträge für internetbasierte Geschäftsmodelle. Bereits in seinem Masterstudium an der Kingston University London befasste sich Olaf Botzem mit den internationalen Aspekten des Urheber- und Datenschutzrechts. Als Lehrbeauftragter an der Dualen Hochschule Baden-Württemberg und Dozent referiert Olaf Botzem in den Bereichen Urheber-, E-Commerce- und Datenschutzrecht.

Thorsten Culmsee studierte Jura an der Albert-Ludwigs-Universität Freiburg mit den Schwerpunkten Gewerblicher Rechtsschutz und Urheberrecht und absolvierte einen Masterstudiengang im Kommunikationsrecht an der Université Panthéon-Assas (Paris 2). Er arbeitet seit 2011 als Rechtsanwalt in der Kanzlei Bartsch Rechtsanwälte² mit den Schwerpunkten Medien- und IT-Recht. Er ist zertifizierter Datenschutzbeauftragter (TÜV Süd).

Thorsten Culmsee berät seine Mandanten in allen Fragen des IT- und Medienrechts. Ein Schwerpunkt seiner Tätigkeit liegt in der Vertragsgestaltung, in der datenschutzrechtlichen Beratung und im Führen gerichtlicher Auseinandersetzungen. Als Referent des Onlineseminaranbieters TeleLex, einem Gemeinschaftsunternehmen des Otto Schmidt Verlages und der Datev, hält er Webinare zum IT-Vertragsrecht.

Werner Dorschel ist Gründungsgesellschafter und Geschäftsführer der DPS Engineering GmbH.

Werner Dorschel begleitet seit vielen Jahren die Entwicklung der IT-Industrie und verantwortet zahlreiche anspruchsvolle Großprojekte in vielfältigen Themen der Kreditwirtschaft und des Handels.

Die DPS Engineering GmbH ist ein Software- und Consultinghaus für die Finanzwirtschaft und den Handel. Mit den Schwerpunkten Solutions, Application Management, Projects und Consulting arbeitet die DPS Engineering GmbH seit 25 Jahren erfolgreich für große Adressen des Marktes.

Durch Konzentration auf die wesentlichen bankfachlichen Kernthemen der Kreditwirtschaft und die Optimierung des Cash Managements des Handels ist DPS Engineering GmbH für die beiden Marktsegmente ein hoch spezialisierter Lösungspartner. Langjährige erfolgreiche Kundenbeziehungen sind Nachweis für die Leistungsfähigkeit des Unternehmens.

DPS Engineering GmbH wird je nach Sachgebiet sowohl als Produktanbieter wie auch als Projektpartner für individuelle Lösungen tätig.

Gernot Fels ist Principal Product Marketing Manager bei Fujitsu.

Nach dem Studium der Informatik an der Universität Karlsruhe war Gernot Fels in verschiedenen Funktionen in den Bereichen Training, Beratung, Vertrieb, Business Development und Marketing bei Siemens, Siemens Nixdorf Computers, Fujitsu Siemens

¹ Mehr über Bartsch Rechtsanwälte erfahren Sie im Autorenprofil von Prof. Dr. Michael Bartsch.

² Mehr über Bartsch Rechtsanwälte erfahren Sie im Autorenprofil von Prof. Dr. Michael Bartsch.

Computers und Fujitsu tätig. Dabei kann er mittlerweile auf eine über 30-jährige Erfahrung in der IT-Branche zurückblicken. Aktuell ist er verantwortlich für das Marketing innovativer IT-Infrastruktur-Themen, sowie der damit verbundenen Lösungen und Dienstleistungen.

Fujitsu ist der führende japanische Anbieter von Informations- und Telekommunikations-basierten (ITK) Geschäftslösungen und bietet eine breite Palette an Technologieprodukten, -lösungen und -dienstleistungen, das von Endgeräten über Rechenzentrumslösungen, Betriebs- und Wartungsservices und Cloud-Lösungen bis hin zum Outsourcing reicht.

Mit über 160.000 Mitarbeitern betreut das Unternehmen Kunden in mehr als 100 Ländern.

FUJITSU entwickelt und fertigt in Deutschland Notebooks, Desktops, Thin Clients, Server, Speichersysteme sowie Mainboards und betreibt mehrere hochsichere Rechenzentren. Mit 8000 Channel-Partnern in Deutschland verfügt Fujitsu zudem über eines der leistungsfähigsten Partnernetzwerke der Branche.

Sämtliche für Big Data relevanten Infrastrukturkonzepte werden von Fujitsu unterstützt: Hadoop-Cluster als integriertes System aus Hardware und Software oder Cloud-Lösung mit optionalem Analytic-Framework, eine Complex Event Processing Engine und IMDB-Appliances auf Basis SAP HANA. Somit kann abhängig von Situation und Anforderungen stets der geeignete Technologiemix eingesetzt werden, um für den Kunden die optimale Lösung zu finden. Die für die Infrastruktur erforderlichen Produkte, wie Server, Speichersysteme, Netzkomponenten und Zugangsgeräte werden ebenfalls von Fujitsu bereitgestellt. Big Data-Infrastrukturen von Fujitsu beinhalten Software und Middleware von Fujitsu selbst, aus der Open Source-Welt und von Partnern wie bspw. SAP oder Software AG. Sie sind aber auch offen für Produkte führender ISVs. Ebenso wichtig wie Produkte und Infrastrukturkonzepte ist das Serviceangebot von der Prozess- und Infrastrukturerberatung, über das Design der optimalen Infrastruktur, die Implementierung, die Integration in die bestehende IT-Landschaft bis hin zum ganzheitlichen Support. Für all diese Leistungen werden auch attraktive Finanzierungsoptionen angeboten. Darüber hinaus ist Fujitsu bekannt für seine flexiblen Sourcing-Modelle. Kundenspezifische Lösungen können vom Kunden entweder in Eigenregie oder von Fujitsu betrieben und verwaltet werden. Big Data Services werden auch aus der Fujitsu Cloud bereitgestellt.

Prof. Dr. Ulrich Föhl ist Professor für psychologische Marktforschung an der Hochschule Pforzheim.

In der Lehre vertritt er insbesondere Werbe- und Konsumentenpsychologie sowie empirische Forschungsmethoden und Statistik. Zudem leitet er als Studiendekan den Studiengang Betriebswirtschaftslehre/Media Management und Werbepsychologie. Vor seiner Hochschultätigkeit war er zehn Jahre in der Automobilbranche tätig. Dort entwickelte und evaluierte er Benutzerschnittstellen für Telematiksysteme und leitete

Konsumentenforschungsprojekte zu Fahrzeuginnovationen insbesondere in den Bereichen User Experience und Design.

Bereits während seiner Zeit in der Automobilbranche ging er in zahlreichen Projekten der Frage nach, wie sich aus komplexen Fahrzeugdaten Rückschlüsse auf Konsumentenbedürfnisse ableiten lassen und wie diese für die Weiterentwicklung unterschiedlicher Fahrzeugkomponenten genutzt werden können. Sein Schwerpunkt im Bereich Big Data liegt insbesondere in der Datenanalyse auf Basis von Modellannahmen aus den Sozial- und Verhaltenswissenschaften.

Prof. Dr. Dieter Hertweck studierte Soziologie-, Politik- und Wirtschaftswissenschaften und promovierte im Bereich Wirtschaftsinformatik bei Prof. Dr. Helmut Krcmar an der Universität Hohenheim. Er arbeitete in verschiedenen Forschungsinstituten, wie dem Fraunhofer IAO (Prof. Bullinger), dem Forschungszentrum Informatik (FZI) am KIT Karlsruhe (Prof. Stucky), sowie der Australian Graduate School of Management Sydney (Prof. Yetton). Seit 2004 lehrt er Geschäftsprozess- und IT-Management im Studiengang Wirtschaftsinformatik an der Hochschule Heilbronn.

Seit 2005 ist er leitender Direktor des Electronic-Business-Instituts Heilbronn (EBI). 2007 gewann er den mit 20.000 Euro dotierten KREATEK-Innovationspreis des Landes Baden-Württemberg. Seit 2009 ist er ERASMUS Visiting Professor an der Cracow University of Economics. 2012 war Prof. Hertweck Visiting Professor an der School of IT an der Bond University/Australien, wo er regionale Innovationssysteme im Mittelstand beforschte. 2013 war Prof. Hertweck Gewinner des IHK-Forschungstransferpreises der Region Heilbronn-Franken, mit dem seine Forschung im Bereich IT-Services für den Mittelstand gewürdigt wurde. Hertweck ist Beirat im Digital Business Cloud-Magazin, Juror der IT-Mittelstandsinitiative, sowie Editor der wissenschaftlichen Fachzeitschriften Journal of International Information Management, und Journal of Information Technology for Development.

Jenny Hubertus studierte Rechtswissenschaften an der Universität des Saarlandes mit dem Schwerpunkt „Deutsches und internationales Vertrags- und Wirtschaftsrecht“. Während des Referendariats und in der Folgezeit war Jenny Hubertus in einer auf Markenrecht und gewerblichen Rechtsschutz spezialisierten Kanzlei schwerpunktmäßig im Informationstechnologie- und Telekommunikationsrecht tätig.

Seit 2014 verstärkt Jenny Hubertus das Team von Bartsch Rechtsanwälte³ in Stuttgart. Sie besetzt die Schnittstelle zwischen dem klassischen Gesellschaftsrecht (Unternehmensgründungen, Beteiligungsmodelle, Gesellschafterstreitigkeiten) und der weiterführenden Firmenberatung (e-Commerce, Datenschutz und Web 2.0) und betreut vorwiegend technologieorientierte Jung-Unternehmen bei der Gestaltung und Umsetzung internetbasierter Geschäftsmodelle.

Neben ihrer Anwaltstätigkeit ist Jenny Hubertus als Referentin und Dozentin im Bereich des Internetrechts und der Sozialen Medien tätig.

³ Mehr über Bartsch Rechtsanwälte erfahren Sie im Autorenprofil von Prof. Dr. Michael Bartsch.

Martin Kinitzki begann nach einer Banklehre im Jahr 2012 das Studium der Wirtschaftsinformatik mit den Schwerpunkten IT-Management sowie Entwicklung webbasierter und mobiler IT-Systeme an der Hochschule Heilbronn. Seit zwei Jahren ist er in der Forschungsgruppe von Prof. Hertweck als wissenschaftliche Hilfskraft am Electronic Business Institut (EBI) Heilbronn beschäftigt. Die ersten Berührungen mit Big Data-Methoden erfolgten im Rahmen des Wirtschaftsinformatikstudiums an der Hochschule Heilbronn bei Prof. Lanquillon.

Seit 2013 forscht Martin Kinitzki sehr intensiv im Bereich Service Automatisierung, Smart Services und quantitative Methoden für international agierende IT-Dienstleister. Seine erste Publikation zum Thema Arbeitsplatz und Rechenzentrum der Zukunft erschien im Rahmen des 30-jährigen Unternehmensjubiläums im Kundenmagazin update der Bechtle AG.

Phillip Küller ist studierter Wirtschaftsinformatiker und arbeitet seit 2010 als wissenschaftlicher Mitarbeiter am Electronic Business Institut der Hochschule Heilbronn. Er koordinierte den wissenschaftlichen Teil des Projekts INNOTRAIN IT und leitet derzeit die Projekte KonfIT-SSC und ITSM4SME.

Neben seiner wissenschaftlichen Expertise bringt Herr Küller umfangreiche praktische Erfahrung aus der IT mittelständischer Unternehmen mit, kennt aber durch seine Tätigkeit für die SAP AG in Walldorf auch die Softwareindustrie gut.

Aktuell promoviert Philipp Küller am Lehrstuhl von Prof. Dr. Krcmar an der Technischen Universität München und befasst sich hierbei mit Geschäftsmodellen und Unternehmensarchitekturen in der Energiewirtschaft.

Prof. Dr. Carsten Lanquillon ist Professor für Wirtschaftsinformatik mit Schwerpunkt Business Intelligence und quantitative Methoden.

Seit mehr als 15 Jahren beschäftigt er sich sowohl in der Praxis als auch in der Theorie erfolgreich mit der Entwicklung und Anwendungen von Data-Mining-Lösung und der Durchführung von Datenanalysen für vielfältige fachliche Fragestellungen aus unterschiedlichen Fachbereichen in großen und kleinen Unternehmen.

Neben seiner Professur ist Herr Lanquillon als Berater und Referent für Themen insbesondere aus den Bereichen Big Data, Advanced Analytics und Business Intelligence tätig.

Zusammen mit Herrn Prof. Dr. Sigurd Schacht gründete und betreibt Herr Lanquillon das Data Science & Analytics Lab an der Hochschule Heilbronn, das sich zum Ziel gesetzt hat, Big-Data-Technologien zur Steigerung des Unternehmenswertes zu erforschen und in die Industrie zu transferieren.

Hauke Mallow Hauke Mallow ist Big Data Lead-Architekt bei der T-Systems, mit über 18 Jahren Praxiserfahrung im Business-Intelligence- und Data-Warehousing-Umfeld. Seit 2012 berät Hauke Mallow Kunden unterschiedlichster Branchen, ihre Data Warehouse-Landschaften mit Big Data-Technologien auf neue Anforderungen, wie z. B. das Internet der Dinge, zukunftsfähig auszurichten. Technologisch liegt sein Fokus auf den Aspekten der Integration von klassischen Data Warehouse-Architekturen mit

Technologien des Hadoop-Ökosystems und den Veränderungen, die Big Data im Bereich Analytics mit sich bringt.

Im Zuge der Weiterentwicklung des Leistungsportfolios der T-Systems, gestaltete er die Definition, den Aufbau und die konkrete Ausgestaltung der Big Data-Kompetenz entscheidend mit.

Über T-Systems: Mit einer weltumspannenden Infrastruktur aus Rechenzentren und Netzen betreibt T-Systems die Informations- und Kommunikationstechnik (engl. kurz ICT) für multinationale Konzerne und öffentliche Institutionen. Auf dieser Basis bietet die Großkundensparte der Deutschen Telekom integrierte Lösungen für die vernetzte Zukunft von Wirtschaft und Gesellschaft. Rund 50.000 Mitarbeiter verknüpfen bei T Systems Branchenkompetenz mit ICT-Innovationen, um Kunden in aller Welt spürbaren Mehrwert für ihr Kerngeschäft zu schaffen. Im Geschäftsjahr 2013 erzielte die Großkundensparte einen Umsatz von rund 9,5 Milliarden Euro.

Dr. Ing. Lothar März ist anerkannter Experte auf dem Themengebiet der ereignis-diskreten Simulation in Produktion und Logistik.

Er verfügt über langjährige Führungs- und Beratungserfahrung u. a. bei Fraunhofer IPA, Dürr Schenck Engineering und LOM Innovation.

Als Chief Operations Officer des umsetzungsorientierten Beratungshauses STREMLER AG verknüpft er ganzheitliches Supply Chain-Wissen mit maßgeschneiderten, innovativen IT-Lösungen.

Mit dem Einsatz der von ihm entwickelten Planungs- und Steuerungsplattform LOM.Cubes® in industriellen Prozessen werden selbst dynamische Systeme innerhalb weniger Sekunden analysiert und hochkomplexe Wirkzusammenhänge transparent dargestellt. So können Unternehmen optimal und flexibel auf ihre Kunden- und Marktanforderungen reagieren.

Fouad ben Nasr Omri ist Wissenschaftler am der Karlsruher Institut für Technologie (KIT) und Unternehmer.

Als Gründer der Beratungs- und „Predictive Analytics“-SaaS-Firma Golden Bayes, betreut er international zahlreiche Unternehmen bei dem strategischen Einsatz von Big Data Analysen und Cloud Computing Infrastrukturen. Fouad ben Nasr Omri ist auch Gründer des ersten Cloud-Infrastructure-Provider in Africa, die Firma Safozi.

Als Wissenschaftler am KIT, erforscht er neue Verfahren für die Bewertung der Zuverlässigkeit sicherheitskritischer IT-Infrastrukturen. Er entwickelt neue mathematische Methoden für die effiziente und skalierbare Analyse von Big Data. Fouad ben Nasr Omri hat seine wissenschaftlichen Ergebnisse auf zahlreichen internationalen Konferenzen eingebracht. Er ist auch Gutachter bei mehreren renommierten Konferenzen und Zeitschriften.

Prof. Dr. Sigurd Schacht ist Professor für Wirtschaftsinformatik insbesondere Betriebswirtschaftslehre an der Hochschule Heilbronn.

Seit mehr als 10 Jahren beschäftigt sich Sigurd Schacht sowohl in der Praxis als auch in der Theorie mit betriebswirtschaftlichen Datenanalysen und der Prüfung von SAP

Systemen bei Unternehmen unterschiedlicher Größenordnung. Sein Schwerpunkt liegt hier vor allem in der Einhaltung von Regularien und der automatisierten Durchführung von Kontrollen im internen Kontrollsyste m mittels IT Unterstützung.

Ergänzend zu seiner Tätigkeit als Professor ist Herr Schacht als Berater und Referent tätig. Im Rahmen dieser Tätigkeit begleitete Herr Schacht mehrere Datenanalyseprojekte bei internationalen Konzernen im In- und Ausland.

Zusammen mit Herrn Prof. Dr. Carsten Lanquillon gründete und betreibt Herr Schacht das Data Science & Analytics Lab an der Hochschule Heilbronn, das sich zum Ziel gesetzt hat, Big Data Technologien zur Steigerung des Unternehmenswertes zu erforschen und in die Industrie zu transferieren.

Dr. Fritz Schinkel ist Director Innovation bei Fujitsu.⁴

Er ist als Program-Manager für Cloud-Infrastrukturen und Big Data verantwortlich für innovative Lösungsangebote. Seit 1991 arbeitet er in der IT-Branche in unterschiedlichen Entwicklungs-, Architektur- und Management-Funktionen unter anderem auf den Gebieten Compiler, CASE-Tools, Web-und Portal-Technologie, dynamisches Datacenter-Management, Cloud-Plattform, Private-Cloud- und Big Data-Lösungen. Über 20 Jahre Erfahrung in der IT-Branche sammelte er bei den Firmen Siemens, Nixdorf, Fujitsu Siemens und Fujitsu. Seinen Doktortitel in Reiner Mathematik erhielt Fritz Schinkel von der Leibniz-Universität, Hannover.

Dr. Christian Schulmeyer ist einer von vier geschäftsführenden Gesellschaftern der Empolis Information Management GmbH sowie vereidigter und öffentlich bestellter Sachverständiger für IT und Multimedia. Als Chief Technology Officer entwickelt er Softwaresysteme zur semantischen Verarbeitung und Analyse großer unstrukturierter Datenmengen, die bei internationalen Unternehmen und öffentlichen Institutionen zum Einsatz kommen.

Sein besonderes Augenmerk liegt auf den Herausforderungen, die sich die IT im Rahmen von Industrie 4.0/Internet of Things und dem Konzept der smarten vernetzten Fabrik stellen muss. Auf diesem Weg ins vierte industrielle Zeitalter engagiert sich Dr. Christian Schulmeyer sehr stark im Dialog mit allen beteiligten Akteuren aus Wirtschaft, Wissenschaft und Verbänden.

Dipl.-Ing. Alphonse Stremler ist Unternehmer und Spezialist auf dem Gebiet Unternehmensentwicklung durch Optimierung integrierter Wertschöpfungsketten. Er leitet als Inhaber das Beratungsunternehmen STREMLER AG.

Aus seiner langjährigen Industriearbeit als Geschäftsführer/Vorstand bei der Zumtobel Licht GmbH, Techno Saarstahl GmbH und Klöckner-Humboldt-Deutz AG kennt Alphonse Stremler die Bedürfnisse des Marktes für umsetzungsorientierte Beratung im Bereich Supply Chain- und Prozessoptimierung. Seit 20 Jahren realisiert die STREMLER AG europaweit Supply Chain Engineering Projekte in den verschiedensten Branchen, bei Konzernen und mittelständischen Unternehmen.

⁴ Mehr über Fujitsu erfahren Sie im Autorenprofil von Gernot Fels.

Dabei ist ein Schlüssel zum Erfolg die Vernetzung der Supply Chain Prozesse und die Planung über mehrere Ebenen durch sekundenschnelle IT-Echtzeitsysteme. Mit dem bedarfsgerechten Datenmanagement ermöglicht sie die Flexibilisierung der Planung und höchste Reaktionsfähigkeit eines Unternehmens auf die Anforderungen des Marktes.

Prof. Dr. Elke Theobald absolvierte ihr Studium der Computerlinguistik, Philosophie und Wirtschaftsinformatik in Heidelberg und Mannheim. Weiterhin war sie als Projektleiterin für Multimedia- und Online-Produkte ab 1992 bei der Langenscheidt KG in München tätig. Ab 1994 war sie Marketingleiterin und Etat Directorin in New Media Agenturen. Seit 1998 ist sie Professorin für computergestützte Medien an der Hochschule Pforzheim und Prodekanin der Fakultät für Wirtschaft und Recht. Ihre Forschungs- und Beratungsschwerpunkte liegen in Online-Marketing, eBranding und Marketing Intelligence. Mit der in ihrem Institut entwickelten Software MANAGEMENT MONITOR führt sie Marketing Intelligence Systeme in zahlreichen international tätigen Unternehmen ein.

Dr. Carsten Ulbricht ist auf Internet und Social Media spezialisierter Rechtsanwalt und für den Standort Stuttgart verantwortlicher Partner der Kanzlei Bartsch Rechtsanwälte⁵ mit den Schwerpunkten IT-, Urheber- und Datenschutzrecht. Seine Schwerpunkte liegen dabei auf der rechtlichen Prüfung internetbasierter Geschäftsmodelle, derzeit vor allem im Bereich Cloud Services und Big Data.

Neben diversen Engagements als Vortragsredner und Dozent berichtet Dr. Ulbricht seit dem Jahr 2007 regelmäßig in seinem Weblog zum Thema „Internet, Social Media & Recht“ unter www.rechtzweinull.de nicht nur über neueste Entwicklungen in Rechtsprechung, Diskussionen in der Literatur und über eigene Erfahrungen, sondern analysiert auch entsprechende Geschäftsmodelle und -projekte auf ihre rechtlichen Erfolgs- und Risikofaktoren.

In seinem Ende Oktober 2013 in 2. Auflage erschienenen Buch „Social Media & Recht – Praxiswissen für Unternehmen“ fasst Dr. Ulbricht die wichtigsten rechtlichen Fragen in einem Praxisratgeber zusammen. Das im Haufe-Verlag erschienene Werk beschreibt dabei die verschiedenen rechtlichen Implikationen, die Unternehmen im Rahmen der Umsetzung einer abgesicherten Social Media Strategie beachten sollten.

Wilhelmus van Geenen ist Geschäftsführer und Gesellschafter der DPS Engineering GmbH⁶ und verantwortet dort unter Anderem den Bereich Anwendungsentwicklung. Nach seiner Ausbildung zum Bankkaufmann studierte Wilhelmus van Geenen Wirtschaftsmathematik an der Universität Karlsruhe. Er verfügt über langjährige Erfahrung in IT-Großprojekten in der Kreditwirtschaft. Seine besondere Expertise liegt in den Bereichen Zahlungsverkehr, Multichannel-Banking und Wertpapier-Services.

⁵ Mehr über Bartsch Rechtsanwälte erfahren Sie im Autorenprofil von Prof. Dr. Michael Bartsch.

⁶ Mehr über die DPS Engineering GmbH erfahren Sie im Autorenprofil von Werner Dorschel.

Thorsten Walter ist Rechtsanwalt und Fachanwalt für Arbeitsrecht.

Als Partner der Kanzlei Bartsch Rechtsanwälte⁷ betreut er seit Jahren mittelständische Unternehmen und Führungskräfte in allen Fragen des individuellen und kollektiven Arbeitsrechts.

Schwerpunkt seiner Tätigkeit ist die Schnittstelle zwischen Arbeits- und IT-Recht. In den Bereichen Arbeitnehmerdatenschutz, Compliance und Social Media berät Thorsten Walter Unternehmen bei der Umsetzung rechtlicher Anforderungen und ihrer Berücksichtigung in Betriebsvereinbarungen, Richtlinien und Verträgen und begleitet die Verhandlungen mit dem Betriebsrat. Er betreut zahlreiche Mandanten bei der Implementierung von Regelungen über die Nutzung unternehmenseigener IT und „Bring your own device“.

Der Herausgeber

Joachim Dorschel ist Rechtsanwalt und Unternehmer.

Als Partner der Kanzlei Bartsch Rechtsanwälte⁸ betreut er seit Jahren Mandanten der IT-Branche in Fragen des IT-Rechts.

Als Geschäftsführer und Gesellschafter des Systemhauses DPS Engineering GmbH⁹, realisiert er anspruchsvolle IT-Projekte für die Branchen Finance and Retail.

Hinzu kommen zahlreiche Aktivitäten als Publizist und Dozent im Bereich des Medien-, Internet- und Provider-Rechts.

Die juristische Perspektive von Big Data verfolgt Joachim Dorschel seit mehreren Jahren aus der rechtlichen Beratung und durch fachwissenschaftliche Aufsätze und Vorträge. Auf wirtschaftlich-technischer Seite war Joachim Dorschel an mehreren Big Data-Projekten der DPS Engineering GmbH für namenhafte Kreditinstitute beteiligt.

⁷ Mehr über Bartsch Rechtsanwälte erfahren Sie im Autorenprofil von Prof. Dr. Michael Bartsch.

⁸ Mehr über Bartsch Rechtsanwälte erfahren Sie im Autorenprofil von Prof. Dr. Michael Bartsch.

⁹ Mehr über DPS Engineering GmbH erfahren Sie im Autorenprofil von Werner Dorschel.

Einführung

Werner Dorschel und Joachim Dorschel

1.1 Keynote: „Die Zeit ist reif für Big Data“

Werner Dorschel

Was ist Big Data?

Die Frage „Was ist Big Data“ wurde in zahlreichen Publikationen höchst unterschiedlich, in der Regel unzureichend und damit unzutreffend beantwortet. Dem steht per se die Vielschichtigkeit des Themas entgegen. Deshalb stellen wir diesem Handbuch keinen weiteren Versuch voran, Big Data als eindimensionales Phänomen zu definieren. Wesentlicher Anspruch wird es sein, sich aus unterschiedlichsten Perspektiven dem Thema Big Data zu nähern.

Die IT-Industrie und die IT-affine Community war schon immer kreativ bei der Hervorbringung von „mächtigen“ Begriffsbildungen zur Benennung von technologischen Umbruchphasen. In der Regel werden Insider-Technologien dadurch quasi populärwissenschaftlich in die breite Öffentlichkeit getragen. Dies gilt im besonderen Maße für Big Data.

Spätestens mit dem World Wide Web hat die Nomenklatur der IT-Industrie die hermeneutische Welt der IT-Spezialisten hinter sich gelassen und wurde Teil einer nicht nur IT-interessierten Öffentlichkeit. Stark gefördert durch die New Economy, bei der IPOs von Internet-Unternehmen mit visionären technologischen Innovationen das Interesse des allgemeinen Publikums fand – nicht zuletzt beflogt von Aktienkurs-Phantasien.

Werner Dorschel
Stuttgart, Deutschland

Joachim Dorschel ✉
Karlsruhe, Deutschland
e-mail: jd@bartsch-rechtsanwaelte.de

Big Data ist in diesem Sinne im besonderen Maße eine äußert facettenreiche Erscheinung im technischen, kommerziellen, rechtlichen und sozialen Kontext.

Neu ist mit Big Data die explizite Assoziation mit soziokulturellen und politökonomischen Veränderungen der Geschäftsmodelle der öffentlichen und privaten Unternehmungen und der sich rasant verändernden sozialen und ökonomischen Bedingungen aller Marktteilnehmer.

Denn eine bedeutende Facette des Phänomens **Big Data** ist das exponentiell zunehmende Misstrauen der breiten Öffentlichkeit in den entwickelten Industriestaaten gegenüber der mit **Big Data** assoziierten „digitalen Macht“ großer IT-Konzerne und staatlicher Institutionen. Spätestens mit der NSA-Affäre in den Jahren 2013/2014 erscheint manchen **Big Data** als technologische Fundierung eines Orwell'schen Big Brother-Ansatzes. **Big Data** insinuiert geradezu eine Namensanalogie.

Diese Technologie-, Politik-, und Grundwertedebatte soll dem öffentlichen bzw. veröffentlichten Diskurs vorbehalten bleiben und wird von den Autoren dieses Handbuchs nicht vertieft.

Allerdings würde es der öffentlichen politischen Diskussion um das Phänomen Big Data substantiell sehr nützlich sein, sich mit den in den nachfolgenden Kapiteln diskutierten technischen, wirtschaftlichen und rechtlichen Rahmenbedingungen von Big Data zu beschäftigen. Ohne eine angemessene Bemühung um die Mehrdimensionalität des Begriffes **Big Data** bleiben viele Publikationen eher romantische Belletristik als denn ein sachdienlicher Beitrag zur Klärung offensichtlicher Veränderungen der IT-Industrie.

Warum dieses Handbuch?

Ohne Zweifel ist **Big Data** ein technologisches und kommerzielles Phänomen und ein Indiz für einen Paradigmenwechsel in der Nutzung von zunehmenden Datenmengen durch alle Marktteilnehmer.

Die häufig zu lesende Trivialerklärung von Big Data als die Verarbeitung von großen Datenmengen – dies in großer Geschwindigkeit und semi- bzw. unstrukturierter Vielfalt – wirft einen äußerst unzureichenden Blick auf die relevanten Zusammenhänge.

Natürlich ist die bereits im vielzitierten Mooreschen Gesetz beschriebene dramatische Kapazitätserweiterung von modernen Computersystemen eine infrastrukturelle Notwendigkeit zur Bearbeitung von schnell wachsenden Datenmengen. Jedoch subsummiert sich unter dem Kunstbegriff Big Data eine Vielfalt von

- Technologien,
- Analytischen Methoden,
- Modellierungs- und Designverfahren,
- Kommerzielle Konzepte,
- Rechtliche Rahmenbedingungen.

Nur durch eine Gesamtschau des Phänomens **Big Data** können Nutzer und Entscheider als Mitgestalter des technologischen und wirtschaftlichen Wandels **Big Data** zielführend einordnen und ggf. nutzen.

Deshalb haben Herausgeber und Autoren dieses Handbuchs ein Kaleidoskop der mit den genannten Aspekten verbundenen Fragestellungen in ihren Beiträgen aufgefächert.

Big Data: Aspekte eines Phänomens

Um dem Leser dieses Handbuchs **Big Data** eine Vorstellung von den inhaltlichen Schwerpunkten der Autorenbeiträge zu geben, folgen hier einige Anmerkungen zu Aspekten des Themas.

Die Gesamtschau erweist sich damit sehr schnell als eine Dekonstruktion eines Kunstbegriffes, der sich letztlich in durchaus heterogene fachliche Einzelbetrachtungen zerlegen muss.

Big Data hat keine Skala

Die bereits zitierte Trivialdefinition von **Big Data** greift zu kurz. Die Verarbeitung großer Datenmengen an sich ist kein neues Phänomen. Losgelöst von der Eigenschaft der Strukturiertheit legt die technologische Infrastruktur von Big Data die Voraussetzung für eine nicht transaktionale aber analytische Sicht der Daten. In der Weiterentwicklung konventioneller Business Intelligence (BI)-Verfahren soll Big Data eine neue Dimension der gegenwarts- und zukunftsbezogenen Analyse von beliebigen Datenmengen unterstützen. Die in tera, peta und ggf. noch höheren Dimensionen zu betrachtenden Datenmengen werden durch den technologischen Wandel der verwendeten Speicher- und Prozessortechnologie deutlich ausgeweitet. Ab welcher Dimension Daten zu **Big Data** werden, ist nicht objektiviert.

Big Data verändert die IT-Industrie

Die grundlegenden Technologien von **Big Data** sind in der IT-Community über Open Source-Komponenten schon seit Jahren verfügbar. **Big Data** existierte also schon lange bevor es in den öffentlichen Raum eintrat. So kann mit Hadoop, MapReduce und kooperierenden Werkzeugen und Methoden die verteilte Speicherung von NoSQL-Daten horizontal skaliert und effektiv verarbeitet werden. Unternehmen wie Google und Amazon haben dies eindrucksvoll in der operativen Praxis genutzt und die Validität dieser technologischen Infrastrukturen und Methoden bewiesen. Das Verständnis dieser verteilten Datenhaltungs- und Verarbeitungsstrukturen ist elementar um zu einer sinnhaften Einordnung von **Big Data** zu kommen.

Big Data und Datenbankmanagement-Systeme

Zur Herstellung einer operativen Arbeitsplattform werden **Big Data**-Konstrukte durch hoch performante Datenbanksysteme für Analyse und sonstige Applikationen bereitgestellt. Diese Datenbanksysteme folgen anderen Effizienz- und Performancekriterien als konventionelle relationale Datenbanken. Durch Nutzung von In-Memory-Technologie

und spaltenorientierter Speicherungs- und Zugriffstechnik entstehen massive Performancesteigerungen und damit Verarbeitungsgeschwindigkeiten, insbesondere für On-Demand-Applikationen. Realtime-Beauskunftung wird dadurch in kommerziellen, industriellen und weiteren Anwendungs-Environments ermöglicht.

Dabei wird mit Interesse zu beobachten sein, wie die drei maßgeblichen Datenbankwettbewerber IBM, ORACLE und SAP ihre Marktpositionen über Big Data-affine Datenbankmanagement-Systeme neu bestimmen werden.

Big Data ist auch eine analytische Methode

Big Data Analytics umfasst Methoden, Verfahren und Werkzeuge zur Konfiguration und Modellierung von vielfältigen Analyseprozessen zu Vergangenheit, Gegenwart und Zukunft. Diese Verfahren können einerseits für naheliegende Optimierungen industrieller Prozesse, in der Entwicklung von interaktiven Spielen, sozialen Netzwerken bis hin zu nachrichtendienstlichen Zwecken eingesetzt werden.

Bei Nutzern von **Big Data Analytics**-basierten Projekten haben sich mittlerweile arbeitsteilige Berufsbilder zur Durchführung von Big Data-Vorhaben herausgebildet.

Big Data als Motor des Wettbewerbs

Betrachtet man die kommerziellen Aspekte bei der Nutzung von **Big Data** erweist sich die schnelle Nutzung von Markt-, Kunden- und Nutzerdaten zunehmend als wichtiger Wettbewerbsfaktor.

Die kunden- bzw. personenindividuelle Ansprache von potenziellen Nachfragern vielfältiger Produkte wird für den Internethandel nahezu überlebenswichtig. Die Nutzung von smarten und mobilen Devices ist ohne **Big Data** für die Produktanbieter nicht mehr denkbar. Diese Digitalisierung der Produktions- und Absatzprozesse ist der kommerziell entscheidende Mehrwert des **Big Data**.

Die Wettbewerbsverschärfung ist nicht nur kennzeichnend für Unternehmen in nationalen und supranationalen Märkten. Der Technologie- und Marktvorsprung – insbesondere amerikanischer Internet-Konzerne – erweist sich dabei zunehmend als limitierend für die Erhaltung des schnellen und freien Marktzugangs neuer Anbieter.

Insofern induziert **Big Data** bedeutende Impulse für die Neudeinition von Produktions- und Absatzmärkten. Man denke hier an das Eindringen von Google und Co. in die klassischen Finanzmärkte.

Wendet man den Blick hin zu der industriellen Implementierung von **Big Data** Systemen z. B. zur Steuerung und Überwachung von hochdifferenzierten Fertigungs- und Produktionssystemen ist die Installation von Big Data ein Muss im Sinne der Modernisierungs- und Innovationsstrategie.

Big Data bedarf erweiterter rechtlicher Rahmenbedingungen

Mit Ausbreitung des World Wide Web für kommerzielle und soziale Nutzungen ergaben sich in den letzten Jahren vielfältige neue Rechtsprobleme, die durch die Gesetzgeber und die Rechtsprechung nicht immer nur erfolgreich gelöst wurden. Man denke an Ur-

heberrecht, Fragen des Datenschutzes oder die Beauskunftung von Datentransfer und Nutzerverhalten.

Neue Verfahren der kommerziellen Nutzung von Outsourcing und ASP-Services – dies zumal im grenzüberschreitenden Datenverkehr – stellen bereits heute gesteigerte Anforderungen an eine angepasste Rechtslage dar.

Vielfältige Rechtsgebiete einschließlich Arbeits- und Wettbewerbsrecht werden tangiert.

Setzt man nun **Big Data** mit dem ebenso prominenten **Cloud Computing** (auch ein mächtiger Begriff der IT-Community) in Verbindung, potenzieren sich die damit drohenden Verletzungen von Rechtsgütern.

Welche Leser profitieren von diesem Handbuch Big Data?

Wie bereits dargestellt, will dieses Handbuch keinen weiteren Beitrag zum politischen Diskurs über die Gefahren von **Big Data** für die Freiheit des Bürgers und des Wettbewerbs sein. Auch ist es kein Manual für Nerds.

Die Intention dieses Buches ist die Information von Lesern mit professionellen Interessen. Es richtet es sich an Verantwortliche und Entscheidungsträger der IT-Industrie und deren Umfeld, die sich einen schnellen Überblick über die wesentlichen technischen, wirtschaftlichen und rechtlichen Aspekte des Phänomens Big Data verschaffen wollen. Dabei sind als Leser durchaus auch Experten in den genannten Teilgebieten willkommen, die sich ihrerseits über die komplementären Fragestellungen von **Big Data** kurSORisch informieren wollen.

Da dieses Handbuch natürlich nicht als vollständiges Curriculum des Phänomens **Big Data** konzipiert ist und verstanden werden will, kann die Lektüre der unterschiedlichen Autorenbeiträge selektiv erfolgen. In der Regel führen die einzelnen Beiträge in ihre Themen top-down ein. So kann also jeder Leser seinen Detaillierungslevel bei der Nutzung des Handbuchs **Big Data** für sich selbst herausfinden.

Ich wünsche den Lesern einen aufschlussreichen, nutzbringenden und letztlich doch auch spannenden Zugang zum Phänomen **Big Data**.

1.2 Einführung und Überblick

Joachim Dorschel

Wie in der Keynote zutreffend ausgeführt, lässt sich die Frage „Was ist Big Data“ nicht in einem Satz beantworten – jedenfalls nicht so, dass alle Aspekte dieses Phänomens Berücksichtigung finden.

Es handelt sich bei Big Data weder um eine Technologie noch um ein Geschäftsmodell. Gleichwohl gibt es Technologien und Geschäftsmodelle, die unter den Begriff Big Data subsumiert werden.

Es wäre aber falsch, aus der Schwierigkeit einer einfachen begrifflichen Abgrenzung und Kategorisierung den Schluss zu ziehen, Big Data sei nicht mehr als ein Modebegriff und eine inhaltsleere Floskel. Denn die wirtschaftliche und technische Nutzung von Big Data, wie sie etwa Internetkonzerne wie Amazon und Google praktizieren, sind technisch und wirtschaftlich real (ausführlich hierzu Feinleib 2014, 4 ff.).

Es gibt wirtschaftliche, technische, politische, soziologische und rechtliche Sachverhalte und Forschungsfragen, denen das Label Big Data angeheftet wird. Je nach Zielsetzung und Fragestellung wird der Begriff Big Data anders definiert und abgegrenzt.

Dieses Handbuch will dem vorhandenen Kaleidoskop treffender und wenig treffender Definitionen keine weitere hinzufügen. Es ist auch nicht der Anspruch dieses Handbuchs, das Phänomen Big Data neu oder besser oder zum ersten Mal richtig zu erklären. Die einzelnen Beiträge analysieren verschiedene – aus Sicht der Autoren besonders bedeutsame – Aspekte von Big Data in der Unternehmenspraxis. Die Autoren nehmen dabei unterschiedliche Sichtweisen ein, was zwangsläufig dazu führt, dass der eine Aspekt mehr, der andere weniger betont wird. Soweit zur Eingrenzung des Betrachtungsgegenstands eine Abgrenzung des Begriffs Big Data erforderlich ist, wird auf die gängigen Definitionen zurückgegriffen, die in der Regel auf Alliterationen des Buchstabens „V“ basieren (hierzu sogleich unter Abschn. 1.2.1).

Man kann Big Data aus drei Perspektiven betrachten: gesellschaftspolitisch, ökonomisch und technisch (hierzu im Folgenden unter Abschn. 1.2.2). Dieses Handbuch will den Leser beim Einsatz von Big Data im Unternehmen unterstützen. Die gewählte Perspektive ist somit allein die ökonomische. Technische und rechtliche Fragen sind aus dieser Perspektive Aufgaben, die gelöst werden müssen.

Diese Herangehensweise will nicht ignorieren, dass mit dem Phänomen Big Data schwierige gesellschafts- und rechtspolitische Fragen verbunden sind, die auch die Möglichkeiten und Grenzen der Datennutzung im Unternehmen betreffen. Diese mit der ihnen gebührenden Tiefe zu betrachten, würde den Rahmen dieses Handbuchs sprengen. Die in dem Handbuch gegebenen Empfehlungen stehen daher unter der Prämisse, dass jedweder Einsatz von Big Data den heute in Deutschland geltenden Gesetzen genügen muss, deren rechtspolitischer und rechtsökonomischer Sinn oder Unsinn im Einzelfall nicht hinterfragt werden soll.

1.2.1 Definitionen

Es gibt unzählige Definitionen von Big Data (eine gute Auswahl bietet King 2014, S. 34 ff.). Die gängigsten Definitionen weisen Big Data drei oder vier spezifische Eigenschaften zu: Volume, Velocity, Variety und Veracity und werden gemeinhin als die drei oder vier V's bezeichnet (vgl. u. a. Kudyba und Kwatinetz 2014, S. 2 ff.; Page et al. 2012, S. 11; Schroeck et al. 2012, S. 4 ff.).

Kritiker halten diese Definition für zu Informatik-orientiert und wollen über alternative Definitionen (z. B. mittels drei F's: fast, flexibel und focused, vgl. Freytag 2014, S. 100) die nutzerorientierte Sicht in den Vordergrund stellen.

Für die Zwecke eines praxisorientierten Handbuchs haben die vier V's neben ihrer weiten Verbreitung in der fachwissenschaftlichen Literatur den großen Vorteil, dass sie Charakteristika von Big Data betonen, aus denen sich unmittelbare technische und funktionale Anforderungen an Big Data-Anwendungen und Projekte ergeben. Die Autoren des vorliegenden Handbuchs legen daher, wo erforderlich, diese Definition zugrunde.

1.2.1.1 Volume

Abhandlungen zum Thema Big Data zitieren häufig das Moore'sche Gesetz (vgl. Moore 1965). Sehr vereinfacht verdoppelt sich hiernach die Rechenleistung von IT-Systemen innerhalb eines bestimmten Zeitraums, der je nach Interpretation zwischen 12 und 24 Monaten angegeben wird. Diese vor 50 Jahren formulierte Regel gilt bis heute.

Eine Folge ist, dass immer kleinere und immer leistungsfähigere Computerchips in immer mehr Lebensbereiche vordringen, dort Steuerungsaufgaben übernehmen und dabei digitale Daten erzeugen. Dabei nimmt auch der Grad der Vernetzung ständig zu. Unter dem Begriff „Internet der Dinge“ wird das Phänomen beschrieben, bei dem Computer zu unsichtbaren Bestandteilen von Geräten und Alltagsgegenständen werden und unbemerkt menschliche Tätigkeiten unterstützen.

Das exponentielle Wachstum der Rechenkapazitäten steht in unmittelbarer Relation zu dem weltweit vorhandenen Datenvolumen. Studien zufolge verdoppelt dieses sich alle zwei Jahre (Rudolph, Linzmajer 2014, S. 13).

Dieses Phänomen gilt als eine der Grundlagen von Big Data: Es stehen immer mehr Daten zur Analyse zur Verfügung, die Aussagen über immer neue Lebensbereiche zulassen.

Allerdings sollte „Volume“ nicht als notwendige oder hinreichende Bedingung einer einzelnen Big Data-Anwendung verstanden werden. Werden in einem Forschungsprojekt Literaturquellen, Beiträge aus fachwissenschaftlichen Foren, technische Messdaten und dergleichen maschinell verarbeitet, handelt es sich auch dann um Big Data, wenn das in Byte gemessene Datenvolumen wegen der Spezialisierung des Themas nicht allzu groß ist. Umgekehrt ist die Verwaltung strukturierter Daten in einer relationalen Datenbank als solche keine Big Data-Anwendung, auch wenn die Datenbank mehrere Petabyte umfasst.

1.2.1.2 Velocity

Das Merkmal „Velocity“ wird nicht einheitlich interpretiert. Teilweise wird hierunter die Geschwindigkeit verstanden, mit der neue Daten entstehen (Fasel 2014, S. 389; McAfee und Brynjolfsson 2014, S. 7). Teilweise wird auch von der Geschwindigkeit gesprochen, mit der Daten produziert und verändert werden müssen (King 2014, S. 35). Wieder andere beziehen dieses Merkmal auf die Anforderungen, die Big Data an die Verarbeitungsgeschwindigkeit von IT-Systemen stellt (Bendler et al. 2014, S. 279).

Die Merkmale Velocity und Volume stehen in unmittelbarer Wechselbeziehung. Je schneller gerechnet wird, desto mehr Daten werden in immer kürzerer Zeit produziert. Zugleich steht das Merkmal Velocity für die kurze Halbwertzeit des Erkenntniswertes von Daten. Die durch digitale Daten beschriebenen Sachverhalte ändern sich immer schneller,

sodass dem vorhandenen Datenbestand neue oder veränderte Daten hinzugefügt werden müssen.

1.2.1.3 Variety

Das Merkmal „Variety“ kennzeichnet die Heterogenität der Datenquelle und Datenformate. IT-Anwendungen in verschiedenen Lebensbereichen können Daten zu höchst unterschiedlichen Zwecken speichern. Dementsprechend vielfältig sind die möglichen Dateninhalte und Datenformate. Das denkbare Spektrum reicht von technischen Messdaten über Social Media-Inhalte bis zu Video-Streams.

Viele der für Big Data relevanten Daten sind unstrukturiert, passen also nicht in ein vordefiniertes Datenmodell. Ein wichtiges Beispiel sind hier Texte und Audio-Daten in natürlicher Sprache.

Übergreifende Erkenntnisse aus unterschiedlichen Quellen und Formaten zu gewinnen, ist eine der Kernaufgaben von Big Data. Dabei geht Big Data über Data-Warehousing und Business Intelligence-Konzepte hinaus, bei denen unterschiedliche Daten über ETL-Prozesse in eine strukturierte Form gebracht und danach analysiert werden. Ziel von Big Data ist es, dass der Computer die Dateninhalte selbst versteht und in der Lage ist, diese zu interpretieren.

1.2.1.4 Veracity

Veracity wird in einigen Big Data-Definitionen als weiteres Merkmal hinzugefügt (vgl. u. a. Bendler et al. 2014, S. 279). Dieses Merkmal bezieht sich auf die Richtigkeit, Vollständigkeit und Verlässlichkeit der Dateninhalte. Es ist kennzeichnend für Big Data-Anwendungen, auch solche Daten einzubeziehen, deren objektiver Erkenntniswert nicht sicher messbar ist. Prominentestes Beispiel sind hier Social Media-Daten. User-generierte Texte in sozialen Netzwerken sind geprägt von subjektiven Empfindungen und unterschiedlichen zeitlichen und inhaltlichen Kontexten. Eine zentrale Aufgabe bei der Nutzung von Big Data ist es, diese Faktoren bei der Planung, Durchführung und Bewertung von Analysen zu berücksichtigen.

1.2.2 Perspektiven

In der Beschäftigung mit Big Data lassen sich drei Fragestellungen unterscheiden:

Welche Gefahren drohen der Gesellschaft und dem Einzelnen, wenn digitale Daten immer präzisere und detailliertere Aussagen und Prognosen über Individuen erlauben?

Welchen Nutzen können Wirtschaft und Gesellschaft aus dem vorhandenen Datenschatz ziehen?

Welche technischen Möglichkeiten und Grenzen gibt es bei der Verarbeitung von Big Data?

1.2.2.1 Gesellschafts- und rechtspolitische Sicht

Die gesellschafts- und rechtspolitische Auseinandersetzung mit dem Phänomen Big Data war in der Vergangenheit wesentlich geprägt durch den so genannten „NSA-Skandal“, also die anlass- und unterschiedslose Speicherung jedweder verfügbarer Daten durch Geheimdienste. Im Kern geht es hier um die Frage, welche Gefahren das in Daten manifestierte Wissen über die Lebensdetails eines Menschen in der Hand eines anderen für diesen Menschen birgt.

Das Bundesverfassungsgericht führt in seinem Volkszählungsurteil vom 15.12.1983 hierzu aus:

Wer nicht mit hinreichender Sicherheit überschauen kann, welche ihn betreffende Informationen in bestimmten Bereichen seiner sozialen Umwelt bekannt sind, und wer das Wissen möglicher Kommunikationspartner nicht einigermaßen abzuschätzen vermag, kann in seiner Freiheit wesentlich gehemmt werden, aus eigener Selbstbestimmung zu planen oder zu entscheiden. Mit dem Recht auf informationelle Selbstbestimmung wären eine Gesellschaftsordnung und eine diese ermöglichte Rechtsordnung nicht vereinbar, in der Bürger nicht mehr wissen können, wer was wann und bei welcher Gelegenheit über sie weiß. Wer unsicher ist, ob abweichende Verhaltensweisen jederzeit notiert und als Information dauerhaft gespeichert, verwendet oder weitergegeben werden, wird versuchen, nicht durch solche Verhaltensweisen aufzufallen

(Bundesverfassungsgericht, Urteil v. 15.12.1983 – 1 BvR 209/83).

Diese Entscheidung wird auch als Geburtsstunde des deutschen Datenschutzrechts bezeichnet (Schröder 2012, Kapitel 1.1.a). Das heute geltende Bundesdatenschutzgesetz aus dem Jahr 1995 basiert in wesentlichen Teilen auf den Befunden dieser Gerichtsentscheidung.

An der Systematik des Bundesdatenschutzgesetzes zeigen sich auch die zwei Dimensionen der Daten-induzierten Gefahrenlage: Es geht um personenbezogene Daten sowohl in der Hand staatlicher Autoritäten als auch privater Unternehmen. Wie real und eng verbunden beide Bedrohungsszenarien mit dem Phänomen Big Data sind, zeigen die Überwachungspraxis westlicher, demokratisch kontrollierter Gemeindienste auf der einen und die Datennutzung durch die vor allem US-amerikanisch geprägte Internetwirtschaft auf der anderen Seite.

Eng verbunden mit der Diskussion um Big Data ist die Kritik an intelligenten, selbstlernenden Algorithmen, mit Hilfe derer immer häufiger Entscheidungen durch Maschinen getroffen werden (siehe hierzu insbesondere Hofstetter 2014).

1.2.2.2 Ökonomische Sicht

Aus einer anderen Perspektive lässt sich Big Data als Chance und Möglichkeit begreifen, bessere und schnellere Entscheidungen zu treffen. Aus dieser Perspektive ist das wachsende Datenvolumen ein Rohstoff, der nutz- und gewinnbringend eingesetzt werden kann.

Der Branchenverband Bitkom hat bereits im Jahr 2012 einen Leitfaden vorgelegt, der die Einsatzbeispiele von Big Data aufzeigt.¹

¹ [http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online\(1\).pdf](http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online(1).pdf).

Dieser zeigt vielfache Einsatzszenarien für Big Data-Anwendungen in Unternehmen unterschiedlicher Branchen. Studien zufolge hat ein überwiegender Anteil der deutschen Unternehmen bereits erste Big Data-Projekte umgesetzt oder ist damit beschäftigt, Möglichkeiten und Nutzungsszenarien zu prüfen (Gluchowski 2014, S. 401 ff.).

Inwieweit heute bereits der Nachweis geführt werden kann, dass Big Data die Ertrags- oder Wettbewerbssituation von Unternehmen tatsächlich verbessert, ist nicht sicher. Eine Umfrage des MIT zeigte, dass Unternehmen, die sich selbst als „datengetrieben“ einschätzen, tendenziell bessere finanzielle und betriebliche Kennzahlen aufweisen (McAfee und Brynjolfsson 2014, S. 8).

In der fachlichen Auseinandersetzung mit den Potenzialen von Big Data lassen sich zwei Themenstellungen unterscheiden. Einerseits ist zu fragen, in welchen Branchen und Unternehmensbereichen Big Data-Anwendungen zu Verbesserungen führen können. Zum anderen sind Methoden und Best Practices zu entwickeln, solche Anwendungen zu planen, einzuführen und in die Unternehmensorganisation zu integrieren. Beide Themenbereiche werden in dem vorliegenden Handbuch ausführlich behandelt.

1.2.2.3 Technische Sichtweise

Aus technischer Sicht bezeichnet Big Data Hardware, Software und Methoden, die zur Verarbeitung und Analyse großer Mengen und unstrukturierter Daten genutzt werden können. Hierzu zählen insbesondere die folgenden Themenbereiche:

- Die Bewältigung großer Datenmengen durch Verteilung der Datenspeicherung und der Datenverarbeitung im Wege der verteilten Parallelverarbeitung.
- Die Beschleunigung von Datenbankzugriffen durch eine Datenhaltung im Arbeitsspeicher (sogenannte In-Memory-Technologien).
- Die Optimierung von Zeilen- und spaltenorientierter Datenspeicherung in Datenbanken zur Unterstützung transaktionsorientierter (OLTP) und analyseorientierter (OLAP) Anwendungen sowie die Verbindung derselben.
- Nicht-relationale Datenbankkonzepte zur Verarbeitung großer und unstrukturierter Datenmengen (NoSQL).
- Realtime-Analyse von Datenströmen durch Complex Event Processing.
- Inhaltliches Verständnis unstrukturierter Daten durch semantische Technologien.

Die genannten Themenbereiche sind in dem vorliegenden Handbuch im Einzelnen dargestellt.

Selbstverständlich kommen diese Techniken auch für andere Aufgabenstellungen als Big Data zum Einsatz. Der Einsatz einer bestimmten Technik macht eine Anwendung nicht zu Big Data. Es handelt sich hier um Elemente der technischen Infrastruktur, die eine effiziente Nutzung von Big Data erst möglich machen (McAfee und Brynjolfsson 2014, S. 12).

1.2.3 Gegenstand dieses Handbuchs

Ziel dieses Handbuchs ist es, dem Leser eine Orientierung über die Möglichkeiten und die Anforderungen an die Nutzung von Big Data im Unternehmen zu bieten. Gesellschaftspolitische und -kritische Fragestellungen wurden in diesem Handbuch bewusst nicht thematisiert, da eine fundierte Analyse nicht nur den Rahmen des Handbuchs gesprengt hätte, sondern auch dessen Charakter als praxisorientierter Leitfaden zuwiderliefe.

Von diesem Handbuch soll gleichwohl nicht die Botschaft ausgehen, jedwede Datenutzung sei gerechtfertigt, solange sie ökonomischen Nutzen bringt. Ausgangspunkt ethischer Fragestellungen ist vielmehr die geltende Gesetzeslage. Die ausführlichen juristischen Darstellungen sollen dem Leser Möglichkeiten aufzeigen, die wirtschaftlichen Potenziale von Big Data zu nutzen, ohne dabei Persönlichkeitsrechte der Bürger zu verletzen.

Die IT-Wirtschaft ist wie kaum eine andere Branche international. Die in diesem Handbuch vermittelten Erkenntnisse sind daher grundsätzlich nicht auf Deutschland beschränkt. Hier bildet der rechtliche Teil eine Ausnahme. Vor allem im Datenschutzrecht gibt es zwischen den Rechtsordnungen der für die IT-Wirtschaft relevanten Länder erhebliche Unterschiede. Substantielle Erläuterungen können daher nur für spezifische nationale Rechtssysteme gegeben werden. Die Darstellungen im rechtlichen Teil dieses Handbuchs beziehen sich daher nur auf Deutschland und sind nicht ohne weiteres auf Big Data-Projekte in anderen Ländern übertragbar.

Dieses Handbuch hat drei Teile: einen betriebswirtschaftlich-organisatorischen, einen rechtlichen und einen technischen.

Im betriebswirtschaftlichen Teil wird zum einen betrachtet, wie datenorientierte Entscheidungsprozesse und Architekturen gestaltet sein können und sich erfolgreich in Unternehmen implementieren lassen (Abschn. 2.1 und 2.2). Zum anderen wird dargestellt, welche operativen Prozesse unter welchen Voraussetzungen durch Big Data unterstützt werden können, wobei auf die übergreifenden Methoden Analyse, Simulation und Planung (Abschn. 2.3 und 2.4) und auf branchenspezifische Anwendungsszenarien (Abschn. 2.5–2.9) eingegangen wird.

Die Autoren Hertweck und Kinitzki zeigen in Abschn. 2.1 auf, unter welchen Voraussetzungen Big Data-Technologien tatsächlich einen positiven Einfluss auf die Entscheidungsprozesse im Unternehmen haben können. Neben der Qualität der Daten und der Leistungsfähigkeit der Auswertungssysteme sind nicht zuletzt auch weiche Faktoren wie die Unternehmenskultur und Bereitschaft der Mitarbeiter zu wissensorientiertem Handeln von entscheidender Bedeutung.

In Abschn. 2.2 erläutern die Autoren Schacht und Küller, wie Big Data als Bestandteil eines Enterprise Architecture Management in Unternehmen implementiert werden kann. Auch aus dem Blickwinkel von Informations- und Technologiearchitekturen und deren Veränderung wird deutlich, dass Big Data weder als isolierte Anwendung noch als neues Feature bestehender Systeme begriffen werden darf. Vielmehr gilt es, datenorientierte Prozesse mit Hilfe eines strukturierten und planvollen Enterprise Architecture Management dauerhaft im Unternehmen zu verankern.

Abschnitt 2.3 beschäftigt sich mit der Nutzung von Daten zu Analysezwecken. Die Autoren Lanquillon und Mallow erläutern die wichtigsten Analyse-Aufgaben und stellen anhand des CRISP-DM Ansatzes die veränderten Anforderungen dar, die sich für ein Advanced Analytics von auf Grundlage großer, heterogener Datenbestände ergeben.

In Abschn. 2.4 erweitert und konkretisiert der Autor März diese Betrachtungen für den Bereich der Planung und Simulation am Anwendungsbeispiel der Fahrzeugfertigung.

Abschnitt 2.5. gibt einen kurzen Überblick über einige wesentliche aktuelle und potenzielle zukünftige Einsatzszenarien von Big Data. Dieser allgemeine Überblick wird in den darauffolgenden Kapiteln durch Analysen für ausgewählte Wirtschaftsbereiche vertieft:

In Abschn. 2.6 erläutern die Autoren Theobald und Föhl, welche grundsätzlichen Herausforderungen beim Einsatz von Big Data in der Marktforschung beachtet werden müssen. Die genannten Autoren geben in Abschn. 2.7. einen Einblick in die Einsatzmöglichkeiten von Big Data im Electronic Commerce und damit einem der prädestinierten Einsatzbereiche für Big Data.

In Abschn. 2.8. stellen die Autoren W. Dorschel, van Geenen und J. Dorschel die Potentiale, Notwendigkeiten und Anforderungen im Zusammenhang mit Big Data im Finanzsektor dar.

Abschnitt 2.9. der Autoren März und Stremler erweitert diesen Blick auf Industrieunternehmen.

In Teil 3 geben die Autoren Bartsch, Botzem, J. Dorschel, Hubertus, Ulbricht und Walther einen Überblick über die rechtliche Dimension von Big Data. Hierbei sind vor allem zwei Bereiche zu unterscheiden: Der Schutz *vor* Daten und der Schutz *von* Daten. Schutz vor Daten (Abschn. 3.1) meint die datenschutzrechtlichen Grenzen, die das europäische und deutsche Recht zum Schutze der Betroffenen der Datenverarbeitung zieht. Schutz von Daten (Abschn. 3.2 und 3.3) betrachtet Daten als Wirtschaftsgut. Hier geht es darum, wie und unter welchen Voraussetzungen Daten Schutz vor unberechtigter Verwertung oder Eingriffen in die Datenintegrität genießen. Abschnitt 3.4 widmet sich der Frage, welche Grenzen das Recht der Beschaffung von Daten zieht. Abschnitt 3.5 stellt einige wesentliche Vertragstypen dar, die im Rahmen von Big-Data-Projekten typischerweise relevant sind.

Teil 4 wendet sich zu der technischen Seite des Phänomens Big Data. Hierzu zählen die Infrastruktur und die Anwendungen, die eine betriebswirtschaftliche Nutzbarmachung großer Datenbestände erst ermöglichen.

In Abschn. 4.1 erläutern die Autoren Lanquillon und Mallow die Architektur klassischer Business Intelligence-Anwendungen und zeigen anhand der bekannten Kennzeichnungselemente von Big Data, Volume, Variety, Velocity und Veracity deren Grenzen. Abschnitt 4.2 definiert sodann wesentliche Anforderungen an Big Data-Anwendungen.

Abschnitt 4.3 gibt einen technischen Überblick über die für Big-Data-Anwendungen eingesetzten technischen Infrastrukturen. Die Autoren Fels und Schinkel gehen auf die Skalierung der IT-Infrastruktur durch verteilte Parallelverarbeitung wie auch auf die Verbesserung von Verarbeitungsgeschwindigkeiten durch den Einsatz von NoSQL-Datenbanken und In-Memory-Technologien ein. Ein eigenes Kapitel widmet sich der

Betrieb der entsprechender Ressourcen und mithin der Frage, unter welchen Voraussetzungen ein Outsourcing sinnvoll ist.

In Abschn. 4.4 erläutert der Autor Schulmeyer die technischen Möglichkeiten intelligenter Suchen, die eine weitere Voraussetzung effektiver Auswertungen und mithin wirtschaftlicher Nutzbarmachung großer, heterogener Datenbestände sind. Der Fokus des Beitrags liegt dabei auf der Analyse unterschiedlicher Daten im Textformat mit linguistischen und semantischen Methoden. Der Beitrag zeigt, dass die bloße Erstellung von Indizes und die Auswertung anhand von Keywords nicht ausreichen, um bei unstrukturierten Daten unterschiedlicher Formate mit angemessenem Aufwand sinnvolle Auswertungsergebnisse zu erzielen. Hierfür ist es vielmehr erforderlich, dass die Software die Daten inhaltlich versteht und miteinander in Verbindung setzt.

Literatur

- Bendler, J., Wagner, S., Brandt, T., Neumann, T. (2014): Taming Uncertainty in Big Data. *Business & Information Systems Engineering* 5/2014 S. 279–288, Wiesbaden: Springer Fachmedien
- Fasel, D. (2014): Big Data – eine Einführung. *HMD Praxis der Wirtschaftsinformatik* 04/2014 S. 386–400, Wiesbaden: Springer Fachmedien
- Feinleib, D. (2014): Big Data Bootcamp. What Managers Need to Know to Profit from the Big Data Revolution. Berkely, CA, USA: Apress.
- Freytag, J (2014): Grundlagen und Visionen großer Forschungsfragen im Bereich Big Data. *Informatik-Spektrum* 37/2014, S. 97–104. Berlin, Heidelberg: Springer-Verlag.
- Gluchowski, P. (2014): Empirische Ergebnisse zu Big Data. *HMD Praxis der Wirtschaftsinformatik* 04/2014 S. 401–411, Wiesbaden: Springer Fachmedien
- Hofstetter, Y. (2014): Sie wissen alles. Wie intelligente Maschinen in unser Leben eindringen und warum wir für unsere Freiheit kämpfen müssen. C. Bertelsmann. München: 2014
- King, S. (2014): Big Data. Potential und Barrieren der Nutzung im Unternehmenskontext, Wiesbaden: Springer Fachmedien.
- Kudyba, S., Kwatinetz, M. (2014): Introduction to the Big Data Era. In S. Kudyba (Ed.), *Big Data, Mining, and Analytics* (S. 1–15). Boca Raton (FL): CRC Press.
- McAfee, A., Brynjolfsson, E. (2014): Besser entscheiden mit Big Data. *Harvard Business Manager* 4/2014, S. 6–12. Hamburg: manager magazin Verlagsgesellschaft
- Moore, G. (1965): Cramming more components into integrated circuits. *Electronics* Volume 28 Number 8, 1965, S. 114–117.
- Page, C., Campbell, R., Coggshall, S., Gillespie, E., Johnson, R., Olson, M. & Perkins, P. (2012): Demystifying Big Data: A Practical Guide To Transforming The Business of Government Microsoft. Washington, DC. Verfügbar unter <http://www-304.ibm.com/industries/publicsector/fileserve?contentid=239170>.
- Rudolph, T., Linzmajer, M. (2014): Big Data im Handel, *Marketing Review St. Gallen*, 10/2014, S. 12–24.
- Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. & Tufano, S. (2012): *Analytics: The real-world use of big data*. New York, USA: Somers.
- Schröder, G. (2012): *Datenschutzrecht für die Praxis*, München: Deutscher Taschenbuch-Verlag.

Joachim Dorschel, Werner Dorschel, Ulrich Föhl, Wilhelmus van Geenen,
Dieter Hertweck, Martin Kinitzki, Philipp Küller, Carsten Lanquillon,
Hauke Mallow, Lothar März, Fouad Omri, Sigurd Schacht,
Alphonse Stremler und Elke Theobald

2.1 Datenorientierung statt Bauchentscheidung: Führungs- und Organisationskultur in der datenorientierten Unternehmung

Dieter Hertweck und Martin Kinitzki

Big Data wird aktuell als einer der Haupttrends der IT-Industrie diskutiert. Big Data d.h. auf Basis großer Mengen unterschiedlich strukturierter Daten die Entscheidungen in Echtzeit oder prognostisch zu treffen. Von hochleistungsfähigen, schnell verfügbaren Prognoseverfahren erhofft man sich eine Risikominimierung für unternehmerische Entscheidungen in hochvolatilen Märkten.

Joachim Dorschel ✉ · Fouad Omri
Karlsruhe, Deutschland
e-mail: jd@bartsch-rechtsanwaelte.de

Werner Dorschel
Stuttgart, Deutschland

Prof. Dr. Ulrich Föhl · Prof. Dr. Elke Theobald
Pforzheim, Deutschland

Wilhelmus van Geenen
Hamburg, Deutschland

Prof. Dr. Dieter Hertweck · Martin Kinitzki · Philipp Küller · Prof. Dr. Carsten Lanquillon ·
Prof. Dr. Sigurd Schacht
Heilbronn, Deutschland

Hauke Mallow
Leinfelden-Echterdingen, Deutschland

Dr. Ing. Lothar März · Dr. Ing. Alphonse Stremler
Lindau/Bodensee, Deutschland

Mit der Übergabe von Entscheidungsgewalt an Informationssysteme ändern sich auch die Regeln des Entscheidens und für die Entscheider. In der Big Data Ära müssen Unternehmensziele aktiver innerhalb des Unternehmens kommuniziert werden. Vorgesetzte werden künftig stärker an der Qualität ihrer Entscheidungen messbar sein. Um sensibel auf Marktänderungen zu reagieren, müssen Mitarbeiter kreativer, kritischer und proaktiv an der permanenten Überarbeitung von Teilzielen und Entscheidungsmodellen beteiligt werden.

Big Data wird deshalb nur dort erfolgreich eingesetzt werden, wo es eine Abkehr von Bauchentscheidungen durch Führungskräfte und eine Hinwendung zur permanent hinterfragten, datengetriebenen Entscheidungskultur gibt. Dies bedeutet, dass Führung zukünftig sehr viel komplexer wird und neue Formen der kooperativen, formalen Modellierung von Entscheidungsgrundlagen erfordert. Das Thema „Enterprise Architecture Modelling and Management“ wird dabei zum Schlüsselthema und im nächsten Buchkapitel detaillierter beleuchtet.

Schenkt man jedoch den Heilsversprechen der IT-Industrie zu große Beachtung und glaubt, man kauft sich mit einer neuen Technologie eine unternehmerische Lösung ein, dann wird der soziotechnische Charakter von Big Data Systemen in gleicher Weise verkannt, wie dies in den vergangenen Dekaden bei Wissensmanagement- und/oder BI-Anwendungen der Fall war.

Gefährlich kann dies in Verbindung mit Internet of the things (IOT) Systemen werden, wenn auf Grund prognostisch getroffener, nicht hinterfragter Entscheidungen automatisierte Transaktionen mit hohem Schadenspotenzial (nicht-) getätigten werden, wie z. B. in den analytischen, datengetriebenen Echtzeitsystemen im Militär, Banken- oder Energiesektor. So wurden im Falkland-Krieg die anfliegenden Exocet-Raketen der Argentinier von den britischen Fregatten nicht beschossen, weil sie als „britische Waffensysteme“ erkannt wurden und die Regel, dass auch der Feind solche Waffen gebrauchen könnte, nicht hinterlegt war. Im Rahmen der Finanzkrise 2008 wurden fatale Wirkungen von regelbasierten Informationssystemen auf die Geschwindigkeit des Zusammenbruchs des Weltfinanzsystems konstatiert. In Deutschland wurde der Fall einer 300 Millionen Euro Überweisung durch ein System der KFW-Bank an die kurz vor der Insolvenz stehende Bank Lehman-Brothers öffentlich (Lieven 2009, S. 219 f.).

Die Gefahren des unreflektierten, nicht professionell gesteuerten Gebrauchs analytischer Systeme hat der österreichische Technikphilosoph Günter Anders bereits 1956 auf den Punkt gebracht:

„Um der letzten Gefahr eines Gewissensrufes vorzubeugen, hat man sich Wesen konstruiert, auf die man die *Verantwortung abschieben* kann, Orakelmaschinen also, technische Gewissens-Automaten – denn nichts anderes sind kybernetische Computingmaschinen, die nun, Inbegriff der Wissenschaft [...], *schnurrend die Verantwortung übernehmen*, während der Mensch danebensteht und, halb dankbar und halb triumphierend, seine Hände in Unschuld wäscht.“

Die Frage ob das Ziel, das durch Hebelkombinatorik eingeschaltet wird, *verantwortbar*, nein auch nur *sinnvoll* ist, spielt natürlich für denjenigen, der die Apparatur bedient oder

bedienen lässt, bereits im *Augenblick, da der Apparat zu rechnen beginnt, überhaupt keine Rolle mehr*, nein, die Frage ist überhaupt vergessen. Und der Antwort misstrauen, hieße im Prinzip der Wissenschaft zu mißtrauen; und wo käme er hin, wenn er einen solchen Präzedenzfall schüfe.“ (Anders 2010, S. 245 f.).

Daraus ergeben sich bis zum heutigen Tage für Entscheidungen auf Basis soziotechnischer Systeme, wie dem Big Data getriebenen Unternehmen, folgende Anforderungen:

1. Die unternehmerische Sinnhaftigkeit von Big Data getriebenen Entscheidungen sollte sich in einem transparenten Zielsystem für alle wiederspiegeln.
2. Nach innen belegbare Fakten und deren Wirkung auf ein Prognoseergebnis ersetzen Bauchentscheidungen und erhöhen die Verantwortung *aller* Beteiligten für eine Entscheidung. Das heißt die finale Zielverantwortung darf nicht an eine „Hebelkombinatorik“ abgegeben werden.
3. Es benötigt eine Kultur der Kreativität bei den Mitarbeitern. Der Volatilität der Märkte geschuldet, müssen Big Data getriebene Zielsysteme und Entscheidungsgrößen permanent angepasst und überarbeitet werden. Dies beinhaltet die proaktive Mitarbeit aller Beschäftigten im Unternehmen. Mitarbeiter dürfen auf keinen Fall – wie bei Anders beschrieben – faktengläubig neben schnurrenden Computern stehen.
4. Um eine analytisch erhobene Veränderung operativ umzusetzen bedarf es einer hohen Kommunikations- und Informationskompetenz innerhalb der Organisation, sowie einer fachlichen und mentalen Veränderungskompetenz der Mitarbeiter.

Aus diesen vier Anforderungen ergibt sich in letzter Konsequenz eine fünfte, die besagt:

5. Wenn in Organisationsfragmenten Teilziele des Zielsystems der Unternehmung permanent neu generiert und integriert werden müssen, wenn es immer zentraler wird, wichtige von unwichtigen Informationen zu trennen, dann wird Führung im Big Data Zeitalter anspruchsvoller und komplexer. Eine transparente, kreative Führung bedarf deshalb der Modellierung und Kommunikation der Entscheidungsgrundlagen.

Um ein besseres Verständnis für den Wandel in der Entscheidungs- und Umsetzungskultur des Unternehmens zu erhalten, werden die zuvor beschriebenen fünf Anforderungen an Big Data getriebene Unternehmensführung detaillierter betrachtet.

2.1.1 Unternehmerische Sinnhaftigkeit von Big Data Entscheidungen

Ob Wissensmanagement-, Business-Intelligence- oder nun Big Data-Anwendungen, die Verlockungen der Softwareindustrie sind vielfältig. Allerdings existieren noch immer zu viele technisch hochperformante Systemruinen, die das Faktum widerlegen, dass man

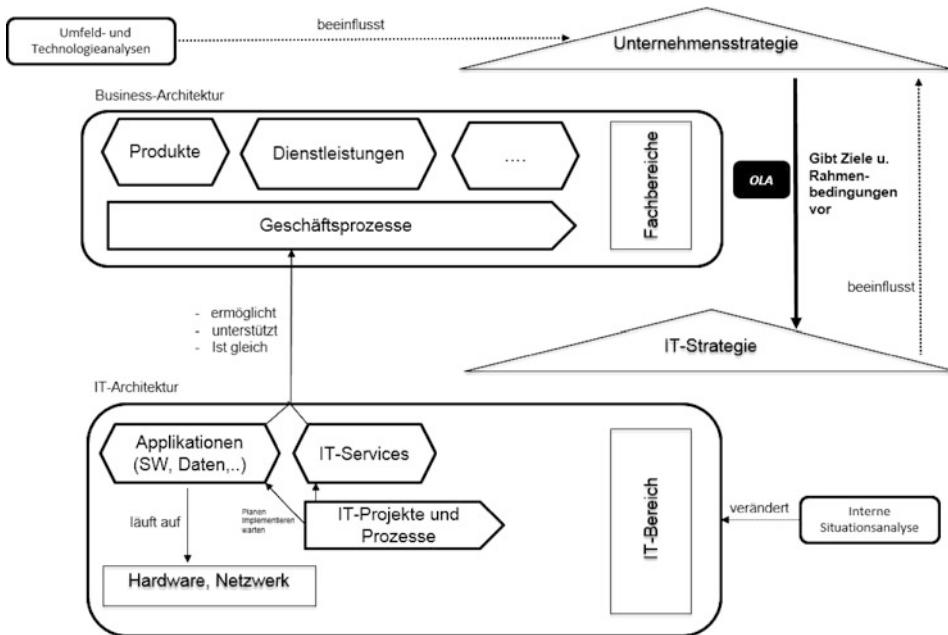


Abb. 2.1 Unternehmensarchitektur in Anlehnung an INNOTRAIN IT (www.innotrain-it.eu)

sich mit Applikationen Problemlösungen erkaufen kann. Dies gilt auch für die Heilsversprechen der Big Data Technologie-Anbieter, die das Auffinden neuer, marktrelevanter Zusammenhänge und daraus ableitbarer Unternehmensziele quasi aus der „Datenursuppe“ als mögliche Methode anpreisen.

Ein erfolgreicher Weg der Nutzung verläuft dagegen umgekehrt, d. h. von der Geschäftsstrategie als oberstes Zielsystem des Unternehmens zur Big Data Strategie und damit zu Big Data getriebenen Entscheidungen. Zum besseren Verständnis sei ein Architekturemodell empfohlen, das sich im Rahmen unserer Forschung im Bereich IT-Management bewährt hat (Abb. 2.1).

Danach werden aus der Unternehmensstrategie sowohl die Business-Architektur des Unternehmens als auch die IT-Strategie nebst IT-Architektur abgeleitet. Deshalb sollte immer ein Business Case vorliegen, der die Einführung von Big Data-Anwendungen in Folge einer Big Data Strategie als sinnvoll erscheinen lässt. Dies wird dann der Fall sein, wenn Echtzeitinformationen, die sich aus strukturierten und unstrukturierten Daten zusammensetzen, einen Mehrwert fürs Unternehmen liefern, z. B. durch

- schnellere und bessere Entscheidungen,
- Senkung unternehmerischer Risiken in Folge präziser Prognoseinformationen,
- Produktdifferenzierung mittels in Echtzeit bereitgestellter, kundenprofilgerechter Informationsprodukte (z. B. Software zu einer vom Kunden im Shop bestellten Hardware, die kundenspezifisch gewünschte Funktionalitäten erschließt).

Um den Wert von Informationen zu bestimmen, unterscheidet Krcmar (Krcmar 2005, S. 42) zwischen folgenden Wertkategorien, die bei der Identifikation von Business Cases von Nutzen sein können:

- *den Normativen Wert* einer Information: der Vergleich der Entscheidungsqualität ohne und mit der benötigten Information (gleich gute Umsetzung der Entscheidung durch die Handelnden vorausgesetzt).
- *den Realistischen Wert* der Information: der empirisch messbare Gewinn, der bei Nutzung der Information durch Entscheider entsteht (monetärer Gewinn, Zeitgewinn, höhere Präzision der Handlung in Folge der Information).
- *den Subjektiven Wert*: der aus der subjektiven Einschätzung des Entscheiders gefühlten Wert, der sich z. B. auf einer Prozent-Skala abschätzen lässt (Bauchgefühl).

Bezüglich der unternehmerischen Sinnhaftigkeit des Einsatzes von Big Data wird im Folgenden vor allem auf den Realistischen Wert der Information Bezug genommen, da er die Interpretation von Daten zu Informationen und die durch die Information ausgelöste Entscheidung sowie die nachgelagerte unternehmerische Handlung miteinander verknüpft. Dies veranschaulicht ein Praxisbeispiel:

Während der Hurrikan Saison in den USA konnten Einzelhändler immer wieder beobachten, dass sich das für gewöhnlich nachgefragte Standardwarensortiment in Erwartung eines Hurrikans spontan verändert. So wurden ihrer Meinung nach signifikant häufiger Konserven, Taschenlampenbatterien, und andere Güter eingekauft, die im Falle eines Wirbelsturms von großem Kundennutzen erscheinen.

Um diese Beobachtung zu validieren, werteten sie die Vergangenheitsdaten der Waren sortimente im ERP-System aus und fanden ihre ursprüngliche Ausgangshypothese bestätigt.

Um in der Hurrikan-Region aus der validierten Hypothese einen geschäftlichen Nutzen zu ziehen, bedarf es im nächsten Schritt eines formal zu modellierenden Entscheidungsmodells, dass erklärt, wie sich in Erwartung eines Hurrikans das Kaufverhalten verändert, und welche Waren sortimentsänderung als Reaktion darauf den höchsten Deckungsbeitrag erzielt.

Würde es im nächsten Schritt gelingen, die Hurrikanerwartung der Kunden zu prognostizieren, und den logistischen Prozess einer Sortimentsumstellung in kürzester Zeit zu bewerkstelligen, so könnte daraus ein wesentlicher Vorteil für das Handelsunternehmen entstehen, weil es z. B. bei weniger verdorbener Frischware, mehr Konserven verkaufen, und einen höheren Sortimentsdeckungsbeitrag erzielen könnte. Die Herausforderung in diesem „einfachen“ Handelsbeispiel besteht in der Operationalisierung des Konstrukt „Hurrikan-Erwartung“ beim Kunden.

Es ist ein mehrdimensionaler Faktor in den Variablen, wie die aktuelle Wettersituation (Klimadaten, Wetterdaten), die Wahrnehmung der Gefährdungsstufe durch den Konsum enten (z. B. aktuelle Thematisierung des Hurrikans in sozialen Netzwerken), sowie

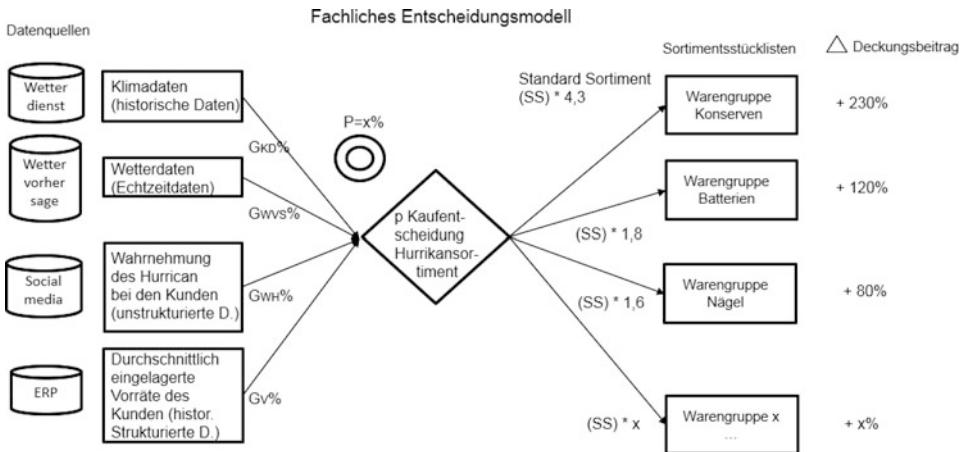


Abb. 2.2 Fachliches Entscheidungsmodell (Eigene Darstellung)

deren bereits eingelagerte Vorräte eingehen. Das Entscheidungsmodell einer Sortimentsumstellung könnte wie folgt aussehen (s. Abb. 2.2).

Die auf Basis von Big Data Erkenntnissen formulierte Veränderung des Zielsystems bei Eintritt des Ereignisses „Hurrikan“ ist unternehmerisch sinnvoll formuliert und transparent. Das heißt man erwartet durch den Einsatz der Big Data Anwendung eine Veränderung von Deckungsbeiträgen in den verschiedenen Teilsortimenten und Warengruppen – so etwa im Nahrungsmittel sortiment eine Steigerung des Deckungsbeitrags der Warengruppe Konserven um 230 %.

Bei dieser Bewertung handelt es sich im Sinne Krcmars (s. o.) um den realistischen Wert der Information, da das Modell weitere Leistungskennzahlen impliziert.

Das heißt die 230 % Steigerung des Deckungsbeitrags bei Konserven werden nur dann erreicht, wenn die Lieferanten für Konserven innerhalb einer bestimmten Frist (t_1) eine Meldung über den benötigten Zusatzbedarf erhalten, und deren Logistiker in einer zuvor vereinbarten Lieferzeit ($t_2 + t_3$) die Waren am „Point of Sales (POS)“ in den Regalen zur Verfügung stellen. (s. Abb. 2.3).

Ein zentraler Aspekt aus Sicht der Führung ist, dass das Zielsystem mit all seinen Abhängigkeiten und Annahmen transparent ist und jeder Mitarbeiter, vom General Management bis zur Fachabteilung, seinen Beitrag zur Gewinnrealisierung kennt. Dies bedeutet eine Abkehr von Bauchentscheidungen und einer nachträglich subjektiven Bewertung eingelegter Maßnahmen durch das Management.

Big Data eröffnet über Analytics Methoden neue Unternehmerische Freiräume und Markt anpassungsprozesse, und verändert mit formalen Entscheidungsmodellen die Entscheidungskultur. Entscheidungen werden auf jeder Ebene der Organisation messbarer, die Ergebnisverantwortung transparenter.

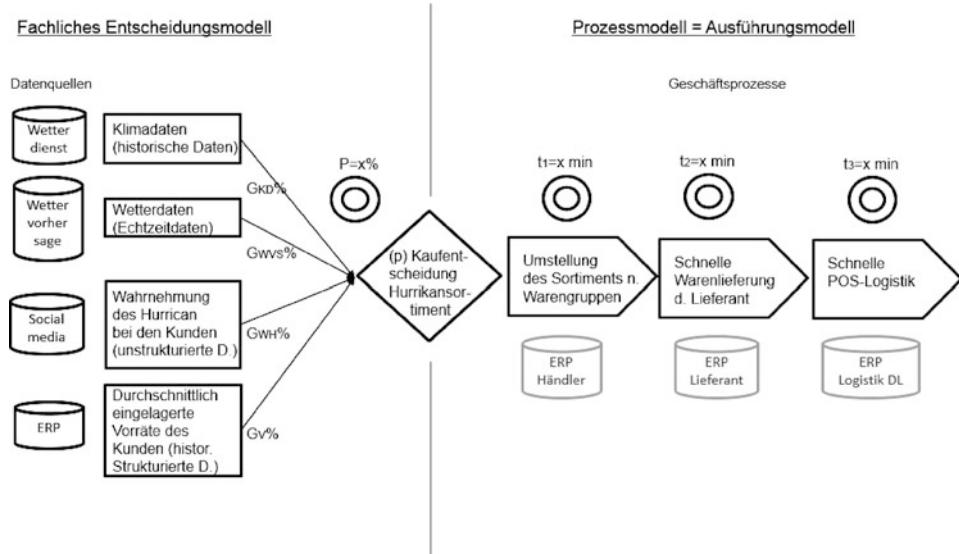


Abb. 2.3 Big Data Entscheidung und Umsetzung (Eigene Darstellung)

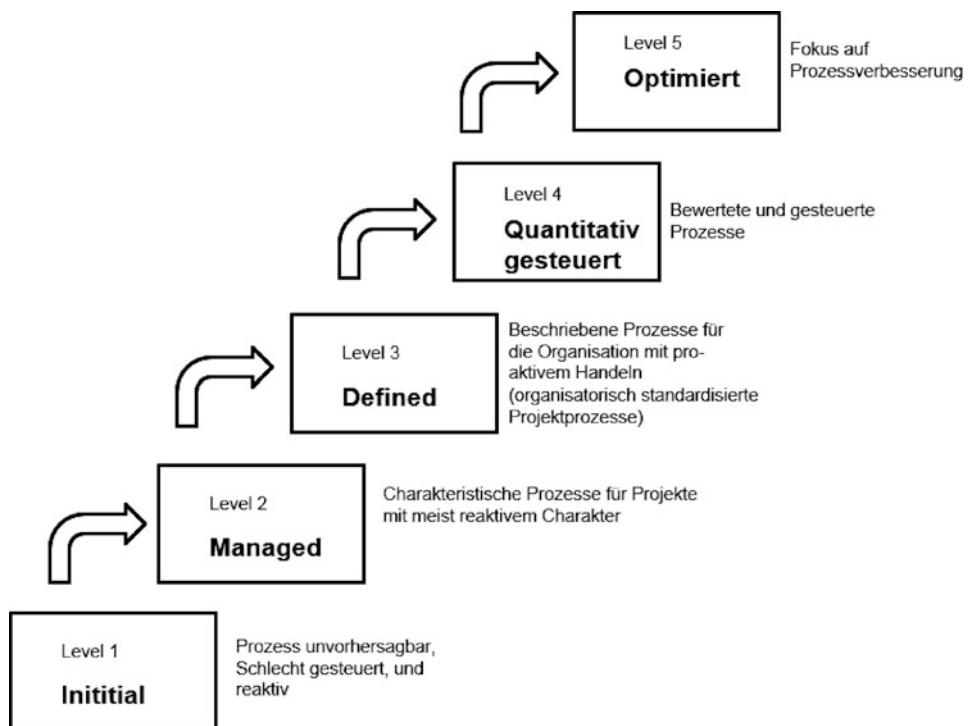


Abb. 2.4 CMMI – Maturity Model für Geschäftsprozesse nach (Hüner et al. 2009)

Mittelfristig kann es sich daher lohnen den Big Data Reifegrad einer Organisation, d. h. die Güte der Big Data Prozesse und Governance z. B. mit Hilfe des CMMI-Maturity-Models zu bestimmen, um daraus Entwicklungserspektiven für das Big Data System des Unternehmens abzuleiten (Hüner et al. 2009, S. 231–238). Das CMMI-Model unterscheidet den Reifegrad von Prozessen entlang folgender Dimensionen (Abb. 2.4).

2.1.2 Fakten erhöhen die Verantwortlichkeit der einzelnen Akteure

Das mit der Datenmenge wachsende Domänenwissen führt dazu, dass die Prädiktoren der Entscheidungsmodelle zunehmend detaillierter werden, was zu einer schrittweisen Erhöhung der Entscheidungssicherheit führt. Da bei der Entwicklung von Prognosemodellen alle Mitarbeiter mit ihrem jeweiligen Fachwissen einbezogen werden, werden auch die Entscheidungsprozesse des Top-Managements für die Mitarbeiter transparenter. Wenn jeder Mitarbeiter weiß, nach welchen Regeln die Potenziale aus den Daten berechnet werden, kann er auch einschätzen, wie hoch die Differenz zwischen dem normativ möglichen und realistisch nach Maßnahmen eingetretenen Mehrwert ist. Damit steht das Top-Management wesentlich stärker unter qualifizierter Beobachtung als bisher. Dies gilt umgekehrt aber auch für den einzelnen Mitarbeiter im Fachbereich, z. B. in der Lebensmittel- oder der Elektroabteilung des oben beschriebenen Handelsunternehmens. Er wird wesentlich stärker in die kontinuierliche Verbesserung des Entscheidungsmodells eingebunden und muss sorgfältig die Einhaltung zuvor spezifizierter Performance-Indikatoren (z. B. Lieferzeit seines Logistikers) überwachen.

Die Bereitstellung bereichsspezifischer Datenpflege- und Analyseverfahren, sowie eine Schulung in der Interpretation erhobener und überwachter Daten sind dabei unerlässlich. Das Unternehmen zieht nur Nutzen aus Big Data-Anwendungen, wenn der Mitarbeiter in der Lage ist, Daten in kürzester Zeit richtig zu interpretieren, um sie in für die Geschäftsprozesse wichtige Information umzuwandeln.

Ferner werden Mitarbeiter vermehrt für die Datenqualität ihres Fachbereichs verantwortlich sein, und dafür sorgen müssen, dass nur zuvor validierte Daten nach einem zuvor definierten, Qualitätssichernden Redaktionsprozess in die Entscheidungsmodelle eingehen.

Welche Auswirkungen nichtvalidierte Daten auf Entscheidungen haben, beschreibt Ahituv/Neumann (Anitav et al. 1994, S. 45) in einem Klassiker der Wirtschaftsinformatik am Beispiel von Ölbohrungen. Dabei wird einem Ölunternehmen ein seismischer Test angeboten, der mit 90 %-iger Wahrscheinlichkeit vorhersagt, ob ein Erfolgs- bzw. Misserfolgssignal (das selbst zu 10 % mit Fehlern behaftet ist) im Falle einer Probebohrung richtig oder falsch ist. Durch die Überlagerung der unscharfen Informationen (10 % Unschärfe Test, 10 % Unschärfe Erfolgs-/Misserfolgssignal) bei gegebenen Kostenstrukturen, würde ein der Bohrung vorgelagerter Test mit neunzigprozentiger Sicherheit auf die Richtigkeit des Erfolgssignals, nachweislich eine Verschlechterung der Prognose im Vergleich zur reinen Zufallsbohrung bedeuten.

Aus diesem Beispiel wird deutlich, welche Verantwortung für die Güte der aus den Big Data-Anwendungen gewonnenen Informationen künftig die Mitarbeiter unterschiedlicher Fachbereiche tragen, wenn diese in komplexe, kollaborativ konstruierte Entscheidungsmodelle eingehen, in denen sich Informationen überlagern.

Der unternehmerisch erfolgreiche Einsatz von Big Data ist nur dann gewährleistet, wenn die involvierten Mitarbeiter Verantwortung für die Qualität der zu verarbeitenden Daten übernehmen (Data Governance), sich auf dem Gebiet analytischer Verfahren weiterbilden und als wichtig erkannte Veränderungen in den Entscheidungsmodellen durch aktive Kommunikation und Handlungen in der Organisation thematisieren.

Eine steigende Anzahl von Personen in professionellen, sozialen Netzwerken, die Information- oder Data Governance in ihrer Berufsbezeichnung stehen haben, scheint ein Beleg dafür zu sein, dass beide Themen in den Unternehmen ankommen (Nadler 2014, S. 5).

2.1.3 Kreativität der Mitarbeiter als Teil einer Big Data freundlichen Unternehmenskultur

Die Volatilität der Märkte erfordert eine stetige Anpassung der zuvor beschriebenen, formalen Entscheidungsmodelle. Da die meisten Menschen dazu neigen, bewährte Handlungsmuster aus dem Gefühl des Vertrauten heraus fortzusetzen (Hertweck 2003, S. 13), bedarf es einer besonders innovations- und veränderungsfreundlichen Unternehmenskultur, um gemeinsam entwickelte und abgestimmte Entscheidungsmodelle permanent in Frage zu stellen. Ist diese Ebene erreicht, spricht North von Individuen, die in der Lage sind, über Kompetenzen hinaus einen Wettbewerbsvorteil für das Unternehmen aufzubauen. Welche Rolle in diesem Transformationsprozess die Unternehmenskultur spielt, sei am Beispiel unseres Einzelhandelsgeschäfts beschrieben (Abb. 2.5).

Aus den analysierten externen und internen Daten und deren Einbindung in formale Entscheidungsmodelle (z. B. Grad der Hurrikan Vermutung des Konsumenten), werden aus den Daten Informationen. Diese werden interpretiert und mit dem Anwendungsbezug (Notwendigkeitskriterien zur Sortimentsumstellung) vernetzt, woraus sich Wissen ergibt.

Ob dieses Wissen in eine für das Unternehmen gewinnbringende Handlung mündet (reale Sortimentsumstellung) hängt vom Können und Wollen der Mitarbeiter in zweiter Linie von der Performanz der Systeme ab.

Wird Können durch wiederholtes richtiges Handeln in Wertschöpfung umgesetzt, so spricht man von einer Kompetenz, die den Unterschied im Wettbewerb ausmacht, einer sogenannten Kernkompetenz des Unternehmens.

Der zentrale Schritt zum Wettbewerbsvorteil bleibt in dieser Kette der Schritt vom Können zum Handeln, der über das Wollen läuft. Wann aber sind Mitarbeiter bereit, ihr Können zum Wohle des Unternehmens aktiv einzusetzen, d. h. woher kommt die Motivation des Wollens?

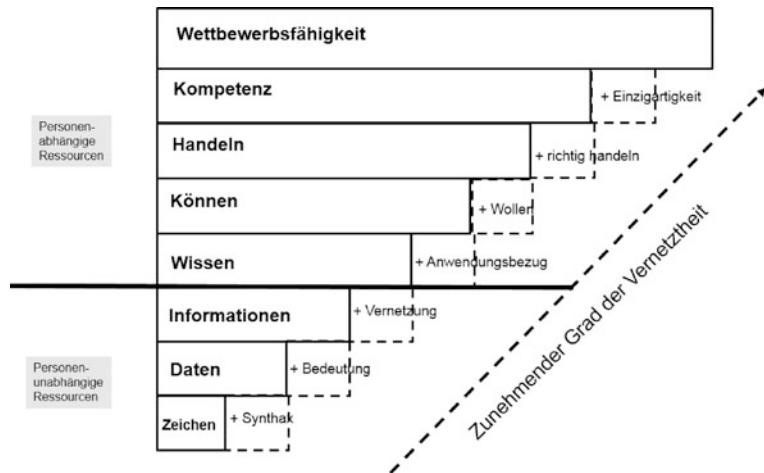


Abb. 2.5 Entstehen von Wissen, Können und Kompetenzen nach North (Völker 2007, S. 60)

Rein empirisch betrachtet erhält man über verschiedenste Studien (Völker 2007, S. 83) hinweg immer wieder eine ähnliche Reihenfolge von drei Hauptkriterien, die zu 90 % erklären, weshalb Mitarbeiter bereit sind Ihr Können in unternehmerisches Handeln überzuführen:

- innovationsfreundliche Unternehmenskultur (Mensch), 44 %,
- Strukturen und Prozesse (Organisation), 24 %,
- Informationstechnologie (Technik), 23 %.

Was kennzeichnet nun eine gute, innovations- und wissensfreundliche Unternehmenskultur?

Unter der Kultur eines Unternehmens versteht man die in der Organisation vorherrschenden Normen, Werte, Regeln und Motivatoren, die das Verhalten der Mitarbeiter prägen und die durch tägliches Handeln rekursiv bestätigt werden. Unternehmenskultur ist nicht kurzfristig gestaltbar, sondern das Sediment tradiert Wertvorstellungen, Denkweisen, Handlungsrouterinen und Erfolgsgeheimnisse.

Typische Werte, die dem mikropolitischen Zurückhalten von Können (Ortmann 1990, S. 30) entgegenwirken und einen positiven Einfluss auf das „Wollen“ haben, sind in Abb. 2.6 zu sehen:

- Ein Klima von Vertrauen und Offenheit: Das bedeutet, dass der Mitarbeiter davon ausgeht, dass das Unternehmen sinnvoll wirtschaftet und ihm ein gewisses Maß an ökonomischer Sicherheit garantiert. Im Gegenzug ist der Mitarbeiter bereit, einen maximal wertschöpfenden Beitrag zum Unternehmenserfolg zu leisten und er erwartet, dass dieser auch als sein Beitrag anerkannt wird (und sich z. B. nicht der Vorgesetzte mit den Leistungen des Mitarbeiters schmückt).

Abb. 2.6 Elemente einer Wissensfreundlichen Unternehmenskultur (Völker 2007, S. 100)



- Ein Klima der Fehlerfreiheit: In zahlreichen Studien konnte bisher belegt werden, dass die Möglichkeit für Mitarbeiter einen Fehler zu begehen, ohne gleich mit Sanktionen rechnen zu müssen, ein Hauptförderer für Innovationen und proaktives Handeln ist.
- Lernbereitschaft: Wenn in einer Organisation aus den Handlungsspielräumen und Fehlern gelernt und die stetige Weiterentwicklung als Wert gelebt wird, stärkt dies den Willen von Mitarbeitern, die Richtigkeit des Handelns auf den unternehmerischen Erfolg bei sich und anderen stets kritisch zu hinterfragen.
- Freiräume und Handlungsspielräume: Kreativität heißt nichts anderes als bekannte Beobachtungen mit neuen Erkenntnissen zu etwas Neuem zu vernetzen. Um neue Erkenntnisse zu gewinnen bedarf es der Freiräume, sich mit Themen zu beschäftigen, die vielleicht nicht im unmittelbaren Umfeld der Arbeitsrealität zu finden sind. Kreative Lösungen erfolgen oft erst durch die Verknüpfung neuer, meist von außerhalb des Unternehmens kommender Erkenntnisse oder Technologien mit bereits Bekanntem und Erlebten.
- Führungsstil: Ein vertrauensvoller Führungsstil, der eher das Ergebnis als den Prozess vorgibt, ermöglicht Mitarbeitern ihre Handlungsspielräume für alternative Lösungen sinnvoll (z. B. Prozessinnovationen) zu nutzen.

Übertragen auf eine Big Data freundliche Unternehmenskultur bedeutet dies, dass Unternehmen ihren Mitarbeitern Freiräume zugestehen, sich mit den existierenden Entscheidungsmodellen regelmäßig und aktiv auseinanderzusetzen. Mitarbeiter der Fachabteilungen, die aktiv an der Entwicklung von Teilentscheidungsmodellen beteiligt sind, erhalten in einer Organisation mit gutem Lernklima eine für Ihre Arbeit sinnvolle Datenpflege, -aufbereitungs- und Analyseausbildung, sowie einen offenen Zugang zu den relevanten Systemen.

2.1.4 Informations- und Kommunikationskompetenz und Veränderungskompetenz als Basis schneller Reaktionszeiten

Um aus Big Data gewonnene Erkenntnisse in schnelles Handeln umzusetzen, bedarf es einer hohen Informations- und Kommunikationskompetenz im und zwischen Unternehmen. Jeder Mitarbeiter muss die Big Data Aktivitäten und Maßnahmen des Unternehmens, sowie die zu Grunde liegende Entscheidungsmodelle kennen. Die von Maßnahmen direkt Betroffenen müssen zusätzlich Detailkenntnisse über die Aktivitäten der angestoßenen Geschäftsprozesse besitzen.

Eingangslagermitarbeiter müssen im oben geschilderten Einzelhandelsunternehmen darüber benachrichtigt werden, dass eine Sortimentsumstellung auf Grund einer Hurrikan-Vorhersage stattfindet, um die entsprechenden Lagerkapazitäten und Zugangswwege für jene Waren freizuhalten, die über die Bestückung am Point of Sale hinaus vorgehalten werden.

Für Lieferanten könnte die Information wichtig sein, wie lange sie auf Grund der Wettervorhersagedaten noch gefahrlos welche Warenmenge an den Einzelhändler liefern können. Schnelle Reaktions- und Lieferzeiten sind für den Business Case eine Grundvoraussetzung. Eine gute informationelle Durchdringung entlang des Big Data Business Cases wird erreicht, wenn alle beteiligten Organisationsbereiche inkl. der Zulieferer an der Entwicklung gemeinsamer Entscheidungsmodelle beteiligt werden.

Betroffene zu Beteiligten machen und im Planungsprozess von deren Kompetenzen profitieren, ist ein wesentlicher Erfolgsfaktor für eine offene Kommunikationskultur.

Kommunikationskompetenz stellt sich im Rahmen kollaborativer Entscheidungsmodell-Entwicklung und Geschäftsprozessvisualisierung ein (Schermann et al. 2008, S. 1582). Spätere Fachtermini der Organisation, hinter denen sich meist komplexe, gemeinsam besprochene Handlungsvorschriften verbergen, werden in der Modellierungsphase gebildet und gehen über Sozialisation in den Wissensvorrat der Beteiligten über.

Es werden Begriffe gebildet, Kontexte selektiert, Entscheidungsgrößen identifiziert und in ihrer Ausprägung bewertet. Die Technik der gemeinsamen Visualisierung von geführten Diskussionen, die auf dem individuellen Erleben der Organisation durch die Mitarbeiter beruhen, erfüllt die Kriterien des Mehrdimensionalen Lernens (Ammon 1985, S. 99–110).

Eine Herausforderung besteht darin, auch Nichtbeteiligte an diesem Lernprozess teilhaben zu lassen. Dies kann durch gezieltes Trainieren der zuvor entwickelten Geschäftsprozesse geschehen, was kurze Reaktionszeiten fördert und Verbesserungspotenziale offenlegt, da Unbeteiligte in einer innovationsfreundlichen Unternehmenskultur oft jene Phänomene hinterfragen, die den am Entwicklungsprozess Beteiligten als unumstößliche Realität erscheint (Betriebsblindheit). Werden die oben beschriebenen Normen und Werte einer wissensfreundlichen Organisationskultur gelebt, sollte ein permanentes, Hinterfragen der Entscheidungsmodelle durch die Fachbereiche ein erwünschtes Ergebnis sein, dass die Anpassungsfähigkeit der Organisation sichert.

So könnten in unserem Beispiel gesundheitsschädliche Konservierungsstoffe in bestimmten Konserven entdeckt und öffentlich werden, so dass Kunden – trotz Hurrikan-

Warnung – die besagte Warengruppe nicht mehr oder in geringeren Mengen kaufen. In diesem Fall müssten Einkäufer und Mitarbeiter der Lebensmittelabteilung den Skandal und dessen Auswirkung auf das Hurrikan-Warensortiment zeitnah thematisieren. Für mögliche Änderungen sollte es in der Organisation einen zentralen Anlaufpunkt geben, an dem Veränderungen des Entscheidungsmodells entgegengenommen und bewertet werden. Ferner sollte es zyklische oder prioritätsgetriebene Anlässe geben, zu denen gesammelte Änderungsanträge diskutiert und in das Modell aufgenommen werden.

Beim Management des Informationslebenszyklus kann man auf bewährte Konzepte aus dem IT-Service-Management zurückgreifen. Eine erkannte Störung der Informationsversorgung (Incidentmanagement) könnte man an einen Service-Desk melden. Dort wird die Störung priorisiert und als vorrübergehendes Phänomen mit Workaround oder als substantielles Problem mit längerfristiger Wirkung eingestuft (Problemmanagement). Im Problemfall muss mit einer zeitnahen Veränderung des formalen Entscheidungsmodells reagiert (Change Management), die Veränderung samt Ursachen in einer Datenbank dokumentiert (CMDB) werden. Über den gleichen Service Desk und mit vergleichbaren Prozessen ließen sich Data-Governance-Themen wie Datenqualität, Datenschutz und Datensicherheit bearbeiten. Auch die Frage des Umgangs mit großen Mengen an „Altdaten“ ließe sich analog zu ITIL mit Daten-Lebenszyklusmodellen behandeln.

Abschließend lässt sich zum Thema kommunikations-/informations- und veränderungsfreundliche Organisations- und Führungskultur folgendes festhalten:

- Organisationale Veränderungskompetenz ist eine Grundvoraussetzung, um auf veränderte Marktbedingungen und veränderte Big Data Entscheidungsmodelle mit angemessenen Maßnahmen zu reagieren.
- Die Veränderungskompetenz betrifft die geistige Flexibilität der Mitarbeiter. Ein einmal erstelltes und kommuniziertes Entscheidungsmodell muss durch proaktive Verarbeitung neuster Informationen permanent in Frage gestellt werden.
- Veränderungskompetenz bedeutet auch in kürzester Zeit neue Maßnahmen und Kennzahlen auf veränderte Umgebungsbedingungen zu finden und diese in performanten Geschäftsprozessen zu erfüllen.
- Um Mitarbeitern die Möglichkeiten zu geben, wahrgenommene Veränderungen in Handlungen umzusetzen, bedarf es organisatorischer Rahmenbedingungen. Dazu zählen eine wissens- und innovationsfreundliche Unternehmenskultur und etablierte Rollen und Prozesse im Bereich Data und Information Governance.
- Zu guter Letzt braucht es Führungskräfte, die ein permanentes Hinterfragen einmal aufgestellter Entscheidungsmodelle nicht als Kritik an ihrer Person, sondern als Zeichen gelebter Innovationskultur betrachten.

Dies ist umso bedeutungsvoller, als in einer informierten Organisation vermehrt der Fall auftritt, dass dezentrale Stellen temporär über Informationsvorsprünge verfügen und Führungskräfte zunehmend mit Informationsasymmetrien zu ihren Ungunsten umgehen müssen. (Altmann 2013, S. 22)

2.1.5 Führung wird komplexer und bedarf der Unternehmensmodellierung, sowie des aktiven Managements der Unternehmensarchitektur

Aus den oben genannten organisatorischen Veränderungen hin zur Daten- und Wissensgetriebenen Organisation werden unterschiedlichste neue Anforderungen an Führungskräfte offensichtlich, wie:

1. Förderung sensibler Mitarbeiter, die noch stärker als bisher das Ohr am Markt haben. Dies schließt Freiräume für Aktivitäten in sozialen Netzwerken mit ein, aus denen das Unternehmen aktuelle, unstrukturierte Datenquellen (z. B. sozialer Netzwerke) beziehen kann, um Innovationen oder Marktveränderungen zeitnah zu erkennen.
2. Wichtiges von Unwichtigem zu trennen: Bei der gesamten Flut strukturierter wie unstrukturierter Daten ist es wichtiger denn je, den geschäftsbezogenen Überblick zu behalten, d. h. geschäftskritische von unwichtigen Daten zu trennen.
3. Analyse- und Interpretationskompetenz: Es bedarf Führungskräften mit einer Vorstellung für die Wirkeffekte unternehmerischer Prozesse und für Kennzahlen, mit denen sich der Erfolg von Maßnahmen steuern lässt.
4. Selbtkritik: Das permanente Hinterfragen von Entscheidungsprämissen durch Mitarbeiter sollte nicht als ein Hinterfragen der Kompetenz der Führungskraft missverstanden werden.
5. Management des Daten-Lebenszyklus: Daten müssen von der Selektion der Datenquellen, über die Validierung und Aufbereitung, bis zur Sicherung und Entsorgung professionell gemanagt werden.
6. Schaffung geeigneter Entwicklungsmethoden für eine kommunikationsfreundliche Organisation, wie die Partizipation von Mitarbeitern bei der Entscheidungsmodellentwicklung, die Schaffung eines fehlertoleranten und lernfreudigen Arbeitsklimas, die Honorierung eines konstruktiv kritischen Unternehmensklimas.
7. Zu guter Letzt bedürfen all die beschriebenen Führungsaufgaben einer abgestimmten Koordination zwischen der Geschäftsführung, den Fachbereichen und der IT.

All diese Anforderungen sinnvoll und zum Wohle des Unternehmens zu steuern geht heutzutage kaum noch ohne ein professionelles Enterprise Architecture Management nebst unterstützendem System, auf das alle Beteiligten entsprechenden Zugriff haben sollten (Hanschke 2011, S. 150 f.). Ein solches System, dass die Abhängigkeiten zwischen der Unternehmensstrategie und den Bereichsstrategien, zwischen Produkten, Prozessen und Informationssystemen transparent und für jeden verständlich macht, ist ein weiterer wesentlicher Schritt zur Etablierung einer kommunikativen und offenen Unternehmenskultur. In Abb. 2.7 ist die Unternehmensarchitektur unseres Einzelhandelsbeispiels zu sehen.

In Abb. 2.7 lässt sich gut erkennen, wie die externen Faktoren Hurrikan-Meldung in den Wetternachrichten und deren Diskussion zu einer kurzfristigen Strategischen Änderung (Sortimentsumstellung, Punkt 1) führen. Die Grundlagen dieser Änderung basieren auf einem Entscheidungsmodell, das zuvor von allen beteiligten Bereichen ausgearbeitet

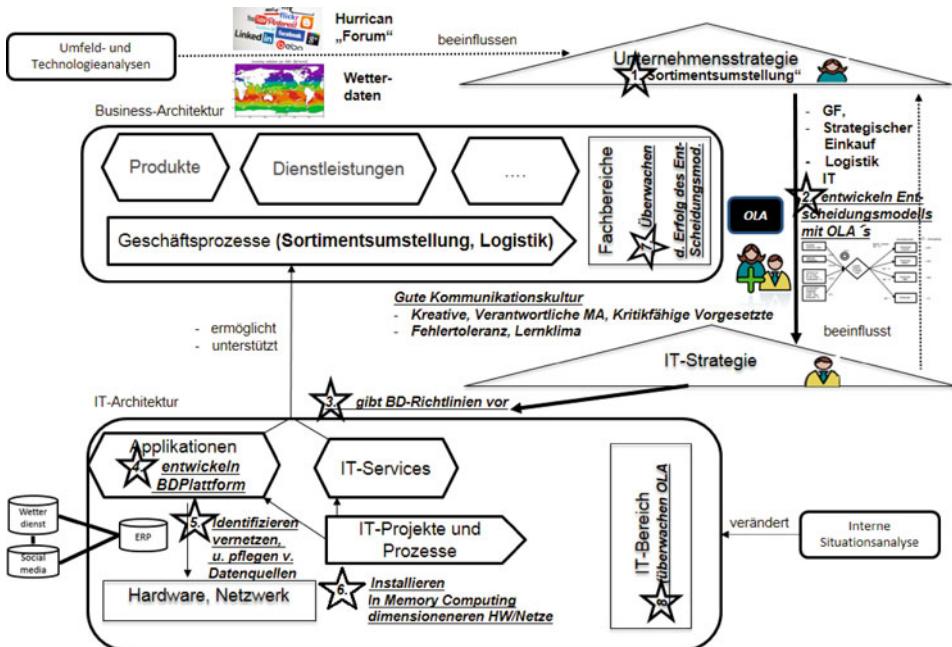


Abb. 2.7 EAM – Fallbeispiel Einzelhandel (eigene Darstellung)

wurde (Punkt 2) und endet in einem Organisational Level Agreement (OLA) zwischen den Fachabteilungen und dem IT-Bereich, aus dem die wichtigsten Leistungsziele für den IT-Bereich folgen. Mit diesen Vorgaben erstellt der IT-Bereich seine Big Data Strategie aus IT-Sicht (Punkt 3) und entwickelt daraus das Big Data System, welches aus vernetzten internen und externen Datenquellen unterschiedlichster Datenarten und Zugriffszeitvorgaben besteht (Punkt 4 und 5).

Um diese Leistungsvorgaben zu erfüllen, bedarf es neuer Investitionen in die Hardware (in Memory-Technologien, Netzwerkdimensionierung, ... s. Punkt 6). Nach Inbetriebnahme des Systems und Schulung der Anwender, werden die Fachbereiche darauf achten, ob die zuvor gemeinsam definierten Zielvorgaben eingehalten werden können bzw. steuern – wenn nötig – dagegen. Diese Abweichungen müssen vom IT-Bereich erfasst und mit Veränderungen der IT-Architektur berücksichtigt werden (Punkt 8).

Um Abweichungen und mögliche Änderungen während des Big Data Betriebs zu thematisieren, bedarf es der permanenten Pflege einer wissensfreudlichen Kommunikationskultur zwischen den Mitarbeitern und der Unternehmensführung. Aus der Vielfalt der zu koordinierenden Führungsmaßnahmen und den Abhängigkeiten von Mensch, Technik und organisationaler Besonderheiten, wird die Notwendigkeit des Einsatzes von EAM Werkzeugen offensichtlich, dem sich das nächste Buchkapitel detaillierter widmet.

2.1.6 Zusammenfassung: Tipps für Entscheider, die es bei der Einführung einer datengetriebenen Entscheidungskultur zu beachten gibt

Die beschriebene Komplexität des Managements von Big Data sollte weder Unternehmen noch Entscheider davon abhalten, sich intensiv damit zu beschäftigen. Mit In-Memory-Datenbanken und endlos verfügbarer Rechenleistung aus virtualisierten Infrastrukturen war es technisch gesehen noch nie einfacher und kostengünstiger, Echtzeit- oder prognostische Entscheidungen mit hoher Business-Wirkung zu treffen.

Die in Big Data getätigten Investitionen werden nur dann zum Verlust, wenn den Aufwänden keine Erträge entgegenstehen, weil:

- man sich aus den Daten Antworten verspricht, die die Daten ohne einen konkreten Business-Case selten hergeben,
- auf Grund veralteter oder fehlerhafter Entscheidungsmodelle hocheffizient und automatisiert falsche Entscheidungen getroffen werden,
- die laufenden Aufwände zur Datenwartung zu hoch werden, wenn es kein vernünftiges oder sogar ein gesetzlich mangelhaftes Data Governance Konzept gibt (Zuständigkeiten, Datenbereinigung, Compliance ...).

Um Big Data für das Unternehmen zu erschließen, sind folgende Hinweise hilfreich:

1. Start smart: Es geht darum, nicht der Versuchung zu unterliegen von Beginn an zu große Entscheidungsmodelle aus großen Datenmengen zu extrahieren.

Wie in Abschn. 2.1.1 gezeigt werden konnte, bedarf es recht gut recherchiert und modellierter Informationen, um eine kausale Wirkung auf eine Entscheidung gesichert belegen zu können. Viele Unternehmen machen bei der Einführung von Neuerungen den Fehler, dass sie zu viel zur gleichen Zeit einführen wollen. Insbesondere bei datengetriebenen Entscheidungsmodellen kann dies zu unhinterfragten Abhängigkeiten zwischen Daten und ihrer Merkmalsausprägungen führen, die sich mit zunehmender Komplexität negativ auf die Entscheidungsgüte auswirken.

Deshalb sollte man stets mit der Business-Strategie beginnen und anschließend iterativ die unterschiedlichsten Big Data Wirkfaktoren in die Big Data Strategie einbringen, um sie in ihrer Business Wirkung zu testen (z. B. Klimadaten, Wetterdaten, Social Media Daten).

2. Den ökonomischen und sicheren Umgang mit Daten planen

Sobald die relevanten Datenbasen identifiziert und den Wirkmechanismen zugeordnet wurden, bedarf es eines Planes, wie man gesetzentreu und ökonomisch mit der Datenpflege verfährt. Dies beinhaltet die Aufbereitung, die Qualitätssicherung der Daten, die Regelung des Zugangs (Datensicherheit), sowie die Archivierung/Entsorgung nicht mehr

benötigter Datenbestände. Wichtig ist, dass sich zu den angesprochenen Dimensionen ge-regelte Zuständigkeiten und Rollen innerhalb der Organisation finden.

3. Die Aktualität von Entscheidungs- und Datenmodellen sicherstellen

Daten sind unternehmerisch nur so viel wert, wie der unternehmerische Kontext, den sie repräsentieren. Ergeben sich offensichtliche Änderungen von Seiten des Marktes oder von den internen Zielsystemen des Unternehmens, müssen diese offen und zeitnah thematisiert werden. Während es mit No-SQL-Datenbanken oder In Memory Computing technisch kein Problem ist, Daten in Echtzeit auszuwerten, können dies allerdings Daten zu nicht mehr relevanten Zielsystemen sein. Um Zielsysteme rechtzeitig in Frage zu stellen und sie zu korrigieren, braucht es vor allem kritikfähige Mitarbeiter, entscheidungs-starke Führungskräfte und eine kommunikationsfreundliche Unternehmenskultur.

4. Eine informations- und kommunikationsfreundliche Unternehmenskultur fördern

Eine informations- und kommunikationsfreundliche Unternehmenskultur zeichnet sich durch eine professionelle Fehlerkultur aus, in der Führungskräfte und Mitarbeiter Fehler begehen dürfen, ohne dass sie postwendend dafür sanktioniert werden. Mitarbeiter sollen in die Entwicklung der Datenmodelle und Zielsysteme des Unternehmens einbezogen und für die neuen Aufgaben im Bereich der Daten-Analyse und im Umgang mit Tools qualifiziert werden.

5. Alignment der Big Data Strategie mit der IT-Strategie und IT-Architektur

Wenn das Paradigma einer kommunikationsfreundlichen Unternehmenskultur auch das Verhältnis zwischen Geschäftsführung und IT-Leitung beschreibt, dann wird die Geschäftsführung mit den von der Big Data Aktivität betroffenen Fachabteilungen und der IT-Abteilung frühzeitig zusammensitzen, um gemeinsam über notwendige Entscheidungsmodelle, Datenqualitäten und Umsetzungszeiten zu beraten. Aus diesen Geschäftsvorgaben bzw. Organisational Level Agreements des Big Data Business Cases lassen sich die notwendige Konfigurationsänderungen der IT-Architektur ableiten. Diese Änderungen können dabei unterschiedlichste Architekturelemente bzw. Configuration Items betreffen, wie etwa Datenbanken, Schnittstellen, Analysesoftware oder Hardware- und Netzwerkerweiterungen, aber auch organisatorische, wie neue Big Data Pflege- und Wartungsprozesse oder neue Big Data Rollen der IT-Mitarbeiter.

Empfehlenswert ist die Dokumentation und Unterstützung des Enterprise-Architecture-Managements und seiner Releases mit einem standardisierten Vorgehensmodell wie z. B. TOGAF (Matthes 2011, S. 150 f.) und professionellen Management-Werkzeugen, wie z. B. AdoIT von der BOC GmbH oder Alfabet Enterprise Architecture Management von der Software AG. Für kleinere und mittlere Unternehmen kann auch das IT-Service- und Enterprise Architecture Management-Portal INNOTRAIN-IT (www.innotrain-it.eu)

empfohlen werden, auf dem zahlreiche Best Practice Materialien und Fallstudien zum Thema Business-/IT-Alignement und Enterprise Architecture Management mehrsprachig verfügbar sind.

Abschließend lässt sich festhalten, dass ein datengetriebenes Vorgehen bei Entscheidungen im Unternehmen wesentlicher infrastruktureller Voraussetzungen bedarf:

Die erste ist Transparenz – d. h. für alle zugängliche, gemeinsam entwickelte und in die Unternehmung kommunizierte, formale Entscheidungs-, Kennzahlen- und Maßnahmenmodelle.

Die zweite ist eine offene, kommunikationsfreundliche Unternehmenskultur, in der Mitarbeiter Veränderungen in Umwelt und Organisation wahrnehmen und daraufhin bestehende Entscheidungs- und Handlungsmodelle in Frage stellen.

Die dritte Voraussetzung sind kritikfähige Führungskräfte, die das in-Frage-stellen zuvor entwickelter Entscheidungsmodelle nicht als Kritik an der Person, sondern an der Sache verstehen.

Die vierte ist ein professioneller Umgang mit Daten in puncto Verantwortlichkeiten, Lebenszyklusmodelle, Datenschutz und Datensicherheit.

Dies sind Rahmenbedingungen, die in klassisch, bürokratisch geprägten Organisationsformen mit starker Abgrenzung zur Umwelt nur schwer umzusetzen sind. Das betrifft insbesondere die Führungsebene, für die selbstgestaltete Informationsasymmetrien zur Absicherung von Bauchentscheidungen wertlos werden.

Die Führung im Big Data sollte die offene, transparente Organisation kultivieren, geschäftskritisch wahrgenommene Umweltbedingungen durch Mitarbeiter honorieren und zeitnah umsetzen.

Dies bedarf Mitarbeiter die in verschiedenste Netzwerke auch außerhalb des Unternehmens eingebunden sind, die sensibel auf Veränderungen reagieren und sich stark mit dem Unternehmen identifizieren. Fähigkeiten, die die kommende Social Media und Mobile Computing geprägte Generation durchaus mitbringt. Gleichzeitig bedarf es aber auch eines neuen Führungstyps, der offensiv Informationen teilt, professionell managt und kognitiv mit permanent auftretenden Veränderungen derart umgehen kann, dass er diese nicht als Kritik an einmal getroffenen Entscheidungen oder gar an seiner Person interpretiert – wahrscheinlich die größte Herausforderung für die Führung im Big Data.

2.2 Enterprise Architecture Management und Big Data

Sigurd Schacht und Philipp Küller

2.2.1 Enterprise Architecture Management und Big Data

Die Studie Big Data Analytics 2014 des BARC Instituts zeigt, dass der wahre Mehrwert von Big Data nicht etwa rein die Menge der analysierten Daten ist, sondern vielmehr in den nahezu in Echtzeit bereitgestellten Analyseergebnissen liegt, die eine zeitnahe Ableitung von Handlungsanweisungen ermöglicht. Dies führt zur Schaffung einer er-

höhten Transparenz im Unternehmen und wirkt sich damit vor allem im Bereich der Geschäftsprozessverbesserung sowie in der Identifikation neuer Geschäftsmodelle aus (Bange and Janoschek 2014, S. 5).

Der Bundesverband für Informationswirtschaft, Telekommunikation und neue Medien e. V. (Bitkom) sieht aufgrund der zunehmenden Digitalisierung der betriebswirtschaftlichen Abläufe Daten als vierten Produktionsfaktor. Der momentane Hype um Big Data, zunächst bezogen auf die technologische Machbarkeit, nunmehr erweitert auf ein strategisches methodisches Vorgehen, wird als Anzeichen für eine Steigerung der Bedeutung von Daten in Unternehmen gesehen (Bitkom 2012, S. 7).

Gemäß dem Motto „Wo Licht ist, ist auch Schatten“ werden vereinzelt auch Bedenken geäußert. So führt Gartner in einer Studie auf, dass bis 2016 große Firmen an Big Data verzweifeln werden. Das liegt daran, dass laut Gartner 85 % der 500 umsatzstärksten Unternehmen nur schlecht auf die Komplexität, Masse und Vielschichtigkeit der Daten vorbereitet sein werden, die notwendig sind, um den Verantwortlichen zeitnahe und zielorientierte Analysen für die Entscheidungsfindung zur Verfügung stellen zu können (Kröger 2013). Nur die Transparenz über die vorhandenen Daten unterschiedlichster Art sowie den vorhandenen Informationsflüssen, Wechselbeziehungen und Verantwortlichkeiten ermöglicht den Unternehmen das Potenzial, das vor allem durch die Kombination von unterschiedlichen Datenquellen entsteht, bewältigen und heben zu können. Die notwendige Transparenz muss nicht nur auf der Ebene der Informationen – also welche Daten stehen zur Verfügung und welche neuen Kenntnisse können hiermit gehoben werden – geschaffen werden, sondern vielmehr – wie in dem vorhergehenden Artikel aufgeführt – in allen Ebenen von der Unternehmensstrategie, über die Unternehmensprozesse hin zu den im Unternehmen vorhandenen Informationsflüssen und -architekturen. Basis für die Schaffung der notwendigen Transparenz kann das Enterprise Architecture Management sein, das die Big Data Strategie in Form von datenbasierten Capabilities im Unternehmenskontext abbildet und damit eine Abstimmung dieser auf die Unternehmensstrategie gewährleistet.

2.2.1.1 EAM ein kurzer Überblick

Die Unternehmensarchitektur, Enterprise Architecture genannt, hat ihren Ursprung in einem 1987 von John Zachmann entwickelten Framework, das die Wichtigkeit der ganzheitlichen Betrachtung von Architekturen auf die Unternehmensebenen beschreibt. Das Framework ist Ausgangspunkt für viele andere Frameworks, wie das hier verwendete TO-GAF der Open Group und stellt Beschreibungskonzepte zur Darstellung der vielfältigen Schnittstellen von Informationssystemen sowie der Integration dieser in die Unternehmenswelt zur Verfügung. Ziel der Aktivitäten von Zachmann war, die steigende Komplexität der Unternehmenswelt durch geeignete Beschreibungskonzepte transparent und beherrschbar zu halten (Hanschke 2013, S. 68 f.).

Enterprise Architecture Management als ganzheitlicher Ansatz hilft die unterschiedlichsten Ebenen des Unternehmens in Form von Modellen auf unterschiedlichsten Ebenen aggregiert aufzunehmen und zu analysieren (vgl. Abb. 2.8).

Abb. 2.8 Ebenen und Gestaltungsobjekte der Unternehmensarchitektur (in Anlehnung an Aier et al. 2008b)

Strategieebene	<ul style="list-style-type: none"> • Strategische Unternehmensziele • Strategische Vorhaben/Projekte • Marktsegmente
Organisations- ebene	<ul style="list-style-type: none"> • Geschäftsprozesse • Organisationseinheiten • Rolle/Verantwortlichkeiten • Capabilities/Fähigkeiten
Integrations- ebene	<ul style="list-style-type: none"> • Applikationen • Fachliche Services • Informationsobjekte • Schnittstellen
Softwareebene	<ul style="list-style-type: none"> • Softwarekomponenten • Datenstruktur
IT-Infrastruktur- eneme	<ul style="list-style-type: none"> • Hardwarekomponenten • Netzwerkkomponenten • Software-Plattformen

Die so erfasste Enterprise Architecture beschreibt das Zusammenspiel zwischen den Modellen der Geschäftsprozesse, der Anwendungen und korrespondierenden Daten sowie deren wechselseitige Abhängigkeiten und liefert eine gemeinsame fachliche Sprache zwischen den IT- und Geschäftsverantwortlichen (Hanschke 2012, S. 1) (Aier et al. 2008b, S. 292) (Bitkom 2011, S. 6). Darüber hinaus ermöglicht die Enterprise Architecture eine Ausrichtung der Analysen auf die Unternehmensstrategie.

Mit steigender Unternehmensgröße bzw. bei einer sehr heterogenen Systemlandschaft nimmt ihre Bedeutung rapide zu, da sie die notwendige Transparenz im Unternehmen schafft (Hanschke 2013, S. 144 f.) (Aier, Riege, and Winter 2008b, S. 294). Genau diese Transparenz über die Informationsflüsse, die dokumentierte Ableitung von Capabilites des Unternehmens ist eine wichtige Voraussetzung für erste Big Data-Projekte, um auf der einen Seite die richtigen Fragestellungen aufzustellen und auf der anderen Seite die richtigen Daten zu identifizieren und zeitnah erheben zu können.

Ähnlich wie eine Landkarte hilft, den kürzesten Weg zwischen zwei Orten zu zeigen, kann ein EAM aufzeigen, welche wesentlichen Strategien, Capabilities, Prozesse, Systeme und Informationen von einer gewünschten Fragestellung durch Big Data tangiert werden. Auf diese Weise können die notwendigen Systeme, Daten und Verantwortlichkeiten für aus der Unternehmensstrategie abgeleitete Fragestellungen standardisiert identifiziert und umgesetzt werden. Mittels einer Analyse über die EAM Datenbasis und über die Visualisierungen der einzelnen Architekturebenen ergeben sich sich Fragestellungen wie (Hanschke 2013, S. 145 f.):

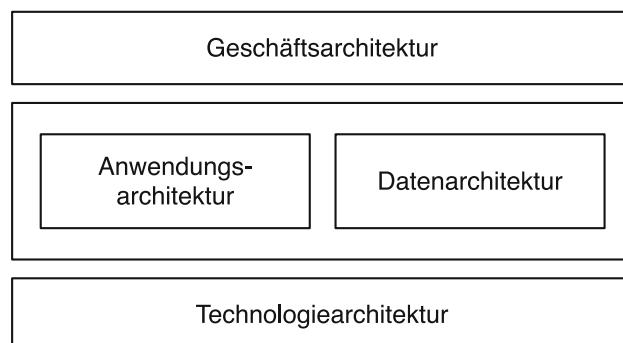
- Welche Geschäftsprozesse sind bei der Big Data-Analyse betroffen?
- Welche Systeme führen die notwendigen Daten für die Analyse?
- Wer sind die verantwortlichen Datenowner, die bei einer Analyse zur Einhaltung der Compliance mit eingebunden werden müssen?

Mittelpunkt des Enterprise Architecture Management ist die Unternehmensarchitektur, die die unterschiedlichen Fragestellungen der Stakeholder beantworten soll, indem mittels Listen und Visualisierungen Transparenz über die Abläufe und Abhängigkeiten im Unternehmen geschaffen wird (Hanschke 2013, S. 185 f.). Bei der Erarbeitung einer Enterprise Architecture wird im Rahmen der Geschäftsarchitektur eine Prozesslandkarte des Unternehmens in Form eines Big Pictures erstellt. Diese Prozesslandkarte bildet in einer Übersichtsdarstellung alle relevanten Elemente, wie Prozesse, Capabilities und Services eines Unternehmens und deren Wechselbeziehungen ab (Bayer 2013, S. 39 f.). Diese Vorgehensweise ermöglicht einerseits einen raschen Überblick über alle Unternehmenselemente und andererseits können nun von diesem Punkt aus die Detaillierungen der jeweiligen Architekturebenen je nach Nutzen-Kosten Relation vorgenommen werden (Aier et al. 2008a). Denn bei EAM-Projekten gilt, wie auch bei Big Data-Projekten: Eine Einführung sollte nicht global sondern eher agil, je nach Bereich, in dem der größte Nutzen gesehen wird, vorgenommen werden (Hanschke 2012, S. 22). Wesentliche Strukturelemente einer Enterprise Architecture, auch Domänen genannt, sind die Geschäftsarchitektur, die Informationsarchitektur, aufgeteilt in Anwendungsarchitektur und Datenarchitektur, sowie die Technologiearchitektur. In diesen Domänen werden die unterschiedlichen Ebenen des Unternehmens als Architekturschichten überführt.

Die Ebenen des Enterprise Architecture Managements

Das Enterprise Architecture Management dient zum Einen der standardisierten Darstellung und zum Anderen der Verwertbarkeit der dokumentierten Abläufe im Unternehmen. Damit die Erstellung des EAMs ebenfalls standardisiert vorgenommen werden kann, bietet es sich an, auf eines der vielzähligen Frameworks wie z. B. das „The Open Group Architecture Framework (TOGAF)“ zurückzugreifen (Hanschke 2013, S. 188 ff.) (Matt-

Abb. 2.9 Architekturebenen nach TOGAF (in Anlehnung an Niemann 2005)



hes 2011, S. 188 ff.). TOGAF bietet einen methodischen Rahmen zur Entwicklung der unterschiedlichen Unternehmensarchitekturen auf Basis von vordefinierten Komponenten sowie von definierten Vorgehensmodellen an. Basis der Unternehmensarchitektur nach TOGAF sind die bereits genannten vier Architekturebenen: Geschäftsarchitektur, Anwendungsarchitektur, Datenarchitektur und Technologiearchitektur (Hanschke 2013, S. 189) (Matthes 2011, S. 188 ff.).

Geschäftsarchitektur Die Geschäftsarchitektur beschreibt im Wesentlichen die betriebswirtschaftlichen Elemente eines Unternehmens. Darunter fallen die Strategie, die Capabilities, die Governance, die Organisation und die Geschäftsprozesse des Unternehmens.

Bei der Anwendung und Aufnahme der Geschäftsarchitektur sind vor allem die fachlichen Mitarbeiter bis hin zum Management beteiligt. Gemäß einem Top-Down Ansatz ist es ratsam, die Objekte der Geschäftsarchitektur in einzelnen zunächst wenig granulierten Schritten in Form von „Landkarten“ als Modell aufzunehmen, um durch einen iterativen Prozess eine Detaillierung vornehmen zu können. Wesentliche Hilfestellung für dieses Vorgehen ist die Schaffung der im vorhergehenden Abschnitt aufgeführten Prozesslandkarte, die Teil dieser Architekturebene ist. Sie zeigt die auf ein Minimum reduzierte Quintessenz des Unternehmens in abstrakter und zwischen den Objekten abhängiger Form auf (Aier et al. 2008a, S. 3).

Die Aufnahme der Geschäftsarchitektur sollte zunächst den Ist-Zustand der Unternehmensstrukturen und Prozesse abbilden und kann weiter als Basis für eine Modellierung eines Soll-Zustandes für Transformationsprojekte herangezogen werden.

Dabei ist zu beachten, dass nicht zwingend alle möglichen Elemente eines Frameworks wie TOGAF in seiner Gänze angewandt werden müssen. Es ist ebenso ausreichend, anhand der Zielrichtung des EAM-Einführungsprojektes im Vorfeld die notwendigen Gestaltungsobjekte zu identifizieren und dementsprechend einen schlanken Einführungsansatz zu wählen.

Die so erstellte Geschäftsarchitektur inklusive der Prozesslandkarte kann für Entscheidungsfindungen auf strategischer, operativer und taktischer Ebene, für Dokumentations- und Schulungszwecke, sowie zu Prozessverbesserungen und Prozessautomatisierungen herangezogen werden (Aier et al. 2008a, S. 4 f.).

Die Geschäftsarchitektur wird mit den unterliegenden Architekturebenen wie der Anwendungsarchitekturebene verknüpft und stellt ein zentrales Koordinationswerkzeug für die konsequente Ausrichtung der IT auf die Geschäftsanforderungen dar (Business/IT-Alignment) (Aier et al. 2008a, S. 3).

Informationsarchitektur Die Informationsarchitektur beinhaltet sowohl die Anwendungsarchitektur sowie die Datenarchitektur.

Die Anwendungsarchitektur zeigt die im Unternehmen vorhandenen und notwendigen Unternehmensanwendungen und deren Abhängigkeiten und Beziehungen in Form von Schnittstellen untereinander sowie zu den Kerngeschäftsprozessen auf. Die Verknüpfung zwischen den Unternehmensprozessen und den sie unterstützenden Applikationen stellt

den Link zwischen der Geschäftsarchitektur und der Anwendungsarchitektur dar (Niemann 2005, S. 77 f.).

Eine Kategorisierung der Anwendungen wird anhand ihrer fachlichen Funktionalität, wie z. B. die Kategorie rechnungslegungsrelevante Systeme, sowie der durch sie zu verarbeiten Informationen, wie z. B. Produktionsdaten, vorgenommen (Hanschke 2012, S. 180).

Ein echter Mehrwert kann zusätzlich entstehen, wenn die im Rahmen von EAM erhöhen Objekte in einer Datenbank zur späteren Verwaltung, Suche und Dokumentation gespeichert werden (Keuntje and Barkow 2010, S. 192 f. u. 272). Dies erlaubt zusätzlich die Nutzung dieser Daten in anderen Anwendungen, beispielsweise Monitoringsystemen und Configuration Management Datenbanken in der IT.

Auch hier gilt: In einem ersten Schritt wird eine Ist-Aufnahme der Anwendungsarchitektur vorgenommen, die dann zur Optimierung in eine Soll-Architektur überführt werden kann. Eine so verwendete Soll-Architektur wird auch als Roadmap für die Einführung, Erweiterung oder Ablösung von Applikationen verwendet (Matthes 2011, S. 192).

In der Datenarchitektur werden alle Daten inklusive ihrer Beziehungen untereinander dargestellt, die für die Durchführung der Geschäftsprozesse benötigt werden. Wichtig bei der Erfassung dieser Objekte ist eine stabile, vollständige, konsistente und verständliche Darstellungsform (Matthes 2011, S. 192 f.).

Auf Basis beider Architekturen können nun die Informationen gemäß Adressaten und Bedürfnissen gruppiert werden und den verschiedenen Rollen mit gleichem Informationsbedarf zugeordnet werden. Gerade die Informationsarchitektur ist ein wesentlicher Ansatzpunkt für Big Data-Projekte, da erstens, über die Geschäftsarchitektur eine Verknüpfung zu der Informationsarchitektur geschaffen werden kann, und zweitens, über die Zuordnung und Gruppierung der Daten zu Informationsgruppen weitere Fragestellungen, die mittels diesen Informationen ausgewertet werden sollen, identifiziert werden können.

Technologiearchitektur Die letzte Ebene des TOGAF stellt die eigentliche physische Ebene der IT dar. Hier werden alle Architekturelemente zum technischen Aufbau und Betrieb der IT Infrastruktur ermittelt und dokumentiert (Matthes 2011, S. 193). Die Technologiearchitektur wird dabei mit den Anwendungen und den Daten verknüpft, was Aussagen über die physikalische Verteilung der Anwendungen und Daten ermöglicht. Weiterhin erlauben Abhängigkeitsanalysen die Darstellung des Einflusses einer Hardwarekomponente auf alle darüber liegenden Ebenen und stellen daher ein adäquates Hilfsmittel der Risikoanalyse und der Investitionsplanung dar.

EAM als Teil der Unternehmensentwicklung

Eine Aufnahme der drei Architekturebenen im Unternehmen kann zu einer Unterstützung der Unternehmensentwicklung führen. Unter Unternehmensentwicklung versteht Bleicher (Bleicher 2011, S. 457 f.) (Marek 2010, S. 15) die Evolution eines ökonomisch orientierten sozialen Systems im Spannungsfeld von Forderungen und Möglichkeiten der Um- und Inwelt. Er sieht vor allem die Schaffung eines höheren Nutzens durch die Inanspruch-

nahme von strategischen Erfolgspositionen als wesentlichstes Element der Unternehmensentwicklung. Somit stellen die Erfolgspositionen einen wesentlichen Aspekt für eine Weiterentwicklung des Unternehmens dar. Wenn in einem EAM die wertschöpfenden Bereiche und Aspekte des Unternehmens mittels geeigneter Visualisierungswerzeuge transparent dargestellt werden und vor allem auch die eigenen Fähigkeiten und Potenziale des Unternehmens, werden Verbesserungspotenziale ersichtlich und führen über einen Transformationsprozess von der Ist- zur Soll-Architektur zu möglichen neuen Erfolgspositionen als maßgebliche Treiber für eine Veränderung des Unternehmens (Thielscher 2010).

Big Data-Analysen unterstützen das Heben und Entwickeln von neuen gewinnversprechenden Geschäftsmodellen und die Weiterentwicklung des Unternehmens dann, wenn bestimmte Fähigkeiten in der Unternehmens-DNA und damit in allen Abteilungen implementiert werden. Zu den Fähigkeiten gehören vor allem technische und methodische Fähigkeiten im Umgang mit Daten im Generellen und mit großen Datenmengen im Spezifischen. Dies bedeutet, dass die bisherige durch EAM vorangetriebene Optimierung und Standardisierung der internen Abläufe durch eine wesentlich stärkere Fokussierung auf die Datenebene für eine datenaustauschende Kultur, soweit gesetzliche Grenzen dies zulassen, ergänzt werden muss. Teil dieses Anpassungsprozesses sind vier wichtige Strategiebereiche, die im Rahmen der Entwicklung zu einem datenversierten Unternehmen zwingend beachtet werden müssen (Newman 2012):

1. Strategie: Schaffung einer datenorientierten Business Strategie.
2. Kultur: Veränderungen der Unternehmenskultur, die den Datenaustausch und das Vertrauen untereinander fördert.
3. Personal: Schaffung des notwendigen Humankapitals, um Kompetenzlücken in der Anwendung von Big Data schließen zu können.
4. Technologie: Schaffung spezifischen Wissens von Big Data Architekturen in der IT, aber auch bei den EA Architekten. Dabei sollte nicht nur eine technische Fokussierung auf der Auswertungsebene betrachtet werden, sondern vielmehr auch geeignete technische Visualisierungsaspekte der Ergebnisse mit berücksichtigt werden.

2.2.1.2 Competitive Advantage durch Big Data

Betrachtet man die Bedeutung der IT innerhalb der Organisationen in den letzten 30 Jahren zeigt sich ihre enorme Entwicklung vom unterstützenden Prozess hin zu einem steuernden Prozess (Gadatsch 2012, S. 1618).

Kaum ein Prozess, kaum ein Unternehmen kann heute ohne IT auskommen. Die Integration dieser als Wechselbeziehung zwischen der IT und den Businessbereichen in Form des Business IT Alignements ermöglicht den Unternehmen ein enormes Potenzial im Bereich der Kostenersparnis und der Identifikation neuer Geschäftsmodelle auf Basis der IT. Gerade die Möglichkeit, nun in allen Prozessen auf IT gestützte Abläufe zurückgreifen zu können und die Tatsache, dass die Kosten für Speicherplatz rapide gesunken sind, ermöglicht es, enorme Datenmengen über die internen Abläufe und über seine Interakti-

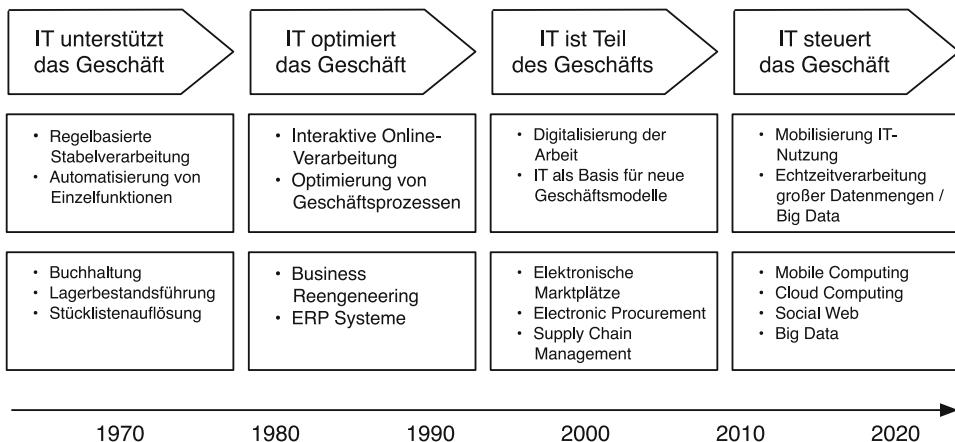


Abb. 2.10 Entwicklungsphasen der IT (in Anlehnung an Gadatsch 2012)

onspartner außerhalb des Unternehmens, seien es Lieferanten, Kunden, Kapitalgeber usw., zu speichern und gegebenenfalls für das Unternehmen auswertbar zu machen. Die IT entwickelt sich somit aus der reinen Teifunktion im Business-IT-Alignement hin zum klaren Money Maker. Sie durchdringt das Unternehmen nicht nur durch die reine unterstützende Funktion in den Prozessen sondern ist maßgeblicher Treiber des Geschäftes auf strategischer, taktischer und operativer Ebene. Durch die Durchdringung aller Bereiche des Unternehmens mit IT in den letzten 30 Jahren wurde somit erst die Möglichkeit geschaffen, Daten als vierten Produktionsfaktor zu entwickeln.

Einen Nutzen in Form eines Wettbewerbsvorteils gegenüber den konkurrierenden Unternehmen wird nur derjenige erreichen, der es schafft, seine internen und extern zur Verfügung stehenden Daten zielorientiert und zeitnah verarbeiten zu können und die Ergebnisse in angemessener Form und „at the Fingertips“ (Gates 1994) zu präsentieren. Die durch Big Data geschaffenen technischen Möglichkeiten liefern die Basis, Daten in Echtzeit zu verarbeiten und adressatengerecht zu visualisieren. Nutzenbringend für das Unternehmen wird dies aber nur, wenn die Möglichkeiten von Big Data in Kombination mit den internen Fähigkeiten für unternehmensweite strategische, taktische und operative Fragestellungen genutzt wird (Gadatsch 2012). Dabei geht es nicht nur darum, Big Data einmalig auf spezifische Fragestellungen anzuwenden, sondern es ist eher ein iterativer Prozess, der durch die Ergebnisse einer Iterationsstufe weitere bzw. spezifischere Fragen aufwirft und somit die unternehmensinterne Big Data-Methodik stetig weiter entwickelt (Conrads 2013). Die für diese Entwicklung notwendigen Fähigkeiten werden in der Unternehmensarchitektur mittels datenorientierten Capabilities abgebildet und unterstützt.

Identifikation datenbasierter Capabilities

Um als Unternehmen eine Differenzierung und damit einen Wettbewerbsvorteil zu den anderen Marktteilnehmern zu erreichen, müssen Unternehmen ihre Produkte, Dienstleis-

tungen und Prozesse im Vergleich zur Konkurrenz mit höherer Qualität, schneller oder kostengünstiger erbringen. Da sich durch die permanente Weiterentwicklung aller Marktteilnehmer diese Vorteile aufgrund von Imitatoren ständig annähern, müssen Unternehmen Fähigkeiten besitzen, die notwendigen Wettbewerbsvorteile stetig weiter- bzw. neue entwickeln zu können (Bitkom 2011, S. 9). In Bezug auf das Enterprise Architecture Management nennt man diese Fähigkeiten Capabilities, die Basis zur Erreichung der Geschäftsziele sind. Sie stellen eine feste fachliche Struktur dar und sind losgelöst zu den Geschäftsprozessen (Hanschke 2012, S. 9). Capabilities sind somit keine Elemente eines Prozesses oder Produktes sondern die Fähigkeiten im Unternehmen, Prozesse effektiv zu betreiben, Produkte zu verbessern und neue Methoden im Unternehmen einzuführen (Freitag, Matthes, and Schulz 2011). Entscheidend ist, dass Capabilities nicht den eigentlichen Wettbewerbsvorteil darstellen, sondern diesen erst ermöglichen (Thielscher 2010).

In einer Analyse des Fraunhofer Instituts für Intelligente Analyse- und Informati-onssysteme in Bezug auf das Innovationspotenzial wurden vielseitige Ziele aufgeführt, die mittels Big Data unterstützt werden. Darunter fallen branchenabhängig Steigerung von Umsätzen durch Fokussierung auf die Bereiche Produktqualität, Marketing, Vertrieb und Kundenbetreuung und die Einsparung von Kosten durch Verbesserungen der internen Abläufe (Fraunhofer IAIS 2012). All diese Ziele dienen dem Zweck der Wettbewerbsverbesserung. Um diese Ziele nun im eigenen Unternehmen erreichen zu können, müssen die eigenen Fähigkeiten im Unternehmen an Big Data angepasst werden. Es sind somit datenorientierte Capabilities zu identifizieren und gegebenenfalls aufzubauen, denn nicht alle Fähigkeiten müssen zwangsläufig im Unternehmen schon vorhanden sein. Die für Big Data notwendigen Fähigkeiten sind eine Mischung aus technischen aber auch betriebswirtschaftlichen Kenntnissen, wie Mathematik, Statistik, Kenntnisse über betriebliche Abläufe sowie analytische Kompetenz.

Zunächst gilt es im Unternehmen bei der Modellierung der Unternehmensarchitektur eine Geschäftsfähigkeitslandkarte (Business Capability Map, BCM) aufzustellen, die alle Ist-Fähigkeiten des Unternehmens widerspiegelt und somit den Rahmen für eine Organisationsanpassung liefert. Die so erstellte Landkarte zeigt dem Unternehmen grob-granuliert die vorhandenen Fähigkeiten auf, die zunächst unabhängig von den jeweiligen Prozessen, Anwendungen und Informationen sind. Um nun eine datenbasierte analytische Denkweise im Sinne von Big Data im Unternehmen zu implementieren, sollten die vorhandenen Capabilities bewertet werden und um fehlende notwendige datenorientierte Capabilities ergänzt werden. Auf diese Weise wandelt sich die Ist-Landkarte in eine Soll-Landkarte, die dem Unternehmen ermöglicht, Kompetenzlücken zu identifizieren und mittels Maßnahmen wie Schulungen, Aufklärungen und ähnlichem im Unternehmen die notwendige DNA für Big Data zu implementieren. Das Unternehmen kann sich somit zu einem datenversierten bzw. -orientierten Gebilde entwickeln und sich permanent kooperative Wettbewerbsvorteile erarbeiten und somit auch den maximalen Vorteil langfristig aus dem vierten Produktionsfaktor Daten ziehen.

Ergänzend bildet die BCM ein Referenzmodell in Bezug auf die Fähigkeiten im Unternehmen und stellt die Basis für Know-How Transfer in der gesamten Organisation dar (Keuntje and Barkow 2010, S. 353 f.).

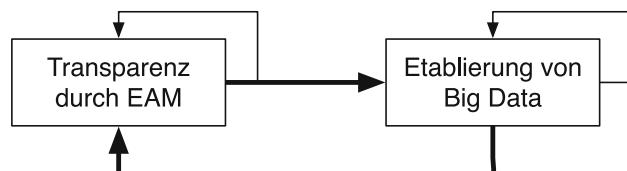
Um Big Data als ein Mittel zur Wettbewerbsdifferenzierung im Unternehmen zu implementieren ist es zunächst notwendig Klarheit, über das eigene Unternehmen mittels dem Aufbau einer Ist-Architektur zu schaffen. Darauf aufbauend kann der Wandel zur einer gewünschten Soll-Architektur vorangetrieben werden.

2.2.2 EAM als Ausgangspunkt für die Etablierung von Big Data im Unternehmen

Die Einführung von neuen Konzepten oder Technologien stellt bei verflochtenen Gebilden wie Unternehmen oftmals eine finanzielle, kulturelle und organisatorische Herausforderung dar, was sich insbesondere in der Komplexität dieser Aufgabe niederschlägt. Dabei beeinflussen zwei Faktoren diese Komplexität maßgeblich: Dynamik und Intransparenz der Aufgabe. Während die Dynamik – also Veränderungen des Einführungsprojektes – stark von den gesetzten Rahmenbedingungen abhängt und nur wenig beeinflusst werden kann, ist es möglich, den zweiten Faktor Intransparenz positiv zu beeinflussen (Dörner 2009, S. 58 ff.). Dabei kann Enterprise Architecture Management als Werkzeug zur Schaffung des notwendigen Überblicks und zur Reduktion der Intransparenz fungieren (Aier, Riege, and Winter 2008b, S. 299). Weiterhin legt EAM die Struktur des Unternehmens offen und sorgt somit für das benötigte Strukturwissen, also das Wissen, über die Art und Weise, wie die Elemente des Unternehmens zusammenhängen und wie sie sich gegenseitig beeinflussen (Dörner 2009, S. 64.). Für die erfolgreiche Einführung von Big Data im Unternehmen ist es folglich sinnvoll, zunächst die notwendige Transparenz in Form einer ganzheitlichen Unternehmensarchitektur herzustellen und sich anschließend der Integration von Big Data in die DNA des Unternehmens zu widmen (vgl. Abb. 2.11).

Bei der Einführung gilt zu bemerken, dass sowohl bei der Einführung von EAM und Big Data ein agiles und evolutionäres Vorgehen (zyklischer Prozess mit kontinuierlichen Verbesserungen) zu bevorzugen ist (Lux, Wiedenhöfer, and Ahlemann 2008, S. 27) (Hanschke 2012, S. 35) (Bitkom 2012, S. 29). Big Data beeinflusst jedoch auch auf allen Ebenen die Unternehmensarchitektur (Vgl. Baron 2013, S. 181 f.) und so ist auch zwischen EAM und Big Data ein wechselseitig iteratives Vorgehen angebracht, um jeweils die notwendige Transparenz (wieder-)herzustellen und so eine aktuelle Sicht für die folgen-

Abb. 2.11 Einführung von EAM und Big Data



den Iterationen bereitzustellen. Die nachfolgenden Unterkapitel stellen sowohl die EAM als auch die Big Data- Einführung vor und versuchen beide miteinander zu verknüpfen. Insbesondere die Big Data-Einführung wird später in diesem Band im Detail vorgestellt.

2.2.2.1 Einführung und Entwicklung einer Unternehmensarchitektur

Die Einführung und Entwicklung einer Unternehmensarchitektur stellt einen großen Aufwand und damit verbunden eine Investitionsleistung des Unternehmens dar (Matthes 2011, S. 25). Entsprechend kann es als sinnvoll erachtet werden, sich über die Nutzung der eingangs erwähnten Rahmenwerke wie TOGAF Gedanken zu machen. Diese bieten oftmals eine gute Basis, sollten jedoch auch auf die Anforderungen des Unternehmens bzw. der jeweiligen Domäne angepasst werden (Vogt et al. 2011, S. 4) (Lux, Wiedenhöfer, and Ahlemann 2008, S. 19).

Neben einem adäquaten Architekturrahmen sollte jedoch auch ein geeignetes Vorgehensmodell für die Einführung von EAM Beachtung finden. Richtig eingesetzt können Vorgehensmodelle die Einführung beschleunigen, die Qualität der Unternehmensarchitektur erhöhen und die nachfolgenden Einführungsrisiken reduzieren (Lux, Wiedenhöfer, and Ahlemann 2008, S. 19 f.):

- Probleme der Dokumentation durch die hohe Komplexität und Kompliziertheit der Unternehmensarchitektur,
- Aufwendige Beschaffung der Informationen über die Architektur des Unternehmens,
- Mangelnde Qualität der erhobenen Informationen,
- Probleme beim Zusammenwirken von IT- und Fachabteilung,
- Mangelndes Commitment der Beteiligten (Nutzen nicht ausreichend bekannt, Widerstand durch Machtverlust, etc.).

Bei der Betrachtung der Vorgehensmodelle gilt es generische und spezifische Vorgehensmodelle für einzelne Rahmenwerke zu unterscheiden. Letztere lassen in der Regel eine allgemeine Anwendbarkeit vermissen, sind jedoch dann für die Einführung ideal, wenn im Anschluss das dazugehörige Rahmenwerk genutzt werden soll. Beispielhaft wäre hier die TOGAF Architecture Development Method (ADM) zu nennen. Diese besteht aus neun Phasen und ist unternehmens- und brachenneutral gestaltet (vgl. Abb. 2.12). Jede Phase an sich ist intern iterativ gestaltet und zudem sind alle Phasen in einem iterativen Modell angeordnet.

Einer Vorauswahl eines Architekturrahmens soll sich an dieser Stelle jedoch entzogen werden und somit wird nachfolgend das Vorgehen auf eine generische Art und Weise vorgestellt. Die in der Literatur verfügbaren Ansätze unterscheiden sich hierbei oftmals nur oberflächlich und zeigen grundsätzlich einen sehr ähnlichen Pfad auf (z. B. Lux, Wiedenhöfer, and Ahlemann 2008; Hanschke 2012). Angelehnt an Hanschke (vgl. Hanschke 2012) werden entsprechend nachfolgend die drei Phasen

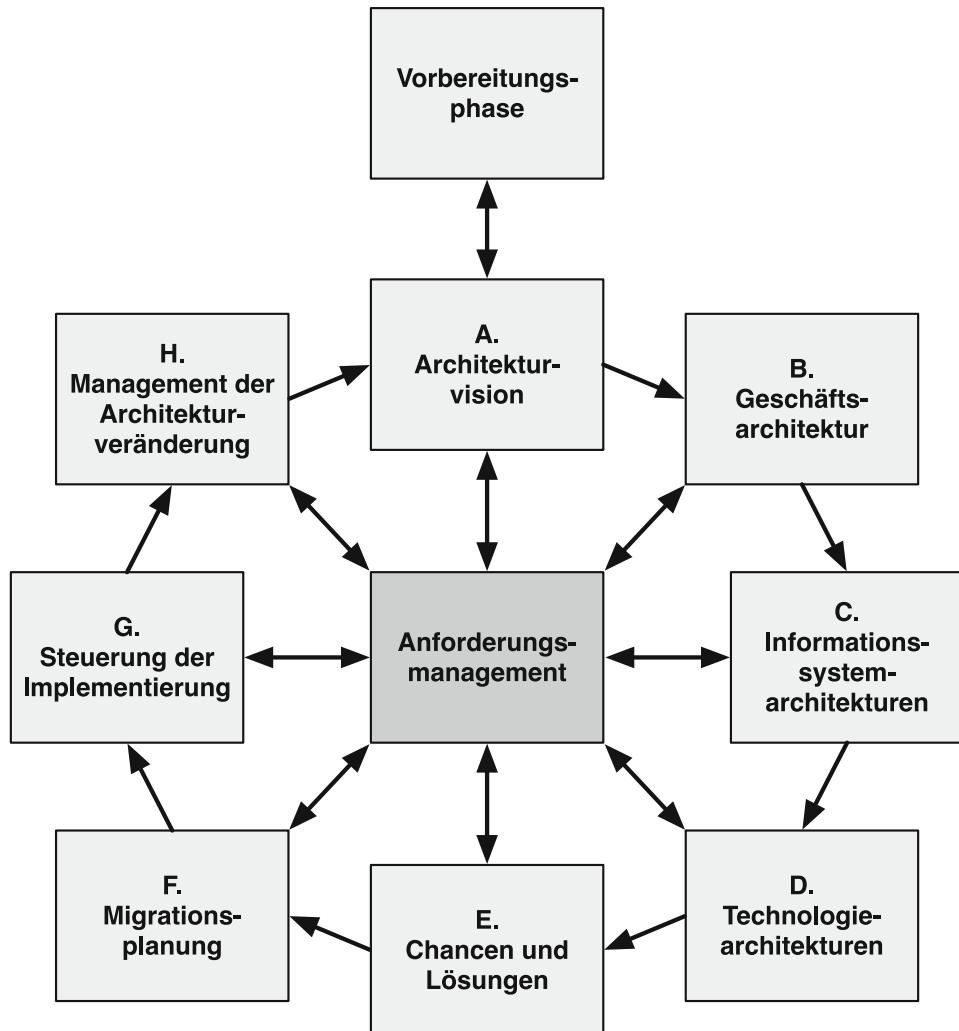


Abb. 2.12 TOGAF ADM (in Anlehnung an The Open Group 2010)

1. Konzeptionierung,
2. Pilotierung und
3. organisatorische Implementierung

vorgestellt und die wichtigsten Aktivitäten nach Hanschke und Lux et al. (Lux et al. 2008; Hanschke 2012) am Beispiel des fiktiven Unternehmens Stadtwerke Frankenbronn erläutert. Dabei gilt auch hier eingangs zu erwähnen, dass die Phasen nicht einmalig durchlaufen werden, sondern in einer agilen Vorgehensweise bei Bedarf – auch mehrmals – wiederholt werden (vgl. Abb. 2.13). Man spricht hierbei von sogenannten Iterations- oder Ausbaustufen.

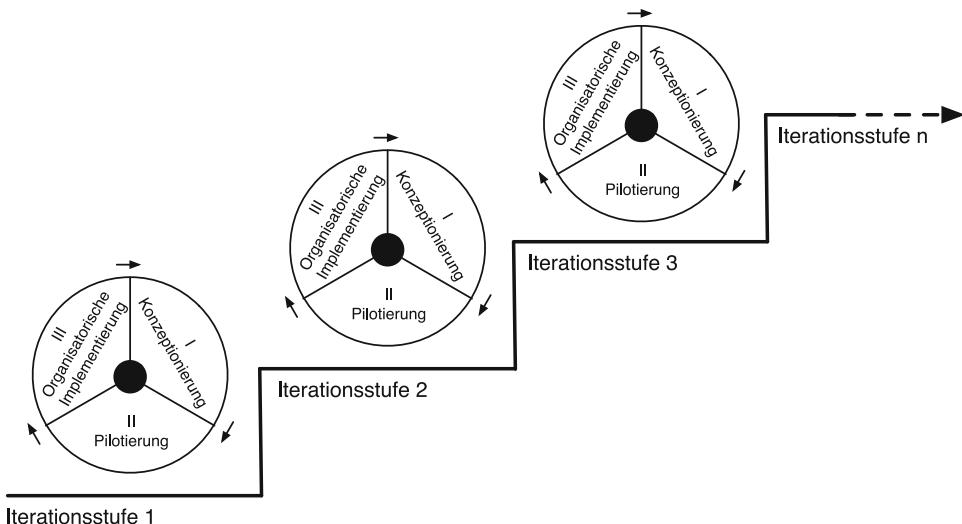


Abb. 2.13 Vorgehensmodell für die Einführung von EAM (in Anlehnung an Hanschke 2012)

Fallbeispiel: Stadtwerke Frankenbronn

Die Stadtwerke Frankenbronn sind ein zukunftsorientierter, mittelständischer Energiedienstleister, der regional rund 50.000 Kunden mit Strom, Fernwärme, Gas und Wasser versorgt. Daneben werden im kommunalen Auftrag das Facility Management der Stadt Frankenbronn durchgeführt und Parkierungseinrichtungen, Erdgas- und Stromtankstellen, sowie mehrere Hallen- und Freibäder betrieben. Mit derzeit rund 420 Beschäftigten erwirtschaften die Stadtwerke Frankenbronn einen Umsatz von ca. 230 Mio. Euro p. a. Mit innovativen Konzepten, zielgerichteten Einkaufs- und Vertriebskooperationen, sowie der Realisierungen gemeinsamer Kraftwerksprojekte im Bereich erneuerbarer Energien konnte die Marktposition des Energiedienstleisters in den vergangenen Jahren kontinuierlich gestärkt werden. Diese Bemühungen wurden mit zahlreichen Auszeichnungen wie Energiekommune, Energiemanager des Jahres oder dem deutschen Rechenzentrumspreis honoriert.

Als Betreiber des regionalen Stromnetzes (Bilanzkreisverantwortlicher) übernehmen die Stadtwerke Frankenbronn eine weitere anspruchsvolle Aufgabe: Um die Stabilität des Stromnetzes zu gewährleisten, müssen die Stadtwerke dafür Sorge tragen, dass die eingespeiste Strommenge zu jeder Zeit dem nachgefragten Stromkonsum der Verbraucher entspricht. Dafür wurde in der Vergangenheit der Stromverbrauch am Tag zuvor für jede Stunde prognostiziert und bei Bedarf Strom an der Börse eingekauft. Mit der Energiewende in Deutschland hat sich diese Situation verschärft. Viele Konsumenten produzieren heute auch Strom innerhalb des regionalen Netzes. Somit beschränkt sich die Prognose nicht mehr nur wie bisher auf den Verbrauch, sondern muss nun auch die Produktion der dezentralen Stromerzeuger berücksichtigen (Küller und Hertweck 2013).

Phase 1: Konzeptionierung

Ziel der Konzeptionierung ist unter anderem die iterative Ableitung der Unternehmensarchitektur und der EAM-Governance. Wie zuvor erwähnt, bietet das agile Vorgehen dabei im Gegensatz zu einer Big-Bang-Einführung den Vorteil vieler Feedback-Schleifen. Durch die gezielte Einbindung aller am EAM-Projekt beteiligten Personen wird dabei sichergestellt, dass das nötige Fachwissen zur Bewertung des Vorgangs herangezogen werden kann. Nachfolgende Auflistung zeigt die wichtigsten Aktivitäten dieser Phase in Anlehnung an Hanschke und Lux et al. (Lux, Wiedenhöfer, and Ahlemann 2008; Hanschke 2012) auf. Dabei müssen diese Aktivitäten nicht zwangsweise sequentiell ablaufen, sondern erfordern teilweise sogar zwingend eine parallele oder iterative Vorgehensweise.

I – Aufsetzen des EAM-Projektes

- Identifikation der Projektsponsoren und Unternehmensarchitekten
- Definition der Soll-Vision und (langfristige) EAM-Ziele
- Definition der Rolle von EAM in Planungs- und Steuerungsprozessen
- Projektorganisation (Aufgaben, Mitarbeiter, Beziehungen)

II – Kontext und Stakeholder Analyse

- Ermittlung der Stakeholder (Betroffene, Nutznießer), sowie Fach- und Machtpromotoren
- Ermittlung der Ziele und Fragestellungen der Stakeholder
- Ermittlung des Bedarfs an Visualisierungen

III – Abgleich mit der Realität

- Definition von Metriken zur EAM-Erfolgsmessung
- Bestimmung der Ausgangslage (IST-Situation) unter Anwendung der definierten Metriken
- Analyse der Datenbeschaffung (Woher bekommt man welche Daten?)

IV – EA-Konzeption

- Anforderungsanalyse
- Festlegung der Einführungsstufen
- Priorisierung der Fragestellungen
- Ermittlung von Architekturprinzipien aus den langfristigen EAM-Zielen und der IT-Strategie
- Optional: Sichtung und Auswahl eines EA-Frameworks
- Definition eines Metamodells für die Unternehmensarchitektur (Elemente, Umfang, Detailierung der Visualisierung) und Dokumentation des Metamodells mit verständlichen Beispielen

- Definition von Auswahlkriterien für die Werkzeugauswahl, Sichtung verfügbarer Lösungen, Wahl und Beschaffung des Werkzeugs

V – EAM-Governance

- Festlegung der Analyse-, Planungs- und Steuerungsinstrumente
- Definition von Richtlinien (Modellierung, Visualisierung, Pflege)
- Organisation innerhalb des EAM festlegen
- Entwicklung von operativen und strategischen EAM-Prozessen (Prozessschritte, beteiligte Rollen, In-/Output)

VI – Validierung

- Etablierung der Werkzeugunterstützung und testweise Datenbefüllung
- Validierung der Konzeption anhand eines repräsentativen Beispiels

Fallbeispiel: Stadtwerke Frankenbronn

Der Wandel der Energiewirtschaft hat auch starke Auswirkungen auf die Organisation, die Abläufe und die IT des mittelständischen Energiedienstleisters. Der Bedarf an einer ganzheitlichen Transparenz wird von der IT als auch der Geschäftsführung gleichermaßen erkannt und soll nun durch ein Projektteam unter der Leitung des CIO adressiert werden. Im Rahmen von mehreren Workshops und diversen Interviews erarbeitet sich das EAM-Team ein erstes Konzept, das sich an gemeinsam definierten Zielen und Visionen ausrichtet. Man einigt sich auf ein Rahmenwerk und gemeinsam mit einem Softwarehersteller konnte ein geeignetes Metamodell (vgl. Abb. 2.14) entwickelt und in die Modellierungsumgebung implementiert werden. Das entwickelte Metamodell orientiert sich dabei an bekannten Referenzmodellen, wurde aber speziell auf die Bedürfnisse der Stadtwerke angepasst und dabei teilweise auch stark reduziert. So enthält das Metamodell heute kein Modelltyp „Projekte“ mehr, da diese im Projektportfolio verwaltet werden, jedoch bietet das Metamodell die Möglichkeit, Capabilities als „datenbasierte Capabilities“ zu kennzeichnen. Bei der Auswahl des Werkzeuges spielte neben der Usability auch die Möglichkeit der Erfassung von Metadaten, des Exports der Daten in andere Systeme, die Analysefunktionen und auch die Möglichkeiten der Veröffentlichung der Modelle für „normale Nutzer“ eine große Rolle.

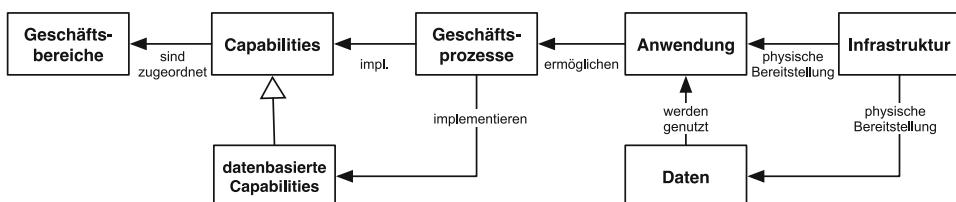


Abb. 2.14 Beispiel: Vereinfachte Darstellung eines Metamodells

Phase 2: Pilotierung

Bevor eine Ausbaustufe auf das Unternehmen angewendet werden kann, sollte im Rahmen der Pilotierung die jeweilige Stufe auf einzelne Teile des Unternehmens angewendet werden, um das definierte Konzept zu erproben, zu verfeinern und weitere Vorgehens- sowie Optimierungsschritte abzuleiten. Zu beachten ist hierbei, dass es sich um ein repräsentatives Unternehmenssegment (ausreichende Größe, realistische Komplexität) handelt, welches die nötigen Informationen zur Bewertung der betrachteten Ausbaustufe liefert.

I – Pilotierung

- Erprobung der Konzeption (Ausbaustufen, Visualisierung, Governance, etc.) der aktuellen Ausbaustufe durch
 - Erhebung der Geschäfts- und IT-Architektur mithilfe von Analysetechniken (Pilot)
 - Modellierung der erfassten Elemente
 - Erfassung ergänzender Informationen
 - Aufbau von Beziehungen zwischen den Elementen der EA
- Überprüfung der gewonnenen Ergebnisse auf deren Tauglichkeit in Projekten, dem Projektportfoliomanagement oder der strategischen Planung.

II – Optimierung

- Optimierung der Konzeption, insbesondere der Unternehmensarchitektur, aber auch der Prozesse und Richtlinien der EAM-Governance.

Fallbeispiel: Stadtwerke Frankenbronn

Nach abgeschlossener Konzeptionsphase folgt bei den Stadtwerken Frankenbronn ebenfalls die Pilotphase. Für diese Phase fokussiert man sich auf den Geschäftsbe- reich Strom als „Pilot“, da dieser in der strategisch geplanten Big Data Einführung ebenfalls im ersten Schritt adressiert werden soll. Die relevanten Strukturen werden in Workshops mit externer Unterstützung erhoben und anschließend durch Interviews mit relevanten Stakeholder verfeinert. Dabei verfolgt man einen Top-Down-Ansatz und bearbeitet die folgenden Ebenen (vgl. Abb. 2.15) der Unternehmensarchitektur nacheinander:

Geschäftsarchitektur Die Geschäftsarchitektur der Stadtwerke Frankenbronn wird zunächst mit drei relevanten Objekttypen aufgestellt: Unternehmensbereiche, Capabilities und Geschäftsprozesse. Dabei gilt, dass die identifizierten Capabilities einem oder mehreren Geschäftsbereich(en) angehören und jeweils durch einen oder mehrere Geschäftsprozess(en) implementiert werden. So gehört die Capability „Bilanzkreismanagement“ dem Geschäftsbereich „Strom“ an und wird von den Prozessen „Prognose“ und „Steuerung“ implementiert.

Informationsarchitektur Bei der Anwendungsarchitektur konzentriert man sich bei den Stadtwerken Frankenbronn auf jene Anwendungen, die für die Ausführung der Geschäftsprozesse unabdingbar sind. Diese Anwendungen werden untereinander in

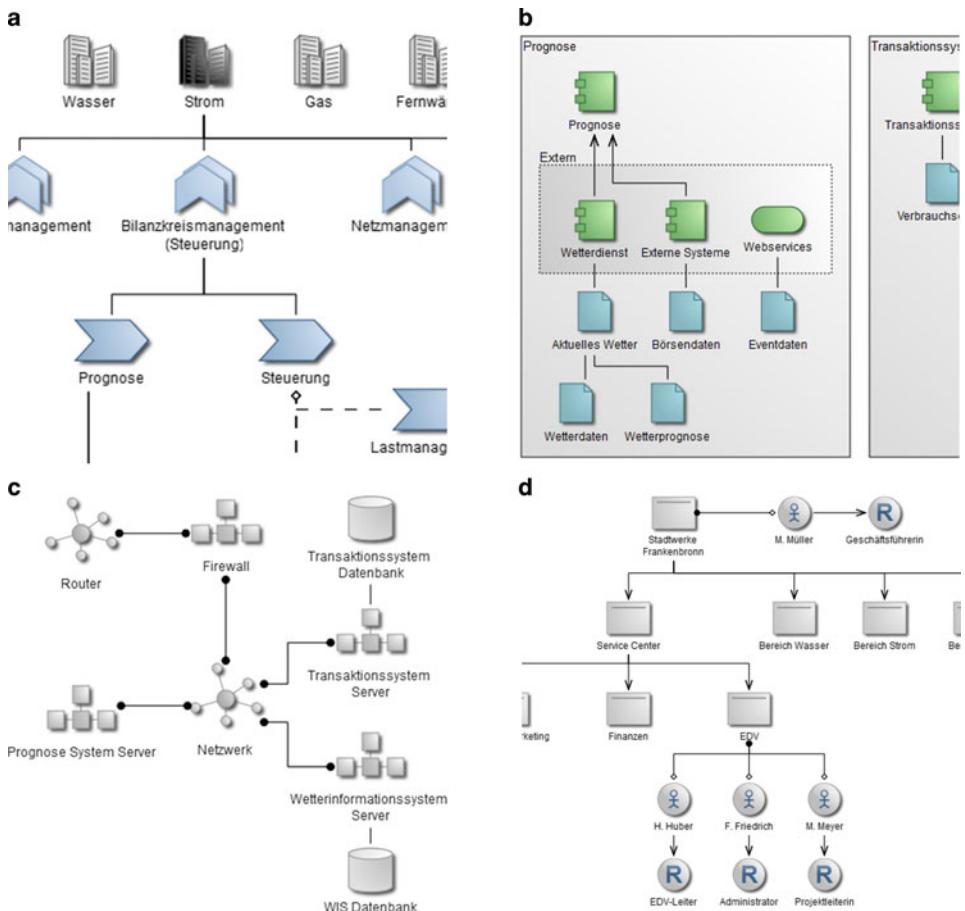


Abb. 2.15 Ausschnitte aus der Unternehmensarchitektur der Stadtwerke Frankenbronn, **a** Geschäftsarchitektur; **b** Informationsarchitektur; **c** Technologiearchitektur; **d** Organisationsstruktur

Form von Schnittstellen verknüpft. Zudem wird eine Verbindung zu den Geschäftsprozessen hergestellt und damit die Abhängigkeiten zwischen Geschäftsprozessen und Anwendungen aufgezeigt. In einem zweiten Schritt werden die jeweiligen Daten des Stadtwerks in grober Granularität (z. B. „Adressdaten“, „Verbrauchsdaten“) als Datenarchitektur modelliert und mit den Anwendungen verknüpft, die diese Daten verwalten oder auch nutzen. Ein Benefit dieser grafischen Modellierung ist dabei, dass sich klar zeigt, dass etliche Daten an mehreren Stellen innerhalb der Organisation vorhanden sind. Diese mögliche Fehlerquelle soll nun mittelfristig im Rahmen des Stammdatenmanagements behoben werden.

Technologiearchitektur Durch die IT-Abteilung werden schlussendlich alle physischen als auch virtuellen Server, sowie die verwendeten Netzwerkkomponenten in-

nerhalb des Rechenzentrums als Architektur der Infrastruktur hinzugefügt. Auf die Erfassung der sehr homogenen Rechnerlandschaft auf Seiten der Anwender verzichtet man in dieser ersten Phase bewusst. Jedoch erfasst man die „inneren Werte“ der jeweiligen Komponenten recht detailliert, da eine Liste der Infrastruktur bis dato nur sehr dürftig gepflegt wurde und zukünftig das Modell diese ersetzen soll.

Zur Herstellung einer übergreifenden Sicht werden die Elemente zwischen den einzelnen Ebenen (z. B. Prozess nutzt Anwendung) miteinander, aber auch mit den Mitarbeiterrollen (z. B. Administrator als Verantwortlicher für einen Server) der vorhandenen Organisationsstruktur der Stadtwerke verknüpft. Die Stakeholder selbst können dabei über ein Onlineportal ergänzende Informationen (z. B. Beschreibungen von Aktivitäten oder relevante Dokumente) den einzelnen Elementen hinzufügen, aber auch die Ergebnisse validieren und Fehler der Modellierung an das EAM-Team melden. Als Ergebnis der Pilotphase können nun fünf Modelle als „Durchstich“ zusammen mit einem aggregierten Modell präsentiert werden. Dies erlaubt die Überprüfung des Konzeptes an einem überschaubaren Teilbereich des Unternehmens und ermöglicht zudem die Identifikation von zahlreichen Verbesserungsvorschlägen.

Phase 3: Organisatorische Implementierung

Nach den im Projektformat erfolgreich abgelaufenen Konzeptionierungs- und Pilotierungsphasen erfolgt letztendlich die Umsetzung (Roll-Out) der Vision im gesamten Unternehmen.

I – Planung der Einführung

- Entscheidung über Einführungsstrategie (z. B. Step-by-Step oder Big Bang)
- Entwicklung eines Schulungskonzeptes und -unterlagen
- Bereitstellung und Betrieb des EAM-Werkzeuges (Administration, Wartung, Support, etc.)

II – Kommunikation

- Erstellung eines Kommunikationsplans (Medien, Aktionen, Zeitvorgaben, Zuständigkeiten)
- Definition von Schlüsselbotschaften
- Etablierung eines Feedbackprozesses

III – Datenerhebung und Modellierung

- Erhebung der Geschäfts- und IT-Architektur mithilfe von Interviews und Dokumentenanalyse (unternehmensweit)
- Modellierung der erfassten Elemente
- Erfassung ergänzender Informationen
- Aufbau von Beziehungen zwischen den Elementen der EA

IV – Publikation der EA

- Veröffentlichung der Unternehmensarchitektur (komplett oder selektierte Teile) für die Nutzer

V – Veränderungsmanagement

- Begleitung der Veränderungsprozesse auf sozialer und kultureller Ebene (weiche Faktoren)
- Schaffung einer EAM-Akzeptanz und Umgang mit Widerständen
- Etablierung technischer und organisatorischer Veränderungsprozesse (Change Management)

VI – Erfolgskontrolle

- Messung des neuen IST-Zustandes und Abgleich mit den zuvor erhobenen Werten (in Phase 1)

Exkurs: Unternehmensmodellierung

Veränderungen im Unternehmen erfordern ein grundlegendes *Verständnis* der Strukturen und Abläufe in einem Unternehmen. Dieses Verständnis kann insbesondere durch die Schaffung von Unternehmensmodellen – also der *Visualisierung von Prozessen und Strukturen* – herbeigeführt werden. Die Erhebung und Visualisierung der Ausgangslage (Ist-Situation) ist dabei oftmals der Grundstein für Veränderungen und Basis für die Ermittlung von Veränderungspotentialen.

Die *Werkzeugunterstützung* stellt einen wichtigen Erfolgsfaktor für die Unternehmensmodellierung dar (Sandkuhl, Wißotzki, and Stirna 2013, S. 69). Die Auswahl des richtigen Werkzeugs ermöglicht erst die langfristige Akzeptanz und damit einhergehend auch die Nutzung der Werkzeuge und Modelle. Nach Sandkuhl et al. (Sandkuhl, Wißotzki, and Stirna 2013, S. 69) hängt die Auswahl von 1. Absichten des Unternehmens (z. B. Weiterverwendung der Modelle?), 2. Situation (z. B. Personal, Ressourcen), sowie 3. Anforderungen an das Werkzeug ab. Dabei lassen sich Werkzeuge in reine Zeichenwerkzeuge und Modellierungsumgebungen unterteilen. Zeichenwerkzeuge bieten dem Nutzer zwar eine größere Flexibilität im Rahmen der Modellerstellung und oftmals günstige Anschaffungskosten, zeigen aber in Bezug auf andere wichtige Aspekte starke Schwächen: Bereitstellung von Metamodellen, Versionierung, Archivierung, Analysefunktionen, Validierung, Erstellung von Teilmodellen, Sichten und Schnittstellen.

Unter Einbindung der Fachexperten und Entscheider im Unternehmen bieten sich verschiedene qualitative und quantitative *Analysetechniken* zur Erhebung der aktuellen Situation an. Gemeinsamer Nenner aller Ansätze ist, den Ist- bzw. Soll-Zustand sowie den Kontext zu erfassen:

Analyse von Dokumenten und Systemen: Bereits existierende Informationen werden aus Dokumenten extrahiert oder aus Systemen (automatisiert) ausgelesen.

Befragung: Informationserhebung durch die Befragung von Personen(-gruppen) (z. B. per Fragebogen oder persönlich), sodass die Antworten dokumentiert und ausgewertet werden können.

Teilnehmende Beobachtung: Systematische Erfassung und Dokumentation von Verhaltensweisen und Abläufen im gewohnten Kontext (z. B. am Arbeitsplatz).

Moderierter Workshop: Kooperative und moderierte Fokussierung einer kleinen Personengruppe, um ein gemeinsames Thema zu bearbeiten.

Selbstaufschrieb: Die befragten Personen ermitteln Informationen im Rahmen ihrer regulären Tätigkeit und stellen diese in vorbereiteten Formularen zur Verfügung.

Partizipative oder grafische Modellierung: Im Rahmen eines Workshops mit Fachkräften werden Fragen gemeinsam diskutiert und die Antworten darauf „live“ grafisch visualisiert (vgl. Jahnke, Herrmann, and Prilla 2008).

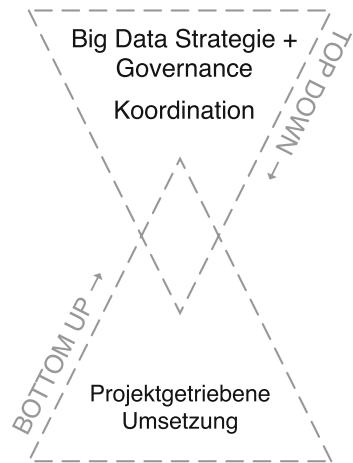
Der Modellierer benötigt neben seinen Modellierungsfertigkeiten auch Kenntnisse über die zu modellierende Domäne und ist daher – abhängig von Umfang und Komplexität – auf die Expertise aus der jeweiligen Fachabteilung angewiesen. Entsprechend eignen sich insbesondere Ansätze, wie die partizipative Modellierung, die verschiedene Perspektiven in die Modellierung einfließen lassen und die gewonnenen Einblicke auch kritisch reflektieren können. Zu beachten ist hierbei, dass unterschiedliche Erfahrungs- und Wissensstände, wie auch fehlende Empathiefähigkeit zu Konflikten beim Modellieren führen können (vgl. Wolff 2008, S. 128). Diese Konflikte gilt es frühzeitig zu adressieren. Dabei stellt die Schaffung einer offenen und innovationsfreudlichen Unternehmenskultur einen wesentlichen Schritt auf dem Weg zur stärkeren Partizipation dar (mehr dazu in Kapitel „Datenorientierung statt Bauchentscheidung: Führungs- und Organisationskultur in der datenorientierten Unternehmung“).

2.2.2.2 Einführung von Big Data unter besonderer Beachtung der Unternehmensarchitektur

Think big. Start small. Act now. Die motivierenden Worte von Barnabas Suebu können wunderbar auf die Einführung von Big Data im Unternehmen übertragen werden:

Think big. Es ist wichtig sich zuerst Gedanken über die eigene Big Data-Vision zu machen. Big Data kann nicht als ein isoliertes Konzept im Unternehmen gesehen werden, sondern muss in alle Bereiche des Unternehmens integriert werden. Diese Integration benötigt eine Vision, eine Strategie und definierte, übergeordnete Rahmenbedingungen. Dabei ist es sinnvoll, dass die Big Data-Strategie sich nahtlos in die IT-Strategie einfügt und dabei der Unternehmensstrategie folgt (vgl. Conrads 2013, S. 127 f.).

Abb. 2.16 Einführungsstrategie



Start small. Wie bereits zuvor beschrieben, stellen Unternehmen komplexe Gebilde dar und verhindern damit quasi den „großen Wurf“ bei der Einführung von Big Data. Konsequenter Weise sollte iterativ mit kleinen Projekten gestartet werden und so Big Data peu à peu im gesamten Unternehmen ausgerollt werden. Quick Wins statt Big Bang und dabei die übergeordnete Vision bei keinem Teilprojekt aus dem Auge verlieren.

Act now. Auf die Frage, wann man mit Big Data beginnen sollte, liefert der Autor Pavlon Baron (Baron 2013) die knackige Antwort: „Gestern“. Für Unternehmen ist es wichtig, früher als der Wettbewerb den Nutzen aus den vorhandenen Daten zu ziehen und sich so Wettbewerbsvorteile zu verschaffen.

„Think big. Start small.“ impliziert eine sogenannte hybride Einführungsstrategie (vgl. Abb. 2.16) und kombiniert dabei die „zentrale Führung“ wie sie bei einer Top-Down-Strategie zum Einsatz kommt, mit dem dezentralen, Projekt-getriebenen Ansatz der Bottom-Up-Strategie. Die Gesamtsicht der Big Data-Einführung wird durch eine zentrale Einheit (z. B. durch ein Big Data Competence Center) gewahrt. Diese Einheit liefert Architekturvorgaben, Richtlinien und koordiniert die einzelnen Teilprojekte (Stichwort: Redundanzvermeidung) der Big Data-Einführung. Diese zentrale Einheit sollte jedoch nicht für die Realisierung verantwortlich sein, denn diese wird im Rahmen von notwendigen Projekten dezentral durchgeführt. Somit kann mit geringem zusätzlichem Budget die Gesamtsicht erhalten bleiben und die Realisierung evolutionär als Teil von notwendigen Projekten erfolgen.

Bei der Implementierung werden sowohl in der Aufbau- als auch in der Ablauforganisation – teilweise gravierende – Umstrukturierungen vorgenommen. Der Grad des organisatorischen Eingriffs hängt davon ab, ob bestehende Systeme und Strukturen vollständig ersetzt oder lediglich angepasst werden müssen. Interdependenzen zwischen den Systemen müssen identifiziert und entsprechend gehandhabt werden. Cramer und Dietze stellen dabei sechs Fragen, die man sich im Rahmen einer Big Data-Strategieentwicklung stellen sollte (Cramer and Dietze 2012):

- Welche Herausforderungen soll die Datennutzung lösen?
- Warum sollen diese Herausforderungen gelöst werden? Wie sieht der Business-Case aus?
- Welche Daten benötigt das Unternehmen dafür?
- Welche Daten liegen heute in welchen Systemen vor? Ist der Detailgrad ausreichend?
- Welche der erforderlichen Daten werden heute noch nicht systematisch erfasst?
- Können die fehlenden Daten als Nebenprodukt bestehender Prozesse erzeugt werden oder sind neue Erfassungswege dafür erforderlich?

Die Ist-Modelle der Unternehmensarchitektur bilden jedoch nicht ausschließlich die Basis zur Beantwortung dieser Fragen. Vielmehr sind sie der Ausgangspunkt für die Entwicklung der zukünftigen Unternehmensarchitektur, also der sogenannten Soll-Architektur. So können Prozesse an die neuen Daten-getriebenen Anforderungen angepasst und vor der Umstellung validiert werden, Anwendungen können für die Datenhaltung definiert (Welches ist das führende System?) und Infrastrukturelemente und deren Einbindung können visualisiert werden. Am vorherigen Beispiel der Stadtwerke Frankenbronn stellt sich eine solche Überführung von Ist nach Soll wie folgt dar:

Die Informationsarchitektur (vgl. Abb. 2.17) berücksichtigt neben der zusätzlichen Big Data-Lösung nun auch die weiteren Datentöpfe und Veränderungen an vorhandenen Systemen. Die Veränderungen auf der Hardwareebene in Form eines zusätzlichen Servers und einer ergänzenden Datenbank werden in der Technologiearchitektur (vgl. Abb. 2.18) visualisiert. Auf Basis dieser Modelle können konkrete Aktivitäten abgeleitet werden, um die IT-Landschaft vom heutigen Ist-Zustand in den finalen Zielzustand zu überführen.

Werden die Ziele der Big Data-Einführung an den Zielen des Unternehmens ausgerichtet, so ist auch sicherzustellen, dass die neu geschaffenen Möglichkeiten durch Big Data nun auch in den Geschäftsprozessen eingesetzt werden können (vgl. Conrads 2013, S. 129). Eine entsprechende Anpassung der Geschäftsprozesse und damit die Aktuali-

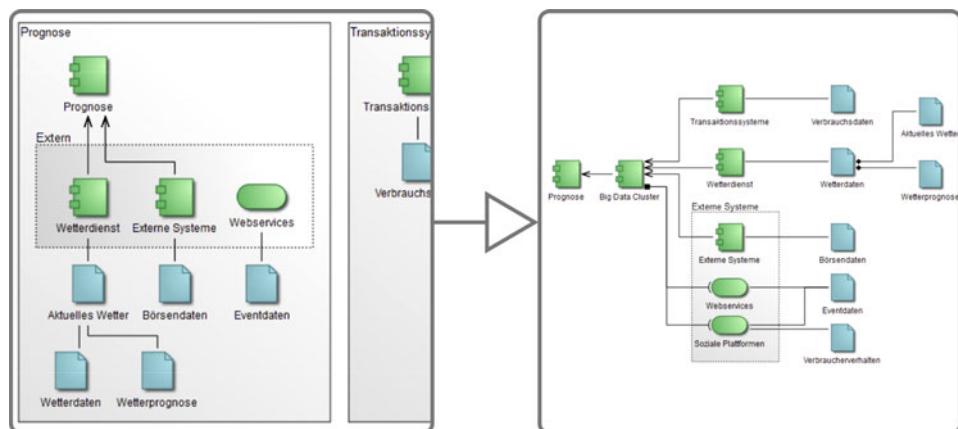


Abb. 2.17 Beispielhafte Anpassung der Informationsarchitektur

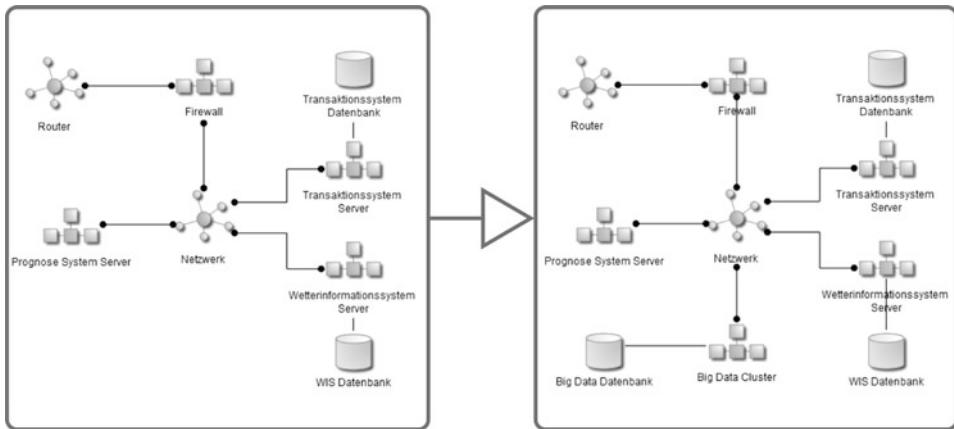


Abb. 2.18 Beispielhafte Anpassung der Technologiearchitektur

sierung der Unternehmensarchitektur ist somit, wie bereits eingangs erwähnt (vgl. Abb. 2.11), die sinnvolle Konsequenz. In diesem Kapitel wurde die Big Data-Einführung aus einer EAM-Perspektive erläutert. In den nachfolgenden Kapiteln widmen sich weitere Autoren den technischen und organisatorischen Aspekten der Einführung von Big Data im Unternehmen.

2.2.3 Fazit

Angesichts der Möglichkeiten die Big Data liefert, sollte sich kein Unternehmen dem entziehen. Eine Einführung sollte aber bedacht vorgenommen werden, indem zunächst die Transparenz von der Strategie ausgehend über die Fähigkeiten bis hin zu den Prozessen und deren Informationsflüsse geschaffen wird. Mittels EAM wird diese Transparenz durch die Aufnahme der Ist-Situation und der damit verbundenen einheitlichen Sprache bzw. Dokumentation geschaffen. Durch die Anwendung des im zweiten Teil des Artikels vorgestellten partizipativen Einführungsmodells wird schon aufgrund der Auseinandersetzung der Mitarbeiter mit den Fähigkeiten, Prozessen und Systemen ein Änderungsprozess gestartet, der hilft das Unternehmen in eine datenversierte Einheit zu verwandeln. Dieser Veränderungsprozess wird weiter verstärkt werden, wenn ein Transfer von der aufgenommenen Ist-Architektur zur Soll-Architektur, abgeleitet aus der Unternehmensstrategie, vorangetrieben wird. Wie in dem vorhergehenden Artikel beschrieben, wird ein Großteil der Veränderung die Unternehmenskultur betreffen. Unternehmen sind nun mal soziotechnische Systeme, bei denen die Interaktion zwischen den Akteuren, in der Regel Menschen, im Mittelpunkt steht. Dies bedeutet aber auch, dass ein wesentlicher Fokus auf den datenbasierten Capabilities liegen muss, damit Big Data nicht nur als kurzzeitiger Hype sondern als langfristige Systemänderung zur Steigerung der Wettbewerbsfähigkeit angewandt werden kann.

2.3 Advanced Analytics mit Big Data

Carsten Lanquillon und Hauke Mallow

Das reine Sammeln und Speichern großer Datenmengen hat noch keinen wirtschaftlichen Wert. Ihre Auswertung zur Erzeugung von Erkenntnissen mithilfe geeigneter Analysemethoden, also Big Data Analytics, ist einer der wichtigsten Bausteine zur Schaffung eines wirtschaftlichen Nutzens. Aber entscheidend ist letztlich, dass ausgehend von diesen Erkenntnissen auch Maßnahmen folgen, die ein Unternehmen im Rahmen ausgewählter Anwendungsfälle voranbringt (Franks 2012, S. 6).

2.3.1 Begriffsdefinitionen und Varianten

Es soll zunächst geklärt werden, was hinter dem oft verwendeten Begriff Advanced Analytics steckt. Da Big Data nicht die betrachteten Fragestellungen, sondern eher die Art der Datenverarbeitung während der Analyse stark verändert, werden gängige Analyseaufgaben und ein Prozessmodell für die Datenanalyse vorgestellt. Darauf aufbauend wird untersucht, was bei der Datenanalyse im Kontext von Big Data in besonderem Maße zu beachten ist.

2.3.1.1 Analyse und Analytics

Oft werden die Begriffe Analyse und Analytics unsauber verwendet oder verwechselt. Daher soll zunächst kurz der Unterschied erläutert werden. In diesem Kontext werden die Begriffe Analyse und Analytics ausschließlich im Sinne von Datenanalyse und Data Analytics betrachtet und verwendet.

Eine *Analyse* ist eine systematische Untersuchung einer Sache. Durch Untergliederung oder Zerlegung des Untersuchungsgegenstands – bei der Datenanalyse also der Daten – in seine Bestandteile sollen z. B. Strukturen, Auffälligkeiten, Regelmäßigkeiten oder Zusammenhänge aufgedeckt werden. Demnach ist die Analyse ein Prozess, bei dem aus Daten Informationen (Erkenntnisse) gewonnen werden.

Dagegen bezeichnet *Analytics* (zu Deutsch die Analytik) die Lehre oder Kunst des Analysierens also der Durchführung von Datenanalysen. So wird der Begriff Analytics auch unmittelbar für die Menge aller Analysemethoden verwendet. Er umfasst dann als Obermenge insbesondere Methoden aus den Bereichen Statistik und Data Mining bzw. dem maschinellen Lernen. Neben den Methoden werden oft auch die den Analyseprozess unterstützenden Technologien und Werkzeuge mit dem Begriff Analytics assoziiert.

Ähnlich wie mit dem Begriff *Statistik* nicht nur die Fachdisziplin, sondern auch das Ergebnis einer statistischen Untersuchung, also etwa eine berechnete Kennzahl oder allgemein eine Zusammenstellung von Daten in geeigneter Form, gemeint sein kann, werden mit dem Begriff Analytics gelegentlich auch die Analyseergebnisse selbst bezeichnet.

2.3.1.2 Analytics-Varianten

Das Wort Analytics wird oft mit weiteren Begriffen als Zusatz kombiniert, um einen speziellen Teilbereich oder eine Fokussierung bzw. genauere Charakterisierung zu kennzeichnen. So wird Analytics im hier relevanten Kontext unternehmerischer Fragestellung mit dem Fokus der Entscheidungsunterstützung auch als Business Analytics bezeichnet. Insbesondere wenn die Daten überwiegend unternehmensintern und strukturiert sind, lässt sich Analytics in Bereich Business Intelligence (BI) einordnen (vgl. dazu auch den Abschn. 4.1). Im Folgenden werden einige typische Unterscheidungen im Kontext von Business Analytics beschrieben. Die meisten Analytics-Varianten kennzeichnen besondere Eigenschaften, die sich untereinander kombinieren lassen.

Unterscheidung nach der adressierten Fragestellung

Die oft im Bereich Business Intelligence anzutreffende Unterscheidung nach der Art der adressierten Fragestellung ist sicherlich die am weitesten verbreitete und wichtigste. Die daraus resultierenden Analytics-Varianten sind in Abb. 2.19 zusammengefasst und sollen im Folgenden genauer beschrieben werden.

Descriptive Analytics Methoden zur beschreibenden Analyse haben das Ziel, allgemeine Beschreibungen oder eine Zusammenfassung eines Sachverhalts zu erzeugen. Die Frage „Was ist geschehen?“ steht dabei im Vordergrund und wird oft ergänzt durch weitere charakteristische Fragen, beispielsweise nach dem Wann und Wo. Viele Fragen dieser Art

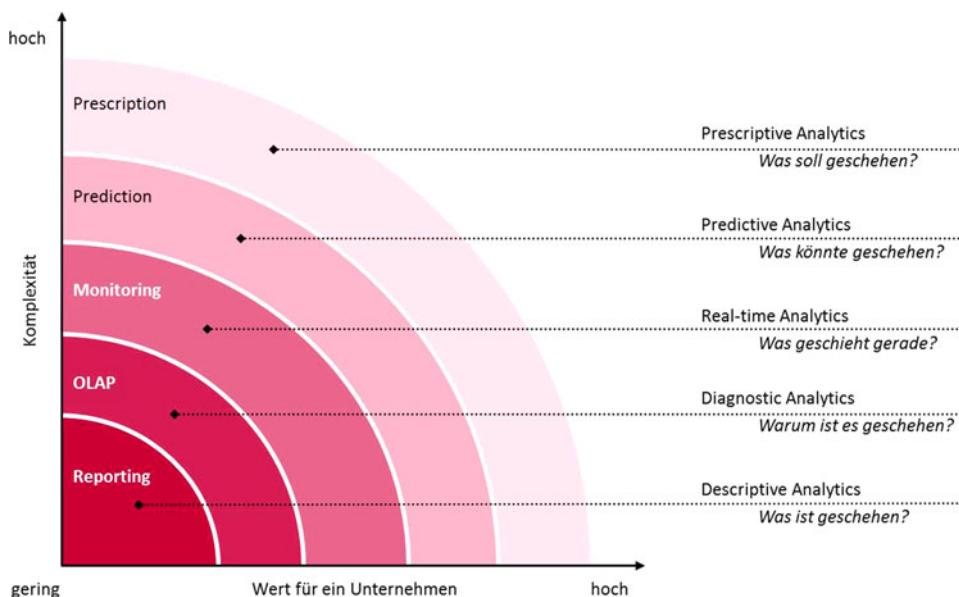


Abb. 2.19 BI-Analysespektrum: Fragestellungen im Kontext von Business Intelligence (BI) in Anlehnung an Eckerson (2007)

lassen sich mit methodisch einfachen Werkzeugen, wie dem Berichtswesen und OLAP, beantworten. Wie lässt sich aber z. B. die Beschreibung eines kreditwürdigen Kunden erstellen? Für beschreibende Fragestellungen dieser Art bedarf es komplexerer Methoden (siehe auch die Diskussion zum Begriff *Advanced Analytics*). In Verbindung mit Echtzeit-Verarbeitung (siehe *Real-time Analytics*) rückt die Fragestellung „Was geschieht gerade?“ in den Vordergrund. Dies ist im Anwendungsbereich der Überwachung und Kontrolle (Monitoring) von besonderer Bedeutung.

Diagnostic Analytics Mithilfe dieser Methoden wird eine Ursachenanalyse betrieben, um die Frage nach dem Warum zu beantworten. Aus dem vorliegenden Datenmaterial, das oft für einen anderen Zweck gesammelt wurde, können zwar Korrelationen, aber kaum kausale Zusammenhänge abgeleitet werden. Dennoch können die entdeckten Zusammenhänge wichtige Erkenntnisse für Fachexperten bei der Suche nach Abhilfe bei Problemen sein. Der Begriff Diagnostic Analytics ist weniger geläufig, sodass diese Fragestellung oft auch im Bereich Descriptive Analytics angesiedelt wird.

Predictive Analytics Prädiktive (vorhersagende) Analysen sollen einen Blick in die Zukunft ermöglichen. Variationen der Frage „Was könnte geschehen?“ stehen im Vordergrund. Es ist allerdings zu beachten, dass die Vorhersage möglicher Zustände nicht nur auf die Zukunft beschränkt ist. Allgemein beziehen sich Vorhersagemodelle auf die Vorhersage unbekannter Werte. Natürlich sind alle zukünftigen Werte von Zielgrößen unbekannt. Aber auch ein vergangener oder gegenwärtiger Wert einer Zielgröße kann unbekannt sein, wie etwa die Vorhersage der Stimmung eines vorliegenden Textes bei der Sentimentanalyse.

Prescriptive Analytics Die präskriptive (vorschreibende) Analyse soll schließlich Antwort darauf geben, mit welchen Handlungen (Schritten) ein Geschäftsziel am besten erreicht werden kann. Es geht demnach um die Frage „Was soll geschehen?“. Aus Perspektive der Anwendung stellt diese Art der Analyse die höchste Form der Entscheidungsunterstützung in einem Unternehmen dar. Im operativen Bereich beantworten Empfehlungssysteme etwa die Frage nach dem nächsten besten Angebot. Wann kann aus der Vergangenheit eine Empfehlung für die Zukunft abgeleitet werden? Von ganz besonderer Bedeutung sind Handlungsempfehlungen in neuen, unerwarteten Situationen. Allerdings stößt man hier mit datengetriebenen Methoden, für die nur Daten aus der Vergangenheit bis zur Gegenwart vorliegen können, schnell an Grenzen der Analysemöglichkeiten. Und es wird stets seltene Ereignisse geben, die nicht vorhergesagt werden können und gravierende Auswirkungen haben. Methoden der deskriptiven und prädiktiven Analyse kommen als Grundbestandteile der präskriptiven Analyse oft in Verbindung mit Optimierungsmethoden oder modellbasierten Simulationstechniken zum Einsatz. Insbesondere auch die Rückkopplung bezüglich des Erfolgs der vorgeschlagenen und umgesetzten Empfehlungen ist ein weiterer wichtiger Aspekt zur iterativen Verbesserung in diesem Bereich (Apte 2010).

Unterstützende Analysemethoden

Die oben vorgestellten Fragestellungen decken analytische Aufgaben sehr gut ab. Die folgenden beiden Analytics-Varianten sollten keine unabhängigen Formen der Analyse darstellen, sondern der Unterstützung anderer Aufgabenstellung dienen.

Exploratory Analytics Zum Aufbau eines Datenverständnisses und zur Erzeugung erster Ideen (Hypothesen) über sinnvolle Zusammenhänge, werden gerne Methoden der explorativen Datenanalyse eingesetzt. Explorative Methoden werden daher oft in einem Zwischenschritt in der Data-Understanding-Phase (siehe Abschn. 2.3.3.2) eingesetzt und können beispielsweise zur Erzeugung von möglichen Hypothesen genutzt werden, die dann im Rahmen der übergeordneten Fragestellung, etwa bei der späteren Modellbildung, verwendet werden können.

Visual Analytics Ein Bild sagt mehr als tausend Worte. Und bei einer guten Visualisierung springen dem Betrachter die Zusammenhänge förmlich ins Auge, denn die menschliche Fähigkeit, in Tabellen oder Zahlenkolonnen Zusammenhänge auszumachen, ist begrenzt. Dagegen unterstützen durchdachte visuelle Darstellungen die Datenexploration, indem sie dem Menschen einen leichteren und oft interaktiven Zugang zu Erkenntnissen ermöglichen. Visualisierungsmethoden können den Analyseprozess auch an anderen Stellen unterstützen. Gerade auch bei der Ergebnispräsentation sind sie ein elementarer und unverzichtbarer Bestandteil.

Unterscheidung nach Art der Daten

Die Unterscheidung zwischen strukturierten, wenig oder gar nicht strukturierten Daten wurde im Rahmen der Variety-Eigenschaft von Big Data bereits diskutiert. Aber auch innerhalb dieser Gruppen gibt es noch große Unterschiede. So zählen etwa Text, Bild, Ton und Video oder allgemein Multimedia-Daten allesamt zur den mengenmäßig dominierenden unstrukturierten Daten.

Die Art der Daten ist für die Datenanalyse von zentraler Bedeutung. Die meisten Standardmethoden der Datenanalyse erwarten strukturierte Daten in Form einer Matrix, in der die betrachteten Untersuchungsobjekte und die zu deren Charakterisierung herangezogenen Merkmale die Dimensionen bilden. Wie oben erkannt, liegen die wenigsten jedoch in dieser Form vor.

Es gibt zwei grundlegende Möglichkeiten, um mit Daten umzugehen, die von der strukturierten Matrixform abweichen. Einerseits können Daten durch eine geeignete Datenvorverarbeitung in eine Matrix transformiert werden, um anschließend Standardmethoden anwenden zu können. Andererseits können spezielle Methoden verwendet werden, die direkt auf anderen Datenformen arbeiten können. Darüber hinaus können spezielle Analysemethoden auch auf besondere Fragestellungen eingehen, die sich aus der Art oder Herkunft der Daten ergeben.

Ein Beispiel für diesen Fall ist *Text Analytics*, das auch als Text Mining bekannt ist. Ziel dabei ist die Gewinnung relevanter Erkenntnisse aus Textdokumenten, wie beispielswei-

se die Identifikation wichtiger Themen oder Konzepte und deren zeitliche Entwicklung (Trends) sowie die Erkennung der Stimmung in Texten (Sentimentanalyse). Letztlich sollen die unstrukturierten Daten auch für eine weitere Verarbeitung in strukturierter Form erschlossen werden, insbesondere auch für eine Integration mit anderen strukturierten internen Daten eines Unternehmens. Texte werden dazu üblicherweise durch geeignete Vorverarbeitungsschritte in sogenannte Dokumentvektoren umgewandelt, deren Elemente jeweils für bestimmte Terme (z. B. Wörter oder Wortgruppen) stehen. Mehrere Dokumentvektoren zusammengefasst bilden die gewünschte Matrix, auf der dann Analysemethoden angewendet werden können. Trotz der erreichten Matrixform gilt es jedoch, bestimmte Eigenschaften zu berücksichtigen, die sich aufgrund der ursprünglichen Art der Daten ergeben, wie etwa die hohe Dimensionalität der Dokumentvektoren oder die Tatsache, dass diese oft nur spärlich besetzt sind. Methoden und Systeme zur semantischen Analyse von Texten im Kontext von Big Data werden im Abschn. 4.4 ausführlicher beschrieben.

Vergleichbar lässt sich dies auf Audio, Bilder oder Video bzw. allgemein Multi-Media-Daten ausführen. Typischerweise kennzeichnet auch die Herkunft der Daten ihre Art und auch gängige Fragestellungen. Dies resultiert beispielsweise in Begriffen wie *Social Media Analytics* oder *Web Analytics*.

Location Analytics Wenn die betrachteten Daten, gleich welcher Art, zusätzlich einen Orts- oder Raumbezug haben, dann lässt sich dieser mit geeigneten Methoden berücksichtigen. Insbesondere visuelle Schnittstellen (Visual Analytics) sind dabei zur Unterstützung einer Analyse sehr wertvoll. Außerdem besteht die Möglichkeit, die betrachteten Daten mit orts- oder raumbezogenen Informationen anzureichern.

Hervorhebung zeitlicher Aspekte

Bis ausgehend von einem Ereignis oder einer Transaktion eine Maßnahme tatsächlich durchgeführt wird, vergeht eine gewisse Zeit. Die Abb. 2.20 zeigt, wie sich diese sogenannte Aktionszeit oder auch Latenz in die vier Abschnitte Daten-, Analyse-, Entscheidungs- und Umsetzungslatzen unterteilen lässt. Je nach Anwendung werden unterschiedlich lange Verzögerungen in den verschiedenen Phasen geduldet.

Real-time Analytics In einigen Anwendungen, wie etwa der Platzierung personalisierter Werbung auf Webseiten, bei persönlichen Empfehlungen im Online-Handel, bei der Erkennung betrügerischer Transaktionen oder bei Anwendungen im Wertpapierhandel, ist die verfügbare Zeit zwischen dem auslösenden Ereignis und der erforderlichen Ausführung einer Maßnahme sehr gering. Geht die tatsächlich benötigte Gesamtlatenz gegen Null, spricht man von einer Echtzeit-Anwendung. Methoden für diese Szenarien werden mit dem Begriff *Real-time-Analytics* zusammengefasst (Anforderungen von Echtzeit-Anwendung und Möglichkeiten der technischen Umsetzung werden in Abschn. 4.3 ausführlicher betrachtet). In vielen Fällen ist jedoch lediglich eine rechtzeitige Verfügbarkeit der Daten, der Analyseergebnisse oder der Aktionen gefordert. Nicht real-time, sondern right-time ist an dieser Stelle dann die zutreffendere Lösung.

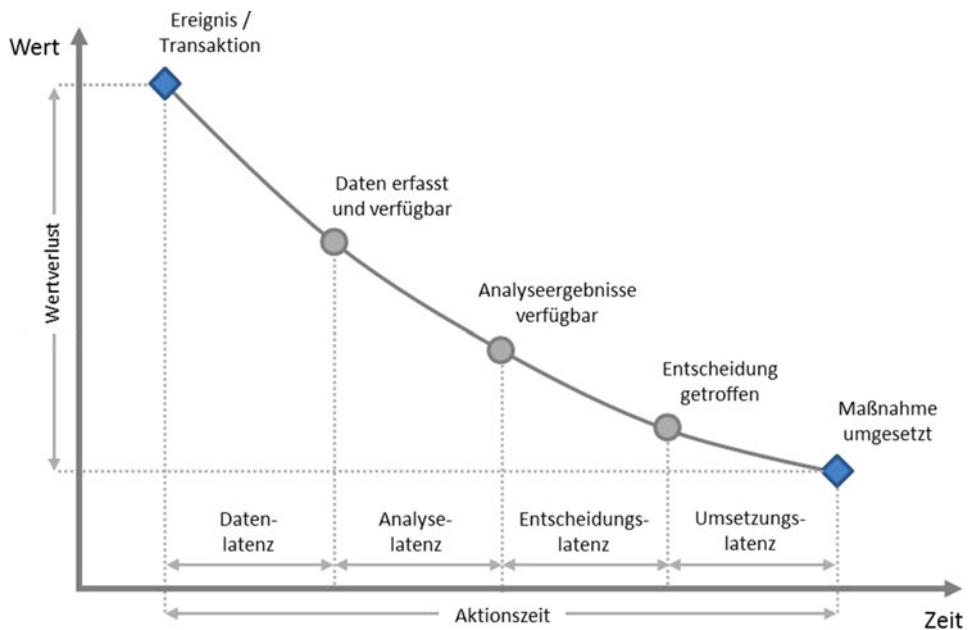


Abb. 2.20 Latenzzeiten vom Ereignis bis zur Maßnahme in Anlehnung an Kemper, Baars, and Mehanna (2010), S. 91

Obwohl bei Echtzeit-Anwendungen die Gesamtlatzen verschwindend gering sein sollte, liegt der Fokus oft lediglich auf der Minimierung der Daten- und Analyselatenz. Systeme werden so gestaltet, dass aktuelle Daten ohne nennenswerten Zeitverzug für Analysen bereitstehen und diese werden dann in extrem kurzer Zeit ausgeführt. Dabei ist jedoch anzumerken, dass dabei mit Analyse in der Regel nicht die Modellerstellung selbst, sondern nur die Anwendung eines bestehenden Modells auf neue Daten gemeint ist, wie etwa der Einsatz bestehender Regeln zur Erkennung von Betrugsfällen. Es ist nicht selten, dass viel Aufwand betrieben wird, um Entscheidungsträgern Analyseergebnisse in Echtzeit zur Verfügung zu stellen, während diese sich dann für die Entscheidungsfindung und Umsetzung von Maßnahmen viel Zeit lassen. Der Mensch in der Entscheidungskette ist oft der kritische Faktor.

Embedded Analytics Um Entscheidungen in Echtzeit zu treffen und umzusetzen, bedarf es einer stärkeren Automatisierung der Entscheidungsprozesse. Dazu muss wiederum eine Einbettung in die bestehenden Geschäftsprozesse erfolgen. In diesem Zusammenhang wird von *Embedded* oder auch *Operational Analytics* gesprochen. Es ist zu bedenken, dass das Spektrum möglicher Entscheidungen äußerst vielfältig ist. Gerade im operativen Bereich lassen sich einige Entscheidungen automatisieren. Im strategischen Bereich ist eine Automatisierung nur schwer vorstellbar. Dennoch können auch dabei Analyseergebnisse besser in den Entscheidungsprozess eingebunden werden. Eine Berücksichtigung von Kollaborationsplattformen sei an dieser Stelle hervorgehoben.

Stream Analytics Eine andere Art der Unterscheidung hinsichtlich der Daten ist die Frage, ob diese vor der Analyse gespeichert werden (*Data-at-Rest*) oder direkt als Datenstrom ohne vorherige Speicherung (*Data-in-Motion*) verarbeitet werden müssen. In klassischen Analyseszenarien liegen die Daten in gespeicherter Form vor. Deshalb ist es möglich und durchaus üblich, dass Analysemethoden in mehreren Iterationen auf die Daten zugreifen. Man spricht dann von *Batch Analytics*. Mit zunehmendem Datenvolumen und zunehmender Geschwindigkeit, mit der die Daten erzeugt werden und verarbeitet werden sollen, bekommt die Analyse von Datenströmen eine immer größere Bedeutung. Zur Kennzeichnung von Analysemethoden für Datenströme wird der Begriff *Stream Analytics* verwendet. Auch der Begriff *Complex Event Processing (CEP)* ist für diesen Aufgabenbereich üblich. Es wird angenommen, dass bei Datenströmen höchstens eine ausgewählte Teilmenge persistiert werden kann. Dies begründet eine starke Einschränkung für Analysemethoden, denn die Daten im Datenstrom können jeweils nur einmal, nämlich sofort, betrachtet werden (Rajaraman, Leskovec, and Ullman 2010, S. 129). Außerdem müssen die Daten aufgrund der Geschwindigkeit, in der sie erzeugt werden, in der Regel in Echtzeit verarbeitet werden. Schließlich ist Echtzeitverarbeitung ein Muss, wenn eine Analyse bereits Ergebnisse liefern soll, während ein Ereignis stattfindet, um dieses zum Vorteil eines Unternehmens zu beeinflussen. Technische Aspekte einer CEP-Lösung werden beim Thema IT-Infrastrukturen für Big Data in Abschn. 4.3.5 detaillierter dargestellt.

Hervorhebung des Analyseorts

In-Database Analytics Das klassische Vorgehen bei der Datenanalyse sieht vor, dass relevante Daten aus ihren Quellsystemen extrahiert und auf dedizierten Analyseservern oder Arbeitsplatzrechnern gespeichert werden, sodass die Berechnungen im Rahmen des Analyseprozesses dort durchgeführt werden können. In Anbetracht stetig wachsender Datenvolumina wird dieses Vorgehen immer weniger praktikabel. Geschickter ist es dann, nicht die Daten dorthin zu bringen, wo die Algorithmen sind, sondern die Algorithmen zu den Daten. Dieses Vorgehen wird schon seit längerer Zeit durch spezielle Datenbankmanagementsysteme unterstützt. Man spricht in diesem Fall von *In-Database Analytics*. Ein großer Nachteil dieser Variante ist, dass oft eine begrenzte Auswahl an Standardalgorithmen zur Verfügung steht, für die eine effiziente Umsetzung innerhalb der Datenbank-Software existiert. Änderungen oder Erweiterungen an den Analysemethoden sind oft problematisch, wenn überhaupt möglich. Auch die Unterstützung des Analyseprozesses (siehe Abschn. 2.3.3) ist weit weniger komfortabel als in gängiger Analyse-Software. Moderne Datenbanken bieten inzwischen oft Schnittstellen zu Statistik-Programmen wie R. Wie auch an anderer Stelle bereits diskutiert, muss jedoch berücksichtigt werden, ob über die Schnittstelle lediglich erstellte Modelle angewendet werden können oder ob auch die Modellerstellung möglich ist.

In-Memory Analytics Der Trend zu In-Memory-Datenbanken mit ihrem klaren Geschwindigkeitsvorteil durch wahlfreie Zugriffe auf die Daten und Vermeidung zeitintensi-

ver IO-Operationen ist unverkennbar. Und insbesondere der Vorteil, der sich dadurch für Analysen mit ihren meist unvorhersehbaren und deshalb kaum planbaren Zugriffsmustern auf die Daten ergibt, ist enorm, selbst interaktive Analysen auf großen Datenmengen werden dadurch möglich. Der dafür oft verwendete Begriff *In-Memory Analytics* ist jedoch genau genommen irreführend, denn die Berechnungen für eine Analyse finden letztlich immer in einem Prozessor statt und nicht im Speicher, unabhängig davon, ob dafür Festplatten oder der Hauptspeicher zum Einsatz kommen (vgl. auch Abschn. 4.3.4).

Unterscheidung nach Anwendungen und Einsatzbereichen

Ebenso gibt es Spezialisierungen und Anwendungsbereiche nach konkreten Themen bzw. Einsatzbereichen, wie etwa *Customer Analytics*, *Risk Analytics* oder *Fraud Analytics*. Entscheidend ist, dass zum Methodenwissen der allgemeinen Datenanalyse jeweils fachliche Expertise hinzukommt. Diese berücksichtigt neben den anwendungsspezifischen Fragestellungen beispielsweise konkrete Annahmen und Erfahrungen bei der Entwicklung, der Anpassung und dem Einsatz von Methoden zur Datenaufbereitung und Datenanalyse, die dann in der Regel den Standardmethoden weit überlegen sind. Weiterhin implizieren Anwendungsbereiche oft bestimmte Eigenschaften oder Anforderungen. Beispielsweise erfolgt die Erkennung betrügerischer Transaktionen üblicherweise in Echtzeit.

Advanced Analytics: Wann ist eine Analyse fortschrittlich?

Wann sind die bei einer Analyse eingesetzten Methoden so fortschrittlich, dass wir von Advanced Analytics sprechen? Und wenn es fortschrittliche Analysemethoden gibt, welche Methoden sind dann so einfach, dass sie nur als Analytics oder als *Basic*, *Core* oder *Simple Analytics* gelten?

Häufig werden die als zukunftsorientiert charakterisierten Varianten Predictive und Prescriptive Analytics als Advanced Analytics zusammengefasst. Entsprechend gehörten dann die vergangenheitsorientierten Varianten Descriptive und Diagnostic Analytics zu den einfachen Varianten. Für einfache Fragestellungen mit einfachen Zusammenfassungen von Daten ist dies nachvollziehbar. Aber wer möchte eine tiefere Analyse, die beispielsweise automatisch verständliche Beschreibungen zur Erklärung von Sachverhalten wie dem Kündigungsverhalten von Kunden aus Daten ableiten soll, als einfach charakterisieren?

Geeigneter scheint daher eine Einteilung nach der Komplexität der eingesetzten Methoden und Durchführungsform der Analysen. Erkenntnisse, die sich mit wenigen SQL-Abfragen notfalls auch durch eine manuelle Ausführung zur Beantwortung explizit formulierter Fragen erzeugen lassen, gehen auf einfache Formen der Analyse zurück. Analysen dieser Art bezeichnet man auch als hypothesengetrieben. Aus BI-Perspektive umfasst *Basic Analytics* das klassische Berichtswesen, OLAP mit seinen dynamischen „Berichten“ und das kontinuierliche Überwachen von Kennzahlen (Monitoring).

Unabhängig von der adressierten Fragestellung lassen sich komplexere statistische Methoden und Methoden des Data Minings bzw. des maschinellen Lernens als *Advanced Analytics* zusammenfassen. Mithilfe dieser datengetriebenen Methoden sollen üblicher-

weise automatisch oder halb-automatisch Erkenntnisse zur Beantwortung implizierter Fragestellungen gewonnen werden. Im Fokus stehen Erkenntnisse oder Zusammenhänge, die nicht durch ein paar einfache SQL-Abfragen erschlossen werden können.

2.3.1.3 Analytics trifft auf Big Data

Big Data Analytics ist der Einsatz von Analysemethoden im Kontext von Big Data. Vordergründig bieten zusätzliche Datenquellen, detailliertere Daten sowie deren schnellere Verfügbarkeit durch geeignete Analysen ein größeres Potenzial bei der Beantwortung fachlicher Fragestellungen im Unternehmen. Anwendungsszenarien für Big Data bzw. Big Data Analytics, wie etwa ein umfassenderes bzw. ergänztes Kundenprofil und Produktfeedback anhand von Daten aus sozialen Netzwerken oder detaillierte Einblicke in Geschäftsprozesse beispielsweise bei der Produktion oder generell bei der Nutzung von Infrastruktur durch vielfältige Sensor-Messwerte und Log-Dateien, werden in den folgenden Kapiteln beschrieben.

Betrachtet man die fachlichen Zielsetzungen, so zeigt sich bei den meisten Anwendungsszenarien, dass es sich ob mit oder ohne Big Data um vergleichbare Fragestellungen handelt, für deren Verständnis und Lösung sich zunächst bewährte Analysemethoden anbieten. Im Folgenden sollen daher klassische Analyseaufgaben und ein allgemeines Prozessmodell für die Datenanalyse vorgestellt werden. Im Anschluss wird darauf aufbauend betrachtet, an welchen Stellen des Analyseprozesses und in welcher Weise Big Data besondere Überlegungen erfordert und Änderungen mit sich bringt.

2.3.2 Analyseaufgaben

Eine wichtige Aufgabe bei der Datenanalyse für ein Anwendungsproblem ist die Abbildung der fachlichen Fragestellung auf wohl definierte, kanonische Analyseaufgaben. Dies ist eine zentrale Aufgabe in der ersten Phase des unten vorgestellten Analyseprozesses. Für die bekannten Analyseaufgaben gibt es jeweils eine ganze Reihe bekannter Standardverfahren. Es ist meist ratsam, zunächst Standardverfahren heranzuziehen, weil ihre Funktionsweisen und Einsatzbedingungen in der Regel besser bekannt sind. Sollten die Standardverfahren zur Lösung eines Fachproblems nicht genügen, mögen speziellere oder gar maßgeschneiderte Verfahren zum Einsatz kommen. Im Folgenden werden die gebräuchlichsten Analyseaufgaben kurz vorgestellt. Der Fokus liegt hier auf datengetriebenen Verfahren. Hypothesengetriebene Verfahren, wie sie im Bereich der klassischen induktiven Statistik üblich sind, werden an dieser Stelle nicht berücksichtigt, auch wenn insbesondere das Testen von Hypothesen zur Bewertung der statistischen Signifikanz von Analyseergebnissen eine wichtige Rolle im Analyseprozess spielt.

Bei den Analyseaufgaben wird üblicherweise zwischen beschreibenden und vorher sagenden (prädiktiven) Aufgaben unterschieden. Bezogen auf die Analyseergebnisse ist diese Trennung nicht so eindeutig, wie die Einteilung der Aufgaben vermuten lässt. Ist ein Vorhersagemodell vom Menschen lesbar und interpretierbar, dann mag es als ver-

ständliche Beschreibung von Zusammenhängen in den Daten genutzt werden. Oft wird sogar verlangt, dass die Vorhersagemodelle interpretierbar sein sollen, und die Analysemethoden werden dementsprechend ausgewählt. Ebenso ist es möglich, dass Modelle, die Zusammenhänge (oft auch als Muster bezeichnet) in den Daten beschreiben, letztlich in einem folgenden Schritt auch zur Vorhersage genutzt werden. Weiterhin kommt es sehr häufig vor, dass die verschiedenen Analyseaufgaben kombiniert werden, um eine fachliche Aufgabenstellung zu lösen. Beispielsweise wenn bei einer Marktsegmentierung verständliche Beschreibungen für die Gruppen ähnlicher Objekte gesucht werden, die das Ergebnis einer Clusteranalyse sind.

2.3.2.1 Prädiktive Analyseaufgaben

Zu den klassischen prädiktiven Analyseaufgaben zählen die *Klassifikation* und die *numerische Vorhersage*, die auch als *Regression* bezeichnet wird. Außerdem sollen an dieser Stelle das *Ranking* von Objekten und die *Zeitreihenanalyse*, die beide oft als Spezialfälle der numerischen Vorhersage betrachtet werden, aufgrund der zunehmenden Bedeutung gerade in Anwendungen im Kontext von Big Data als eigenständige Analyseaufgaben ergänzt werden. Charakteristisch für alle diese Aufgaben ist die Existenz einer Zielgröße bzw. eines Zielattributs. Für ausgewählte (in der Anwendung neue) Objekte sollen die unbekannten Werte des Zielattributs vorhergesagt werden. Auch wenn das primäre Ziel natürlich die Vorhersage ist, so können je nach Verständlichkeit der Modelle diese auch zur Beschreibung der Zusammenhänge in den zu Grunde liegenden Daten verwendet werden und somit auch zu einem Erkenntnisgewinn beitragen.

Die Erstellung von Vorhersagemodellen wird auch *Predictive Modeling* und die dafür eingesetzten Methoden als *Predictive Analytics* bezeichnet. Sollen geeignete Vorhersagemodelle aus Daten erstellt werden, kommen üblicherweise sogenannte überwachte Lernverfahren zum Einsatz. Im Englischen nennt man diese Art der Modellerstellung *Supervised Learning*. Überwachung in diesem Zusammenhang bedeutet, dass für die Modellerstellung Lernbeispiele (Trainingsdaten) benötigt werden, für die die tatsächlichen (wahren) Werte des Zielattributs bekannt sind, sodass beim Lernvorgang für die Trainingsdaten bestimmt werden kann, ob die vorhergesagten Werte korrekt sind.

Klassifikation Das Ziel der Klassifikationsaufgabe ist die Erstellung eines Modells, des sogenannten Klassifikators, das Objekte anhand beobachtbarer Merkmale zu vorgegebenen Klassen oder Kategorien zuordnen kann. Dabei kann ein Objekt zu keiner, genau einer oder sogar mehreren Klassen gehören. Welche Variante sinnvoll ist, hängt vom Einsatzszenario ab. In der Theorie werden sehr häufig binäre Klassifikationsprobleme betrachtet, bei der ein Objekt genau einer von zwei Klassen zugeordnet wird. Auch das bekannte Erlernen eines Konzepts (Concept Learning), bei dem eine Boolesche Funktion angibt, ob ein Objekt zu einem Konzept gehört oder nicht, lässt sich als binäres Klassifikationsproblem interpretieren. Andere Klassifikationsprobleme mit mehr Klassen oder mehrfachen Zuordnungen zu Klassen lassen sich ohne Einschränkung aus binären Entscheidungen ableiten. Dazu werden einfach mehrere Klassifikatoren kombiniert. Beispielsweise kann die

Support Vector Machine (SVM), ein weit verbreitetes, überwachtes Lernverfahren, das für Klassifikationsprobleme geeignet ist, in der Grundversion nur binäre Entscheidungen treffen. Dagegen können die ebenso sehr verbreiteten Entscheidungsbäume prinzipiell mit einer beliebigen Anzahl an Klassen gleichzeitig umgehen.

Numerische Vorhersage Die numerische Vorhersage, auch als Regression oder Scoring bekannt, hat zum Ziel, ein numerisches Zielattribut anhand anderer beobachtbarer Merkmale vorherzusagen. Klassifikation und numerische Vorhersage bilden als Aufgaben den Kern der prädiktiven Modellierung bzw. von Predictive Analytics. Der Unterschied liegt im Skalenniveau des Zielattributs. Bei der Klassifikation ist das Zielattribut nominal und somit diskret, während das Zielattribut bei der numerischen Vorhersage metrisch ist und zumindest als Zwischenergebnis meist als stetig angenommen wird. Neben den klassischen Methoden zur Regressionsanalyse aus der Statistik kommen beispielsweise auch Regressionsbäume oder künstliche neuronale Netze häufig zum Einsatz.

Aufgrund der Ähnlichkeit von Klassifikation und numerischer Vorhersage können viele Analysemethoden eventuell mit geeigneter Anpassung oder Interpretation der Ergebnisse für beide Aufgabentypen verwendet werden. Beispielsweise werden im Rahmen einer Klassifikationsaufgabe oft Wahrscheinlichkeiten (oder allgemeiner Score-Werte) für die Zugehörigkeiten zu den gegebenen Klassen als numerische Zielgröße vorhergesagt. Basierend auf diesen Werten kann im Nachgang eine diskrete Klassifikationsentscheidung, zum Beispiel durch Mehrheitsentscheidungen oder durch einen Vergleich mit Schwellwerten, vorgenommen werden.

Ranking Die Aufgabe, Objekte in eine Reihenfolge nach einem vorgegebenen Kriterium zu bringen, korrespondiert mit einem Zielattribut, das ordinal skaliert ist. Da für diese Aufgabe oft Verfahren zur numerischen Vorhersage verwendet werden (Scoring), wird dieser Aufgabentyp in der Literatur oft gar nicht separat aufgeführt. Bei der Evaluierung der Ergebnisse gilt es wenigstens zu beachten, dass es nicht unmittelbar auf die Korrektheit der vorhergesagten Werte ankommt, sondern vielmehr auf die Beziehung zwischen jeweils zwei Werten, sodass sich möglichst die richtige Reihenfolge ergibt. Die Rankingaufgabe hat beispielsweise beim Information Retrieval und beim Collaborative Filtering eine große Bedeutung.

Zeitreihenanalyse Bei der Zeitreihenanalyse wird die Entwicklung einer Variablen über die Zeit untersucht. Zunächst kann dabei ein Verlauf in seine wesentlichen Bestandteile zerlegt und beschrieben werden. Häufig soll aber die zukünftige Entwicklung vorhergesagt werden. Da es dabei meistens um numerische Zielgrößen geht, handelt es sich letztlich um eine spezielle Variante der numerischen Vorhersage. Die besondere Form der Daten, über die Zeit aufgezeichnete Werte eines Attributes, gebietet jedoch auch die Anwendung spezieller Analysemethoden.

2.3.2.2 Beschreibende Analyseaufgaben

Alle weiteren Analyseaufgaben werden oft im weiteren Sinn als *beschreibend* gekennzeichnet. Zwar gibt es auch die konkrete Aufgabe der *Beschreibung* bzw. *Generalisierung*, aber auch die Zusammenfassung einer größeren Datenmenge durch die Angabe von Auffälligkeiten oder Strukturen kann als Beschreibung betrachtet werden. Beschreibende Verfahren lassen sich gut mit Methoden zur Vorhersage kombinieren. Entdeckte Strukturen können einerseits zum Beispiel verwendet werden, um daraus geeignete Ausprägungen eines Zielattributs für eine anschließende Klassifikationsaufgabe abzuleiten. Andererseits können bekannte, aber im Analyseprozess zunächst nicht verwendete Zielattribute bei der Auswahl oder Bewertung von entdeckten Zusammenhängen nützlich sein. Bei beschreibenden Analyseaufgaben können neben überwachten Methoden insbesondere auch unüberwachte Methoden zum Einsatz kommen. Diese zeichnen sich dadurch aus, dass sie ohne Verwendung eines konkreten Zielattributs arbeiten.

Konzeptbeschreibung Bei dieser Analyseaufgabe soll für eine Teilmenge bestimmter Objekte eine verständliche Beschreibung erzeugt werden. Die relevanten Objekte gehören in weiterem Sinn einem Konzept an, d. h. es kann sich auch um eine bestimmte Klasse oder ein Cluster von Objekten handeln.

Wenn auch Objekte bekannt sind, die dem Konzept nicht angehören, man spricht in diesem Fall von einer Kontrastmenge, dann können überwachte Lernmethoden zum Einsatz kommen, um charakteristische bzw. diskriminierende Merkmale zu identifizieren. Somit ist diese Aufgabe der Klassifikationsaufgabe sehr ähnlich. Und in der Tat werden sehr oft Klassifikatoren erstellt, die für Menschen leicht verständlich und interpretierbar sind, wie beispielsweise Entscheidungsbäume und Regelmengen. Der Unterschied gegenüber der Klassifikationsaufgabe besteht jedoch darin, dass Verständlichkeit wichtiger ist als die erwartete Prognosegüte und dies so weit wie möglich während des Lernprozesses berücksichtigt wird. Gibt es keine Kontrastgruppe, muss auch ohne das Prinzip der Diskriminierung zwischen Objekten, die dem Konzept zugehören bzw. nicht zugehören, eine Beschreibung durch Generalisierung über die gegebenen Objekte hinaus erzeugt werden.

Eine besondere Form der Konzeptbeschreibung ist die Erkennung und Beschreibung von Subgruppen mit dem Ziel, Gruppen von Objekten mit einer interessierenden (ausgezeichneten) Eigenschaft zu entdecken und zu beschreiben. Die relevante Eigenschaft, die die Objekte auszeichnet, lässt sich als Konzept verstehen. Allerdings wird nicht erwartet, dass für dieses Konzept eine allgemeine Beschreibung gefunden werden kann. Vielmehr wird angenommen, dass sich die relevanten Objekte auf verschiedene Gruppen mit unterschiedlichen Beschreibungen aufteilen, die es aufzudecken gilt. In diesem Sinne gibt es Gemeinsamkeiten mit der Clusteranalyse, allerdings kann bei der Gruppeneinteilung die Kenntnis des Konzepts ausgenutzt werden, es liegt in gewisser Weise eine Form der Überwachung vor. Ein typisches Beispiel für diese Aufgabe ist die Ursachenanalyse aus dem Bereich Diagnostic Analytics. Beschreibungen von Produkten mit bestimmten Qualitätsproblemen werden dabei als Hinweise auf mögliche Ursachen gedeutet. Unter der Annahme, dass die Qualitätsprobleme unterschiedliche Ursachen haben können, müssen verschiedene Gruppen mit unterschiedlichen Beschreibungen entdeckt werden.

Clusteranalyse Ziel der Clusteranalyse bzw. des Clusterings ist eine datengetriebene Segmentierung einer Menge von Objekten. Im Gegensatz zur Klassifikation, bei der die Klassen vorgegeben sind, stellen die entdeckten Cluster (auch Segmente, Gruppen oder Klassen genannt) bei der Clusteranalyse das Analyseergebnis dar. Daher bezeichnet man Verfahren zur Clusteranalyse auch als strukturentdeckend. Die formale Definition eines Clusters ist nicht einfach und es gibt vielfältige Varianten, die alle ihre Berechtigung für spezielle Anwendungsszenarien haben mögen. Typischerweise wird gefordert, dass Objekte in einem Cluster möglichst ähnlich zueinander sein sollen, während Objekte, die verschiedenen Clustern angehören, möglichst unähnlich sein sollen. Man spricht in diesem Zusammenhang von Homogenität innerhalb eines Clusters und Heterogenität zwischen verschiedenen Clustern. Dies macht deutlich, dass eine Clusteranalyse auf einer angemessenen Bestimmung von Ähnlichkeiten oder Abständen zwischen Objekten beruht, die im Rahmen einer Anwendung sinnvoll interpretiert werden kann. Da es bei der reinen Clusteranalyse kein vorherzusagendes Zielattribut gibt, ist die Evaluierung eines Clusterergebnisses weitaus schwieriger als die Bewertung bei der Klassifikation oder der numerischen Vorhersage. Es gibt zwar Gütekriterien für Cluster, aber diese mit den Zielen der Fragestellung in Einklang zu bringen, ist oft nicht trivial. Letztlich muss die entdeckte Struktur für die Anwendung nicht sinnvoll sein. Durch geeignete Attributauswahl, Ähnlichkeitsmaße und Art oder Form der Cluster, die ein Clusterverfahren erzeugen kann, muss versucht werden, relevante Strukturen zu finden. Zu beachten ist allerdings auch, dass die Clusteranalyse immer ein Ergebnis liefert, unabhängig davon, ob es überhaupt eine Struktur in den vorliegenden Daten gibt oder nicht.

Abhängigkeitsanalyse: Bei dieser Analyseaufgabe werden Abhängigkeiten, d. h. Beziehungen oder Zusammenhänge, zwischen den Attributen der betrachteten Objekte oder zwischen den Objekten selbst untersucht. Neben den klassischen Methoden der Statistik zur Zusammenhangsanalyse, wie die Korrelationsanalyse für metrische Attribute oder die Kontingenzanalyse für nominale Attribute, ist hier insbesondere die Assoziationsanalyse von Bedeutung. Bei der Betrachtung von Beziehungen zwischen Objekten spielt die Linkanalyse eine zunehmend größere Rolle. Diese untersucht Beziehungen zwischen Objekten, die als Knoten in einem Graphen aufgefasst werden, wie etwa die Verbindungen zwischen Webseiten oder Mitglieder in sozialen Netzwerken. Die Assoziationsanalyse ermittelt, welche Ereignisse in einer Menge von sogenannten Transaktionen, d. h. Mengen in irgendeiner Form zusammenhängender oder gebündelter Ereignisse, häufig gemeinsam auftreten. Diese Art der Analyse ist insbesondere durch den Einsatz bei der Warenkorbanalyse bekannt, bei der bestimmt wird, welche Produkte häufig zusammen gekauft werden. Soll die Reihenfolge der Ereignisse, d. h. die zeitliche Entwicklung, berücksichtigt werden, handelt es sich um eine Sequenzanalyse.

Die Berechnung der Stärke der Abhängigkeit zwischen Objekten oder Merkmalen ist im Einzelfall, basierend auf geeigneten Maßen, einfach. Die Schwierigkeit liegt vielmehr in der Bestimmung aller „interessanten“ Abhängigkeiten bereits bei mäßig großen Datenbeständen. Zum einen besteht die inhaltliche Herausforderung zu definieren, was

interessant ist, und zum anderen ist es aufgrund der Kombinationsvielfalt eine rechnerische Herausforderung, alle relevanten Kombinationen zu testen. Um die Komplexität der Berechnung in den Griff zu bekommen, werden üblicherweise Schwellwerte für die Häufigkeit definiert und im Rahmen einer intelligenten Suchstrategie ausgenutzt. Bekannte Verfahren in diesem Zusammenhang sind beispielsweise der Apriori-Algorithmus oder FP-Growth-Algorithmus.

Üblicherweise arbeitet die Abhängigkeitsanalyse ohne Zielattribut und daher unüberwacht und wird verwendet, um Einblicke in die Struktur der Daten zu gewinnen. Die Methoden lassen sich jedoch auch um die Verwendung eines Zielattributs erweitern, um etwa die Richtung der Abhängigkeiten vorzugeben oder die Menge der ermittelten Abhängigkeit zu filtern, in dem beispielsweise nur solche Abhängigkeiten betrachtet werden, die das Zielattribut betreffen. Diese lassen sich dann auch zur Vorhersage von Werten des Zielattributs verwenden.

Abweichungsanalyse: Ziel bei der Abweichungsanalyse ist die Identifikation von Objekten, bei denen Ausprägungen bei ein oder mehr Attributen beobachtet werden, die den meisten anderen nicht entsprechen, die also vom Normalfall oder von der Erwartung abweichen. Dabei kann es sich um Ausreißer handeln, die in einem gewissen Umfang bei hinreichend großen Datenbeständen normal sind, oder aber um Anzeichen für eine Veränderung im Umfeld der Anwendung. Eine typische Anwendung ist die kontinuierliche Beobachtung von Attributen (Kennzahlen) über die Zeit, das sogenannte Monitoring. Verfahren dieser Art werden intensiv im Rahmen der statistischen Qualitätskontrolle behandelt.

Formal ließe sich diese Analyseaufgabe auch als binäre Klassifikationsaufgabe mit dem Konzept „Abweichung liegt vor“ oder „Ist Ausreißer“ definieren. Sind dann auch noch bestimmte Fälle als Abweichungen oder Ausreißer bekannt, dann lässt sich diese Aufgabe mithilfe überwachter Lernverfahren lösen. Das Problem bei dieser Aufgabe stellt vielfach die extreme Ungleichverteilung der Klassen dar, da Abweichungen oder Ausreißer naturgemäß selten sind, denn sonst spiegeln sie eher den Normalfall wider. Klassische Lernalgorithmen für die Klassifikation haben jedoch oft Schwierigkeiten, wenn die Klassenverteilung sehr schief ist.

2.3.3 CRISP-DM: Ein Prozessmodell für Analyseprozesse

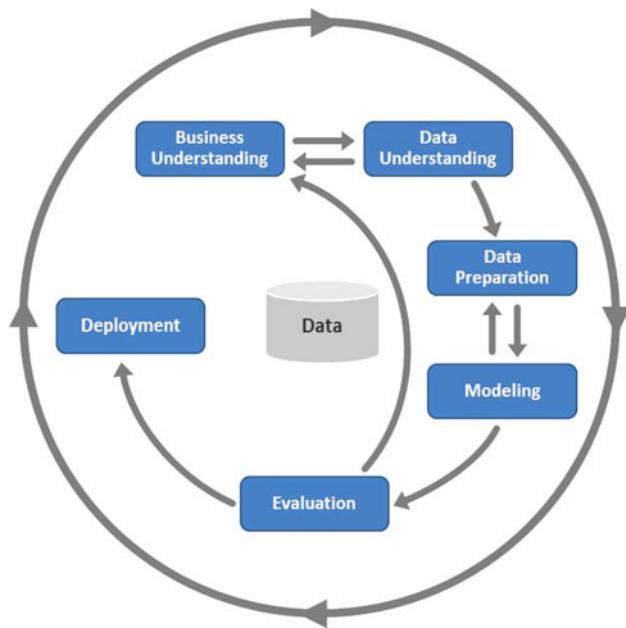
Die Entdeckung nützlicher Zusammenhänge oder Muster in Daten zur Lösung fachlicher Aufgabenstellungen sollte systematisch mithilfe eines Prozessmodells mit wohldefinierten Phasen erfolgen. Ein standardisiertes und strukturiertes Vorgehen erleichtert den Einstieg in die Durchführung intelligenter Datenanalysen und kann diese beschleunigen, hilft bei der Planung und Verwaltung der erforderlichen Aktivitäten und erhöht die Validität und Verlässlichkeit. Darüber hinaus fördert ein derartiges Verfahren die Wiederholbarkeit und Nachvollziehbarkeit aller tatsächlich durchgeführten Verarbeitungsschritte und Entschei-

dungen, erleichtert die Wiederverwendung von Erfahrungen und hilft letztlich durch ein gemeinsames Verständnis über den allgemeinen Ablauf einer Datenanalyse bei der Kommunikation zwischen den beteiligten Parteien.

Im Kontext von Big Data Analytics werden bereits erste „neue“ Analyse-Prozessmodelle beschrieben. Es sollte aber zunächst bedacht werden, dass es im Bereich KDD (Knowledge Discovery in Databases) bzw. Data Mining bereits bewährte Ansätze gibt. Die Prozessschritte im Kern einer Analyse sind bei den meisten Prozessmodellen naturgemäß sehr ähnlich. Ein wichtiges Unterscheidungsmerkmal ist daher die konkrete Einbettung in den Unternehmenskontext. Mit Blick auf den Einsatz im Kontext von Big Data Analytics verwenden wir hier die Bezeichnung *Analyseprozess* anstatt der sonst auch üblichen Bezeichnung *Data-Mining-Prozess*.

Ein sehr weit verbreitetes und ausgereiftes Prozessmodell ist der *Cross Industry Standard Process for Data Mining*, kurz CRISP-DM (Shearer 2000). Wir gehen davon aus, dass die angestrebte Datenanalyse in Form eines Projekts durchgeführt wird. CRISP-DM bricht ein Analyseprojekt hierarchisch auf vier Abstraktionsebenen herunter. Auf der obersten Ebene wird der Analyseprozess, wie in Abb. 2.21 dargestellt, durch sechs Phasen vollständig abgedeckt. Dabei wird insbesondere die Einbettung in den Unternehmenskontext ausgehend von einer fachlichen Zielsetzung bis hin zur Anwendung der Ergebnisse berücksichtigt. Die nächste Ebene beschreibt generisch, d. h. unabhängig von konkreten Problemstellungen und Methoden, die Aufgaben der einzelnen Phasen. Daher ist das Prozessmodell ohne Anpassung auch auf neue Herausforderungen wie etwa Big Data Analytics anwendbar. Erst in der dritten Ebene wird auf die mögliche Umsetzung

Abb. 2.21 CRISP-DM: Ein Standard-Prozessmodell zur Durchführung von Datenanalysen in Anlehnung an Chapman et al. (2000)



für verschiedene Situationen bei konkreten Analyseproblemen eingegangen. Die letzte Ebene, eine Instanziierung des Prozessmodells für eine konkrete Problemstellung, dient letztlich der Aufzeichnung der geplanten Aufgaben, der Entscheidungen, der tatsächlich ausgeführten Schritte oder Maßnahmen und der Ergebnisse. Insbesondere eine ausführliche Dokumentation ist essentiell, da eine Datenanalyse stets verlässlich und vor allem reproduzierbar sein sollte. Es sollte nicht nur dokumentiert werden, was am Ende erfolgreich war und umgesetzt wurde, sondern auch Wege, die nicht erfolgreich waren. Dies kann in späteren Iterationen sehr viel Arbeit und somit Zeit und Kosten sparen.

Die folgende Beschreibung konzentriert sich auf die fachlichen und analytischen Aspekte der Phasen. Für eine detaillierte Beschreibung sei der Leser an die ausführliche CRISP-DM-Dokumentation verwiesen (Chapman et al. 2000). Der Analyseprozess beginnt beim Business Understanding, folgt insgesamt aber keinem streng linearen Ablauf. Neue Erkenntnisse während der Analyse machen es oft erforderlich, dass Phasen mehrfach besucht werden. Ein Sprung von einer Phase zu einer beliebigen anderen ist stets möglich, wenn dies inhaltlich erforderlich ist. Die Pfeile im Schaubild sind lediglich als die in einem Analyseprojekt am häufigsten zu beobachtenden Sprünge zu interpretieren. Der äußere Kreislauf deutet an, dass der gesamte Analyseprozess oft mehrfach durchlaufen wird (z. B. auch bei Folgeprojekten) und dass zuvor gemachte Erfahrungen bei späteren Iterationen berücksichtigt werden sollten.

2.3.3.1 Business Understanding

Aus Unternehmenssicht ist die Durchführung einer Datenanalyse nur sinnvoll, wenn sie durch eine relevante Problemstellung begründet wird. Der Analyseprozess beginnt mit der Phase des *Business Understanding*. Hier sollen als erstes der fachliche Kontext und die Rahmenbedingungen des Anwendungsszenarios verstanden und die fachlichen Anforderungen und Ziele mit angemessenen Erfolgskriterien bestimmt werden. Warum ist das Projekt für das Unternehmen wichtig und wie sollen die Ergebnisse verwendet werden? Es sei daran erinnert, dass Ziele messbar und quantifizierbar sein sollten. Andernfalls ist eine spätere Evaluierung der Analyseergebnisse problematisch.

Ein ganz zentraler Schritt ist die Abbildung der fachlichen Ziele auf geeignete Analyseaufgaben (siehe Abschn. 2.3.2). Für die Lösung einer fachlichen Problemstellung kann auch eine Kombination mehrerer der kanonischen Analyseaufgaben erforderlich sein. Außerdem müssen entsprechend der fachlichen Erfolgskriterien Kennzahlen für die Evaluierung der Analyseergebnisse festgelegt und die Vorgehensweise geplant werden. Erfahrungen aus möglicherweise bereits bestehenden Problemlösungen im Unternehmen sind unbedingt zu berücksichtigen.

Die Erwartungshaltung der Auftraggeber sollte in Abhängigkeit von Randbedingungen und verfügbaren Ressourcen wie Daten, Rechenleistung und Fachexperten entsprechend angepasst werden. Hervorzuheben ist an dieser Stelle, dass insbesondere weiche Aspekte wie eine ungenügende Verfügbarkeit und Mitarbeit der beteiligten Fachbereiche sowie die gelebte Politik einzelner Beteiligter oder ganzer Bereiche kritische Projektrisiken darstellen.

2.3.3.2 Data Understanding

In der Data-Understanding-Phase erfolgt eine erste Sichtung der zur Verfügung stehenden Daten. Es werden hinsichtlich der Analyseziele relevante und geeignete Datenquellen identifiziert und erste Erkenntnisse über die Daten basierend auf einfachen Zusammenfassungen, Visualisierungen und auch explorativer Datenanalyse gewonnen. Es ist hervorzuheben, dass die explorative Analyse lediglich der ersten Erkundung der Daten und dem Vertrautwerden dient und nicht das Ziel selbst ist.

In dieser Phase sollen insbesondere Probleme mit den Daten ans Licht kommen. Mangelnde Datenqualität und oftmals unzureichende Möglichkeiten, diese entscheidend zu verbessern, stellen in vielen Projekten ein hohes Risiko für eine erfolgreiche Analyse dar. Ist die Datenbasis hinsichtlich Relevanz, Menge oder Qualität unzureichend, sind Maßnahmen zur Erhöhung der Datenqualität einzuleiten oder alternative Datenquellen zu erschließen. Durch einen Sprung in die erste Phase können auch Ziele und Erwartungshaltungen korrigiert werden.

2.3.3.3 Data Preparation

Ziel der Data-Preparation-Phase ist die Konstruktion eines Datensatzes in einem Format, das für die nachfolgend eingesetzten Methoden in der Modellierungsphase geeignet ist. Elementare Schritte der Datenaufbereitung sind die Auswahl der zu verwendenden Datensätze (Zeilen) und Attribute, Bereinigung, Transformationen wie etwa Aggregationen, Normalisierungen oder das Ableiten neuer Merkmale, das Zusammenfügen von Daten aus verschiedenen Tabellen oder Datenquellen (Integration) und die Formatierung. Während durch Bereinigung und Transformation semantische Änderungen an den Daten vorgenommen werden, geht es bei der Formatierung ausschließlich um Änderungen syntaktischer Natur, um die Daten entsprechend der vorgesehenen Analysemethoden und Werkzeuge aufzubereiten.

Da verschiedene Modellierungsmethoden sehr unterschiedliche Anforderungen an die Daten haben können, gibt es nicht die eine geeignete Aufbereitung der Daten. Die meisten Standardverfahren der Datenanalyse arbeiten mit Daten in Form einer Matrix. Aber z. B. die Reihenfolge der Attribute und akzeptierte Datentypen und deren Kodierungen mögen variieren. Da in einem Analyseprojekt oft verschiedene Analysemethoden angewendet werden, erfordert dies ein ständiges Wechselspiel zwischen der Datenvorbereitung und der folgenden Modellierungsphase. Deshalb und insbesondere auch dann, wenn die zu verwendenden Daten nicht bereits integriert und bereinigt aus einem Data-Warehouse kommen, gehört die Datenvorbereitung zu den zeitaufwendigsten Phasen. Erfahrungsgemäß werden häufig etwa 70–80% der Gesamtprojektzeit dafür veranschlagt.

2.3.3.4 Modeling

Die Modellierungsphase gilt auch als Data-Mining im engeren Sinne im Prozess der Wissensentdeckung in Datenbanken. In der Praxis wird jedoch meist der gesamte hier dargestellte Prozess als Data-Mining bezeichnet. Ziel der Modellierungsphase ist die Auswahl und Anwendung geeigneter Analysemethoden, die je nach Kontext auch als Mo-

dellierungstechniken oder Lernverfahren bezeichnet werden. Schließlich sollte bereits in dieser Phase eine Bewertung der Analyseergebnisse insbesondere zur Vermeidung von Overfitting, d. h. der Überanpassung eines Modells, erfolgen. Eine Evaluierung unter Berücksichtigung des unternehmerischen Kontexts und der fachlichen Fragestellung wird in der folgenden Phase stattfinden.

Mithilfe der Analysemethoden soll aus den Daten ein „Modell“ erzeugt werden. Im weitesten Sinne ist mit einem Modell ein Analyseergebnis gemeint. Was sich konkret dahinter verbirgt, hängt einerseits von den Analyzezielen und andererseits von den gewählten Methoden ab. Bei einer Klassifikationsaufgabe handelt es sich bei dem Modell um einen Klassifikator. Je nach Lernverfahren und der damit verbundenen Sprache zur Repräsentation der Lernergebnisse können sich dahinter zum Beispiel Entscheidungsbäume, Regelmengen oder die Parameter eines bestimmten Funktionstyps verbergen. Bei anderen Analyseaufgaben mag der Begriff des Modells zur Beschreibung der Ergebnisse weniger geläufig sein. Aber auch die Einteilung eines Datensatzes in Cluster oder die verständliche Beschreibung von Daten oder Konzepten sei in diesem Kontext als Modell zu verstehen und deren Erzeugung aus Daten als Modellerstellung bezeichnet.

Da es je nach Analyseaufgabe nicht das eine Lernverfahren gibt, das konsistent bei allen Anwendungen das beste ist, müssen geeignete Lernverfahren ausgewählt werden. Bei der Auswahl gibt es neben der eigentlichen Analyseaufgabe verschiedene wichtige Einflussfaktoren, wie etwa die Beschaffenheit der Daten, die Erfahrung der Modellierer und die Verfügbarkeit von Lernverfahren in entsprechenden Analyse-Werkzeugen. Empfehlenswert ist es, anfangs einfache Lernverfahren zu verwenden, deren Funktionsweise gut verstanden wird. Komplexere Lernverfahren können später zum Einsatz kommen, wenn mehr Verständnis für das vorliegende Problem mit seinen Daten aufgebaut wurde und die erreichte Qualität der Ergebnisse nicht genügt. Komplexere Lernverfahren haben meist mehr Stellschrauben (Parameter), die bei falscher Wahl zu deutlich schlechteren Ergebnissen führen können als einfache Lernverfahren (Domingos 2012).

Beim Lernen aus Daten stellt das sogenannte Overfitting sicherlich das größte Risiko für einen erfolgreichen operativen Einsatz dar. Overfitting tritt auf, wenn ein Modell zufällige Eigenarten der Daten (Rauschen) anstatt der zu Grunde liegenden Zusammenhänge (das Signal) beschreibt. Eine zu hohe Modellkomplexität ist dabei die Hauptursache. Bei Vorhersagemodellen bedeutet Overfitting, dass die Modellgüte bei neuen Daten deutlich schlechter ist als auf den Trainingsdaten. Diese Problematik sollte bereits in der Modellierungsphase erkannt und möglichst gebannt werden, sodass keine überangepassten Modelle als Analyseergebnisse an die Evaluierungsphase weitergegeben werden. Zur Erkennung von Overfitting muss die Modellgüte auf nicht zum Lernen verwendeten Daten bewertet werden. Dies erfordert das Aufteilen der verfügbaren Daten in eine Trainingsmenge und eine Test- bzw. Validierungsmenge. Durch Kreuzvalidierung kann dem Problem einer zu kleinen Trainingsmenge begegnet werden. Neben der Erkennung stellt die Vermeidung von Overfitting ein wichtiges Konzept dar. Ein frühzeitiges Beenden des Lernvorgangs oder das Bestrafen unnötiger Komplexität, beispielsweise durch Regularisierung, sind übliche Wege dafür.

Häufig werden nicht mehr nur einzelne Modelle als Ergebnis verwendet, sondern sogenannte *Model Ensembles*. Wie bei einem Expertengremium soll die Kombination mehrerer Meinungen (Ergebnisse) im Mittel deutlich treffendere Resultate ergeben. Unterschiedliche Ansätze, wie z. B. Bagging oder Boosting, erreichen durch gezielte Variation der Trainingsdaten mit demselben Lernverfahren unterschiedliche Ergebnisse, die kombiniert werden können (Domingos 2012). Aber auch die Kombination von Ergebnissen verschiedener Lernverfahren kann vorteilhaft sein, da jedes Verfahren Stärken in unterschiedlichen Lernsituationen haben kann.

2.3.3.5 Evaluation

Wenn nur intensiv genug gesucht wird, lassen sich in nahezu allen Datenbeständen auffällige Muster finden, selbst wenn die zugrunde liegenden Daten zufällig erzeugt wurden. Diese können sogar statistisch signifikant sein. Statistische Signifikanz ist wichtig, geht aber nicht zwingend mit Relevanz und Anwendbarkeit für ein Unternehmen einher. Eine geeignete Evaluierung ist daher essentiell, bevor ein Analyseergebnis zur Lösung einer fachlichen Fragestellung kommuniziert und zur Anwendung gebracht werden sollte.

Ziel der Evaluierung ist die Überprüfung des Nutzens der Analyseergebnisse im unternehmerischen Kontext und letztlich die Auswahl des Modells (oder des Modell-Ensembles), das die Aufgabenstellung unter Berücksichtigung der Erfolgskriterien am besten löst. Dies setzt natürlich voraus, dass die Ziele und Erfolgskriterien in der Phase Business Understanding entsprechend definiert wurden. Außerdem sollte in dieser Phase der gesamte bisherige Analyseprozess einem Review unterzogen werden.

Die Evaluierungsmethoden und die Evaluierungskriterien hängen stark von der Analyseaufgabe ab. Damit beispielsweise ein Prognosemodell für die Auswahl in Frage kommt, sollte es besser sein als ein Vergleichsmodell, das sogenannte Null-Modell. Letzteres kann z. B. durch Verwendung einfacher Regeln, wie etwa die Klassifikation anhand der häufigsten Klasse oder eines bedeutsamen (relevanten) Merkmals, gebildet werden. Liegen Ergebnisse aus einer bestehenden Lösung vor, sollten diese für den Vergleich herangezogen werden.

2.3.3.6 Deployment

In der letzten Phase des Prozessmodells werden Berichte und Abschlusspräsentationen zur Dokumentation des Analyseprozesses und Kommunikation der Ergebnisse bereitgestellt. Die Evaluierungsphase stellt sicher, dass die Analyseergebnisse den Anforderungen der geplanten Anwendung genügen. Für einen erfolgreichen Abschluss eines Analyseprojekts genügt das jedoch noch nicht. Ebenso wichtig sind eine verständliche und nachvollziehbare Dokumentation und eine überzeugende Präsentation gegenüber den Auftraggebern und beteiligten Fachbereichen.

Namensgeber für die Phase ist die Planung und Umsetzung der Anwendung der Ergebnisse, das sogenannte *Deployment*. Die Spannbreite der möglichen Nutzung kann dabei von einem einmaligen Erkenntnisgewinn durch Interpretation und Verstehen der Ergebnisse über eine einmalige eher manuelle Anwendung eines Modells auf einen ausgewählten

Datenbestand bis hin zur regelmäßigen automatischen Anwendung durch Integration in die betroffenen Geschäftsprozesse erfolgen. Bei einer regelmäßigen Anwendung sollte auch eine Kontrolle und Wartung eines Modells erfolgen, um auf Veränderungen im Unternehmenskontext oder der Umwelt gezielt reagieren zu können.

2.3.4 Big Data Analytics: Was ist anders?

Die Datenmenge, deren Vielfalt und die hohe Geschwindigkeit, mit der die Daten erzeugt werden und verarbeitet werden sollen, das sind die primär identifizierten Eigenschaften von Big Data. Diese stellen große Herausforderungen sowohl an das Datenmanagement als auch an die Analyse. Der erzielte technologische Fortschritt ist enorm. Die Big Data-Technologien werden gar als disruptiv eingeschätzt, vergleichbar durchaus mit dem Internet. Für eine erfolgreiche Adaption in Unternehmen liegt der Fokus jedoch noch zu stark bei den Technologien. Dass diese verfügbar und zu vertretbaren Kosten nutzbar sind, sind notwendige Voraussetzungen für die Lösung von Big Data-Problemen. Die erfolgskritischen Faktoren zur Schaffung wirtschaftlichen Nutzens aus Big Data liegen dann aber vor allem an den Fähigkeiten der Nutzer, die Daten problemadäquat analysieren und interpretieren zu können und auf gewonnenen Erkenntnissen auch angemessene Maßnahmen folgen zu lassen.

Bei Datenanalysen spielt aufgrund von Faktoren wie benötigtem Speicherplatz und Rechenzeit die Infrastruktur und somit die Technologie eine sehr wichtige Rolle. Aber auch die Verfügbarkeit und die Eigenschaften des Rohstoffs Daten sowie der Mensch als möglicher Engpass sind als Einflussfaktoren nicht zu unterschätzen (Domingos 2012). Daher werden im Folgenden die Auswirkungen von Big Data auf den Analyseprozess und seine Ergebnisse entlang der Dimensionen *Daten*, *Technologie* und *Mensch* untersucht, um insbesondere Herausforderungen und Risiken zu identifizieren.

2.3.4.1 Einfluss der Daten auf den Analyseprozess

Die Untersuchung der Dimension Daten orientiert sich am zuvor dargestellten Prozessmodell für Analyseprozesse CRISP-DM. Aufgrund der generischen Aufgabenbeschreibung und vollständigen Abdeckung des Analyseprozesses lässt es sich unmittelbar auch auf Big Data-Fragestellungen anwenden. Entscheidend ist, an welcher Stelle im Prozess die zentralen Eigenschaften von Big Data, also Volume, Velocity, Variety und Veracity, einen besonderen Einfluss haben, den es zu beachten gilt. Im darauf folgenden Abschnitt zur Dimension Technologie werden dann die wichtigsten dabei identifizierten Aspekte aufgegriffen.

Business Understanding

Jede sinnvolle und erfolgreiche Datenanalyse erfordert im Vorfeld eine relevante Problemstellung, die im Rahmen eines unternehmerischen Anwendungsszenarios (Business Case) definiert wurde. Eine angemessene Zielsetzung unter Berücksichtigung strategischer Vorgaben und der gegebenen Rahmenbedingungen ist der elementare erste Schritt

im Analyseprozess, ohne den ein Analyseergebnis wertlos ist, daran ändert auch Big Data nichts.

Allerdings wird gerade im Kontext von Big Data Analytics vermehrt die Möglichkeit gepriesen, durch explorative Datenanalyse, insbesondere auch mit fortschrittlichen Visualisierungstechniken, die Daten sprechen zu lassen, um beeindruckende Antworten auf Fragen zu erhalten, die man nicht einmal vorher stellen musste. Im Rahmen einer seriösen Datenanalyse, deren Erfolg anhand definierter Ziele bewertet werden soll, ist diese Vorgehensweise allerdings fragwürdig. Es sollte nicht auf die Daten gehört werden, sondern auf die Analyseergebnisse. Mit einem übergeordneten Ziel im Hinterkopf können diese Wundermittel zur Unterstützung des Analyseprozesses in der Data-Understanding-Phase jedoch sehr wertvoll sein.

Big Data-Technologien führen im Hintergrund zu einer starken Veränderung der Art und Weise der Datenverarbeitung. Eine grundlegende Veränderung hinsichtlich der Fragen, warum Analysen durchgeführt werden, welche Analyseaufgaben also im Kern bearbeitet werden und wie der Erfolg bewertet wird, gibt es allerdings nicht. Wohl aber verschieben sich die Schwerpunkte und die Einschätzung dessen, was basierend auf den verfügbaren Daten und technologischen Möglichkeiten machbar ist. Mit größeren und vielfältigeren Datenmengen lassen sich sicherlich ambitionierte Ziele verfolgen (Dominigos 2012). Es zeigt sich, dass viele der für die relevanten Fragestellungen nützlichen Methoden schon vor der Big Data-Ära entwickelt wurden, nun aber vermehrt und mit deutlich mehr Beachtung zum Einsatz kommen. Dass natürlich eine Weiterentwicklung von Analysemethoden durch aktuelle Trends wie zurzeit Big Data beschleunigt wird, steht außer Frage.

Der Fokus vieler Big Data-Anwendungen liegt deutlich häufiger bei prädiktiven und sogar präskriptiven Aufgaben, als dies bei klassischen BI-Anwendungen der Fall ist. Es ist daher deutlich häufiger der Einsatz fortschrittlicher Analysemethoden gefragt, also Advanced Analytics. Die Auffassung, dass sich Big Data von BI dadurch abgrenzt, dass es prädiktive und präskriptive Analysen ermögliche, während BI ausschließlich vergangenheitsorientierte Analysen böte, ist jedoch falsch, wenn man berücksichtigt, dass mit dem Ziel der Entscheidungsunterstützung auch bei BI alle bekannten Analysemethoden nutzbar sind (vgl. hierzu auch die Diskussion zum Analysespektrum im Abschn. 4.1 über die Grenzen klassischer BI-Lösungen).

Ein zweiter Schwerpunkt liegt in der Analyse unstrukturierter Daten, die meist einen Großteil des gesamten Datenbestands ausmachen. Insbesondere ist zurzeit die Analyse von Textdokumenten hervorzuheben. Aber auch Text Analytics ist keineswegs neu. Eine typische Aufgabe ist zum Beispiel die Sentimentanalyse. Diese lässt sich gerade im Vergleich zur früher oft betrachteten Zuordnung von Textdokumenten zu bekannten Themen als deutlich anspruchsvollere Textklassifikationsaufgabe auffassen.

Aufgrund der großen Datenmengen, die in hoher Geschwindigkeit und breiter Vielfalt auf ein Unternehmen zuströmen, ist es undenkbar, dass diese manuell verwertet werden können. Zudem sollen Daten in immer kleiner werden Analysefenstern und in immer kürzeren Zeitabständen analysiert werden, da sie sonst schon wieder obsolet sind und

möglicherweise relevante Entscheidungen nicht getroffen werden konnten. Daher werden die Echtzeitverarbeitung bei der Analyse (Real-time Analytics) und die Einbettung der Analysen bzw. der Analyseergebnisse in die Geschäftsprozesse (Embedded Analytics) und die Verarbeitung von Datenströmen (Stream Analytics) immer wichtiger. Da Datenströme üblicherweise höchstens in Auszügen persistiert werden, wird das, was und vor allem wann etwas mit den Daten gemacht werden kann, stark eingeschränkt.

Die automatische kontinuierliche Echtzeitüberwachung (Monitoring) von Datenströmen möge als Beispiel für eine typische Analyseaufgabe dienen. Die Welt und die Daten, die diese beschreiben, ändern sich ständig. Unternehmen müssen diese Veränderungen erkennen, um darauf reagieren zu können. Mithilfe eines geeigneten Monitorings kann auf Auffälligkeiten und Veränderungen in den Datenströmen und somit auf mögliche Entscheidungsfelder hingewiesen werden.

Es genügt jedoch nicht, die Datenströme ständig zu überwachen. Es muss auch eine entsprechende Bewusstheit und Einstellung im Unternehmen geschaffen werden, auf derartige Erkenntnisse zu achten und darauf bei Bedarf zu reagieren. Dies gilt natürlich genauso für alle anderen Ergebnisse der Datenanalyse. Analysen müssen daher in geeigneter Weise in Geschäftsprozessen bzw. Entscheidungsprozessen integriert werden (siehe Abschn. 2.3.4.3 und Davenport, Barth, and Bean 2012).

Abschließend sei mit Blick auf die Rahmenbedingungen einer Aufgabenstellung die Einhaltung gesetzlicher Bestimmungen hervorgehoben. Da durch Big Data immer mehr Daten aus einer steigenden Anzahl von Datenquellen zusammengeführt, potenziell gespeichert, verarbeitet und weitergegeben werden, die insbesondere auch die Privatsphäre einzelner betrifft, ist beispielsweise die Einhaltung gesetzlicher Vorgaben zum Datenschutz und zur Datensicherheit ein wichtiger Aspekt und kann bei Nichtbeachtung zum Erfolgsrisiko werden. Für eine detaillierte Betrachtung dieser Themen und anderer rechtlicher Aspekte, wie etwa die Beachtung des Urheberrechts bei der Speicherung und Analyse von Anwenderbeiträgen aus fremden Anwendungen, sei auf den rechtlichen Teil dieses Handbuchs verwiesen. Eine Klärung, was rechtlich möglich und auch was moralisch oder ethisch vertretbar ist, muss so früh wie möglich und vor der Durchführung weiterer Analyseschritte erfolgen.

Data Understanding

In Bezug auf die Data-Understanding-Phase sollen die Aspekte Datenqualität sowie Eigenschaften und Nutzen der Daten betrachtet werden.

Die Daten in einem internen operativen oder analytischen System sind typischerweise sorgfältig modelliert und haben je nach Fragestellung einen gewissen Wert für ein Unternehmen. Mit sinkenden Speicherkosten wird der sorgfältigen Modellierung und Auswahl relevanter Daten immer weniger Bedeutung zugeschrieben. Dagegen ist der überwiegende Anteil der Daten bei Big Data für eine konkrete Aufgabenstellung belanglos. Big Data stellt daher für eine Aufgabe überwiegend Rauschen dar. Und meist ist es eine große Herausforderung, den kleinen relevanten Anteil, also das Signal, zu erkennen und zu nutzen (Franks 2012, S. 16–18).

Die sehr großen Datenbestände prägen namentlich den Begriff Big Data. Auf den ersten Blick sollte angenommen werden, dass mehr Daten beim Lernen aus Daten immer vorteilhaft sind. Zunächst soll überlegt werden, wie die Größe eines Datenbestandes zu stande kommt. Da selbst unstrukturierte Daten mit geeigneter Datenvorverarbeitung in eine strukturierte Form gebracht werden können, was üblicherweise auch genau so geschieht, stellen wir uns Daten der Einfachheit halber als Matrix vor. Dabei ordnen wir die betrachteten Objekte in Zeilen an und assoziieren die verfügbaren Attribute mit den Spalten. In vielen Big Data-Anwendungsszenarien stehen sehr viele, wenn nicht gar alle Objekte einer Grundgesamtheit zur Verfügung. Folglich ist die Anzahl der Zeilen in der Regel sehr groß. Aber aufgrund der Vielfalt möglicher Datenquellen und der Komplexität vielen Daten kann auch die Anzahl verfügbarer oder konstruierter Attribute und daher die Spaltenanzahl sehr groß sein. Natürlich können auch Zeilen- und Spaltenanzahl sehr groß sein.

Mehr Spalten stellen für eine Analyse nicht automatisch einen Vorteil dar. Ganz im Gegenteil, eine größere Spaltenanzahl kann sogar kontraproduktiv sein, denn wenn die zusätzlichen Attribute für eine Analyseaufgabe keine relevanten Informationen liefern und stattdessen aufgrund fehlender oder unzureichender Auswahl relevanter Attribute der sogenannte Fluch der Dimensionalität (*curse of dimensionality*) zuschlägt, können die erzielten Ergebnisse sogar schlechter werden (Domingos 2012). Beispielsweise könnte eine Analysemethode durch sehr viele Attribute „verwirrt“ werden, weil sich Ähnlichkeiten in hochdimensionalen Räumen nicht mehr vernünftig deuten lassen.

Geht es aber um die Anzahl der Zeilen im Datensatz, also die Anzahl der Objekte, die bei der Analyse berücksichtigt werden, dann sollten zusätzliche Daten für die Modellgüte zumindest nicht nachteilig sein. Bei sehr großen Datenmengen kann allerdings der zusätzliche Nutzen von immer mehr Objekten irgendwann verschwindend gering werden. Gleichzeitig ist es wahrscheinlich, dass die zusätzlich benötigte Rechenleistung oder der zusätzliche Speicherbedarf ab einem gewissen Punkt als größerer Nachteil empfunden wird. Spätestens dann sollte überlegt werden, auf einer Stichprobe der Daten zu arbeiten. Big Data mit verfügbarer skalierbarer Technologie auch für die einzelnen Schritte im Analyseprozess mag zwar die Möglichkeit bieten, alle verfügbaren Daten zu bearbeiten, aber dies bedeutet nicht, dass alle Daten tatsächlich verwendet werden müssen. Stattdessen ist die gezielte Verwendung von Stichproben in Erwägung zu ziehen (Franks 2012, S. 200).

Bei einem überwachten Lernszenario werden Trainingsdaten benötigt, für die die wahren Werte des Zielattributs bekannt sind. Bei einigen Aufgaben werden die wahren Werte nach einer gewissen Zeit stets bekannt sein wie etwa bei der Frage, ob eine Transaktion einen Betrugsvorversuch darstellt oder nicht. In vielen Fällen erfordert die Bereitstellung einer Trainingsmenge jedoch eine mühsame manuelle Bestimmung der Werte für das Zielattribut, also einen hohen Aufwand. Genau dann mögen auch die Trainingsmengen im Kontext von Big Data nicht mehr wirklich groß sein, sodass für die Modellerstellung nicht zwingend Big Data-Technologie zum Einsatz kommen muss. Dies wird auch in der Modellierungsphase und bei den technologischen Aspekten thematisiert.

Damit die Auswahl und Verwendung der Daten im Analyseprozess erleichtert wird, ist ein einheitlicher, transparenter Zugriff über eine Art Data-Hub kombiniert mit Kolaborationstechniken sehr sinnvoll (Davenport, Barth, and Bean 2012). Letztere fördern Diskussionen und Kommentare über die Relevanz und Qualität verschiedener Datenquellen und einzelner Datenfelder (Attribute). Gut gepflegte Metadaten sind wichtig, die in Form eines erweiterten Datenkatalogs angeboten werden sollten. Insbesondere sollte auch Auskunft über die Herkunft und Glaubwürdigkeit (Veracity) der Datenquellen und über weitere Datenqualitätskennzahlen zu den einzelnen Attributen idealerweise unter Berücksichtigung der betrachten Fragenstellungen gegeben werden.

Bei realen Anwendungen sind die Daten meistens schmutzig und die Datenqualität ist deshalb immer zu hinterfragen. Die Beurteilung der Datenqualität bezieht sich stets auf den geplanten Einsatzzweck. Anders als bei geplanten und kontrollierten Experimenten sind die relevanten Daten üblicherweise ursprünglich nicht für einen Analysezweck erzeugt und gespeichert worden, sondern für einen anderen, meist operativen Zweck wie etwa die Unterstützung von operativen Geschäftsprozessen. Die verschiedenen Einsatzzwecke gehen in der Regel mit unterschiedlichen Ansprüchen und Anforderungen an die Daten einher. Selbst wenn die Datenqualität für den originären Einsatzzweck akzeptabel ist, für eine geplante Analyse muss dies nicht zutreffen.

Schon beim Data Warehousing als Bestandteil einer traditionellen BI-Lösung ist die Sicherstellung der Datenqualität eine große Herausforderung, die viele Unternehmen nicht in angemessenem Maße bewältigen. Dabei geht es primär um strukturierte Daten, die überwiegend aus internen Datenquellen stammen. Auf diese könnte ein Unternehmen allerdings im Rahmen eines Datenqualitätsmanagements unmittelbar Einfluss nehmen, um Veränderungen zu bewirken.

Im Kontext von Big Data mit zahlreichen externen Datenquellen und einer hohen Dynamik sind die Voraussetzungen für ein erfolgreiches Datenqualitätsmanagement ungleich schwieriger. So mögen maschinengenerierte Daten etwa bis auf gelegentliche Ausfälle von Sensoren oder Unstimmigkeiten bedingt durch Prozess- oder Formatänderungen kein nennenswertes Qualitätsproblem haben. Bei der breiten Masse der sogenannten nutzergenerierten Inhalte in den sozialen Medien sieht dies jedoch anders aus. Zum einen können die Inhalte von zweifelhafter Qualität sein, was auch eine gezielte Manipulation von Daten und Falschangaben beispielsweise bei Produktbewertungen einschließt. Zum anderen lässt sich die Identität der Nutzer nicht immer eindeutig ermitteln und mit den in internen Datenbeständen geführten Kunden in Verbindung bringen. Letztlich lässt sich auf externe Quellen selten Einfluss nehmen, um eine Steigerung der Datenqualität zu bewirken.

Dem Garbage-In-Garbage-Out-Prinzip folgend stellt die oft mangelhafte Datenqualität ein erhebliches Erfolgsrisiko dar. Aber wie stark beeinflusst die Datenqualität tatsächlich die Analyseergebnisse? Relevante Erkenntnisse mögen sich auch trotz Problemen mit der Datenqualität gewinnen, wenn es gelingt, diese in angemessener Weise bei der Interpretation der Ergebnisse zu berücksichtigen. Dies ist umso wichtiger, wenn eine Integration mit Ergebnissen basierend auf verlässlicheren internen Daten angestrebt wird. Wenn sich also die Qualität externer Daten nicht nennenswert beeinflussen lässt und eine Korrektur

nur mit hohem Aufwand sehr begrenzt möglich ist, dann sollte das Qualitäts- und Vertrauensniveau der Datenquellen und der Datenfelder wenigstens entsprechend gekennzeichnet werden.

Abschließend sei die Dualität zwischen Rechtzeitigkeit und Konsistenz bei Echtzeit-Anwendungen in Kombination mit Datenströmen hervorgehoben. Als Variante des CAP-Theorems für verteiltes Datenmanagement (siehe auch Abschn. 4.3.3) lässt sich für die Echtzeitverarbeitung das SCV-Theorem (Speed-Consistency-Volume) formulieren (Mohanty, Jagadeesh, and Srivatsa 2013, S. 225). Geht man davon aus, dass das Datenvolumen als solches gesetzt ist, da potenziell alle verfügbaren Daten verarbeitet werden sollen, dann muss es stets einen Kompromiss zwischen der Konsistenz der Daten sowie den daraus abgeleiteten Analyseergebnissen und der Verarbeitungsgeschwindigkeit geben. Wird daher zusätzlich eine Echtzeitverarbeitung gefordert, so geht dies im Allgemeinen zu Lasten der Konsistenz, etwa weil die relevanten Daten noch nicht vollständig oder zum Teil schon nicht mehr verfügbar sind, aber auch weil nicht ausreichend Zeit für einen umfassenden Prozess zur Bereinigung und Integration zur Verfügung steht.

Data Preparation

Erfahrungsgemäß entfallen häufig 70–80% der Bearbeitungszeit auf die Datenaufbereitung. Im Kontext von Big Data wird dieser Anteil deutlich größer werden, da viele Daten als Konsequenz des Schema-on-Read-Ansatzes beim Lesen erst noch richtig interpretiert werden müssen. Auch das Identifizieren und Extrahieren der Datenelemente innerhalb der Big Data-Quellen, die für einen Anwendungsfall tatsächlich einen Wert haben könnten, und die Transformation unstrukturierter Daten in eine strukturierte Form durch das Konstruieren relevanter Attribute hat in der Regel einen erheblichen Aufwand zur Folge.

Der Schema-on-Read-Ansatz bietet die Möglichkeit, die Rohdaten ohne Verluste durch irgendwelche Qualitätsprüfungen zu speichern. Gerade in Anbetracht einer dynamischen Welt, in der Datenquellen und Datenformate Änderungen unterliegen, die nicht zwingend kommuniziert werden, ist das eine große Hilfe. Die Interpretation der Daten kann auf die Zeit verschoben werden, in der die Daten tatsächlich verwendet werden sollen. Und da jede Nutzung der Daten abhängig vom Einsatzzweck anders aussehen kann, können jeweils auch andere Maßstäbe an die Aufbereitung und Interpretation der Daten gelegt werden. Dafür stehen dann stets die unverfälschten und vollständigen Rohdaten zur Verfügung. Das ist ein sehr großer Vorteil, der aber zu einem hohen Preis erkauft wird. Das Vorgehen birgt die Gefahr, dass zahlreiche Schema-Varianten kursieren und verwendet werden. Inkonsistenzen bei der Schema-Interpretation sind die sehr wahrscheinliche Konsequenz. Dies erinnert an die Entwicklung von Datenbankanwendungen, bevor durch die sogenannte ANSI-SPARC-Architektur eine Trennung verschiedener Beschreibungsebenen für Datenbankschemata ein hohes Maß an Datenunabhängigkeit erreicht wurde. Im Kontext von Big Data müssen Werkzeuge weiterentwickelt und bereitgestellt werden, die die Verwendung nutzbarer und sinnvoller Schemata für den Schema-on-Read-Ansatz unterstützen und überwachen. Dies könnte beispielsweise im erweiterten Datenkatalog verankert werden.

Die Integration von Daten aus verschiedenen Quellen ist eine weitere große Herausforderung. Der Nutzen für ein Unternehmen wird deutlich größer sein, wenn Big Data mit internen Daten kombiniert werden kann (Franks 2012, S. 21). Oft ist aber schon das Verbinden semantisch zusammenhängender Daten aus verschiedenen internen Quellen schwierig, da beispielsweise unterschiedliche Primärschlüssel definiert wurden und eindeutige Zuordnungen nicht immer bekannt sind. Kommen nun auch noch externe Daten hinzu, gibt es oft mangels Kenntnis der Modellierung und Einflussmöglichkeiten auf die Systeme noch mehr Schwierigkeiten.

Modeling

Die größten Herausforderungen hinsichtlich der mit großen Datenmengen einhergehenden Anforderungen an die Rechenleistung stammen neben der Datenvorbereitung von der Modellerstellung aus der Modellierungsphase und der Modellanwendung (Deployment). Dieser Sachverhalt ist neben anderen Kriterien auch bei der Auswahl geeigneter Analysemethoden zu berücksichtigen. Dabei ist es ganz elementar, stets genau zwischen Modellerstellung und Modellanwendung zu unterscheiden, insbesondere dann, wenn es um Echtzeit-Anwendungen und die Verarbeitung von Datenströmen geht. Diese Unterscheidung wird nicht in allen Bereichen deutlich gemacht, obwohl sie einen entscheidenden Einfluss auf die erforderlichen Technologien und Methoden im Rahmen der jeweiligen Phasen des Analyseprozesses hat. In vielen Fällen ist es nämlich ausschließlich die Modellanwendung, die mit Blick auf große Datenmengen und dem Einsatz in Echtzeit-Anwendungen kritisch ist.

Bei der Modellerstellung gibt es in Abhängigkeit der Analyseaufgabe verschiedene Kriterien für die Auswahl geeigneter Analysemethoden (Lernverfahren). Wie bereits an anderer Stelle erwähnt, spielen Verständlichkeit und Interpretierbarkeit der Analyseergebnisse im Unternehmenskontext oft eine ganz zentrale Rolle. Innerhalb dieser Vorgaben gibt es oft die Wahl zwischen einfachen und komplexeren Lernverfahren. Da die einfachen Verfahren oftmals in ihrer Funktionsweise und ihren Anwendungsvoraussetzungen besser verstanden werden und meist weniger Parameter haben, die den Analyseerfolg erheblich beeinflussen können, ist es empfehlenswert, zunächst mit einfachen Verfahren zu beginnen. Bei einfachen Verfahren, die Modelle mit geringer Komplexität erzeugen, ist die Gefahr des Overfittings geringer. Stattdessen ist der sogenannte Bias oft hoch, wobei es sich um unerwünschte Eigenschaften eines Modells handelt, die auf der gewählten Lernmethode basieren. Diese unerwünschten Eigenschaften lassen sich durch komplexere Modelle auf Kosten höherer Varianz reduzieren. Varianz bezeichnet in diesem Zusammenhang unerwünschte Variationen eines Modells basierend auf Artefakten, also Eigenschaften oder Besonderheiten, in den Trainingsdaten, die nicht für die Gesamtheit der Daten repräsentativ sind. Bei hoher Varianz führen kleinere Veränderungen in den Trainingsdaten zu großen Unterschieden bei den Analyseergebnissen. Die Verfahren oder Ergebnisse gelten dann nicht als robust und es besteht die Gefahr von Overfitting. Wenn Modelle direkt in Geschäftsprozesse eingebettet werden sollen und die menschlichen Kontrollmöglichkeiten sehr begrenzt sind, dann stellt dies ein großes Risiko dar. Zur

Vermeidung werden beispielsweise Techniken wie Regularisierung oder die Kombination mehrerer einfacher Modelle (Ensemble-Methoden) eingesetzt.

Ein anderer Weg für die Vermeidung von Overfitting ist die Verwendung von mehr Trainingsdaten. Hier kommt Big Data ins Spiel. Die Verfügbarkeit großer Trainingsmengen weckt die Hoffnung, dass auch komplexere Modelle mit geringer Gefahr des Overfittings erstellt werden können. Dabei ist jedoch Vorsicht geboten. Denn liegen gleichzeitig auch sehr viele beschreibende Attribute für die betrachteten Objekte vor, ist der Datenraum also hochdimensional, wird man selbst mit den großen Datenmengen bei Big Data oft nur einen sehr kleinen Teilbereich der möglichen Kombinationen der beschreibenden Attribute abdecken können. Der vermeintliche Vorteil besteht dann also gar nicht.

Ein anderer Aspekt ist die Tatsache, dass komplexere Lernverfahren oft weniger gut skalieren. Dies führt dazu, dass angesichts großer Datenmengen dann doch wieder auf einfachere Lernverfahren zurückgegriffen wird, für die skalierbare Implementierungen existieren. Dies ist eine paradoxe Situation: Mit mehr Daten können zwar prinzipiell komplexere Modelle erstellt werden, bei Verfügbarkeit großer Datenmengen werden wegen der extremen Rechenzeiten aber oftmals einfache Lernverfahren angewendet (Domingos 2012). Die Weiterentwicklung skalierbarer Lernmethoden ist hier gefragt (siehe Skalierbare Modellanwendung in Abschn. 2.3.4.2).

Bedeutet Big Data aber wirklich immer gleich große Datenmengen, auf denen Lernverfahren operieren? Wie groß sind die Datenmengen bei der Modellerstellung wirklich? Der für eine Big Data-Anwendung relevante Anteil des gesamten Datenbestandes ist oft gering. Was bleibt nach einer sinnvollen Auswahl und nach einer angemessenen Datenvorverarbeitung tatsächlich an Daten übrig? Wenn es sich um überwachte Lernaufgaben handelt, wie groß ist der Anteil der Daten, für den die Werte des Zielattributs bekannt sind? Und wie groß ist somit die Trainingsmenge? Es wird Situationen geben, in denen die Trainingsmenge sehr groß ist und skalierbare Lernmethoden unverzichtbar sind. Es wird aber auch Anwendungsszenarien geben, bei denen die Trainingsmengen letztlich überschaubar bleiben und Analysemethoden auf traditionelle Art und Weise anwendbar sind, also ohne dass Big Data-Technologien erforderlich wären. Allenfalls könnten dann geeignete Konnektoren zum einfachen Zugriff auf Daten aus dem Big Data-Umfeld die Arbeit erleichtern. Ein Beispiel dafür ist die Sentimentanalyse. Das Bereitstellen von Texten, für die eine Einschätzung des Sentiments manuell (durch Menschen) vorgenommen wurde, ist sehr zeitaufwendig und in der Regel teuer, wenn nicht gerade günstige Crowd-Sourcing oder Out-Sourcing Möglichkeiten bestehen.

Gerade die Problematik, die sich in der Anwendung komplexer Lernverfahren bei sehr großen Datenmengen aufgrund der Herausforderungen bei der Parallelisierung der Abläufe ergibt, mag erklären, warum der Einsatz sogenannter Ensemble-Methoden weiter an Beliebtheit zunimmt (siehe auch Abschn. 2.3.4). Durch Ensemble-Methoden lassen sich die teilweise mäßigen Ergebnisse einfacher Lernverfahren zu sehr guten Gesamtergebnissen kombinieren. Abgesehen von der Modellgüte und der Stabilität der Ergebnisse liegt der entscheidende Vorteil in der einfachen Parallelisierbarkeit der Erstellung der verschiedenen Modelle. Solange keine Abhängigkeiten durch iteratives Anpassen von

Trainingsdaten entstehen, kann die Modellerstellung jedes Modells im Ensemble unabhängig voneinander erfolgen.

Evaluation

Bezüglich der Evaluationsphase soll auf Risiken bei der Modellauswahl und der Modellinterpretation eingegangen werden. Zu den Interpretationsrisiken zählen insbesondere die irrtümliche kausale Deutung von Zusammenhängen, die selektive Auswahl bestimmter Analyseergebnisse und das multiple Testen. Diese Risiken gibt es nicht erst seit der Big Data-Ära, aber in großen Datenbeständen treten aufgrund der kombinatorischen Vielfalt einige Sachverhalte absolut gesehen häufiger auf. Dadurch stehen diese Risiken bei Big Data-Anwendungen deutlich stärker im Fokus.

Eine große Gefahr besteht darin, zur Unterstützung einer bestimmten Meinung oder geplanten Maßnahme sich genau die Analyseergebnisse herauszusuchen, die dafür gerade geeignet sind, während die anderen Ergebnisse bewusst missachtet werden. Dies wird auch als Rosinenpickerei (Cherry-Picking) bezeichnet. Schließlich lässt sich durch gezieltes Auswählen fast alles begründen. Richtig wäre es stattdessen, auf Basis aller relevanten Ergebnisse die beste Maßnahme zu bestimmen (Franks 2012, S. 189).

Auch die Problematik des multiplen Testens bzw. multipler Vergleiche bei Lernverfahren ist zu berücksichtigen. Bei einem klassischen Hypothesentest wird genau ein Sachverhalt untersucht. In vielen Lernszenarien werden viele Sachverhalte wie etwa die Auswahl eines Attributes durch Bestimmung der Stärke des Zusammenhangs zum Zielattribut auf derselben Datengrundlage evaluiert, um dann den Sachverhalt auszuwählen, bei dem die gewünschte Eigenschaft am stärksten ausgeprägt ist. Da es regelrecht normal ist, dass bei einer großen Anzahl getester Sachverhalte einige dabei sind, die mit statistischer Signifikanz auffallen, obwohl die Zusammenhänge faktisch nicht bestehen (zufällige Konzidenzen), kann dieses Vorgehen zu unbrauchbaren Ergebnissen führen. Eine mögliche Abhilfe ist die Bonferroni-Korrektur (Jensen and Cohen 2000).

Obwohl immer wieder betont wird, dass Korrelation keine Kausalität impliziert, werden dennoch Analyseergebnisse basierend auf Beobachtungen oft genug, möglicherweise auch unbewusst, fälschlicherweise im Sinne kausaler Zusammenhänge interpretiert und präsentiert. Wurden diese Ergebnisse zum Zweck des Erkenntnisgewinns erzeugt, dann ist dieser Fehler unbedingt zu vermeiden. Wenn das Ziel jedoch die Vorhersage eines Zielattributs ist, dann mag die korrekte Deutung des zu Grunde liegenden statistischen Zusammenhangs eine dem Prognoseziel untergeordnete Rolle spielen. Schließlich sind nach Aussage des Statistikers Georg Box alle Modelle falsch, aber einige immerhin nützlich. Und solange ein Prognosemodell gute Ergebnisse liefert, mag es für den Anwender keine Rolle spielen, ob die Zusammenhänge kausaler Natur sind oder nicht.

Im Rahmen von Big Data nimmt der Anteil der Modelle zu, die automatisiert und eingebettet in Geschäftsprozesse zur Anwendung kommen. Beispiele dafür sind Empfehlungssysteme oder die Platzierung von Werbung auf Webseiten. Dabei spielen Verständlichkeit und Interpretierbarkeit der Modelle eine untergeordnete Rolle. Modellgüte, also etwa die Korrektheit von Vorhersagen, Robustheit und auch eine schnelle Berechenbar-

keit bei der Modellanwendung sind dagegen in der Regel deutlich wichtiger. Bei vielen anderen, insbesondere interaktiven Analyseprozessen, die mit Erkenntnisgewinn einhergehen sollen und bei denen Menschen verschiedener Bereiche über Analyseergebnisse diskutieren, sind Verständlichkeit und Interpretierbarkeit mindestens ebenso wichtig wie die Modellgüte. Auch die Berücksichtigung des notwendigen Einsatzes menschlicher Arbeitskraft und Interaktion im Analyseprozess ist wichtig. An diesen grundsätzlichen Überlegungen hat sich durch Big Data jedoch nichts verändert.

Deployment

Wie in den vorausgehenden Phasen des Prozessmodells bereits angesprochen, haben die typischen Anwendungsszenarien einen großen Einfluss auf die Gestaltung der Nutzung der Analyseergebnisse. Die Einbettung der Analyseergebnisse in die Geschäftsprozesse erfordert geeignete Austauschformate für Modelle, die auch Beschreibungen der notwendigen Schritte für die Datenvorverarbeitung enthalten. Bei einem regelmäßigen Einsatz sind angemessene Möglichkeiten für die Überwachung und Wartung der Modelle notwendig. Dafür bietet sich eine Modell-Management-Komponente innerhalb einer Big Data-Architektur an. Diese könnte dann auch die Kommunikation und Diskussionen über die Modelle im kollaborativen Sinne unterstützen (Davenport, Barth, and Bean 2012).

Mit Blick auf die Technologie ist gerade bei einer Echtzeit-Anwendung auf kurze Berechnungszeiten im Rahmen der Modellanwendung zu achten. Allein dieser Aspekt ist oftmals gemeint, wenn von Echtzeit-Analyse die Rede ist. Eine vollständige Echtzeit-Modellbildung ist dagegen selten erforderlich und auch kaum realisierbar, solange Menschen die Analyseergebnisse vor der Anwendung kontrollieren sollen. Ist eine Kontrolle nicht erforderlich, dann ist jedoch eine Modellanpassung wie etwa beim aktiven Lernen (*Active Learning*) oder bei der Berücksichtigung von Anwenderrückmeldungen zur Relevanz oder Korrektheit der Ergebnisse auch in Echtzeit realistisch.

2.3.4.2 Technologische Aspekte

Bei der Betrachtung von Algorithmen in der Informatik stellen Speicherplatz und Zeit traditionell die zwei wichtigsten begrenzten Ressourcen dar. Die eine lässt sich üblicherweise nur auf Kosten der anderen optimieren. Durch Big Data-Technologien scheint der Speicherplatz kaum noch eine Rolle zu spielen. Aber Zeit ist der kritische Faktor. Wie können riesige Datenmengen effizient verarbeitet werden?

Aus technologischer Perspektive ist Skalierbarkeit der Schlüssel zum Erfolg bei Big Data. „Teile und herrsche“ lautet die altbekannte Strategie, um mit großen Problemen umzugehen. Die massiv parallele Verarbeitung (Processing), kurz MPP, erlaubt das Aufteilen von Teilschritten einer Berechnung auf viele Rechner, sogenannten Knoten in einem Cluster. Das inzwischen sehr prominente Map-Reduce-Programmiermodell kommt hier sehr oft zum Einsatz. Details zu Map-Reduce insbesondere im Kontext von Hadoop werden in Abschn. 4.3.2 erläutert. Aber auch Weiterentwicklungen und Alternativen dazu, die etwa versuchen die benötigten Daten im Hauptspeicher zu halten oder mehr Operatoren als das rudimentäre „Map“ und „Reduce“ bereitzustellen, kommen inzwischen zur Anwen-

dung. Die Ideen der parallelen Datenverarbeitung sind im Kern nicht neu. Aber dennoch kommen durch Big Data-Technologien starke Veränderungen in der Art und Weise der Datenverarbeitung auf breiter Front daher. Technologien und Werkzeuge, insbesondere zur Lösung der Volume-Velocity-Variety-Herausforderung, machen den Umgang mit Big Data zu vertretbaren Kosten inzwischen möglich. Das Datenmanagement scheint inzwischen beherrschbar. Im Folgenden sollen technologische Aspekte der Analyse von Big Data betrachtet werden.

Datenvorverarbeitung, Modellerstellung und Modellanwendung wurden zuvor als die elementaren Phasen des Analyseprozesses identifiziert, bei denen Skalierbarkeit der eingesetzten Methoden eine entscheidende Rolle spielt. Lösungen zu diesen Herausforderungen sollten, verpackt in Form geeigneter Big Data-Analytics-Werkzeuge, eine transparente Nutzung erlauben, mit der die im Rahmen von Datenanalysen erforderlichen Berechnungen effizient dort durchgeführt werden, wo sich die Daten befinden.

Skalierbare Datenvorverarbeitung und Modellanwendung

Ein großer Teil der rechenintensiven Schritte fällt im Rahmen der Datenvorverarbeitung an (siehe Abschn. 2.3.3.3). Sehr häufig handelt es sich um einfache Operationen, die zeilenweise auf den Attributen durchgeführt werden, oder um einfache Aggregationen mehrerer Zeilen. Die meisten Berechnungen lassen sich ohne Probleme auf verschiedene Knoten verteilen und somit parallelisieren. Genügend verfügbare Rechenleistung vorausgesetzt, lässt sich dadurch eine enorme Beschleunigung erreichen.

Skalierbare Modellerstellung

Bei der Anwendung von Lernverfahren ist eine zentrale Frage, ob sie für große Datenmengen geeignet sind. Die Entwicklung und Anwendung skalierbarer Lernverfahren ist eine der größten Herausforderungen beim Lernen aus großen Datenmengen, also bei Advanced Analytics mit Big Data.

Üblicherweise durchläuft ein Lernverfahren die Trainingsdaten, um Aggregationen (meist Summen) und andere Kennzahlen zu berechnen. Diese werden dann verwendet, um die Parameter eines zugrunde liegenden Modells zu bestimmen oder zu optimieren (Wu et al. 2014). In vielen Fällen sind mehrere Iterationen dieses Vorgangs notwendig und die Abhängigkeiten zwischen den Berechnungen können komplex sein. Soll auf Basis großer Datenmengen ein Modell erstellt werden, muss sorgfältig auf die Laufzeit- und Speicherkomplexität geachtet werden.

Ein einfaches Beispiel soll demonstrieren, wie schnell dies bei großen Datenmengen zu Problemen führen kann. Ältere hierarchische Clusterverfahren aus der klassischen Statistik verwenden zur Konstruktion der Objekthierarchie paarweise Ähnlichkeiten zwischen allen Objekten. Dies bedeutet, dass der Rechenaufwand quadratisch von der Anzahl der Objekte abhängt. Schon bei mäßig großen Datenmengen werden diese Verfahren praktisch unausführbar.

Verfahren, bei denen die Komplexität stärker als linear mit dem Datenvolumen wächst, können bei realen Anwendungen mit sehr großen Datenmengen oft nicht mehr in akzeptab-

bler Zeit ausgeführt werden. Dies schränkt die Wahl der Lernverfahren stark ein und ist der Grund dafür, dass im Kontext von Big Data zunächst eher einfache Verfahren zum Einsatz kommen, denen man geringe Anforderungen bezüglich der erforderlichen Berechnungen zuschreibt. Wie bereits feststellt, ist dies eine paradoxe Situation, denn auf der anderen Seite wurde für die Anwendung komplexer Lernverfahren mit potenziell komplexen Analyseergebnissen (Modellen) geäußert, dass mit mehr Daten verlässlichere und genauere Ergebnisse erzielt werden können. Nun liegen Daten in größerer Menge vor und es werden einfache Lernverfahren empfohlen. Ein Ziel ist daher die Verbesserung der Skalierbarkeit für Lernverfahren, damit auch komplexere Verfahren angewendet werden können.

Welche Eigenschaften entscheiden darüber, ob ein Verfahren skalierbar ist? Lernverfahren, die nur eine Iteration benötigen, sind für eine parallele Umsetzung und einen Einsatz bei großen Datenmengen sehr gut geeignet, insbesondere dann, wenn sich die Objekte über ihre Schlüssel gleichmäßig auf die Knoten zur Berechnung verteilen lassen. Naive-Bayes oder k-Nächste-Nachbar-Verfahren sind Beispiele für diesen Fall. Auch mehrere Iterationen stellen kein Problem dar, wenn die erforderliche Kommunikation zwischen den Knoten, also etwa der Austausch von Zwischenergebnissen, von Iteration zu Iteration begrenzt ist. Der klassische k-Means-Algorithmus für die Clusteranalyse zählt zu dieser Gruppe. Viele Iterationen mit einem hohen Kommunikations- und Synchronisationsaufwand lassen sich dagegen weniger gut parallelisieren. Als Beispiel für diesen Fall sei die Support Vector Machine (SVM) genannt. Auch eine sehr schiefe Schlüsselverteilung ist schädlich, da sich dann die Rechenlast nicht gut auf die verschiedenen Knoten verteilen lässt. Für eine ausführlichere Betrachtung dieser Thematik sei der Leser beispielsweise an Beschreibungen der Umsetzung von Lernverfahren im Kontext von Map-Reduce verwiesen (Chu et al. 2006).

Skalierbare Modellanwendung

An verschiedenen Stellen wurde bereits darauf hingewiesen, dass bei Echtzeit-Anwendungen oftmals nur die Modellanwendung in Echtzeit erfolgen muss und nicht die Modellerstellung. Ein bereits bestehendes Modell einschließlich der Beschreibung erforderlicher Datenvorverarbeitungsschritte muss in geeigneter Form bereitgestellt werden, sodass es als Modellinstanz in einem beliebigen Knoten zum Einsatz kommen kann. Bei allen Objekten, auf die ein Modell angewendet werden soll, müssen zunächst die relevanten Datenvorverarbeitungsschritte angewendet werden, um es in die vom Modell erwartete Struktur bzw. Form zu bringen.

Die Anwendung von Prognosemodellen auf neue Objekte, das sogenannte Scoring, wird sehr oft durchgeführt und lässt sich sehr gut automatisieren und parallelisieren, da es keine Abhängigkeiten zu berücksichtigen gilt. Jedes Objekt kann unabhängig von den anderen bewertet werden.

Bei anderen Analyseaufgaben kann die Modellanwendung auch schwieriger zu parallelisieren sein. Als Beispiel sei bei der Verarbeitung von Datenströmen das Complex Event Processing (CEP) genannt. Komplexe Regeln können dabei Zusammenhänge zwischen verschiedenen Objekten (Ereignissen) berücksichtigen. Aufgrund der gegebenen Abhängigkeiten gibt es stärkere Einschränkungen bei der Parallelisierung der Regelverarbeitung.

Big Data-Analytics-Werkzeuge

Es ist nicht ungewöhnlich, dass Standard-Analyse-Software versucht, die relevanten Daten im Hauptspeicher zu halten. Mit der Strategie gelangt man mit großen Datensätzen schnell an die Grenzen des Analysierbaren. Bei großen Datensätzen sollten die Berechnungen dort durchgeführt werden, wo sich die Daten befinden, wie dies auch schon bei In-Database-Analytics und In-Memory-Analytics umgesetzt wird. Einige Software-Projekte, wie beispielsweise *Apache Mahout* (Owen et al. 2012) oder das verteilte Machine-Learning-System *MLbase* im Umfeld von *Apache Spark* (Kraska et al. 2013), Machine-Learning-Bibliotheken wie *MADlib* aber auch spezielle Pakete der Statistik-Software *R* und einige gängige Programmiersprachen wie Python oder Java, ermöglichen dies.

Werkzeuge für Big Data Analytics werden bestrebt sein, die Komplexität, die sich durch die Anforderung nach Skalierbarkeit ergibt, weitgehend zu verstecken. Eine möglichst automatische Parallelisierung und Optimierung der Verarbeitungsschritte wird dabei angestrebt werden. Letztlich sollen deklarative AnalyseSprachen das Umsetzen von Lernverfahren oder die Analysen selbst erleichtern. Auch ein automatisches Evaluieren verschiedener Modelle und Parameter und Vorgabe von zeitlichen Grenzen ist möglich. Wie stark sich der vollständige, generische und interaktive Analyseprozess aber insgesamt durch geeignete Werkzeuge automatisieren lässt, bleibt abzuwarten. Letztlich kann durch eine gute Datenaufbereitung häufig mehr erreicht werden als durch den Versuch eines noch geeigneteren Lernverfahrens zu finden. Die Frage ist also, inwieweit sich die Datenaufbereitung automatisieren lässt oder die Datenaufbereitung eine Kunst bleibt, die den Menschen letztlich als kritischen Erfolgsfaktor benötigt.

2.3.4.3 Der Mensch im Unternehmen als Einflussfaktor

Die größte Hürde bei der erfolgreichen unternehmensweiten Anwendung und Nutzung von Big Data ist weniger im Umfeld von Daten und Technologie zu sehen, sondern in der Unternehmenskultur und im Management (vgl. auch Abschn. 2.1). Außerdem mangelt es oft an Verständnis dafür, wie mithilfe von Analysen Verbesserungen im Unternehmen erzielt werden können (LaValle et al. 2011). Abschließend soll daher kurz erläutert werden, wo und in welcher Weise der Mensch zum kritischen Erfolgsfaktor wird.

Dass der Mensch eine wesentliche Rolle in IT-Systemen spielt oder zumindest spielen sollte, ist keine neue Erkenntnis. Auch Erfahrungen aus dem Bereich BI zeigen, dass Projekte letztlich eher am Faktor Mensch als an unzureichender Technologie scheitern. Eine Ursache dafür ist sicherlich in der fachbereichsübergreifenden und unternehmensweiten Tragweite der meisten BI-Lösungen und deshalb auch der meisten Big Data-Vorhaben zu sehen.

Top-Management-Unterstützung

Wesentliche Erfolgsfaktoren bei BI- und Big Data-Vorhaben sind die Aufmerksamkeit und die Unterstützung durch das Top-Management in einem Unternehmen. Die übergeordneten Analyseziele und die Mehrheit der eingesetzten fortschrittlichen Analysemethoden

sind nicht neu, allerdings fristeten letztere außerhalb des akademischen Umfelds und einiger spezieller Anwendungen in ausgewählten Branchen wie etwa dem Finanzsektor oft mangels Beachtung ein Nischendasein neben den dominant und omnipräsent erscheinenden statischen und dynamischen Berichtsmöglichkeiten vieler BI-Lösungen. Die deutlich stärke Beachtung der technologischen Möglichkeiten und die sich dadurch eröffnenden Einsatzmöglichkeiten unter Berücksichtigung der Unternehmensstrategie können letztlich zu einem notwendigen Wandel in der Unternehmenskultur führen, der für den Erfolg von Big Data-Vorhaben erforderlich ist.

Unterstützung auf anderen Ebenen im Unternehmen

Unterstützung ist auf allen Ebenen eines Unternehmens notwendig, und zwar sowohl in den Fachbereichen als auch in den IT-Bereichen. Gerade in unteren und mittleren Führungsebenen und auf der fachlichen Ebene sind Akzeptanz und Unterstützung oftmals nicht in ausreichendem Maße gegeben. Stattdessen spielen persönliche Befindlichkeiten sowie abteilungs- oder bereichspolitisch motiviertes Verhalten eine zu große Rolle. Auch dort muss eine entsprechende Einstellung zu Gunsten der Unterstützung durch analytische Systeme wie etwa Big Data-Anwendungen etabliert werden. Andernfalls entsteht ein nicht zu unterschätzendes Erfolgsrisiko.

Aber auch auf einer inhaltlichen Ebene fordert Big Data ein Umdenken im Unternehmen, insbesondere was die Ausrichtung von Fachbereichen und IT betrifft. Traditionell sind die Themen Datenmanagement und Analyse-Systeme in IT-Bereichen eines Unternehmens verankert und diese, geleitet von strategischen Vorgaben zur Kostensenkung und spezifische fachliche Anforderungen oft ignorierend, richten den Fokus auf Standardisierung, Stabilität und Skalierbarkeit. Die Exploration und Analyse von Daten im Kontext von Big Data benötigt jedoch Agilität. Eine in den Fachbereichen verankerte Verantwortung für die Analysemöglichkeiten kann dies unterstützen. Langfristig dürfen IT-Systeme nicht länger nur der Automatisierung von Geschäftsprozessen dienen, sondern die Gewinnung von Erkenntnissen muss ein integraler Bestandteil der Systeme sein (Davenport, Barth, and Bean 2012).

Der Engpass Mensch in der Rolle des Data Scientists

Es gibt zwar Analysen, die völlig automatisiert erstellt und angewendet werden können, zumindest wenn sich der Analyseprozess nach anfänglicher Überwachung als robust und zuverlässig gezeigt hat und wenn die Auswirkungen von Fehlentscheidungen zumindest kurzfristig nicht erfolgskritisch sind, in den meisten Fällen jedoch wird der Mensch im Analyseprozess eine wichtige Rolle einnehmen. Verlässliche und vertrauenswürdige Analysen können nicht einfach per Knopfdruck erstellt werden. Vielfach möchten Softwareanbieter ihre Produkte zwar mit diesem Versprechen anpreisen, doch die Realität sieht anders aus. So vielversprechend und angenehm Werkzeuge mit vermeintlich leicht zu erstellenden Analyseergebnissen sind, gibt es doch eine extrem gefährliche Kehrseite der Medaille: Mithilfe dieser Tools ist es auch leichter das Falsche zu tun, wenn entsprechendes Hintergrundwissen über den Analyseprozess und die Analysemethoden fehlt (Franks 2012,

S. 241). Der Analyseprozess ist hochgradig interaktiv und benötigt intensive menschliche Kontrolle. Es mögen sich die Berechnungen im Analyseprozess auch für große Datens Mengen mithilfe geeigneter Big Data-Technologie angemessenen beschleunigen lassen, es bleibt dann jedoch der Mensch mit angemessenen Fähigkeiten der größte Engpass im Analyseprozess (Domingos 2012).

Für die Begleitung des Analyseprozesses bedarf es der Fähigkeiten und Kenntnisse aus den jeweils betroffenen Fachbereichen, aus der Datenanalyse und im Einsatz von Computertechnologie zur Durchführung der notwendigen Verarbeitungsschritte. Zusätzlich wird ein hohes Maß an Kommunikations- und Präsentationskompetenz verlangt. Dies ergibt das Anforderungsprofil des sogenannten Data Scientist, das für Unternehmen eine entscheidende Rolle spielen wird (Davenport, Barth, and Bean 2012). Da die Anzahl hochqualifizierter Mitarbeiter, die diese Kompetenzen in ausreichendem Maße mitbringen, oft nicht sehr groß ist, sollten zumindest geeignete Teams oder Abteilungen in Unternehmen aufgebaut werden, in denen die erforderlichen Kompetenzen gebündelt werden.

2.3.5 Zusammenfassung und Ausblick

Mit den inzwischen verfügbaren Big Data-Technologien ist bereits sehr viel erreicht. Es existieren Lösungen, die die Herausforderungen von Big Data mit seinen großen Datens Mengen, die in großer Geschwindigkeit und Vielfalt auf ein Unternehmen zuströmen, adressieren und zu bändigen helfen.

Einen Nutzen können Unternehmen aus Big Data allerdings nur durch eine geeignete Analyse ziehen. Deshalb ist Big Data Analytics mit seinen zahlreichen Varianten ein entscheidender Baustein zur Schaffung wirtschaftlichen Nutzens.

Da im Kern von Big Data Analytics dieselben analytischen Fragen stehen, die auch schon vor der Big Data-Ära gestellt wurden, sollte von der Erfahrung bezüglich typischer Analyseaufgaben und deren Lösungen im Rahmen eines standardisierten Analyseprozesses profitiert werden.

Dennoch sind einige Dinge zu beachten. Insbesondere Probleme, die sich aus mangelnder Datenqualität und inkonsistenter Datenintegration ergeben, bilden eine zentrale Herausforderung. Auch die Weiterentwicklung skalierbarer Analysemethoden und die Integration in geeigneten Werkzeugen ist ein wichtiger Aspekt. Die größte Hürde zum Erfolg von Big Data-Projekten wird aber letztlich der menschliche Faktor sein.

Der Mensch muss im Kontext von Big Data wesentlich stärker in den Vordergrund gerückt werden. Dies erfordert einerseits die Verfügbarkeit von Mitarbeitern mit geeigneten Fähigkeiten für die Umsetzung von Big Data-Vorhaben und andererseits auch eine angemessene Berücksichtigung und Unterstützung des Menschen in allen relevanten Prozessen in einem Unternehmen. Letztlich muss aber auch die Einstellung der Mitarbeiter und Führungskräfte auf allen Ebenen eines Unternehmens zum Thema Datenanalyse und dem Einsatz der Analyseergebnisse offen sein, aber natürlich nicht frei von kritischer Reaktion.

Ein größerer Mehrwert durch Big Data wird sich insbesondere dann ergeben, wenn die internen Daten eines Unternehmens im Rahmen einer relevanten Fragestellung mit Big Data verknüpft und angereichert werden können. Um eine solide Integration beider Welten zu erreichen, wird eine ganzheitliche, unternehmensweite Datenstrategie erforderlich sein, die sowohl Daten aus traditionellen operativen und analytischen Systemen als auch Big Data-Quellen berücksichtigt (Franks 2012, S 22).

2.4 Simulation: Neue Einsatzfelder durch Big Data

Lothar März

2.4.1 Einführung

Neben der Qualität der Produkte und Leistungen sind es zunehmend Reaktionsfähigkeit und Flexibilität, die darüber entscheiden, wie erfolgreich Unternehmen sich am Markt behaupten. Exzenter Lieferservice ist ein wesentlicher Wettbewerbsvorteil und schafft zufriedene und treue Kunden. Durch die Synchronisation der Marktanforderungen mit der Leistungserbringung wird es möglich, Produktqualität und Lieferservice zu steigern, während Produktionskosten und Working Capital minimiert werden (vgl. hierzu eingehender Abschn. 2.9.2).

Die Zielsetzung in Produktion und Logistik war und ist es, die Effizienz von Prozessen weiter zu steigern und Funktionen zu optimieren. Mit der kurzfristigen Verfügbarkeit und Verarbeitbarkeit von Daten zu Informationen ergeben sich neue Anwendungsfälle für die zeitnahe Planung und Steuerung. Die Informations- und Kommunikationstechnologien sind hierbei der Hebel. Komplexe ERP- bzw. PPS-Systeme mit einer zentralen MRP-Planung kommen allerdings in optimierten, dezentralen Strukturen an ihre Grenzen. Die realen Einflussgrößen in Fertigung und Montage sind in den Planungssystemen zumeist nur teilweise berücksichtigt, sodass sich die Fertigungssteuerung mithilfe von isolierten Einzelanwendungen (zumeist abgebildet in Excel) die notwendige Planungssicherheit selbst erarbeitet. Ansätze zur selbstgesteuerten Produktion, wie beispielsweise Kanban, können helfen, der Komplexität zu begegnen. Dennoch sind spätestens bei Themen wie Verfolgung der Anlageneffizienz, Qualitätsdatenerfassung und -verwaltung oder Rückverfolgbarkeit, Methoden ohne IT-Unterstützung nicht mehr vorstellbar.

Durch die Verteilung von Planungs- und Steuerungsinformationen und -wissen entstehen Medienbrüche zwischen planender und ausführender Ebene. Daten müssen redundant gepflegt werden und die Mitarbeiter sind gezwungen, zeitaufwändige buchungstechnische Korrekturen zu betreiben. Im Zuge der weiter fortschreitenden innerbetrieblichen Integration und der Orientierung an den internen und externen Wertschöpfungsketten entstand der Bedarf an höher integrierten Softwarelösungen, um den steigenden Anforderungen der produzierenden Unternehmen gerecht zu werden. Voraussetzung zur Anwendung weitergehender Planungsansätze ist die Verfügbarkeit von Daten, die

- jederzeit aktuell verfügbar sind,
- semantisch eindeutig und in ein konsistentes Kennzahlensystem eingebunden sind und
- all die zusätzliche Informationen transportieren können, die für die Anwendung von neuen Planungsmethoden notwendig sind.

Die Verarbeitung großer Datenmengen ist die Grundlage für Anwendungen, die zum Ziel haben, zu jedem Zeitpunkt optimale Vorschläge zur Planung und Steuerung zu erstellen. Im Kontext von Analysemethoden stellt die vorgestellte Anwendung ein Verfahren des Advanced Analytics mit Big Data dar. Die Fragestellungen „Was wird passieren“ und „Was ist zu tun?“ sollen mit dem vorgestellten Entscheidungsunterstützungssystem beantwortet werden und ordnen die Applikation in die Verfahren von Predictive sowie Prescriptive Analytics ein (siehe Abschn. 2.3). Im nachfolgenden Beitrag wird am Beispiel der Montageplanung in der Automobilindustrie aufgezeigt, welche Möglichkeiten und Potenziale innovative Planungsansätze bieten, die es verstehen, die verfügbare Datenflut für solcherart Fragestellungen zu nutzen. Dazu wird zunächst die Planungsaufgabe eingegrenzt und die Methodik der simulationsgestützten Planung umrissen. Daran schließt sich die konkrete Beschreibung der Zielsetzung, des Ablaufes und der Datenanforderungen an.

2.4.2 Planungsablauf in der Fahrzeugindustrie

Die Fahrzeugindustrie zeichnet sich durch stark kundenindividuelle und somit variantenreiche Produkte aus. Durch die hohen Lohn- und Lohnnebenkosten in Deutschland und Europa müssen die vorhandenen Ressourcen bei wechselnden Anforderungen so eingeplant werden, dass die Mitarbeiter gleichmäßig ausgelastet werden. Diese variantenreiche Produktion erfordert somit eine sorgfältige Programm- bzw. Sequenzplanung. Die Programmplanung verteilt die zu montierenden Fahrzeuge im Tagesverlauf dergestalt, dass die eingesetzten Mitarbeiter gleichmäßig ausgelastet und Unterauslastungen sowie Kapazitätsspitzen vermieden werden. In der variantenreichen Serienfertigung haben sich sequenzierte Linien als Produktionssystem durchgesetzt, bei denen die Produkte taktgebunden eingesteuert werden (Boysen et al. 2007). Durch die arbeitsintensive Produktionsstruktur und der niedrigen Automatisierung ist die Montage im Vergleich zu Rohbau und Lackierung der Bereich mit dem höchsten Personalaufwand. Damit ist die Montage der Bereich, der neben dem Einkauf die höchsten Einsparpotenziale verspricht (Niederprümm und Sammer 2012). Parallel dazu wächst der Trend zur Verlagerung des Variantenentstehungspunktes in die Montage, um einerseits die Auftragsspezifizierung zu einem späten Zeitpunkt zu ermöglichen und um andererseits in den vorgelagerten Produktionsstufen eine vereinfachte Produktstruktur abbilden zu können. Dennoch bleiben die Forderungen erhalten, die Montage hochflexibel zur Abbildung unterschiedlicher Varianten bei gleichzeitiger Rationalisierung der Prozesse auszulegen (KPMG 2011).

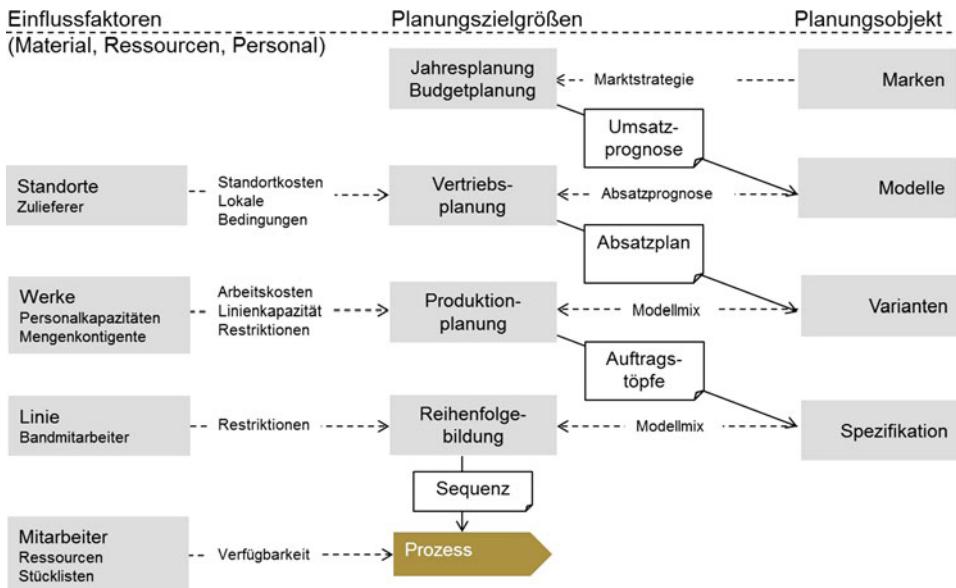


Abb. 2.22 Planungskaskade in der Automobilindustrie

Die Abstimmung von Auftragslast mit den verfügbaren Produktionsfaktoren erfolgt über eine Planungskaskade, an deren Ende die Auftragsreihenfolge steht, mit der die Fahrzeuge in der Montage eingetaktet werden (Abb. 2.22, vgl. Auer et al. 2011). In Abhängigkeit des Planungshorizonts werden in den aufeinander abgestuften Planungsaufgaben die Produktionsfaktoren von Material, Ressourcen und Mitarbeiter (-kapazität) in verschiedenen Aggregationen berücksichtigt; zu Beginn in grober (Volumen-)Abschätzung bis zum physischen Wertschöpfungsprozess in der Montage. Die kapazitiven Grenzen der Produktionsfaktoren werden durch kontextbezogene Restriktionen gezogen. Der Abgleich der Anforderungen mit den Leistungsangebot der Produktionsfaktoren muss ständig auf allen Ebenen erfolgen, um in der Umsetzung keine unlösbar Engpässe bzw. inakzeptable Unterauslastungen zu erhalten.

Die Markenstrategie des Unternehmens, die üblicherweise auf Marktanalysen beruht, bildet den Ausgangspunkt der Planung. Die Marktstrategie führt zu jährlichen Budgetplanungen mit Vertriebsvorhersagen für die nächsten 7 bis 10 Jahre, die rollierend aktualisiert werden. Die Vertriebsplanung konkretisiert die Absatzzahlen anhand von Hauptkriterien, wie beispielsweise Motorisierung, Karosserieformen, Getriebe etc. und weist Produktionsvolumina möglichen Produktionsstandorten zu. Die Entscheidung für oder gegen einen Produktionsstandort fällt zu diesem Zeitpunkt und wird hauptsächlich getrieben von betriebswirtschaftlichen sowie lokalen Bedingungen existierender oder geplanter Standorte und Zulieferer.

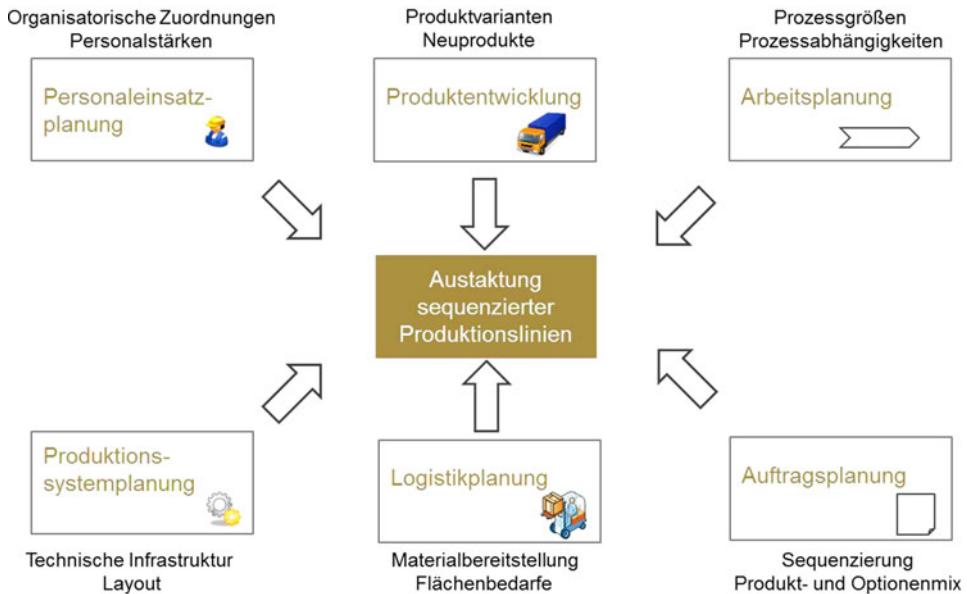


Abb. 2.23 Einflussfaktoren sequenziert produzierender Linien

Die Programmplanung nutzt Absatzprognosen, Einbauraten und in der Jahresplanung vereinbarte monatliche Produktions- bzw. Absatzmengen, um für einzelne Fertigungssperioden über die Art und Menge der herzustellenden Varianten aus dem gegebenen Variantenportfolio zu entscheiden. Dabei muss einerseits der von der Fließbandabstimmung vorgegebene kapazitative Rahmen der Fertigungstakte eingehalten und andererseits die Bauteilverfügbarkeit beachtet werden. Dazu wird der Auftragsbestand auf einzelne Tages- und Schichtprogramme sowie auf die Materialbedarfe herunter gebrochen. Der Abgleich erfolgt anhand der Planungsprozesse des Slotting, Balancing und Sequencing. Die Arbeitsplanung legt die Grundlagen für die Austaktung der Linien, da sie u. a. die Zuordnung von Prozessen zu Mitarbeitern definiert und maßgeblichen Einfluss auf die Harmonisierungsmöglichkeiten (Line Balancing) des Personaleinsatzes hat.

Die Einflussgrößen von sequenzierten Produktionslinien gehen über die reine Personalbetrachtung hinaus. In Abb. 2.23 sind die wichtigsten Einflussfaktoren aufgezeigt, die bei der Austaktung von sequenzierten Linien zu berücksichtigen sind.

Am Beispiel der Personaleinsatzplanung wird im folgenden Kapitel die Komplexität der Planungsaufgabe aufgezeigt und der Einsatz einer simulationsgestützten Anwendung motiviert.

2.4.3 Herausforderungen an die Planung

2.4.3.1 Erhöhung der Planungsgenauigkeit

Die aufgezeigte Planungskaskade mit einer sich anschließenden Sequenzplanung verhindert nicht, dass es immer wieder zu dynamischen Überlastfällen kommt, wenngleich in der Praxis die flexible Mitarbeiterorganisation einen hohen Anteil der Überlastfälle abfängt. Die Sequenzplanung versucht anhand von Fahrzeugkriterien und Reihenfolgeregeln solche Engpässe zu vermeiden. Dies ist oftmals nicht hinreichend, da die Regeln nicht anhand einer Vorschau erstellt werden, sondern aufgrund der Erfahrungen an der Linie gebildet wurden. Erst wenn es in einer Mitarbeitergruppe zu übermäßigen Belastungsspitzen kommt, die sich aufgrund der Folge von Fahrzeugtypen oder Ausstattungsvarianten ergeben, können durch Hinterlegung einer Regel diese Überlastungen zukünftig vermieden werden. Da es mannigfaltige Belastungsspitzen gibt, sind nicht alle möglichen und zu Spitzen führenden Fahrzeugkombinationen zu berücksichtigen, da sonst die Berechnungszeiten zu lang würden oder keine Lösung mehr gefunden würde. Doch selbst bei vollständiger Berücksichtigung der Arbeitsanforderungen je Takt verhindert die Sequenzbildung nicht, dass es immer wieder zu dynamischen Belastungsspitzen in Mitarbeitergruppen kommt. Dies ist darauf zurückzuführen, dass einerseits die Glättung der Belastungen auf Basis der Fahrzeugkriterien über eine Mittelwertbetrachtung erfolgt (statisch). Die Prozessanforderungen werden lediglich summarisch mit den Mitarbeiterkapazitäten verglichen und vernachlässigen Prozessabhängigkeiten, die eine Teilung bzw. parallele Bearbeitung nicht ermöglichen. Zudem werden all diejenigen erhöhten Prozesszeitanforderungen nicht erkannt, die sich erst im Laufe der zukünftigen Sequenzbildung ergeben. Gründe hierfür sind selten auftretende Fahrzeugfolgen oder Verschiebungen im Anteil von Produkttypen oder Ausstattungsvarianten.

Vor dem Hintergrund einer zunehmenden Erhöhung der Produkt- und damit Prozesszeitvarianz in der Montage und einer Verschärfung der Effizienzbestrebungen sind genauere Planungsverfahren gefragt, die eine Analyse, Bewertung und kontinuierliche Anpassung der Kapazitäten an den Lastanforderungen ermöglicht. Deutsche Automobilhersteller investieren kontinuierlich hohe Summen in die Aus- und Weiterbildung der Mitarbeiter, was zu einer im internationalen Wettbewerb überdurchschnittlichen Flexibilität des Produktionspersonals führt. Diese Personalflexibilität wird derzeit zumeist nur als reaktives Hilfsmittel eingesetzt, um das geplante Produktionsprogramm umsetzen zu können. So wird eine kostenintensive Personalflexibilität vorgehalten, statt den Personaleinsatz, die Logistik und das Produktionsprogramm integrativ zu planen und so ein Gesamtoptimum zu erreichen. Denn es gibt noch hohes Potenzial im Einsatz des Personals und der kosteneffizienten Sicherstellung der Materialversorgung. Die hierzu notwendigen Daten liegen in den Planungssystemen vor. In Verbindung mit aktuellen Analyse- und Auswertungsfunktionalitäten können proaktiv die Auswirkungen anstehender Produktionsprogramme bewertet und mit alternativen Konfigurationen verglichen werden. Dabei spielen folgende Aspekte eine Rolle:

Hohe Datenqualität

Im Falle der Auftragsdaten müssen diese korrekt sein. Im Falle der Vorschaudaten ist zu prüfen, welchen Informationsgehalt die Plandaten haben und auf welche Planobjekte sie wirken.

Vollständige und korrekte Abbildung der Wirkzusammenhänge

Die quantitativen Zusammenhänge zwischen den Planungsbereichen sind präzise und in Bezug auf die Einflusskriterien vollständig zu erfassen. Die Ignoranz einer Einflussgröße kann zu fehlerhaften Aussagen führen. Daher ist eine hinreichende Validationsphase zur Überprüfung der Übertragbarkeit der Ergebnisse auf den Realzustand unabdingbar.

Berücksichtigung des zeitlichen Fortschritts

Die dynamischen Zusammenhänge bedingen eine Berechnung der Parameteränderungen unter Berücksichtigung des Zeitfortschritts. Aufgrund der Komplexität der Zusammenhänge in sequenzierten Produktionslinien sind analytische Verfahren hierfür zu aufwendig. Aufgrund des stückgutorientierten Planungsbereiches bietet sich die Problemlösungsmethode der logistischen Simulation an.

Der Einsatz einer simulationsgestützten Anwendung erlaubt die Optimierung hinsichtlich

- Verbesserte Austaktung der Produktionslinien durch die Nutzung eines interaktiven Szenarienplanungssystems zur Vermeidung von Unterauslastungen und Engpässen,
- Optimierung des Ressourceneinsatzes unter Berücksichtigung von Kapazitäten und Restriktionen,
- Konflikterkennung unter Berücksichtigung der dynamischen Abhängigkeiten in Produktion und Logistik,
- Absicherung der logistischen Versorgungsketten durch Parametrierung und Überprüfung der Versorgungsketten,
- Erweiterung der Planungsfunktionalität durch den Vergleich von Szenarien mit veränderten Stellgrößen (Kapazitäten, Bedarfe),
- Erhöhen der Reaktionsfähigkeit bei Änderungen durch schnell verfügbare Ergebnisdaten und -reports,
- Einbindung in operative Planungsprozesse durch definierte Datenübernahmen z. B. anhand von ERP- bzw. PPS-Schnittstellen.

2.4.3.2 Einsatz der Simulation in der Planung

Mit der Simulation ist es möglich, die Auslastungsverläufe des Personals je Fahrzeug und Station zu prognostizieren und somit Belastungsspitzen auf den Taktzeitpunkt vorherzusagen. Zur Abbildung der Auslastungen mittels Simulation boten sich den Unternehmen bislang nur zwei Möglichkeiten: Aufbau eigener Simulationsexpertise oder Einkauf von Simulations-Know-how. Aufgrund der Kostenstruktur entscheiden sich viele Unternehmen erst zum Einsatz der Simulation, wenn kontinuierliche Aufgaben anstehen oder bei



Abb. 2.24 Hohes Potenzial an simulationsgestützter Planung

Planungsaufgaben, die einen hohen Einflussgrad auf die laufenden Kosten bzw. Investitionskosten haben. Bei einem Einflussgrad der Simulation auf die Kosten von ca. 10 % (VDI 3633 Blatt 1 2010), ist der Einsatz der Simulation im Einzelfall oder für seltene Anwendungen zu teuer. Das Potenzial für Anwendungen im Bereich der taktischen und strategischen Planung ist demgegenüber groß.

Mit der Einführung von Systemen zur simulationsbasierten Planung und Steuerung in Echtzeit (März et al. 2012) eröffnen sich dem Unternehmen neue Möglichkeiten zur kontinuierlichen Auslegung der Produktionsfaktoren am optimalen Betriebspunkt. Der Kostenvorteil im Vergleich zur Etablierung eines Teams von Simulationsexperten liegt nicht nur in der günstigeren Investition, sondern auch in den deutlich niedrigeren Kosten je Simulationsstudie. So entfallen die Kosten von Simulationsexperten, die bei einer Anstellung im Unternehmen eine hinreichend große Anzahl an Studien durchführen müssten (mind. 5 Studien pro Jahr und Person), um einerseits das Know-how der Modellerstellung aufrecht zu erhalten und andererseits dauerhaft ausgelastet zu sein. Da diese Personalkosten entfallen und lediglich die Personalaufwendungen des Planers anfallen, sind deutlich geringere Kosten je Simulationsstudie möglich. Damit erhält der Anwender ein Werkzeug, welches eine kontinuierliche, simulationsgestützte Planung im operativen, taktischen und strategischen Planungsbereich erlaubt. Damit kann eine hierarchische Planung etabliert werden, die Maßnahmen in Abhängigkeit des Planungshorizonts anhand von Szenarienanalysen bewertet und das Unternehmen bei seiner Aufgabe unterstützt, seinen Resourceneinsatz und die Produktionslogistik optimal einzustellen. Unter dem Begriff der Echtzeit-Anwendung wird verstanden, dass die Latenzzeit zur Datenermittlung, Datenanalyse und -auswertung sowie Entscheidungsumsetzung gering ist; üblicherweise genügt die Anforderung, dass eine Entscheidung im Sinne der Verfolgung der Unternehmensziele rechtzeitig getroffen werden kann (siehe Abschn. 2.3.1.2).

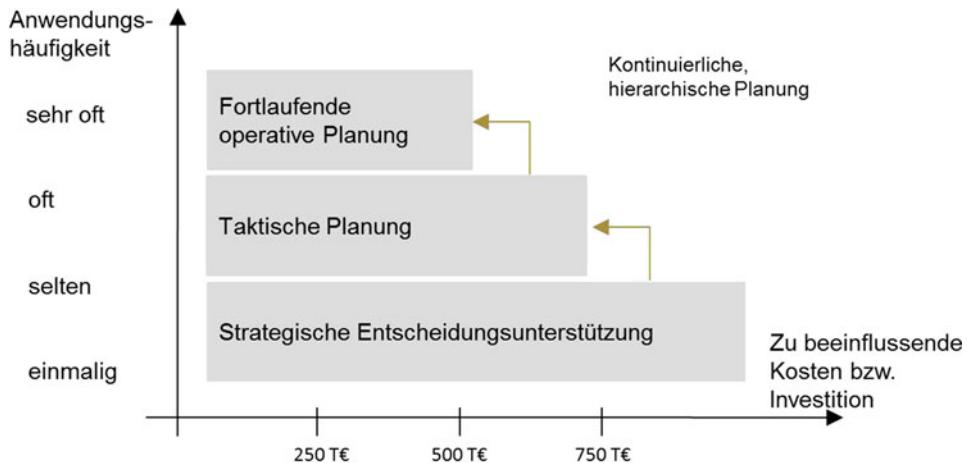


Abb. 2.25 Kontinuierliche, hierarchische Planung

2.4.3.3 Simulationsgestützte Planung

Die Planungsunterstützung durch eine simulationsgestützte Anwendung zur selbstständigen Durchführung von Simulationsstudien durch den Planer setzt hohe Ansprüche an die Ergonomie in der Bedien- und Auswertelogik voraus. Eine enorme Robustheit ist gegenüber beliebigen Eingaben der Anwender gefordert, um einerseits die Anwendung mit hohen Freiheitsgraden zu versehen und um andererseits bei Plausibilitätskonflikten eine konstruktive Rückmeldung an den Planer zu geben. Dies setzt eine Architektur voraus, die eine kontrollierbare Verarbeitung von Eingangs-, Berechnungs- und Ausgangsdaten ermöglicht. Diese Logikbausteine müssen jederzeit erweiterbar und hinsichtlich der Kundenankopplungspunkte konfigurierbar gestaltet werden. Insgesamt lassen sich folgende thematisch zusammengefasste Anforderungen an die Einführung einer simulationsgestützten Planung in den operativen Planungsprozessen unterscheiden.

Bedienlogik Oberfläche

Die Bedienlogik umfasst die Mensch-Maschine-Schnittstelle und enthält Funktionen zur Eingabe, Steuerung und Ausgabe der Anwendung.

- Navigation
- Einstellung von Parametern
- Bearbeitungsfunktionalitäten zur Eingabe
- Ansteuerung des Moduls zur Berechnung durch Simulation
- Ansteuerung von Reporting-Funktionalitäten (Export)

Integrationslogik

Die Wirkzusammenhänge zwischen den Objekten von Ressourcen, Produkte und Prozesse im Allgemeinen sowie im Speziellen die Wirkzusammenhänge in sequenzierten Produk-

tionslinien bedingen Abhängigkeiten und Wechselwirkungen, die im Modell abzubilden sind. Die gegenseitigen Restriktionen und Randbedingungen sind im Falle von Eingaben bzw. Konfigurationen zu berücksichtigen und ggfs. für kundenspezifische Anpassungen parametrisierbar zu halten.

Die Integrationslogik greift insbesondere im Zusammenspiel mit der Funktionalität, im Modus der Ergebnisanalyse Prozessbausteine editieren zu wollen und diese Prozesse in Form alternativer Eingangsdatensätze für weitere Simulationen nutzen zu wollen. Der Wechsel von Ergebnis- zu Eingangsdatum erfordert die Prüfung auf Plausibilität und Machbarkeit, ermöglicht jedoch im Gegenzug weitreichende Konfigurationsmöglichkeiten in der interaktiven Austaktung von Produktionslinien. Dazu sind neben der Überprüfung auf Restriktionen und Wirkzusammenhänge auch Informationen über die Tragweite der Änderungen, z. B. Auswirkungen auf die Anzahl an Produkten, bereitzustellen.

Berechnung durch Simulation

Die Berechnung der Auslastungen der Mitarbeiter in sequenzierten Produktionslinien erfolgt anhand einer ereignis-diskreten Simulation. Die Simulationslogik ist so auszulegen, dass die unterschiedlichen Ausprägungen einer Linie hinsichtlich Produktbewegung und -pufferung, Linienstruktur, Personalorganisation, Prozessabhängigkeiten etc. berücksichtigt werden können. Die dazu notwendigen Simulationsdaten werden identifiziert und als Grundlage zur Simulation aus den Gesamtdaten extrahiert. Im Zusammenhang mit der Logik zur Verteilung von Simulationsläufen auf verteilte Rechner kann dem Ziel nach schnellen Simulationsläufen Rechnung getragen werden. Die Verteilung von Belehrungsaufgaben auf mehrere Prozessoren und die Skalierbarkeit der Anwendung ist aus technologischer Sicht ein Schlüssel zum Erfolg (Abschn. 2.3.4.2).

Auswertelogik

Die Auswertefunktionalität der Ergebnisse muss eine Reihe von Diagrammen und Tabellen umfassen und eine Vielzahl von Analysen ermöglichen. Die Struktur und die Anordnung der Auswertediagramme entspringen der Überlegung, dass die Auswertung der durch die Produkte verursachten Auslastungen in Bezug auf die Zyklen und die Organisationseinheiten erfolgen muss. Darüber hinaus stehen Diagramme zur Analyse der Prozesszeitgrößenverteilungen sowie Variantenspreizungen der Produktanforderungen zur Verfügung. Filterfunktionen über Produkte, Produkttypen, Ausstattungsmerkmale und Zyklen ermöglichen die gezielte Auswertung von produktspezifischen Ausprägungen und Prozesszeitanforderungen.

2.4.3.4 Erhöhte Datenanforderungen

Der Übernahme, Handhabung und der Speicherung der Daten kommt eine zentrale Bedeutung in der Anwendung zu. Die Daten für eine simulationsgestützte Anwendung setzen sich aus den Stamm-, Plan- und Bewegungsdaten zusammen, die aus den proprietären ERP-Systemen entnommen werden. Aufgrund der zu erwartenden hohen Datenvolumina ist ein besonderes Augenmerk auf die Schnelligkeit der Datenübernahme und der

Verfügbarkeit der Daten zu legen. Grundsätzlich lassen sich folgende Aspekte bei der Manipulation und Ablage von Daten unterscheiden.

Übernahme von Daten

Darunter werden Aufgaben der Extraktion, Transformation, Integration, Prüfung, Validierung, Bereinigung und Transport von Daten adressiert. Häufig treten hierbei Prozesse mit ähnlichen Aufgabenstellungen auf, deren programmtechnische Umsetzung durch Adaptation von Musterlösungen (Patterns) möglich ist.

Die übernommenen Daten sind für eine Übernahme und Interpretation durch den Simulator aufzubereiten. Im Falle von gängigen ERP-Systemen wie beispielsweise SAP erleichtern Übernahmeprotokolle in Form von definierten Schnittstellen die Anbindung. Bei der Erstellung dieser Extraktions-, Transformations- und Ladeprozesse (ETL) sind auch Aspekte der Bereitstellung (Deployment) und des Betriebs zu berücksichtigen.

Notwendige Daten zur Simulation

Für die Durchführung von Simulationen ist eine konsistente Verwaltung inklusive Plausibilitätsprüfungen von Eingangs-, Simulations- und Ergebnisdaten notwendig. Die Eingangsdaten fassen alle zur Abbildung des Modells und der Ablauflogik notwendige Informationen zusammen. Die Simulationsdaten definieren einen Simulationslaufs, so z. B. die Laufzeitlänge und den Abbildungsrahmen der Simulation.

Daten zur Auswertung

Die Sammlung und Aufbereitung relevanter Daten erfolgt im Data Warehouse und dient zum performanten Zugriff für die Erstellung von Analysen und für das Reporting. Im Gegensatz zu den Eingangsdaten der Simulation, die redundanzfrei und für die Verarbeitung im Simulator stringent auf wenige Informationen reduziert werden, ist der Ansatz im Data Warehouse entgegengesetzt: Hier sind die Daten für die unterschiedlichen Berichte und Auswertungen bereits datentechnisch vorzubereiten, um im Falle einer Abfrage kurzfristig auf die relevanten und vorkonfigurierten Datentabellen zugreifen zu können.

2.4.4 Praxisbeispiel Automobilendmontage

2.4.4.1 Zielsetzung der Anwendung

Am konkreten Beispiel einer variantenreichen Serienfertigung soll nachfolgend aufgezeigt werden, wie durch die Nutzung relevanter Daten in Kombination mit Simulationsalgorithmen neue Anwendungsfelder eröffnet werden. Das Prinzip der variantenreichen Serienfertigung ist heute Grundlage vieler Industriezweige. Eine Ursache für die Ausweitung des Variantenangebotes ist, dass sich die Unternehmen durch ein möglichst umfangreiches Variantenangebot von der Konkurrenz absetzen wollen, um auch in stagnierenden Märkten neue Kunden zu gewinnen. Darüber hinaus führt die Internationalisierung des

Absatzes dazu, dass die Produkte an die technischen, kulturellen und rechtlichen Rahmenbedingungen im Ausland angepasst werden müssen. Ein weiterer Aspekt ist, dass sich die Individualisierung der Produkte auch durch den Ansatz des Mass Customization auf Marktbereiche ausgeweitet hat, die bisher der klassischen Massenproduktion vorbehalten waren. Daher wird hier auch von einer Variantenfertigung gesprochen. Obwohl flexible Produktionstechnologien sowie leistungsfähige Planungs- und Steuerungssysteme zur Verfügung stehen, führt die Angebotsausweitung zu einer Komplexitätserhöhung an vielen Stellen des Produktionsprozesses, die eine effiziente Produktion erschwert. In der Montage, üblicherweise der letzten Stufe der Produktion, herrscht in der Variantenproduktion in der Regel die Losgröße 1 vor. Dies liegt darin begründet, dass aufgrund der hohen Variantenzahl und Kundenindividualität eine Produktion auf Lager nicht praktikabel oder aufgrund der hohen Kapitalbindung in Fertigprodukten unerwünscht ist. Eine Möglichkeit, bei hohen Stückzahlen eine effiziente Montage zu ermöglichen, ist die Verwendung des Fließproduktionsprinzips.

In einer solchen Fließmontage, in der nur Produkte hergestellt werden, für die ein expliziter Kundenauftrag vorliegt, treten dabei aufgrund der Variantenvielfalt verschiedene Probleme auf. Dazu gehören vor allem die Materialbereitstellung der Variantenteile und die schwankende Kapazitätsauslastung der Montagestationen. Während für die Materialbereitstellung leistungsfähige Logistikkonzepte zur Verfügung stehen, ist die Ausnutzung der Produktionskapazität weiterhin ein unzureichend gelöstes Problem. Durch die unterschiedlichen Bearbeitungszeiten für die Produkte an den einzelnen Arbeitsstationen kommt es zu Über- und Unterauslastungen der Werker. Die Überlastungen können zu schlechterer Arbeitsausführung und damit zu Qualitätsproblemen führen oder einen Unterstützer-(Springer-)Einsatz notwendig machen. Im Fall von Unterauslastungen kommt es dagegen zu Wartezeiten. Die durch Belastungsschwankungen verursachten Kosten in Fließproduktionslinien können auch unter dem Begriff der Modell-Mix-Verluste zusammengefasst werden.

Der Planung von sequenzierten Produktionslinien kommt daher eine hohe Bedeutung zu. Sie muss vor dem Hintergrund unscharfer Randbedingungen (Auftragsmix) eine Produktion ermöglichen, die zu jedem Zeitpunkt effizient und ohne Störungen abläuft. Dabei hat die Planung eine Reihe von Randbedingungen und Restriktionen zu berücksichtigen, die sich aufgrund technischer, technologischer, räumlicher oder prozessbedingter Ursachen ergeben. Unterschiedliche Planungsfelder haben Auswirkungen in der Auslegung der Produktionslinie und müssen teilweise gleichzeitig geplant werden. Eine Änderung beispielsweise in der Zuordnung von Prozessen zu einer Station hat Auswirkung auf die Bereitstellung der für diesen Prozess notwendigen Montageteile.

Durch die Bereitstellung der simulationsbasierten Anwendung soll beispielsweise den Planern aus verschiedenen Fachabteilungen die Möglichkeit eröffnet werden, die bestehende Struktur der Produktionslinie sowie alternative Konfigurationen zu analysieren, zu bewerten und zu gestalten. Die Anwendung berücksichtigt alle relevanten Einflussgrößen und bildet das dynamische Ablaufverhalten durch ereignisdiskrete Simulation ab. Das Ziel der Simulation ist es, die Prozesszeitanforderungen der anliegenden Sequenz sowie die

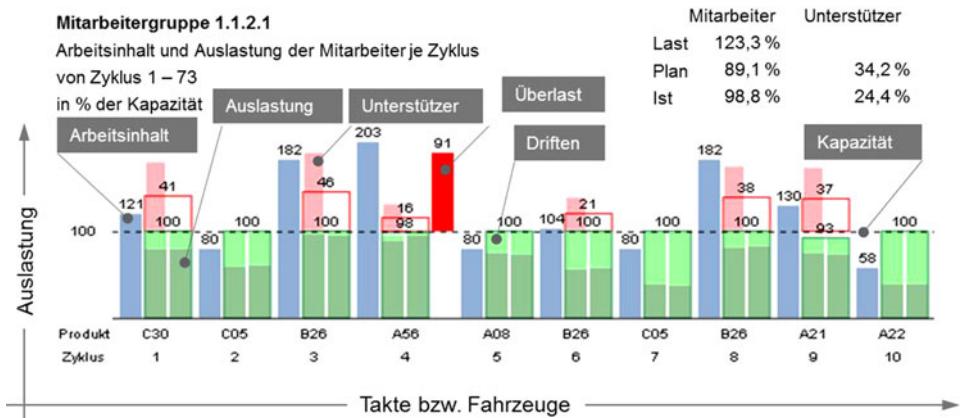


Abb. 2.26 Auslastungsdiagramm Personaleinsatz

Auslastungen der Mitarbeiter an der Linie mit hoher Genauigkeit vorherzusagen, um den Personaleinsatz von Mitarbeiter und Springer vorherzusagen. In Abb. 2.26 ist eine Gegenüberstellung der Bedarfsanforderungen je Takt mit dem Kapazitätsangebot (Mitarbeiter) beschrieben.

Damit kann bereits im Vorfeld der Montagetätigkeiten der Bedarf an Kapazitäten erkannt und Maßnahmen zur bedarfsgerechten Bereitstellung von Mitarbeitern getroffen werden. Der Einsatz zur Personaleinsatzplanung ist sowohl im Tagesgeschäft (Schicht) als auch mittelfristig (Wochenvorschau) und strategisch angedacht.

2.4.4.2 Ablauf einer Anwendung

Um eine Produktionslinie durch iterative Veränderung von Stellparametern (Anzahl Mitarbeiter, Zykluszeit, Zuordnung von Prozesszeitbausteinen zu Mitarbeitergruppen, etc.) zu optimieren, werden die Ergebnisse analysiert und zielgerichtete Veränderungen an den Eingangsdaten vorgenommen. Dazu unterstützt die Anwendungsplattform den Planer in jeder Phase:

- Bei der Eingabe durch kontextbezogene Informationen,
- bei den Simulationsläufen durch die Parallelisierung und schichtübergreifende Abbildung von Produktionsszenarien und
- bei der Ergebnisanalyse. Hier kann der Planer in den Ergebnisdiagrammen die Planungsobjekte (entspricht den Stellparametern) anwählen und eine neue Eingangskonfiguration vornehmen. So lassen sich beispielsweise bei der Analyse der Prozesszeitzuordnungen einzelne Prozesse auswählen und per Drag and Drop anderen Mitarbeitergruppen zuordnen. Dabei können Wirkzusammenhänge von z. B. Vorranggraphen oder technischen Restriktionen berücksichtigt und dem Planer mitgeteilt werden.

Simulationsstudien Integrierte Produktionsprogrammplanung- und Personaleinsatzplanung						
Studien		Stellgrößen				Ergebnisse
		Produktionsprogramm (Anzahl & Typ Fahrzeuge, Reihenfolge)	Personalorganisation (Zuordnung zu Teams, Teamstärke)	Prozessplanung (Zuordnung Prozesse zu Stationen & Teams, Prozesstellung)	Taktzeit	
1	Machbarkeit	variabel	fix	fix	fix	kurzfristige Vorschau der Kapazitätsauslastungen zur Disposition des Sprungereinsatzes
2	Personaleinsatz	fix (repräsentative Belastungsszenarien)	variabel	fix	fix	Analyse von Personaleinsatzszenarien zur Identifikation der optimalen Personalzuordnung
3	Prozessplanung	fix (repräsentative Belastungsszenarien)	fix	variabel	fix	Analyse der Auswirkungen von Prozessplanänderungen auf die Personalauslastungen
4	Produktivität	fix	fix	fix	variabel	Überprüfen der Auswirkungen von Produktivitätserhöhungen bzw. - reduzierungen auf das Personal

Abb. 2.27 Simulationsstudien und Stellgrößen

Die Visualisierungen zur Konfiguration, Animation und Ergebnisanalyse, ergo der Ein- und Ausgabefunktionalitäten, verschmelzen miteinander und werden für den Planer nicht mehr unterschieden; der Planer erhält ein integratives Planungssystem zur interaktiven Austaktung seiner Produktionslinien.

Das Ziel des Planers ist es, Hinweise und Ansatzpunkte zur weitergehenden Glättung der Belastungsverläufe zu erhalten. Hierzu sind die Stellgrößen zu identifizieren, die unter Berücksichtigung wechselnder Produktionsprogramme den größten Hebel auf eine Glättung der Arbeitsinhalte an den Stationen haben (Abb. 2.27).

Aufgrund der Vielzahl an Einflussgrößen und Ausprägungen ist eine Planungslogik notwendig, die den Wirkzusammenhängen der Planungsobjekte sowie die Wechselbeziehungen zur Laufzeit einer Simulation Rechnung trägt. In Bezug auf sequenzierte Produktionslinien sind vielfältige Beziehungen zwischen den Planungsfeldern auszumachen (siehe Abb. 2.23).

Der Ablauf einer Anwendung durchläuft die Phasen Modell, Szenario, Berechnung und Auswertung. Das Modell setzt sich aus den Stamm- und den operativen Daten (Sequenz-, Prozess- und Attributdaten) zusammen. Mit der Festlegung eines Szenarios durch Auswahl von Stammdaten und operativen Daten sowie der abzubildenden Zeitperiode wird ein Simulationsszenario definiert. Die Berechnung des Personaleinsatzes erfolgt durch die Simulation. Im Anschluss daran stehen unterschiedliche Analysefunktionen zur Interpretation der Ergebnisse zur Verfügung (vgl. Abb. 2.28).

2.4.4.3 Datenanforderungen

Die vollständige Erfassung aller Einflussgrößen und ihrer Wirkzusammenhänge ist die Voraussetzung, um einerseits entscheidungsrelevante Ergebnisse der dynamischen Austaktung zu erhalten, und andererseits, um Änderungen interaktiv und ohne Programmier-

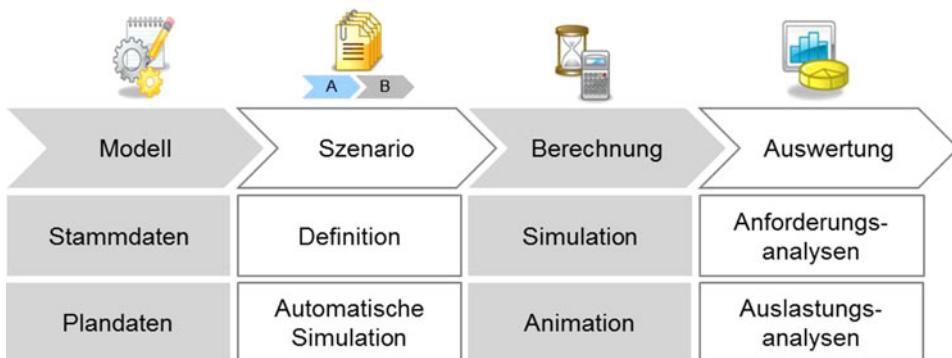


Abb. 2.28 Ablauf einer Simulationsstudie

Abb. 2.29 Stamm- und Plandaten zur Personaleinsatzplanung

Beschreibung der Modelldaten		
Planung	Stammdaten	Plandaten
Auftragslast	Aufträge	Auftragszuordnung
	Prozesse	Sequenzplanung
Produktionssystem	Ressourcen	Ressourcenplanung
	Stationen	Arbeitsplanung

tätigkeiten vornehmen zu können. Die Einflussgrößen entsprechen Stamm- und Planungsdaten, die Planungsobjekten zugeordnet sind.

In Abb. 2.29 sind die Stamm- und Planungsdaten der Planungsbereiche dargestellt, die für die Simulation des Personaleinsatzes erhoben werden. Ein Auftragslastszenario definiert sich aus jeweils einem Stammdatensatz Aufträge und Prozesse sowie den Planungsdatensätzen zur Auftragszuordnung und Sequenzplanung. Analog hierzu wird das Produktionssystem durch die Stammdatensätze zu den Ressourcen und Stationen sowie den Informationen zur Ressourcen- und Arbeitsplanung definiert.

Zeitgemäße Planungssysteme weisen eine Client-Server-Architektur aus, die einen ubiquitären Zugriff auf die Anwendungsfunktionalität und den Auswertungen durch jeden berechtigten Anwender in Echtzeit erlaubt. Dazu ist eine Systemarchitektur zu schaffen, die unterschiedliche Fragestellungen unabhängig voneinander und jederzeit ermöglicht. Dies setzt einerseits eine verteilte Anwendung sowie eine extrem leistungsfähige Simulation im Sinne von Laufzeitverhalten voraus. In Abb. 2.30 ist die grundlegende Anwendungsarchitektur zur Online-Leistungssteuerung aufgezeigt, die von unterschiedlichen Anwendergruppen zeitgleich verwendet werden kann.

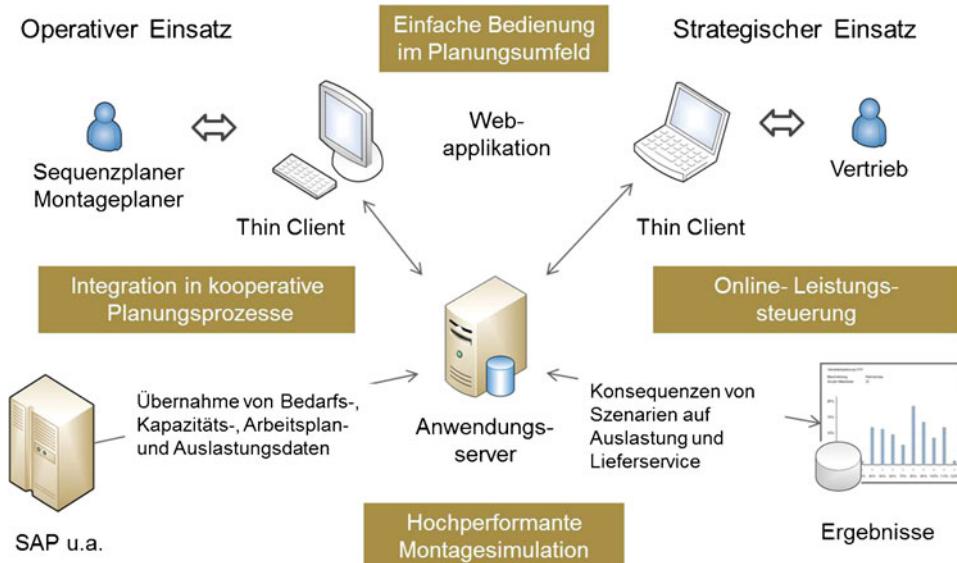


Abb. 2.30 Anwendungsarchitektur zur Online-Leistungssteuerung

2.4.5 Fazit und Ausblick

Mit Big Data werden neben großen Datenmengen auch oftmals die Technologien umrissen, die es erlauben, diese Datenmengen zu sammeln und auszuwerten. Die sinnvolle Aggregation der Daten zu Informationen, aus denen der Anwender Rückschlüsse ziehen kann, ob aufgrund der vorliegenden Datenlage Handlungen abzuleiten sind oder eben nicht, ist eine Kernfunktionalität, die mit Big Data Anwendungen verfolgt werden. In dem beschriebenen Kontext zur Planung und Steuerung einer Endmontagelinie käme es den Informationen gleich, die aufzeigen, wann und in welchem Maße Engpässe respektive Unterauslastungen an der Linie entstehen. Der Beitrag hat darüber hinaus aufgezeigt, dass unter Zuhilfenahme der logistischen Simulation eine weitergehende Nutzung der Daten neue Anwendungsfelder eröffnet, die mit den Technologien zur schnellen Erhebung und Verarbeitung großer Datenmengen ermöglicht werden. Die aktuell abgerufenen Ist- und Plandaten werden in dynamischen Modellen mit prozessrelevanten Informationen angereichert und simuliert. Damit werden aus Big Data weitere Daten produziert, die für die Entscheidungsunterstützung in der Produktionsplanung herangezogen werden können: Smart Data.

2.5 Big Data-Analysen: Anwendungsszenarien und Trends

Fouad Omri

Daten werden als ein neuer Produktionsfaktor der Wirtschaft gesehen. Daten sind ein Rohstoff, der durch Analysen veredelt und als Baustein für Geschäftsentscheidungen bereitgestellt wird. Big Data-Analysen können erheblichen wirtschaftlichen und strategischen Mehrwert für das Unternehmen generieren. Der Einsatz von Big Data-Analysen führt nicht nur zur Verbesserung der betrieblichen Effizienz verschiedener Funktionsbereiche eines Unternehmens, sondern kann auch Möglichkeiten schaffen Produkt- und Serviceangebote zu erweitern.

Dieses Kapitel widmet sich der Verbreitung konkreter Anwendungsszenarien für Big Data-Analysen. Obwohl viele Anwendungsbeispiele für Big Data-Analysen bekannt sind, fehlen den Unternehmen häufig überzeugende Anwendungsszenarien, um Big Data-Analysen wirtschaftlich nutzbringend einzusetzen. Der Nutzen von Big Data-Analysen lässt sich in einigen Funktionsbereichen besonders belegen. Dieses Kapitel präsentiert Anwendungsszenarien für Big Data-Analysen basierend auf sieben ausgewählten Funktionsbereichen: Marketing und Vertrieb, Forschung und Entwicklung, Risikomanagement, IT, Produktion, Logistik, und Kundenservice. Einige dieser Bereiche werden in den folgenden Kapiteln näher betrachtet.

Die Anwendungsszenarien für Big Data-Analysen lassen sich grob in Klassifikations- und Regressions-Analysen unterteilen. Auf der einen Seite zielen Klassifikationsanalysen auf den Ist-Zustand, um die Transparenz der Entscheidungsfindung zu erhöhen und damit

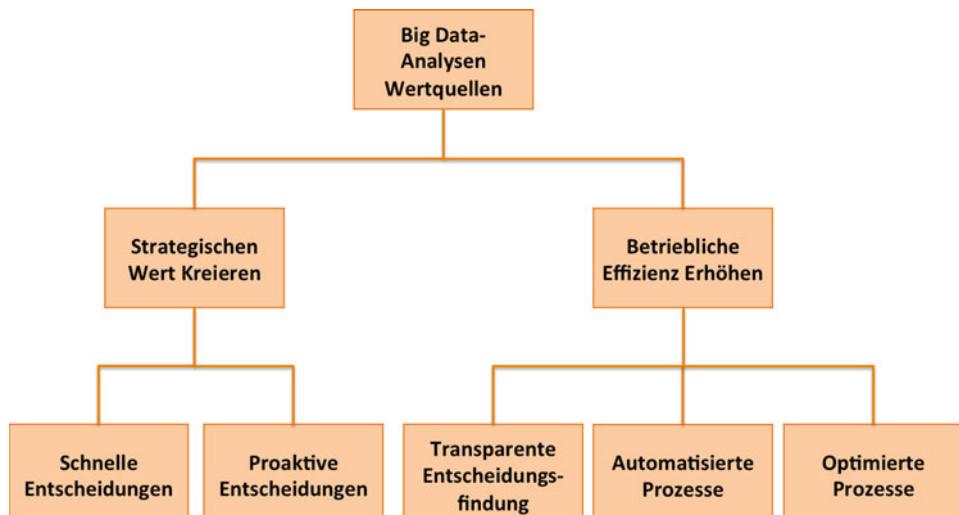


Abb. 2.31 Betriebliche Wertquellen von Big Data-Analysen

	Klassifikationsanalysen	Regressionsanalysen
Ziel	<ul style="list-style-type: none"> ○ Transparente Sicht auf den Geschäftsverlauf ○ Identifikation von Performanz-Treibern ○ Effiziente Prozesse 	<ul style="list-style-type: none"> ○ Komplexe wirtschaftliche Zusammenhänge vorhersagen ○ Proaktive Entscheidung ○ Geschäftsrisiken vorhersagen
Business Fragen	<ul style="list-style-type: none"> ○ Was ist geschehen? ○ Warum ist das geschehen? ○ Wo ist das Problem? ○ Wie viel, wie oft, wo? 	<ul style="list-style-type: none"> ○ Was kann am besten passieren? ○ Was wäre der Worst-Case? ○ Was würde passieren, wenn...? ○ Was wird als nächstes passieren?

Abb. 2.32 Klassifikations- vs. Regressions-Analysen

die Effizienz von Unternehmensprozessen zu steigern. Auf der anderen Seite sind Regressionsanalysen auf die Zukunft gerichtet, um die Wirkung von Entscheidungen vorherzusagen und Innovationen sowie zukünftige Leistungstreiber ausfindig zu machen (näher zu den Möglichkeiten der Analyse Abschn. 2.3).

2.5.1 Big Data-Analysen: Anwendungsszenarien

Im Folgenden wird nicht gesondert auf Business Intelligence (BI) Anwendungsszenarien eingegangen. Der Grund ist, dass die BI Analysen als Spezialisierung der Big Data-Analysen gesehen werden können. Der Hauptunterschied zwischen BI Analysen und Big Data-Analysen ist die Ausrichtung auf die Daten, die analysiert werden sollen. BI Analysen setzen strukturierte und konsistente Daten voraus, wohingegen Big Data-Analysen auf unstrukturierte und möglicherweise inkonsistente Daten optimiert sind (s. hierzu auch Abschn. 2.3).

2.5.1.1 Marketing und Vertrieb

Die häufigsten Anwendungsszenarien im Marketing und Vertrieb zielen auf die Erhöhung der Transparenz und Effizienz von Produkt und Service-Angeboten.

Big Data-Analysen im Marketing und Vertrieb ermöglichen eine individuelle und abgestimmte Kundenansprache. Produkt- und Service-Angebote können auf Kundensegmente oder einzelne Kunden zugeschnitten werden, um Streuverluste im Marketing zu verringern. Der Erfolg von Marketing- und Vertrieb-Entscheidungen sowie Werbekampagnen kann gemessen werden: wie und ob der Umsatz sich erhöht, wenn bestimmte Maßnahmen getroffen werden.

Mit Hilfe von Big Data-Analysen eröffnen sich Cross- und Up-Selling Potenziale. Kundendaten wie Transaktionen, demographische Daten und Standortdaten können analysiert werden, um abgestimmte Angebote zur richtigen Zeit dem richtigen Kunden zu unterbreiten. Kaufmuster können aus den Daten identifiziert werden, um eine granulare Kundensegmentierung zu erreichen und um die Werbekampagnen möglichst individuell

Klassifikationsanalysen	Regressionsanalysen
<ul style="list-style-type: none"> ○ Granulare Kundensegmentierung nach Wert/Potenzial ○ Effiziente Marketingkampagnen ○ Bewertung der Kunden-Zufriedenheit/Beschwerden ○ Effiziente Vertriebsplanung ○ Preisgestaltungen ○ Optimierung von Angeboten ○ Wettbewerbsanalyse 	<ul style="list-style-type: none"> ○ Kundenverhalten vorhersagen ○ Cross- und Up-Selling-Potenzialen identifizieren ○ Marktrends Vorhersagen ○ Preisgestaltungen ○ Optimierung von Angeboten ○ Optimierung von Marketingkampagnen

Abb. 2.33 Anwendungsszenarien in Marketing und Vertrieb

und mit weniger Streuverlusten als bisher zu gestalten. Kundenabwanderungen können vorhergesagt werden, um proaktiv Gegenmaßnahmen einzuleiten (z. B. Rabattpakete). Die Analyse von Kundendaten ist immer im Spannungsfeld zwischen Datengewinnung und dem Datenschutz zu verstehen.

Auch Unternehmens-, Produkt- und Marktstrategien lassen sich mit Big Data-Analysen deutlich verbessern. Social-Media-Inhalte aus Blogs, Foren oder Facebook sowie Informationen aus Lokal- und Fachpresse oder auch Ergebnisse von Suchmaschinen können ausgewertet werden. Aus der Analyse der unstrukturierten und strukturierten Daten entstehen Berichte über Wettbewerber und Märkte, die genauer und aktueller sind als gewöhnliche Berichte.

2.5.1.2 Forschung und Entwicklung

Big Data-Analysen können überwiegend für innovationszwecke eingesetzt werden. Neue Trends und neue Produktideen können frühzeitig durch die Analyse von Patendatenbanken erkannt werden. Nutzermeinungen aus Social-Media-Kanälen oder Foren fließen werden bei der Neuentwicklung und Verbesserung von Produkten und Dienstleistungen aufgenommen. Die Auswertung solcher Daten ermöglicht auch die Auswertung der Markenwahrnehmung.

Die Analyse und die Aggregation von Messdaten aus verschiedenen Quellen (z. B. Forschungseinheiten, Sensordaten, usw.) führen zu genaueren Erkenntnissen über die Marktreife und Qualität des Produktes, verringern die Forschungs- und Entwicklungs-Kosten und verkürzen die Time-to-Market.

Klassifikationsanalyse	Regressionsanalyse
<ul style="list-style-type: none"> ○ Analyse von Messdaten für Test Zwecken ○ Markenwahrnehmung analysieren ○ Produkten/Verfahren verbessern ○ Neue Produktideen ○ Kundenbedürfnissen identifizieren 	<ul style="list-style-type: none"> ○ Qualitätsvorhersagen ○ Marktreife vorhersagen ○ Strategische Entwicklungsziele planen ○ Neue Produktideen

Abb. 2.34 Anwendungsszenarien in der Forschung und Entwicklung

2.5.1.3 Kundenservice

Produktangebote werden immer komplexer und Innovationszyklen werden immer kürzer. Servicekräfte werden schnell überfordert Kundenanfragen effizient zu beantworten. Die Bereitstellung von schnellen Datenanalysen, aggregiert über verschiedene Datenquellen, ermöglicht den Service-Mitarbeitern schnell eine passende Antwort auf konkrete Probleme zu finden. Durch die Auswertung von Serviceberichten können Experten lokalisiert werden, die bei besonders schwierigen Problemen Instruktionen geben können.

Zudem führt die Analyse Diagnosedaten im laufenden Betrieb und die Aggregation solche Daten mit Informationen aus dem Service und früheren Produktstörungen zur Optimierung der Kundenservicemodele zusammen. Störungen lassen sich frühzeitig feststellen, bevor sie Schaden ausrichten. Damit werden ungeplante Stillstände verhindert, Stillstandzeiten werden verkürzt, Wartungen werden vorausschauend durchgeführt, und Wartungskosten werden eingespart. Die Effizienzsteigerung des Kundenservice verbessert die Markenwahrnehmung und stärkt die Bindung der Kunden an das Produkt- und Service-Angebot.

2.5.1.4 Produktion

Produzierende Unternehmen streben nach der Optimierung ihrer Fertigungsprozesse. Dafür werden Daten bezüglich Produkten, Produktionsketten und Lieferketten erfasst. Mit Hilfe von Big Data-Analysen werden solche Daten verknüpft und analysiert, um einen ganzheitlichen Blick auf die Prozesse zu bekommen.

In vielen Fertigungsindustrien stammen die Produktkomponenten von unterschiedlichen Herstellern. Daten aus der Fertigung, von CAD-Systemen, und Daten über die Produktkomponenten können aggregiert werden, um Ursachen und Wirkung von Qualitätsmängeln zeitnah auszuwerten.

2.5.1.5 Logistik

Kurze Produktzyklen, komplexe Produkte, sinkender Eigenfertigungsanteil, weltweite Beschaffungsmöglichkeiten, volatile Märkte, Naturkatastrophen, knappe Rohstoffe und niedrige Lagerbestände: die Optimierung der Lieferketten eines Unternehmens wird im-

Klassifikationsanalysen	Regressionsanalysen
<ul style="list-style-type: none"> ○ Analyse von Kundenreklamationen ○ Unterstützung von Servicemitarbeitern ○ Kundensegmentierung und Priorisierung ○ Abwanderungsrisiko von Kunden identifizieren ○ Effiziente Planung der Verfügbarkeit von Ersatzteilen ○ Optimierung von Wartungs- und Reparatur-Intervalle 	<ul style="list-style-type: none"> ○ Trends in Kundenanfragen prognostizieren ○ Gewährleistungsanalysen ○ Vorausschauende Wartung ○ Effiziente Planung der Verfügbarkeit von Ersatzteilen

Abb. 2.35 Anwendungsszenarien im Bereich Kundenservice

Klassifikationsanalysen	Regressionsanalysen
<ul style="list-style-type: none"> ○ Qualitätsanalysen der Produktion ○ Produktionsprozessoptimierung ○ Effiziente Produktionsplanung ○ Fehlerursachen erkennen ○ Energieeffizienz 	<ul style="list-style-type: none"> ○ Vorausschauende Instandhaltung ○ Qualitätsvorhersage der Produkte

Abb. 2.36 Anwendungsszenarien in der Produktion

mer komplexer. Big Data-Analysen erlauben eine schnelle Verarbeitung hoher Volumina an unstrukturierten Daten, aggregiert aus den zuvor erwähnten Faktoren/Informationen. Aggregierte Analysen liefern ein übergreifendes Lagebild der Lieferkette für transparentere und effizientere Entscheidungen. Lager können mit Hilfe von Simulationen und Szenarien-Bildung überwacht und optimal verwaltet werden. Mögliche Störungen der Lieferkette werden früh erkannt. Damit wird der Geschäftsbetrieb aufrecht gehalten und die Kundenzufriedenheit optimiert.

Klassifikationsanalysen	Regressionsanalysen
<ul style="list-style-type: none"> ○ Optimaler Lagerbestände ermitteln ○ Lieferketten optimieren ○ Lieferungen überwachen 	<ul style="list-style-type: none"> ○ Risiken von Lieferketten vorhersagen und simulieren ○ Lieferanfragen vorhersagen

Abb. 2.37 Anwendungsszenarien in der Logistik

Klassifikationsanalysen	Regressionsanalysen
<ul style="list-style-type: none"> ○ Performanz-Probleme identifizieren ○ Effizientes Hardware Sizing ○ Zuverlässigkeit der IT-Infrastruktur bewerten ○ IT-Architektur einpassen ○ Sicherheit der IT-Infrastruktur bewerten 	<ul style="list-style-type: none"> ○ Zuverlässigkeit der IT-Infrastruktur vorhersagen ○ Cyber-Attacken vorhersagen ○ Hardware Ressourcen proaktiv und optimal planen

Abb. 2.38 Anwendungsszenarien in der IT

2.5.1.6 IT

Log-Dateien werden bislang nur für begrenzte Zeit gespeichert und unsystematisch ausgewertet. Gründe sind die hohen Volumina der Logs und die unstrukturierte Natur der Daten.

Big Data-Technologien bieten eine kosteneffiziente Archivierung der Logs. Big Data-Analysen ermöglichen es Muster und Abhängigkeiten in den Logs zu identifizieren, um Probleme der IT-Landschaft zu erkennen. Defekte in der Software-, Hardware- oder Netzwerk-Landschaft des Unternehmens können proaktiv erkannt werden, bevor es zu Dienstunterbrechungen kommt.

Die Analyse der Logs über längere Zeiträume ermöglicht es Performance-Probleme rechtzeitig zu entdecken. Hardware- und Netzwerk-Ressourcen können präventiv geplant werden. Damit werden Kosten gespart ohne die Qualität der IT Infrastruktur zu beeinträchtigen.

IT-Sicherheit ist ein Paradebeispiel für den Mehrwert von Big Data-Analysen. Die steigige Digitalisierung wirtschaftlicher Aktivitäten führt zum kontinuierlichen Wachstum der sicherheitsrelevanten Daten.

Bislang bezogen IT-Sicherheit-Warnsysteme nur firmeninterne Logs aus Firewalls, Netzwerk-Logs, Betriebssystemen und „Intrusion-Detection“ und „Intrusion-Prevention“

Klassifikationsanalysen	Regressionsanalysen
<ul style="list-style-type: none"> ○ Leistungsfähige Werttreibermodelle erstellen ○ Kreditwürdigkeit prüfen ○ Transparente Risikoanalysen führen ○ Kundenabwanderung erkennen ○ Betrugsfälle erkennen 	<ul style="list-style-type: none"> ○ Risiko-Vorhersagemodelle erstellen ○ Szenarien und Prognosen bilden ○ Betrugsfälle vorhersagen

Abb. 2.39 Anwendungsszenarien in Risikomanagement

Systemen, in ihre Analysen ein. Informationsquellen, die außerhalb des Monitoring-Bereichs IT-Sicherheit-Warnsysteme wie Listen von Hacker-IP-Adressen, oder Informationen von anderen Firmen können in die Sicherheitsanalyse miteinbezogen werden. Intelligente Frühwarnsysteme mit Präventionsmechanismen können dann entstehen: das System schlägt Alarm bei einem auffälligen sicherheitsrelevanten Verhalten.

2.5.1.7 Risikomanagement

Zu den wichtigsten Anwendungsszenarien von Big Data-Analysen im Bereich Risikomanagement ist die Verschaffung einer informierten und transparenten Entscheidungsfindung indem auffällige Geschäftsrisiken früh erkannt werden. Effiziente Vorhersagemodelle entstehen, wenn Daten aus den verschiedenen Abteilungen und Informationsquellen intelligent verknüpft und analysiert werden. Banken können die Kreditwürdigkeit der Kunden aus dessen Social Media Verhalten prüfen. Unternehmen bekommen die Möglichkeit, kundenindividuelle Angebote und Dienstleistungen anzubieten bevor die Kunden zur Konkurrenz abwandern.

2.5.2 Big Data-Analysen: Trends

Im folgenden werden einige Trends vorgestellt, die aus dem Einsatz Big Data-Analysen entstehen können.

2.5.2.1 Trends im Rechtswesen

Das Rechtswesen ist datengesteuert. Juristische Entscheidungen werden auf Grundlage der geltenden Rechtsvorschriften, vorhandener Rechtsprechung, technischen und regulatorischen Leitlinien, Normen und Standards sowie einer Vielzahl weiterer Quellen getroffen. Urheber der Rechtsquellen sind verschiedenste öffentliche und private Institutionen auf internationaler, europäischer, nationaler, landes- und kommunaler Ebene. Dateninhalte- und -formate sind entsprechend heterogen.

Vernetzte Recherchemöglichkeiten über einzelne Rechtsquellen und -urheber hinweg eröffnen neue Möglichkeiten der Rechtsfindung. Dies kann die Qualität von Entscheidungen und deren Begründung verbessern, Kosten sparen und Prozesse beschleunigen.

2.5.2.2 Trends im Transportwesen

Immer mehr Fahrzeuge werden vernetzt. Viele Fahrzeuge liefern mittlerweile Daten über ihren technischen Zustand und ihren Standort.

Toyota hat in Zusammenarbeit mit Microsoft ein System entwickelt, dass die Standortdaten sowie andere Messdaten wie die Geschwindigkeit der Fahrzeuge auswertet mit dem Ziel die Verkehrslage zu analysieren um den Verkehrsfluss zu erleichtern.

Das deutsche Mauterhebungssystem liefert wertvolle Daten dessen Analyse eine effiziente Verkehrssteuerung ermöglicht um Staus und entsprechende Zeitverluste zu vermeiden.

Solche Daten und Analysen bieten beispielweise neue Effizienzsteigerungsmöglichkeiten für die Transportlogistiker. Transportlogistiker agieren in einem engen Markt mit einer immer komplexeren Kostenspirale: die Verkehrsdichte nimmt ständig zu, die Kosten für Fahrzeuge (Treibstoff, Wartung, Betrieb, Personal) steigen und eng kalkulierte Lieferzeiten. Mit Hilfe Verkehrsdaten können Routen und Beladungen geplant werden um Zeitverzögerungen und Leerfahrten zu vermeiden. Daten über den technischen Zustand der Fahrzeuge ermöglichen eine optimale Planung der Wartungs- und Reparaturintervalle um Stillstandzeiten zu verringern.

Die Analyse der Verkehrs- und Fahrzeugs-Daten hat die Entwicklung Fahrerlose Fahrzeuge vorangetrieben. Solche Fahrzeuge sollen das menschliche Versagen auf den Straßen verringern. Außerdem Fahrerlose Fahrzeuge bieten neue Möglichkeiten für die Transportlogistiker um Personalkosten zu optimieren.

Straßen und Autobahnen-Betreiber können auch von Big Data-Analysen profitieren. Messdaten wie die Vibration, die Geschwindigkeit, der Standort und weitere Daten über die rollenden Fahrzeuge (z. B. Reifendruck, Fahrzeugtyp, usw.) können ausgewertet werden um die Straßenqualität zu bewerten und entsprechende Reparaturen zu planen.

2.5.3 Trends im Sozialen Sektor

Die Vereinten Nationen (UN) hat das Projekt Global Pulse¹ initiiert um das gesellschaftliche Wohlergehens zu analysieren und mögliche gesellschaftliche Probleme vorherzusagen. So könnte beispielsweise eine auf die Anzahl der Tweets, die Bezug auf den Reispreis hatten, basierende Analyse die Preiserhöhung der Reis in Indonesien prognostizieren.

Daten aus unterschiedlichen Quellen wie Social Media Daten und Handydaten sowie historische Daten können aggregiert werden um Maßnahmen im Fall von Katastrophen optimal zu leiten. Ferner, erlaubt die Analyse des Kaufverhaltens und des Sparverhaltens einer Population potentielle wirtschaftliche Krisen vorherzusagen.

¹ <http://www.unglobalpulse.org/>.

2.5.4 Trends im Gesundheitswesen

Einer der möglichen Trends im Gesundheitswesen ist die Individualisierung der medikamentösen Behandlung der Patienten. Die Analyse der DNA-Daten verschiedener Patienten soll es ermöglichen zugeschnittene Medikamente zu entwickeln. Außerdem, soll die aggregierte Analyse der DNA-Daten und der Patientenakten neue Erkenntnisse über Krankheitsbildungen zu prognostizieren und proaktiv Gegenmaßnahmen zu treffen.

Die Telemedizin hat als Ziel die Behandlung der Patienten möglichst zu dezentralisieren. Dabei werden verschiedene Werte wie Blutdruck, Herzschlag, Temperatur, Bewegung, usw. überwacht und zu dem behandelnden Arzt online geleitet. Die intelligente Verknüpfung der Messdaten mit den Patientenakten und anderen archivierten Daten (z. B., Krankheitsbild Symptome) erlaubt eine effiziente dezentrale Patientenvorsorge anzubieten.

2.6 Big Data wird zu Smart Data – Big Data in der Marktforschung

Elke Theobald und Ulrich Föhl

2.6.1 Big Data in der Marktforschung – Goldgrube oder Datengrab?

Zu Beginn eines Beitrags über Big Data in der Marktforschung soll die grundlegende Frage aufgeworfen werden, inwieweit Big Data ein relevantes Thema für die Marktforschung ist. Diese Frage muss vor dem Hintergrund der Aufgaben der Marktforschung und der Relevanz von und der Erkenntnisse durch Big Data beantwortet werden. Relevant ist das Thema dann, wenn das durch Big Data generierte Wissen einen Vorteil bei der Erfüllung der Aufgaben der Marktforschung bringt – sei es zum Beispiel durch einen neuen Erkenntnisgewinn, durch die schnellere Informationsverfügbarkeit oder durch eine kostengünstigere Datenerhebung.

Marktforschung hat in den Unternehmen die Aufgabe, Wissen über Märkte, Konsumenten, Wettbewerber und weitere für die Unternehmensführung relevante Wissensbereiche (wie Technologietrends, gesellschaftliche Entwicklungen) zu schaffen und diese dem Unternehmen bei konkreten Fragestellungen zur Verfügung zu stellen bzw. beratend tätig zu sein. Welches relevante Wissen kann Big Data der Marktforschung liefern? Da Big Data als eine fest definierte Datensammlung nicht existiert, muss überlegt werden, wo Daten in einem großen Umfang entstehen, die bei Aussagen über die Untersuchungsobjekte der Marktforschung behilflich sein könnten.

Fast alle Unternehmen sind bereits heute im Besitz von Big Data. In unterschiedlichsten internen Datensystemen schlummern aufgezeichnete Kundengespräche, E-Mail-Kommunikation mit Kunden und Lieferanten, es existieren CRM-Systeme mit Kundenhistorien und Außendienstberichten, Social-Media-Präsenzen werden betrieben und viele weitere Datenquellen sind schon seit Jahren fester Bestandteil der IT-Landschaft der Un-

ternehmen. Parallel dazu kaufen die Unternehmen Informationen von Drittanbietern zu wie z. B. von Wirtschaftsdatenbanken oder von Marktforschungsinstituten. Allein aus diesen bereits existenten Datensammlungen ergeben sich unzählige Möglichkeiten der gewinnbringenden Analyse durch die Marktforschung. Sei es zum Beispiel indem die Informationen über die Wettbewerber und die Märkte aus der Beschaffungsabteilung, der Vertriebsabteilung und dem Produktmanagement zusammengeführt oder die Stärken und Schwächen der eigene Produkte aus den Call-Center-Protokollen extrahiert werden.

Neben diesen internen Datenquellen existieren im Internet durch die Digitalisierung der Kundeninteraktionen und -transaktionen auf Websites und in sozialen Medien weitere relevante Daten, die zum Beispiel über die Bedürfnisse der Kunden, ihre Kaufabsichten oder ihre Bewertungen von Produkten Auskunft geben. Die täglich bei sozialen Medien entstehenden Datenmengen sind unvorstellbar groß. Neben den Social-Media-Beiträgen produziert aber auch die klassische Webanalyse kontinuierlich Daten ohne weiteres Zutun z. B. durch die Nutzungsprotokolle von Webangeboten in Logfiles oder bei der Speicherung und Auswertung von Bewegungsprofilen z. B. in Customer Journey- oder Clickstreamanalysen, durch Web oder Mouse Tracking. In Zukunft dürfte diese Datenmenge durch die Informationen von mobilen Endgeräten und das Internet der Dinge noch beachtlich wachsen. All diese von Menschen mittelbar oder unmittelbar erzeugten Daten können ausgewertet und daraus Erkenntnisse für Fragestellungen der Forscher und der Unternehmen gezogen werden. Damit kann Big Data einen Beitrag zu den Forschungsfragen der Unternehmen liefern und diese teilweise sogar in Echtzeit wie bei Social-Media-Analysen zur Verfügung stellen.

Die Fülle an möglichen Datenquellen ist schier unerschöpflich. Sie lassen sich nach ihrem Ursprung grob in drei Gruppen unterteilen (Dapp 2014, S. 9):

- Maschinell generierte Daten, z. B. Sensor- oder Logdaten, Klickstatistiken.
- Von Menschen generierte Daten, z. B. Beiträge in Social Networks, Korrespondenz, Publikationen/Patente, Bilder, Freitext, Formulare, Protokolle, Open Data/Web Content, Sprache/Audio/Video.
- Unternehmensdaten (Geschäftsdaten), z. B. Stamm- und Falldaten zu Produkten oder Kunden, CRM- und Transaktionsdaten.

Grundsätzlich kann die eingangs aufgeworfene Frage positiv beantwortet werden: Es gibt sowohl interne als auch externe Datenquellen, die für eine ganzheitliche Sicht auf die Erkenntnisobjekte der Marktforschung zu einer Datenbasis zusammengeführt werden können, sodass der Begriff Big Data für den entstehenden Quellenkorpus gerechtfertigt ist. Diese Daten können relevante Informationen in der erforderlichen Qualität beinhalten, sodass sie eine mögliche Erkenntnisquelle für die Marktforschung darstellen mit der Chance, neue Erkenntnisse zu generieren und dies aufgrund der Digitalisierung und Automatisierung vielleicht effizienter und schneller als mit „klassischen“ Methoden.

2.6.2 Der Marktforschungsprozess bei Big Data

Marktforschung als Servicefunktion im und für Unternehmen ist kein Selbstzweck, sondern immer geleitet von einer Forschungsfrage. Fragestellungen sind dabei die Generierung von Consumer Insights, Markt- und Wettbewerbsanalysen, Werbetests, Preisforschung oder Positionierungsanalysen, um nur einige Beispiele aus der Praxis zu nennen. Ausgehend von der Forschungsfrage operationalisiert der Forscher die Fragestellung in sinnvolle Einheiten und entwickelt ein Forschungsdesign. Es folgt die Feldphase mit der Durchführung der Erhebung oder der sekundärstatistischen Analyse und die Auswertung der Ergebnisse. Auf Basis der Ergebnisse gibt der Marktforscher eine Antwort auf die Forschungsfrage. Sehr häufig werden Kommunikationsmittel für die Weitergabe des generierten Wissens erstellt, häufig in Form von Präsentationen oder Studienberichten. Abbildung 2.40 visualisiert den Ablauf bei einem klassischen Marktforschungsprojekt.

Der nachfolgende Beitrag zeigt systematisch auf, welche grundsätzlichen Herausforderungen bei der Verwendung von Big Data in der Marktforschung berücksichtigt werden müssen.

2.6.2.1 Die Forschungsfrage

Bei der Entdeckung der Forschungsfragen könnte einer der wesentlichen Paradigmenwechsel durch Big Data entstehen. Die klassische Marktforschung wird stark durch konkrete Forschungsaufträge geprägt und natürlich können auch klassische Forschungsfragen

Abb. 2.40 Der klassische Marktforschungsprozess (Eigene Darstellung)



an Big Data als Datenquelle gestellt werden. Auch entstehen im Rahmen der klassischen Recherche bzw. der Erhebung in der Regel zusätzliche Fragestellungen, dennoch könnte sich durch Big Data eine neue Qualität bei der Definition der Forschungsfrage ergeben. Durch die Quellenkorpora von Big Data können im Rahmen der kontinuierlichen Beobachtung und Analyse der Daten neue Erkenntnisse entstehen, ohne dass eine konkrete Fragestellung vorab definiert wurde. Gerade die Entdeckung neuer Zusammenhänge, das Data Mining, wird als eine der großen Chancen von Big Data begriffen. Durch das Zusammenführen bislang unabhängiger Datenquellen entstehen neue Möglichkeiten des Erkenntnisgewinns. Im Rahmen der Datenanalyse von Big Data müssen in Zukunft nicht immer konkret definierte Fragestellungen verfolgt werden, sondern der Forscher kann emergent nach Zusammenhängen in den Daten suchen, um neue Erkenntnisse, Ideen und Konzepte zu generieren und damit Insights zu gewinnen, ohne dass ein konkreter Rechercheauftrag gegeben war. Damit ergeben sich verstärkt Einsatzmöglichkeiten der Marktforschung im Rahmen eines Frühwarnsystems oder der Trendbeobachtung durch das kontinuierliche Monitoring von Big Data.

Mit dieser grundlegend neuen Herangehensweise geht die Einschätzung der Autoren Voss und Sylla einher, die in Big Data eine disruptive Technologie für die Marktforschung sehen. Die Daten werden nicht mehr durch Umfragen erzeugt, sondern liegen potenziell schon online vor (Voss und Sylla 2014, S. 40). Teile des klassischen Marktforschungsprozesses würden somit neu definiert.

2.6.2.2 Das Forschungsdesign

Marktforscher legen im Rahmen des klassischen Forschungsdesigns neben der Bestimmung der Erhebungsmethoden bei Primärerhebungen auch die zu befragende Stichprobe fest, um ein bestmögliches Abbild der Realität zu erreichen. Das Forschungsdesign beim Einsatz von Big Data beschäftigt sich zunächst mit der **Identifikation der relevanten Quellen** – dieser Schritt kann mit der Festlegung der Stichprobe gleichgesetzt werden, auch wenn im Rahmen von Big Data die Möglichkeiten bestehen, eine viel größere Grundgesamtheit zu nutzen, um die Erkenntnisse abzusichern und Prognosen mit mehr Informationen anzureichern (Dapp 2014, S. 13). Nach der Quellenauswahl folgen in der Big Data-Forschung für den Marktforscher die **Aufbereitung der relevanten Daten** aus diesen Quellen und die **Wahl der analytischen Methoden**, mit denen die Daten zusammengeführt und ausgewertet werden können. Dieser Prozess kann mehrfach durchlaufen werden, bis der gewünschte Erkenntnisgewinn eingetreten ist.

Die Auswahl der Daten ist ein kritischer Meilenstein bei einem Big Data-Projekt – hier sollte nicht die Menge der Daten im Vordergrund stehen, vielmehr ist die Selektion der relevanten Daten entscheidend – also Smart Data und nicht unbedingt Big Data (Kary 2014, S. 26). Bei der Identifikation der relevanten Quellen und der Bewertung der darin enthaltenen Daten beschäftigen sich die Forscher mit folgenden zentralen Fragestellungen:

- Relevanz der Daten: Liefern die Daten relevante Informationen zur Forschungsfrage bzw. können die Daten zu einem Erkenntnisgewinn beitragen?

- Qualität der Daten: Wie aktuell sind die Daten und bilden die enthaltenen Informationen die Realität ab? Damit verbunden entstehen auch Fragen zu Redundanzen von Informationen.
- Struktur der Daten: Sind die Daten so strukturiert, dass sie verwendet werden können?
- Zugänglichkeit der Daten: Ist der Datenzugriff öffentlich möglich und wie lassen sich die Daten extrahieren?
- Vergleichbarkeit der Daten: Können Daten aus verschiedenen Quellen kombiniert werden, gibt es Ordnungskriterien, die die Daten vergleichbar machen (z. B. gleiche Währungseinheiten, gleiche Zielgruppenmodelle, gleiche Zeiteinheiten und vieles mehr).

Für die nachfolgenden Schritte der Datenaufbereitung und der Analyse müssen in der Konzeptphase folgende Entscheidungen getroffen werden:

- Selektion: Welche konkreten Daten aus den Quellen sollen zur Verwendung in der Analyse herangezogen werden (z. B. Selektion der Datensätze der letzten 2 Jahre)? Gegebenenfalls folgt die Datenextraktion.
- Bereinigung bzw. Vereinheitlichung: Welche nicht benötigten Informationen können gelöscht werden (zum Beispiel Uhrzeit eines Kaufs). Wie müssen die Daten vereinheitlicht werden, um die Vergleichbarkeit der Informationen zu erreichen (zum Beispiel gleiches Datumformat, gleiche Währungen u. v. m.)?
- Verdichtung: Wie können Daten verdichtet werden? Nach welchen Kriterien können Informationen für den Erkenntnisgewinn zusammengefasst werden?
- Kombination: Wie können Daten aus unterschiedlichen Quellen kombiniert werden, um den angestrebten Informationsgewinn zu erzielen? Erzeugt die Kombination der Quellen valide Datensätze und neue Erkenntnisse?
- Problemhandling: Welche Quelle und welche Daten haben Priorität bei widersprüchlichen Aussagen in den Quellen? Wie werden Probleme in den Quellen behandelt (z. B. Nachrecherche, Ausschluss)?
- Analyse: Welche analytische Verfahren nutzen wir zur Erkenntnisgenerierung aus den Daten, z. B. kommen hier Korrelationsanalysen, Multivariate Analysen, Clusteranalysen, Assoziationsanalysen, Regressionsanalysen, Trendanalysen, Klassifikationsverfahren, Textmining, Zeitreihenanalysen, Häufigkeitsauswertungen und viele andere Analyseverfahren in Betracht.

Abbildung 2.41 fasst die wichtigsten Entscheidungskriterien für die Datenbewertung und die Datenaufbereitung zusammen.

2.6.2.3 Die Erhebungsphase: Die Nadel im Heuhaufen

„Es stellt sich aber die Frage, ob das nun heißt, dass es mehr Nadeln im Heuhaufen gibt, oder der Heuhaufen nur größer geworden ist.“ (Kary 2014, S. 22)²

² zitiert wird der Big-Data Experte Stefan Rüping, Fraunhofer Institut.

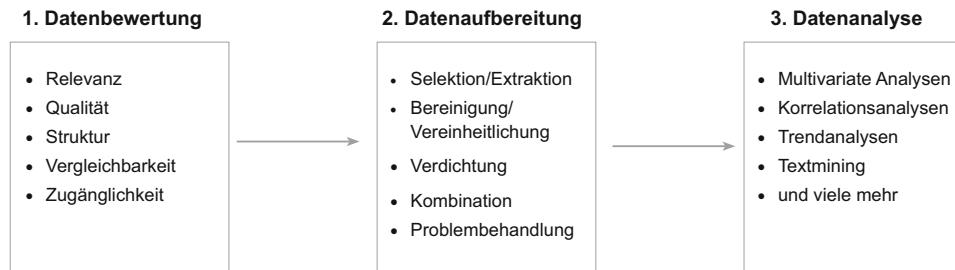


Abb. 2.41 Datenbewertung und Datenaufbereitung bei Big Data (Eigene Darstellung)

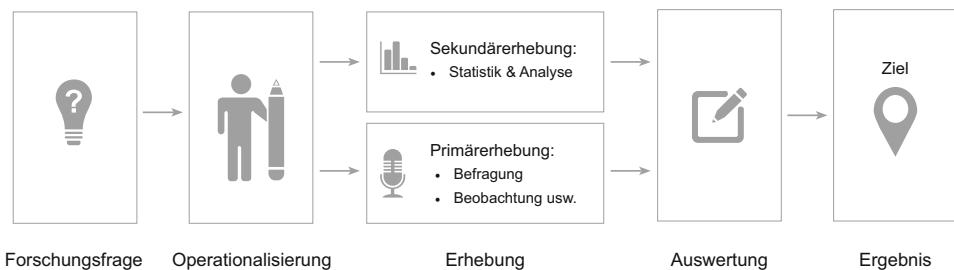


Abb. 2.42 Der Analyseprozess bei Big Data (Quelle: Eigene Darstellung)

Vor der Auswertungsphase müssen die Daten in eine auswertbare Form gebracht werden. Hier arbeiten Informatiker und Marktforscher eng zusammen und der Marktforscher wird zum Data Analyst. Die Quelldaten werden zu neuen Datensätzen transformiert und kombiniert und die Analysen auf dem neuen Quellenkorpus durchgeführt. Die Daten werden vorbereitet, Analysen und Suchläufe durchgeführt, Muster gesucht, Hypothesen auf den Daten getestet, die Ergebnisse geprüft und interpretiert, gegebenenfalls die Algorithmen der Analyse oder das Datenmaterial angepasst und der Kreislauf beginnt von vorne. Der Marktforscher überwacht die kontinuierlichen Analysen und führt den Prozess so lange durch, bis plausible Ergebnisse mit neuem Erkenntnisgewinn generiert werden können. Einmal erhobene Datensätze können dabei zu unterschiedlichen Zwecken mehrfach und immer wieder ausgewertet werden. Abbildung 2.42 veranschaulicht den Prozess.

2.6.3 Aktuelle Herausforderungen für den Big Data Einsatz in der Marktforschung

Der Einsatz von Big Data in der Marktforschung ist mit einigen zentralen Herausforderungen konfrontiert, die nachfolgend vorgestellt werden sollen.

2.6.3.1 Datenzugänglichkeit und Repräsentativität

Die auszuwertenden Informationen liegen im Internet, aber auch unternehmensintern häufig in unterschiedlichen Anwendungen und damit auch Datenformaten vor (Hofmann 2012, S. 141)³. Die Heterogenität der Formate ist eine Herausforderung für die Kombination und die Analyse der Daten. Dieser Umstand kann dazu führen, dass bestimmte Quellen nicht aufgenommen werden können – einfach weil dazu bislang die Schnittstellen nicht programmiert wurden, weil die Informationen nur in geschlossenen Nutzergruppen oder Datenbanken zu finden sind, oder weil der Zugriff durch Captcha geschützt ist. Wir müssen davon ausgehen, dass nur ein bestimmter Prozentsatz an Informationen im Internet mit Standardmethoden automatisiert zugänglich ist. Grundsätzlich lauert in dieser Tatsache die Gefahr, dass die technische Zugänglichkeit der Quellen das Entscheidungskriterium für die Datenaufnahme in die Forschung ist – und nicht die inhaltlich beste Quelle gewählt wird. Einfach zugängliche Daten werden vielleicht eher für Big Data-Analysen gewählt, und vielleicht werden auch die Auswertungsmethoden anhand der technischen Möglichkeiten der Datenquellen bestimmt (Boyd 2010).

Für die Marktforschung stellt sich darüber hinaus die grundsätzliche Frage, inwiefern die zugänglichen Daten alle Meinungen und Fakten wiederspiegeln und wirklich ein Abbild der Realität bilden. Anders gesagt: Genügt die Datenlage in Big Data den Gütekriterien der Marktforschung? Dieser Herausforderung muss sich jeder Forscher im Big Data-Umfeld bewusst sein und die Frage bestmöglich beantworten bzw. entsprechende Vorkehrungen treffen, indem z. B. ergänzende Erhebungsmethoden eingesetzt oder zusätzliche Analysen zur Absicherung der Methoden und der Erkenntnisse durchgeführt werden.

Probleme bestehen bezüglich der Unzuverlässigkeit, der Fehleranfälligkeit und der Unvollständigkeit großer Datensätze, denn Big Data kann grundsätzlich nicht mit der Vollständigkeit aller Informationen gleichgesetzt werden (o. V. 2014)⁴. Diese grundlegenden Probleme werden bei der Kombination unterschiedlicher Datenquellen potenziert, denn mit jeder Verknüpfung steigt die Komplexität. Wenn beispielsweise 5 Datenpunkte aus der Warenkorbanalyse mit 5 Fakten aus abgeschlossenen Verkäufen verknüpft werden, resultieren am Ende nicht 10 neue Datensätze, sondern sie multiplizieren sich auf bis zu 25 Items, die neu interpretiert werden können. Dabei besteht bei steigender Komplexität die Gefahr, im Rahmen der Analyse Muster erkennen zu wollen, die gar nicht existieren. Dies könnte im schlimmsten Fall zu falschen Entscheidungen führen (Dapp 2014, S. 33).

2.6.3.2 Herausforderung Text Mining und Social-Media-Analyse

Das digital zugängliche Material ist naturgemäß nicht nach der Forschungsfrage strukturiert und zeichnet sich bei natürlich sprachlichen Texten durch eine große sprachliche

³ Olaf Hofmann redet in diesem Zusammenhang von „Blended Data“ und „Blended Samples“, die Daten/Informationen aus unterschiedlich(st)en Quellen nutzen und zielgerichtet für ein bestimmtes Erkenntnisinteresse auswerten.

⁴ Als größte Hürde bei der praktischen Umsetzung nennen die Befragten die Sicherheit der Daten. So sehen die Entscheider aus Industrieunternehmen Datensicherheit (50 %), Datenqualität (44 %) und Datenschutz (42 %) als größte Hürden.

Ausdrucksvarianz aus. Viele Analyseprozesse im Textmining setzen auf das Vorkommen spezieller Schlagwörter, um Antworten auf die Forschungsfragen zu finden. Doch Synonyme, Homonyme, Dialekte und Fehlschreibweisen erschweren das Finden relevanter Beiträge im Internet und in internen Datenbanken. Wirklich alle Quellen zu finden, diese zu homogenisieren und daraus gegebenenfalls eine wie auch immer geartete sinnvolle Stichprobe zu ziehen (z. B. nach Quellenarten oder nach Nutzertypen) oder die Daten intelligent zu kombinieren bleibt deshalb nach wie vor und für jede Forschungsfrage wieder neu eine spannende Herausforderung und bestimmt wesentlich die Qualität der nachfolgenden Forschung.

Neben der umfassenden Erschließung der relevanten Quellen und validen Methoden zur Stichprobengenerierung müssen gerade bei Big Data notwendige Maßnahmen zur Qualitätssicherung im Quellenkorpus durchgeführt werden. So gilt es z. B. Doppelnenigungen herauszufiltern (da User identische Beiträge in Twitter, Facebook und Google Plus posten oder sie über mehrere parallele Konten verfügen) oder gefälschte bzw. bezahlte Einträge zu identifizieren.

In der Analysephase muss festgelegt werden, welche automatisierten Analysen der Informationen sinnvoll durchführbar sind. Bei strukturierten Informationen, wie Kundendatenbanken oder der Auswertung von Warenkörben, sind automatisierte Analysen der quantifizierbaren Daten auf statistischer Basis einfacher zu identifizieren und zu entwickeln als zum Beispiel beim Text Mining im Rahmen von Social-Media-Analysen. Gerade Inhaltsanalysen in sozialen Medien stellen eine große Herausforderung dar. Sollen z. B. Produkterfahrungen oder Einstellungen analysiert werden, ist eine semantische Analyse notwendig. Hier hat die Computertechnologie in den letzten Jahren Fortschritte gemacht, eine vollautomatische korrekte Erkennung ist jedoch in den seltensten Fällen möglich. Softwareprogramme können bei der Analyse die Arbeitsschritte unterstützen, durchführen sollte sie aber ein Mensch, um die Qualität der Analyse abzusichern. Die Analysen erfordern die Ausarbeitung eines differenzierten Codeplans, um den Analyseprozess nachvollziehbar und valide zu machen. Standardkorpora und Wörterbücher helfen hier nur bedingt, da z. B. der Begriff „lang“ im Zusammenhang mit Reparaturzeiten als negativ zu werten ist, aber im Zusammenhang mit der Lebensdauer einer Batterie in der Regel eine positive Eigenschaft darstellt. Analytische Programme können uns bei textbasierten Big Data-Korpora vor allen Dingen helfen, die Daten zu strukturieren und passende Beiträge zu spezifischen Themen zu identifizieren – sie bereiten uns die Quellenkorpora auf und machen sie zugänglich. Am Ende des Big Data-Prozesses gilt es, die richtigen Schlüsse aus der Kombination von Daten und Methodik zu ziehen (o. V. 2014)⁵.

2.6.3.3 Pluralität der Meinungen

Im Wissensmanagement ist es eine altbekannte Weisheit, dass man nicht alles, was man weiß, sagen kann und dass nicht alles, was man sagen kann, auch aufgeschrieben wird

⁵ In der Erhebung von PwC sehen die Nutzer die größten Gefahren in technischen Problemen und der möglichen Fehlinterpretation der Daten.

(Angioni 2004, S. 252)⁶. Dieser Grundsatz gilt ebenso für das Wissen, das in Unternehmen generiert wird, und für Aussagen der Verbraucher im Internet. Studien über die Nutzung von Internet- und Web 2.0-Anwendungen bestätigen bis heute, dass nur bestimmte Zielgruppen aktiv Online-Beiträge verfassen. Auf der Rezeptionsebene wird User Generated Content von sehr vielen Internetnutzern verwendet, doch das aktive Verfassen von Beiträgen ist nach wie vor auf eine spezifische Zielgruppe beschränkt. Durch diese Tatsache bleibt nach wie vor die zentrale Frage bestehen, ob zum Beispiel über User Generated Content die Pluralität der Meinungen in der Bevölkerung erfasst werden kann: Big Data sind zwar viele Daten, aber sie enthalten vielleicht doch nicht alle möglichen Meinungsbilder und vielleicht auch gerade nicht die der eigenen Zielgruppe.

Doch trotz aller Herausforderungen: Big Data ist auf jedem Fall auch eine Fundgrube. Der Forscher hat die Chance, authentische Meinungsäußerungen zu finden, ohne diese gesondert zu erheben. Diese können gut als Original-Töne z. B. bei einer Produktneueinführung oder einer neuen Werbekampagne genutzt werden und die strukturierten Analysen durch Consumer Insights ergänzen.

2.6.3.4 Interpretation multimedialer Daten

Bei den Datenmengen im Internet handelt es sich nicht nur um strukturierte Daten (Gogia 2012)⁷ und auch nicht nur um reine Textdaten. Zunehmend spielen weitere Informati-onsträger wie z. B. Videos und Bilder eine bedeutende Rolle und sollen in die Analysen einbezogen werden. Im Kontext von Big Data bedeutet dies, dass z. B. alleine bei YouTube pro Minute 100 Std. Film hochgeladen werden (Googlewatchblog.de). Hier die relevanten Beiträge z. B. zu den eigenen Produkten oder Themenkreisen zu recherchieren, ist in einem gewissen Rahmen möglich (vorausgesetzt, die Medien tragen gut recherchierbare und treffende Metainformationen). Sollen alle Medien in die Analyse einbezogen werden, muss mit unterschiedlichen Analysemethoden für Bildanalyse, Textanalyse und Videoanalyse gearbeitet werden, die nur in einem sehr begrenzten Umfang automatisiert möglich sind.

2.6.3.5 Der Kontext macht den Unterschied

Bei der Analyse von Meinungen und Einstellungen und bei der Bewertung von Aussagen spielt häufig der Kontext der Aussage eine große Rolle bei der Interpretation. Im Rahmen der Datenextraktion (z. B. bei Social-Media-Analysen) oder der Datenfusion werden Aussagen und Fakten vom Kontext getrennt, damit fehlen dem Forscher teilweise wichtige Interpretationshinweise. Betrachtet der Forscher während der Analyse nur einzelne Beiträge, ohne den Kontext der Ursprungsquelle zu berücksichtigen, existiert eine Ursache der möglichen Fehl- oder Überinterpretation. Gerade bei Big Data entstehen sehr viele Informationen, die für die Analyse isoliert in den Quellenkorpus übernommen werden.

⁶ Das sogenannte implizite Wissen.

⁷ Mehr als die Hälfte der Umfrageteilnehmer nannten die Verwaltung von und den Zugriff auf unstrukturierte Daten (57 %) und die wachsende Datenmenge (51 %) als größte Herausforderungen.

Oft diskutieren beispielsweise Blogger Beiträge aus befreundeten Blogs, indem sie deren Beiträge zitieren – ohne Kontextinformation kann dies zu einer Fehleinschätzung des Beitrags führen. Auch bleibt die Identität der Autoren im Internet häufig im Verborgenen. Somit ist zum Beispiel unklar, ob ein Blogger im Auftrag einer Organisation oder eines Unternehmens handelt.

2.6.3.6 Von Korrelationen und Kausalitäten

Big Data liefert in erster Linie Korrelationen, statistische Zusammenhänge in den Daten, aber keine Kausalitäten (Anderson 2008). Die entstehenden Datenreihen können durch Analysetools auf Korrelationen untersucht werden, doch nicht immer sind diese durch Algorithmen gefundenen Zusammenhänge logisch oder logisch erkläbar. Die Kausalität der identifizierten Zusammenhänge kann nicht automatisiert belegt oder erklärt werden, sondern dies obliegt nach wie vor dem Forscher. Die Algorithmen stellen die logischen Kombinationen dar und zeigen mögliche Zusammenhänge in den Daten, die wir ohne die Analyse nicht entdeckt hätten. Aber die Analyse zeigt nicht auf, warum es diese Zusammenhänge gibt. Mit Danah Boyd gesprochen: Was und Warum sind zwei unterschiedliche Fragen (Boyd 2010). Erst der Forscher stellt fest, welche Antworten die Daten sinnvoll geben können, die Datenanalyse liefert lediglich die Entscheidungsbasis hierfür. Allerdings zeigt diese Betrachtung bereits eine Gefahr, die Big Data inhärent ist: Die Analysen könnten den Forscher dazu verleiten, die Daten für sich sprechen zu lassen und nicht mehr modellgeleitet Zusammenhänge zu entdecken. In diesem Vorgehen liegen Chancen und Risiken zugleich, weshalb sie vom Forscher ganzheitlich abzuwegen sind. Zum einen besteht die Chance auf Erkenntnisse zu stoßen, die im Rahmen einer Modellbildung nicht in Betracht gezogen worden wären, zum anderen besteht die Gefahr, dass die menschliche Intuition und Kreativität beschränkt bzw. in eine bestimmte Richtung geleitet wird (Dapp 2014, S. 34; Streif 2013, S. 19)⁸.

2.6.3.7 Topaktuell und doch Schnee von gestern

Internetbasierte Daten haben den Vorteil, dass bei aktuellen Themen zeitnah Bewertungen und Argumente der Konsumenten gefunden werden können. Trendthemen lassen sich somit teilweise frühzeitig durch die Analyse von Big Data erkennen. Google Insight for Search macht es vor: Durch die Analyse von Millionen Suchanfragen lässt sich die Ausbreitung von Grippeviren ebenso prognostizieren wie die Entwicklung von Börsenkursen oder die Sieger des European Song Contest (Reips 2009, S. 129). Die Ergebnisse der Prognosen sind inzwischen auch im Unternehmensumfeld beachtlich. So schafft es das US-Handelsunternehmen Target mit einem Prognoseverfahren auf Basis der Vergangenheitsdaten künftige Konsumentenbedürfnisse erschreckend präzise vorherzusehen und entsprechende Angebote im Vorfeld zu unterbreiten⁹. Amazon sendet per Anticipatory

⁸ Streif (2013, S. 19) „Fast alles ... ist stark datengetrieben aber selten erkenntnisgetrieben.“

⁹ Berühmt sind folgende Beispiele: Target gelingt es vorherzusagen, welche Kunden in nächster Zukunft ein Baby erwarten werden. Ein Telekommunikationsunternehmen schafft es, die Kündigung der Kunden vorhersehen noch bevor diese durchgeführt wurde und ergreift präventiv Gegenmaß-

Package Shipping Waren schon vor der Bestellung in die Verteilcenter – Basis der Verteilung ist die datenbasierte Wahrscheinlichkeit für künftige Bestellung. Apple hat ein Patent für einen Algorithmus angemeldet, der Werbebotschaften in Abhängigkeit von der Stimmung des Nutzers an bestimmte Endgeräte sendet (Campillo-Lundbeck 2014, S. 17).

Allerdings darf nicht vergessen werden, dass mit Big Data sehr häufig Daten der Vergangenheit ausgewertet werden. Sei es das Klick- und Surfverhalten im Internet, um Rückschlüsse auf Interessen und Bedürfnisse der Konsumenten zu ziehen oder um zu erkennen, in welcher Phase des Kundenlebenszyklus sie sich gerade befinden. Was wir auf Basis dieser Analyse nicht wissen, ist, welche Interessen und Bedürfnisse der Konsument morgen haben wird. Die Analysen des bisherigen Verhaltens sind wertvoll, basieren jedoch stets auf Vergangenheitswerten oder im besten Fall auf Echtzeit-Informationen. Zwar können durch die Interpretation dieser Fakten Entwicklungen und Trends identifiziert und Annahmen über zukünftiges Verhalten getroffen werden, allerdings in der Regel nur innerhalb der bekannten Szenarien. Neue Strukturen, Visionen oder Wirkfaktoren oder gar disruptive Entwicklungen können nicht prognostiziert werden (Berchtenbreiter 2013).

2.6.4 Die Zukunft von Big Data in der Marktforschung

Data is the next intel inside (Tim O'Reilly).

Welche Erkenntnisse können wir mit Hilfe von Big Data gewinnen? Die Antwort ist immer abhängig von der Qualität der Daten, ihrem Kontext, der Konsistenz, der Möglichkeiten der Kombination und letztendlich der Fähigkeit des Data Analysten, sie in unternehmensrelevante Zusammenhänge zu bringen (Bloching et al. 2012, S. 55). Letztendlich wird die Bedeutung digitaler Datenkorpora in Zukunft durch immer neu entstehende Datensammlungen wachsen. Das Berufsbild des Marktforschers wird sich in den nächsten Jahren immer mehr zum Data Analyst wandeln – hoffentlich zu einem Data Analyst, der auch fundierte methodische Kenntnisse in der Marktforschung besitzt. Der Marktforscher muss neben maschinellen Algorithmen auch die Kunst der Interpretation der Daten beherrschen.

Auch bei Big Data-Analysen müssen die Qualitätsprinzipien der Marktforschung angelegt werden und es darf immer die Frage gestellt werden, ob alles Machbare auch wirklich sinnvoll ist. Sinnvoll sind Anwendungen dann, wenn sie zu neuen Erkenntnissen führen (z. B. durch die Kombination bislang getrennter Daten) oder sie Prozesse vereinfachen. Aber gerade bei der Vereinfachung von Prozessen muss sichergestellt werden, dass die bislang erreichten Qualitätsstandards auch beibehalten bleiben. Marktforschungsprojekte werden sich ändern. Neben die Ad-hoc-Forschung werden verstärkt permanente Analysen treten. Monitoring, Dashboards und Visualisierungen werden durch Big Data an Bedeutung gewinnen (Voss und Sylla 2014, S. 40).

nahmen. Ein Kreditkartenunternehmen weiß auf Basis der Analyse der Kreditkartenabrechnung, welche Paare sich scheiden lassen werden. Weitere Beispiele finden sich in Bloching et al. (2012).

Durch die Verknüpfung der Erkenntnisse aus Marktforschung und Datenanalyse können die Unternehmensprozesse frühzeitig auf veränderte Marktbegebenheiten angepasst werden und garantieren so vielleicht den Unternehmenserfolg (Berchtenbreiter 2013). Dadurch könnte sich Big Data als Quelle für Innovationen und Kreativität erweisen und idealerweise in neue Geschäftsideen, Produkte oder Dienstleistungen münden (Dapp 2014, S. 3).

2.7 Big Data und Electronic Commerce – Neue Erkenntnisse zur Customer Journey

Ulrich Föhl und Elke Theobald

2.7.1 Einleitung

Bei der Frage nach der Bedeutung des Themas Big Data für den Electronic Commerce ist zunächst zu klären, inwieweit charakteristische Merkmale von Big Data im E-Commerce zum Tragen kommen.

Die charakteristischen Merkmale von Big Data werden üblicherweise in den sogenannten „Drei V's“ zusammengefasst (Kudyba und Kwatinetz 2014, S. 2 f.): Velocity, Volume und Variety. Velocity bezieht sich dabei einerseits auf die Geschwindigkeit, mit der Daten entstehen und sich verändern, andererseits auf die Geschwindigkeit, in der Daten verarbeitet werden können. Die Geschwindigkeit hat unmittelbare Auswirkungen auf die Datenmenge (Volume). Variety steht schließlich für die Vielfalt der Daten, also ihre unterschiedlichen Formate und Kanäle, aus denen sie stammen.

Welche Rolle diese Aspekte im E-Commerce spielen, soll durch einen Blick auf aktuell wichtige Themen im Onlinehandel sowie eine Betrachtung der unterschiedlichen Datentypen, die sich daraus ergeben, herausgearbeitet werden. Im Anschluss daran wird ausgeführt, welche Anwendungsmöglichkeiten daraus resultieren, wobei ein ganzheitlicher Blickwinkel eingenommen wird, der die Customer Journey, also das Kundenverhalten im Rahmen eines Kaufprozesses, in den Mittelpunkt stellt. Ziel dabei ist es, aus „Big Data“ „Smart Data“ zu generieren, wofür am Ende des Kapitels einige zentrale Empfehlungen abgeleitet werden.

2.7.2 Aktuelle Themen im E-Commerce

Traditionell steht im E-Commerce eine Fülle leicht erfassbarer Informationen zur Verfügung. So liefert bereits die klassische Webanalyse umfangreiches Material, wie Nutzungsprotokolle oder Bewegungsprofile. Inzwischen erweitern allerdings aktuelle Themen, die teilweise technologisch getrieben sind und teilweise auf einem geänderten Verhalten der Konsumenten basieren, nicht nur Datenvolumen und Geschwindigkeit, sondern verändern

auch stark die Datenvariabilität. Im Folgenden wird eine Reihe von Themen vorgestellt, die den E-Commerce aktuell und in Zukunft in besonderem Maße prägen und beschäftigen werden (Heinemann 2014, S. 11–15).

Eine aktuelle Tendenz im Onlinehandel stellt das sogenannte **Multi-Screening** dar, womit der Umstand zum Ausdruck gebracht wird, dass Internetnutzer in verschiedenen Situationen oder auch parallel unterschiedliche Geräte nutzen. So zeigen Studien von Google, dass etwa 65 % der Onlinekäufe über ein mobiles Endgerät initiiert und zu 61 % an einem Desktop-Rechner abgeschlossen werden (Heinemann 2014, S. 11). Mobile Geräte stellen somit häufig einen ersten Kontakt zwischen Kunde und Onlineshop her, der dann im weiteren Verlauf auf anderen Kanälen fortgeführt wird. Der rapide wachsende Tablet-Markt setzt diesen Trend hin zur Nutzung mehrerer Informationskanäle weiter fort. Während Nutzungsdaten vor einiger Zeit noch nahezu ausschließlich aus Daten klassischer Webseiten analysiert werden konnten, besteht die Herausforderung zunehmend darin, Daten aus mehreren Kanälen miteinander zu verbinden und ganzheitlich aufzubereiten und zu analysieren.

Mit **internetfähigen Fernsehgeräten** zeichnet sich eine Erweiterung des Multi-Screenings ab. So arbeiten Unternehmen wie Google oder Apple an Lösungen für Smart-TV, wodurch sich Internet und Fernsehen in einem Gerät verbinden lassen. Damit ergibt sich die Chance, Internet- und Fernsehverhalten gemeinsam zu betrachten, was bislang durch die stärkere Trennung der Medien erschwert war.

Ein weiteres zentrales Thema, das bereits vor einiger Zeit aufkam, dessen Bedeutung aber immer noch anhält, stellen **Social-Media-Daten** dar. Aufgrund der zunehmenden Social-Media-Nutzung und -verbreitung stellt sich zum einen die Frage, welche Social-Media-Aktivitäten für das jeweilige Unternehmen und seine Zielgruppen empfehlenswert sind und wie Social-Media-Daten der Konsumenten von Unternehmen genutzt werden können. Zum anderen müssen diese Aktivitäten im Kontext der Nutzung verschiedener Kanäle adressiert werden, was die Notwendigkeit einer **Cross-Media-Strategie** verdeutlicht.

Die parallele Nutzung verschiedener Kanäle, die Verbreitung internetfähiger mobiler Endgeräte sowie die angestiegene Social-Media-Nutzung begünstigen die sogenannte **SoLoMo**-Vernetzung, eine Vernetzung sozialer, lokaler und mobiler Daten, die neue Möglichkeiten der Vermarktung schafft, wenn die Daten integriert analysiert werden (Heinemann 2014, S. 14). Aufgrund der mobilen Verfügbarkeit von Internet und sozialen Netzwerken entstehen beispielsweise Kommunikation zu Marken und Produkten sowie neue Absatzchancen unter Einbindung lokaler Angebote bzw. Händler. Umgekehrt lassen sich aufgrund der Möglichkeiten zur Lokalisierung durch mobile Geräte Schlussfolgerungen für den lokalen Handel ableiten, wodurch Offline- (stationärer) und Online-Handel näher zusammenrücken.

Im Zuge der aktuellen technologischen Entwicklungen hat sich auch die Einstellung des Konsumenten in Kaufsituationen deutlich verändert. Aus einem eher passiven Rezipienten von Produkt- und Werbeinformationen wurde ein aktiver Konsument mit einem hohen Bedürfnis nach **Selbstbestimmung** in der Interaktion mit Unternehmen und ihrer

Kommunikation (Spieß 2013, S. 126). Daraus ergibt sich ein höheres Maß an Kommunikation zwischen Konsumenten sowie zwischen Konsument und Unternehmen, beispielsweise auf Social-Media-Kanälen, wodurch vermehrt Daten generiert werden. Zudem werden zunehmend digitale Informationsangebote im Vorfeld eines Produktkaufs genutzt, was wiederum den Unternehmen wertvolle Daten zu den Bedürfnissen der Konsumenten liefert.

Amazon mit einer Sortimentsbreite von über 2,5 Mio. Produkten reagiert darauf, indem das Unternehmen Produktsuche und Handel stark zusammenrückt, was sich mit dem Begriff **Search-Commerce** umschreiben lässt und auch beispielsweise von Google im Rahmen seiner Bücherplattform „Google eBooks“, bei Produktsuchen sowie der Entwicklung eigener Bezahlsysteme aufgegriffen wird.

Multiple Produktdigitalisierung als weiteres relevantes Zukunftsthema bezeichnet den Trend der zunehmenden Digitalisierung von Produkten, die grundsätzlich digitalisierbar sind wie etwa Musik, Zeitschriften, Bücher oder auch Tickets. Dadurch verlagern sich weitere Käufe oder Serviceleistungen in den Bereich des Electronic Commerce, was wiederum die verfügbare Datenmenge steigert.

Die einzelnen Trends verdeutlichen klar einen Anstieg des Datenvolumens, der sich etwa aus der vermehrten Nutzung digitaler Informationsquellen im Rahmen des Kaufprozesses, aufgrund der durch Konsumenten und Unternehmen betriebenen Social-Media-Aktivitäten, die zunehmende Internetverfügbarkeit durch mobile Endgeräte, die auch selbst Daten beispielsweise zum Standort generieren, oder die Zunahme digitaler Informationen durch digitales Fernsehen oder Digitalisierung vieler Produktgruppen ergibt. All diese Informationen entstehen in Echtzeit- neben Volumen und Geschwindigkeit steigt durch die sehr unterschiedlichen Datenquellen, die inzwischen neben klassischen Webseitendaten für den E-Commerce nutzbar sind, auch die Variabilität der verfügbaren Daten und Datenstrukturen. Der E-Commerce ist somit klar von allen 3 „V's“ betroffen, sodass sich die Frage stellt, wie mit diesen großen und stetig wachsenden Datenmengen gewinnbringend umgegangen werden kann.

2.7.3 Daten und Datenstrukturen

Die Daten, die heute und in Zukunft im E-Commerce nutzbar sind, gehen weit über die klassische Webanalyse hinaus. Die Webanalyse zielt darauf ab, den Erfolg von Onlineaktivitäten zu erfassen und betrachtet dazu Daten wie Klickraten auf Seiten oder etwa Werbebanner, die auf bestimmten Seiten verbrachte Zeit, Conversions- und Interaktionsraten oder die Navigationsroute innerhalb eines Angebots (Haberich 2013, S. 50; Kudyba 2014, S. 147f.). Allein diese Daten wachsen kontinuierlich an und werden noch nicht konsequent und umfassend genutzt. So ergab eine Econsultancy-Studie unter Marketing-Experten, dass weniger als die Hälfte der verfügbaren Daten für die Weiterentwicklung des Geschäfts genutzt wird (Haberich 2013, S. 56).

Darüber hinaus stehen Unternehmen weitere Datenmengen im Bereich der Business Intelligence zur Verfügung, also Daten zu Geschäftsabläufen mit Informationen zu Warenwirtschaft, CRM oder Einkaufsdaten. Häufig liegen diese in unterschiedlichen Systemen im Unternehmen vor und müssen zunächst zusammengeführt werden. Eine integrierte Betrachtung von im E-Commerce relevanten Daten kann sich folglich auf folgende Datenquellen stützen (in Anlehnung an Grüger 2013, S. 260 f.; Kudyba 2014, S. 153–155):

- Daten aus Online-Marketing-Kanälen wie etwa Google AdWords,
- Tracking-Daten: Bestellungen, Umsätze, Artikel,
- Backend-Daten: Retouren, Stornierungen, Lieferzeiten,
- Technische Daten: Devices, Browser, IP-Adressen,
- CRM-Daten,
- Einkaufsdaten: Bestand, Marge,
- Daten aus dem Produktinformationsmanagement (PIM): Anzahl Produktbilder, Testberichte, Kundenbewertungen,
- Mobile Daten (Apps, Sensordaten wie GPS etc.),
- Daten aus Social Media-Kanälen (Likes, Texte, Bilder etc.).

Es ist davon auszugehen, dass die Zahl digitaler Datenkanäle in nächster Zeit noch weiter anwachsen wird. So werden etwa durch Wearables, also tragbare Geräte wie Uhren, die vielfältige Sensoren zum Beispiel zur Erfassung von Körperfunktionen aufweisen, sowie das Internet der Dinge noch weitere Datenmengen hinzukommen, wodurch sich Webanalyse und Business Intelligence zunehmend zu Digital Analytics weiterentwickeln, das durch Datenvolumen und -variabilität in steigendem Maß mit Big Data konfrontiert sein wird (Haberich 2013, S. 50).

Dadurch nimmt auch die Schwierigkeit bei der Verknüpfung der sehr unterschiedlichen strukturierten und unstrukturierten Datentypen zu. So wird die Betrachtung der Customer Journey, also die Beobachtung des Kaufprozesses eines Kunden von der Identifikation eines Bedürfnisses bis hin zum abgeschlossenen Kauf, schon allein dadurch erschwert, dass nicht alle Touchpoints eindeutig einem Konsumenten zuzuordnen sind. So werden zwar schon bei kleineren Aktionen im Web viele unterschiedliche Cookies gesetzt. Zum Beispiel führt das Betreten der Eingangsseite des Onlinehändlers Priceminister zur Speicherung von 44 Cookies unterschiedlicher Unternehmen, der Klick auf ein Produktangebot zu weiteren etwa 40 Cookies (Eudes 2014). Hier stellen sich aber zunehmend datenschutzrechtliche Fragen, zudem wachsen die Möglichkeiten, die Speicherung von Cookies zu unterdrücken. Eine Alternative zu Cookies stellen Fingerprintverfahren dar, bei denen Geräte auf Basis technischer Merkmale sowie der Softwarekonfiguration ein eindeutiges Profil erhalten.

Weitere Probleme ergeben sich aus der parallelen Nutzung verschiedener Geräte. So führt der Wechsel des Gerätes (zum Beispiel von der Erstinformation auf dem mobilen Endgerät hin zu einer ausführlicheren Beschäftigung am Desktop-Rechner oder Note-

book) innerhalb einer Customer Journey zu fragmentierten Daten, sofern die Daten beider Geräte nicht aufgrund eines Login-Prozesses klar einem User zuzuordnen sind.

Wenn es gelingt, die Daten innerhalb eines sowie zwischen verschiedenen Geräten einem Nutzerprofil zuzuordnen, ist eine wesentliche Voraussetzung geschaffen, um aus den entstandenen Big Data einen Mehrwert zu schaffen, indem sich das Konsumentenverhalten lückenloser und kanalübergreifend dokumentieren lässt.

Eine ähnliche Herausforderung stellt die Verbindung von Social-Media-Daten mit Informationen aus der Webanalyse dar. Profile in sozialen Netzwerken sind zunächst einmal nicht mit anderen Seiten, etwa denen eines E-Shops, verbunden. APIs (Application Programming Interfaces) ermöglichen die Kommunikation verschiedener Softwarekomponenten beziehungsweise verschiedener Webseiten. Dadurch können Webseitenbetreiber Zugang zu den Daten ihrer Nutzer aus sozialen Netzwerken erhalten, also etwa zu Informationen über soziale Objekte wie Interessen und Aktivitäten, die sich für Marketingzwecke nutzen lassen.

Insgesamt stehen Unternehmen bei der Nutzung dieser umfangreichen Datenmengen vor mehreren Herausforderungen: Zum einen sind bereits die Datenmengen einzelner Kanäle, etwa der Webanalyse, so umfangreich, dass viele Unternehmen sie noch nicht umfassend gewinnbringend einsetzen. Zum anderen wird das Potential der Verbindung verschiedener Datenquellen noch nicht voll ausgereizt, was unter anderem an der Schwierigkeit liegt, diese technisch miteinander zu verbinden. Für diese Herausforderungen stehen zunehmend Lösungen bereit, die allerdings auch zu Fragen des Datenschutzes in Konflikt stehen.

2.7.4 Umfassende Verhaltensanalyse im Rahmen der Customer Journey

Die differenzierte Analyse des Verhaltens auf Basis der Verfügbarkeit vielfältiger Interaktions- und Beobachtungsdaten ist als zentrale Chance von Big Data-Analysen zu sehen (Bachmann et al. 2014, S. 32–34). Auch im E-Commerce ist ein tiefes Verständnis des Kunden entscheidende Voraussetzung für den Unternehmenserfolg. Folglich ist es konsequent, möglichst umfassend Konsumentendaten zu erfassen und zueinander in Beziehung zu setzen, um Kundenverhalten bestmöglich vorhersagen zu können und sich ideal darauf einzustellen. Digitale Datenquellen gehen dabei in Menge und Differenziertheit über die Möglichkeiten klassischer Konsumenten- und Marktforschungsmethoden hinaus, was ihre Nutzung besonders attraktiv macht.

Noch immer stellt es für den E-Commerce eine Herausforderung dar, den Konsumenten so gut kennenzulernen, dass ihm zum richtigen Zeitpunkt das richtige Produkt angeboten werden kann. Ein guter Verkäufer eines lokalen Shops in der Nachbarschaft kann hier durchaus als Vorbild betrachtet werden. Er kennt seine Kunden von früheren Käufen, kann ihnen aus früherer Erfahrung und durch effizientes Nachfragen eine überschaubare Anzahl passgenauer Produktalternativen präsentieren, dazu für den Kunden nützliche Informationen zusammenstellen und nach erfolgtem Kauf auf weitere Produkte aufmerksam machen,



Abb. 2.43 Stufen der Customer Journey (Quelle: Eigene Darstellung auf Basis von Foscht und Swoboda 2011, S. 26)

die nicht nur zum gerade gekauften Produkt, sondern auch zu weiteren Gewohnheiten des Kunden passen. Wenn dieser hohe Grad an Personalisierung auch im E-Commerce erreicht werden könnte, hätte dies Auswirkung auf die Conversion. Bislang gelangen rund 90 % der Besucher eines Onlineshops nicht bis zum letzten Schritt in der Customer Journey, an denen sich die finale Frage nach dem Kauf des Produkts stellt. Und etwa die Hälfte derer, die bis zu diesem Punkt gelangen, brechen hier ab (Morys 2013, S. 372). Sicherlich liegt in diesen Zahlen eine gewisse Unschärfe, weil nicht beliebig genau abschätzbar ist, ob der Kauf eventuell auf einem anderen Kanal oder zu einem anderen Zeitpunkt abgeschlossen wird, dennoch scheinen viele Interessenten auf dem Weg bis zur finalen Kaufentscheidung im E-Commerce verlorenzugehen.

Konkrete Möglichkeiten, wie sich die Conversion durch die Nutzung von Big Data-Analysen steigern lässt, sollen im Folgenden im Rahmen eines Customer Journey-Modells diskutiert werden, das die einzelnen Stationen von der Identifikation eines Bedürfnisses bis hin zum Produktkauf beschreibt. Die einzelnen Stufen sind in Abb. 2.43 dargestellt.

2.7.4.1 Bedarfs-/Mangelerkennung

In der ersten Phase der Customer Journey geht es darum, dass ein Konsument den Bedarf für ein neues Produkt erkennt oder mit einem Mangel konfrontiert wird, der die Motivation erzeugt, sich mit dem Kauf eines Produktes zu beschäftigen.

Unspezifische Werbestrategien wie das beliebige Einblenden von Werbebannern auf Webseiten funktionieren zunehmend weniger. Konsumenten sind einem immensen Angebot oft gleichwertiger Produkte sowie einer Überflutung mit Werbereizen auf vielen Kanälen ausgesetzt. Die erste große Herausforderung in der Customer Journey besteht folglich darin, Aufmerksamkeit für Produkte des eigenen Unternehmens zu erzeugen. Stark personalisierte Empfehlungen haben hier eine höhere Chance, auf Interesse zu stoßen. Dabei kann es sich um Angebote handeln, an denen generelles oder aber auch situationsspezifisches Interesse besteht.

Als zentrale Basis für personalisierte Produktempfehlungen werden vom Nutzer selbst erzeugte Daten wie zuletzt betrachtete Produkte, bisherige Suchanfragen oder ähnliches

genutzt. Zudem lassen sich Vergleiche mit dem Such- und Kaufverhalten anderer Nutzer anstellen, die zu differenzierteren Empfehlungen sowie zu Cross-Selling führen können.

Des Weiteren bietet die Möglichkeit der Einbindung von Daten aus sozialen Netzwerken weitere Optionen der Personalisierung. So lassen sich Social-Media-Daten zu Interessen und Aktivitäten für Produktempfehlungen nutzen. Auch können Informationen aus sozialen Medien mit Shopmetriken, also Daten aus der Customer Journey innerhalb des Shops, kombiniert werden (Völcker 2013, S. 285). So lässt sich beispielsweise herausarbeiten, welche Produktkategorien verstärkt von Nutzern gekauft werden, die über ein Netzwerk wie etwa Facebook einen Shop betreten haben. Eine solche Information lässt sich beispielsweise nutzen, um auch anderen Usern, die aus diesen Netzwerken heraus den Shop betreten, ebenfalls Produkte aus dieser Kategorie anzubieten. Ebenfalls könnten Käufer von Produkten aus dieser Kategorie motiviert werden, den Kauf in dem entsprechenden Netzwerk zu teilen. Schließlich lässt sich durch die Verbindung dieser Daten auch das Targeting im Social-Media-Marketing verbessern. So kann Wissen aus Shopmetriken eingesetzt werden, um zielgerichtet Werbung auf Social-Media-Kanälen zu schalten.

Zunehmend ergeben sich auch Möglichkeiten, situativ Bedürfnisse oder Bedarfe von Konsumenten zu erfassen. So liegt beispielsweise nahe, dass Kleidung auch an Wetterbedingungen orientiert wird, weshalb Zalando eine wetterabhängige Outfitberatung ankündigte (Zimmer 2014, S. 16). Auf Basis einer Analyse des Kaufverhaltens in Abhängigkeit des Wetters ließ sich ableiten, dass eine Kombination der Wetterkonditionen sowie der Temperatur gute Prädiktoren für Auswahl und Kauf bestimmter Kleidungsstücke darstellen. Wetterdaten werden hierbei über eine API ausgelesen und sinnvoll mit Informationen zur Passung bestimmter Kleidungsstücke untereinander kombiniert. Wetterinformationen werden somit als Indikatoren genutzt, um bestimmte Bedürfnisse oder naheliegende Verhaltensweisen abzuleiten und darauf Kaufvorschläge aufzubauen.

Auch mobile Daten erlauben es, situativ Bedürfnisse zu schaffen oder zu antizipieren, wodurch sich Möglichkeiten ergeben, Online- und stationären Handel stärker miteinander zu verbinden. Casino und SAP gelang es beispielsweise mithilfe einer im stationären Handel genutzten Smartphone-App, intelligente Produktvorschläge zu generieren, die auf der Kaufhistorie des Kunden oder etwa seinem aktuellen Standort im Laden basieren. In den betroffenen Testmärkten konnte eine Umsatzsteigerung von über 10 Prozent beobachtet werden, wobei insbesondere hohes Cross-Selling-Potential aufgezeigt werden konnte (Bloching et al. 2012, S. 126).

Welche hohe Präzision die Nutzung mobiler Daten in Verbindung mit weiteren Datenquellen haben kann, zeigt ein Projekt des MIT Media Lab. Hierbei analysierte eine Forschergruppe Standortdaten von Mobiltelefonen in Kombination mit Wetterdaten und demografischen Daten im Umfeld eines bestimmten stationären Händlers und konnte damit dessen Verkaufszahlen an einem konkreten Tag schneller vorhersagen als der Händler selbst. Solche Daten lassen sich beispielsweise zur Steuerung der Lagerbestände oder des Personals nutzen (Kudyba und Kwatinetz 2014, S. 11 f.).

2.7.4.2 Suche

Die große Stärke bei der Suche nach Produkten und Informationen im Onlinehandel besteht im meist umfangreichen Produktangebot und der Systematik, mit der gesucht werden kann. Die Vielzahl an Produkten und Suchmöglichkeiten führt allerdings schnell zu einer Überforderung des Kunden, der sich bei vielen Produktaufen meist nicht auf eine aufwendige Informationssuche einlassen will. Der stationäre Handel ermöglicht durch einen guten Verkäufer eine schnellere Eingrenzung auf wenige relevante Produktalternativen. Die Analyse vielfältiger Datenquellen im E-Commerce bietet Ansätze, um die Onlinesuche zu optimieren.

Bislang wird die Reihenfolge der in einem Onlineshop dargestellten Produkte maßgeblich durch Kauf- und Klickverhalten des Kunden gesteuert, was in vielen Fällen eine unzureichende Ergebnisliste zur Folge hat (Grüger 2013, S. 264–266). Eine Einbindung weiterer Datenquellen verbessert das Ergebnis deutlich. So könnten Informationen zum Lagerbestand dazu genutzt werden, ausverkaufte Produkte oder solche mit langen Lieferzeiten nicht beziehungsweise eher am Ende der Liste anzuzeigen. Produkte, die in mehreren Varianten (z. B. Größen oder Farben) vorliegen, könnten intelligent behandelt werden, indem nicht alle einzelnen Ausprägungen untereinander in einer Ausgabeliste angezeigt werden, was den Überblick über die Suchergebnisse erschwert. Zusätzlich können weitere Kennzahlen wie etwa PIM-Daten zur Verfügbarkeit von Produktbildern zum jeweiligen Artikel oder Daten über die Retourenquote bei der Bestimmung der Reihenfolge in der Ergebnisliste aufgegriffen werden.

Eine weitere Maßschneidierung der Suchergebnisse kann sich etwa durch die systematische Nutzung von CRM-Daten ergeben. So lassen sich bei männlichen Kunden bei einer Suche nach Schuhen direkt Männerschuhe auf den vorderen Listenpositionen ausgeben.

Schließlich unterstützt eine an den verwendeten Browser und das jeweilige Gerät optimierte Ausgabe die Übersichtlichkeit der Suchergebnisse. So können etwa bei mobilem Zugriff die Ergebnislisten entsprechend abgekürzt werden.

2.7.4.3 Bewertung

Bei der Wertung verschiedener Produktalternativen als Vorbereitung der finalen Kaufentscheidung wählen Konsumenten unterschiedliche Strategien. Eine oft wichtige Rolle spielt der Produktpreis. Konsumenten stehen online inzwischen viele leicht zugängliche Möglichkeiten des Preisvergleichs zur Verfügung, die im E-Commerce bislang noch nicht umfassend auf Händlerseite genutzt werden (Grüger 2013, S. 268 f.). Eine kontinuierliche Sammlung von Wettbewerberpreisinformationen zum Beispiel durch die Beobachtung der Wettbewerber-Websites kann hier einen strategischen Vorteil bringen. Diese Informationen können für dynamisches Pricing verwendet werden, wodurch die eigenen Produkte bei Preisvergleichen attraktivere Plätze einnehmen.

Andere Nutzergruppen wählen andere Strategien bei der Bewertung unterschiedlicher Produktalternativen. Die tiefe Integration von Social-Media-Kanälen erlaubt auch hier eine umfassendere Beschreibung des Verhaltens von Nutzern in der Phase der Bewertung. Erfolgen beispielsweise an dieser Stelle in der Customer Journey verstärkt Wechsel zu

sozialen Medien, um dort die in Erwägung gezogenen Produkte genauer zu prüfen, lassen sich diese Daten zur Kundensegmentierung nutzen. Social-Media-affinen Kunden könnten dadurch künftig stärker personalisierte Social-Media-Informationen zu den Produkten in der engeren Wahl angeboten werden.

2.7.4.4 Kauf und Nachkaufphase

Einer der letzten Hinderungsgründe am Produktkauf im Onlineshop kann eine vermutete längere Lieferzeit darstellen, was durchaus zum Abbruch eines Kaufprozesses führen kann. Amazons Ansatz des sogenannten Anticipatory Shippings bietet hier Ansätze, Lieferzeiten noch weiter zu verkürzen (Spiegel et al. 2013). So ließ sich das Unternehmen ein Patent sichern, in dem aus Informationen zu früheren Käufen, Wunschlisten, Umtausch von Waren Wahrscheinlichkeiten für die bevorstehende Bestellung von Produkten berechnen lassen. Diese ermöglichen es, Produkte frühzeitig, also sogar vor der eigentlichen Bestellung, in nahegelegenen Versandcentern bereitzustellen. Auch hier schafft die Kombination unterschiedlicher Verhaltensdaten, die über Konsumenten erhoben werden, neue Wege, künftiges Verhalten zu antizipieren und die Logistik darauf abzustimmen.

Auch der Bezahlvorgang lässt sich durch die Einbindung verschiedener Daten optimieren. So ist es schon seit längerer Zeit üblich, die Kreditwürdigkeit eines Kunden auf Basis verschiedener Datenquellen zu optimieren. Dieses Kunden-Scoring wird dann beispielsweise dazu genutzt zu bestimmen, welche Zahlungsarten angeboten werden. Neben persönlichen Informationen zu Anschrift und Alter, bisherigen Einkäufen, dem Bezahlverhalten oder der bisherigen Retourenquote lassen sich künftig auch weitere Daten wie etwa aus sozialen Netzwerken nutzen. Je mehr Datenkanäle genutzt werden, desto präziser lässt sich das Verhalten eines Kunden durch den Vergleich mit einer bestehenden Kundenbasis prognostizieren (Grüger 2013, S. 262; zum Scoring bei der Bonitätsprüfung durch Kreditinstitute vgl. auch Abschn. 2.8.3.3; zur datenschutzrechtlichen Einschätzung Abschn. 3.1.9).

Nach dem Produktkauf schaffen Möglichkeiten der Bewertung im Onlineshop oder auch auf Social-Media-Kanälen eine Datenbasis für künftige Kaufprozesse. So können Produktbewertungen genutzt werden, um künftige Empfehlungen zu steuern oder die Positionen in der Ergebnisliste im Rahmen der Produktsuche anzupassen.

Durch die zunehmende Digitalisierung von Produkten wie etwa E-Books liegen immer mehr Nutzungsdaten nach dem Zeitpunkt des Kaufes vor. So ermöglichen E-Reader nicht nur eine Identifikation des Nutzers über ein Nutzerkonto, sondern auch die Aufzeichnung von Daten über Lesehäufigkeit oder gesetzte Lesezeichen. Daraus lässt sich noch detaillierteres Wissen über den jeweiligen Nutzer generieren, das über die bisherigen Informationen zu gekauften oder bewerteten Produkten deutlich hinausgeht.

Die Erfassung vielfältiger Datenpunkte in der Customer Journey, die teilweise auch aus unterschiedlichen Kanälen stammen können, schafft des Weiteren eine bessere Voraussetzung zur Bewertung von Werbekampagnen. Daten zum Produktkauf lassen sich mit Werbekontakten auf verschiedenen Kanälen (z. B. Social Media, mobil, Desktop-Rechner) verknüpfen. Die Wirksamkeit von Werbekampagnen wird dadurch auch über verschiedene

Kanäle hinweg besser bewertbar, sofern die Daten dem jeweiligen Konsumenten eindeutig zugeordnet werden können. Somit lassen sich auch die kanalspezifischen Maßnahmen präziser steuern.

2.7.5 Wie aus „Big Data“ „Smart Data“ wird

So umfangreich die für den E-Commerce nutzbaren Datenmengen und so vielfältig die sich daraus ergebenden Möglichkeiten sind, so schwierig gestaltet sich die konkrete Umsetzung und gewinnbringende Nutzung.

Insbesondere besteht die Gefahr, das Thema Big Data zu technikorientiert zu betrachten. So geben nach einer Befragung des Magazins „isreport“ 78 % der Unternehmen an, die Big Data-Strategie in die Verantwortung des IT-Bereichs legen zu wollen (Bachmann et al. 2014, S. 234). Dieser Ansatz greift allerdings zu kurz, weil dadurch wesentliche Perspektiven außer Acht gelassen werden.

Beim Einsatz von Big Data handelt es sich aus messtheoretischer Sicht um eine konsequente Umsetzung eines Multi-Trait-Multi-Method-Ansatzes, wie er von Campbell und Fiske (1959) aus dem Bereich der Sozialwissenschaften und der Psychologie vorgeschlagen wurde. Dieser zielt darauf ab, mit vielfältigen Methoden viele verschiedene Aspekte zum Beispiel des Verhaltens von Konsumenten zu erfassen. Bei sachgerechter Anwendung entstehen dadurch aufgrund vieler unterschiedlicher Arten von Informationen aus unterschiedlichen Quellen valide, reliable und differenzierte Insights in alle Facetten des Konsumentenverhaltens (Church und Dutta 2013, S. 25). Bloching et al. (2012, S. 73) vergleichen das Ergebnis der gewinnbringenden Nutzung von Big Data mit einer Marktaufnahme in HD-Qualität, in die mit hoher Schärfe gezoomt werden kann. Über ein differenziertes Bild des einzelnen Konsumenten wird so der Gesamtmarkt erschließbar, zudem lässt sich durch die Gesamtbetrachtung auch das Verhalten des einzelnen Konsumenten besser verstehen. Die Konsequenz daraus lautet aus Sicht der Autoren: „Marketers und Vertrieb werden nicht alles anders machen, aber vieles deutlich besser.“ (Bloching et al. 2012, S. 88). Big Data zielt somit nicht darauf ab, ganz neue Fragen zu beantworten. Vielmehr ermöglicht der smarte Einsatz großer Datenmengen differenziertere Einsichten, die mit bisherigen Methoden in diesem Ausmaß noch nicht erfass- und damit auch nicht nutzbar waren.

Um allerdings aus „Big Data“ „Smart Data“ zu schaffen und die sprichwörtlichen Nadeln im Heuhaufen finden zu können, sollten einige Grundätze beachtet werden. Neben der Betrachtung der Qualität der jeweiligen Datenbasis (siehe dazu Abschn. 2.6, Marktforschung) ist die Einbindung von Ansätzen aus den Sozialwissenschaften zielführend. So helfen Kenntnisse von Modellen des Konsumentenverhaltens, gezielt Daten für die Analyse auszuwerten oder auch korrelativ erhaltene Ergebnisse besser kausal erklären zu können. Zudem ermöglicht die Fokussierung auf die Customer Journey eine kanalübergreifende Betrachtung des Verhaltens, was den tatsächlichen aktuellen Nutzungsgewohnheiten der Konsumenten entspricht und sich durch die kombinierte Betrachtung vieler Datenkanäle mit Big Data-Ansätzen besser untersuchen lässt als mit früheren Methoden.

Wenn allerdings eine einseitige Konzentration auf technische Aspekte sowie das Erheben großer Datenmengen zum Selbstzweck vermieden werden soll, hat dies auch starke Auswirkungen für die Akteure im Unternehmen, die sich mit Big Data befassen. Neben technischem Know-how bei der Verbindung und Aufbereitung des komplexen strukturierten und unstrukturierten Datenmaterials sowie Expertise in der Analyse großen Datenmengen benötigt eine zielführende Big Data-Strategie im E-Commerce eine starke Einbindung der Fachabteilungen wie etwa des Marketings und Experten aus Markt- und Konsumentenforschung (Kudyba und Kwatinetz 2014, S. 13). In einem iterativen Prozess lassen sich dadurch einerseits die richtigen Fragen stellen, andererseits Analyseergebnisse besser verstehen und in konkrete E-Commerce-Maßnahmen übersetzen.

Durch die hohe Aufmerksamkeit, die neue Methoden oder Technologien häufig gewidmet wird, besteht die Gefahr, bisherige Daten, Standards oder Best Practices zu vernachlässigen. Stoever (2014, S. 42), CEO des Business Intelligence Dienstleisters Minubo, formuliert pointiert: „Doch statt sich Gedanken darüber zu machen, wie man Wetterdaten mit den eigenen Umsatzzahlen verknüpfen kann, sollte zunächst einmal geklärt werden, was Umsatz im eigenen Unternehmen eigentlich bedeutet.“ Der professionelle Umgang mit Big Data erfordert eine klare Priorisierung der wesentlichen Fragestellungen. Hier sollte stets mit Fragen begonnen werden, die auch in der Zeit vor Big Data bedeutsam waren: „Was sind die Kennzahlen, die wichtig sind für meinen Shop? Wie lautet die unternehmensweit verbindliche Definition dieser Kennzahlen? Welche Zahlen muss ich mir monatlich, welche wöchentlich, welche täglich ansehen, um die Übersicht über meine Geschäftsprozesse zu behalten und ihre Entwicklung effizient steuern zu können?“ (Stoever 2014, S. 42).

Bei allen Chancen, die sich durch Big Data ergeben, sollte beachtet werden, dass die Analyse bestehender, auch umfangreicher Datenmengen stets eine Momentaufnahme darstellt, also Aussagen zu Bedürfnissen und Verhaltensweisen von Konsumenten in Vergangenheit und Gegenwart macht. Ergebnisse ermöglichen zwar Vorhersagen über künftiges Verhalten, diese berücksichtigen aber mögliche größere Veränderungen in der Zukunft noch nicht (Berchtenbreiter 2013).

Insgesamt zeigen innovative Ansätze im E-Commerce auf Basis der kombinierten Analyse großer Datenmengen sehr deutlich die Potentiale von Big Data auf. Die Betrachtung verschiedener Datenkanäle und -arten wird dem Verhalten des heutigen Konsumenten gerecht und liefert tiefere Einblicke in sein Verhalten, als es mit früheren Ansätzen möglich war. Dadurch wird es möglich, sich noch besser auf Kundenbedürfnisse einzustellen und sich dadurch Wettbewerbsvorteile zu verschaffen. Big Data erweitert das methodische Repertoire für Optimierungen im E-Commerce und hilft dabei, Online- und Offline-Aspekte des Konsumentenverhaltens integriert zu betrachten. Entscheidend für den Erfolg sowie eine effiziente und effektive Nutzung von Big Data ist eine interdisziplinäre Herangehensweise an das Thema. Wenn technisches mit fachlichem und sozialwissenschaftlichem Know-how kombiniert wird, wird das Potential von Big Data in idealer Weise ausgeschöpft und mündet in E-Commerce-Anwendungen, die noch vor wenigen Jahren in sehr weiter Ferne zu sein schienen.

2.8 Big Data in der Kreditwirtschaft

Wilhelmus van Geenen, Werner Dorschel und Joachim Dorschel

2.8.1 IT in der Kreditwirtschaft

Banken und verbundene Finanzdienstleister gehören zu den ersten nicht-staatlichen Einrichtungen, die seit den 1950er-Jahren Informationstechnologie bei der Abwicklung ihrer Geschäftsprozesse einsetzten.

2.8.1.1 Abgrenzung

Die IT-Infrastruktur einer Bank hängt maßgeblich von deren Größe, dem Leistungspotfolio den durch das Institut bedienten nationalen und internationalen Märkten und dessen Einbindung in Verbundorganisationen ab. Die nachfolgenden Ausführungen beziehen sich auf eine deutsche Universalbank. In den für den deutschen Markt wichtigen Sparkassen und genossenschaftlichen Instituten (z. B. Volks- und Raiffeisenbanken) werden viele IT-Aufgaben durch Verbundorganisationen¹⁰ wahrgenommen, sodass sich die Frage der Nutzung von Big Data auch dort und nicht nur in den Instituten selbst stellt.

Eine besondere Rolle nehmen in der deutschen Bankenlandschaft auch die Landesbanken ein, die einerseits als Universalbanken am Markt auftreten, andererseits auch öffentliche Aufgaben haben und Serviceleistungen für Sparkassen erbringen.

2.8.1.2 Mainframe, Batch, Dialog und Multichannel

Die zu Beginn der Entwicklung der kommerziellen Datenverarbeitung verfügbaren zentralen Großrechnertechniken waren vor allem für die Abbildung von Geschäftsvorfällen mit hohem Repetitionsgrad geeignet (Moch 2011, S. 15). Zu den Kernanforderungen der Banken gehören bis heute die persistente Speicherung von Kunden- und Kontendaten sowie deren geschäftsvorfallsbezogene und periodische Fortschreibung. Hierbei handelte es sich um hochstandardisierte Prozesse, die in Batchläufen verarbeitet wurden und die mithin für eine Automatisierung durch Großrechner gut geeignet waren.

Heute sind die Produktionsprozesse in Banken in der Regel vollständig durch Informationstechnologie abgebildet. IT-Systeme in der Finanzbranche sind traditionell auf Massenverarbeitung auch großer Datenmengen ausgerichtet. Die Batchverarbeitung spielt hierbei nach wie vor eine große Rolle. Daher ist die Finanzwirtschaft bis heute ein Einsatzgebiet von Host- bzw. Mainframesystemen, die in der Lage sind, hohe Transaktionszahlen performant und verarbeitungssicher abzuwickeln, und die bis heute das Zentrum der Systemlandschaft einer Bank bilden. Ergänzt werden diese Prozesse im Retail und im Corporates-Bereich durch eine Vielzahl moderner Sachbearbeiterdialoge und Multikanal-Services.

¹⁰ Bei den Sparkassen sind dies die Finanz Informatik GmbH & Co. KG, bei den genossenschaftlichen Instituten die GAD eG und die FIDUCIA IT AG.

Die Datenverwaltung erfolgt in relationalen oder hierarchischen Datenbanken wie DB2 oder IMS-DB. Neben dem Mainframe existieren vielfältige dezentrale Systeme, wobei die eingesetzten Client-Server-Modelle, Netzwerktechniken und Serversysteme sich nach dem Stand im Lebenszyklus der jeweiligen Anwendung stark unterscheiden.

2.8.1.3 Legacy-Systeme und Standardisierung

Die IT-Systemlandschaft einer deutschen Universalbank ist historisch gewachsen. Es existieren funktionsmächtige Legacy-Systeme, die wesentliche Geschäftsprozesse der Bank abbilden und in ihrer Funktionalität tief in die Aufbauorganisation, Verarbeitungsstrukturen und Prozesse der Bank eingebunden sind. Daneben etablieren sich in zunehmendem Maße Standardapplikationen, die Funktionalitäten für spezifische Geschäftsbereiche oder Organisationseinheiten bereitstellen. Im Sinne der Modernisierung der IT-Infrastruktur, einer Verbesserung der Time-to-Market und einer langfristigen Senkung der Gesamtbetriebskosten sind die meisten deutschen Banken bestrebt, ihre IT-Landschaft im Sinne einer IT-Industrialisierung auf standardisierte Lösungen zu transformieren (Hüthig 2013, S. 8 ff.).

2.8.1.4 Core-Banking-Systeme und Fachanwendungen

Das Zentrum der Anwendungslandschaft einer Großbank bildet ein Core-Banking-System, das die Kernprozesse der Bank wie Kontenführung und Kundendatenverwaltung abbildet. Der deutsche Markt für Kernbankensysteme ist geprägt von Eigenentwicklungen, den Lösungen der Verbundorganisationen¹¹ und Standardanwendungen wie SAP Banking Services. Um das Kernbankensystem herum gruppieren sich differenzierte Anwendungen zur Abwicklung der universalbanktypischen Standardprozesse, die wiederum in ein unternehmensweites Rechnungs- und Meldewesen sowie das Risikomanagement einfließen müssen.

Zu den Fachanwendungen zählen Front- und Backoffice-Systeme zur Abbildung elementarer Tätigkeitsbereiche der Bank, etwa des Zahlungsverkehrs, des Kreditwesens, des Handels, der Abwicklung von Wertpapiergeschäften, des Risikomanagements und des Controlling und ergänzender Funktionen wie Vertrieb, Beratung und Multichannel-Services. Anders als in Industrie und Handel haben sich in der Bankenwelt bislang keine omnifunktionalen ERP-Systeme etabliert, die alle wesentlichen Geschäftsprozesse mit einer einzigen, branchen- und kundenspezifisch angepassten Unternehmenssoftware abbilden.

2.8.1.5 Datenverwaltung, IDV und Business Intelligence

Die heterogene Anwendungslandschaft hat zwangsläufig zu einer siloartigen Struktur der Datenverwaltung in den Transaktionssystemen geführt. Die Datenstrukturen sind traditionell an Konten und Depots orientiert. Anders als in Anwendungen des eCommerce (hierzu

¹¹ Die deutschen Sparkassen setzten die Anwendung „OSPlus“ der Finanz Informatik GmbH & Co. KG ein, die Volks- und Raiffeisenbanken „bank21“ der GAD eG oder „agree“ der FIDUCIA IT AG.

ausführlich in Abschn. 2.7) steht der Kunde als Individuum nicht im Zentrum der Datenverwaltung. Banksysteme kennen ihren Kunden in der Regel nur insoweit, als er an einem in dem jeweiligen System abgebildeten System beteiligt ist. Für eine auf den Kunden zentrierte zentrale Sicht sind Bankanwendungen typischerweise nicht ausgelegt.

Banken haben zunächst versucht, die systemimmanen fehlenden kundenbezogenen Auswertungsmöglichkeiten durch individuelle Datenverwaltung zu ersetzen. Bis heute sind in vielen Fachbereichen auf Excel-, Access- und anderen Office-Tools basierende und mit Hilfe von selbstgeschriebenen Makros erstellte Auswertungen und Analysen im Einsatz.

Um dem wachsenden Bedarf nach Anwendungs- und geschäftsbereichsübergreifenden Datenanalysen insbesondere für die Gesamtbanksteuerung Rechnung zu tragen, haben die meisten Kreditinstitute parallel und oft auch mit hoher Redundanz Data-Warehousing- und BI-Konzepte entwickelt. Diese sind in der Regel auf Basis relationaler Datenbanken realisiert und werden über individuelle Schnittstellen aus den Liefersystemen versorgt. Der Aufwand zur Sicherstellung von Datenqualität und Datenkonsistenz in und über die verschiedenen Data Warehouses und Systemdatenbanken ist allerdings erheblich, was den Druck zur Entflechtung und Neuorganisation der gewachsenen Datensammlungskonstrukte stetig steigert.

2.8.1.6 Aktuelle Herausforderungen

Auch ohne Big Data steht die Bank-IT heute vor Herausforderungen, welche die Ressourcen der internen IT und die finanziellen Mittel in hohem Maße binden:

- IT-Projekte sind heute zu einem wesentlichen Teil durch regulatorische Anforderungen induziert. Die Taktzahl, mit der Gesetzgeber und Aufsichtsbehörden die Banken mit neuen Anforderungen und Regelwerken konfrontieren, steigt stetig und bindet einen großen Teil der für Änderungsprojekte verfügbaren Mittel.
- Banken sind bestrebt, ihre Legacy-Systeme zumindest partiell zu ersetzen. Die Migration auf Standardsoftware und das hiermit verbundene Downsizing der Altsysteme erfordert aufwendige Transformationsprojekte. Dies gilt umso mehr, wenn, wie in der Praxis häufig, mit der Standardisierung ein Outsourcing der jeweiligen Anwendungen verbunden ist.
- Fusionen der Banken und der Verbundorganisationen machen aufwendige Harmonisierungs- und Konsolidierungsprozesse erforderlich, die durch die fehlende Standardisierung und die Heterogenität und Individualität der Anwendungslandschaften der jeweiligen Fusionspartner zusätzlich erschwert werden.
- Die Internationalisierung der Eigentümerstruktur kann mit einer Internationalisierung der aufsichtsrechtlichen Anforderungen einhergehen. Zusätzliche Anforderungen wie eine unternehmensweite Mehrwährungsfähigkeit greifen tief in die Systeme und Prozesse der Bank ein.

2.8.2 Big Data bewegt die Bank-IT

Die Durchdringung finanzwirtschaftlicher Unternehmen wie auch die großen Datenmengen, die dort verarbeitet werden, machen es nicht verwunderlich, dass diese Branche in der öffentlichen Debatte um das wirtschaftliche Potential von Big Data eine herausgehobene Rolle einnimmt. In der brancheninternen und publizistischen Wahrnehmung stechen zwei Aspekte besonders hervor: Die neu und intensiver empfundene Digitalisierung des Retail-Banking und die immer weiter um sich greifenden regulatorischen Anforderungen, welche stets neue Notwendigkeiten der Überwachung und Analyse schaffen.

2.8.2.1 Digitalisierung der Kundenbeziehung

Die deutschen Banken verfügen schon seit den Zeiten von BTX über Angebote zur elektronischen Nutzung ihrer Produkte – noch bevor sich nach Ende der sogenannten .com-Blase in den 1990er-Jahren erste Geschäftsmodelle der heutigen Internetwirtschaft etablierten. Gleichwohl ist seit einigen Jahren ein wachsender Druck erkennbar, die Digitalisierung der Kundenbeziehung zu verbreitern und zu vertiefen (vgl. hierzu etwa Hüthig 2014, 12 ff.). Die Argumentation lässt sich wie folgt zusammenfassen:

- Während das Retail-Geschäft in vielen Großbanken zu Boomzeiten des Investment-Banking unter Ertragsgesichtspunkten eine untergeordnete Rolle spielte, gewinnt es nach der durch die Finanzkrise erzwungenen, am Risikoabbau orientierten Neuausrichtung an Bedeutung. Durch die aktuelle Niedrigzinsphase wird dieser Effekt noch verstärkt.
- Die digitale Wirtschaft entdeckt Finanzdienstleistungen für sich. Dies gilt sowohl für etablierte Internet-Konzerne wie Ebay mit dem Online-Bezahlsystem PayPal, Apple mit dem mobilen Bezahlsystem Apple Pay oder Google mit der E-Geld-Anwendung Google Wallet als auch für eine wachsende Zahl von Startup-Unternehmen, sogenannte „Fintechs“, die mit immer neuen Produkten etwa im Bereich des Zahlungsverkehrs klassische Bankdienstleistungen überflüssig machen wollen (Wiebe 2014a).
- Kunden sind immer seltener bereit, für die Inanspruchnahme von Bankprodukten eine Filiale aufzusuchen. Das persönliche Beratungsgespräch verliert für die Kundenbindung an Bedeutung.
- Die Bindung der Kunden an ihre Hausbank nimmt ab. Kunden sind häufiger bereit, die Bank zu wechseln.
- Die Bereitschaft der Kunden nimmt zu, für alltägliche Finanzgeschäfte wie Bezahlvorgänge und Überweisungen Leistungen von Nicht-Banken in Anspruch zu nehmen.

Die Forderung, den Kunden in das Zentrum der Bank-IT zu stellen, bedeutet für die Kreditinstitute einen Paradigmenwechsel. Wie in Abschn. 2.8.1.5 dargestellt, sind Bank-Anwendungen an Konten orientiert. Die vorhandenen IDV- und Data Warehousing-Konzepte haben in Bezug auf den Kunden und die Möglichkeiten seiner individualisierten und interaktiven Ansprache nicht die Leistungsfähigkeit moderner E-Commerce-

Anwendungen. Die individualisierte Sicht auf den Kunden haben in einer Bank nicht die IT-Systeme sondern der Berater, der mit Hilfe unterstützender Tools die Daten der einzelnen Anwendungssysteme auswertet und sich über assoziatives Wissen ein ganzheitliches Bild des Kunden schafft.

Für eine kundenzentrierte Sicht sind neben einer Auswertung der zentralen Konto- und Depotsysteme eine Einbeziehung sämtlicher Erkenntnisquellen einschließlich aller vorhandenen Kanäle und deren Vernetzung erforderlich.

Banken stehen dabei vor dem Dilemma, in ihrem vertrieblichen Auftritt mit Unternehmen zu konkurrieren, bei denen die Auswertung und Kommerzialisierung von Kundendaten seit jeher zum Geschäftsmodell gehört, zugleich aber dem strengen deutschen Datenschutzrecht und dem Bankgeheimnis wie auch dem eigenen Anspruch an Seriosität und Diskretion unterworfen zu sein.

2.8.2.2 Transparenzanforderungen durch die Bankenaufsicht

Die Finanzwirtschaft unterlag ob ihrer gesamtwirtschaftlichen Bedeutung schon immer einer besonderen staatlichen Aufsicht. Infolge der Finanzkrise wurde diese noch einmal verschärft. Hervorzuheben sind in diesem Zusammenhang hier die folgenden Regelwerke:

- Richtlinien des Basler Ausschusses für Bankenaufsicht für effektive Risikodatenaggregation und Risikoberichterstattung (BCBS 239),
- Reformpaket des Basler Ausschusses für Bankenaufsicht zur bestehenden Bankenregulierung (Basel III),
- die überarbeitete Finanzmarktrichtlinie (MiFID II) sowie die korrespondierende Verordnung (MiFIR) der Europäischen Union,
- Verordnung über den außerbörslichen Handel mit Derivat-Produkten (EMIR) der Europäischen Union,
- der einheitliche europäische Aufsichtsmechanismus (SSM) und der EU-weite Banken-Stresstest,
- das Gesetz zur Vermeidung von Gefahren und Missbräuchen im Hochfrequenzhandel,
- der Dodd-Frank Wall Street Reform and Consumer Protection Act und Foreign Account Tax Compliance Act (FATCA) der Vereinigten Staaten von Amerika.

Banken müssen gegenüber Aufsicht und Prüfern in gesteigertem Maße auskunftsfähig sein und zugleich über ein jederzeit aktuelles und vollständiges Bild der eigenen Risikosituation verfügen.

Die Heterogenität der von den jeweiligen regulatorischen Anforderungen betroffenen Anwendungssysteme und die hierdurch bedingte siloartige und in Teilen hochredundante Datenhaltung machen es schwer, aus immer neuen Blickwinkeln und in immer neuen Verknüpfungen auf die eigenen Datenbestände zu blicken. Darüber hinaus sind vor allem im Risikocontrolling zunehmend Realtime-Analysen erforderlich. Auch Data Warehouses stoßen hier an ihre Grenzen. Big Data-Konzepte und -Methoden werden hier als ein Weg gesehen, die hinsichtlich der Datensammlung- und Datenhaltungskonzepte erforderlichen

Restrukturierungsmaßnahmen umzusetzen und die Voraussetzungen zu schaffen, den Anforderungen der Regulierung zu genügen und Aufwand und Kosten unter Kontrolle zu halten.

2.8.3 Einzelne Geschäftsbereiche

Ungeachtet der in Abschn. 2.8.2 dargestellten Entwicklungen werden Big Data-Anwendungen und Technologien erst langsam in die Systemlandschaft der Kreditinstitute integriert. Nachrichten über erfolgreich abgeschlossene Big Data-Projekte sind in diesem Bereich noch selten. Analysen und Thesen zum Thema Big Data im Bankbetrieb stammen im Wesentlichen von Anbietern entsprechender Lösungen,¹² Beratern und aus der fachwissenschaftlichen Literatur.

Um die tatsächlichen Potentiale von Big Data im Bankbetrieb auszuloten, ist es erforderlich, die wesentlichen Geschäfts- und Leistungsbereiche einer Bank einzeln zu betrachten.

Einige besonders wichtige Bereiche sind nachfolgend überblicksweise dargestellt.

2.8.3.1 Zahlungsverkehr

Die Abwicklung des Zahlungsverkehrs mit seinen Basisprozessen Überweisung und Lastschrift ist eine typische Bankdienstleistung.

Im Zahlungsverkehr zeigt sich deutlich, wie in der Bank-IT Veränderungs- und Beharrungskräfte gleichermaßen wirken. Einerseits wurde mit SEPA eine europaweite Normierung des Zahlungsverkehrs und mithin ein hoher Grad an Harmonisierung erreicht. Andererseits spielt der beleghafte Zahlungsverkehr trotz der vielfältig vorhandenen elektronischen Angebote nach wie vor eine wesentliche Rolle¹³.

Zahlungsverkehrssysteme werden in der Fachwelt nicht zu den prädestinierten Einsatzgebieten von Big Data gezählt. Allerdings legen stetig steigende Anforderungen an Datenvolumina und Verarbeitungsgeschwindigkeit, die Möglichkeiten eines Realtime-Monitoring sowie die Auswertung von Transaktionsdaten den Einsatz von Big Data-Technologien nahe.

Transaktionszahlen und Datenvolumina

Die SEPA-Einführung hat zu merklichen Veränderungen im Markt geführt. Die wesentlichen Beobachtungen sind:

- Konzentration der Abwicklungsdienstleister,
- Vervielfachung des Datenvolumens gegenüber den nationalen Formaten,

¹² Hervorzuheben ist hier insbesondere SAP, die mit HANA gezielt den internationalen Bankenmarkt adressiert.

¹³ Im Jahr 2013 wurden noch ca. 816 Mio. Überweisungen beleghaft eingereicht, vgl. hierzu Deutsche Bundesbank: Zahlungsverkehrs- und Wertpapierabwicklungsstatistiken der in Deutschland 2009–2013, Stand: Juli 2014, abrufbar unter bundesbank.de.

- Verkürzung der Ausführungsfristen und
- Reduzierung der Gebühren.

Viele Zahlungsverkehrsabwickler haben das mit der SEPA-Umstellung verbundene Risiko und die nicht unerheblichen Kosten gescheut und ihren Zahlungsverkehr ganz oder in Teilen ausgelagert. Dieser Trend ist auch weiter zu beobachten.

Die Konzentration auf Seiten der Abwicklungsdienstleister führt dazu, dass die Anzahl der Transaktionen für den einzelnen Dienstleister stetig steigt. Bis zu 100 Millionen Transaktionen an einem Tag müssen verarbeitet werden. Mit SEPA hat sich außerdem der Speicherbedarf gegenüber den bisherigen nationalen Formaten um den Faktor 3 bis 4 gesteigert. Der Speicherbedarf einer deutschen DTA-Datei mit 100.000 Transaktionen liegt bei ungefähr 35 MB, bei einer SEPA-Datei je nach Ausprägung zwischen 100 und 150 MB.

Die Vorhaltung der Daten für mögliche Nachfolgeprozesse wie Rückrufe und Rücklastschriften über mehrere Tage im operativen Datenbestand und die Archivierung zur Gewährleitung der gesetzlichen Vorhaltefristen stellen höchste Ansprüche an den Speicherbedarf in diesem Umfeld. Die Anforderungen an die Performance, bedingt durch die regulatorische Vorgabe der taggleichen Verarbeitung sind eine große Herausforderung im Massenzahlungsverkehr. Die Cut-Off-Zeiten der einzelnen Prozessbeteiligten führen zu sehr engen Verarbeitungszeitfenstern. Traditionelle Tagesendeverarbeitungen mit zahlreichen Batchprozessen und Reportauswertungen während einer stillgelegten Verarbeitung sind kaum mehr darstellbar. Die 24 × 7- und Multizeitzonen-Verarbeitung ist vor allem für international am Markt agierende Zahlungsverkehrsdiensleister unausweichlich.

In wie weit es sich bei der Zahlungsverkehrsabwicklung um Big Data handelt, mag zu diskutieren sein. Diese als „Big“ einzuordnen, ist jedenfalls unbestreitbar. Dementsprechend werden für Zahlungsverkehrssysteme Technologien eingesetzt, die dem Bereich Big Data zuzurechnen sind. Der Ausbau der horizontalen und vertikalen Skalierung ist ein gängiger Weg, der Anforderung zu genügen, immer mehr Transaktionen in kürzer Zeit zu verarbeiten.

Realtime-Monitoring

Zusätzliche technische Anforderungen stellen sich, wenn Zahlungsverkehrsanwendungen neben der Zahlungsabwicklung auch die eine Realtime-Beauskunftung erlauben und Analyseaufgaben übernehmen sollen.

Eine recht neue Anforderung an ist die Möglichkeit einer Realtime-Verfolgung von Zahlungsverkehrs-Transaktionen im End-to-End-Prozess. Dies erfordert ein Echtzeit-Monitoring in der Massentransaktionsabwicklung. Hier können der Einsatz von In Memory-Datenbanken und Hybridmodellen aus zeilen- und spaltenorientierter Datenhaltung Vorteile bieten.

Analyse von Transaktionsdaten

Die Transaktionsdaten des Zahlungsverkehrs haben große Aussagekraft über Vorlieben und Verhaltensweisen von Bankkunden. Die Analyse dieser Daten ist zur Unterstützung einer individualisierten Kundenansprache gut geeignet.

Ist der Bankkunde eine natürliche Person, verbietet der Zweckbindungsgrundsatz des Datenschutzrechts (vgl. Abschn. 3.1.1.2) entsprechende Auswertungen. Denkbar sind hier allein anonymisierte Analysen, etwa zur Primärdatenerhebung im Bereich der Marktforschung. In Rechtsordnungen mit weniger strengem Datenschutzrecht wird durchaus über personalisierte Analysen auf Basis von Zahlungsverkehrsdaten und Kreditkartenabrechnungen zur Vorhersage künftiger Lebenssituationen und Entwicklungen berichtet.

Geschäftliche Relevanz hat die Analyse des Zahlungsverkehrs vor allem für Unternehmenskunden. Zahlreiche Kreditinstitute arbeiten an Lösungen, die Beratung ihrer Kunden bei Themen wie der Absicherung von Währungsrisiken und der Liquiditätssteuerung durch eine gezielte Analyse des Zahlungsverkehrs zu unterstützen.

Inwieweit solche Anwendungen den Einsatz von Big Data-Technologien erforderlich machen, ist eine Frage des Einzelfalls. Anders als beim Realtime-Monitoring wird es für einzelne Abfragen häufig ausreichen, auf die historischen Daten zuzugreifen. Sollen dagegen Zahlungsverkehrsdaten mit Daten anderer Systeme verknüpft werden, bedarf es entsprechender Vernetzungen und Datenbanktechniken, um in angemessener Zeit zu verwertbaren Ergebnissen zu kommen.

2.8.3.2 Handel

Geschwindigkeit gehört zu den entscheidenden Faktoren im Handelsbereich. Dies gilt für die Durchführung des eigentlichen Handelsgeschäftes und für die Einschätzung der Auswirkungen eines Geschäftes auf die Risikoposition im Rahmen von Portfoliosimulationen und der Verarbeitung von Marktinformationen.

Hochfrequenzhandel

Ein besonders prominentes – weil auch umstrittenes – Beispiel für die Möglichkeiten von Big Data in der Finanzwirtschaft ist der Hochfrequenz- oder Algorithmushandel.

Beim Hochfrequenzhandel handelt es sich um einen vollständig automatisierten Handel, bei dem sehr leistungsfähige Banksysteme über Hochgeschwindigkeitsnetze mit den jeweiligen Börsensystemen verbunden sind. Über selbstlernende Algorithmen werden kleinste Arbitragekonstellationen über verschiedene Börsenplätze erkannt und durch Roboter-generierte Kauf- und Verkaufsorders ausgenutzt. Voraussetzung hierfür sind ein Echtzeit-Monitoring und eine Echtzeitverarbeitung der Börsenaktivitäten und der Marktnachrichten. Big Data-Methoden und -Technologien erlauben hier weitere Beschleunigung und eine Verbesserung der Entscheidungsalgorithmen und Analyseverfahren. Kritiker sehen in diesem Bereich allerdings systemische Risiken und Einfallstore für manipulative Eingriffe (Hofstetter 2014, S. 177 ff.).

Portfoliosimulation

Der Handlungs- und Entscheidungsspielraum der Händler wird durch vorab definierte Risikolimits begrenzt. Jeder Deal muss in seinen Auswirkungen auf die Risikostruktur des eigenen Handelsportfolios des Händlers und/oder eines übergeordneten Portfolios geprüft werden. Um kurzfristig entscheidungsfähig zu sein, muss der Händler die Möglichkeit haben, diese Veränderungen vor Handelsabschluss zu simulieren. Je schneller eine Simulation durchgeführt werden kann, umso aktiver kann der Händler am Markt agieren. Anforderungen für Simulationen in Echtzeit, vor allem auch auf übergeordnete Portfolios sind heute bereits Standard.

Big Data-Methoden und -Technologien bieten Banken hier eine Chance, effizient am Markt zu agieren, ohne durch eine nicht erkannte oder regelwidrig in Kauf genommene unangemessene Kumulation von Risiken schädliche Verwerfungen für das Institut oder gar das gesamtwirtschaftliche Gefüge zu verursachen.

Marktdatenversorgung

Die zeitnahe Verarbeitung von Marktdaten, wie z. B. aktuelle und historische Kurse oder Marktnachrichten wie Adhoc-Meldungen, ist die Voraussetzung einer zutreffenden Einschätzung der Gesamtmarktsituation. Aufgrund der Flut von Daten sind auch hier bereits Methoden und Techniken im Einsatz, die heute unter Big Data subsumiert werden.

Auch in diesem Bereich werden neue Verfahren erprobt um der zunehmenden Anzahl von Informationsdaten Herr zu werden. Neuartige Algorithmen sind nicht nur darauf ausgerichtet, die Informationen schnell zu verarbeiten, sondern filtern auch in Echtzeit aus der großen Menge unstrukturierter Textdaten genau die für den Händler entscheidenden Informationen heraus.

2.8.3.3 Kreditgeschäft

Die Kreditvergabe gehört zum Kernbereich des Bankgeschäfts, das Kreditausfallrisiko zu den essentiellen Risikoarten einer Bank. Ein funktionsfähiger Kreditmarkt ist eine zwingende Voraussetzung für eine intakte Volkswirtschaft. Vor diesem Hintergrund sind die hohen regulatorischen und bankinternen Anforderungen an das Kreditgeschäft nachvollziehbar. Die daraus resultierenden Prozesse sind aufwendig und kostenintensiv. Im Kreditgeschäft sind Big Data-gestützte Anwendungen naheliegend, da auf umfassende Datenanalysen gestützte Unternehmens- und Persönlichkeitsprofile häufig eine belastbare Bonitätsprüfung des Kunden erlauben als die subjektive Einschätzung des Beraters.

Strukturierung von Krediten und Kreditderivaten

Strukturierte Kreditprodukte waren einer der Auslöser der Finanzkrise. Die Intransparenz dieser Produkte und die hieraus folgende mangelhafte Risikoabschätzung haben es verhindert, dass Banken Risiken rechtzeitig erkannt und Gegenmaßnahmen eingeleitet haben.

Zur Vermeidung künftiger Fehlentwicklung wird eine verbesserte Informationsversorgung entlang der gesamten Verbriefungskette gefordert (Hamerle und Plank 2010, S. 27 ff.). Big Data bietet hier die Möglichkeit, relevante Informationen wie volkswirt-

schaftliche Daten oder Marktdaten in die Bewertung einfließen zu lassen und so in Echtzeit zu verlässlicheren Einschätzungen und Risikovorhersagen zu gelangen.

Individuelles Scoring auf Basis von Big Data

Das Scoring ist das Ergebnis der individuellen Kreditwürdigkeitsprüfung des Kunden. Es ist somit Ausdruck des potentiellen Kreditausfallrisikos und entscheidet letztendlich darüber, ob und ggf. zu welchen Konditionen die Kreditgewährung erfolgen kann. Das Scoring ist eine Bonitätsnote, die sich aus verschiedenen Merkmalen wie Beruf, Einkommen, Sicherheiten, Wohnort, familiäre Situation, Erfahrungen aus der Kundenbeziehung etc. ergibt. Eine wesentliche Erkenntnisquelle für die Bonitätsprüfung ist in Deutschland die Schufa-Auskunft.

Die Bank hat ein Interesse daran, möglichst viele Informationen über einen Kreditantragsteller zu erhalten. Dies gilt in besonderem Maße für (potentielle) Neukunden, für die im Gegensatz zu Bestandskunden keine Daten aus einer bestehenden Kundenbeziehung in den hauseigenen Systemen vorhanden sind.

Maschinelle Auswertungen öffentlich zugänglicher Daten und Information, die der Kunde der Bank zur Verfügung stellt, bieten in diesem Bereich neue Möglichkeiten, mit wenig Aufwand und großer Genauigkeit schnell zu einer verlässlichen Bewertung zu kommen. Über den „Digital Footprint“ des Kreditnehmers werden unterschiedlichste Daten aus sozialen Netzwerken wie z. B. Facebook und Xing (sog. „Social Scoring“), Onlineplattformen wie eBay und Amazon oder auch lokale Daten von den Endgeräten der Nutzer gesammelt und verarbeitet. Die ausgewerteten Daten sind zum einen solche, die für jeden frei im Internet zur Verfügung stehen wie auch Daten, zu denen der Kunde der Bank Zugang gewährt. Diese Daten können nicht nur dazu dienen, neue zusätzliche Informationen in die Ermittlung einfließen zu lassen, sondern auch um Widersprüche in den Bestandsdaten zu erkennen¹⁴.

Bonitätsratings auf Basis von Big Data, die sich auf natürliche Personen beziehen, sind in Deutschland vor allem durch rechtliche Bestimmungen beschränkt. Das Datenschutzrecht erlaubt es nur in begrenztem Umfang, Entscheidungen und Bewertungen allein auf maschinelle Prozesse zu stützen (vgl. §§ 6a, 28b BDSG, hierzu näher in Abschn. 3.1.9). Eine Einwilligung des Betroffenen in die Verarbeitung seiner Daten oder gar den Zugriff auf nicht öffentlich zugängliche Informationen (z. B. zugangsgesicherte Profildaten aus Sozialen Netzwerken und Web Shops, lokale Daten von PC's und mobilen Endgeräten) müssen den gesetzlichen Anforderungen genügen. Die Einwilligung muss freiwillig erteilt

¹⁴ Einer der bekanntesten Dienstleister, der ein Bonitätsscoring ausschließlich auf Basis von Big-Data-Analysen anbietet, ist das Hamburger Startup-Unternehmen Kreditech (kreditech.com). Das Scoring erfolgt ausschließlich auf Basis eines selbstentwickelten Algorithmus, der aus bis zu 10.000 online verfügbaren Indikatoren einen Kreditscore berechnet. Ein weiteres Beispiel ist das Startup-Unternehmen Kabbage, das Onlineshops mit Working Capital versorgt. Händler können Kabbage Zugang zu Kundenfeedbacks, Social Media Daten und weitere nicht öffentlich zugängliche Daten zur Verfügung stellen, um den Prozess der Kreditgewährung zu beschleunigen und/oder bessere Konditionen zu erhalten.

werden, was schon dann zweifelhaft ist, wenn diese von der Bank zur Voraussetzung für einen Vertragsabschluss gemacht wird. Die Einwilligung darf den Betroffenen auch nicht unangemessen benachteiligen (vgl. § 307 BGB). Eine unangemessene Benachteiligung liegt etwa vor, wenn die Einwilligungserklärung zu offen formuliert ist, dem Anbieter einen zu weitgehenden Zugriff auf die Privatsphäre des Betroffenen ermöglicht oder die Auswertungsmöglichkeiten sich zu weit von den Wertungen der Datenschutzgesetze entfernen.

Mit Blick auf die restriktive Haltung, die die Aufsichtsbehörden gegenüber entsprechenden Lösungen einnehmen sowie die öffentliche Kritik (Schulzki-Haddouti 2014) sollten Banken und Anbieter auf eine sorgfältige technische und juristische Ausgestaltung eines auf Big Data gestützten Scoring achten.

Das Internet als neue Kreditplattform

Mit der Möglichkeit, Scoringergebnisse und damit eine Prognose auf die Kreditwürdigkeit auch außerhalb klassischer Bankprozesse zu ermitteln, drängen über den Kanal Internet neue Player in den Kreditmarkt. Zwar handelt es sich hier in der Regel um den Markt der Kleinkonsumentenkredite zwischen 1000 € und 5000 € aber auch der Markt für Klein-selbstständige und Firmengründer mit Kreditvolumen bis 50.000 € wird hier bedient. Zu unterscheiden sind hier im Wesentlichen folgende Erscheinungsformen:

- Etablierte Banken, die das Internet als weiteren Vertriebskanal über ihre Online-Auftritte oder Vermittlungssportale nutzen.
- Online-Plattformen, die nach dem Peer-to-Peer-Prinzip agieren und Ratenkredite durch Privatpersonen an Privatpersonen oder andere Formen der Finanzierung (z. B. Crowd Funding) vermitteln.
- Neue Unternehmen, die sich z. B. auf Online-Kreditvergaben in Ländern ohne flächen-deckendes Privatkundescoring spezialisiert haben.

Allen Varianten gleich ist der Bedarf nach einer möglichst umfassenden Bonitätsprüfung des Kreditantragstellers in Echtzeit.

2.8.3.4 Gesamtbanksteuerung

Die Aufgabe der Gesamtbanksteuerung ist die übergreifende Steuerung der Bank und aller ihrer Geschäftsfelder hinsichtlich der strategischen Ziele, des Ertrag, des Wachstums, des Risikos und der gesetzlichen Meldepflichten. Dies führt zu einem umfangreichen Datenbedarf aus allen Bereichen der Bank und weiteren Datenquellen wie z. B. Markt-informationen.

Die Verfahren sind historisch gewachsen. Dies gilt auch für die Datenhaltung. Datenbanken pro Geschäftsbereich stellen eher die Regel als die Ausnahme dar. Im Gesamtbanksteuerungsumfeld kommen typischerweise Data Warehouse-Konzepte zum Einsatz. Inkonsistente Daten auf Gesamtbankebene und schwerfällige Analyseprozesse sind zwangsläufige Folgeerscheinungen. Gerade in der Finanzkrise, in der Geschwindigkeit

der Analyse, Korrektheit der Daten und Flexibilität hinsichtlich neuer Auswertungen von essentieller Bedeutung waren, haben die Mängel offenkundig werden lassen.

Die Gesamtbanksteuerung ist daher einer der Bereiche, die von den in Abschn. 2.8.2.2 beschriebenen Anforderungen am stärksten betroffen sind. Anbieter von Big Data-Anwendungen für Banken adressieren gezielt Aufgabenbereiche der Gesamtbanksteuerung¹⁵.

Spätestens unter den Anforderungen aus Basel III ist ein zentraler Datenhaushalt, der bereichsübergreifend sämtliche Unternehmensdaten in konsistenter Form bereitstellt, unverzichtbar für ein effizientes Risikomanagement. Die Aktualität des zentralen Datenbestandes und die Performance der Analyse-, Simulations- und Reportprozesse trägt maßgeblich dazu bei, mit welcher Güte fundierte Entscheidungen getroffen werden können. Aufgrund der sich ständig an das Umfeld anpassenden regulatorischen Anforderungen und der Notwendigkeit, ad hoc auf Krisensituationen reagieren zu können, ist auch weiterhin mit häufigen Änderungen in den Auswertungsanforderungen und damit verbunden an den Datenhaushalt zu rechnen. In der aktuell durch Datensilos geprägten heterogenen Dateninfrastruktur sind Änderungen und Erweiterungen oder gar Architekturwechsel im laufenden Betrieb nur mit hohem Aufwand und erheblichen Risiken zu bewerkstelligen. Die Herausforderung muss also darin bestehen, Big Data-Technologien in die bestehenden Infrastrukturen zu integrieren und nicht pauschal zu ersetzen.

2.8.3.5 Vertrieb und Multichannel Services

Der in Abschn. 2.8.2.1 dargestellte Trend der Digitalisierung der Kundenbeziehung legt es Banken nahe, sich in ihren vertrieblichen Aktivitäten gegenüber bestehenden und potentiellen Kunden auf datengestützte Profile und hieraus gewonnene Erkenntnisse zu stützen, und sich damit ähnlich aufzustellen wie die etablierten Player im Bereich des Electronic Commerce (hierzu näher Abschn. 2.7). Da Banken über den heute üblichen Multichannel-Ansatz in der Interaktion mit ihren Kunden in Gestalt von Debit- und Kreditkarten, Geldautomaten, SB-Terminals, PoS, Internet-Banking, Mobile Devices, Apps etc. schon diverse elektronische Schnittstellen zur Verfügung stehen, ist die technische Realisierung einer kundenindividualisierten Kommunikationsstrategie nicht schwierig.

Neben den rechtlichen Grenzen der Datennutzung (für Zahlungsverkehrsdaten vgl. Abschn. 2.8.3.1) stellt sich für Banken allerdings die Frage, inwieweit die Vorteile kundenindividualisierter Ansprache durch die Skepsis der Kunden gegenüber „Datenkraken“ aufgehoben werden. Es bedarf sorgfältiger Prüfung, ob und inwieweit Banken ihren Vorsprung gegenüber Internet-Konzernen in Bezug auf Diskretion und Seriosität, der trotz der durch die Finanzkrise ausgelösten Verwerfungen in der Wahrnehmung der Kunden vorhanden ist (vgl. etwa Hüthig 2014, S. 18), durch eine Übernahme von Vertriebsstrategien aus dem Bereich der Internettwirtschaft verspielen.

¹⁵ So war die SAP Liquidity Risk Management Lösung eine der ersten bankfachlichen Anwendungen der SAP auf Basis der Big Data-Plattform HANA.

Die Resistenz des Bankensektors und seiner Kunden gegenüber einer allumfassenden Digitalisierung zeigt sich am dauerhaften Fortbestand des Bargelds. Studien zufolge ist der Anteil der Barzahlungen im Einzelhandel gegenüber Kartenzahlungen zwar leicht rückläufig, überwiegt diese jedoch immer noch signifikant (Krüger und Seitz 2014, S. 18). Die Anonymität des Bargelds wird als einer der Gründe für dessen fortgesetzte Bedeutung im Zahlungsverkehr angeführt (Krüger und Seitz 2014, S. 67).

2.8.4 Big Data, Outsourcing und Cloud Computing

Kreditinstitute suchen seit Jahren verstärkt nach IT-Governance-Modellen, die signifikante Kostensenkungen ermöglichen. Neben den bereits erwähnten Maßnahmen zur Kostenreduktion

- Übergang zu Standardapplikationen,
- Downsizing der Mainframe-Infrakstruktur,
- Installation von Near- und Offshore Projekten

zeigt sich das IT-Outsourcing als weitere wesentliche Handlungsoption für die Optimierung der Wertschöpfungsketten und einer damit einhergehenden Kostendegression.

Ein IT-Outsourcing erfolgt bei komplexen Groß-IT-Strukturen in der Regel nur in Teilkomponenten. Dabei ist zu unterscheiden, inwieweit IT-Services und/oder die darauf aufbauenden kompletten Betriebsprozesse einem Outsourcing unterworfen werden.

Wir wollen an dieser Stelle die grundlegenden Fragen des IT-Outsourcings nicht weiter vertiefen. Hierzu wird auf die umfangreich verfügbare Fachliteratur verwiesen.

Für unser Thema Big Data in der Kreditwirtschaft ist das IT-Outsourcing in zweierlei Hinsicht von Interesse:

- Gefahr der Datendesintegration in Outsourcing-Projekten,
- Managed Services für Big Data in der Cloud.

2.8.4.1 Gefahr der Datendesintegration

Die Verlagerung von IT-Prozessen zu spezialisierten Dienstleistern wird zunehmend zum Standard für Bankgeschäftsprozesse, die zum einen geringes Differenzierungspotenzial für den Bankkunden und zum anderen nur geringe Deckungsbeiträge generieren. Typisch sind dafür Prozesse des beleghaften und beleglosen Zahlungsverkehrs, der Wertpapierorder-Abwicklung und Depotführung bis hin zu hochstandardisierten Core-Banking-Eigenschaften wie Kontoführung.

Outsourcing als Instrument zur Kostendegression wird je nach Größe des Kreditinstituts und seines Geschäftsmodells allerdings äußerst unterschiedlich modelliert.

Allen Ausprägungen eines IT-Outsourcing ist jedoch die Gefahr der Desintegration der ganzheitlichen Datensicht gemeinsam.

Dieser Aspekt war in der Vergangenheit nicht im Fokus der Planer und Architekten von IT-Outsourcing. Umso wichtiger ist es in Hinblick auf die veränderten Anforderungen an Big Data-Verfahren, angemessen für die Datenintegration Sorge zu tragen. Der Outsourcing-Service-Provider muss in die Überlegungen für eine Big Data Infrastruktur einbezogen werden. Gemeinhin werden in den gängigen Outsourcing-Projekten lediglich triviale Formate und Datenprotokolle festgelegt, die den Datenaustausch aus buchhalterischer und formaler Sicht regelt.

Mit der zunehmenden Virtualisierung von IT-Strukturen wird die Komplexität einer unternehmensweiten Datenbevorratung und -analyse eine schnell wachsende Herausforderung an die IT-Architekten.

2.8.4.2 Managed Services für Big Data in der Cloud

Eine gleiche Umkehrung des Ansatzes ist die Auslagerung von Big Data-Prozessen zu Anbietern von Managed-Services für Big Data. Dies ist insbesondere für alle die Unternehmen von Interesse, denen die Einführung von Big Data eine enorme Anstrengung zur Bereitstellung des notwendigen Skills und Infrastrukturen bedeuten würde.

Von großem Interesse ist dabei das Angebot von Managed-Services für Big Data nicht nur von den marktführenden IT-Spezialisten wie IBM & Co, sondern auch von den Internet-Innovatoren à la Google und Amazon. So bietet z. B. Amazon mit AWS Big Data einen Service für Infrastruktur und skalierbare Big Data-Verfahren in der Amazon-Cloud an.

Hier kommen nun für die Kreditinstitute fast schmerzhafte Effekte zusammen:

Diejenigen, die zunehmend in die Geschäfte der Kreditwirtschaft hineindrängen, sind gleichzeitig innovative Hightech-Anbieter für skalierbare Cloud-Anwendungen zur Bereitstellung, Analyse und Visualisierung von Big Data.

Der Innovationsvorsprung von Amazon, Google & Co. in der Virtualisierung von Cloud-Ressourcen und der darauf basierenden Bereitstellung von Big Data-Managementverfahren ist objektiv eine sehr große Herausforderung für die konservativen IT-Infrastrukturen der Kreditwirtschaft.

Setzt man nun das Cloud-basierte Big Data-Serviceangebot der Internetunternehmen ins Verhältnis zu den selbstgeschöpften Daten aus den eigenen kommerziellen und sozialen Prozessen, bleibt der Fantasie des Beobachters beliebig viel Raum.

2.8.5 Fazit

Fokussieren wir die Betrachtung primär auf die in der DACH-Region agierenden Kreditinstitute, sehen sich diese Finanzdienstleister einem enormen Druck zur Innovation ausgesetzt. Dieser Druck wird durch die historisch bedingten Hemmnisse im gewachsenen IT-Environment erschwert.

Der spätestens mit der Finanzkrise sichtbar gewordene Reformbedarf stellt enorme Anforderungen an die Elastizität und Leistungskraft der Bank-IT. Es treten die Internet- und

FinTech-Anbieter zunehmend ernsthaft in den Wettbewerb mit den Banken um ureigene Leistungsangebote für den Privat- und Firmenkunden.

Big Data zeigt sich dabei als ein unverzichtbares Element der Wettbewerbsstrategie. Ohne Big Data-Management ist eine dynamische und erfolgreiche Time to Market-Politik zukünftig nahezu ausgeschlossen.

Die Geschwindigkeit der Veränderungen im Angebots- und Bezahlssystem, die Kreditierung von Privattransaktionen und die hochpersonalisierte individuelle Kundenansprache gehören noch lange nicht zum täglichen Instrumentarium des Marktauftritts von Kreditinstituten. Die Einführung eines Big Data-Management kann jedoch nicht darauf warten, bis die notwendigen Modernisierungsarbeiten der historisch gewachsenen IT-Infrastrukturen abgeschlossen sind.

Hier sind dynamischere Umsetzungsstrategien zu entwickeln, die eine deutlich schnellere Einführung und Nutzung von Big Data sicherstellen.

2.9 Chancen und Herausforderungen von Big Data in der Industrie

Alphonse Stremler und Lothar März

2.9.1 Unternehmerische Ziele zur Erhöhung der Wertschöpfung

2.9.1.1 Anforderungen in Produktion und Logistik

Weltweit werden sich die Märkte immer stärker anpassen. Die Kunden erwarten bei Produkten und Leistungen individuelle Auswahlmöglichkeiten. Die Unternehmen agieren global. Das Geschäft ist an der Schnittstelle zum Kunden lokal, sogar kundenindividuell. Individualisierung neben Standard- und Massengeschäft erfordern sowohl auf der Produktseite als auch auf der Produktions- und Informationsmanagementseite neue Lösungen.

Dies führt in den nächsten Jahren zu konfigurierbaren, am Verkaufspunkt auf den einzelnen Kunden abgestimmte Lösungen. Produkte und Leistungen sind systematisch, modular aufgebaut. Die kundenindividuellen Lösungen erfordern andere Prozesse und Produktionsschritte als Standard- oder Massenprodukte.

Die klassischen Produktentwicklungs- und Produktionsstrategien unterscheiden zwischen Manufaktur und Serienproduktion. Diese Einteilung stimmt längst nicht mehr mit der realen unternehmerischen Wirklichkeit überein. Typische Beispiele für Hybridstrategien sind:

- Plattformstrategie der Automobilhersteller,
- Kundenindividuell konfigurierte LKWs,
- Taylormade Anzüge und Kleider in den oberen Segmenten der Bekleidungsindustrie,
- Maschinen- und Anlagenbau.

Die Produkt- und Leistungskonfiguration erfolgt in Zukunft in Echtzeit. Am Point of Sale gilt es, die Kundenerwartungen hinsichtlich Leistung und Lieferumfang zu erfüllen. Durch die Simulation von Szenarien und Echtzeitplanung wird der verbindliche Liefertermin ermittelt und zugesagt. Kurze, verbindliche Reaktionszeiten werden in Zukunft Bestandteil der Leistung und des Kundeservice sein. Diese sind, im Gegensatz zu heute, durch Datenmodelle abgesichert. Flexible Anpassungen sind jederzeit durch Simulation möglich.

Das bedeutet, dass sich die Unternehmen durch die Fortschritte in der Echtzeit-Informationstechnologie neu aufstellen werden und die immer größere Komplexität der Produkt- und der integrierten Prozesswelt wird schrittweise beherrschbar.

2.9.2 Effizienzsteigerung durch integriertes Realtime-Informations- und Datenmanagement in der integrierten Supply Chain

Die wesentlichen Effizienzsteigerungen einer Supply Chain sind nicht alleine in der Produktions- und Beschaffungslogistik zu erzielen. Die Eingangsinformationen an der Schnittstelle zum Kunden sowie die Leistungs- und Produktkonfigurationen definieren die Leistungen über die einzelnen Wertschöpfungsstufen und „ziehen“ über den Pull-Effekt die Supply Chain. Die Beherrschung des Datenmanagements ist in Zukunft ausschlaggebend für die Kosteneffizienz und die Mittelbindung in der Lieferkette.

Aufbauend auf integrierte Realtime-Informationsmanagement-Konzepte wird in den nächsten Jahren die Online-Produktkonfiguration zum Standard für Best-in-Class-Unternehmen werden. Komplexe, innovative Produkte, vom LKW bis hin zum Maschinenbau, werden online konfiguriert werden, mit gleichzeitiger Darstellung der Kosten- und Preis-sensitivität bei der gemeinsamen Ausarbeitung der Leistungen und Lieferumfänge an der Schnittstelle zum Kunden.

Auch die Verfügbarkeitsprüfungen von Produkten, Aggregaten und Komponenten entlang der gesamten Supply Chain wird in Echtzeit erfolgen. Durch diese Daten-Management-Technologien kann direkt ein verbindlicher Liefertermin festgelegt werden, da die Abstimmung durch Planung und Simulation online mit den Planungen der Werke und Zulieferanten erfolgt. (s. hierzu Abschn. 2.4)

Die Unternehmen werden wesentliche Investitionen in neue Methoden zur Steigerung der Flexibilität bei der Anpassung der Ressourcen und zur Beherrschung der Produktkomplexität tätigen. Innovative Planungssysteme und Datenmanagement-Methoden werden die gesamte Industrie revolutionieren.

Entscheidend für einen Spaltenplatz im Wettbewerb wird auch in Zukunft sein, die Leistungsreserven im Unternehmen hinsichtlich Reaktionsfähigkeit und Geschwindigkeit sowie Best-in-Class Performance in der Produktion, beim Lieferservice und der Innovation zu realisieren – und das bei niedrigem Working Capital und optimalem Ressourceneinsatz.

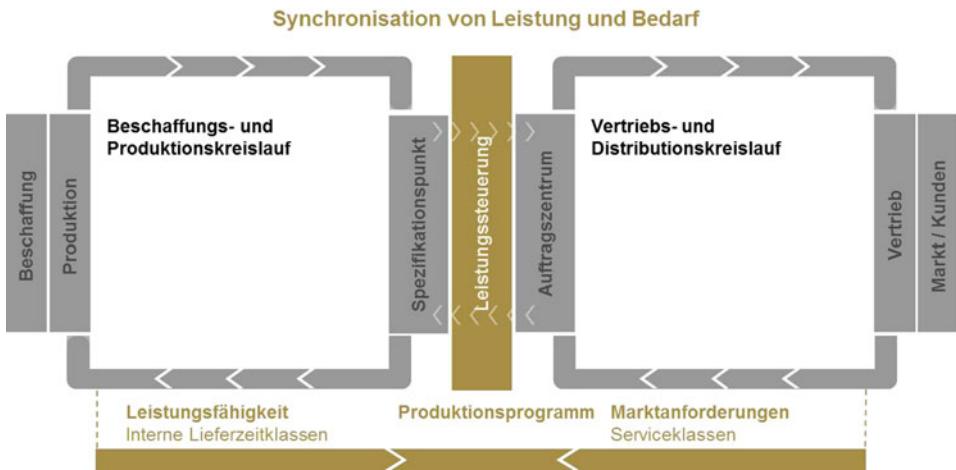


Abb. 2.44 Zweikreis-Modell der Produktion

Wertschöpfung bedeutet Leistungserstellung zu optimalen Kosten und zu einem optimalen Preis. Gesteigert wird diese durch die Erzielung eines maximalen Kapitalumschlags realisierbar durch Echtzeit Steuerung über alle Stufen der Wertschöpfungskette.

2.9.3 Ein Modell der Produktion

Im täglichen Betriebsablauf sind die Zielsetzungen der einzelnen Abteilungen eines Unternehmens oft divergierend. Basierend auf dem Pull-Prinzip, beschreibt das Zweikreis-Modell der Produktion die optimale Verzahnung der Regelkreise „Beschaffung und Produktion“ sowie „Vertrieb und Distribution“ (Abb. 2.44). Gegenläufige Anforderungen, die sich aus der Unstetigkeit des Marktes und der angestrebten Stetigkeit der Produktion ergeben, werden harmonisiert und synchronisiert. Eine über die Zeit dargestellte Bedarfsstruktur löst als realer Marktbedarf den „Pull“ auf Beschaffung und Produktion aus.

2.9.4 Leistungssteuerung in Echtzeit für maximale Reaktivität der Supply Chain

Marktanforderungen und Reaktivität der Lieferkette über mehrere Stufen stellen ein Spannungsfeld dar, das gezielt über die Leistungssteuerung ausgeglichen werden muss. Die Planung und Steuerung ist von entscheidender Bedeutung, um die Leistung zu maximieren und das Gesamtsystem effizienter zu gestalten. Mit innovativen Systemen ist die Steuerung der Produktion sogar in Echtzeit möglich.

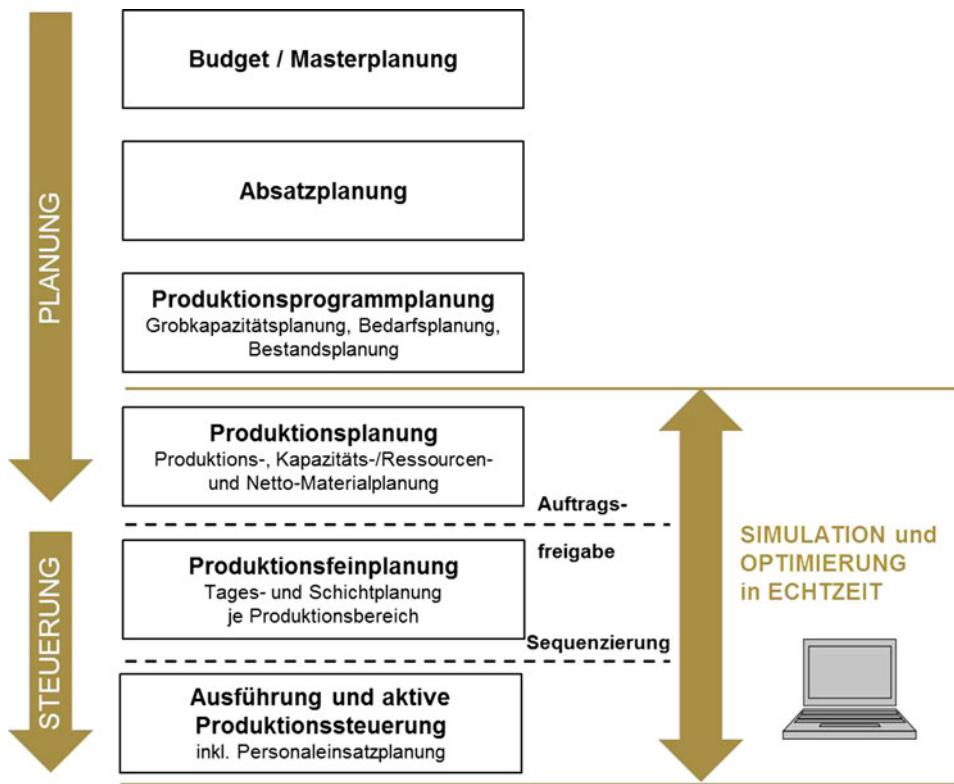


Abb. 2.45 Planung, Simulation und Optimierung in Echtzeit

Der Fokus des Supply Chain Engineering sollte dabei auf einer hohen Dynamik und Transparenz in der Abbildung und Bewertung von Änderungen in Echtzeit liegen:

- Angebot und Nachfrage werden kontinuierlich aufeinander abgestimmt.
- Die Planung bereitet Operations auf bekannte und wahrscheinliche Ereignisse vor.
- Die Steuerung kann bei Eintritt von unerwarteten Ereignissen sofort reagieren.

Bisher nicht gekannte Reaktionsfähigkeit kann darüber hinaus mit der Installation eines Echtzeit-Planungs- und Simulationssystem auf Basis von Ist-Daten zur sekundenschaffenden Planung, Steuerung sowie Optimierung des Ressourceneinsatzes bis auf Shopfloor-Ebene erzielt werden (Abb. 2.45). Selbst in hochkomplexen Systemen ist dadurch eine kontinuierlich im Leistungsmaximum gesteuerte Produktion möglich.

Zur Sicherstellung der Leistungssteuerung exakt am Marktbedarf wird ein Realtime-Führungsmonitoring mit hoher Informationsqualität installiert. Damit können Leistungs- und Bestandsabweichungen gegenüber dem realen Marktbedarf bewertet, online korrigiert und in kürzester Zeit angepasst werden.

2.9.5 Ebenen und Stufen der Planung

Die ERP-Plattform bildet nach wie vor das Herzstück jedes Industrieunternehmens. Durch eine integrierte und effiziente Planung und Steuerung der Produktionsaufträge nach Kundenbedarf gelingt es den Unternehmen, ihren Kunden exzellenten Lieferservice zu bieten und gleichzeitig Kosten und Kapitalumschlag zu optimieren.

Täglich ist der Planer jedoch mit kurzfristigen Änderungen der Auftragssituation oder Störungen im Produktionsablauf konfrontiert und soll das scheinbar Unmögliche möglich machen. Unerwartete Kundenaufträge mit hoher Priorität, kurzfristige Stornierungen, der ungeplante Ausfall einer Maschine oder eines Mitarbeiters oder auch Fehlteile führen in der Regel zu erheblichen Mehraufwendungen. Zusatzkosten durch Umrüstungen, Überstunden sowie Sondertransporte und negative Auswirkungen auf die Lieferperformance entstehen.

Optimierungspotenzial durch laufend aktualisierte Reihenfolgeplanung realisieren
 Die Erfahrung zeigt, dass es in vielen Unternehmen ein erhebliches Optimierungspotenzial gibt, das durch eine laufend an die aktuelle Situation angepasste Reihenfolgeplanung genutzt werden könnte. Ab einer gewissen Komplexität der Lieferketten, einer zunehmenden Produkt- und Variantenvielfalt und immer größeren Anforderungen an Flexibilität und Servicegrad, kann dieses Potenzial nur mit entsprechenden, innovativen Softwaresystemen realisiert werden (Abb. 2.46).

Eine besonders komfortable Bedienung, sekundenschnelle Erstellung von Szenarien sowie hohe Transparenz für den Planer über den aktuellen Belegungs- und Terminzustand

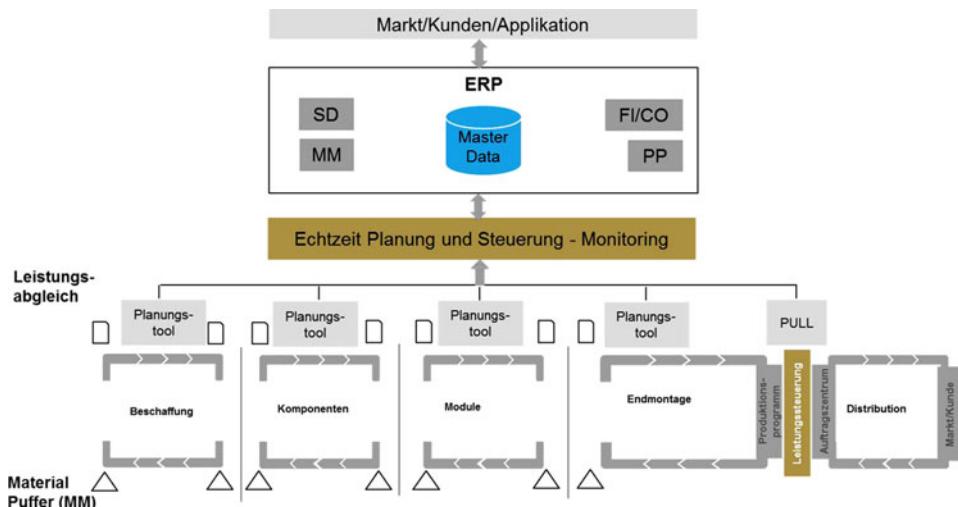


Abb. 2.46 Reihenfolgeplanung in Echtzeit

in der Produktion bietet hier eine kundenspezifisch angepasste Simulations- und Optimierungssoftware.

Die Simulation berücksichtigt zuvor definierte, arbeitsvorgangsbezogene Prioritätsregeln, Vorranggraphen sowie durch Ressourcen, Logistik und Kapazitäten gegebene Randbedingungen und Abhängigkeiten, wie sie dem Verständnis und der Anschauung des Produktionsplaners entsprechen. Dies ermöglicht dem Planer, Maschinenbelegungspläne oder z. B. den Einsatz von Mitarbeitern auf Machbarkeit zu überprüfen und alternative Szenarien fast spielerisch zu vergleichen.

Darstellung von Engpässen und Unterauslastung

Engpässe oder auch die Unterdeckung von Ressourcen werden vom System dargestellt und die Produktion nahe am optimalen Betriebszustand gefahren. Ein innovatives System ist dabei so performant, dass auch hochgradig komplexe Systeme interaktiv und ohne Wartezeiten, in Sekundenschelle simuliert und optimiert werden können.

Mit der Unterstützung durch diese Planungsassistenten sind wir auf dem Weg in den Aufbau von intelligenten Produktionssystemen. Jede Veränderung der Aufträge und jede neue Anforderung wirkt bei den derzeitigen Systemen wie eine zuerst nicht einmal wahrgenommene Störung. Durch einen systembasierten Ansatz, der sowohl deterministisch als auch stochastisch die neuen Anforderungen vorwärts denkt und Wirkzusammenhänge bei gegenseitigen Beeinflussungen und Abhängigkeiten berücksichtigt, wird ein intelligentes, proaktives System geschaffen. Dieses ermöglicht der Organisation die Wertschöpfung, sich immer wieder an die neuen, jeweils vom Markt definierten Bedingungen anzupassen. Die Angleichung erfolgt schrittweise und systembasiert. Wie ein menschliches Hirn sich immer wieder anpasst, richten sich die Organisation und technologische Prozesse an den neuen Gegebenheiten aus. Dies erfolgt in einem kontinuierlichen Prozess. Das ist der große Durchbruch: Intelligente Produktionssysteme.

Eine an die konkreten Bedürfnisse des Kunden angepasste PPS-Add-on-Lösung kann als Client-Server-Applikation im internen Netzwerk integriert und einfach mittels eines normalen Browsers, wie zum Beispiel dem Internet Explorer, bedient werden. Auf diese Art ist eine Einbindung in die operative Planung durch Datenübernahme aus z. B. SAP, Navision, Oracle und anderen Datenbanken schnell realisierbar. Software- und standortunabhängig kann der Planer von verschiedenen Terminals oder auch von mobilen Geräten auf den Planungsassistenten zugreifen.

Die Zukunft der Produktionsplanung und -Steuerung wird bereitet durch Echtzeitplanungs-Systeme. Diese innovativen Systeme schaffen die Plattform für eine verbesserte und effizientere Planung, die Optimierung des Ressourceneinsatzes sowie eine wesentlich erhöhte Reaktionsfähigkeit, die insgesamt zu einer Stärkung des Unternehmens im Wettbewerb führen. In den nachfolgenden Kapiteln werden die Aspekte zur Datenhaltung und Manipulation sowie die Erfolgsfaktoren von Big Data aufgezeigt.

2.9.6 Daten als Schlüsselfaktor des unternehmerischen Erfolges

2.9.6.1 Kundenindividuelle Produkte und Leistungen konfigurieren

Der zielgerichtete Umgang mit Markt- und Produktdaten ist der entscheidende Hebel für Unternehmen, die Innovationsführerschaft halten oder anstreben. Durch die kontinuierliche Analyse und Betrachtung der technologischen und produkttechnischen Anforderungen in der Anwendung (Monitoring/Scanning) ist zu prüfen, ob die Komponenten oder das Kernprodukt von neuen Anforderungen betroffen sind bzw. ob die Weiterentwicklungen ausreichend sind, um Wettbewerbsvorteile zu erzielen. Genauso ist die Frage ständig zu stellen, ob nicht Neuentwicklungen angestoßen werden müssen, um die Wettbewerbsfähigkeit zu sichern.

Die Zusammenführung von Markt-, Produkt- und Technologieinformationen stellt hohe Anforderungen an die Erfassung, Aufbereitung und Verknüpfung von Daten aus unterschiedlichen Fakultäten. Darauf aufbauend kann ein Technologie-Roadmap aufsetzen, welches durch definierte Quality Gates die Entwicklungen verfolgt und durch ständigen Abgleich von Technologiekalender und Marktentwicklung diejenigen Programme filtert, die letztendlich als Innovation am Markt bestehen können.

Die zunehmend kundenindividuelle Produktgestaltung steht im krassen Widerspruch zu der kostengetriebenen Forderung in Produktion und Logistik, die Varianten von Prozessen und Bauteilen zu begrenzen. Die Vielfalt an angebotenen Leistungen kann durch das Ausschöpfen von Synergien und der Einführung von einheitlichen Modulbaukästen reduziert werden. Dabei spielt das Datenmanagement eine bestimmende Rolle, denn Neuentwicklungen sind in die bestehende Applikationslandschaft zu integrieren. Dazu ist zu prüfen, inwieweit die Anforderungen des Neuproduktes durch die standardisierten Kern- und Komponentenfunktionalitäten vorhandener Produktbaureihen und weitergehender Applikationsentwicklung erfüllt werden können. Die Variabilität des Leistungsspektrums sollte weitgehend durch Konfiguration der Grundmodule und einer im Wertschöpfungsprozess späten kundenindividuellen Anpassung erfolgen. Die Beherrschung dieser Komplexität setzt die Beherrschung der Dateninterpretation über Produktkonfiguration und -varianz voraus.

2.9.6.2 Transparenz schaffen

In Produktionsunternehmen liegen Stamm-, Plan- und Bewegungsdaten informationstechnisch vor. Damit stehen sie grundsätzlich zur Darstellung und Weiterverarbeitung dem Unternehmen zur Verfügung. Um Transparenz über die aktuelle Situation und zu erwartende Entwicklungen zu erhalten, sind die vorhandenen Daten in einer Form zu erheben und zu verarbeiten, die eine Information transportiert: Besteht Handlungsbedarf oder bewegt sich das Unternehmen innerhalb der Planvorgaben? Dazu muss zunächst klar sein, wann das Unternehmen seinen optimalen Betriebspunkt erreicht hat und anhand welcher Kennzahlen dies abzulesen ist. Erst mit der Kenntnis über den zulässigen Betriebsbereich ist eine Aussage möglich, ob Korrekturmaßnahmen zu ergreifen sind oder nicht. Der wirtschaftlich ideale Betriebsbereich unterliegt den gegenläufigen Zielsetzungen von

Produktivität und niedrigen Beständen einerseits sowie kurzen Durchlaufzeiten und hoher Termintreue andererseits. Einen Einfluss auf den optimalen Betriebspunkt von Produktion und Logistik haben insbesondere die Produktionsstruktur, die Auftragsabwicklungsprozesse, die Losgrößen und die Planungs- und Steuerungsverfahren. Erst der Vergleich zwischen aktuellem und zulässigem Zustand der Produktion liefert die relevante Information und Transparenz, um Entscheidungen treffen zu können.

Zudem verkürzt die Verfügbarkeit von Informationen Such- und Prozesszeiten. Ein Beispiel ist der Entwicklungsprozess, bei dem die generierten Daten in den Fachabteilungen von Forschung und Entwicklung, der Arbeitsvorbereitung und der Produktionsplanung in Form einer integrierten Datenhaltung kontinuierlich aktualisiert und zugänglich gemacht werden können. Dadurch erhalten die involvierten Planungsinstanzen Rückmeldungen über den Planungsstatus und können Einflüsse auf die eigenen Planungen antizipieren.

Grundsätzlich ist die Verfügbarkeit von Wissen in Unternehmen von elementarer Bedeutung. Erfolgreiche Unternehmen verstehen es, Wissen organisational zu verankern. Wissen basiert auf Informationen, die als Daten gespeichert werden. Die Interpretationsfähigkeit von Daten gibt den Daten eine Bedeutung und wird somit eine Information für uns. Wissen entsteht dadurch, dass wir Informationen verarbeiten, Relationen zwischen Informationen herstellen und die gewonnenen Erkenntnisse speichern. Zur Verankerung von Wissen in einem Unternehmen ist somit die Erhebung von transparenten Daten eine der Grundvoraussetzungen. Transparent bezeichnet in diesem Zusammenhang, dass die Daten korrekt, vollständig und eindeutig sind. Transparente Daten lassen sich interpretieren und zu Informationen verknüpfen. Die Informationen wiederum können als Erfahrungswerte dem Unternehmen zugänglich gemacht werden, womit institutionalisiertes Wissen entsteht. Wissen ist Macht und für ein Unternehmen ein Wettbewerbsvorteil.

2.9.6.3 Reaktionsfähigkeit erhöhen

Die in Abschn. 2.9.1.1 dargestellten Trends in Richtung volatiler Märkte erschwert die Planbarkeit in den kurz- und mittelfristigen Zeithorizonten. Es wird daher immer wichtiger, die Abstimmungszyklen zwischen den Anforderungen der Auftragsabwicklung und der Leistungserbringung auf jeder Leistungsstufe der Produktion und Logistik zu verkürzen. Voraussetzung hierfür ist es, Status und Änderungen der Auftragslast als auch die der eigenen Organisation zeitnah zu kennen. Daten müssen transparent und quasi in Echtzeit verfügbar sein.

Die Verfügbarkeit von aktuellen Daten ist somit die Voraussetzung, um die richtigen Entscheidungen auf Basis der vorliegenden Informationen zu treffen. Nicht mehr gültige bzw. nicht verfügbare Informationen können zu Entscheidungen führen, die möglicherweise bei Kenntnis der tatsächlichen Sachlage nicht getroffen worden wären. Zudem ist eine schnelle Reaktion erst möglich, wenn Änderungen von Bestands-, Belegungs- und Verfügbarkeitsdaten zum Zeitpunkt der Entstehung ohne Latenzen signalisiert werden.

Echtzeitdaten erlauben eine direkte Reaktion auf vorliegenden Gegebenheiten. Dabei geht es nicht nur darum, potenzielle Engpässe in der Auftragsabwicklung zu detektie-

ren, sondern auch, um bei Leistungsverlusten aufgrund von Unterauslastungen oder bei überhöhten Bestandsentwicklungen rechtzeitig entgegen zu steuern. Unternehmen, die es verstehen, jederzeit auf ihre aktuellen Daten zugreifen zu können und diese zur Analyse und Entscheidungsfindung zu nutzen, verbessern damit nachhaltig ihre Leistung und ihren Service.

2.9.6.4 Entscheidungen durch Lösungsvorschläge unterstützen

Die Bereitstellung von Status-Informationen über den aktuellen Betriebspunkt hilft, um einen Handlungsbedarf zu identifizieren. Allerdings ist damit noch ungeklärt, in welcher Form angemessen auf die Planabweichung zu reagieren ist. Werden die Änderungen die gewünschten positiven Effekte aufweisen oder haben sie negative Auswirkungen auf andere Bereiche? Wird beispielsweise eine Kapazitätserhöhung an Engpassressourcen den Gesamtdurchfluss erhöhen oder wird dadurch nur ein nächster Engpass sichtbar? Oder wie wirkt sich das Vorziehen eines Auftrages auf die Liefertermintreue der nachfolgenden Aufträge aus?

Damit der Planer die richtigen Entscheidungen treffen kann, muss er die Auswirkungen von Maßnahmen auf die Gesamtsituation abschätzen können. Die Komplexität der Wirkzusammenhänge von Produktionssystemen lassen die Auswirkungen von Entscheidungen allerdings schwer vorhersagen. Vor diesem Hintergrund bieten sich speziell Expertensysteme an, die anwenderspezifisches Wissen über ein begrenztes Fachgebiet speichern und die Schlussfolgerungsfähigkeit von Planern in begrenzten Aufgabenfeldern rekonstruieren. Damit erlauben Expertensysteme eine Trennung zwischen Lösungswissen und Anwendung des Wissens. Ein weiteres Charakteristikum eines Expertensystems ist es, dass der Anwender die Problemlösung verstehen muss, insofern, dass er die erzeugten Alternativen hinsichtlich ihrer Lösungsgüte beurteilen kann. Dabei kann der Einsatz visueller Medien helfen, die sich besonders als Informationsübermittler eignen und somit das Verständnis der Zusammenhänge erleichtern.

Expertensysteme lassen sich in die Komponenten Wissensbasis, Problembeschreibung und -lösung sowie Dialog unterteilen (Abb. 2.47). Die Wissensbasis hält die Fakten und Regeln des Anwendungsbereichs fest. Die Problembeschreibung transformiert die Daten aus der Wissensbasis dergestalt, dass eine Interpretation und Problemlösung durch das Lösungsverfahren möglich wird. Die Dialogkomponente hat sowohl zur Wissensbasis (Eingabe von Daten) und zur Problemlösungskomponente (Lesen der Ergebnisdaten) eine Schnittstelle.

2.9.6.5 Neue Produktionskonfigurationen und Produkteinführungen durch Szenarien absichern

Die Ermittlung von Lösungsvorschlägen durch Expertensysteme unterstützt bei der Fragestellung, wie auf eine vorliegende Problemsituation reagiert werden kann. Oftmals sieht sich der Unternehmer aber auch Fragestellungen gegenüber, die sich mit der kurz-, mittel- und langfristigen Planung befassen. So sind beispielsweise folgende exemplarische Aufgabenstellungen zu lösen:

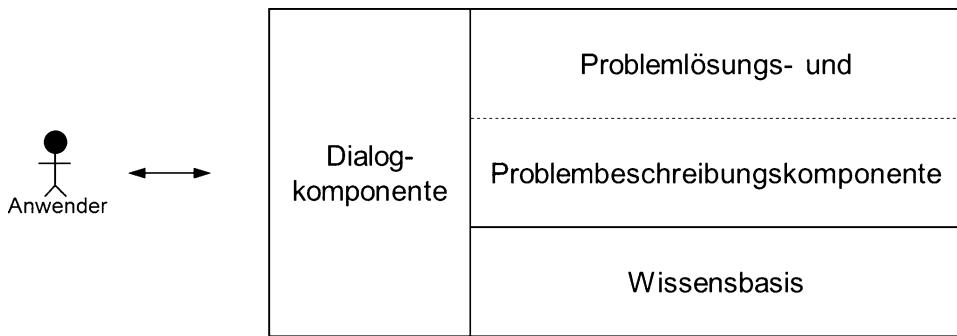


Abb. 2.47 Expertensysteme

- Kann ein avisierte Großauftrag mit den geforderten Lieferterminen zugesagt werden?
- Welche Ersatzinvestition sollte für einen bestehenden Maschinenpark getätigt werden, um maximale Flexibilität zu erhalten?
- Welche Kapazitäten sollten erhöht werden, um das anstehende Produktionsprogramm zu meistern?
- Welche Maßnahmen sind zu ergreifen, um ein Neuprodukt reibungslos in die laufende Produktion einzuschleusen?

Zur Beantwortung dieser Fragestellungen sind alternative Handlungsoptionen abzubilden und zu bewerten. Üblicherweise sind hierzu aufwendige Datenanalysen und Auswertungen notwendig, um unterschiedliche Szenarien vergleichen zu können. Aufgrund der Vielzahl an Einflussgrößen (Auftragslast, Produktionssystem) und der permanenten Veränderung der Datengrundlagen sind manuell erstellte Entscheidungsunterlagen zumeist bereits zum Zeitpunkt der Präsentation schon nicht mehr aktuell. Änderungen müssen mit hohem Aufwand nachgerechnet werden.

Mit der Szenarien-Technik erhöhen sich die Anforderungen an Datenzugriff und -verarbeitung, da neben aktuellen, statischen Daten nunmehr Informationen über die Wirkzusammenhänge in Form von dynamischen Modellen hinterlegt werden müssen. Der Aufwand allerdings lohnt sich: Mit Was-wäre-wenn-Planspielen lassen sich real nicht umsetzbare Alternativen im Vorfeld der Umsetzung bewerten.

Mit der Verfügbarkeit aller im Unternehmen verfügbaren Daten und der sinnhaften Interpretation der selbigen sieht der Unternehmer damit nicht nur wo er sich gerade befindet, sondern es wird ihm möglich sein, auch einen Blick auf ausgewählte Ausschnitte in der Zukunft zu werfen. Mit diesem Scheinwerfer nach vorne kann er frühzeitig den für ihn richtigen Weg erkennen und Stolpersteine ausweichen.

2.9.7 Erfolgsfaktoren zum Ausschöpfen der Potenziale von Big Data

2.9.7.1 Umgang mit Daten

Der Zugriff auf Daten zur Verarbeitung in der Planung und Steuerung von Produktionen findet seine Grenzen, wenn Aspekte der Privatsphäre, der Sicherheit und des geistigen Eigentums betroffen sind. So sind beispielsweise die Erhebung personenbezogener Arbeitszeitdaten oftmals problematisch (hierzu näher Abschn. 3.1). Personenbezogene Daten sind Informationen, die dazu genutzt werden können, die Identität eines Mitarbeiters zu erfahren. Arbeitszeitdaten und Projektzeitdaten sind personenbezogene Daten und unterliegen als solche besonderen Datenschutzbestimmungen. Diese Angaben müssen auf besonders geschützten Servern gespeichert werden, d. h. der Zugriff darauf ist nur wenigen besonders befugten Personen möglich, die mit der technischen oder kaufmännischen Betreuung der Server befasst sind. Zur Nutzung solcher Daten für die Planung und statistischen Auswertung sind die Datensätze anonymisiert zu verwenden.

Neben der Wahrung der Privatsphäre dient die Datensicherheit auch zum Schutz vor unbefugten Zugriff durch Dritte. Grundsätzlich gilt, dass Daten bzw. Informationen schützenswerte Güter sind. Der Zugriff auf diese sollte beschränkt und kontrolliert sein. Folgende Ziele sollten mit dem Umgang mit Daten verfolgt werden (Eckert 2012):

- **Vertraulichkeit:** Daten dürfen lediglich von autorisierten Benutzern gelesen bzw. modifiziert werden, dies gilt sowohl beim Zugriff auf gespeicherte Daten wie auch während der Datenübertragung.
- **Integrität:** Daten dürfen nicht unbemerkt verändert werden. Alle Änderungen müssen nachvollziehbar sein.
- **Verfügbarkeit:** Verhinderung von Systemausfällen. Der Zugriff auf Daten muss innerhalb eines vereinbarten Zeitrahmens gewährleistet sein.

In Bezug auf die Wahrung des geistigen Eigentums spielt die Datensicherheit ebenfalls eine maßgebliche Rolle, denn jedes Produkt, das in Form, Inhalt und Funktionalität nicht ausreichend geschützt ist, kann kopiert, verkauft und verwendet werden. Viel zu häufig unterschätzen besonders KMU die Bedeutung geistiger Schutzrechte. Geistige Schutzrechte oder Intellectual Property Rights (IPRs) schaffen bei entsprechender Verwertung nachhaltiges Wachstum und bringen in einem globalisierten Wirtschaftssystem finanzielle und strategische Vorteile. Der Weg von der Erfindung bis zur Veröffentlichung und Markteinführung sollte daher umsichtig beschritten werden.

2.9.7.2 Technologien

Die technologischen Fortschritte in der Erfassung, Verarbeitung und Speicherung von Daten sind Treiber für die Nutzung von Big Data in der Planung und Steuerung von Produktionsunternehmen. Die Infrastruktur zur Erfassung von Daten liegt in den meisten produzierenden Unternehmen vor: Mittels BDE (Betriebsdatenerfassung) und MDE (Maschinendatenerfassung) können zeitnah der Status eines jeden Auftrags und der Ressour-

cen und Mitarbeiter nachvollzogen werden. Durch Scan-Technologien (Barcode, RFID) in Zusammenhang mit einer leistungsfähigen IT-Infrastruktur sind die Aufwände zur Erfassung gering und ermöglichen eine lückenlose Verfolgung der Informations- und Logistikflüsse. Der Trend zu immer kostengünstigeren und vernetzten Sensoren hilft bei der Erhebung zusätzlicher, qualitativer Daten.

Die kontinuierliche Erhöhung der Leistung von Mikrochips gemäß dem Moore'schen Gesetz (Blau 2009) bildete eine wesentliche Grundlage der Digitalen Revolution. Die stetig steigenden Kapazitäten zur Speicherung und Veränderung von Signalen erlauben es, große Datenmengen miteinander zu verarbeiten. Die In-Memory-Technik auf dem Arbeitsspeicher eines Rechners bietet zudem wesentlich höhere Zugriffsgeschwindigkeiten als Festplattenlaufwerke und die Algorithmen für den Zugriff sind einfacher (hierzu näher Abschn. 4.3.4). Deshalb sind In-Memory-Datenbanken wesentlich schneller und ihre Zugriffszeiten sind besser vorhersagbar als die von auf Festplatten zugreifenden Datenbankmanagementsystemen. So nutzt beispielsweise SAP Hana die Potenziale der In-Memory-Technik zur Analyse großer Datenmengen, um ein Abbild der Produktion in Echtzeit zu liefern.

Zudem erleichtern Entwicklungen zu virtualisierten, parallelen und verteilten Anwendungen die Verarbeitung großer Datenmengen. Als Beispiel sei Apache Hadoop (vgl. Abschn. 4.3.2) genannt, ein freies, in Java geschriebenes Framework für skalierbare, verteilt arbeitende Software mit hohen Anforderungen an intensiver Rechenleistung.

2.9.7.3 Analysetechniken und Algorithmen

Zur systematischen Analyse und Auswertung der verfügbaren Daten bieten sich Business Intelligence (BI) Lösungen an. In der Praxis versteht man darunter in den meisten Fällen die Automatisierung des Berichtswesens, um Transparenz zu erhalten. Dabei werden die in den ERP-Systemen anfallenden Unternehmensdaten genutzt, um unter verschiedenen Blickwinkeln die Situation des Unternehmens zu analysieren und ggf. zu bewerten.

Die Analyse erfolgt üblicherweise nicht in den ERP-Systemen, sondern in einer davon getrennten Datenbasis. Dazu liegt hierfür ein Data-Warehouse (DWH) vor. Gründe dafür sind z. B. ungeeignete Strukturierung der Daten im ERP-System, keine Auswertungsmöglichkeit über mehrere ERP-Systeme hinweg bzw. unzulässige Belastung des ERP-Systems durch analytische Auswertungen.

Neben der Datensammlung im DWH greifen über die BI-Anwendung gehende Analyse- und Expertensysteme zumeist auf eigene Datenspeicher bzw. Datenbanksysteme zurück, in denen sie zusätzliche, nicht in den ERP-Systemen gespeicherte Daten halten, um die verwendeten Algorithmen zur Problemlösung anwenden zu können. Die Bandbreite an Anwendungsbeispielen ist sehr groß. Das am meisten verwendete Analysewerkzeug in der Planung ist MS Excel, bei dem zumeist die Daten im Tabellenkalkulationsprogramm selbst gehalten und verarbeitet werden. Mit den steigenden Anforderungen an Problemlösungskompetenz sowie an Reaktions- und Prognosefähigkeit finden zunehmend Optimierungs- und Simulationsanwendungen Eingang in die operativen Planungs- und Steuerungsprozesse (März und Weigert 2011).

2.9.7.4 Datenzugriff

Die Zugriffsmöglichkeiten von Client-Server-Architekturen erlauben einen allseits verfügbaren Zugriff auf zentral gespeicherte Daten. Mit der Unabhängigkeit von der verwendeten Hardware auf Client-Seite (Notebook, Pads, Smartphone, etc.) können Informationen jederzeit und von überall abgerufen werden. Voraussetzung für die Verfügbarkeit des Zugriffs sind Kommunikationsnetze, die eine stetige Online-Fähigkeit ermöglichen. Mit der Technologie des WLAN (Wireless Local Area Network) steht eine mobile Kommunikation zur Verfügung, die mehrere Endgeräte in einem räumlich begrenzten Gebiet per Funk vernetzen und mit dem Intranet bzw. Internet verbinden kann.

2.9.7.5 Organisationale Transformation und Führung

Mit dem Aufkommen großer Datens Mengen stellt sich allerdings die Frage, ob alle Daten zentral gehalten werden müssen und ob es nicht Daten gibt, die nur dezentral verfügbar sein müssen. Die Beantwortung dieser Frage geht einher mit der Überlegung, dass eine dezentrale Informationsbereitstellung gleichermaßen mit der Verlagerung auf dezentrale Planungsverantwortung, von einem transaktionalen zu einem demokratischen Auswerteparadigma wechselt. Eine vollständig zentrale Planung ist aufgrund der hohen Komplexität und vielzähligen Wechselbeziehungen nur schwer aktuell zu halten. Das Nutzen des dezentralen Wissens in der Planung und Steuerung der Wertschöpfungsstufen und dem kontinuierlichen Abgleich mit den benachbarten Planungsinstanzen fordert sowohl dezentrale Planungskompetenz als auch eine übergeordnete Planungskoordination mit vollständiger Vernetzung der involvierten Leistungseinheiten auf der Basis relevanter Schnittstelleninformationen.

Das heutige Verständnis für Planung und Steuerung ist noch stark reaktions-orientiert. Mit den aufgezeigten Technologien zur Nutzung von Daten in der Planung und Steuerung ist ein Trend in Richtung dezentraler Planungsverantwortung erkennbar, die in einem Gesamtkontext eingebunden eine szenario-basierte Steuerung des Unternehmens ermöglichen wird.

2.9.8 Fazit

Der Weg vom auslastungsorientierten Anbieter zum marktorientierten Käufer erfordert ein Umdenken in der Gestaltung und dem Betrieb von Industrieunternehmen. Unternehmen müssen die vom Markt geforderten Anforderungen in eine Produkt- und Organisationstrategie umsetzen. Auf Dauer haben nur die Unternehmen Erfolg, die es verstehen, die markt- und technologiegetriebenen Innovationen in ihrer Organisation abzubilden. Der Hebel liegt in der engen Verflechtung der strategischen, prozessbezogenen und der informationstechnischen Ebene zur Auftragsabwicklung. Das Wissen im Umgang mit den unternehmerischen Daten, der Nutzung von Big Data zur proaktiven Planung und Steuerung in Echtzeit, kann hierbei Wettbewerbsvorteile sichern.

Literatur

Literatur zu 2.1

- Altmann, G. (2013): Neue Entscheidungskultur. *Personalmagazin*, 3, 22–23.
- Ammon, G. (1985): Der mehrdimensionale Mensch. *Dynamische Psychiatrie*, 18, 99–110.
- Anders, G. (2010): *Die Antiquiertheit des Menschen: Über die Seele im Zeitalter der zweiten industriellen Revolution* (S. 245). C.H.Beck.
- Anitav, N., Riley, H. N., & Ahituv, N. (1994): Principles of Information Systems for Management, 45.
- Hanschke, I. (2011): *Strategisches Management der IT-Landschaft: Ein praktischer Leitfaden für das Enterprise Architecture Management*. Carl Hanser Verlag.
- Hertweck, D. (2003). *Escalating Commitment als Ursache gescheiterter D.V.-Projekte: Methoden und Werkzeuge zur Deeskalation*. Deutscher Universitäts-Verlag GmbH.
- Hüner, K. M., Ofner, M., & Otto, B. (2009): Towards a maturity model for corporate data quality management. In *Proceedings of the 2009 ACM symposium on Applied Computing – SAC '09* (S. 231–238). New York, New York, USA: ACM Press.
- Krcmar, H. (2005). *Informationsmanagement*. Heidelberg: Springer; Auflage: 4., überarb. u. erw. Aufl.
- Lieven, P. (2009). *Der Werdegang der Krise*. (R. Elschen & T. Lieven, Eds.) (S. 219–236). Wiesbaden: Gabler.
- Matthes, D. (2011). *Enterprise Architecture Frameworks Kompendium: Über 50 Rahmenwerke für das IT-Management*. Springer; Auflage: 2011.
- Nadler, U. (2014): Kann man „Big Data“ managen? Wie passt „Big Data“ in Information Governance Konzepte? Präsentation GI-Regionalgruppe Hamburg, 29. Januar http://www.hbt.de/fileadmin/media/GI_272_Big_Data_Governance.pdf, (zugegriffen am 8. April 2014)
- Ortmann, G. (1990). *Computer und Macht in Organisationen: Mikropolitische Analysen* (S. 652). VS Verlag für Sozialwissenschaften; Auflage: 1990.
- Schermann, M., Prilla, M., Krcmar, H., & Herrmann, T. (2008). Bringing life into references process models : A participatory approach for identifying , discussing , and resolving model adaptations. In *Multikonferenz Wirtschaftsinformatik 2008* (S. 1577–1588). München.
- Völker, R. (2007). *Wissensmanagement im Innovationsprozess*. Heidelberg: Physica-Verlag HD.

Literatur zu 2.2

- Aier S., Maletta F., Riege C., Stucki K., Frank A (2008a). Aufbau und Einsatz der Geschäftsarchitektur bei der AXA Winterthur – Ein minimal invasiver Ansatz. In: DW2008: Synergien durch Integration und Informationslogistik, Gesellschaft für Informatik Köllen, St. Gallen
- Aier S., Riege C., Winter R (2008b). Unternehmensarchitektur – Literaturüberblick und Stand der Praxis. *Wirtschaftsinformatik* 50(4):292–304
- Bange C., Janoschek N (2014). Big Data Analytics 2014 – Auf dem Weg zur datengetriebenen Wirtschaft. Tech. rep., BARC-Institut, Würzburg
- Baron P (2013). Big Data für IT-Entscheider. Carl Hanser Verlag, München
- Bayer F (2013). Prozessmanagement für Experten. Springer-Verlag, Berlin, Heidelberg
- Bitkom (2011). Enterprise Architecture Management – neue Disziplin für die ganzheitliche Unternehmensentwicklung. Tech. rep., Bitkom e. V., URL http://www.bitkom.org/de/publikationen/38337_67462.aspx

- Bitkom (2012). Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Tech. rep., Bitkom e. V., Berlin
- Bleicher K (2011). Das Konzept integriertes Management, 8th edn. Campus Verlag, Frankfurt am Main
- Conrads R (2013). In sieben Schritten zum erfolgreichen Big-Data-Projekt. Informatik-Spektrum 37(2):127–131
- Cramer C., Dietze A (2012). Vom Hype zur Umsetzung – Checkliste für die Big-Data-Strategie. URL <http://bit.ly/1mAhdU6>
- Dörner D (2009). Die Logik des Mißlingens: Strategisches Denken in komplexen Situationen, 8th edn. Rowohlt Taschenbuch Verlag, Reinbeck bei Hamburg
- Frauenhofer IAIS (2012). BIG DATA – Vorsprung durch Wissen Innovationspotenzialanalyse. URL http://www.fraunhofer.de/content/dam/zv/de/forschungsthemen/kommunikation/bigdata/Innovationspotenzialanalyse_Big-Data_Fraunhofer-IAIS.pdf
- Freitag A., Matthes F., Schulz C (2011). A METHOD FOR BUSINESS CAPABILITY DEPENDENCY ANALYSIS. URL http://wwwmatthes.in.tum.de/file/100js250slddo/sebis-Public-Website/Team/Andreas-Freitag/Final_INNOV_Capabilities_2011.pdf
- Gadatsch A (2012). Big Data. wisu – Das Wirtschaftsstudium 41. Jahrga(WISU 12/12):1615–1621
- Gates B (1994). Information at your Fingertips. In: Comdex, Las Vegas
- Hanschke I (2012). Enterprise Architecture Management – Einfach und Effektiv. Carl Hanser Verlag, München
- Hanschke I (2013). Strategisches Management der IT-Landschaft, 3rd edn. Carl Hanser Verlag, München
- Jahnke I., Herrmann T., Prilla M (2008). Modellierung statt Interviews? Eine „neue“ qualitative Erhebungsmethode. In: Herczeg M., Kindsmüller MC (eds) Mensch und Computer 2008. 8. fachübergreifende Konferenz für interaktive und kooperative Medien, Oldenbourg Verlag, München, S. 377–386
- Keuntje J., Barkow R (2010). Enterprise Architecture Management in der Praxis, 1st edn. Symposium Publishing, Düsseldorf
- Kröger K (2013). 10 Gartner-Trends bis 2017. URL <http://www.cowo.de/a/2501630>
- Küller P., Hertweck D (2013). Bedeutung von Services in einer dezentralen Energieversorgung. HMD – Praxis der Wirtschaftsinformatik 50(291):60–70
- Lux J., Wiedenhöfer J., Ahlemann F (2008). Modellorientierte Einführung von Enterprise Architecture Management. HMD – Praxis der Wirtschaftsinformatik 45. Jg.(262):19–28
- Marek D (2010). Unternehmensentwicklung verstehen und gestalten – Eine Einführung, 1st edn. Gabler, Wiesbaden
- Matthes D (2011). Enterprise Architecture Frameworks Kompendium. Springer-Verlag, Heidelberg
- Newman D (2012). Big Data Disruptions Tamed With Enterprise Architecture – 27 March 2012. URL <http://www.gartner.com/resId=1964716>
- Niemann KD (2005). Von der Unternehmensarchitektur zur IT-Governance, 1st edn. Springer Fachmedien, Wiesbaden
- Sandkuhl K., Wißotzki M., Stirna J (2013). Unternehmensmodellierung. Springer-Verlag, Berlin, Heidelberg
- The Open Group (2010). TOGAF Version 9 – Ein Pocket Guide. Van Haren Publishing, Zaltbommel

- Thielscher J (2010). Enterprise Architecture Management Capabilities entwickeln. In: AK Professional Services vom 9. Juni 2010, Bitkom e. V., Bad Homburg
- Vogt M., Hertweck D., Küller P., Hales K (2011). Adapting IT Governance Frameworks Towards Domain Specific Requirements : Examples of the Domains of Small & Medium Enterprises and Emergency Management. In: Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan August 4th–7th 2011, Detroit
- Wolff F (2008). Ökonomie multiperspektivischer Unternehmensmodellierung: IT-Controlling für modell-basiertes Wissensmanagement. Gabler Verlag, Wiesbaden

Literatur zu 2.3

- Apte C (2010) The Role of Machine Learning in Business Optimization. In: Proceedings of the 27th International Conference on Machine Learning
- Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) CRISP-DM 1.0, Step-by-step data mining guide. URL <http://www.the-modeling-agency.com/crisp-dm.pdf>
- Chu C, Kim S, Lin Y, Yu Y, Bradski G, Ng A, Olukotun K (2006) Map-Reduce for Machine Learning on Multicore. In: Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS'06), S. 281–288
- Davenport T, Barth P, Bean R (2012) How ‘Big Data’ is Different. MIT Sloan Management Review 54(1):22–24
- Domingos P (2012) A Few Useful Things to Know About Machine Learning. Communications of the ACM 55(10):78–87
- Eckerson W (2007) Predictive Analytics: Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report, TDWI Research
- Franks B (2012) Taming the Big Data Tidal Wave. Wiley
- Jensen D, Cohen P (2000) Multiple Comparisons in Induction Algorithms. Machine Learning 38:309–338
- Kemper HG, Baars H, Mehanna W (2010) Business Intelligence – Grundlagen und praktische Anwendungen. Vieweg+Teubner
- Kraska T, Talwalkar A, Duchi J, Griffith R, Franklin M, Jordan M (2013) MLbase: A Distributed Machine-learning System. In: 6th Biennial Conference on Innovative Data Systems Research (CIDR’13)
- LaValle S, Lesser E, Shockley R, Hopkins M, Kruschwitz N (2011) Big Data, Analytics and the Path from Insights to Value. MIT Sloan Management Review 52(2):21–31
- Mohanty S, Jagadeesh M, Srivatsa H (2013) Big Data Imperatives. Apress
- Owen S, Anil R, Dunning T, Friedman E (2012) Mahout in Action. Manning Publications Co.
- Rajaraman A, Leskovec J, Ullman J (2010) Mining of Massive Datasets. Standford University, URL <http://infolab.stanford.edu/~ullman/mmds/book.pdf>
- Shearer C (2000) The CRISP-DM Model: The new blueprint for data mining. Journal of Data Warehousing 5(4):13–22
- Wu X, Zhu X, Wu GQ, Ding W (2014) Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering 26(1):97–107

Literatur zu 2.4

- Auer, S., März, L., Tutsch, H., Sihn, S. (2011): Classification of Interdependent Planning Restrictions and their Various Impacts on Long-, Mid- and Short Term Planning of High Variety Production. New Worlds of Manufacturing. 44th CIRP International Conference on Manufacturing Systems 2011, edited by N. A. Duffie, Madison, Wisconsin, Omnipress.
- Boysen, N., Malte, F., Scholl, A. (2007): Produktionsplanung bei Variantenfließfertigung. Planungshierarchie und Hierarchische Planung. Jenaer Schriften zur Wirtschaftswissenschaft. Hrsg.: H.-W. Lorenz, Scholl A. Wirtschaftswissenschaftliche Fakultät, Friedrich-Schiller-Universität Jena, 22/2006
- KPMG (2011) KPMG's Global Automotive Executive Survey 2011. Publ. No. 101205.
- März, L., Pröpster, M., Röser, S. (2012): Simulationsgestützte Bewertung getakteter Linien. *wt Werkstatttechnik online* 102 (2012) 3: 146–151
- Niederprüm, M., Sammer, K. (2012): Generation of car body variants via late configuration, Konferenz Montagesysteme 2012 Automotive Circle International, 28.–29. Februar 2012, Bad Nauheim
- VDI 3633 Blatt 1 (2010) Simulation von Logistik-, Materialfluß- und Produktionssystemen; Grundlagen.

Literatur zu 2.6

- Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 16.7.2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
- Angioni, G. (2004). Doing, Thinking, Saying. In C. Sanga & T. Ortalli (Eds.), *Nature Knowledge* (S. 249–261). New York-Oxford: Berghahn Books.
- Berchtenbreiter, S. (2013). *BIG DATA und die Implikationen für die Marketingforschung*. Marktforschung.de. <http://www.marktforschung.de/information/fachartikel/marktforschung/big-data-und-die-implikationen-fuer-die-marketingforschung/>, zugegriffen am 14.05.2013.
- Bloching, B., Luck, L., & Rame, T. (2012). *Data unser*. München: Redline Verlag.
- Boyd, D. (2010). *Privacy and Publicity in the Context of Big Data*. WWW 2010. <http://www.danah.org/papers/talks/2010/WWW2010.html>
- Campillo-Lundbeck, S. (2014). Allmächtiger Algorithmus. *Horizont*, 5/2014, 17.
- Dapp, T. F. (2014). Big Data. Die ungezähmte Macht. *Deutsche Bank Research*, 04.03.2014.
- Faulbaum, F., Stahl, M., & Wiegand, E. (2012). *Qualitätssicherung in der Umfrageforschung. Neue Herausforderungen für die Markt- und Sozialforschung*. Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute. Wiesbaden: Springer VS.
- Gogia, S. (2012). Was „Big Data“ so wichtig für Kundeninteraktion macht. *Forrester Research*, 01.06. 2012.
- Googlewatchblog <http://www.googlewatchblog.de/2013/05/jahre-youtube-nutzer-stunden/>, zugegriffen am 19.06.2014.
- Hofmann, O. (2012). Entwicklungen in der Online-Marktforschung. Vom ungeliebten Kind zum Allheilmittel. In F. Faulbaum, E. Stahl, & Wiegand, E. (Eds.), *Qualitätssicherung in der Umfrageforschung* (S. 139–146). Wiesbaden: Springer VS.
- Kary, J. (2014). Datenüberflutet. *Markt und Mittelstand*, 3/2014, 20–26.

- o.V., *Big Data ist bei Industrieunternehmen als Thema angekommen*. Marktforschung.de, <http://www.marktforschung.de/information/nachrichten/marktforschung/big-data-ist-bei-industrieunternehmen-als-thema-angetreten/backpid/3279/>, zugegriffen am 11.03.2014.
- Reips, U.-D. (2009). Schöne neue Forschungswelt – Zukunftstrends. In C. König, M. Stahl, & E. Wiegand (Eds.), *GESIS-Schriftenreihe Band 1: Nicht-reaktive Erhebungsverfahren* (S. 129–138). GESIS: Bonn.
- Streif, S. (2013). Daten wie Sand am Meer, *acquisa*, 10/2013, 19.
- Voss, A. & Sylla, K.-H. (2014). Innovationspotenzialanalyse Big Data – Ergebnisse für das Marketing. *Marketing Review St. Gallen*, 1/2014, 36–45.

Literatur zu 2.7

- Bachmann, R., Kemper, G., & Gerzer, T. (2014). *Big Data – Fluch oder Segen?* Heidelberg: Mitp.
- Berchentbreiter, S. (2013). *BIG DATA und die Implikationen für die Marketingforschung*. Marktforschung.de. <http://www.marktforschung.de/information/fachartikel/marktforschung/big-data-und-die-implikationen-fuer-die-marketingforschung/>, zugegriffen am 05.06.2014.
- Bloching, B., Luck, L., & Ramge, T. (2012). *Data unser*. München: Redline Verlag.
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-mimic method matrix. *Psychological Bulletin*, 56, 81–105.
- Church, A. H. & Dutta, S. (2013). The promise of big data for OD: Old wine in new bottles or the next generation of data-driven methods for change? *OD Practitioner*, 45, 23–31.
- Eudes, Y. (10.04.2014). *Comment notre ordinateur nous manipule*. LeMonde.fr. http://www.lemonde.fr/technologies/article/2014/04/10/big-brother-ce-vendeur_4399335_651865.html, zugegriffen am 05.06.2014.
- Foscht, T. & Swoboda, B. (2011). *Käuferverhalten*. Wiesbaden: Gabler.
- Grüger, O. (2013). Webanalyse und Business Intelligence als Basis für die aktive Steuerung von Webshops. In R. Haberich (Ed.), *Future Digital Business* (S. 259–273). Heidelberg: Mitp.
- Haberich, R. (2013). Digital Intelligence. In R. Haberich (Ed.), *Future Digital Business* (S. 49–70). Heidelberg: Mitp.
- Heinemann, G. (2014). *Der neue Online-Handel*. Wiesbaden: Springer Gabler.
- Kudyba, S. (2014). Mining and analytics in E-Commerce. In S. Kudyba (Ed.), *Big Data, Mining, and Analytics* (S. 147–163). Boca Raton (FL): CRC Press.
- Kudyba, S. & Kwatinetz, M. (2014). Introduction to the Big Data Era. In S. Kudyba (Ed.), *Big Data, Mining, and Analytics* (S. 1–15). Boca Raton (FL): CRC Press.
- Morys, A. (2013). Fünf Regeln, damit aus einer kleinen Kennzahl ein großer Deckungsbeitrag wird. In R. Haberich (Ed.), *Future Digital Business* (S. 371–399). Heidelberg: Mitp.
- Spiegel, J. R., McKenna, M. T., Lakshman, G. S., & Nordstrom, P. G. (2013). Method and system for anticipatory package shipping. U.S. Patent 8,615,473 B2, Dec 24, 2013
- Spieß, E. (2013). *Konsumentenpsychologie*. München: Oldenbourg.
- Stoever, L. (2014). Verschieben Sie Big Data! *Internet World Business*, 7/14, 42.
- Völcker, T. (2013). Schürfen nach Gold: Der Wert von Social Media-Daten. In R. Haberich (Ed.), *Future Digital Business* (S. 275–288). Heidelberg: Mitp.
- Zimmer, D. (2014). Outfit für jedes Wetter. *Internet World Business*, 6/14, 16.

Literatur zu 2.8

- Hamerle, A., Plank, K. (2010), Intransparenzen auf Verbriefungsmärkten. Auswirkungen auf Risikoanalyse und Bewertung. Informatik-Spektrum 10/2010. Springer. Wiesbaden 2010
- Hofstetter, Y. (2014), Sie wissen alles. Wie intelligente Maschinen in unser Leben eindringen und warum wir für unsere Freiheit kämpfen müssen. C. Bertelsmann. München: 2014
- Hüthig, S. (2013). Die Entdeckung der Standardsoftware. Bankmagazin 10/2013: S. 8.
- Hüthig, S. (2014). Digitalisierung. Viel mehr als ein Projekt. Bankmagazin 10/2014: S. 12.
- Krüger, M., Seitz, F. (2014). Kosten und Nutzen des Bargelds und unbarer Zahlungssysteme. Studie im Auftrag der Deutschen Bundesbank (20.08.2014) abrufbar unter www.bundesbank.de
- Moch, D. (2011). Strategischer Erfolgsfaktor Informationstechnologie. Wiesbaden: Gabler.
- Schulzki-Haddouti, C. (2014). Zügelloses Scoring. CT 21/2014, S. 38.
- Wiebe, F. (2014a). Angriff der Computer-Nerds. Handelsblatt vom 12.08.2014.
- Wiebe, F. (2014b). Die Angst vor dem Silicon Valley. Handelsblatt vom 13.10.2014.

Literatur zu 2.9

- Blau, D. (2009): Das Moore'sche Gesetz. München: GRIN-Verlag
- Eckert, C. (2012): IT-Sicherheit. Konzepte – Verfahren – Protokolle. München: Oldenbourg Wissenschaftsverlag, Auflage:4
- März, L., Weigert, G. (2011): Simulation und Optimierung. Heidelberg: Springer; Hrsg.: März L, Krug W, Rose O, Weigert G: Simulation und Optimierung in Produktion und Logistik. Praxisorientierter Leitfaden mit Fallbeispielen. Reihe VDI-Buch.

Michael Bartsch, Olaf Botzem, Thorsten Culmsee, Joachim Dorschel,
Jenny Hubertus, Carsten Ulbricht und Thorsten Walter

3.1 Datenschutz

3.1.1 Prinzipien des Datenschutzrechts

Thorsten Culmsee

3.1.1.1 Einleitung

In einer Welt des Ubiquitous Computing nehmen Datenmengen und Quellen, aus denen Daten erhoben werden können, permanent und dynamisch zu (Finsterbusch und Knop 2012). Big Data-Anwendungen vermögen extrem große, ungeordnete Mengen heterogener Daten unterschiedlichster Herkunft zu strukturieren, unter verschiedenen Gesichtspunkten miteinander zu kombinieren und auszuwerten und so aus den vorhandenen Daten qualitativ neue Daten zu synthetisieren. Die Qualität der Ergebnisse steigt dabei mit der Größe der Datenmenge, die analysiert wird (Roßnagel 2013, S. 562 und 564; Weichert 2013, S. 251 f.).

Von unkontrollierten Datenbeständen, die mithilfe von Big Data-Anwendungen analysiert werden, kann eine diffuse Bedrohlichkeit ausgehen: Der Einzelne weiß nicht, was welcher Big Data-Verarbeiter – gleich ob staatliche Behörde oder privates Unternehmen – über ihn weiß, weiß aber, dass der Big Data-Verarbeiter vieles, auch Höchstpersönliches über ihn wissen kann (so das BVerfG zur Vorratsdatenspeicherung: BVerfG, Urt. v. 02.03.2010, Az.: 1 BvR 256/08 u. a., RdNr. 241 f. – zitiert nach juris).

Prof. Dr. Michael Bartsch · Olaf Botzem · Thorsten Culmsee · Joachim Dorschel ✉ ·
Thorsten Walter
Karlsruhe, Deutschland
e-mail: jd@bartsch-rechtsanwaelte.de

Jenny Hubertus · Dr. Carsten Ulbricht
Stuttgart, Deutschland

Kann der einzelne Mensch aber nicht überschauen, welche Informationen in seiner sozialen Umwelt über ihn bekannt sind, kann er „*in seiner Freiheit wesentlich gehemmt werden, aus eigener Selbstbestimmung zu planen und zu entscheiden*“ (BVerfG NJW 1984, S. 419 [422]). Das Grundrecht auf informationelle Selbstbestimmung, hergeleitet aus dem allgemeinen Persönlichkeitsrecht (Art. 2 Abs. 1 GG) in Verbindung mit dem Schutz der Menschenwürde (Art. 1 Abs. 1 GG), schützt den Einzelnen daher gegen die unbegrenzte Erhebung, Speicherung, Verwendung und Weitergabe seiner persönlichen Daten und gewährleistet, dass jeder Mensch grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten und damit darüber bestimmen kann, „*wer was wann bei welcher Gelegenheit über ihn weiß*“ (BVerfG NJW 1984, S. 419 [422]). Das Grundrecht auf informationelle Selbstbestimmung ist damit ein grundlegender Baustein einer auf der Freiheit und Autonomie des Einzelnen beruhenden Kommunikationsordnung. Es weist zudem einen funktionalen Bezug zum demokratisch verfassten Gemeinwesen auf, denn Selbstbestimmung ist „*eine elementare Funktionsbedingung eines auf Handlungs- und Mitwirkungsfähigkeit seiner Bürger begründeten freiheitlich demokratischen Gemeinwesens*“ (BVerfG NJW 1984, S. 419 [422]).

Das Recht auf informationelle Selbstbestimmung gewährt dem Einzelnen aber kein absolutes und uneingeschränktes Recht auf seine personenbezogenen Daten, sondern ist seinerseits vom Gesetzgeber mit verfassungsrechtlich legitimen öffentlichen Interessen und konfligierenden Grundrechten Dritter, etwa der Datenverarbeiter in Ausgleich zu bringen (BVerfG NJW 1984, S. 419 [422]; Roßnagel et al. 2001, S. 48 ff.).

Hierbei orientiert sich der Gesetzgeber an grundlegenden Prinzipien des Datenschutzrechts. Diese finden sich in den allgemeinen Datenschutzgesetzen wie dem BDSG oder den jeweiligen Landesschutzgesetzen, aber auch in unterschiedlicher Ausformung in bereichsspezifischen Regelungen (Witt 2010, S. 73).

3.1.1.2 Prinzipien des Datenschutzrechts

Verbot mit Erlaubnisvorbehalt

Das deutsche Datenschutzrecht folgt konstruktiv dem Prinzip des Verbots mit Erlaubnisvorbehalt. Das heißt: Die Erhebung, Verarbeitung oder Nutzung personenbezogener Daten ist grundsätzlich verboten, es sei denn, dass der Betroffene hierin eingewilligt hat oder ein Gesetz oder eine andere Rechtsvorschrift die datenschutzrechtlich relevante Handlung erlaubt (vgl. § 4 Abs. 1 BDSG, § 12 Abs. 1 TMG).

Im Rahmen einer Big Data-Anwendung ist jeder einzelne Verarbeitungsschritt – beginnend mit der Datenerhebung über die weitere Verarbeitung und die Nutzung der Analyseergebnisse – gesondert auf seine rechtliche Zulässigkeit zu prüfen (Weichert 2013, S. 255).

Die Einwilligung ist hierbei kein taugliches Instrument, um Big Data-Anwendungen datenschutzkonform zu gestalten (Katko und Babaei-Beigi 2014, S. 362 ff.).

Der Einwilligende muss über die verarbeitende Stelle, Art und Umfang der Datenverarbeitung und ihren Zweck informiert sein und hierzu ausdrücklich, freiwillig und formgerecht seine Zustimmung erteilt haben (§ 4 a, § 28 Abs. 3 a BDSG). Das ist im Zu-

sammenhang mit Big Data-Anwendungen, bei denen extrem große Mengen heterogener Daten unterschiedlichster Herkunft verarbeitet werden, wegen der Vielzahl an Betroffenen bereits verwaltungstechnisch nicht handhabbar (Weichert 2013, S. 255). Eine Einwilligungslösung wird häufig auch daran scheitern, dass die erhobenen Daten für eine Vielzahl von Zwecken analysiert werden sollen oder im Zeitpunkt der Einwilligung der Zweck der Big Data-Analyse noch nicht hinreichend bestimmt ist (Roßnagel 2013, S. 564).

Aufgrund welcher gesetzlichen Tatbestände eine Big Data-Anwendungen erlaubt ist, kann nur mit Blick auf die konkrete Anwendung und den einzelnen Verarbeitungsschritt geprüft und beantwortet werden. Für private Stellen kommen hier zunächst §§ 27 ff. BDSG und insbesondere § 28 Abs. 1 Nr. 2 BDSG als Erlaubnistatbestände infrage (hierzu näher im Folgenden unter Abschn. 3.1.2).

Entscheidend ist stets, dass die Interessen des Datenverarbeiters gegenüber den schutzwürdigen Interessen des Betroffenen überwiegen. Da es im Rahmen von Big Data-Anwendungen aufgrund der Vielzahl von Betroffenen nicht möglich ist, die Interessen jedes Einzelnen zu berücksichtigen, dürfen die Betroffeneninteressen bei Abwägungsentscheidungen allgemein und ohne nähere Differenzierung einbezogen werden (Roßnagel 2013, S. 564; Weichert 2013, S. 257).

Unzulässig bleiben aber generell Big Data-Anwendungen, die den Einzelnen zum Objekt der Datenverarbeitung machen. Big Data-Analysen, die darauf abzielen, einen Menschen in seiner ganzen Persönlichkeit zu erfassen und ein Persönlichkeitsbild von ihm zu erstellen, sei es auch in der Anonymität einer statistischen Erhebung, oder die in den Kernbereich der privaten Lebensführung vordringen, sind daher verboten (BVerf NJW 1969, S. 1707; Roßnagel 2013, S. 565). Gleches gilt, wenn die Big Data-Anwendung zur Folge hat, dass der Einzelne zum Objekt automatisierter Entscheidungen degradiert wird (vgl. § 6 a Abs. 1 BDSG).

Zweckbindung

Nach dem Zweckbindungsgrundsatz dürfen personenbezogene Daten nur zu dem Zweck verarbeitet und genutzt werden, zu dem sie ursprünglich erhoben worden sind.

Das Recht, grundsätzlich selbst über die Preisgabe und Verwendung persönlicher Daten bestimmen zu können, setzt die Kenntnis ihrer beabsichtigten Verwendung voraus (BVerfG NJW 1984, S. 419 [422]). Der Verwendungszweck ist daher bereits bei der Erhebung personenbezogener Daten in Hinblick auf den gesamten Verarbeitungsprozess konkret festzulegen (§ 28 Abs. 1 S. 2 BDSG). Er muss legitim sein. Ohne die Festlegung eines legitimen konkreten Verarbeitungszwecks kann die Erforderlichkeit einer Datenverarbeitung nicht geprüft und können zu treffende Abwägungsentscheidungen nicht vorgenommen werden (Weichert 2013, S. 256). Der konkrete Verarbeitungszweck ist dem Betroffenen mitzuteilen (§ 4 Abs. 3 S. 1 Nr. 2, § 4 a Abs. 1 S. 2, § 33 Abs. 1 S. 1 BDSG). Personenbezogene Daten auf Vorrat zu unbestimmten oder noch nicht bestimmmbaren Zwecken zu sammeln, ist unzulässig (BVerfG NJW 1984, S. 419 [422]).

Aus dem Prinzip der Zweckbindung folgt weiter, dass personenbezogene Daten grundsätzlich nur für den Zweck, für den sie erhoben wurden, im erforderlichen Umfang ver-

wendet und für die Dauer seines Bestehens gespeichert werden dürfen (Roßnagel 2013, S. 564 f.; Weichert 2013, S. 256). Eine spätere Zweckänderung ist nichtöffentlichen Stellen nur unter den engen Voraussetzungen des § 28 Abs. 3 BDSG erlaubt. Ursprünglich für einen spezifischen Zweck gespeicherte Daten dürfen nicht für alle denkbaren Zwecke ausgewertet werden (Roßnagel 2013, S. 565). Aus der Zweckbindung folgt auch das Verbot, Daten, deren ursprüngliche Zwecke miteinander nicht vereinbar sind und sich gegenseitig ausschließen, zusammenzuführen (BVerfG NJW 1984, S. 419 [427]). Im Umkehrschluss bedeutet das allerdings, dass die Zusammenführung und gemeinsame Auswertung von Daten durch eine verantwortliche Stelle möglich ist, wenn die in die Analyse einfließenden Daten einem gemeinsamen, legitimen und konkreten Zweck zugeordnet werden können (Weichert 2013, S. 256).

Big Data-Anwendungen geraten leicht in rechtliche Konflikte zum Zweckbindungsgrundsatz, denn sie zeichnen sich gerade dadurch aus, dass sie heterogene Datenmassen unterschiedlichster Herkunft losgelöst von den ursprünglichen Erhebungs-, Verarbeitungs- und Nutzungszwecken zusammenführen, strukturieren und analysieren, wobei die neuen Verarbeitungszwecke zunächst noch nicht feststehen, sondern erst aus dem Analyseergebnis resultieren.

Datenschutzrechtliche Gestaltungsspielräume für Big Data-Analysen können sich jedoch im Zusammenhang mit öffentlich zugänglichen Daten ergeben (§ 28 Abs. 1 S. 1 Nr. 3, Abs. 2 Nr. 1, § 29 Abs. 1 S. 1 Nr. 2 BDSG).

Öffentlich zugänglich sind Daten, die allgemein zugänglich gemacht werden dürfen oder die ihrer Intention und technischen Gestaltung nach nicht auf den Zugriff durch einen eingeschränkten Nutzerkreis beschränkt sind (Gola und Schomerus 2015, S. 436 f.; Weichert 2013, S. 257). Hierzu zählen allgemein zugängliche Quellen wie etwa Zeitungen, öffentlich zugängliche Register wie das Handelsregister (vgl. § 9 Abs. 1 S. 1 HGB) oder Internetdaten, soweit der Zugriff auf sie für jedermann eröffnet ist, was bei Zugriffs-schranken in Form von Registrierungs- oder Login-Verfahren nicht der Fall ist (Gola und Schomerus 2015, S. 436 f.; Weichert 2013, S. 257). Nicht hierzu zählen amtliche Informationen oder Open Government Data, zu denen ein Anspruch auf Zugang beispielsweise nach dem IfG oder UIG besteht (vgl. § 1 Abs. 1 S. 1, § 5 IfG; § 1 Abs. 1, § 9 Abs. 1 S. 1 Nr. 1 UIG; § 1 Abs. 3 IWG).

Handelt es sich um frei zugängliche Massendaten, dürfen die Betroffeneninteressen in zu treffende Abwägungsentscheidungen etwa nach § 28 Abs. 1 S. 1 Nr. 3, Abs. 2 Nr. 1, § 29 Abs. 1 S. 1 Nr. 2 BDSG allgemein und ohne nähere Differenzierung einbezogen werden (Weichert 2013, S. 257). Eine Pauschalisierung der Betroffeneninteressen verbietet sich aber bei einem Widerspruch des Betroffenen (§ 35 Abs. 5 BDSG) oder wenn besondere Arten personenbezogener Daten gemäß § 3 Abs. 9 BDSG wie beispielsweise Gesundheitsdaten betroffen sind (Weichert 2013, S. 257).

Prinzip der Transparenz

Um sein Recht auf informationelle Selbstbestimmung wirksam ausüben zu können, muss der Einzelne seinerseits wissen können, „*wer was wann bei welcher Gelegenheit über ihn weiß*“ (BVerfG NJW 1984, S. 419 [422]; vgl. insofern auch Erwagungsgrund 38 der

RL 95/46/EG vom 24.10.1995 – Datenschutzrichtlinie). Das daraus folgende Gebot der Transparenz ist in verschiedenen datenschutzrechtlichen Vorschriften umgesetzt. Der Betroffene kann Einsicht in das von der verantwortlichen Stelle zu führende Verfahrensverzeichnis nehmen (§ 4 g Abs. 2 S. 2 BDSG) und Auskunft über die Erhebung, Verarbeitung und Nutzung seiner personenbezogenen Daten (§ 34 BDSG) und – bei automatisierten Einzelentscheidungen – über den logischen Aufbau des Verarbeitungsverfahrens (§ 6 a Abs. 3 BDSG) verlangen. Daneben bestehen Benachrichtigungspflichten (§ 33 BDSG), die im Falle einer Big Data-Anwendung allerdings schon aus praktischen Gründen wegen der unübersehbaren Zahl von Betroffenen gemäß § 33 Abs. 2 Nr. 7 lit. a) BDSG entfallen dürften. Auch Informationspflichten, etwa im Rahmen der Datenerhebung (§ 4 Abs. 3 S. 1 BDSG) oder im Falle einer Datenschutzpanne (§ 42 a BDSG), dienen der Transparenz von Datenverarbeitungsvorgängen.

Das Prinzip der Transparenz verlagert sich bei Big Data-Anwendungen von der individuellen auf die institutionelle Ebene. Da Big Data-Anwendungen besondere Risiken für die Rechte der Betroffenen aufweisen, unterliegen sie häufig vor Beginn der Verarbeitung der Vorabkontrolle (§ 4 d Abs. 5 S. 1 BDSG). Zuständig für diese ist der Datenschutzbeauftragte (§ 4 d Abs. 6 S. 1 BDSG), der sich in Zweifelsfällen an die zuständige Aufsichtsbehörde zu wenden hat (§ 4 d Abs. 6 S. 3 BDSG). Mit dem Verfahren der Vorabkontrolle verbindet sich die Chance, Big Data-Anwendungen prozedural abzusichern und rechtlich zu legitimieren (vgl. hierzu auch Katko und Babaei-Beigi 2014, S. 363 f.).

Unmittelbarkeit der Datenerhebung

Das Transparenzprinzip spiegelt sich auch im Direkterhebungsgrundsatz. Danach sind personenbezogene Daten grundsätzlich beim Betroffenen zu erheben (§ 4 Abs. 2 S. 1 BDSG).

Werden im Rahmen von Big Data-Anwendungen aus den vorhandenen Daten qualitativ neue Daten synthetisiert, dann kann der Betroffene die Preisgabe von Angaben über sich selbst nicht mehr kontrollieren. Der datenverarbeitende Dritte weiß in diesem Moment mehr über ihn als er selbst.

Beispielhaft lässt sich dies am Falle der US-amerikanischen Warenhauskette Target illustrieren. Target hatte anhand der Auswertung von Facebook- und Twitter-Daten und der statistisch nachgewiesenen Korrelation der Aufgeschlossenheit im dritten und vierten Monat schwangerer Frauen für neue, ihnen bis dahin unbekannte Markenprodukte ermittelt, welche Nutzerinnen der Dienste wahrscheinlich gerade schwanger waren und diese gezielt beworben (vgl. Jüngling 2013).

Das Vorgehen von Target wäre nach deutschem Datenschutzrecht unzulässig, da die Ausnahmeregelung in § 4 Abs. 2 S. 2 Nr. 2 lit. b BDSG) wegen der besonderen Sensibilität der betroffenen Daten nicht greift.

Datensparsamkeit

Der Grundsatz der Datenvermeidung und Datensparsamkeit greift das Postulat datenschutzfreundlicher Technikgestaltung auf und ist in § 3 a S. 1 BDSG und weiteren bereichsspezifischen Datenschutzregelungen normiert (vgl. etwa § 78 b SGB X).

Der Grundsatz der Datenvermeidung und Datensparsamkeit verlangt, Datenverarbeitungssysteme technisch so zu gestalten, dass personenbezogene Daten so wenig wie möglich oder gar nicht erhoben, verwendet oder genutzt werden (Gola und Schomerus 2015, S. 103 ff.). Schon im Rahmen der Konzeptionierung von IT-System ist daher zu prüfen, wie die Erhebung, Verarbeitung und Nutzung personenbezogener Daten vermieden oder möglichst eingeschränkt werden kann.

Big Data-Anwendungen verhalten sich diametral zum Grundsatz der Datenvermeidung und Datensparsamkeit, da ihre Analyseergebnisse von der Größe und Qualität der Basisdaten abhängen. Es ist das Ziel von Big Data-Anwendungen, möglichst große und aussagekräftige Datenmengen zu verarbeiten.

Obwohl ein grundlegendes Prinzip des modernen Systemdatenschutzes hat der Grundsatz der Datenvermeidung und Datensparsamkeit nur den Rang eines Programmsatzes oder einer Zielvorgabe. Der Verstoß gegen diese Zielvorgabe führt daher nicht zur materiellen Unzulässigkeit der Datenverarbeitung. Der Grundsatz der Datenvermeidung und Datensparsamkeit ist im BDSG auch nicht bußgeld- oder strafbewehrt (vgl. §§ 43, 44 BDSG). Die Aufsichtsbehörden können diesen Grundsatz nicht zwangsweise nach § 38 Abs. 5 S. 1 und 2 BDSG durchsetzen, sondern allenfalls beratend auf ihre Einhaltung hinwirken (Gola und Schomerus 2015, S. 104; Scholz 2011, S. 411). Gegenüber öffentlichen Stellen kann ein Verstoß gegen § 3 a S. 1 BDSG allerdings zu einer Beanstandung führen (Schulz 2013, Rn. 105).

Der Grundsatz der Datenvermeidung und Datensparsamkeit entfaltet jedoch mittelbare Wirkung, etwa im Rahmen von Abwägungsentscheidungen nach § 28 Abs. 1 S. 1 Nr. 2 BDSG, Vergabeverfahren der öffentlichen Hand oder als Konkretisierung des Maßstabs der im Verkehr erforderlichen Sorgfalt nach § 276 Abs. 2 BGB (Schulz 2013, Rn. 103; Grüneberg 2014, Rn. 18).

Technische Konkretisierungen des Grundsatzes der Datenvermeidung und Datensparsamkeit sind die Anonymisierung und die Pseudonymisierung personenbezogener Daten (§ 3 a S. 2 BDSG).

Von besonderem Interesse in Hinblick auf Big Data-Anwendungen ist die Anonymisierung, die dazu führt, dass Daten nicht mehr in den sachlichen Anwendungsbereich des BDSG fallen (Gola und Schomerus 2015, S. 96).

Anonymisierung setzt voraus, dass personenbezogene Daten derart verändert werden, dass sie nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können (§ 3 Abs. 6 BDSG).

Eine Person ist bestimmt oder bestimbar, wenn sie direkt oder indirekt identifiziert werden kann. Das ist beispielsweise der Fall, wenn sie sich von anderen Personen einer Gruppe eindeutig, etwa über bestimmte Merkmale unterscheiden lässt. Hierbei ist auch Zusatzwissen zu berücksichtigen, das die verantwortliche Stelle oder Dritte, die die personenbezogenen Daten verarbeiten oder nutzen, haben (Dammann 2014, S. 312; a. A. Gola und Schomerus 2015, S. 96).

Entscheidend für eine wirksame Anonymisierung ist, ob eine Reidentifizierung mit vertretbarem Aufwand möglich ist. Ob ein Anonymisierungsverfahren als ausreichend zu erachten ist, ist im Zeitpunkt der Anonymisierung für den gesamten Zeitraum zu prognostizieren, in dem die Daten gespeichert, verarbeitet und genutzt werden sollen (Roßnagel 2013, S. 563). Verbesserungen der Auswertungsmöglichkeiten sind also zu beachten (Dammann 2014, S. 315 f.).

Eine technische Variante wirksamer Anonymisierung ist es, Datensätze in Teile zu zerlegen und so durch zu würfeln, dass eine Deanonymisierung ausgeschlossen ist. Als eine valide Form der Anonymisierung wird es auch angesehen, eine so großen Zahl von Einzeldaten zu aggregieren und als einheitlichen Gruppendatensatz beispielsweise zu statistischen Zwecken so weiterzuverarbeiten, dass eine einzelne Person nicht mehr reidentifiziert werden kann (Poppenhäger 2003, S. 1627 ff.; Gola und Schomerus 2015, S. 96; Weichert 2013, S. 258 f.). Unzureichend ist dagegen das Löschen einzelner Identifikatoren wie etwa des Namens des Betroffenen oder einer identifizierenden Nummer, weil der Personenbezug schon bei wenigen Merkmalen möglich ist.

Für Big Data-Anwendungen ist Anonymisierung ein mögliches Konzept zur datenschutzkonformen Ausgestaltung. Mit Menge und Qualität der Daten und der Möglichkeit, anonymisierte Datensätze mit Big Data-Analysewerkzeugen auszuwerten, steigt jedoch das Reidentifizierungsrisiko (Weichert 2013, S. 254, 257). Die Anforderungen an Anonymisierungsverfahren sind für verantwortliche Stellen, die Big Data anwenden, daher besonders hoch (Katko und Babaei-Beigi 2014 S. 361 f.).

3.1.1.3 Fazit

Dass Big Data „*konträr zu den Prinzipien des Datenschutzes*“ liegt (so Roßnagel 2013, S. 562, 567), ist so allgemein formuliert nicht richtig. Das Datenschutzrecht eröffnet auch Big Data-Anwendungen rechtliche Spielräume, innerhalb derer diese datenschutzkonform gestaltet werden können (so auch Weichert 2013, S. 258).

Als technisch-organisatorische Maßnahmen wird hier zum Beispiel die Anreicherung von Datenbeständen mit datenschutzrechtlichen Metadaten vorgeschlagen, die Vorgaben zur Datennutzung enthalten (Weichert 2013, S. 259). Weiterhin kann der Big Data-Verarbeiter durch eine gute Strukturierung und Dokumentation seiner Anwendung für ein transparentes Datenmanagement sorgen (Dorschel und Nauerth 2013, S. 38). Wie das jüngst ergangene Google-Urteil des EuGH zeigt, steht der Einzelne Big Data-Verarbeiter auch nicht schutzlos gegenüber (EuGH NVwZ 2014, 857 ff.). An das Urteil anknüpfende rechtliche Vorkehrungen und Verfahren zur Durchsetzung der Betroffenenrechte können ebenfalls zur Datenschutzkonformität der Anwendungen beitragen.

Über das Individuum hinaus haben Big Data-Anwendungen Bedeutung für das freiheitlich-demokratische Gemeinwesen, denn von ihnen geht das Risiko nicht nur des gläsernen Menschen, sondern der durchleuchteten und gläsernen Gesellschaft aus (Schirrmacher 2013, S. 190). Big Data-Anwendungen können nicht nur zur statistischen Diskriminierung einzelner gesellschaftlicher Gruppen führen (Roßnagel 2013, S. 566), sondern auch dazu verführen, Entscheidungen an technisch-determinierte Prognosen zu

knüpfen und damit nicht mehr primär von demokratisch-legitimierte Verfahren abhängig zu machen. Dem Staat kommt daher gegenüber Big Data-Verfahren eine Gewährleistungsverantwortung für das demokratische Gemeinwesen zu, die über das subjektive Recht auf informationelle Selbstbestimmung hinausweist. Die datenschutzrechtlichen Aufsichtsbehörden, die über das Verfahren der Vorabkontrolle für rechtliche und demokratische Legitimation von Big Data-Anwendungen sorgen können, sind daher frühzeitig in Entscheidungen über die Anwendung von Big Data-Verfahren einzubeziehen.

3.1.2 Gesetzliche Erlaubnistarbestände und Interessenabwägung

Carsten Ulbricht

Die Verarbeitung von Big Data unterliegt, soweit personenbezogene Daten gespeichert, verarbeitet oder weitergegeben werden, dem bereits genannten Verbot mit Erlaubnisvorbehalt.

Der Begriff der personenbezogenen Daten umfasst alle Informationen, die über eine bezogene Person etwas aussagen oder mit ihr in Verbindung zu bringen sind. Das sind nicht nur klassische Daten, wie etwa der Name oder der Geburtsort, sondern auch Äußerungen, die sich auf einen bestimmten oder bestimmbaren Betroffenen beziehen, die Wiedergabe von mündlichen und schriftlichen Aussagen eines Betroffenen und die Darstellung des privaten oder des dienstlichen Verhalten eines Betroffenen (vgl. Gola und Schomerus 2015, BDSG, § 3 RdNr. 2 ff.; Dammann 2014, BDSG, § 3 RdNr. 7 ff. BDSG, § 3 RdNr. 7 ff.), aber auch Nutzungs- und Geo- und Bewegungsdaten.

Wenn und soweit nicht nur anonyme oder pseudonyme Daten (siehe Abschn. 3.1.7) verarbeitet werden sollen, ist die Verarbeitung von Big Data nach deutschem Recht also nur zulässig, wenn ein gesetzlicher Erlaubnistarbestand oder eine ausdrückliche Einwilligung des Betroffenen diese legitimiert.

Damit ist die Prüfung der möglichen Erlaubnistarbestände, beziehungsweise die dem jeweiligen Erlaubnistarbestand entsprechende Modellierung, zentraler Bestandteil eines jeden Big Data Projekts. Aus der Erfahrung ist daher eine frühzeitige datenschutzrechtliche Prüfung und Bewertung zu empfehlen, um eine spätere Verzögerung oder sogar einen Stopp des Projekts aufgrund einer Fertigstellung nachfolgenden Prüfungs- und Freigabeprozesses zu verhindern.

Bei der Frage nach etwaigen gesetzlichen Erlaubnistarbeständen ist im Hinblick auf die Art der zu bearbeitenden Daten und im Hinblick auf die Herkunft der Daten zu differenzieren. Gesetzliche Legitimationstarbestände finden sich im Telemediengesetz (TMG), im Telekommunikationsgesetz (TKG) und im Bundesdatenschutzgesetz (BDSG).

Da sich die Legitimationstarbestände und auch die jeweiligen Rechten und Pflichten der Gesetze im Hinblick auf die Erhebung und Verwendung von Daten unterscheiden, sind die Gesetze gegeneinander abzugrenzen. Erschwert wird die Abgrenzung teilweise durch die Überschneidung der Gesetze, insbesondere im Falle der Zugangsvermittlung oder

Datengewinnung über das Internet. Darüber hinaus differenzieren die Gesetze teilweise zwischen verschiedenen Arten von Daten, nämlich zwischen Bestands-, und Nutzungs- bzw. Verkehrsdaten.

3.1.2.1 Anwendungsbereiche und Abgrenzungen von TMG, TKG und BDSG

Das Telemediengesetz (TMG), das seit dem Jahr 2007 das Teledienstegesetz (TDG) und den Staatsvertrag über Mediendienste (MDStV) ersetzt, gilt für den Bereich der Telemedien. Dies sind nach § 1 Abs. 1 TMG alle elektronischen Informations- und Kommunikationsdienste, die keine Telekommunikationsdienste (§ 3 Nr. 24 TKG), telekommunikationsgestützte Dienste (§ 3 Nr. 25 TKG) oder Rundfunk (§ 2 Rundfunk Staatsvertrag) darstellen. Zu den umfassendsten Diensten zählen dabei insbesondere solche der Individualkommunikation (zum Beispiel E-Mail und Datendienste). Das TMG findet damit Anwendung auf die Datengewinnung durch Anbieter entsprechender Telemedien, einschließlich der öffentlichen Stellen. Der Dienstanbieter wird in § 2 Abs. 1 Nr. 1 TMG definiert als eine „natürliche oder juristische Person, die eigene oder fremde Telemedien zur Nutzung bereithält oder den Zugang zur Nutzung vermittelt“.

Das Telekommunikationsgesetz (TKG) hingegen ist der Nachfolger des früheren Fernmeldeanlagengesetzes (FAG) und regelt Telekommunikationsdienste. Diese sind nach der Definition von § 3 Nr. 24 TKG „in der Regel gegen Entgelt erbrachte Dienste, die ganz oder überwiegend in der Übertragung von Signalen über Telekommunikationsnetze bestehen, einschließlich Übertragungsdienste in Rundfunknetzen“. Das TKG betrifft damit im Wesentlichen die Telekommunikationsinfrastruktur sowie die hierüber erbrachten Telekommunikationsdienstleistungen.

Das Bundesdatenschutzgesetz (BDSG) enthält die datenschutzrechtlichen Grundsätze und umfangreichsten Regelungen zur Erhebung, Speicherung und Verarbeitung personenbezogener Daten. Es dient als Auffangtatbestand, ist insofern aber gegenüber den oben genannten Spezialgesetzen als subsidiär, also nachrangig anzusehen.

Bei der Gestaltung eines Big Data Projektes muss je nach der Art der Datenerhebung (z. B. über Telekommunikations- oder Telemediendienste) also grundsätzlich unterschieden werden, wann jeweils die bereichsspezifischen Regelungen des TKG oder TMG greifen und wie diese wiederum vom BDSG abzugrenzen sind.

Dazu hat sich das sog. Schichtenmodell (Schaar 2002, Rn. 247 ff.) herausgebildet. Danach ist zu unterscheiden, welche Ebene funktionell betroffen ist. Die Inhaltsebene regelt das BDSG, das TMG ist auf die Transportbehälterebene anzuwenden und der Datentransport selbst unterliegt dem TKG.

Wie oben bereits dargestellt, findet das BDSG in seiner Rolle als Auffanggesetz darüber hinaus Anwendung, wenn TKG und TMG keine Spezialregelungen treffen.

3.1.2.2 Der Legitimationstatbestand der Einwilligung

Gemeinsame Legitimationsmöglichkeit aller vorgenannten Gesetze ist die Einwilligung. Um diese Wirkung entfalten zu können, muss sie auf dem freien Willen des Betroffenen beruhen (§ 4a Abs. 1 BDSG).

Die Einwilligung setzt weiter voraus, dass der Betroffene über den Gegenstand der Einwilligung informiert ist und die Tragweite der Entscheidung überblicken kann. § 4a BDSG enthält deshalb Vorgaben zur Information des Betroffenen. So ist der Betroffene auf den vorgesehenen Zweck der Erhebung, Verarbeitung oder Nutzung seiner Daten, auf den sich die Einwilligung bezieht, ausdrücklich hinzuweisen. Dementsprechend muss die Information so formuliert sein, dass sie verständlich und vollständig den Umfang und die Zwecke umschreibt, für die die Daten verwendet werden sollen. Darüber hinaus muss der Betroffene im Rahmen der Einwilligung der entsprechend erteilten Datenschutzinformation in aller Regel aktiv zustimmen (Opt-In).

Auch wenn die Legitimation über eine Einwilligung der Betroffenen in einigen Fällen ein gangbarer Weg sein kann, dürften sich die allermeisten Big Data Modelle mangels Kontaktmöglichkeit beziehungsweise Zugang zu den betroffenen Personen als wenig realistisch darstellen. Teilweise wird sich die Einholung einer Einwilligung auch angesichts der Menge und Heterogenität der Daten schon grundsätzlich als schwierig darstellen. Für die Daten, für die bei der Erhebung eine Einwilligung eingefordert wurde, müsste diese informiert, freiwillig, ausreichend bestimmt und formgerecht gegeben worden sein. Dies dürfte in vielen Fällen zweifelhaft sein, weil die Zwecke bei Big Data in vielen Fällen unbestimmt sind, die Folgen nicht absehbar waren und als Alternative zur Einwilligung oft nur angeboten wird, auf die gewünschte Dienstleistung zu verzichten.

Ob eine Einwilligung eine mögliche Datenverwendung rechtfertigen kann ist stets am Einzelfall zu prüfen. In vielen Fällen wird die Einwilligung aber kein taugliches Instrument sein, um ein Big Data-Projekt datenschutzrechtlich zu legitimieren.

Ob Big Data-Analysen auch ohne Einwilligung zulässig sind, hängt von den Daten, von der jeweiligen Phase der Verarbeitung und vom Verarbeitungszweck ab. Für alle entsprechenden Gestaltungen, bei denen die betroffenen Personen nicht in die jeweilige Erhebung, Speicherung und Verarbeitung ihrer Daten eingewilligt haben, sollen deshalb nachfolgend die weiteren gesetzlichen Erlaubnistratbestände dargestellt werden, über die Datenerhebung, -speicherung oder -verarbeitung legitimiert werden kann.

3.1.2.3 Weitere Befugnisse zur Datenverarbeitung nach TMG und TKG

In den Fällen in denen das TMG, etwa bei einer Zugangsvermittlung zu den Daten über das Internet, als bereichsspezifische Vorschrift eingreift, legt § 12 Abs. 1 TMG fest, dass ein Dienstanbieter personenbezogene Daten – also Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (§ 3 Abs. 1 S. 1 BDSG) – zur Bereitstellung von Telemedien nur erheben und verwenden darf, soweit es gesetzliche Vorschriften, die sich ausdrücklich auf Telemedien beziehen, dies erlauben oder der Nutzer eingewilligt hat.

Sobald der Datenverarbeiter Daten über ein Telemedium erheben will, hat er den Nutzer nach § 13 Abs. 1 TMG zu Beginn des Nutzungsvorgangs über Art, Umfang und Zweck der Erhebung und Verwendung der personenbezogenen Daten in allgemein verständlicher Form zu unterrichten, sofern eine solche Unterrichtung nicht bereits erfolgt ist. Bei Verwendung eines automatisierten Verfahrens, das eine spätere Identifizierung des Nutzers

ermöglicht und eine Erhebung oder Verwendung personenbezogener Daten darstellt, ist der Nutzer zu Beginn des Verfahrens zu unterrichten. Der Inhalt der Unterrichtung muss für jeden Nutzer jederzeit abrufbar sein. Bezüglich der Datenverarbeitung differenziert das TMG zwischen Bestands- und Nutzungsdaten.

Bestandsdaten sind Daten, die für Begründung, inhaltliche Ausgestaltung oder Änderung eines Vertragsverhältnisses zwischen dem Dienstanbieter und den Nutzer über die Nutzung von Telemedien erforderlich sind (§ 14 Abs. 1 TMG), beispielsweise Namen, Benutzerkennung oder Anschrift. Die Speicherung ist danach am Zweck der Erhebung und der Erforderlichkeit auszurichten. Die Weitergabe solcher Bestandsdaten bestimmt sich nach § 14 Abs. 2 TMG und beschränkt sich im Wesentlichen auf Auskünfte gegenüber öffentlichen Stellen in Bereichen wie Strafverfolgung und Gefahrenabwehr.

Nutzungsdaten dürfen verarbeitet werden, wenn dies erforderlich ist, um die Inanspruchnahme von Telemedien zu ermöglichen und abzurechnen (§ 15 Abs. 1 TMG). Als Beispiele werden Merkmale zur Identifikation des Nutzers, Angaben über Beginn und Ende sowie des Umfangs der jeweiligen Nutzung und Angaben über die von Nutzern in Anspruch genommenen Telemedien aufgeführt. Nutzungsdaten dürfen etwa zu Abrechnungszwecken verwendet werden.

Schließlich darf der Dienstanbieter nach § 15 Abs. 3 TMG zu Werbe – und Forschungszwecken oder zur bedarfsgerechten Gestaltung der Telemedien pseudonyme Nutzungsprofile erstellen. Entsprechend pseudonymisierte Nutzungsdaten dürfen zum Zwecke der Marktforschung schließlich auch an andere Dienstanbieter übermittelt werden (§ 15 Abs. 5 S. 3 TMG).

Wenn und soweit im Rahmen von Big Data Daten über Telemedien erhoben werden, sind diese gesetzlichen Vorgaben des Telemediengesetzes einzuhalten.

Die Zulässigkeit der Erhebung und Verarbeitung grosser Datenmengen hängt also zunächst davon ab, welche Daten erforderlich sind, um die Inanspruchnahme des Telemediums zu ermöglichen. Dies hängt von der konkreten Anwendung ab. So muss z.B. eine Gesundheitsapp auch gewisse Gesundheitsdaten speichern und verarbeiten, um ihren eigentlichen Zweck zu erfüllen. Außerhalb der zur Inanspruchnahme notwendigen Daten, scheint vor allem die Möglichkeit der pseudonymisierten Verarbeitung und Weitergabe von Nutzungsdaten nach § 15 Abs.3 TMG geeignet, um Big Data Analysen über Telemedien gewonnene Daten zu ermöglichen.

3.1.2.4 Weitere Befugnisse zur Datenverarbeitung nach dem TKG

Sollen große Datenmengen aus der Telekommunikationsinfrastruktur beziehungsweise im Rahmen der hierüber erbrachten Telekommunikationsdienstleistungen erhoben werden, ist die Rechtskonformität auf Grundlage des Telekommunikationsgesetz (TKG) zu prüfen. Denkbare Anwendungsszenarien sind etwa die Auswertung von Standortdaten oder Bewegungsmuster der Nutzer von Mobiltelefonen über die Telekommunikationsinfrastrukturen.

Das TKG differenziert zwischen Bestandsdaten, also die Informationen die für die Begründung, inhaltliche Ausgestaltung, Änderung oder Beendigung eines Vertragsverhältnisses erhoben werden (§ 3 Nr. 3 TKG) und Verkehrsdaten, die bei der Erbringung

eines Telekommunikationsdienstes erhoben, verarbeitet oder genutzt werden (§ 3 Nr. 30 TKG).

Grundsätzlich sind Verkehrsdaten nach Beendigung der Verbindung unverzüglich zu löschen. Auf Grundlage einer Einwilligung des Betroffenen dürfen entsprechende Daten zur Vermarktung, zur bedarfsgerechten Gestaltung oder zur Bereitstellung von Diensten mit Zusatznutzen auch weitergehend verwendet werden. Dann muss dem Betroffenen allerdings Zweck und Dauer der Verarbeitung mitgeteilt werden und der Zeitraum der Speicherung darf das erforderliche Maß nicht überschreiten. Die weitergehende Ausnahmen zur Speicherung von Verkehrsdaten aus § 97 ff. TKG sind sehr eng formuliert und werden im Hinblick auf die jeweilige Erforderlichkeitsschwelle im Bereich Big Data wohl nicht als Ermächtigungsgrundlage herangezogen werden können.

Während die Speicherung von großen Datenmengen im Bereich des Telekommunikationsrechts noch teilweise legitimiert werden kann, wird sich eine weitergehende Verarbeitung oder Auswertung der so gewonnenen Daten in den meisten Fällen nur über eine spezifische Einwilligung datenschutzkonform gestalten lassen. Den Telekommunikationsbetreibern eröffnet sich im Rahmen des Vertragsschlusses mit ihren Kunden allerdings eine Möglichkeit, die Voraussetzungen einer Einwilligung zu erfüllen.

3.1.2.5 Weitere Befugnisse zur Datenverarbeitung nach dem BDSG

Wenn bei einem Big Data-Projekt die bereichsspezifischen Regeln des TMG oder TKG nicht eingreifen, so ist die Erhebung, Verarbeitung und Nutzung personenbezogener Daten nach § 4 Abs. 1 BDSG zulässig, wenn das BDSG die Datenverarbeitung erlaubt oder der Betroffene eingewilligt hat. Das BDSG differenziert hinsichtlich der Erlaubnistatbestände zwischen öffentlichen und nicht-öffentlichen Stellen, enthält dementsprechend unter einer Gesetzesüberschrift gleichsam zwei unterschiedliche Gesetze.

Im öffentlichen Bereich unterscheidet das Gesetz zwischen der Erhebung, (§ 13 BDSG), der Speicherung, Veränderung und Nutzung (§ 14 BDSG), sowie der Übermittlung an öffentliche und private Stellen (§§ 15 bzw. 16 BDSG). Zulässig ist der Datenumgang in allen Fällen nur, soweit dies zur Erfüllung der Aufgaben der Stelle erforderlich ist. Dies muss sich aus spezialgesetzlichen Regelungen ergeben, insoweit greift das BDSG die allgemeinen Anforderungen an die Gesetzmäßigkeit der Verwaltung (Art. 20 Abs. 3 GG) auf. §§ 13 und 14 BDSG enthalten zusätzlich differenzierte Regelungen für den Umgang mit besonderen personenbezogenen Daten, § 14 Abs. 2 BDSG normiert zudem Voraussetzungen einer Verwendung für andere Zwecke als den der Erhebung.

Bei der Datenverarbeitung im nicht-öffentlichen Bereich differenziert das BDSG zwischen dem Datenumgang für eigene Zwecke (§ 28 BDSG) und dem geschäftsmäßigen Umgang zum Zwecke der Übermittlung (§ 29 BDSG) und speziell zur Übermittlung in anonymisierter Form (§ 30 BDSG). Demgemäß ist bei der Gestaltung von Big Data-Projekten auch danach zu differenzieren, ob der Datenumgang eigenen oder geschäftsmäßigen Zwecken der Übermittlung dient.

Ermächtigung der Datenerhebung, -speicherung und -verarbeitung nach § 28

Abs. 1 BDSG

Im Rahmen des den Datenumgang für eigene Zwecke regelnden § 28 Abs. 1 Satz 1 BDSG ist eine Datenerhebung, -speicherung und -verarbeitung zulässig, soweit dies der Zweckbestimmung des Vertragsverhältnisses dient (Nr. 1), es zur Wahrung berechtigter Interessen der verantwortlichen Stelle erforderlich ist (Nr. 2) oder die Daten allgemein zugänglich sind und kein Grund zur Annahme besteht, dass schutzwürdige Interessen des Betroffenen überwiegen (Nr. 3).

Die Zweckbestimmung des Vertrages (§ 28 Abs. 1 Satz 1 Nr. 1 BDSG) rechtfertigt den jeweiligen Datenumgang nur, wenn dieser für die Begründung, Durchführung oder Beendigung eines rechtsgeschäftlichen oder rechtsgeschäftsähnlichen Schuldverhältnisses erforderlich ist. Dabei wird die Erforderlichkeit in der Regel restriktiv interpretiert. Insofern sind nur in einigen wenigen Fällen eine Legitimation über § 28 Abs. 1 Satz 1 Nr. 1 BDSG denkbar, in denen die Big Data-Analyse quasi der eigentliche Inhalt des Vertrages ist. Ansonsten wird der Zweck die Zweckbestimmung des Vertrages Big Data in aller Regel nicht rechtfertigen können.

Im Gegensatz dazu kommt es bei dem Legitimationstatbestand des § 28 Abs. 1 Satz 1 Nr. 2 BDSG auf eine Abwägung berechtigter Interessen der datenverarbeitenden Stelle mit entsprechend schutzwürdigen Interessen des Betroffenen an.

Hierzu hat der Bundesgerichtshof (BGH NJW 1986, 2505) ausgeführt: „Der wertausfüllende Begriff der „schutzwürdigen“ Belange verlangt eine Abwägung des Persönlichkeitsrechts des Betroffenen und des Stellenwerts, den die Offenlegung und Verwendung der Daten für ihn hat, gegen die Interessen der speichernden Stelle und der Dritten, für deren Zweck die Speicherung erfolgt. Dabei sind Art, Inhalt und Aussagekraft der beanstandeten Daten an den Angaben und Zwecken zu messen, denen ihre Speicherung dient. Nur wenn diese am Verhältnismäßigkeitsgrundsatz ausgerichtete Abwägung, die die speichernde Stelle vorzunehmen hat, keinen Grund zur Annahme bietet, dass die Speicherung der in Frage stehenden Daten zu dem damit verfolgten Zweck schutzwürdige Belange des Betroffenen beeinträchtigt, ist die Speicherung zulässig.“

Problematisch scheint insofern aber das Erfordernis, dass die entsprechende Datenanalyse den berechtigten Interessen eines auswertenden Unternehmens nicht nur dienlich, sondern hierfür nach dem Gesetz erforderlich sein muss (Gola und Schomerus 2015, BDSG, § 28 Rn. 14). Erforderlich meint in diesem Zusammenhang, dass es zu der konkreten Verwendung keine objektiv zumutbare Alternative gibt. Dies soll nach der überwiegenden Literaturmeinung jedoch nicht auf Fälle beschränkt sein, in denen die Verarbeitung zwingend erforderlich ist (Simitis 2014, BDSG, § 28 Rn. 108).

Mangels konkreter Beschränkung im BDSG können grundsätzlich, sowohl ideelle, als auch wirtschaftliche Interessen von der verantwortlichen Stelle geltend gemacht werden. So wurde von der Rechtsprechung das Interesse von Unternehmen über sog. „harte“ Bonitätsdaten ihrer Vertragspartner (z. B. Abgabe einer eidesstattlichen Versicherung u. ä.) informiert oder gewarnt zu werden, als so gravierend angesehen, dass dieses Interesse regelmäßig entgegenstehenden Interessen des Betroffenen überwiegt. Insofern sind Ansätze

denkbar, Big Data-Analysen über berechtigte Interessen des Unternehmens zu rechtfertigen. In Anbetracht der jeweils individuell vorzunehmenden Interessenabwägung und dem grundsätzlich festzulegenden spezifischen Verarbeitungszweck, ist die die jeweilige Systematik der Auswertung großer Datenmengen auf Grundlage von § 28 Abs. 1 Satz 1 Nr. 2 BDSG stets einer spezifischen Prüfung zu unterziehen, die der nachfolgend dargestellten Systematik folgt.

Grundsätzliches Prüfungsmodell für eine datenschutzrechtliche Interessenabwägung

Auf Grundlage der am 9.4.2014 von der europäischen „Article 29 Data Protection Working Group“ veröffentlichten Stellungnahme „Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC“ sollte sich eine datenschutzrechtliche Interessenabwägung im Rahmen einer europarechtskonformen Auslegung an den nachfolgenden Schritten orientieren:

SCHRITT 1: Definition des Interesses der datenverarbeitenden Stelle

SCHRITT 2: Prüfung der datenschutzrechtlichen Legitimität des Interesses

SCHRITT 3: Prüfung der Erforderlichkeit der konkreten Datenverarbeitung im Hinblick auf das Interesse der datenverarbeitenden Stelle

SCHRITT 4: Vorläufige Prüfung eines etwaigen Überwiegens der Interessen des Betroffenen

Hierbei ist das Interesse der datenverarbeitenden Stelle (wirtschaftliche, öffentliche oder andere Interessen) ebenso einzubeziehen, wie die Sensibilität der zu verarbeitenden Daten, die Person des Betroffenen (Minderjährige, Arbeitnehmer etc.), die Art der Datenverarbeitung (Big Data, Profiling, Veröffentlichung etc.) und etwaige (positive wie negative) Folgen für den Betroffenen.

SCHRITT 5: Finale Prüfung eines etwaigen Überwiegens der Interessen des Betroffenen bei Berücksichtigung spezifischer Maßnahmen zur Umsetzung der Datenschutzanforderungen (sog. „safeguards“)

In diesem Schritt sollen organisatorische oder technische Maßnahmen identifiziert werden, die die Erfüllung der Datenschutzanforderungen unterstützen beziehungsweise ermöglichen. Denkbare Maßnahmen sind (frühzeitige) Anonymisierungs- oder Pseudonymisierungs- oder Aggregationstechniken, Informations- oder Widerspruchsmöglichkeiten, Maßnahmen zur Förderung der Datensparsamkeit (Datenlöschung unmittelbar nach Verwendung etc.)

SCHRITT 6: Dokumentation der Einhaltung der Datenschutzanforderungen (Schritte 1–5) und Förderung von Transparenz

SCHRITT 7: Gewährleistung der Betroffenenrechte

Dieses Prüfungsmodell kann bei jedweder datenschutzrechtlichen Interessenabwägung auch im Rahmen der nachfolgenden Legitimationstatbestände herangezogen werden.

Ist die jeweilige Verarbeitung von Big Data unter Zugrundelegung der vorgenannten Prüfung zur Wahrung der Interessen der datenverarbeitenden Stelle erforderlich, so ist diese nach § 28 Abs. 1 Satz 1 Nr. 2 BDSG legitimiert.

Als davon unabhängige, weitere Legitimationsgrundlage ermöglicht § 28 Abs. 1 Satz 1 Nr. 3 BDSG die Verwendung von allgemein zugänglichen Daten, wenn nicht schutzwürdige Interesse des Betroffenen an dem Ausschluss der Verarbeitung oder Nutzung gegenüber dem berechtigten Interesse der verantwortlichen Stelle offensichtlich überwiegen.

Allgemein zugänglich sind sämtliche Informationsquellen, „die sich sowohl ihrer technischen Ausgestaltung als auch ihrer Zielsetzung nach dazu eignen, einem individuell nicht bestimmbaren Personenkreis Informationen zu vermitteln“ (Simitis 2014, BDSG, § 28 Rn. 151). Dazu gehören damit Fernsehen, Zeitungen oder Radio, sowie öffentliche Register und auch Informationen aus dem Internet, soweit der Zugriff für jedermann, unter anderem durch Suchmaschinen eröffnet sein soll. Bezuglich etwaiger Daten aus sozialen Netzwerken wird gemeinhin danach unterschieden, ob der jeweilige Bereich für die Allgemeinheit freigegeben oder nur einem beschränkten Nutzerkreis zugänglich gemacht wird. In letzterem Fall fehlt es an der für diesen Legitimationstatbestand (Eichler und Kamp 2014, § 28 Rn. 91) erforderlichen allgemeinen Zugänglichkeit.

Für Daten aus diesen allgemein zugänglichen Quellen ist eine Verarbeitung nur ausnahmsweise unzulässig, nämlich dann, wenn ein schutzwürdiges Interesse des Betroffenen am Ausschluss der Verarbeitung oder Nutzung gegenüber dem berechtigten Interesse der verantwortlichen Stelle offensichtlich überwiegt. Das Gewicht der Interessen der Betroffenen muss deutlich höher sein als das der Interessen der verarbeitenden Stelle. Weiter muss das Überwiegen leicht erkennbar sein, das heißt die Verletzung der Interessen ist für einen unvoreingenommenen verständigen Beobachter ohne weiteres wahrzunehmen (Gola und Schomerus 2015, BDSG, § 28 Rn. 31). Die verantwortliche Stelle muss für entsprechende Daten also nur eine summarische Prüfung vornehmen, ob sich ein offensichtliches Interesse des Betroffenen quasi aufdrängt. Eine nennenswerte Einzelfallprüfung ist nicht geboten.

Dieser aus der Informationsfreiheit folgende Tatbestand ist damit eine wichtige denkbare Grundlage um Big Data-Analysen aus den genannten Quellen zu ermöglichen.

Im Hinblick auf die genannte Interessenabwägung sollte die Big Data Verwendung dennoch so konfiguriert werden, dass die Datenerhebung sich ganz im Sinne des Grundsatzes auf Datensparsamkeit auf das Notwendige beschränken und nicht nur die Erhebung, sondern auch die Zusammenführung solcher Datenbestände ebenfalls die Voraussetzungen erfüllt.

Des Weiteren setzen die im Datenschutzrecht verankerten Prinzipien der Zweckbegrenzung und -bindung einer personenbezogenen Datenerhebung und -auswertung deutliche Grenzen. Auch wenn einer der Mehrwerte von Big Data-Analysen gerade die unbegrenzte Auswertung von Daten ist, um entsprechende Erkenntnisse zu gewinnen, sollte versucht werden, den Zweck der jeweiligen Big Data-Analyse gerade auch im Rahmen der Datengewinnung aus öffentlich zugänglichen Quellen (z. B. Internet) möglichst konkret zu definieren und einzuschränken und dann die entsprechende Datenerhebung und auch -auswertung an diesem Zweck auszurichten.

Ermächtigung der Datennutzung und -verarbeitung nach § 28 Abs. 2 BDSG

Sämtliche vorgenannten Erlaubnistatbestände des § 28 Abs. 1 BDSG unterliegen dem bereits dargestellten Grundsatz der Zweckbindung. Mithilfe dieses Grundsatzes sollen Einschränkungen des Rechts auf informationelle Selbstbestimmung auf das Unvermeidbare reduziert werden. Durch den Zweckbindungsgrundsatz wird gefordert, dass das Erheben von personenbezogenen Daten nur für festgelegte eindeutige und rechtmäßige Zwecke erfolgen darf und dass eine Weiterverarbeitung (Nutzen, Verarbeiten) der personenbezogenen Daten nur auf eine Weise erfolgt, die mit dem Grundsatz der Zweckbestimmung vereinbar ist.

Eine Ausnahme vom generellen Zweckbindungsgrundsatz stellt § 28 Abs. 2 S. 1 BDSG dar. Dieser lässt eine Übermittlung oder Nutzung personenbezogener Daten auch für andere Zwecke bei der Wahrung berechtigter Interessen der verantwortlichen Stelle (Nr. 2) und der Verwendung allgemein zugänglicher Daten (Nr. 3) zu.

Mit der Wahl eines „anderen Zwecks“ kann die verantwortliche Stelle das ursprüngliche Verwendungsziel in den vorgenannten Fällen also ändern. Sie ist aber natürlich dennoch im Rahmen des § 28 BDSG verpflichtet, die Daten ausschließlich für „eigene Geschäftszwecke“ zu verarbeiten.

Diese Vorschrift ermöglicht im Rahmen von Big Data also, dass Daten, die im Rahmen eines anderen Zwecks auf Grundlage von § 28 Abs. 1 S. 1 Nr. 2 oder Nr. 3 BDSG erhoben worden sind, zu Zwecken der Datenanalyse übermittelt oder genutzt werden.

Die Zulässigkeitsgrenzen des § 28 Abs. 1 S. 1 Nr. 2 BDSG beziehungsweise Nr. 3 bleiben gleichsam bestehen. Überwiegende schutzwürdige Interessen der Betroffenen schließen also eine Verwendung aus, die berechtigte Interessen der verantwortlichen Stelle sichern soll, offensichtlich überwiegende Interessen der Betroffenen eine Anknüpfung an Daten, die allgemein zugänglich sind oder veröffentlicht werden dürfen (Simitis 2014, BDSG, § 28 Rn. 173).

Demnach ist auch bei einer etwaigen Datenweitergabe eine entsprechende Interessenabwägung vorzunehmen.

Ermächtigung der Datennutzung und -verarbeitung zu Zwecken der Werbung und des Adresshandels nach § 28 Abs. 3 BDSG

§ 28 Abs. 3 BDSG regelt die Verarbeitung und Nutzung personenbezogener Daten für Zwecke des Adresshandels oder der Werbung.

Neben verschiedenen genauer definierten Optionen zur datenschutzrechtlichen Legitimation über eine Einwilligung in Werbung oder Adresshandel findet sich in dieser Vorschrift insbesondere das sogenannte „Listenprivileg“. Voraussetzung für das Listenprivileg ist, dass die Daten über die entsprechende Personengruppe listenmäßig zusammengefasst sind und aus den einzelnen Datensätzen kein schutzwürdiges Interesse der jeweiligen Person verletzt wird.

Da das Listenprivileg sich nur auf listenmäßig zusammengefasste Daten bezieht, die zudem noch enumerativ aufgezählt sind, scheint dieser Erlaubnistaatbestand nicht geeignet, Big Data-Analysen, die ja gerade mit großen und oft unstrukturierten Daten umgehen, zu rechtfertigen.

Ermächtigung des Scoring auf Grundlage von § 28b BDSG

Big Data-Analysen sind kein Selbstzweck, sondern dienen in aller Regel einer bestimmten Zielsetzung. Typische Zielsetzungen im Bereich Big Data sind die Klassifikation, Bewertung beziehungsweise einer Segmentierung.

So bietet § 28b BDSG eine spezifische Grundlage im Rahmen der Begründung, Durchführung oder Beendigung eines Vertragsverhältnisses einen Wahrscheinlichkeitswert für ein bestimmtes zukünftiges Verhalten des Betroffenen zu erheben oder zu verwenden (näher hierzu in Abschn. 3.1.9).

Ermächtigung der Datenerhebung und -speicherung auf Grundlage von § 29 Abs. 1 BDSG

In Abgrenzung zu § 28 BDSG regelt § 29 BDSG das geschäftsmäßige Erheben, Speichern, Verändern oder Nutzen personenbezogener Daten zum Zweck der Übermittlung. Die Erlaubnisnorm bezieht sich damit auf Datenverarbeitungsvorgänge, die zum Zwecke der Datenübermittlung an Dritte stattfinden. Wenn also die Datenübermittlung an Dritte der eigentliche Geschäftsgegenstand der jeweiligen Big Data-Analyse ist, so ist der Datenumgang nur zulässig, wenn eine der Voraussetzungen des § 29 Abs. 1 Nr. 1 bis 3 BDSG vorliegt.

Die Optionen des § 29 Abs. 1 Nr. 1 und Nr. 2 BDSG entsprechen weitgehend den Tatbestandsvoraussetzungen des § 28 Abs. 1 S. 1 Nr. 2 und Nr. 3 BDSG. Insoweit sind hier die bereits oben erläuterten Prüfungsschritte vorzunehmen.

§ 29 Abs. 1 Nr. 2 BDSG bestimmt, dass Daten aus allgemein zugänglichen Quellen entnommen werden können oder wenn die verantwortliche Stelle sie veröffentlichen durfte, es sei denn, dass das schutzwürdige Interesse des Betroffenen an dem Ausschluss der Erhebung, Speicherung oder Veränderung offensichtlich überwiegt. Für öffentlich zugängliche Daten ist eine entsprechende Datenverwendung also auch zum Zwecke der Übermittlung als zulässig anzusehen, soweit keine schutzwürdigen Interessen des Betroffenen offensichtlich überwiegen.

Der ausfüllungsbedürftige Begriff des „schutzwürdigen Interesses“ verlangt auch hier eine Abwägung der Interessen des Betroffenen an dem Schutz seiner Daten und des Stellenwerts, den die Offenlegung und Verwendung für ihn hat, mit den Interessen des

jeweiligen Datenanbieters für deren Zwecke die Speicherung oder Verarbeitung erfolgt. Bei der Abwägung sind dann entsprechend Art, Inhalt und Aussagekraft der beanstandeten Daten an den Aufgaben und Zwecken zu messen, denen die Datenerhebung und Speicherung dienen (vgl. Gola und Schomerus 2015, § 29 Rdnr. 11). Kann die datenerhebende Stelle darlegen und erforderlichenfalls beweisen, dass sie die Daten zur Erreichung des angestrebten rechtlich zulässigen Zwecks braucht, darf sie die Daten erheben, solange entgegenstehende schutzwürdige Interessen des Betroffenen nicht erkennbar sind.

Ermächtigung der Datenweitergabe auf Grundlage von § 29 Abs. 2 BDSG

Für die Übermittlung personenbezogener Daten an Dritte ist darüber hinaus § 29 Abs. 2 BDSG zu prüfen.

Diese Vorschrift sieht vor, dass die Übermittlung im Rahmen des Zweckes nach Abs. 1 nur dann zulässig ist, wenn erstens der Dritte, dem die Daten übermittelt werden, ein berechtigtes Interesse an ihrer Kenntnis glaubhaft dargelegt hat und zweitens kein Grund zu der Annahme besteht, dass der Betroffene ein schutzwürdiges Interesse an dem Ausschluss der Übermittlung hat.

Wenn die Daten rechtskonform erhoben worden sind, dürfte ein schutzwürdiges Interesse wohl nur wenigen Fällen entgegenstehen. Allerdings dürfte sich die weitere Voraussetzung der Darlegung eines berechtigten Interesses des Dritten gemäß § 29 Abs. 2 Nr. 1 als problematisch darstellen.

Eine Weitergabe ist danach nur zulässig, wenn der betreffende Dritte ein hinreichendes berechtigtes Interesse an der Übermittlung darlegt. In Fällen der Übermittlung nach § 29 BDSG wird also eine entsprechende Abfrage der berechtigten Interessen des Dritten integriert werden müssen.

Diese Abfrage muss zumindest stichprobenmäßig verifiziert werden, um den Anforderungen, die das BDSG an die datenverarbeitende Stelle anlegt, gerecht zu werden.

Ermächtigung zur Datenerhebung und -verarbeitung auf Grundlage § 30a BDSG

Nach § 30a BDSG ist eine Big Data Analyse zu Zwecken der „geschäftsmäßigen“ Markt- und Meinungsforschung als zulässig anzusehen, wenn keine schutzwürdigen Interessen des Betroffenen an dem Ausschluss der Datenverarbeitung bestehen, oder die Daten aus allgemein zugänglichen Quellen stammen beziehungsweise das Marktforschungsunternehmen diese veröffentlichen durfte und schutzwürdige Interessen des Betroffenen nicht „offensichtlich“ entgegenstehen. Diese Sonderregelung wurde nach der Gesetzesnovelle im Jahre 2009 mit ins Gesetz aufgenommen, um den Besonderheiten der Markt- und Meinungsforschung für „eine nachhaltige demokratische und wirtschaftliche Entwicklung in der Bundesrepublik“ gerecht zu werden.

Gemäß § 30a BDSG sind das Erheben und insbesondere das Nutzen von personenbezogenen Daten erlaubt, wenn sie allgemein zugänglich sind und kein schutzwürdiges Interesse des Betroffenen an dem Ausschluss der Datenverwendung offensichtlich überwiegt. Diesbezüglich kann auf die vorgehenden Ausführungen referenziert werden.

Ist die Erhebung von Daten nach den obenstehenden Ausführungen zulässig, so dürfen diese Daten auch gespeichert und verarbeitet werden (indizierende Wirkung der Erhebung, vgl. Ehmann 2014, BDSG § 30a RdNr. 124), wobei allerdings immer das Gebot der frühestmöglichen Anonymisierung zu beachten ist. Diese schreibt § 30a BDSG zum Schutz der gespeicherten Daten vor.

Für eine hinreichende Anonymisierung müssen dann sämtliche Angaben gelöscht werden, aus denen Rückschlüsse auf den Betroffenen gezogen werden können (weiterführend siehe Abschn. 3.1.3.1). Ab dem Zeitpunkt der Anonymisierung unterfallen die Daten schließlich nicht mehr dem BDSG und können daher auch frei an die Auftraggeber übermittelt werden. Eine nicht-anonymisierte Übermittlung ist nach § 30a BDSG nicht erlaubt. Hierfür müssen bereits bei der Erhebung die strengeren Regelungen des § 29 BDSG beachtet werden.

§ 30a BDSG verbietet zudem im Rahmen einer gestuften Zweckbindung, dass die Daten aus allgemein zugänglichen Quellen für einen anderen – der Markt- oder Meinungsforschung – fremden Zweck verarbeitet oder genutzt werden, wenn sie zuvor nicht so anonymisiert worden sind, dass ein Personenbezug nicht mehr – auch nicht durch den Empfänger – hergestellt werden kann. Damit soll u. a. ausgeschlossen werden, dass die Daten für forschungsfremde Werbezwecke übermittelt oder genutzt werden, da hierfür in aller Regel ein Personenbezug notwendig ist.

Zusammenfassend kann eine Erhebung oder Verarbeitung von Daten im Rahmen von Big Data-Projekten über § 30a BDSG gerechtfertigt sein. Dies ist jedoch stets anhand einer individuellen Interessenabwägung und durch möglichst frühzeitige Beachtung des Anonymisierungsgebots sicherzustellen.

3.1.3 Anonymisierung und Pseudonymisierung; Verschlüsselung

Carsten Ulbricht

Werden mit Big Data statistische, nicht personenbezogene Erkenntnisse angestrebt, so sind Möglichkeiten einer frühzeitigen Anonymisierung oder Pseudonymisierung zu prüfen, die – wie obenstehend bereit dargestellt – aus der Anwendbarkeit des BDSG herausführen können. Die frühzeitige und richtige Anonymisierung oder Pseudonymsieierung von Datenmengen sind damit sehr relevante Gestaltungsoptionen, um eine Big Data-Analyse zulässig zu gestalten.

Soweit keine personenbezogene Daten im Sinne des § 3 Abs. 1 BDSG, also keine Informationen betroffen sind, die einer bestimmten oder bestimmbaren natürlichen Person zuzuordnen sind, greifen die vorgenannten datenschutzrechtlichen Grundsätze nämlich nicht ein.

Datenschutzrechtliche Restriktionen können folglich durch Anonymisierung (siehe Abschn. 3.1.3.1) oder (zum Teil) durch Pseudonymisierung (Abschn. 3.1.3.2) ausgeschie-

den werden. In diesem Rahmen kann auch die Verschlüsselung (Abschn. 3.1.3.3) eine Rolle spielen.

3.1.3.1 Anonymisierung

Nach dem BDSG sind nur die Daten geschützt, die einen hinreichenden Personenbezug im Sinne des § 3 Abs. 1 BDSG aufweisen – eine natürliche Person also bestimbar ist. Die Bestimmbarkeit einer natürlichen Person ist dabei weit auszulegen. Es darf keinerlei Personenbezug bestehen.

Eine Anonymisierung gemäß § 3 Abs. 6 BDSG erfordert deshalb, dass keinerlei Rückschlüsse auf die bezogene Person mehr möglich sind bzw. solche nur mit einem unverhältnismäßig großen Ressourcenaufwand zu realisieren sind. Ziel der Anonymisierung ist es somit, jede Möglichkeit des Rückschlusses auf die Person zu unterbinden.

Zu den Anforderungen, die das Datenschutzrecht an den Umgang mit personenbezogenen Daten stellt, wird in der überwiegenden Kommentarliteratur vertreten, dass Daten grundsätzlich dann als anonym anzusehen sind, wenn die Herstellung des Personenbezuges ohne unverhältnismäßigen Aufwand nicht mehr möglich ist. Die Möglichkeit einer Zusammenführung von (bislang anonymen) Daten mit einem zur Personenbestimmung geeigneten Zusatzwissen führt auch dann nicht zur Bestimmbarkeit der Personen, wenn rechtliche Regelungen sie wirksam dauerhaft mit so hoher Wahrscheinlichkeit ausschließen, dass das Risiko praktisch vernachlässigt werden kann (vgl. Dammann 2014). Potenzielle Anonymisierungsverfahren müssen den bereits erwähnten Anforderungen genügen. Wird die Bestimmbarkeit einer Person praktisch ausgeschlossen und ist auch die Rückgängigmachung der Anonymisierung entsprechend unwahrscheinlich, so kann nicht (mehr) von personenbezogenen Daten ausgegangen werden. Dementsprechend sind datenschutzrechtliche Vorgaben zu vernachlässigen.

Personenbezogene Daten, bei denen die Rückgängigmachung der Anonymisierung entsprechend der obenstehenden Anforderungen ausgeschlossen ist, sind mithin keine personenbezogenen Daten mehr.

Dementsprechend unterliegen sie auch nicht mehr den Regularien des Datenschutzrechtes, weshalb die verantwortliche Stelle auch bei Big Data Projekten entsprechend ungehindert über sie verfügen darf. Sollte der verantwortlichen Stelle jedoch ohne unverhältnismäßigen Aufwand die Möglichkeit zu Teil werden, die betroffenen Personen bzw. Daten wieder bestimbar zu machen, so sind die Bestimmungen des BDSG vollumfänglich anzuwenden. So ist zur Sicherung der Persönlichkeitsrechte etwa im Bereich der wissenschaftlichen Forschung gem. § 40 Abs. 2 BDSG eine getrennte Speicherung sowie eine Anonymisierung der personenbezogenen Daten grundsätzlich vorgeschrieben, sobald bzw. sofern dies nach dem Forschungszweck möglich ist. Weiterhin ist die Reidentifizierung der Daten beispielsweise nach § 43 Abs. 2 BDSG strafbar (Reidentifizierungsverbot).

Da die Individualisierung eines Datensatzes auf eine Person – je nach Einzelfall – schon bei Vorliegen weniger Merkmale möglich sein kann, ist zu empfehlen, neben den Identifikatoren (z. B. Name, Adresse usw.) auch weitere markante Merkmale zu löschen oder in Aggregatoren aufzulösen, indem Merkmale durch Gruppenbildung generalisiert, einheit-

liche Datensätze durch Variationen diversifiziert und präzise Begriffe durch allgemeine ersetzt werden (Weichert 2013, S. 258). Dabei ist darauf zu achten, dass entsprechende Aggregationen einen Personenbezug auch dann verhindern, wenn die Datenmengen und damit auch die Merkmale pro Person dynamisch wachsen und damit auch die Auswertungsmöglichkeiten entsprechend zunehmen (Roßnagel 2013, S. 566).

Die (frühzeitige) Anonymisierung ist damit ein wichtiges Werkzeug, um Big Data Projekte auch datenschutzkonform umzusetzen. Es ist bei der Modellierung des Datenschutzmodells auf jeder Stufe der Datenverarbeitung zu eruieren, wann im Hinblick auf den Verarbeitungszweck eine Anonymisierung erfolgen kann und sicherzustellen, dass der Personenbezug dann auch im Sinne der obenstehenden Anforderungen gewährleistet beziehungsweise nachhaltig beseitigt wird.

3.1.3.2 Pseudonymisierung

Nach § 3 Abs. 6a BDSG bezeichnet das Pseudonymisieren den Vorgang des Ersetzens des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen, sodass die Bestimmung des Betroffenen ausgeschlossen oder wesentlich erschwert wird.

Pseudonymisiert werden hauptsächlich personenbezogene Daten, sofern über das Pseudonym eine direkte oder indirekte persönliche Zuordnung wieder möglich sein soll. Dieser Vorgang dient ebenso wie die Anonymisierung der Datensparsamkeit (vgl. § 3 a BDSG). Mittels einer Zuordnungsfunktion besteht jedoch, im Gegensatz zur Anonymisierung, weiterhin die Möglichkeit, auch im Nachhinein, die Datensätze zu einer Person zuzuordnen (vgl. Weichert 2008, Rn. 61).

Pseudonymisierung kann vor allem durch Kodierung erreicht werden. Dabei wird eine Zeichenfolge generiert mittels derer eine Zuordnung zu den Daten möglich ist. Gängige Beispiele dafür sind Kundennummern, Mitgliedsnummern oder auch die Sozialversicherungsnummer. Bei der Umsetzung dieses Verfahrens eröffnen sich verschiedene Möglichkeiten. Zum einen können Pseudonyme durch die Erzeugung von Zufallswerten und deren Zuordnung zum Betroffenen bei einer Zuordnungsliste (Referenzliste, Referenztabelle) gebildet werden (Referenzpseudonyme). Zum anderen können Pseudonyme mit Hilfe einer Einwegfunktion, beispielsweise der Generierung eines Hashwertes, aus den personenbezogenen Daten erstellt werden (Einwegpseudonyme). Der klare Vorteil der Erzeugung von Pseudonymen durch ein Hashverfahren ist die Möglichkeit zur dezentralen Generierung. Dadurch müssten keine Zuordnungstabellen aufbewahrt werden.

Neben dem Verfahren der Kodierung ist zur Pseudonymisierung auch die Separierung geeignet. Die Separierung wird durch eine strikte Trennung sensibler, personenbezogener Daten von den Schlüsseldaten erreicht. Die nach § 3a BDSG vorgeschriebene Datensparsamkeit wird in diesen Fällen durch eine Reduzierung des Kreises der Stellen erreicht, die Zugriff auf die Zuordnungsregel haben. Zwischen verschiedenen Stellen besteht die Möglichkeit personenbezogene Daten für unterschiedliche Zwecke zur Verfügung zu stellen, sofern dennoch eine strikte Datentrennung stattfindet. Dies setzt jedoch voraus, dass die Zuordnungsregeln konsequent bei der Stelle verbleiben, die die pseudonymisierten Daten zur Verfügung gestellt hat.

Welches Verfahren bevorzugt anzuwenden ist, muss die verantwortliche Stelle bei genauer Betrachtung des jeweiligen Einzelfalls entscheiden.

Sowohl beim Verfahren der Kodierung als auch beim Verfahren der Separierung kann das zu verwendende Pseudonym durch verschiedene Stellen festgelegt werden. In Frage kommen dabei eine vertrauenswürdige dritte Stelle (sog. Trustcenter) mit Verfügungsmacht über die Zuordnungsfunktion, die verarbeitende Stelle oder die betroffene Person selbst. Leicht identifizierbare Pseudonyme wie beispielsweise eine Zusammensetzung aus Teilen des Namens sowie dem aktuellen Wohnort in Verbindung mit dem Geburtsdatum genügen für eine wirksame Pseudonymisierung in aller Regel nicht (vgl. BeckOK DatenSR/Schild BDSG § 3 Rn. 102–104 und Art. 6 e Satz 1 Europäische Datenschutzrichtlinie). Die Verarbeitung und Nutzung der pseudonymisierten Daten darf anschließend, allgemein formuliert, nur durch Stellen stattfinden, die keinen Zugang auf die Zuordnungsfunktion haben. Andernfalls wäre eine Zuordnung wieder möglich und die Bestimmungen des Datenschutzrechts würden wieder eingreifen.

Situationsabhängig weist die Pseudonymisierung besondere Eigenschaften auf, welche sich gegenüber der Anonymisierung als Vorteil erweisen. So besteht bei Daten unter demselben Pseudonym die Möglichkeit der Verkettung, wodurch Datensammlungen bis hin zu umfassenden Profilen unter einem Pseudonym entstehen können. Die Zuordnungsfunktion ermöglicht außerdem eine gezielte Aufdeckung des Pseudonyms. So kann für bestimmte Zwecke – unter Einhaltung vordefinierter Voraussetzungen – der Personenbezug von der Stelle, die über die Zuordnungsfunktion verfügt, wiederhergestellt und die Person, der das Pseudonym zugeordnet wurde, identifiziert werden. Hat die Stelle, die die Daten verwendet keine Kenntnis von der Zuordnungsfunktion und auch keine Möglichkeit zur Kenntnisnahme, besteht in Bezug auf die Daten kein Unterschied zwischen den personenbezogenen und den anonymisierten Daten (vgl. Arning et al. 2006, S. 702).

Folglich bietet gerade die Pseudonymisierung über die genannten Verfahren unterschiedliche Optionen, Big Data-Projekte unter Wahrung der Datenschutzgrundsätze zu betreiben und gleichzeitig relevante und (mittelbar) auch zuordenbare Erkenntnisse und Ergebnisse zu gewinnen.

3.1.3.3 Verschlüsselung

Seit der Novellierung des BDSG im Jahre 2009 findet sich in dem Gesetz der Grundsatz der Verschlüsselung. Technisch versteht man unter Verschlüsselung den Vorgang, bei dem ein klar lesbarer Text (Klartext) oder auch Informationen anderer Art wie Ton- oder Bildaufzeichnungen mit Hilfe eines Verschlüsselungsverfahrens (Kryptosystem) in eine „unleserliche“, das heißt nicht einfach interpretierbare Zeichenfolge (Geheimtext) umgewandelt wird. Diese Verschlüsselungsvorschrift wird durch eine für jeden Verschlüsselungsvorgang einzeln spezifizierte „Arbeitsanweisung“ (Schlüssel) ergänzt. Wer den Schlüssel kennt, kann den angewendeten Algorithmus rückgängig machen und erhält den Klartext. Die Sicherheit der Verschlüsselung ist also maßgeblich von der Qualität des Algorithmus sowie der Anzahl der Möglichkeiten (Schlüssellänge) abhängig (Ernestus 2014, BDSG, RdNr. 166).

Mit Hilfe von Verschlüsselung (Kryptographie) kann folglich zwar die Rückgängigmachung der Anonymisierung bzw. der Pseudonymisierung durch die zuständigen Stellen die mit den Daten arbeiten sollen nicht verhindert werden, gleichwohl wird aber der Zugriff durch Unbefugte auf die Daten gänzlich verhindert, sofern die Verschlüsselung stark genug ist. Die Kryptographie bezeichnet mithin mathematische Methoden, um Techniken und Algorithmen zu entwickeln, welche die Sicherheit der Daten schützen, indem sie die unbefugte Kenntnisnahme oder absichtliche Manipulationen verhindert (vgl. BSI IT-Grundschutz, M3.23, S. 2330).

Dabei kommt es entscheidend auf die Wirksamkeit der Verschlüsselung an. Ein (zu) einfacher Algorithmus wäre es, jedes Zeichen der personenbezogenen Daten, also des Klartextes, um eine bestimmte Anzahl von Buchstaben im Alphabet zu verschieben. Der Schlüssel beinhaltet in diesem Fall dann die genaue Anzahl und die Richtung, in der die Buchstaben verschoben werden sollen. Da hierbei aber sowohl der Algorithmus als auch der Schlüssel relativ leicht zu erraten sind, werden für die Algorithmen heutzutage aufwendige mathematische Funktionen verwendet. Ohne den richtigen Schlüssel ist die Lösung dann nur durch Ausprobieren aller Möglichkeiten (sog. Brute-Force-Methode) zu finden. Wählt man für den Schlüssel nun eine Kombination aus Groß- Kleinbuchstaben & Zahlen aus, hat man schnell eine Menge verschiedener Möglichkeiten die beim heutigen Stand eines Computers nicht unter 1000 Jahren durch Ausprobieren zu erraten ist.

Während die Daten vor der Verschlüsselung noch personenbezogen waren, kann durch die Verschlüsselung die Bestimmbarkeit für andere als den Schlüsselhabern entfallen.

Der Schlüssel wirkt in diesem Fall wie die Zuordnungsfunktion beim Pseudonymisieren von personenbezogenen Daten. Eine Verschlüsselung kann demnach die personenbezogenen Daten aus Sicht des Schlüsselhabers pseudonymisieren und aus Sicht aller Schlüsselnichthabers sogar anonymisieren. Im BDSG wurde das Verschlüsseln als besondere Maßnahme bisher jedoch noch nicht aufgenommen.

Problematisch aus rechtlicher Sicht ist bei der Verschlüsselung, dass der Personenbezug von entsprechenden Daten nicht wirklich beseitigt wird. Dennoch kann die Verschlüsselung bei der automatisierten Verarbeitung und Nutzung personenbezogener Daten für eine Zugangs-, Zugriffs- und Weitergabekontrolle sorgen (vgl. Ernestus 2014, RdNr. 164–165).

Demgemäß ist auch die Verschlüsselung ein wichtiges Werkzeug, welches im Rahmen von Big Data-Projekten im jeweils geeigneten Stadium der Datenverarbeitung eingesetzt werden kann, um die Datenschutzanforderungen zu erfüllen.

3.1.4 Technologien zur Umsetzung datenschutzrechtlicher Anforderungen

Carsten Ulbricht

Die Einhaltung der vorgenannten datenschutzrechtlichen Rahmenbedingungen kann auch auf technologischer Ebene unterstützt werden. Unter Privacy Enhancing Technologies versteht man Ansätze, bei denen mit entsprechenden Technologien die Datenschutzanforderungen bereits unmittelbar in die Datenanalyse integriert werden.

Beim sogenannten Anonymize-and-Mine Ansatz werden unmittelbar nur anonyme (oder pseudonyme) Daten erhoben. Hierfür müssen bereits bei der Erhebung die Daten weggelassen werden, die den oben dargelegten Anonymitätsanforderungen entgegenstehen. Mit der unmittelbaren Anonymisierung, die ja durch vorherige Definition der auszuscheidenden Kriterien und Informationen gewährleistet werden muss, geht allerdings der Nachteil einher, dass schon bei der Erhebung Daten „verloren gehen“ können, die möglicherweise bei der Datenanalyse ansonsten wichtige beziehungsweise die entscheidenden Erkenntnisse bringen würden.

Vorzugswürdig scheint – je nach Big Data-Projekt – insoweit der Mine-and-Anonymize Ansatz, der eine Erhebung personenbezogener Daten vornimmt, diese Daten aber dann auf Grundlage spezifischer Datenschutzkriterien anonymisiert. Dieser Ansatz erfordert allerdings eine individuelle Ausgestaltung jedes Projektes, da jeweils entschieden werden muss, welche erhobenen Informationen wie auszuscheiden sind, damit eine hinreichende Anonymität gewährleistet ist.

Neben den vorgenannten exemplarisch erläuterten Privacy Preserving Technologies sind diverse andere technologische Ansätze in der Diskussion (z. B. Secure Distributed Computing), die zum effektiven Einsatz im Big Data-Projekt frühzeitig geprüft und bei entsprechender Geeignetheit einbezogen werden können.

3.1.5 Zulässigkeit einzelner Phasen von Big Data-Analysen

Carsten Ulbricht

Aus den vorgehenden Ausführungen folgt, dass Big Data-Analysen nur dann datenschutzkonform sind, wenn die Vereinbarkeit mit den gesetzlichen Vorgaben *auf jeder Stufe* der Datenverarbeitung (Erhebung, Speicherung etc.) gewährleistet ist. Die Datenverarbeitung ist also (z. B. unter Einsatz der vorgenannten Privacy Preserving Technologies) so zu modellieren, dass entweder anonyme, anonymisierte oder pseudonymisierte Daten erhoben bzw. (weiter-)verarbeitet werden oder sich eben für jeden Verarbeitungsschritt ein gesetzlicher Erlaubnistatbestand begründen lässt.

Zur Veranschaulichung sollen nachfolgend anhand der jeweils genannten Datenverarbeitungsschritte geeignete Maßnahmen skizziert werden.

3.1.5.1 Erhebung von Big Data

Zur Gewährleistung der Datenschutzanforderungen sollte bereits die Datenerhebung – nach Möglichkeit – auf das (zwingend) erforderliche Maß beschränkt werden.

Wenn die Zielsetzung der Big Data-Analyse dem nicht entgegensteht, sollten nur anonyme Daten erhoben oder die Daten unmittelbar nach der Erhebung anonymisiert oder pseudonymisiert werden.

Zur Gewährleistung der Zweckbindung und der Nichtverkettbarkeit könnten die Daten schon bei der Erhebung mit dem jeweiligen Zweck gekennzeichnet und die Datenbestände im Hinblick auf eine spätere getrennte Speicherung auch unabhängig voneinander erhoben werden.

Nach Möglichkeit sollte bereits bei oder vor der Erhebung über die Erfüllung der notwendigen Informationspflichten die entsprechende Transparenz hergestellt werden.

3.1.5.2 Speichern von Big Data

Nach der Sammlung muss die verantwortliche Stelle für die weiteren Verarbeitungsschritte eine Befugnis zum Speichern beziehungsweise nachfolgend zum Weiterverarbeiten haben.

Wenn die gespeicherten Rohdaten den Umfang von Big Data erreicht haben, ist in der Regel für sehr viele Daten ein Personenbezug anzunehmen, weil in den großen Datenmengen eine erhebliche Menge an Merkmalen für Analysen verfügbar ist, sodass diese Merkmale zur Identifizierung einer Person zusammengeführt werden könnten. Im Zusammenhang mit der Speicherung sind demnach dezentrale Datenhaltung, Anonymisierungs- oder Pseudonymisierungsmaßnahmen oder anderen Privacy Enhancing Technologies und auch eine frühestmögliche Löschung nicht (mehr) benötigter Daten zu erwägen.

Das Speichern der Daten für einen neuen Zweck könnte entweder dadurch legitimiert werden, dass berechtigte Interessen der verantwortlichen Stelle nach § 28 Abs. 2 Satz 1 Nr. 2 BDSG vorgebracht werden können oder die Daten aus allgemein zugänglichen Quellen stammen (z. B. § 28 Abs. 1 Satz 1 Nr. 3 BDSG), da entsprechende Daten grundsätzlich keiner Zweckbindung unterliegen.

In beiden Fällen ist das Speichern aber nur dann zulässig, wenn es zusätzlich auf einen bestimmten neuen Zweck begrenzt ist und die schutzwürdigen Interessen des Betroffenen nicht überwiegen.

3.1.5.3 Personenbezogene Auswertung von Big Data

Diverse Big Data-Anwendungen zielen darauf ab, für bestimmte Personen eine bestimmte Fragestellung zu beantworten, oder wie beim Profiling, diverse Eigenschaften und Informationen zu einem (nutzbringenden) Profil zusammenzufügen. Typische Beispiele sind das Verfolgen („Tracking“) des Nutzungsverhaltens im Internet oder das Scoring (siehe Abschn. 3.1.5.6).

Da diese Auswertungen oft die Zielsetzung verfolgen, das Verhalten der Person vorherzusagen und beeinflussen zu können und damit in einem besonderen Spannungsverhältnis zum Recht auf informationelle Selbstbestimmung stehen, können entsprechend personen-

bezogene Analysen nur auf Grundlage einiger weniger Legitimationsbestände rechtskonform begründet werden. Die konkrete Ausgestaltung und die Zulässigkeit hängt dabei stets von den Umständen des Einzelfalls ab.

Im Bereich des Profiling wird man sich an § 15 Abs. 3 TMG oder § 28 Abs. 3 BDSG orientieren können und müssen. Für ein etwaiges Scoring sollte sich die verantwortliche Stelle sich streng an die Vorgaben des § 28b BDSG halten.

Die Rechtfertigung einer personenbezogenen Auswertung im Zusammenhang mit Big Data auf Grundlage einer Interessenabwägung aus § 28 Abs. 1 Satz 1 Nr. 2 BDSG dürfte sich in vielen Fällen als schwierig herausstellen, da schutzwürdige Interessen etwa dann überwiegen, wenn – wie in diversen Big Data-Anwendungen – sich die Auswertung für den Betroffenen als nicht vorhersehbar darstellt und/oder zu tief in dessen Persönlichkeitsrechte eindringt (vgl. Roßnagel 2013, S. 564).

Abschließend gilt es bei entsprechenden Big Data-Auswertungen stets darauf zu achten, dass automatische Entscheidungen zum Nachteil des Betroffenen gemäß § 6a BDSG unzulässig sind. Danach dürfen Entscheidungen, die für den Betroffenen eine rechtliche Folge nach sich ziehen oder ihn erheblich beeinträchtigen, nicht ausschließlich auf eine automatisierte Verarbeitung personenbezogener Daten gestützt werden, die der Bewertung einzelner Persönlichkeitsmerkmale dienen. Eine ausschließlich auf eine automatisierte Verarbeitung gestützte Entscheidung liegt insbesondere dann vor, wenn keine inhaltliche Bewertung und darauf gestützte Entscheidung durch eine natürliche Person stattgefunden hat.

3.1.5.4 Auswertung von Big Data

Deutlich unproblematischer stellt sich regelmäßig die abstrakte Auswertung von Daten dar, wenn es darum geht, Entwicklungen, Strukturen und Muster zu erkennen.

Beziehen sich die Ergebnisse auf ausreichend große Gruppen von Betroffenen, sodass einzelne Betroffene nicht erkannt werden können, ist diese Form der Auswertung mangels Personenbezug datenschutzrechtlich zulässig (Roßnagel 2013, S. 565).

Dabei ist durch rechtliche, technische oder organisatorische Maßnahmen sicherzustellen, dass nicht über eine statistische Auswertung doch noch ein Personenbezug hergestellt werden kann.

3.1.5.5 Veröffentlichen von Big Data

Die Veröffentlichung personenbezogener Daten die personenbezogene Daten enthalten, ist daher grundsätzlich unzulässig.

Anders hingegen ist insofern die Veröffentlichung aggregierter oder statistischer Daten, die keinerlei Personenbezug mehr aufweisen beziehungsweise die anonymer, in sehr eng begrenzten Ausnahmefällen auch pseudonymer Daten zu beurteilen. Hier ist in vielen Fällen auch eine Veröffentlichung der entsprechenden Ergebnisse und (abstrakten) Ableitungen denkbar.

3.1.5.6 Zusammenfassung

Die vorgehenden Ausführungen zeigen, dass diverse Gestaltungsmöglichkeiten bestehen, um die Erhebung, Speicherung und Auswertung von Big Data mit den datenschutzrechtlichen Vorgaben in Einklang zu bringen. Je nach Einzelfall oder Werkzeug kann bei Big Data-Projekten über eine entsprechend granulare Ausgestaltung die Erhebung personenbezogener Daten vermieden oder über Anonymisierungs- oder Pseudonymisierungsmechanismen teilweise im Zusammenwirken mit weiteren Maßnahmen auf eine rechtssichere Grundlage gestellt werden.

In Ansehung der wachsenden Bedeutung des Datenschutzes in Zeiten von Internet, Mobiler und Sozialer Medien und der steigenden Sensibilität bei Menschen sind die verantwortlichen Stellen gut beraten, die Datenschutzrechte zu achten und klar zu kommunizieren, welche Daten zu welchen Zwecken wie eingesetzt und verarbeitet werden.

Je nach Rechtslage sollte bei den Betroffenen nach einer hinreichenden Aufklärung die Einwilligung (Opt-In) zur jeweiligen Datenerhebung bzw. die Möglichkeit zur Ablehnung (Opt-Out) gegeben werden. Transparenz und Kontrolle der Betroffenen sind elementare Voraussetzungen um Big Data nicht zu einer unkontrollierbaren Gefahr für die Menschen werden zu lassen, was schlussendlich zu einer Überregulierung führen kann.

Gleichzeitig scheint es sinnvoll, dass der Gesetzgeber die aktuelle datenschutzrechtliche Systematik auch im Hinblick auf Big Data zu überdenkt und unter Berücksichtigung moderner Entwicklungen überarbeitet. Der allgemeingültige Grundsatz der Datensparsamkeit passt kaum noch in eine Welt allgegenwärtiger Datenverarbeitung und ständige wachsender Datenmengen, ebenso wie der aktuell sehr undifferenzierte Ansatz personenbezogener Daten. Vorzugswürdig scheint eine weitergehende Differenzierung nach der „Sensibilität“ von Informationen. Davon sollte abhängig gemacht werden, ob diese Daten ohne weitergehende Voraussetzungen, auf Grundlage eines Opt-Out bzw. erst nach einer ausdrücklichen Einwilligung verarbeitet werden dürfen.

Das Festhalten an einer umfassenden Aufklärung der Nutzer über die Datenverwendung und -weitergabe bei Erhebung personenbezogener Daten und entsprechend personenbezogener Auswertung ist – bis auf nachvollziehbare Ausnahmen im Bereich öffentlich zugänglicher Daten – im Hinblick auf die dringend notwendige Transparenz und das Erfordernis einer informierten Entscheidung der Betroffenen hingegen von besonderer Bedeutung.

3.1.6 Betroffenenrechte

Jenny Hubertus

Zum Schutz der informationellen Selbstbestimmung gewährt § 6 BDSG dem Betroffenen Ansprüche auf Auskunft, Berichtigung, Löschung und Sperrung. Hierbei handelt es sich um unabdingbare Rechte, die der Betroffene auch nicht durch eine Vereinbarung mit der verantwortlichen Stelle oder einem Dritten ausschließen oder einschränken kann (§ 6 Abs. 1 BDSG).

Im Bereich der Datenverarbeitung durch öffentliche Stellen (insbesondere durch Behörden; vgl. § 2 Abs. 1, 2 BDSG) findet sich der Auskunftsanspruch des Betroffenen in § 19 BDSG, flankiert durch die Ansprüche auf Berichtigung, Löschung und Sperrung von Daten in § 20 BDSG.

Gegenüber nicht-öffentlichen Stellen steht dem Betroffenen ebenfalls ein Auskunftsanspruch zu, der sich nach § 34 BDSG richtet und der auch hier über § 35 BDSG hinsichtlich der Rechte des Betroffenen auf Berichtigung, Löschung und Sperrung seiner Daten ergänzt wird.

Im Nachfolgenden sollen ausschließlich die Rechte des Betroffenen gegenüber nicht-öffentlichen Stellen, also insbesondere gegenüber natürlichen Personen und Unternehmen (vgl. § 2 Abs. 4 BDSG) erläutert werden. Die Darstellungen sind auf Ansprüche gegenüber Behörden in weiten Teilen übertragbar.

Die Geltendmachung von Auskunfts- und Korrekturansprüchen gegenüber der verantwortlichen Stelle setzt jedoch voraus, dass der Betroffene zumindest eine grobe Kenntnis von der Erhebung, Verarbeitung und Nutzung seiner Daten hat. Um diese Transparenz zu gewährleisten, stellt das Bundesdatenschutzgesetz den Ansprüchen des Betroffenen auf Auskunft und Korrektur einen Benachrichtigungsanspruch voran:

3.1.6.1 Benachrichtigung des Betroffenen

Werden personenbezogene Daten beim Betroffenen erhoben, so ist dieser nach § 4 Abs. 3 BDSG von der verantwortlichen Stelle über die Identität dieser verantwortlichen Stelle, die Zweckbestimmungen der Erhebung, Verarbeitung und Nutzung seiner personenbezogenen Daten und die möglichen Empfänger zu unterrichten. Werden personenbezogenen Daten nicht unmittelbar beim Betroffenen erhoben, ergibt sich die gleiche Verpflichtung aus § 33 BDSG.

Das Gesetz lässt in § 33 Abs. 2 BDSG zahlreiche Ausnahmen zu, bei deren Eingreifen eine Benachrichtigung des Betroffenen nicht erforderlich ist. Neben der Geheimhaltungsbedürftigkeit (§ 33 Abs. 2 Nr. 3 BDSG), der Gefährdung der öffentlichen Sicherheit oder Ordnung (§ 33 Abs. 2 Nr. 6 BDSG), oder des Vorliegens eines Erlaubnistarbestandes (§ 33 Abs. 3 Nr. 4 BDSG), ist eine Benachrichtigung insbesondere dann entbehrlich, wenn der Betroffene bereits auf andere Art und Weise Kenntnis erlangt hat (§ 33 Abs. 2 Nr. 1 BDSG).

Eine besondere Form der Benachrichtigung schreibt das Gesetz nicht vor, allerdings empfiehlt sich alleine schon aus Gründen der Beweissicherung eine schriftliche Benachrichtigung des Betroffenen.

Die erforderliche individuelle Information des Betroffenen kann auch nicht durch einen Hinweis in Allgemeinen Geschäftsbedingungen ersetzt werden, obgleich ein solcher bewirken kann, dass der Betroffene auf andere Art und Weise von der Speicherung Kenntnis erlangt. Hierzu bedarf es jedoch einer deutlichen Gestaltung des Hinweises in den Allgemeinen Geschäftsbedingungen, da erfahrungsgemäß gerade diese Regelungen von den Betroffenen vielfach nicht, oder nicht mit der nötigen Sorgfalt zur Kenntnis genommen werden (Gola und Schomerus 2015, § 33 Rn. 18).

Auch genügt es nicht, dass die verantwortliche Stelle den Betroffenen durch irgendeine Handlung, beispielsweise einen in eine Werbesendung integrierten Hinweis, wissen lässt, dass sie seine personenbezogenen Daten gespeichert hat. Die Benachrichtigung hat vielmehr ausdrücklich zu erfolgen.

Ein Verstoß gegen die Benachrichtigungspflicht ist nach § 43 Abs. 1 Nr. 8 BDSG eine Ordnungswidrigkeit, die mit einer Geldbuße von bis zu fünfzigtausend Euro geahndet werden kann. Datenverarbeitende Unternehmen tun daher gut daran, die Benachrichtigungspflicht des Betroffenen nicht vorschnell abzulehnen.

Zur Unzulässigkeit der Datenerhebung, -verarbeitung und -nutzung – und in Folge dessen zu einem Löschungsanspruch des Betroffenen – führt die unterlassene Benachrichtigung jedoch nicht.

3.1.6.2 Benachrichtigungspflicht bei Web-Crawling und Screen-Scraping?

Neben den bereits oben genannten Ausnahmen, lässt § 33 Abs. 2 Nr. 7 a) BDSG aber auch eine Ausnahme von der Benachrichtigungspflicht für den Fall zu, dass die Daten für eigene Zwecke aus allgemein zugänglichen Quellen entnommen werden und eine Benachrichtigung wegen der Vielzahl der Betroffenen unverhältnismäßig erscheint.

Der Begriff der allgemein zugänglichen Quelle entspricht dem des § 28 Abs. 1 Nr. 3 BDSG. Hierunter sind Daten aus öffentlichen Registern, Büchern, Zeitschriften und sonstigen Medien, aber auch teils Daten aus Sozialen Netzwerken zu verstehen, sofern diese wiederum öffentlich und nicht nur einem ausgewählten Personenkreis (Mitgliedern, Friends/Freunden, Followern) zugänglich gemacht werden.

Hintergrund dieser Ausnahme ist, dass diese Daten nach § 28 Abs. 1 Nr. 3 BDSG verwendet werden dürfen, wenn das schutzwürdige Interesse des Betroffenen an dem Ausschluss der Verarbeitung oder Nutzung das berechtigte Interesse der verantwortlichen Stelle nicht offensichtlich überwiegt. Bedenkt man bei dieser Abwägung aber, dass der Betroffene gerade bei in Sozialen Netzwerken preisgegebenen Daten selbst (z. B. über die von ihm gewählten Privatsphäreinstellungen) über deren Verbreitung bestimmen kann, wird sein schutzwürdiges Interesse einer Datenerhebung, -verarbeitung und -nutzung nur in den seltensten Fällen entgegenstehen.

Dürfen die Daten hiernach aber ohnehin verwendet werden, ist auch eine Benachrichtigung nicht erforderlich, insbesondere, wenn eine Benachrichtigung auf Grund der Vielzahl der Fälle, etwa auch aus finanzieller Sicht, für die verantwortliche Stelle unverhältnismäßig wäre (Forgó 2015; Wolff und Brink 2014, § 33 BDSG Rn. 66).

Relevant wird diese Ausnahme von der Benachrichtigungspflicht insbesondere bei den heute weit verbreiteten technischen Möglichkeiten, über automatische Suchroboter, z. B. mittels Web-Crawling- oder Screen-Scraping-Tools, zahlreiche Daten aus öffentlich zugänglichen Internetquellen zu durchsuchen, zu speichern und so in den Datenbestand des eigenen Unternehmens zu überführen, wo die Daten dann für eigene Zwecke systematisiert und ggf. Dritten in neu zusammengestellter Form zur Verfügung gestellt werden können.

Auch wenn dieses Vorgehen, insbesondere in urheberrechtlicher Sicht, nicht unumstritten ist (hierzu nachfolgend ausführlicher unter Abschn. 3.2), löst es eine Benachrichtigungspflicht mit den zuvor dargestellten Sanktionen zumindest dann nicht aus, wenn sich das Abgreifen der Daten auf allgemein zugängliche Quellen beschränkt.

3.1.6.3 Auskunftsanspruch des Betroffenen

§ 34 BDSG gewährt dem Betroffenen das Recht, Auskunft über die zu seiner Person gespeicherten Daten zu verlangen. Zudem sind die Herkunft dieser Daten, mögliche Empfänger und der Zweck der Speicherung offenzulegen.

Von diesem Auskunftsverlangen wird heute zunehmend Gebraucht gemacht. Hierbei ist jedoch zu berücksichtigen, dass der Betroffene einen solchen Auskunftsanspruch nicht ohne jegliche Verdachtsmomente, gleichsam „ins Blaue hinein“ geltend machen kann. Der Auskunftsersuchende muss zumindest ausreichend darlegen, dass die zur Auskunft verpflichtete Stelle personenbezogene Daten über ihn gespeichert haben könnte.

Ist diese Voraussetzung erfüllt und weigert sich die verantwortliche Stelle dennoch, die entsprechende Auskunft zu erteilen, so stellt auch dies nach § 43 Abs. 1 Nr. 8 a–c BDSG eine bußgeldbewehrte Ordnungswidrigkeit dar. Zudem hat der Betroffene jederzeit die Möglichkeit, seinen Auskunftsanspruch gerichtlich durchzusetzen. Da der Auskunftsanspruch an keine weiteren Voraussetzungen geknüpft ist, wird einer solchen Klage in der Regel stattgegeben und die verantwortliche Stelle zur Erteilung der verlangten Auskunft verpflichtet.

Derartige Auskunftsverlangen sollten daher – auch wenn sie lästig sein mögen – ernst genommen werden. Ein (ordentlich geführtes) Verfahrensverzeichnis kann helfen, die erforderliche Auskunft schnell und effektiv zu erteilen.

Erteilt die verantwortliche Stelle Auskunft, so hat dies gemäß § 34 Abs. 6 BDSG ausdrücklich in Schrift- bzw. Textform zu erfolgen. Kosten darf die zur Auskunft verpflichtete Stelle hierfür nicht erheben.

Der Auskunftsanspruch nach § 34 BDSG stellt ein höchstpersönliches Recht des Betroffenen dar. Er ist daher nicht auf andere übertragbar.

Indes ist die verantwortliche Stelle nicht in allen Fällen gezwungen, Auskunft zu erteilen. So hat sie z. B. das Recht, die Auskunft über die Herkunft und die Empfänger der Daten zu verweigern, sofern das Interesse an der Wahrung eines Geschäftsgeheimnisses gegenüber dem Informationsinteresse des Betroffenen überwiegt.

Ist der Betroffene bereits nach § 33 BDSG nicht über die Datenerhebung zu benachrichtigen (z. B. wegen Gefährdung der öffentlichen Sicherheit oder Ordnung), entfällt folgerichtig auch der Auskunftsanspruch (§ 34 Abs. 7 BDSG).

Gleches gilt in der oben angesprochenen Konstellation, in der Daten mittels Web-Crawling- oder Screen-Scraping-Methoden aus dem Internet entnommen werden. Soweit es sich um frei zugängliche Daten handelt, besteht nach § 33 Abs. 2 Nr. 7 a) BDSG keine Benachrichtigungspflicht, und in Folge dessen nach § 34 Abs. 7 i. V. m. § 33 Abs. 2 Nr. 7 a) BDSG auch keine Auskunftspflicht gegenüber dem Betroffenen.

Da Auskunftsersuchen jedoch auch in diesen Fällen nicht ausgeschlossen sind, stellt sich für die datenverarbeitende Stelle die Frage nach einem zweckmäßigen Umgang mit dennoch eingehenden Anfragen. Hierbei ist zu berücksichtigen, dass der schlichte formaljuristische Verweis auf § 34 Abs. 7 BDSG den Betroffenen nicht zufriedenstellen wird. Um eine weitergehende Auseinandersetzung zu vermeiden, steht es der verantwortlichen Stelle daher frei, gleichwohl Auskunft zu erteilen, auch wenn eine gesetzliche Auskunftspflicht nicht besteht.

Insbesondere in den Fällen, in denen eine Benachrichtigung zwar unterbleiben kann, weil sie einen unverhältnismäßigen Aufwand erfordern würde (so auch bei § 33 Abs. 2 Nr. 7a), das Auskunftsersuchen aber erfahrungsgemäß die seltene Einzelfallausnahme bildet und folglich eine Auskunft im konkreten Einzelfall unschwer gegeben werden kann, scheint eine Auskunftserteilung auch geboten (Gola und Schomerus 2015, § 34 Rn. 18; Dix 2014, § 34 Rn. 57).

3.1.6.4 Korrekturrechte

Neben dem zuvor behandelten Auskunftsrecht stehen dem Betroffenen aber auch die in § 35 BDSG aufgelisteten Korrekturrechte zu.

Berichtigung

Unrichtig gespeicherte personenbezogene Daten sind zu berichtigen.

Woraus sich die Unrichtigkeit der Daten ergibt, ist für den Berichtigungsanspruch gleichgültig. Neben den schon unrichtig erhobenen Daten werden daher auch solche Daten erfasst, die zunächst „richtig“ gespeichert wurden, durch eine nachträgliche Veränderung des zugrunde liegenden Sachverhaltes jedoch unrichtig geworden sind. Selbst kleinste Unrichtigkeiten sind zu berichtigen; eine Erheblichkeitsschwelle kennt das Gesetz nicht. Schließlich kann die Unrichtigkeit auch daraus resultieren, dass eine Information aus dem Kontext herausgerissen wird und ohne den entsprechenden Sinnzusammenhang einen falschen Rückschluss zulässt.

Lösung

Darüber hinaus sind personenbezogene Daten in den Fällen des § 35 Abs. 2 S. 2 Nr. 1–4 BDSG zu löschen.

Hiernach muss eine Lösung zum einen dann erfolgen, wenn die Speicherung unzulässig ist, beispielsweise weil die Daten selbst unrichtig sind. Zwar besteht hier vorrangig die oben skizzierte Pflicht zur Berichtigung, ist eine Korrektur aber nicht möglich – etwa weil der verantwortlichen Stelle die Kenntnis über den tatsächlich zutreffenden Sachverhalt fehlt – tritt an die Stelle der Berichtigung die Lösung. Zudem hat die verantwortliche Stelle hinsichtlich bestimmter, das Persönlichkeitsrecht des Betroffenen besonders tangierender Daten die Beweispflicht für deren Richtigkeit. Kann die verantwortliche Stelle im Bestreitensfall die Richtigkeit der Daten nicht beweisen, sind diese Daten ebenfalls zu löschen. Für eigene Zwecke gespeicherte Daten sind zudem immer dann zu löschen, wenn ihre Kenntnis für den Zweck der Speicherung nicht mehr erforderlich ist. Dies

kann insbesondere dann der Fall sein, wenn eine die weitere Speicherung legitimierende Zweckbestimmung nicht mehr vorliegt.

Soweit schließlich die Daten geschäftsmäßig zum Zwecke der Übermittlung verarbeitet werden, werden der verantwortlichen Stelle regelmäßige Prüfungspflichten auferlegt. Ergeht eine Prüfung, dass eine fortdauernde Speicherung der Daten nicht länger erforderlich ist, so sind auch diese ab diesem Zeitpunkt zu löschen.

Sperrung

Tritt die Löschung bestimmter Daten in Konflikt mit gesetzlichen, satzungsmäßigen oder vertraglichen Aufbewahrungsfristen und kommt eine Löschung der Daten hiernach nicht in Betracht, dürfen die Daten zwar weiter erhalten bleiben, sie sind jedoch zu sperren.

Auch wenn Grund zu der Annahme besteht, dass durch eine Löschung schutzwürdige Interessen des Betroffenen einträchtig würden, tritt an die Stelle der Löschung die Sperrung.

Hierbei gilt der Grundsatz, dass unzulässigerweise erhobene oder unrichtige Daten stets gelöscht werden dürfen, da hierdurch gerade nicht das schutzwürdige Interesse des Betroffenen beeinträchtigt, sondern geschützt wird.

Ist schließlich eine Löschung wegen der besonderen Art der Speicherung der Daten nicht oder nur mit unverhältnismäßig hohem Aufwand möglich, so darf die datenverarbeitende Stelle ebenfalls anstatt der Löschung zur Sperrung der Daten übergehen.

Personenbezogene Daten sind ferner immer dann zu sperren, wenn ihre Richtigkeit vom Betroffenen zwar bestritten worden ist, sich aber weder ihre Richtigkeit noch ihre Unrichtigkeit hat feststellen lassen.

3.1.6.5 Das „Recht auf vergessen werden“

Die zuvor dargestellten Korrekturrechte gelten auch für in das Internet eingestellte Daten unmittelbar. In der Praxis stellen sie jedoch recht stumpfe Schwerter im Kampf gegen die explodierenden Datenmengen dar.

Gerade Suchmaschinen standen hier vielfach in der Kritik, stellte sich doch bei der bloßen Wiedergabe fremder Inhalte in der eigenen Trefferliste zunehmend die Frage, wer in einem solchen Fall überhaupt als verantwortliche Stelle anzusehen ist, gegenüber der Korrekturansprüche geltend zu machen sind. Unklar war auch, wie eine solche Korrektur rein tatsächlich durchgesetzt werden sollte. Hierzu hat sich nun zwischenzeitlich der EuGH positioniert:

In der Rechtssache C-131/12 stellte der EuGH mit Urteil vom 13.05.2014 fest, dass auch der Betreiber einer Internet-Suchmaschine eine Datenerhebung vornimmt, indem er automatisch, kontinuierlich und systematisch im Internet veröffentlichte Informationen aufspürt. Diese Daten werden sodann durch das Indexierprogramm des Suchmaschinenbetreibers ausgelesen und gespeichert und in Form von Ergebnislisten an andere Nutzer weitergegeben. Der Suchmaschinenbetreiber ist daher als verantwortliche Stelle anzusehen, die in ihrem Verantwortungsbereich und im Rahmen ihrer Befugnisse und Möglichkeiten dafür zu sorgen hat, dass ihre Tätigkeit den Anforderungen des Datenschutzes entspricht.

Der EuGH sah den Suchmaschinenbetreiber daher in der Pflicht, von der Ergebnisliste, die im Anschluss an eine Personensuche angezeigt wird, die Links zu von Dritten veröffentlichten Internetseiten mit Informationen über diese Person zu entfernen. Eine solche Verpflichtung könne laut EuGH selbst dann bestehen, wenn der betreffende Name oder die betreffenden Informationen auf dieser Internetseite nicht vorher oder gleichzeitig gelöscht werden, ggf. auch dann, wenn ihre Veröffentlichung dort als solche rechtmäßig ist. Auch eine ursprünglich rechtmäßige Veröffentlichung sachlich richtiger Daten kann im Laufe der Zeit nicht mehr den Bestimmungen des Datenschutzes entsprechen, wenn die Daten in Anbetracht aller Umstände des Einzelfalls, insbesondere der verstrichenen Zeit, für die Zwecke, für die sie ursprünglich verarbeitet worden sind, nicht mehr erheblich sind (Gerichtshof der Europäischen Union, Urteil v. 13.05.2014 – C-131/12).

Unter gewissen Voraussetzungen billigt der Europäische Gerichtshof daher den Betroffenen ein „Recht auf vergessen werden“ im Internet zu. Betroffene können sich hierzu unmittelbar an den Betreiber einer Internet-Suchmaschine wenden, der wiederum eine sorgfältige Begründetheitsprüfung vorzunehmen hat. So können z. B. besondere Gründe, wie die Rolle der betreffenden Person im öffentlichen Leben, ein überwiegendes Informationsinteresse der Öffentlichkeit begründen und damit eine Beibehaltung der Information in den Ergebnislisten durchaus rechtfertigen. Folgt die verantwortliche Stelle dem Antrag eines Betroffenen nicht, so kann dieser weiterhin eine Kontrollstelle oder das zuständige Gericht anzu rufen.

Der EuGH stützte sich bei dieser Entscheidung auf die EU-Datenschutzrichtlinie (Richtlinie 95/46/EG), auf der auch das deutsche Bundesdatenschutzgesetz basiert. Damit ist die Entscheidung des Europäischen Gerichtshofes zum „Recht auf vergessen werden“ auch in Deutschland unmittelbar zu berücksichtigen.

Die Rechtsprechung des EuGH zum „Recht auf vergessen werden“ zeigt, dass es den derzeitigen rechtlichen Strukturen an einer praktikablen Handhabe fehlt, den stetig wachsenden Datensammlungen zu begegnen. Zugleich offenbaren sich hier aber auch erste Tendenzen zur Einschränkung von Big Data, die in Zukunft sicherlich noch zunehmen werden, um den Schutz des Betroffenen tatsächlich effektiv gewährleisten zu können.

3.1.7 Internationale Datenverarbeitung

Olaf Botzem

Grenzüberschreitende Datenverarbeitung ist mit besonderen datenschutzrechtlichen Problemen behaftet. Grund hierfür ist die unterschiedliche Ausgestaltung der nationalen Datenschutzregime und damit verbunden auch der Schutzniveaus.

3.1.7.1 Anwendbares Recht

Datenschutzrecht ist zwingendes Recht im Sinne von Art. 9 der Verordnung 593/2008/EG des Europäischen Parlaments und des Rates vom 17. Juni 2008 über das auf vertragli-

che Schuldverhältnisse anzuwendende Recht (Rom I) und kann daher im Rahmen einer Rechtswahlklausel nicht vertraglich abbedungen werden.

Nach § 1 Abs. 5 BDSG sind die maßgeblichen Kriterien für die Frage des anwendbaren Rechts bei grenzüberschreitender Datenverarbeitung der Sitz der verantwortlichen Stelle sowie des Ortes der Datenerhebung bzw. -verarbeitung. Diese Regelung basiert auf Art. 4 Abs. 1 der Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr („EU-DSRL“). Wie sich aus der Formulierung von § 1 Abs. 5 BDSG ergibt, ist zwischen grenzüberschreitender Datenverarbeitung innerhalb der EU und solcher in sogenannten Drittstaaten, die nicht Mitgliedstaaten der EU sind, zu unterscheiden.

Datenverarbeitung in der EU

Innerhalb der EU gilt nach § 1 Abs. 5 S. 1 BDSD (vgl. Art. 4 EU-DSRL) das sogenannte Sitzlandprinzip. Danach ist grundsätzlich das nationale Recht des Staates anzuwenden, in dem die verantwortliche Stelle ihren Sitz hat. Der Ort, an dem die Datenverarbeitung stattfindet, ist in diesem Fall nicht maßgeblich. Für die Anwendbarkeit des Sitzlandprinzips ist es ohne Bedeutung, wie das Schutzniveau in den an der Datenverarbeitung beteiligten Staaten ausgestaltet ist.

Das Sitzlandprinzip bietet für Unternehmen den großen praktischen Vorteil, dass sie sich grundsätzlich nicht mit den verschiedenen nationalen Ausgestaltungen des Datenschutzrechts auseinandersetzen und diese nationalen Besonderheiten im Rahmen ihrer Datenschutz-Compliance nicht beachten müssen.

Unternehmen, die in verschiedenen Mitgliedsstaaten Niederlassungen haben, müssen allerdings eine grundsätzliche Einschränkung des Sitzlandprinzips beachten. Wenn die Datenverarbeitung von einer solchen Niederlassung außerhalb des Sitzlandes ausgeführt wird und diese Niederlassung nicht bloß Auftragnehmer (vgl. § 11 BDSG) ist, finden die nationalen Datenschutzvorschriften des Ortes dieser Niederlassung Anwendung.

Datenverarbeitung in Drittstaaten

Im Unterschied zur Datenverarbeitung innerhalb der EU wird bei der Datenverarbeitung in Drittstaaten auf das sogenannte Territorialprinzip abgestellt. Danach ist der Ort der Datenverarbeitung entscheidend. Auch nach dem Territorialprinzip gilt, dass das Schutzniveau der beteiligten Staaten nicht in die Bewertung einfließt.

Als maßgebliches Indiz gilt eine Nutzung im Inland belegener, automatisierter oder sonstiger dateimäßiger Mittel (vgl. Art. 4 Abs. 1 lit. c EU-DSRL). In den für die Praxis wohl relevantesten Fällen, der Datenverarbeitung über das Internet, bedeutet dies, dass auf den Ort des Hostings der Webseite, also des Serverstandorts, abzustellen ist. Wer in einem Staat lediglich Telekommunikations- oder sonstige Übertragungseinrichtungen für sein Angebot nutzt, fällt nicht in den Anwendungsbereich des nationalen Datenschutzrechts dieses Staates. Die bloße Anmeldung eines Nutzers über eine Maske im Webbrowser

reicht daher nicht aus, um den Anbieter der Anmeldemaske dem nationalen Recht des Landes, in dem sich der Nutzer befindet, zu unterwerfen.

Forum Shopping

Nach einer Entscheidung des Oberverwaltungsgerichts Schleswig-Holstein (Beschluss vom 22.04.2013, AZ: 4 MB 11/13) haben internationale Konzerne die Möglichkeit, durch konzerninterne Organisationsstrukturen und Regelungen der Anwendbarkeit besonders strenger, nationaler Datenschutzregime aus dem Weg zu gehen.

In dem Fall, den das Oberverwaltungsgericht Schleswig-Holstein zu entscheiden hatte, hatte Facebook konzernintern eine Regelung getroffen, nach der die irische Niederlassung für die Erhebung und Verarbeitung der personenbezogenen Daten in Deutschland verantwortlich und auch praktisch zuständig sein sollte und nicht die deutsche Niederlassung. Das Oberverwaltungsgericht Schleswig-Holstein hat die Wirksamkeit dieser Gestaltung bestätigt.

3.1.7.2 Voraussetzungen für die rechtskonforme Datenverarbeitung in der EU

Wenn ein in Deutschland ansässiges Unternehmen Daten innerhalb der EU verarbeitet, gelten, wie bereits in Abschn. 3.1.7.1 dargelegt, die Vorschriften des BDSG. Bei der Einschaltung eines Dritten im europäischen Ausland zur Verarbeitung der Daten sind daher die Erfordernisse des § 11 BDSG, der sogenannten Auftragsdatenverarbeitung, zu beachten. Der Auftragnehmer bildet dann rechtlich gesehen eine Einheit mit dem Auftraggeber.

Gemäß § 11 Abs. 1 BDSG ist der Auftraggeber für die Einhaltung der Vorgaben des BGSG durch den Auftragnehmer verantwortlich. Das Gesetz verpflichtet den Auftraggeber zur sorgfältigen Auswahl und Überwachung des Auftragnehmers. Aufträge über die Auftragsdatenverarbeitung sind schriftlich zu erteilen. Es muss detailliert festgelegt werden, wie die Verarbeitungsprozesse aussehen, welche technischen und organisatorischen Maßnahmen der Auftragnehmer zu ergreifen hat und ob er zur Begründung von unter Auftragsverhältnissen berechtigt ist. Der Auftraggeber ist gegenüber dem Auftragnehmer weisungsbefugt. Der Auftragnehmer hat außerdem bestimmte, gesetzlich vorgesehene Schutzmaßnahmen zu ergreifen.

Gegenstand der rechtspolitischen Diskussion ist seit einigen Jahren die Schaffung eines einheitlichen Datenschutzniveaus innerhalb der EU durch eine sogenannte Datenschutz-Grundverordnung („DS-GVO“). Die konkrete Ausgestaltung dieser DS-GVO ist allerdings sehr umstritten, so dass unklar ist, in welchem Zeitraum ein Inkrafttreten erwartet werden kann.

3.1.7.3 Voraussetzungen für die rechtskonforme Datenverarbeitung in Drittstaaten

Eine Übermittlung personenbezogener Daten in Drittstaaten ist nur unter zwei Voraussetzungen zulässig. Zunächst muss die Übermittlung nach dem BDSG oder einem anderen Gesetz erlaubt sein oder der Betroffene muss in die Übermittlung eingewilligt haben.

Ist eine dieser Alternativen erfüllt, ist in einem zweiten Schritt erforderlich, dass kein schutzwürdiges Interesse des Betroffenen an dem Ausschluss der Übermittlung in den Drittstaat vorliegt. Diese zweite Voraussetzung ist insbesondere dann nicht erfüllt, wenn der Empfänger in einem Land ansässig ist, das aus Sicht der EU über kein angemessenes Datenschutzniveau verfügt.

Die europäische Kommission hat bisher lediglich für folgende Staaten verbindlich festgestellt, dass ein angemessenes Datenschutzniveau vorliegt: Andorra, Argentinien, Färöer Inseln, Guernsey, Isle of Man, Israel, Jersey, Kanada, Neuseeland, Schweiz, Uruguay.

Für den in der Praxis wichtigsten Empfängerstaat, die USA, hat die EU-Kommission bisher kein angemessenes Datenschutzniveau festgestellt.

Die Datenübermittlung in solche Drittstaaten wie die USA ist nur zulässig, wenn besondere Voraussetzungen vorliegen. Eine Möglichkeit ist die Genehmigung der zuständigen Aufsichtsbehörde. Diese kann die Genehmigung zur Datenübermittlung erteilen, wenn ausreichende Garantien für den Datenschutz bestehen. Mögliche Fälle sind die Verwendung der sogenannten EU-Standard-Vertragsklauseln, verbindlicher Unternehmensregelungen („binding corporate rules“ oder „BCR“) oder die Geltung des sogenannten Safe-Harbor-Abkommens.

EU-Standardvertragsklauseln

Die Europäische Kommission hat sogenannte Standardvertragsklauseln für die Übermittlung personenbezogener Daten in Drittländer verabschiedet. Die Verwendung dieser EU-Standardvertragsklauseln führt in der Praxis zur Entbehrlichkeit der Genehmigung der zuständigen Aufsichtsbehörde. An die Stelle der Prüfung des konkreten Sachverhalts tritt hier die vorab erfolgte allgemeine Wertung der Europäischen Kommission.

Sollten die EU-Standardvertragsklauseln nicht unverändert übernommen werden, ist zu unterscheiden. Ergänzungen oder Anpassungen, die lediglich die Erfordernisse gemäß § 11 Abs. 2 S. 2 BDSG abbilden, lösen keine Genehmigungspflicht aus (vgl. 23. Bericht der Landesregierung über die Tätigkeit der für den Datenschutz im nicht öffentlichen Bereich in Hessen zuständigen Aufsichtsbehörde – LT-Drs. 18/2942, Ziff. 11.1; vgl. auch Bayerisches Landesamt für Datenschutzaufsicht, 4. Tätigkeitsbericht 2009/2010, Ziff. 11.3 = RDV 2011, 155).

Vereinbaren die Vertragspartner jedoch individuelle Klauseln, die signifikant von den EU-Standardvertragsklauseln abweichen, kann nicht mehr auf die vorweggenommene Wertung der europäischen Kommission abgestellt werden. Hier ist eine Genehmigung durch die zuständige Aufsichtsbehörde erforderlich.

Binding Corporate Rules – BCR

Eine andere Möglichkeit der legalen Datenübermittlung in Drittstaaten ist die Schaffung von Binding Corporate Rules. Datenschutzrechtlich gibt es kein Konzernprivileg. Die Übermittlung von personenbezogenen Daten – insbesondere Mitarbeiterdaten – von einer Konzerngesellschaft zu einer anderen ist daher grundsätzlich nicht zulässig. Durch die Implementierung von BCR wird konzerninterner Datenverkehr legitimiert.

BCR bieten in der Praxis gegenüber EU-Standardvertragsklauseln den Vorteil, dass eine Anpassung an die konkreten Bedürfnisse des Konzerns möglich ist, soweit das erforderliche Schutzniveau durch andere Regelungen und Maßnahmen garantiert wird. Entscheidender Maßstab ist also auch hier die Garantie eines angemessenen Schutzniveaus.

Inhaltliche Voraussetzungen sind insbesondere (i) die Haftung der Muttergesellschaft des Konzerns oder – sofern deren Sitz sich außerhalb der EU befindet – einer benannten Konzerngesellschaft mit Sitz innerhalb der EU für sämtliche Verstöße aller Konzerngesellschaften sowie (ii) die Implementierung angemessener organisatorischer Maßnahmen zur Kontrolle der Einhaltung der BCR. Solche organisatorischen Maßnahmen sind z. B. die Bildung eines zuständigen Mitarbeiterstabes, Schulungen für die Daten verarbeitenden Mitarbeiter und regelmäßige Datenschutzschutzaudits.

Ob im Fall von BCR eine aufsichtsbehördliche Genehmigung erforderlich ist, wird von den Aufsichtsbehörden unterschiedlich bewertet. Jedenfalls aber aus Gründen der Absicherung und Minimierung von Haftungsrisiken ist es ratsam, bei der Implementierung von BCR die zuständigen Aufsichtsbehörden einzuschalten.

Safe Harbor Principles

Die Safe Harbor Principles sind gemeinsam von der europäischen Kommission und dem amerikanischen Handelsministerium als Maßstab für ein angemessenes Datenschutzniveau aufgestellt worden. Hat sich ein Unternehmen zur Einhaltung dieser Prinzipien verpflichtet, reicht das nach Ansicht der deutschen Datenschutzaufsichtsbehörden jedoch allein nicht aus, um den Datentransfer in die USA datenschutzkonform zu gestalten.

Hintergrund ist, dass die Einhaltung der Safe Harbor Principles in den USA nicht staatlich überwacht wird. Deutsche Unternehmen müssen sich daher von den amerikanischen Datenempfängern nachweisen lassen, dass die Safe Harbor Principles tatsächlich eingehalten werden.

Das Safe Harbor Programm wird insgesamt kritisch betrachtet, insbesondere wegen der fehlenden staatlichen Kontrolle, aber auch wegen einer mutmaßlichen Überschreitung der Kompetenzen der Europäischen Kommission durch Aushandlung der Prinzipien. Es bleibt daher abzuwarten, ob das Safe Harbor Programm nach Inkrafttreten der DS-GVO weiter existieren wird.

3.1.7.4 Praxisfall Cloud Computing

Die Nutzung externer Cloud-Services bietet Unternehmen reizvolle Vorteile. Insbesondere bedeutet derartiges Outsourcing von IT-Dienstleistungen in der Regel ein hohes Kostenersparnispotential, da Posten für ungenutzte Kapazitäten und deren Wartung oder die regelmäßige Investition in neue Hard- und Software wegfallen. Datenschutzrechtlich sind derartige Outsourcing-Maßnahmen aber nicht unproblematisch.

Cloud Computing wird mehrheitlich als Auftragsdatenverarbeitung qualifiziert, sodass die in Abschn. 3.1.7.2 bereits angedeuteten Voraussetzungen gemäß § 11 BDSG eingehalten werden müssen. Dies ist in der Praxis mit verschiedenen Problemen verbunden.

Cloud Computing innerhalb der EU

Wie bereits in Punkt I. 1. dargestellt, gilt im innereuropäischen Datenverkehr das Sitzlandprinzip. Dies führt beim Cloud Computing aber zu einer Umkehrung des ursprünglich beabsichtigten Effekts, dass sich Unternehmen nicht mit den nationalen Unterschieden der Datenschutzvorschriften auseinandersetzen müssen.

Da Cloud-Computing mehrheitlich als Auftragsdatenverarbeitung qualifiziert wird, ist der Cloud-Nutzer datenschutzrechtlich die verantwortliche Stelle ist (vgl. Abschn. 3.1.7.2). Für den Cloud-Anbieter hat das zur Folge, dass er durch die verschiedenen Sitzländer seiner Nutzer auch die Einhaltung verschiedener Datenschutzgesetze beachten muss. Hat der Cloud-Anbieter Nutzer aus allen Mitgliedstaaten der EU muss er alle nationalen Datenschutzvorschriften einhalten. Der hiermit verbundene zusätzliche Aufwand kann also signifikant sein, das Haftungsrisiko ebenfalls.

Ob und inwieweit die neue DS-GVO hier Abhilfe schaffen wird, bleibt abzuwarten.

Cloud Computing in Drittstaaten

Beim Cloud-Computing außerhalb der EU treten weitere Besonderheiten auf. Aufgrund der potentiell großen Vielzahl an Rechenzentren in verschiedenen Staaten, die ein Cloud-Anbieter zur Erbringung seiner Leistung nutzt, besteht die Möglichkeit vielfältigen Zugriffs auf die Daten durch nationale Behörden aufgrund von sicherheitsrechtlichen Zugriffsrechten. Dieses Thema ist insbesondere bei Cloud-Anbietern aus den USA oder mit dort befindlichen Rechenzentren von Bedeutung. Der amerikanische Patriot Act gewährt umfangreiche Zugriffsrechte unter dem Vorwand der nationalen Sicherheit. Der Cloud-Nutzer kann nur schwer erkennen und vorab einschätzen, wer in welchen Fällen Zugriff auf seine Daten haben kann.

Ein weiteres Problem ist die Durchsetzung des hohen europäischen Datenschutzniveaus in Drittstaaten. Die faktische Gewährleistung des Datenschutzes erfordert entsprechende Durchsetzungsmechanismen. Auch dies kann bei der Beteiligung einer Vielzahl von Staaten, möglicherweise auch im Rahmen von Unterauftragsverhältnissen, nur schwer sichergestellt werden.

3.1.7.5 Zusammenfassung

Unternehmen müssen bei der grenzüberschreitenden Datenübermittlung zahlreiche rechtliche Vorgaben und Risiken beachten. Dies gilt in besonderem Maße bei der Übermittlung in Drittstaaten außerhalb der EU und beim Cloud Computing.

Um diese Risiken zu managen und zu minimieren, ist die Verwendung der EU-Standardvertragsklauseln zu empfehlen. Darüber hinaus ist es ratsam bei jeder Anpassung oder Änderung dieser EU-Standardvertragsklauseln für einen konkreten Fall die zuständigen Aufsichtsbehörden zu konsultieren, um diese Anpassungen so rechtssicher wie möglich zu gestalten.

3.1.8 Big Data in der Personalabteilung

Thorsten Walter

3.1.8.1 Einführung

Seit den 90er-Jahren erfassen und analysieren Unternehmen Daten zur Unterstützung operativer und strategischer Entscheidungen. Im Supply-Chain-Management und in der Vertriebsanalyse nutzen die Unternehmen die aus der Datenanalyse gewonnenen Erkenntnisse zur Steuerung ihrer Prozesse. Im Filial- und Einzelhandel setzen Unternehmen die Analyseergebnisse zur Personalisierung ihres Waren- und Dienstleistungsangebotes ein. In vielen operativen Bereichen steigern die aus der Datenanalyse gewonnenen Erkenntnisse den Unternehmenserfolg und generieren Wettbewerbsvorteile.

Die immer stärkere Digitalisierung gewerblicher und privater Lebensbereiche führt zu einer stetig wachsenden Datenflut. Mit der wachsenden Menge der Daten wächst das Bedürfnis, diese Daten zu erschließen und nutzbar zu machen.

Während Big Data-Methoden heute in vielen operativen und industriellen Bereichen eingesetzt werden, werden die mit dem Einsatz dieser Technik verbundenen Vorteile im Human-Resources-Bereich kaum genutzt. In den Personalabteilungen sind Prozesse zur Sammlung und Auswertung von Daten selten. Eine strukturierte Auswertung von Daten zur gezielten Verbesserung von Personalentscheidungen ist die Ausnahme.

Neben den technischen Berührungsängsten sind die datenschutz- und betriebsverfassungsrechtlichen Hürden ein Hemmschuh für die Verbreitung von Big Data.

3.1.8.2 Daten, Daten und noch mehr Daten

Neben rein arbeitsplatzbezogenen Daten (z. B. Stellenbeschreibungen, Leistungs- und Anforderungsprofilen) werden in jedem Unternehmen die Stammdaten der Belegschaft (z. B. Anschrift, Alter, Unterhaltpflichten), die Informationen aus dem Bewerbungsprozess (z. B. Zeugnisse, Bewerbungsunterlagen, Zertifikate) und die Daten der Lohnabrechnung (z. B. Steuerklasse, Unterhaltpflichten, Familienstand) sowie alle sonstigen vergütungsrelevanten Daten (z. B. tarifliche Eingruppierung, Vergütung, Boni, Sonderzahlungen, betriebliche Altersversorgung) gespeichert und verarbeitet. Es werden Fehlzeiten und die Arbeitszeit erfasst. Die von den IT-Systemen generierten Log- und Protokolldateien (z. B. Daten biometrischer Zugangskontrollsysteme, Datenbankprotokolle) liefern Informationen, die sich in den Zusammenhang mit dem Mitarbeiter bringen lassen.

Ließen sich diese Informationen mit anderen, externen Datenbeständen, beispielsweise Einträge in Social-Media-Profilen (z. B. XING, LinkedIn oder Facebook), Blog-Einträgen, Kundenprofilen, Krankheits- und Wetterstatistiken kombinieren, ergeben sich ungeahnte Möglichkeiten für den Human-Resources-Bereich.

Beispielweise ermöglicht die strukturierte Auswertung von Bewerberdaten, Social-Media-Profilen sowie Recruiting- und Personalmarketingplattformen verbesserte Personalauswahlentscheidungen bei Neueinstellungen oder bei der (Nach-)Besetzung von Stellen. Die Ergebnisse der Auswertung von berufsbezogenen Social-Media-Plattformen und

der im Unternehmen gespeicherten Informationen zum beruflichen Werdegang und zur Qualifikation von Mitarbeitern lassen sich für ein verbessertes Talent-Management nutzen. Die Auswertung von Informationen zum Kaufverhalten bestimmter Kundengruppen, Veranstaltungsmittelungen und Wetterdaten hilft bei der Ermittlung des Personalbedarfs und der Identifikation von Belastungsspitzen. Die Kombination unternehmensinterner Daten über krankheitsbedingte Fehlzeiten mit Statistiken über Erkrankungen und deren Verlauf verbessert die Personalplanung. Die Auswertung geleisteter Arbeitszeit und Abrechnungsdaten macht die Effizienz einzelner Mitarbeiter, Mitarbeitergruppen oder Abteilungen messbar. Biometrische Zugangskontrollsysteeme ermöglichen die Erstellung von Bewegungsprofilen und lassen Rückschlüsse auf das Sozialverhalten der Mitarbeiter zu.

Durch den Einsatz von Big Data lassen sich Personalmaßnahmen effizienter steuern. Das Recruiting und Talentmanagement profitiert in besonders hohem Maße von der Auswertung dieser Daten. Allerdings wird gerade in diesen Bereichen besonders stark in die Persönlichkeitsrechte der betroffenen Mitarbeiters eingegriffen.

3.1.8.3 Problemstellung

Die unternehmensintern gespeicherten Daten lassen sich mit externen Daten kombinieren und zu aussagekräftigen Profilen über die betroffenen Mitarbeiter zusammenstellen. Gegenstand der Datenanalyse sind die personenbezogenen Daten der Mitarbeiter des Arbeitgeberunternehmens. Diese Daten schützt das Bundesdatenschutzgesetz. Sie können nicht ohne Weiteres gesammelt, gespeichert und verarbeitet werden.

Datenschutzrecht

Im Zusammenhang mit der Verarbeitung von Mitarbeiterdaten ergeben sich zahlreiche Fragestellungen. Dürfen personenbezogene Mitarbeiterdaten überhaupt gesammelt werden? Müssen die Daten gelöscht werden? Und wenn ja, wann? Dürfen die Daten gespeichert, analysiert und ausgewertet werden? Wenn nein, wäre dies zulässig, wenn die Daten soweit anonymisiert sind, dass die hinter dem Datum stehende Person nicht zuordenbar ist?

§ 32 Abs. 1 Bundesdatenschutzgesetz als Erlaubnisnorm

Die (noch) geltende Regelung zum Arbeitnehmerdatenschutz in § 32 Bundesdatenschutzgesetz setzt der Nutzung von Mitarbeiterdaten enge Grenzen, wenn die Nutzung nicht unmittelbar für die Begründung, Durchführung oder Beendigung des Arbeitsverhältnisses erforderlich ist. „Erforderlich“ bedeutet, dass die Verarbeitung von personenbezogenen Mitarbeiterdaten geeignet und zugleich das mildeste Mittel ist, um den unternehmerischen Interessen des Arbeitgebers bei der Begründung, Durchführung und Beendigung von Arbeitsverhältnissen gerecht zu werden (Seifert 2014, § 32 Rn. 11). Dabei muss die Datenverarbeitung geboten, nicht nur nützlich sein (Gola und Schomerus 2015, § 32 Rn. 12). Der Anwendungsbereich von § 32 Bundesdatenschutzgesetz ist eng. Steht die Datenverarbeitung nicht in unmittelbarem Zusammenhang mit dem in § 32 Bundesdatenschutzgesetz geregelten Verarbeitungszweck, ist sie unzulässig.

Die Big Data-Analyse dient primär der Unternehmenssteuerung. Für die Durchführung oder Beendigung eines Arbeitsverhältnisses ist sie nicht erforderlich. Die Erkenntnisse einer Big Data-Analyse könnten für die Begründung des Arbeitsverhältnisses nützlich sein. Die Auswertung von Recruiting- und Personalmarketingplattformen, Social-Media-Profilen und Bewerbungsunterlagen ist der Personalauswahlentscheidung bei der Einstellung neuer Mitarbeiter sicherlich dienlich, zwingend geboten ist sie nicht. Zudem sind Backgroundchecks von Mitarbeitern datenschutzrechtlich unzulässig, weil sie gegen den in § 4 Abs. 2 Bundesdatenschutzgesetz normierten Grundsatz der Direkterhebung verstößen (Bäcker 2015, § 4 Rn. 28 ff.). Danach sind Daten direkt beim Bewerber zu erheben (Gola und Schomerus 2015, § 4 Rn. 19 ff.). Dies gilt für die Big Data-Analyse entsprechend.

§ 32 Bundesdatenschutzgesetz ist keine taugliche Erlaubnisnorm zur Rechtfertigung der Verarbeitung von Mitarbeiterdaten mit Big Data-Methoden.

§ 28 Abs. 1 Satz 1 Nr. 3 Bundesdatenschutzgesetz als Erlaubnisnorm

§ 28 Bundesdatenschutzgesetz erlaubt die Erhebung von personenbezogenen Daten für die Erfüllung eigener Zwecke, wenn dies zur Wahrung berechtigter Interessen des Arbeitgeberunternehmens erforderlich ist und die schutzwürdigen Interessen des Mitarbeiters am Ausschluss der Datenverarbeitung nicht überwiegen.

Die Anwendbarkeit von § 28 Bundesdatenschutzgesetz setzt voraus, dass es sich um Daten handelt, die frei zugänglich sind und der Arbeitgeber ein berechtigtes Interesse an ihrer Erhebung hat. Ob Daten, die in berufsbezogenen Netzwerken wie XING oder LinkedIn gespeichert und nur nach Registrierung und Anmeldung beim Portal zugänglich sind, „allgemein zugänglich“ im Sinne von § 28 Abs. 1 Satz 1 Nr. 3 Bundesdatenschutzgesetz sind, ist umstritten (Thüsing und Forst 2014, § 7 Rn. 8; Schiedermaier 2014, § 25 Rn. 83; Spindler und Nink 2015, § 28 BDSG Rn. 7). Zudem dürfen Daten, die nach dem Allgemeinen Gleichbehandlungsgesetz zu einer Diskriminierung führen würden nicht verarbeitet werden (Thüsing und Forst 2014, § 7 Rn. 1). Gleiches gilt für Daten, deren Erhebung nach den von der Rechtsprechung entwickelten Grundsätzen zum Fragerecht des Arbeitgebers (z. B. Frage nach bestehender Behinderung, Schwangerschaft, Vorstrafen oder Gewerkschaftszugehörigkeit) unzulässig wäre (Linck 2013, § 151 Rn. 17). Ferner ist ein berechtigtes Interesse des Arbeitgebers an der Datenerhebung erforderlich.

Für das Unternehmen bedeutet dies ein hohes Maß an Unsicherheit. Die Grundsätze zum Fragerecht des Arbeitgebers sind einzelfallabhängig und teilweise in der Rechtsprechung stark umstritten. Weitere Unsicherheiten liegen im Streit über die Frage der Zugänglichkeit von Daten in berufsbezogenen Netzwerken und in der notwendigen Interessenabwägung. Für das Unternehmen besteht ein hohes Risiko, dass der Datenzugriff unzulässig ist.

Einwilligung des Betroffenen

Eine Datenverarbeitung ist gemäß § 4a Bundesdatenschutzgesetz mit (wirksamer) Einwilligung des Betroffenen zulässig. Eine Einwilligung des Mitarbeiters ist erst dann wirksam,

wenn sie nach angemessener Information freiwillig erfolgt (Simitis 2014, § 4a Rn. 70). Zudem darf sie nicht pauschal erfolgen. Der Zweck der Datenverarbeitung ist so konkret wie möglich anzugeben (Simitis 2014, § 4a Rn. 72). Die Einwilligung ist jederzeit frei widerruflich.

Die Aufsichtsbehörden stellen im Kontext eines Arbeitsverhältnisses hohe Anforderungen an die Freiwilligkeit der Einwilligung des Mitarbeiters. Für das Unternehmen besteht das Risiko, dass die Einwilligung des Mitarbeiters unwirksam, der Datenzugriff deshalb rechtswidrig ist. Die Einwilligung ist mit einigen rechtlichen Risiken für das Unternehmen verbunden. Eine taugliche Ermächtigungsgrundlage für die Verarbeitung von Mitarbeiterdaten ist sie in aller Regel nicht.

Anonymisierung personenbezogener Mitarbeiterdaten

Die Datenschutzgesetze verbieten die Verarbeitung personenbezogener Daten. Die Person, auf die sich das Datum bezieht, muss bestimmt oder zumindest bestimbar sein. Bestimmbarkeit liegt vor, wenn die Person allein durch die Daten nicht identifiziert werden kann, aus dem Kontext einer Information aber auf die Identität der Person geschlossen werden kann. Ermöglicht die Kombination verschiedener Informationen die Identifikation einer Person, handelt es sich um ein personenbezogenes Datum, das dem uneingeschränkten Schutz der Datenschutzgesetze unterliegt.

Die Anonymisierung personenbezogener Daten löst das datenschutzrechtliche Problem des Arbeitgebers nur dann, wenn durch die Anonymisierung der Personenbezug vollständig und unwiederbringlich aufgehoben wird oder die Information soweit unkenntlich gemacht wird, dass eine Identifikation der hinter den Informationen stehenden Person nur mit unverhältnismäßig großem Aufwand möglich ist. Die Anonymisierung führt nicht zur Vernichtung von Daten, sie entfernt nur diejenigen Informationen, die einen Personenbezug ermöglichen.

Die Besonderheit von Big Data besteht darin, große Datenmengen aus unterschiedlichen Datenquellen zu sammeln und auszuwerten. Dabei werden in einem beliebig großen Datenbestand unerkannt gebliebene Muster, Regeln oder Verbindungen durch den softwaregestützten Einsatz von Algorithmen visualisiert. Die im Anonymisierungsprozess durch das Entfernen einzelner Informationen entstandenen Lücken lassen sich durch andere Merkmale ergänzen und machen einen Personenbezug möglich. Die Wiederherstellung des Personenbezugs von anonymisierten Daten ist durch den Einsatz von Big Data-Methoden in aller Regel ohne unverhältnismäßigen Aufwand möglich.

Fazit

Die hohen datenschutzrechtlichen Hürden werden die Einführung von Big Data im Human-Resources-Bereich erschweren, schlimmsten Falls sogar verhindern. Eine Lösung für dieses nicht auflösbare Dilemma ist nicht in Sicht. Die politischen Bemühungen um eine Novellierung des Arbeitnehmerdatenschutzrechts machen seit Jahren keine erkennbaren Fortschritte. Von politischer Seite ist keine Unterstützung zu erwarten. An technischen Lösungen für eine permanente Daten-Anonymisierung wird mit Hochdruck gearbeitet.

Die Industrie arbeitet an Verschlüsselungsmethoden, die eine dauerhafte Anonymisierung personenbezogener Daten ermöglichen sollen. Ob es gelingen wird, personenbezogene Daten so zu anonymisieren, dass eine Re-Identifikation ausgeschlossen ist und diese Technologie der rasanten technischen Entwicklung in diesem Bereich standhalten wird, ist offen.

Betriebsverfassungsrecht

Neben den datenschutzrechtlichen Hürden erschwert das Betriebsverfassungsrecht den Einsatz von Big Data. Das Betriebsverfassungsgesetz und das Personalvertretungsrecht enthalten bislang keine spezialgesetzlichen Regelungen zur Beteiligung der Mitarbeitervertretungen bei der Einführung automatisierter Personaldatenverarbeitung. Deshalb sind die vorhandenen Beteiligungsrechte auf die Sachverhalte automatisierter Personaldatenverarbeitung anzuwenden.

§ 94 Betriebsverfassungsgesetz – Personalfragebögen

Nach § 94 Betriebsverfassungsgesetz bedürfen Personalfragebögen der Zustimmung des Betriebsrates. Das Mitbestimmungsrecht erfasst die Zusammenstellung von Fragen über persönliche Verhältnisse der Mitarbeiter, einschließlich der Bewerber, ihrer Ausbildung, ihrem Fertigkeiten, ihren beruflichen Werdegang, Krankheiten, Verwandtschaftsverhältnissen usw. Das Mitbestimmungsrecht erfasst nur die einzelnen Fragen, nicht aber die Verwendung der hieraus gewonnenen Erkenntnisse (Kania 2015, § 94 BetrVG Rn. 3). Ob die Datenerhebung bei Dritten, also nicht beim Mitarbeiter selbst, von § 94 Betriebsverfassungsgesetz erfasst wird, ist offen. Offen ist auch, ob § 94 Betriebsverfassungsgesetz das Beteiligungsrecht des Betriebsrates bei der Einführung von Big Data auslöst. Wegen der rechtlichen Unsicherheiten über den Anwendungsbereich von § 94 Betriebsverfassungsgesetz sollte das Unternehmen die Zustimmung des Betriebsrates vorsorglich einholen.

§ 92 Betriebsverfassungsgesetz – Personalplanung

Der Arbeitgeber ist gemäß § 92 Betriebsverfassungsgesetz verpflichtet, die Erkenntnisse aus der Datenauswertung ergebenden Planungsüberlegungen mit dem Betriebsrat zu beraten. „Personalplanung“ im Sinne von § 92 Betriebsverfassungsgesetz erfasst die gesamte Personalpolitik, also die Personalbedarfs-, die Personalbeschaffungs- und Personalentwicklungsplanung (Thüsing 2014, § 92 Rn. 3 ff.). Die Rechtsprechung hat ein Beteiligungsrecht des Betriebsrates für die Fälle der Errichtung, des Ausbaus und der Auswertung von Personalinformationssystemen bereits anerkannt (Bundesarbeitsgericht, Beschluss v. 11.03.1986 – 1 ABR 12/84).

Diese Grundsätze wird man bei der Einführung von Big Data übertragen können. Danach wäre das Unternehmen verpflichtet, die Einführung von Big Data mit dem Betriebsrat zu beraten. § 92 Betriebsverfassungsgesetz normiert aber „nur“ eine Beratungspflicht. Der Zustimmung des Betriebsrates bedarf die Einführung von Big Data danach nicht.

§ 87 Abs. 1 Nr. 1 und § 87 Abs. 1 Nr. 6 Betriebsverfassungsgesetz

Gemäß § 87 Abs. 1 Betriebsverfassungsgesetz hat der Betriebsrat ein erzwingbares Mitbestimmungsrecht in allen Fragen, die die Ordnung des Betriebes betreffen. Nach § 87 Abs. 1 Nr. 6 Betriebsverfassungsgesetz ist die Einführung und Anwendung technischer Einrichtungen zur Überwachung von Verhalten und Leistung der Arbeitnehmer mitbestimmungspflichtig und zwar unabhängig davon, ob eine Überwachung beabsichtigt oder gewollt ist. Die bloße technische Möglichkeit der Überwachung reicht aus, um das Mitbestimmungsrecht des Betriebsrates auszulösen (st. Rspr. seit Bundesarbeitsgericht, Beschluss v. 09.09.1975 – 1 ABR 20/74).

Big Data ist eine technische Einrichtung im Sinne von § 87 Abs. 1 Nr. 6 Betriebsverfassungsgesetz. Eine Leistungsüberwachung ist durch die Nutzung dieser Technologie möglich, wenn nicht sogar gewollt. Der Mitbestimmungstatbestand des § 87 Abs. 1 Nr. 6 Betriebsverfassungsgesetz ist ausgelöst. Die Einflussnahmemöglichkeit des Betriebsrates ist groß. Sie besteht nicht nur bei der Einführung, sondern auch bezüglich der Art und Weise der Durchführung und jeder Änderung. Damit unterliegen alle Änderungen, die die verarbeiteten Daten, die Programmabläufe und den Zugriffsschutz betreffen, der Mitbestimmung.

Fazit

Besteht im Unternehmen ein Betriebsrat, unterliegt die Einführung Big Data der Mitbestimmung des Betriebsrates. Das Mitbestimmungsrecht des Betriebsrates ist zwingend. Der Arbeitgeber kann den Einsatz von Big Data nicht gegen den Willen des Betriebsrates durchsetzen. Lässt sich eine Einigung nicht herstellen, steht dem Arbeitgeber der Weg zur Einigungsstelle offen. Die Einigungsstelle entscheidet verbindlich über die Zulässigkeit der beabsichtigten Maßnahme.

Haben Unternehmen und Betriebsrat sich auf eine Regelung verständigt, wird hierüber eine Betriebsvereinbarung geschlossen. Aufgrund der eingeschränkten Geltung des AGB-Rechts auf Betriebsvereinbarungen ist der Gestaltungsspielraum von Unternehmen und Betriebsrat größer als bei arbeitsvertraglichen Vereinbarungen. Zudem wirken Änderungen einer Betriebsvereinbarung automatisch auf alle Arbeitsverhältnisse, was den administrativen Aufwand gering hält.

3.1.8.4 Zusammenfassung

Der Einsatz von Big Data wird zur Herausforderung für das Unternehmen, wenn Mitarbeiterdaten betroffen sind. Es gibt keinen gesetzlichen Erlaubnistatbestand, der eine Datenverarbeitung im Rahmen eines Big Data-Prozesses rechtsverbindlich zulässig macht. Die Anwendung bestehender gesetzlicher Erlaubnistatbestände ist mit großen Rechtsunsicherheiten verbunden. Dass ein novelliertes Datenschutzrecht den praktischen Anforderungen Rechnung tragen wird, ist unwahrscheinlich. Hoffnung machen technische Lösungen, die die notwendigen Daten soweit anonymisieren, dass der Personenbezug unwiederbringlich aufgehoben wird und die Daten dadurch ohne Verstoß gegen datenschutzrechtliche Grundsätze nutzbar werden. Die hierfür notwendigen Technologien sind noch im Ent-

wicklungsstadium. Ob sie der rasanten technischen Entwicklung dauerhaft Stand halten werden, ist unsicher. Nach der aktuellen Rechtslage ist die Verarbeitung von personenbezogenen Mitarbeiterdaten im Big Data Prozess mit einigen rechtlichen Unsicherheiten und Risiken verbunden.

3.1.9 Automatisierte Entscheidungen und Scoring

Joachim Dorschel

Sinn und Zweck von Big Data ist es insbesondere, Entscheidungen auf Basis von Datenanalysen zu treffen. Je nach Anwendungsbereich können solche Entscheidungen weitreichende Folgen für die Betroffenen haben. Besonders kritisch ist dies in Fällen, in denen die Betroffenen keinen Einfluss auf die Datengrundlage und die angewandten Verfahren haben. Das Datenschutzrecht reglementiert daher die Verfahrensweisen, Entscheidungen allein auf die automatisierte Verarbeitung personenbezogener Daten zu stützen (§ 6a BDSG, hierzu Abschn. 3.1.9.1) wie auch das Scoring (§ 28b BDSG, hierzu Abschn. 3.1.9.2).

3.1.9.1 Automatisierte Einzelentscheidungen

§ 6a BDSG regelt Verfahren, in denen ein Computer auf Grundlage personenbezogener Daten Entscheidungen trifft, die für den Betroffenen von wesentlicher Bedeutung sind. Grundgedanke der Norm ist es, dass Menschen nicht zum bloßen Objekt von Computeralgorithmen werden sollen (von Lewinski 2015, § 6a Rn 1). Die Vorschrift ist damit letztlich Ausdrucksform des in Art. 1 Abs. 1 des Grundgesetzes postulierten Gebots der Menschenwürde, das es grundsätzlich verbietet, Menschen zu bloßen Objekten zu degradieren.

Verbot reiner Computerentscheidungen

Nach § 6a Abs. 1 BDSG ist es verboten, „*Entscheidungen, die für den Betroffenen eine rechtliche Folge nach sich ziehen oder ihn erheblich beeinträchtigen, ... ausschließlich auf eine automatisierte Verarbeitung personenbezogener Daten*“ zu stützen, „*die der Bewertung einzelner Persönlichkeitsmerkmale dienen*“. Dabei soll eine ausschließlich auf eine automatisierte Verarbeitung gestützte Entscheidung insbesondere dann vorliegen, wenn keine inhaltliche Bewertung und darauf gestützte Entscheidung durch eine natürliche Person stattgefunden hat. Die Vorschrift kommt also immer dann zur Anwendung, wenn ein Computer Persönlichkeitsmerkmale bewertet und kein Mensch Einfluss auf die Bewertung und Entscheidung nimmt.

Unsicher ist, welche personenbezogenen Daten Persönlichkeitsmerkmale im Sinne des § 6a Abs. 1 BDSG sind. Die der Vorschrift zugrundeliegende EU-Richtlinie nennt beispielhaft die berufliche Leistungsfähigkeit, die Kreditwürdigkeit, die Zuverlässigkeit oder das Verhalten einer Person. Nicht zu den Persönlichkeitsmerkmalen sollen äußere Parameter gehören, etwa die Entfernung des Wohnorts vom Arbeitsplatz oder das Ranking

einer Internetseite (von Lewinski 2015, § 6a Rn 9). Es liegt auf der Hand, dass in diesem Bereich viele Grenzfälle denkbar sind, so dass Rechtsunsicherheiten nicht ausgeschlossen werden können.

Die Vorschrift betrifft nur solche Entscheidungen, die für den Betroffenen eine rechtliche Folge nach sich ziehen oder erheblich beeinträchtigen. Ein wichtiges Beispiel im Kontext von Big Data ist der Nichtabschluss eines Vertrages, wenn etwa ein Kreditantrag ausschließlich auf Grundlage einer automatisierten Datenanalyse geprüft und bearbeitet wird (siehe hierzu auch Abschn. 2.8.3.3).

Ausnahmen

§ 6a Abs. 2 BDSG regelt zwei Ausnahmen vom Verbot der automatisierten Einzelentscheidung, die der Vorschrift einen großen Teil ihrer Schärfe nehmen.

Gemäß § 6a Abs. 2 Ziff. 1 BDSG gilt das Verbot nicht, wenn die Entscheidung im Rahmen des Abschlusses oder der Erfüllung eines Vertragsverhältnisses oder eines sonstigen Rechtsverhältnisses ergeht und einem Begehrten des Betroffenen stattgegeben wurde. Wird in dem zitierten Beispiel des Online-Kreditantrags dem Antrag entsprochen, ist hiernach eine automatisierte Entscheidung zulässig.

Unsicherheit herrscht allerdings darüber, ob sich die Stattgabe nur auf den Vertragsabschluss selbst oder auch auf die Konditionen beziehen muss. Das Konzept des § 6a BDSG ließe leer, wenn die verantwortliche Stelle beliebige Konditionen, im Kreditbeispiel etwa einen Zinssatz weit über dem marktüblichen, verlangen könnte. Vorgeschlagen wird hier, den Ausnahmetatbestand jedenfalls dann anzuwenden, wenn die angebotenen Konditionen marktüblich sind (von Lewinski 2015, § 6a BDSG, Rn 40).

Gemäß § 6a Abs. 2 Ziff. 2 BDSG ist auch eine ablehnende Entscheidung auf Grundlage eines automatisierten Verfahrens möglich, wenn ein Verfahren vorgesehen ist, durch das die schutzwürdigen Interessen des Betroffenen gewahrt werden und dem Betroffenen die Tatsache einer automatisierten Einzelentscheidung sowie auf sein Verlangen auch die wesentlichen Gründe mitgeteilt werden.

Entscheidende Voraussetzung für die Anwendung dieser Ausnahmebestimmung ist die Wahrung der schutzwürdigen Interessen des Betroffenen. Das Gesetz macht keine Vorgaben, welche Maßnahmen hierzu geeignet sind. In der zugrundeliegenden EU-Richtlinie wird beispielhaft die Möglichkeit des Betroffenen genannt, seinen Standpunkt individuell geltend zu machen. Für die Praxis ist daher zu empfehlen, in Situationen, in denen der Computer zu einer ablehnenden Entscheidung kommt, dem Betroffenen die Möglichkeit aufzuzeigen, seinen Standpunkt individuell darzulegen und die dargelegten Umstände einzelfallbezogen zu berücksichtigen (Gola und Schomerus 2015, § 6a Rn 14b).

3.1.9.2 Scoring

§ 28b BDSG regelt die Erhebung und Verwendung von Wahrscheinlichkeitswerten „für ein bestimmtes zukünftiges Verhalten des Betroffenen“ zum Zweck der Entscheidung über die Begründung, Durchführung oder Beendigung eines Vertragsverhältnisses. Die Vorschrift betrifft damit insbesondere Bonitäts-Ratings, die bei der Kreditvergabe, Ver-

mietungen und der Begründung von anderen Dauerschuldverhältnissen in der Praxis eine wichtige Rolle spielen.

Anders als § 6a BDSG gilt § 28b für manuelle und maschinelle Verfahren. Die Vorschrift regelt bereits die Erhebung des Scoringwertes, gilt also unabhängig davon, ob dieser Wert tatsächlich Einfluss auf eine Entscheidung hat. Werden Scoringwerte bei einer Entscheidung im Sinne von § 6a BDSG verwendet, gelten die Vorschriften nebeneinander.

Ein Scoring ist nach § 28b BDSG unter folgenden Voraussetzungen zulässig:

- Die zur Berechnung des Wahrscheinlichkeitswerts genutzten Daten müssen nachweisbar „unter Zugrundelegung eines wissenschaftlich anerkannten mathematisch-statistischen Verfahrens“ für die Berechnung des Wahrscheinlichkeitswerts erheblich sein.
- Die Nutzung und ggf. Übermittlung der Daten muss datenschutzrechtlich zulässig sein.
- Für die Berechnung des Wahrscheinlichkeitswertes dürfen nicht ausschließlich Anschriftendaten genutzt werden.
- Bei einer Nutzung von Anschriftendaten muss der Betroffene vor Berechnung des Wahrscheinlichkeitswertes über die vorgesehene Nutzung dieser Daten unterrichtet worden sein.

Bei einer Bonitätsprüfung zum Zwecke der Kreditgewährung ist eine Heranziehung der in § 10 Abs. 1 Satz 6 KWG genannten Kriterien jedenfalls als zulässig anzusehen (von Lewinski 2015, § 28b Rn 31.1).

3.2 Leistungsschutz

Carsten Ulbricht und Jenny Hubertus

3.2.1 Urheberrecht an Daten

Die bisherigen rechtlichen Ausführungen zum Datenschutz setzen sich mit der Legitimität der Verwendung von Big Data im Verhältnis zu den sogenannten Betroffenen auseinander, also denen, auf die sich die jeweiligen Daten beziehen und damit (zumindest mittelbar) etwas über diese aussagen (können).

Bei Big Data-Projekten sollten allerdings stets auch die urheber- bzw. datenbankrechtlichen Implikationen bedacht und rechtskonformen Lösungen zugeführt werden. Es geht beim Urheberrecht um den rechtlichen Schutz derer, die ein konkretes Werk (z. B. Texte, Bilder, Videos etc.) erstellt haben. Beim Datenbankrecht werden rechtliche Ansprüche derjenigen, die eine Datenbank unter Einsatz von Aufwand oder Investitionen erstellt haben.

In diesem Abschnitt geht es – untechnisch gesprochen – um die Frage, wem bestimmte Daten und Inhalte „gehören“.

Wer entsprechend geschützte Inhalte oder Daten für eigene Zwecke entnimmt, verletzt damit also unter Umständen Urheber- oder Datenbankrechte. Demgemäß sollten zur Vermeidung rechtlicher Weiterungen bei Big Data-Projekten auch die nachfolgenden rechtlichen Regelungen beachtet werden.

3.2.1.1 Internationales Urheberrecht

Zunächst ist darauf hinzuweisen, dass die Beurteilung urheberrechtlicher Fragen maßgeblich von dem jeweils anzuwendenden Recht abhängt.

Der Geltungsbereich des nationalen Urhebers ist grundsätzlich auf das Gebiet des jeweiligen Staates beschränkt (Territorialitätsprinzip). Das Territorialitätsprinzip gilt auch für das deutsche Urhebergesetz (UrhG). Die Folge des Territorialitätsprinzip ist, dass sich der Urheber ebenso wie der international tätige Verwender nicht einem einheitlich weltweit gültigen Urheber- bzw. Leistungsschutzrecht gegenüber sieht, sondern mit einem Bündel nationaler Regelungen mit zum Teil erheblich unterschiedlichem Gehalt konfrontiert wird.

Aus dem Grundsatz des Territorialitätsprinzips wird in der Regel die Geltung des sogenannten Schutzlandprinzips abgeleitet. Nach dem Schutzlandprinzip soll die Entstehung des Urheberrechts, die erste Inhaberschaft ebenso wie der Inhalt und die Schranken des Urheberrechts nach dem Recht des Landes zu beantworten sein, für dessen Gebiet Schutz beansprucht wird (vgl. BGH, GRUR 2003, 328 – Sender Felsberg; GRUR 1999, 152 (153) – Spielbankaffäre).

Nach dem Schutzlandprinzip kommt es mithin entscheidend darauf an, wer Inhaber der jeweiligen Schutzrechte ist und wo dieser seinen Sitz hat. Mithin beurteilen sich Entstehung, Übertragung, Beendigung, Umfang und Schutzdauer von Urheber- und Leistungsschutzrechten nach dem Recht desjenigen Landes, für dessen Gebiet sie in Rede stehen. Mithin sind nach Fertigstellung des Gutachtens – je nach Bedarf – noch weitergehende Bewertungen für andere Rechtsordnungen (z. B. USA) einzuholen.

3.2.1.2 Urheberrechtliche Schutzfähigkeit von Informationen und Daten

Das Urheberrechtsgesetz (UrhG) schützt grundsätzlich nur Werke im Sinne von § 2 UrhG.

Hierunter fallen die in § 2 Abs. 1 UrhG aufgeführten Werkarten wie

- Sprachwerke (Kap. 1),
- Werke der Musik (Kap. 2),
- Lichtbildwerke (Kap. 5),
- Filmwerke (Kap. 6),
- Darstellung wissenschaftlicher oder technischer Art, wie Zeichnungen, Pläne, Karten, Skizzen, Tabellen und plastische Darstellungen (Kap. 7),

bzw. auch die weitergehenden im UrhG aufgeführten Werke, wie Computerprogramme (§ 69a UrhG) oder Datenbankwerke (§ 87a UrhG).

Die aufgeführten Werkarten sind allerdings nicht abschließend geregelt. Das bedeutet, dass auch solche schöpferischen Leistungen, die nicht ausdrücklich als Werkart genannt sind, grundsätzlich Urheberrechtsschutz genießen können. Hierzu zählen im Prinzip auch einzelne Daten, die jedoch nur dann geschützt sind, wenn sie auch die übrigen Schutzvoraussetzungen erfüllen.

Eine Einschränkung des Urheberrechtsschutzes ergibt sich aus dem Werkbegriff. Geschützte Werke im Sinne des Urheberrechtsgesetzes sind nur „persönliche, geistige Schöpfungen“ (§ 2 Abs. 2 UrhG).

Die vorgenannten Werkarten sind insofern auch nur geschützt, soweit diese die notwendige Schöpfungshöhe erreichen. Die Schöpfungshöhe definiert das Maß kreativer Leistungen, dass eine geistige Schöpfung wie ein Text, eine Musikstück oder ein Computerprogramm aufweisen muss, um Urheberrechtschutz genießen zu können. Ob die Schöpfungshöhe erreicht wird, muss im Einzelfall beurteilt werden. Dies schließt es aus, im Rahmen dieses Gutachtens generelle Grundsätze für die verwendeten Daten und Informationen aufzustellen, an denen man sich in der Praxis leicht orientieren könnte.

Möglich ist lediglich einige Grundsätze zu verinnerlichen und sich darüber hinaus an Erfahrungswerten zu orientieren. Im Weiteren werden deshalb einige Hinweisen zu den einzelnen Voraussetzungen der Schutzfähigkeit von Daten und Informationen gegeben.

Wichtig für die Beurteilung der Schutzfähigkeit eines Datums ist primär die sogenannte Schöpfungshöhe oder Gestaltungshöhe. Mit diesen Schutzvoraussetzungen wird eine gewisse „Bagatellschwelle“ für den Urheberrechtsschutz gewährleistet.

Als allgemeiner Grundsatz gilt, dass die Schöpfungshöhe zum Beispiel für Texte generell niedrig angesiedelt wird. Bei den meisten Werkarten ist auch der „Schutz der kleinen Münze“ anerkannt, gemeint sind Schöpfungen von geringer Individualität bzw. Originalität. Nach der Rechtsprechung ist in der Regel nur das nicht geschützt, „was jeder so gemacht hätte“. Je größer der Gestaltungsspielraum bei der Erschaffung, desto größer ist die Wahrscheinlichkeit, dass das Ergebnis als Werk im Sinne des Urheberrechts geschützt ist. Einzelne Wörter sind daher niemals, kurze Sätze, Satzteile oder Überschriften in aller Regel nicht geschützt. Dies gilt auch für Einzeldaten, deren Urheberrechtschutz wegen ihres naturgemäß geringen Umfangs zumeist daran scheitert, dass sie die für den Schutz konkreter Formulierung erforderliche Schöpfungshöhe nicht erreichen.

Selbst soweit es sich bei den übernommenen Daten bereits um die Beurteilungen oder Bewertungen Einzelner handelt, ist eine urheberrechtliche Qualität fraglich. Geht es doch beim urheberrechtlichen Schutz nicht darum, inwieweit ein Leistungsergebnis auf erheblichem Aufwand z. B. geistig-wissenschaftlicher Art beruht, sondern ausschließlich um das Ergebnis selbst. Erschöpft sich aber das Ergebnis in einer bloßen Zahl oder einem Wert, findet die ggf. dahinter stehende Gedankenführung, die inhaltliche Verarbeitung und Auswahl der Erkenntnisse und eine evtl. Korrektur meist keine gestalterische Darstellung, sodass es auch hier am Werkcharakter fehlt.

Einzelne Daten sind insofern in aller Regel nicht urheberrechtlich geschützt. Sobald eine gewisse Zusammenstellung allerdings Schöpfungshöhe erreicht (z. B. entsprechend individuelle Texte), von einem Urheberschutz ausgegangen werden kann. Bei Fotos, Audio-

oder Videoaufnahmen wird allerdings in der Regel von Urheberrechtsschutz auszugehen sein.

3.2.1.3 Urheberrechtlicher Schutz der Einzeldaten

In der Regel stellt bereits die bloße Übernahme fremder Daten eine Vervielfältigungs- und Veröffentlichungshandlung dar, obwohl § 15 Abs. 1 UrhG dieses Recht zur Vervielfältigung, Verbreitung und Ausstellung von Werken ausschließlich dem Urheber zuweist. Wichtig ist daher zunächst, dass es sich bei den übernommenen Daten nicht um dem urheberrechtlichen Schutz unterliegende Werke gem. § 2 UrhG handelt.

Dies ist insbesondere dann nicht der Fall, wenn es sich ausschließlich um computergenerierte oder sonst maschinell erstellte Daten handelt. Bei diesen fehlt die schöpferische Leistung eines Einzelnen, die § 2 Abs. 2 UrhG zum urheberrechtlichen Schutz voraussetzt, völlig.

Selbst soweit es sich bei den übernommenen Daten bereits um die Beurteilungen oder Bewertungen Einzelner handelt, ist eine urheberrechtliche Qualität fraglich. Geht es doch beim urheberrechtlichen Schutz nicht darum, inwieweit ein Leistungsergebnis auf erheblichem Aufwand z. B. geistig-wissenschaftlicher Art beruht, sondern ausschließlich um das Ergebnis selbst.

Erschöpft sich aber das Ergebnis in einer bloßen Zahl oder einem Wert, findet die ggf. dahinter stehende Gedankenführung, die inhaltliche Verarbeitung und Auswahl der Erkenntnisse und eine evtl. Korrektur meist keine gestalterische Darstellung, sodass es auch hier am Werkcharakter fehlt.

3.2.1.4 Urheberrechtlicher Schutz von computergenerierten Werken

Derzeit ist jedenfalls nach deutschem Recht davon auszugehen, dass computergenerierten Werken mangels persönlicher geistiger Schöpfung urheberrechtlich nicht geschützt sind. In der Regel kann nur der vom Menschen geschaffene Inhalt in urheberrechtlichen Schutz erwachsen.

Weitergehende Ausführungen zu Rechtsfragen an computergenerierten Werken und der aus der mangelnden Schutzfähigkeit erwachsenden Notwendigkeit entsprechender vertraglicher Regelungen finden sich in Abschn. 3.5.3.

3.2.1.5 Urheberrechtlicher Schutz von Sammel- oder Datenbankwerken

Das Urheberrecht schützt in § 4 Abs. 1 UrhG aber auch Sammlungen von Werken, Daten oder anderen unabhängigen Elementen, die aufgrund der Auswahl oder Anordnung der Elemente eine persönliche geistige Schöpfung darstellen, als selbstständige Werke. Gleiches gilt nach § 4 Abs. 2 UrhG für Datenbankwerke, also Sammelwerke, deren Elemente systematisch oder methodisch angeordnet sind und einzeln mit Hilfe elektronischer Mittel zugänglich gemacht werden.

Auch § 4 UrhG setzt nicht zwingend voraus, dass es sich bei den Elementen, die in einer Sammlung oder Datenbank zusammengefasst sind oder werden, wiederum um eigenständige Werke handelt. Daher können auch Daten oder andere unabhängige Elemente, die

nach dem Vorgesagten keine eigenständigen Werke sind, Werkqualität dadurch erhalten, dass sie zu einer Sammlung oder Datenbank zusammengestellt werden. Voraussetzung ist jedoch stets, dass die Kombination der zusammengefassten Elemente eine besondere Struktur in der Auswahl oder Anordnung der Daten erkennen lässt und sich damit das Sammelwerk als solches als eine persönliche geistige Schöpfung darstellt.

Die Auswahl oder Anordnung der einzelnen Daten oder Elemente muss also durch die Individualität des Urhebers des Sammel- oder Datenbankwerkes gekennzeichnet sein. Ihre Gestaltung muss über die rein handwerkliche oder routinemäßige Leistung hinausgehen.

Gerade bei Ergebnislisten oder sonstigen Gegenüberstellungen (z. B. verschiedener Preise oder Angebote) ergibt sich die Auswahl der Informationen und Daten und deren Anordnung, ggf. in Rubriken in alphabetischer und chronologischer Reihenfolge, aber aus der Natur der Sache und ist durch Logik bzw. Zweckmäßigkeit vorgegeben. Wann immer dem so ist, fehlt es folglich an dem maßgeblichen Kriterium der Individualität und damit an einem schutzfähigen Sammel- bzw. Datenbankwerk.

3.2.2 Schutz des Datenbankherstellers

Mit dem Erstellen einer Datenbank sind oft erhebliche finanzielle Investitionen verbunden. Das Urheberrecht gewährt dem Hersteller einer Datenbank daher ein Schutzrecht eigener Art (sui generis), geregelt in den §§ 87 a–e UrhG.

Hier nach obliegt das ausschließliche Recht zur Vervielfältigung, Verbreitung und öffentlichen Wiedergabe der gesamten Datenbank, bzw. wesentlicher Teile davon, einzig und allein dem Hersteller. Der Vervielfältigung, Verbreitung oder öffentlichen Wiedergabe eines nach Art oder Umfang wesentlichen Teils der Datenbank steht nach § 87 b Abs. 1 S. 2 UrhG aber auch die wiederholte und systematische Vervielfältigung, Verbreitung oder öffentlichen Wiedergabe von nach Art und Umfang unwesentlichen Teil der Datenbank gleich, sofern diese Handlungen einer normalen Auswertung der Datenbank zuwiderlaufen oder die berechtigten Interessen des Datenbankherstellers unzumutbar beeinträchtigen.

Dieser Schutz besteht unabhängig von einem eventuellen urheberrechtlichen Schutz des Datenbankwerkes selbst und bezieht sich ausschließlich auf die Investitionen, die im Zusammenhang mit der Beschaffung und Sammlung, der Überprüfung und Aufbereitung, sowie der Darstellung und Veröffentlichung des Datenbankinhaltes angefallen sind.

Folglich ist nur das Datenbankwerk in seiner Gesamtheit geschützt, während die einzelnen Inhalte und die Datenbank selbst, zumindest nach den §§ 87 a–e UrhG, keinen Schutz genießen. Gleichwohl ist für das Datenbankwerk und seine Inhalte ein urheberrechtlicher Schutz nach den allgemeinen Regelungen (§ 2 Abs. 1; § 4 Abs. 2 UrhG) denkbar.

Der Gesetzgeber hat sich jedoch in Umsetzung in der Datenbankrichtlinie 96/9/EG bewusst dafür entschieden, den Schutz des Datenbankherstellers als verwandtes Schutzrecht auszustalten, mit der Folge, dass die Leistungsschutzrechte nach den §§ 87 a–e UrhG

selbst dann bestehen, wenn es sich bei der Datenbank um eine Sammlung als solcher eigenständig nicht schutzfähiger Informationen und Daten handelt.

3.2.2.1 Der Begriff der Datenbank

Eine Datenbank in diesem Sinne ist nach § 87 a Abs. 1 S. 1 UrhG eine Sammlung von Werken, Daten oder anderen unabhängigen Elementen, die systematisch oder methodisch angeordnet und einzeln mit Hilfe elektronischer Mittel oder auf andere Weise zugänglich sind und deren Beschaffung, Überprüfung oder Darstellung eine nach Art oder Umfang wesentliche Investition erfordert hat.

Ob es sich bei den Inhalten der Datenbank selbst um urheberrechtlich geschützte Werke und Leistungen handelt, ist – wie bereits aufgezeigt – gleichgültig. Maßgebend ist einzig und allein, ob durch die Beschaffung, Überprüfung oder Darstellung der Datenbankinhalte eine in qualitativer oder quantitativer Hinsicht wesentliche Investition erbracht wurde. Hierbei muss es sich nicht zwangsläufig um eine Investition finanzieller Art handeln. Der Investitionsbegriff ist weit zu verstehen und umfasst ebenfalls (alternativ oder kumulativ) den Einsatz von Arbeitszeit und Energie.

Die Investition ist jedoch nur insoweit geschützt, als sie zur Beschaffung des Inhalts der Datenbank erforderlich ist und daher der Ermittlung von vorhandenen Elementen und deren Zusammenstellung in einer Datenbank dient. Die finanziellen Mittel, die benötigt werden, um die Elemente, aus denen die Datenbank später bestehen soll, überhaupt erst zu erzeugen, sind nicht erfasst. Diese, durchaus erheblichen Investition, die im Vorfeld unternommen werden, um die Ursprungsdaten bzw. -Elemente überhaupt erst zu erzeugen, bleiben außen vor.

Entscheidend ist daher ausschließlich, ob auch noch nachdem die einzelnen Rohdaten tatsächlich vorliegen, durch die Zusammenstellung und Ordnung dieser Daten eine entscheidende Investition erbracht wird.

Berücksichtigungsfähig ist hierbei zudem, welcher personelle und damit finanzielle Aufwand auf die Pflege, Wartung und Aktualisierung der Datenbank entfällt, welche Kosten für die Bereitstellung der Server und Datenbankinfrastruktur, für die Bereitstellung der Datenaufbereitungs- und Abfragesysteme anfallen.

3.2.2.2 Der Begriff des Datenbankherstellers

Da sich der Schutz hiernach auf die Investitionen beschränkt, ist als Hersteller der Datenbank die (natürliche oder juristische Personen) anzusehen, die die mit der Beschaffung, Überprüfung oder Darstellung des Datenbankinhaltes einhergehende Investition vorgenommen hat, bzw. das damit zusammenhängende Investitionsrisiko trägt.

Hierbei muss es sich nicht zwangsläufig zugleich um denjenigen handeln, der die Ausgangsdaten ermittelt, oder die Datenbank konzipiert hat. Für den Erwerb des Datenbankrechts nach den §§ 87 a–e UrhG kommt es damit gerade nicht auf eine eigene schöpferische Leistung an, wie dies § 4 Abs. 1 UrhG beispielsweise für das Recht am Datenbankwerk voraussetzt.

Gerade in Wirtschaftsunternehmen wird daher vielfach das Recht an der Datenbank als solcher dem Mitarbeiter zustehen, der die Datenbank selbst und die dahinter stehende Infrastruktur entwickelt hat. Demgegenüber trägt ausschließlich der Arbeitgeber die mit der Erstellung der Datenbank verbundene Investition und ist daher als Datenbankhersteller im Sinne der §§ 87 a–e UrhG anzusehen. Um weitergehenden Probleme in diesem Bereich zu begegnen, sollte sich der Dienstherr und Datenbankhersteller daher etwaige Urheberrechte an der Datenbank selbst vertraglich einräumen lassen.

3.2.2.3 Die Rechte des Datenbankherstellers

Dem Datenbankhersteller weist nun § 87 b Abs. 1 UrhG das alleinige Recht zur Vervielfältigung, Verbreitung und öffentlichen Wiedergabe der Datenbanken insgesamt oder wesentlicher Teile zu.

Aber selbst wenn nur unwesentliche Teil der Datenbank vervielfältigt, verbreitet oder öffentlich wiedergegeben werden, kann eine Verletzung der Datenbankschutzrechte des Herstellers vorliegen. Nach § 87 b Abs. 1 S. 2 UrhG steht nämlich der urheberrechtswidrigen Nutzung eines wesentlichen Teils der Datenbank die wiederholte und systematische Vervielfältigung, Verbreitung oder öffentliche Wiedergabe von nach Art und Umfang unwesentlichen Teilen der Datenbank gleich, sofern diese Handlungen einer normalen Auswertung der Datenbank zuwiderlaufen oder die berechtigten Interessen des Datenbankherstellers unzumutbar beeinträchtigen.

Für die Voraussetzungen, unter denen Teile einer Datenbank entnommen und weiterverwendet werden dürfen, kommt es damit entscheidend auf die Frage an, wann ein wesentlicher, wann aber nur ein unwesentlicher Teil der Datenbank vorliegt.

Die ausdrückliche Beschränkung des gesetzlichen Schutzes auf wesentliche Teile sollte einer weitreichenden Monopolisierung von Datenbankinhalten entgegenwirken, die Verwendung von Rohmaterial aus einer Datenbank für die Herstellung von Konkurrenzprodukten in Sekundärmarkten ermöglichen und der durch Art. 10 EMRK geschützten Informationsfreiheit Rechnung tragen. Daher sind im Interesse eines angemessenen Ausgleichs von Investitionsschutz und Informationsfreiheit an die Wesentlichkeit keine allzu geringen Anforderungen zu stellen (vgl. Dreier und Schulze 2013, § 87 b Rn. 5).

Es ist daher maßgeblich darauf abzustellen, wann mit der Entnahme des betroffenen Teils der Datenbank ein qualitativ oder quantitativ erheblicher Schaden für die Investitionen einhergeht, sich die Investition gleichsam in dem entnommenen Teil wiederspiegelt. Hier ist immer eine Entscheidung im Einzelfall und unter Berücksichtigung des Investitionsgedankens erforderlich.

Einzelne Datensätze sind daher jedenfalls für sich gesehen keine wesentlichen Teile einer Datenbank. Ihre Vervielfältigung, Verbreitung und öffentliche Wiedergabe ist daher grundsätzlich ohne Zustimmung des Datenbankherstellers zulässig, sofern eine Verwertung nur im Einzelfall stattfindet und sich diese im Rahmen der normalen Auswertung der Datenbank hält.

Es muss also sichergestellt werden, dass viele, für sich gesehen unwesentlichen Teile nicht derart intensiv genutzt werden, dass diese Nutzung in ihrer Gesamtheit der Entnahme eines wesentlichen Teils der Datenbank gleichsteht.

3.2.2.4 Schranken des Rechts des Datenbankherstellers

Daneben ist nach § 87 c Abs. 1 UrhG die Vervielfältigung auch eines wesentlichen Teils einer Datenbank immer dann zulässig, wenn die Vervielfältigung ausschließlich zum privaten Gebrauch, zu eigenen wissenschaftlichen Zwecken, oder zur Veranschaulichung des Unterrichts erfolgt. Eine Weiterverbreitung oder öffentliche Wiedergabe bleibt aber auch in diesen Fällen unzulässig.

Screen-Scraping

Im Zusammenhang mit der zwischenzeitlich weit verbreiteten Veröffentlichung von Datenbanken im Internet und den rein technischen Möglichkeiten, diese Daten automatisiert auszulesen und für eigene Geschäftszwecke weiterzuverwenden (sog. Screen-Scraping), stellt sich vermehrt die Frage nach der rechtlichen Zulässigkeit solcher Geschäftsmodelle, insbesondere in urheberrechtlicher, aber auch in wettbewerbsrechtlicher Hinsicht.

Bekannt sind derartige Geschäftsmodelle insbesondere im Bereich der Flugreisevermittlung. Betreiber von Internetreiseportalen bedienen sich hierbei der im Internet veröffentlichten Originaldaten verschiedener Fluglinien, lesen deren Onlineangebote maschinell aus und betten die so erhaltenen Daten in die eigenen Seiten ein.

Mit der passenden Technologie sind jedoch den Anwendungsszenarien des Screen-Scrapings keine Grenzen gesetzt, so dass auch noch viele weitere Geschäftsmodelle zur Verwertung von Big Data denkbar sind.

Obwohl der BGH im Falle der Flugvermittlung im Internet (vgl. Urteil des BGH vom 30. April 2014, Az.: I ZR 224/12) von der Zulässigkeit dieses Geschäftsmodells ausgegangen war, lässt sich diese Aussage nicht pauschal auf alle denkbaren Screen-Scraping-Formen übertragen.

Entscheidend ist ausschließlich,

- dass die entnommenen Einzeldaten nicht selbst urheberrechtlich geschützt sind,
- kein eigenständig schutzwürdiges Datenbankwerk betroffen ist,
- keine wesentlichen Teile der Datenbank extrahiert werden,
- sich die Entnahme unwesentlicher Teile im Rahmen der normalen Auswertung einer Datenbank hält und diese nicht systematisch und wiederholt erfolgt,
- keine Nachahmung des ursprünglichen Geschäftsmodells vorliegt und
- keine technischen Schutzvorkehrungen des Ursprungsanbieters überwunden werden.

Entnahme wesentlicher Teile einer Datenbank

Ob schließlich eine nach §§ 87 a ff. UrhG geschützte Datenbank vorliegt, deren Verwertung ausschließlich dem Hersteller obliegen würde, beurteilt sich – wie bereits oben

ausgeführt – ausschließlich danach, ob eine schützenswerte Investition des ursprünglichen Anbieters anzunehmen ist.

Selbst wenn hiernach eine Datenbank vorliegt, ist die Vervielfältigung, Verbreitung und öffentliche Wiedergabe von Teilen einer Datenbank jedoch so lange ohne Zustimmung des Datenbankherstellers zulässig, wie es sich hierbei nur um einen unwesentlichen Teil handelt und die Verwertung einer normalen Auswertung der Datenbank entspricht.

Teilweise wird hier vertreten, dass der Anbieter seinen Nutzern mit Hilfe der Technik des Screen-Scrapings die komplette Datenbank permanent öffentlich verfügbar mache, da die Nutzer ständig auf alle Datensätze der betreffenden Datenbank zugreifen und sich die für sie interessanten Datensätze mindestens online übermitteln lassen könnten. Dabei komme es auch nicht darauf an, wie viele Datensätze den Nutzern aufgrund entsprechender Suchanfragen tatsächlich übermittelt würden. Entscheidend sei nur, dass sich der Screen Scraper mittels des Screen-Scrapings die betreffende Datenbank derart zu eigen mache, dass er sie für den Nutzer öffentlich verfügbar halte. Hiernach würden wesentliche Teile der Datenbank zumindest nach § 87 b Abs. 1 UrhG öffentlich wiedergegeben (vgl. Kahler und Helbig 2012, S. 51 f.).

Die wohl herrschende Ansicht vertritt jedoch die Ansicht, dass einzelne Datensätze, die seitens des Screen Scrapers ausgelesen und auf der eigenen Webseite wiedergegeben werden, nicht als wesentlicher Teil einer Datenbank in diesem Sinne angesehen werden können.

Entnahme unwesentlicher Teile einer Datenbank

Aber selbst wenn nur die Entnahme eines unwesentlichen Teils der Datenbank vorliegt, kommt eine urheberrechtswidrige Nutzung dann in Betracht, wenn diese Nutzung einer normalen Auswertung der Datenbank zuwiderläuft oder die berechtigten Interessen des Datenbankherstellers unzumutbar beeinträchtigt.

Auch hier wird vertreten, der Screen Scraper benutze die Datenbank gerade nicht, um sich selbst Informationen über die in der Datenbank bereit gehaltenen Inhalte zu verschaffen, sondern für eigene kommerzielle Zwecke. Schon dies widerspräche der normalen Auswertung einer Datenbank (vgl. Kahler und Helbig 2012, S. 53).

Dies gelte insbesondere, wenn ein Konkurrenzprodukt aufgebaut werde, welches die Auswertung der Datenbank beeinträchtigen könnte oder wodurch sich der Screen Scraper den Abschluss eines Lizenzvertrages erspare.

Doch selbst wenn die Verwertung nicht für ein direktes Konkurrenzprodukt erfolge, sondern bloß für ein Mehrwertprodukt, das einen anderen Markt bediene, könne eine unzumutbare Beeinträchtigung der berechtigten Interessen des Ursprungsanbieters vorliegen. Hier sei noch stärker auf das Amortisationsinteresse abzustellen, mit der Folge, dass auch neuartige Auswertungsmöglichkeiten in erster Linie dem Datenbankhersteller obliegen. So soll Unzumutbarkeit zumindest dann anzunehmen sein, wenn nicht nur der Gewinn des Ursprungsanbieters verringert, sondern dessen Amortisation ernsthaft gefährdet werde (Wiebe 2013, § 87 a UrhG, Rn. 33).

Überwiegend wird jedoch schlicht darauf abgestellt, ob die Informationen frei im Internet zugänglich sind und potentiell von jedem in dieser Art und Weise genutzt werden können.

Das reine „Abfragen“ stellt nämlich dann eine normale Auswertung der Datenbank und damit keine Verletzungshandlung dar, wenn die Datenbank erst einmal und ohne Einschränkungen öffentlich zugänglich gemacht wurde.

Solange dies der Fall ist, wir die Amortisation der Investition des Datenbankherstellers nicht beeinträchtigt.

Unlautere Nachahmung

Ein Wettbewerbsverstoß in der Gestalt einer Nachahmung nach § 4 Nr. 9 setzt voraus, dass eine fremde Leistung ganz oder teilweise als eigene Leistung angeboten wird. Solange dies nicht der Fall ist, etwa weil auf den fremden Ursprung der Daten hingewiesen wird, scheidet eine Nachahmung aus.

Vielfach werden im Rahmen von Big Data-Projekten auch große Mengen von Daten aus unterschiedlichsten Quellen ausgelesen und verwertet und hierdurch ein neuer Informationsgehalt geschaffen, der alleine nach außen sichtbar ist. Die einzelnen Daten, bzw. deren Herkunft haben dann nach außen hin keine eigenständige Bedeutung, so dass es auch dann an einer wettbewerbsrechtlich relevanten Verletzungshandlung fehlt (vgl. Ziegler und Smirra 2013, S. 421).

3.2.3 Unlautere gezielte Mitbewerberbehinderung

Eine gezielte Behinderung nach § 4 Nr. 10 UWG liegt hingegen vor, wenn die wettbewerblichen Entfaltungsmöglichkeiten des Ursprungsanbieters als Mitbewerber in Bezug auf Absatz, Bezug, Werbung, Produktion, Forschung, Entwicklung, Planung, Finanzierung, Personaleinsatz usw. beeinträchtigt werden. Das Screen-Scraping müsste hierzu also nicht in erster Linie auf die Förderung der eigenen wettbewerblichen Entfaltung, sondern auf die Störung der fremden wettbewerblichen Tätigkeit gerichtet sein.

In diesem Zusammenhang wäre es denkbar, dass es zu gewissen Störungen beim Zugriff auf die Website kommt, wenn durch das automatisierte Abrufen der Daten die Servierfähigkeit spürbar verlangsamt wird.

Ob jedoch derartige Umstände die hohen Anforderungen an eine gezielte Behinderung erfüllen, ist zweifelhaft. Immerhin begibt sich der Betreiber einer Internetseite durch das Zurverfügungstellen der Informationen im Internet zwangsläufig in die Situation, dass Dritte – auch gehäuft – auf die Daten zugreifen können.

Vor allem wird aber gerade nicht gezielt der Zweck verfolgt, die wettbewerbliche Entfaltung des Ursprungsanbieters zu stören. Vielmehr baut das Angebot des Screen Scrapers gerade auf diesem Angebot und damit seiner Funktionsfähigkeit auf (vgl. BGH, Urteil vom 30.04.2014, Az.: I ZR 224/12, Rn. 25.)

Damit kommt eine wettbewerbsrechtlich unzulässige gezielte Mitbewerberbehinderung nur noch dann in Betracht, wenn der Ursprungsanbieter durch das Screen-Scraping daran gehindert wird, seine eigene Leistung am Markt durch eigene Anstrengungen zur Geltung zu bringen. Erforderlich ist insoweit eine Beeinträchtigung der wettbewerblichen Entfaltungsmöglichkeit, die über die mit jedem Wettbewerb verbundene Konkurrenzsituation hinausgeht.

Hierbei führt allein der Umstand, dass sich der Screen Scraper über etwaige Allgemeine Geschäfts- oder Nutzungsbedingungen des Ursprungsanbieters, wonach dieser kein Auslesen der eigenen Webseite mittels Screen-Scraping-Werkzeugen zuzulassen will, hinwegsetzt, nicht zu einer wettbewerbsrechtlichen Unzulässigkeit.

Solange derartige Bedingungen einseitig vom Anbieter vorgegeben werden und nicht im Rahmen eines Anmelde- oder Login-Verfahrens durch Anklicken ausdrücklich akzeptiert werden müssen, handelt es sich hierbei nur um ein einseitiges Angebot. Ohne eine ausdrückliche oder zumindest stillschweigende Annahmeerklärung des Screen Scrapers werden solche Bestimmungen nicht wirksam zum Vertragsinhalt.

Anders ist dies aber, wenn technische Schutzvorkehrungen, mittels derer der Ursprungsanbieter ein Auslesen seiner Angebote verhindern will, gezielt umgangen werden.

Selbst wenn nun aber die Allgemeinen Geschäfts- oder Nutzungsbedingungen des Anbieters eine Erklärung beinhalten sollten, wonach ein Auslesen der Seite mittels Screen Scaping-Tools verboten ist und jeder Nutzer diese Bedingungen durch Setzen eines Häckchens aktiv bestätigen muss, steht dies einer technischen Schutzmaßnahme nicht gleich.

Mittels einer solchen Checkbox (Ankreuzkästchen), die jeder Nutzer von Bestellvorgängen im Internet kennt, will der Anbieter allem voran sicherstellen, dass seine Allgemeinen Geschäfts- oder Nutzungsbedingungen in den zu schließenden Vertrag mit einbezogen werden. Diese primär vertragsrechtliche Maßnahme kann einer Begrenzung der Nutzung der Internetseite durch technische Maßnahmen gegen eine automatisierte Abfrage nicht gleichgesetzt werden. Dies gilt selbst dann, wenn in den zu bestätigenden Bedingungen Screen-Scraping-Anwendungen verboten sind (vgl. BGH, Urteil vom 30.04.2014, Az.: I ZR 224/12, Rn. 38).

Erforderlich sind echte technische Hürden, die ein automatisches Auslesen der Daten unterbinden sollen.

Screen-Scraping-Modelle sind daher auch nach der Entscheidung des BGH nicht ohne Risiko. Es werden hierbei die verschiedensten rechtlichen Problemkreise tangiert, auch wenn diese der Zulässigkeit dieser Anwendungen im Ergebnis nicht zwingend im Wege stehen werden. Doch selbst wenn nach dem Vorgesagten der Zugriff auf fremde Daten in vielen Fällen rechtlich unbedenklich möglich ist, sollten sich Anwender stets in einem frühen Stadium Gewissheit über die rechtlichen Voraussetzungen verschaffen, um das Tool rechtskonform ausgestalten zu können.

Maßgebend ist daher die Anzahl der Abrufe und die Art der abgerufenen Informationen. Zudem muss gewährleistet sein, dass keine wesentlichen Teile der Datenbank entnommen werden und die wiederholte und systematische Verwertung unwesentlicher

Teile sich im normalen Rahmen der Nutzung bewegt und hierbei insbesondere keine technischen Schutzvorrichtungen überwunden werden.

Kann dies sichergestellt werden, ist der Einsatz von Screen Scaping-Tools und damit diese wirtschaftliche Nutzung von Big Data rechtlich in weitreichendem Rahmen möglich.

3.2.4 Sonstige Leistungsschutzrechte

Der zuvor dargestellte Schutz des Datenbankherstellers in den §§ 87 a–e UrhG stellt im Bereich Big Data sicherlich das bedeutendste Leistungsschutzrecht dar.

Für Leistungsschutzrechte (oder auch „verwandte Schutzrechte“) ist – wie auch bereits oben gesehen – allgemein kennzeichnend, dass der Schutz nicht an einer persönlichen geistigen Schöpfung anknüpft – wie dies sonst zum urheberrechtlichen Schutz nach § 2 Abs. 2 UrhG erforderlich ist – sondern an einer Leistung anderer Art, die jedoch der schöpferischen Leistung des Urhebers ähnlich ist oder im Zusammenhang mit urheberrechtlich geschützten Werken erbracht wird.

Im Einzelnen werden unter diesem Begriff die folgenden Schutzrechte zusammengefasst:

- Schutz wissenschaftlicher Ausgaben, § 70 UrhG,
- Schutz nachgelassener Werke, § 71 UrhG,
- Schutz der Lichtbildner, § 72 UrhG,
- Schutz des ausübenden Künstlers, §§ 73 ff. UrhG,
- Schutz des Veranstalters, § 81 UrhG,
- Schutz des Tonträgerherstellers, § 85 UrhG,
- Schutz des Sendeunternehmens, § 87 UrhG,
- Schutz des Presseverlegers, §§ 87 f. UrhG,
- Schutz des Filmherstellers, § 94 UrhG,
- Schutz der Laufbilder, § 95 UrhG.

Anknüpfungspunkt ist hier folglich der Prozess der Leistungserbringung, geschützt wird der damit einhergehende Aufwand. Das Resultat kann eigenständig geschützt sein, sofern ihm Werkqualität nach § 2 UrhG zukommt, der jeweilige Leistungsschutz besteht aber unabhängig davon.

Während das Urheberrecht – als Ausprägung einer persönlichen geistigen Schöpfung – zwangsläufig auch nur einer natürlichen Person zukommen kann, richten sich diese verwandten Schutzrechte regelmäßig auch an juristische Personen, sprich Unternehmen.

3.2.4.1 Schutz des Presseverlegers

Die Vorschriften zum Schutz des Presseverlegers wurden am 01.08.2013 neu in das Urheberrechtsgesetz eingefügt und sind ebenfalls als verwandtes Schutzrecht ausgestaltet. Der

Schutz bezieht sich also nicht primär auf das Presseerzeugnis selbst, sondern kommt dem Hersteller eines Presseerzeugnisses zu gute.

Ursprünglich sollte hierdurch erreicht werden, dass ausschließlich der Presseverleger das Recht hat, bereits kleinste Ausschnitte eines Presseerzeugnisses (sog. Snippets) öffentlich zugänglich zu machen, sodass diese insbesondere von Suchmaschinen in deren Trefferlisten nicht länger unentgeltlich angezeigt werden dürfen.

Verabschiedet wurde jedoch eine Gesetzesfassung, die die Veröffentlichung einzelner Wörter oder kleiner Textausschnitte eines Presseerzeugnisses auch weiterhin ohne Einschränkungen oder eine Entgeltpflicht zulässt.

Die weitere Verwertung dieser Big-Data-Quelle bleibt – zwar im gesetzlich vorgesehenen Rahmen – aber damit immerhin auch weiterhin möglich.

3.3 Integritätsschutz

Joachim Dorschel und Michael Bartsch

Während es im vorangegangen Abschn. 3.2. um den Schutz von Daten vor unzulässiger Ausbeutung durch Dritte ging, also Eingriffe in Daten als geistiges Eigentum, beschäftigt sich das folgende Kapitel mit dem Schutz der Datenintegrität, also dem gewollten oder ungewollten manipulativen Eingreifen in Datenbestände durch Dritte. Solche Eingriffe können durch Datenlöschung, Datenveränderung oder Datenunterdrückung erfolgen.

Das deutsche Recht kennt strafrechtliche und zivilrechtliche Schutzinstrumente. Dies ist zunächst kein Spezifikum des Integritätsschutzes. Auch die vorsätzliche Verletzung der Rechte am geistigen Eigentum ist strafbar. Anders als im Bereich des geistigen Eigentums, in dem die Strafbarkeit nur eine weitere Rechtsfolge einer Schutzrechtsverletzung ist, kennt das Strafgesetzbuch spezifisch auf den Schutz der Integrität von Daten gerichtete Straftatbestände (hierzu näher im Folgenden in Abschn. 3.3.1.) während sich der zivilrechtliche Schutz, soweit er sich nicht aus dem strafrechtlichen Schutz ableitet, im Wesentlichen auf das allgemeine Deliktsrecht stützt (hierzu im Folgenden in Abschn. 3.3.2.).

3.3.1 Strafrechtlicher Schutz der Datenintegrität

Das Strafrecht schützt über die Tatbestände der Sachbeschädigung (§ 303 Strafgesetzbuch, StGB), der Datenveränderung (§ 303 a StGB) und der Computersabotage (§ 303 b StGB) sowohl Datenträger und IT-Systeme als auch die Daten selbst vor manipulativen Eingriffen. Die Tatbestände des Ausspähens von Daten (§ 202 a StGB), des Abfangens von Daten (§ 202 b StGB) und hierauf gerichteter Vorbereitungshandlungen (§ 202 c StGB) sind auf den Schutz von Daten vor unberechtigter Kenntnisnahme gerichtet. Diese Tatbestände unterstützen damit sowohl den Schutz personenbezogener Daten und den Schutz geistigen

Eigentums vor Spionage als auch die Abwehr unbefugten Eindringens in IT-Systeme und mithin die Verteidigung der Integrität der dort verarbeiteten Daten.

Selbstverständlich gelten neben den genannten Spezialnormen auch die allgemeinen Bestimmungen. So kann Weitergabe von Daten strafbar sein, wenn diese fremde Geheimnisse enthalten. Gleches gilt, wenn Daten zum Zwecken des Betrugs, der Urkundenfälschung oder vergleichbaren Straftaten genutzt werden.

3.3.1.1 Sachbeschädigung (§ 303 StGB)

§ 303 StGB stellt die Beschädigung oder Zerstörung fremder Sachen unter Strafe. Unter Sachen sind nur körperliche Gegenstände zu verstehen (Weidemann 2014, § 303 Rn. 4). Tatobjekt können daher IT-Systeme und Speichermedien sein, nicht jedoch Daten selbst. Erforderlich ist ein physischer Eingriff in die Sachsubstanz. Die Tatsache, dass Daten in Form elektromagnetischer Signale auf IT-Systemen verarbeitet und gespeichert werden, führt nicht dazu, dass eine Veränderung dieser Daten zu einem Eingriff in das IT-System im Sinne des Sachbeschädigungs-Tatbestandes wird. Rechtstheoretisch sind hier schwierige Abgrenzungsfragen denkbar. Da die Daten selbst über § 303 a StGB (hierzu sogleich) umfassend geschützt sind, ist die praktische Bedeutung jedoch gering.

Typische Fälle der datenbezogenen Sachbeschädigung sind die Vernichtung von Datenträgern, das Einwirken auf IT-Systeme durch Hitze, Kälte, Feuchtigkeit oder rohe Gewalt sowie die Aktenvernichtung von Computer-Ausdrucken.

Gegenstand der Sachbeschädigung können nur fremde Sachen sein. Die beschädigte Sache darf also nicht im Alleineigentum des Täters stehen und auch nicht herrenlos sein. Die Vernichtung eines eigenen Datenträgers, etwa zum Schutz vor strafrechtlicher Verfolgung, erfüllt nicht den Tatbestand der Sachbeschädigung, kann aber nach anderen Vorschriften, etwa unter dem Gesichtspunkt der Strafvereitelung (§ 258 StGB), strafbar sein.

Strafbar ist nur die vorsätzliche Sachbeschädigung. Der Täter muss die Beschädigung oder Zerstörung also beabsichtigen, sicher vorhersehen oder zumindest billigend in Kauf nehmen. Die fahrlässige Sachbeschädigung, also etwa die versehentliche Vernichtung eines Datenträgers, ist strafrechtlich nicht sanktioniert, kann aber gleichwohl auf zivilrechtlicher Grundlage Schadensersatzansprüche nach sich ziehen (hierzu Abschn. 3.3.2.4).

3.3.1.2 § 303 a Datenveränderung

Gemäß § 303 a StGB wird bestraft, wer rechtswidrig Daten löscht, unterdrückt, unbrauchbar macht oder verändert. Der Begriff der Daten ist in § 202 a Absatz 2 definiert als Informationen, die elektronisch, magnetisch oder sonst nicht unmittelbar wahrnehmbar gespeichert sind oder übermittelt werden. Der Tatbestand der Datenveränderung ergänzt jenen in der Sachbeschädigung: strafbar sind auch solche computerbezogenen Einwirkungen, die nicht zu einem Subtanzeingriff in die Hardware führen, deren Folge gleichwohl der Verlust oder die Beeinträchtigung der Verwendbarkeit von Daten ist (Wieck-Noodt 2014, § 303 a Rn. 1).

Der Tatbestand der Datenveränderung schützt das Recht des Berechtigten, „seine“ Daten bestimmungsgemäß zu nutzen. Es geht also um die Verwendbarkeit der Daten durch den Berechtigten (Hoeren und Völkel 2014, S. 24).

Die Berechtigung an den Daten kann sich aus verschiedenen Rechtgründen ergeben, wobei Berechtigung und Eigentum am Datenträger auseinanderfallen können (hierzu Hoeven und Völkel 2014, S. 26. ff.). Speichert der Eigentümer des Datenträgers dort von ihm selbst generierte Daten, ist er zweifelsfrei Berechtigter. Erfolgt die Speicherung auf einem Server im Rechenzentrum eines Dritten, ist derjenige, der die Daten generiert hat, berechtigt und nicht der Betreiber des Rechenzentrums (Wieck-Noodt 2014, § 303a Rn. 10). Ein Rechenzentrumsbetreiber, der, z. B. im Zuge einer Vergütungsstreitigkeit vorsätzlich Datenbestände seines Kunden löscht, kann sich nach § 303a StGB strafbar machen. Schwierig kann die Bestimmung des Berechtigten sein, wenn Datenbestände im Auftrag generiert und gespeichert werden.

Nicht entscheidend für die Berechtigung an einem Datenbestand ist, wen die Daten inhaltlich betreffen. Dies ist ein wesentlicher Unterschied zu den Bestimmungen des Datenschutzrechts, die stets auf den Betroffenen der Daten abstellen. Für die Anwendung dieses § 303a StGB ist auch nicht entscheidend, ob die Daten einen wirtschaftlichen Wert haben.

Tatobjekt einer strafbaren Datenveränderung können nur Daten sein, an denen ein *anderer* eigene Rechte hat. Wie bei der Sachbeschädigung ist der Tatbestand also nicht erfüllt, wenn der Täter eigene Daten zum Zwecke der Beweisvernichtung löscht oder manipuliert (vergleiche oben Abschn. 3.3.1.1). Der Begriff der Daten umfasst digitale Informationen jedweder Art, also auch Programme die als Source Code oder als Object Code vorliegen.

Der Tatbestand der Datenveränderung kann auch durch Unterlassen erfüllt werden. Dies setzt juristisch eine sogenannte Garantenpflicht voraus, also eine rechtlich relevante Verpflichtung des Täters, den Taterfolg zu verhindern. Diese Variante ist vor allem bei Outsourcing-Sachverhalten relevant, wenn der Outsourcer es unterlässt, einen vorhersehbaren Angriff auf Datenbestände seines Kunden zu unterbinden.

Anders als bei der Sachbeschädigung, die nur die Zerstörung oder Beschädigung eines Gegenstandes unter Strafe stellt, ist bei der Datenveränderung auch die Datenunterdrückung strafbar, also Handlungen, die nicht unmittelbar auf die Daten einwirken, gleichwohl aber dazu führen, dass der Berechtigte Daten nicht wie vorgesehen nutzen kann. Das sogenannte Daten-Kidnapping, also das Einschleusen von Schadprogrammen, die Daten verschlüsseln und nur gegen Zahlung eines Lösegelds wieder freigeben, erfüllt den Tatbestand der Datenveränderung daher ohne Weiteres.

Der Tatbestand der Datenveränderung ist auch erfüllt durch Veränderungen von Datenbeständen, die keinen negativen, einer Beschädigung gleichkommenden Eingriff, sondern lediglich eine Bedeutungsveränderung darstellen. Entscheidend ist allein, dass der Täter den Informationsgehalt oder Aussagewert der Daten verändert (Wieck-Noodt 2014, § 303a Rn. 15).

Wie bei der Sachbeschädigung sind nur vorsätzliche Datenveränderungen strafbar. Versehenhafte Eingriffe in Datenbestände erfüllen den Tatbestand nicht.

Das Einverständnis des Berechtigten schließt eine Strafbarkeit wegen Datenveränderung aus. Ein Dienstleister, der im Auftrag seines Kunden an Datenbeständen dieses Kunden arbeitet, begeht selbstverständlich keine strafbare Datenveränderung. Dies gilt auch dann, wenn der Dienstleister aus Fahrlässigkeit Fehler macht. Strafbarkeit wäre aber dann anzunehmen, wenn der Dienstleister vorsätzlich den ihm zugewiesenen Arbeitsbereich überschreitet, etwa indem er die ihm gegebenen Zugriffsmöglichkeiten nutzt, um Datenbestände zu manipulieren, die ihm nicht zugewiesen sind.

Strafbar ist bereits der Versuch der Datenveränderung. Beim Einschleusen von Malware ist die Schwelle zur Strafbarkeit somit bereits mit der Speicherung der schädlichen Programme auf einem fremden IT-System überschritten, also nicht erst dann, wenn diese Programme auf fremde Datenbestände einwirken. Auch Vorbereitungshandlungen, also etwa das Herstellen und Verbreiten von Malware mit dem Ziel, Datenveränderungen zu begehen, ist gemäß § 303 a Abs. 3 strafbar.

3.3.1.3 Computersabotage (§ 303 b)

§ 303 b StGB enthält verschiedene Qualifikationstatbestände. Die Beschädigung von IT-Systemen und Datenveränderungen mit besonderem Schadenspotenzial werden härter bestraft. Praktisch bedeutsam ist diese Regelung insbesondere bei gezielten Angriffen auf die IT-Systeme von Unternehmen mit dem Ziel, Schaden anzurichten, sowie bei der Verbreitung von Malware zum Zweck der wirtschaftlichen Bereicherung.

3.3.1.4 § 202 a Ausspähen von Daten

Zentrale Norm des Schutzes digitaler Daten vor unberechtigtem Zugriff ist § 202 a StGB, der landläufig auch als „Hacking-Paragraf“ bezeichnet wird. Bestraft wird hiernach, „*wer unbefugt sich oder einem anderen Zugang zu Daten, die nicht für ihn bestimmt und die gegen unberechtigten Zugang besonders gesichert sind, unter Überwindung der Zugangssicherung verschafft*“. Daten im Sinne des Gesetzes sind nach der gesetzlichen Definition solche, „*die elektronisch, magnetisch oder sonst nicht unmittelbar wahrnehmbar gespeichert sind oder übermittelt werden*.“

Anders als bei den Tatbeständen der Sachbeschädigung und Datenveränderung richtet sich der Schutz bei der Ausspähung von Daten nicht nur gegen manipulative Eingriffe, also Angriffe auf die Datenintegrität, sondern gegen jedweden Zugriff auf den Datenbestand, auch wenn dieser im Sinne einer Datenspionage nur der Kenntnisnahme dient. Insoweit ähnelt der Schutzbereich den Tatbeständen des Datenschutzrechts, wobei anders als dort eine Strafbarkeit auch gegeben ist, wenn es sich nicht um personenbezogene Daten geht. Angriffsobjekt einer Datenausspähung können daher auch die von einer industriellen Produktionsanlage erzeugten Messdaten sein, die unter Verwendung von Big Data-Technologien realtime überwacht werden.

Wie bei den Datenveränderungstatbeständen muss ich auch das Ausspähen von Daten gegen fremde Datenbestände richten. Das Gesetz spricht von solchen Daten, die nicht für

den Täter „bestimmt“ sind. Während es aber bei der Datenveränderung darauf ankommt, ob der Täter eigene Rechte an den Daten hat (vgl. oben Abschn. 3.3.1.2) ist beim Ausspähen von Daten die Verfügungsmacht über die Daten entscheidend. Ausspähen kann man mithin nur solche Daten, über die ein Anderer Verfügungsmacht besitzt. Verfügungsmacht bedeutet die faktische Herrschaft über die Daten, die bei der Datenerhebung mit der Speicherung, bei der Übermittlung mit deren Empfang entsteht (Weidemann 2014, § 202a Rn. 8).

Wer die Verfügungsmacht über die Daten innehat, kann sich auch dann nicht wegen Ausspähens von Daten strafbar machen, wenn er die Daten vertrags- oder datenschutzrechtswidrig verwendet (eine Strafbarkeit wegen Verletzung der Datenschutzgesetze ist hiervon freilich unbenommen, vgl. hierzu Abschn. 3.1).

Voraussetzung der Strafbarkeit ist weiterhin, dass die Daten gegen unberechtigten Zugang besonders gesichert sind. Dies setzt Vorkehrungen voraus, den Zugriff auf die Daten auszuschließen oder wenigstens nicht unerheblich zu erschweren (vgl. BGH, Beschluss vom 06.07.2010 – 4 StR 555/09). Solche Vorkehrungen können physikalischer (z. B. die Aufbewahrung in einem verschlossenen Tresor) wie digitaler (Passwörter, Verschlüsselung etc.) Natur sein. Die Effektivität der Zugangssicherung ist nicht entscheidend, sondern der Wille des Berechtigten, die Daten vor unbefugter Kenntnisnahme zu bewahren. Dass also bestimmte marktübliche Sicherungsmechanismen von technisch versierten Angreifern leicht umgangen werden können, ändert an der Strafbarkeit nichts.

Der Datenzugriff muss unbefugt sein. Wo sich also ein Unternehmen im Rahmen der gesetzlichen und vertraglichen Bestimmungen Zugriff auf Datenbestände verschafft, die ein Mitarbeiter durch Passwort gesichert hat, ist eine Strafbarkeit ausgeschlossen. In der Praxis tut sich hier freilich eine Grauzone auf, da die Frage, unter welchen Voraussetzungen und in welchem Umfang Unternehmen auf Datenbestände ihrer Mitarbeiter zugreifen dürfen, arbeitsrechtlich schwierig zu beantworten ist. Dies gilt insbesondere dann, wenn die Datenbestände auch private Daten der Mitarbeiter enthalten (vgl. hierzu näher Abschn. 3.1.8).

Sind die hier genannten Voraussetzungen erfüllt, ist jede Handlung strafbar, mit der sich der Täter Zugang zu den Daten verschafft. Ausreichend ist die bloße Kenntnisnahme der Daten. Es ist also nicht erforderlich, dass der Täter der Daten kopiert.

3.3.1.5 § 202 b Abfangen von Daten

§ 202b StGB ergänzt den strafrechtlichen Datenschutz, indem er das Abfangen von Daten während eines Übermittlungsvorgangs auch dann unter Strafe stellt, wenn keine besonderen Sicherungsmaßnahmen getroffen wurden. Strafbar ist hiernach, „*wer unbefugt sich oder einem anderen unter Anwendung von technischen Mitteln nicht für ihn bestimmte Daten ... aus einer nichtöffentlichen Datenübermittlung oder aus der elektromagnetischen Abstrahlung einer Datenverarbeitungsanlage verschafft.*“

Der Begriff der Daten ist identisch mit dem der Tatbestände der Datenveränderung und Datenausspähnung. Tatobjekt können aber nur Daten sein, die sich zur Zeit des Zugriffs in einem Übermittlungsvorgang befinden. Ist die Übermittlung bereits abgeschlossen und

sind die Daten auf dem Zieldatenträger gespeichert, ist eine Strafbarkeit wegen Abfangens von Daten ausgeschlossen. Ein Hacking-Angriff kann dann nur noch nach § 202a StGB wegen Ausspähens von Daten strafbar sein, wenn die Daten gegen unberechtigten Zugriff besonders geschützt sind (vgl. Abschn. 3.3.1.4).

Der Tatbestand des Abfangens von Daten hat insoweit eine ähnliche Schutzrichtung wie das Fernmeldegeheimnis nach § 88 Telekommunikationsgesetz (TKG), dessen Verletzung nach § 206 StGB ebenfalls strafbar ist. Während das Fernmeldegeheimnis aber nur Telekommunikationsanbieter und andere Telekommunikationsdienstleister und deren Mitarbeiter bindet, kann eine Tat nach § 202b StGB von jedermann begangen werden, der sich Zugang zu übermittelten Daten verschafft.

Tatobjekt können wie bei § 202a StGB nur solche Daten sein, die nicht für den Täter bestimmt sind. Ebenso muss der Zugriff auf die Daten unbefugt erfolgen. Auf die Darstellungen in Abschn. 3.3.1.4 kann verwiesen werden.

Strafbarkeit setzt weiter voraus, dass die Übermittlung der abgefangenen Daten nicht-öffentlich erfolgt. Dies ist der Fall, wenn der Absender die Daten für einen erkennbar beschränkten Personenkreis bestimmt hat. Auf den Inhalt der übermittelten Daten kommt es nicht an (Weidemann 2014, § 202 a Rn. 6). Auch Daten aus dem Internet können daher i.S.v. § 202b StGB abgefangen werden, wenn die Übermittlung sich nicht an einen unbestimmten Empfängerkreis richtet.

Wie bei § 202a StGB ist es für die Strafbarkeit nicht erforderlich, dass der Täter die Daten kopiert. Ausreichend ist auch hier eine bloße Kenntnisnahme, die sich der Täter allerdings unter Anwendung technischer Mittel verschaffen muss. Bei solchen technischen Mitteln kann es sich um Hard- und Software, Codes, Passwörter o. ä. handeln.

3.3.1.6 § 202 c Vorbereiten des Ausspähens und Abfangens von Daten

Der Versuch des Ausspähens und Abfangens von Daten steht nicht unter Strafe. Gleichwohl sind nach § 202c StGB bestimmte Vorbereitungshandlungen strafbar, die vor allem die Herstellung und Verbreitung von Hacking-Tools und Zugangsdaten betreffen. Strafbar ist, „*wer eine Straftat nach § 202a oder § 202b vorbereitet, indem er Passwörter oder sonstige Sicherungscodes, die den Zugang zu Daten ... ermöglichen, oder Computerprogramme, deren Zweck die Begehung einer solchen Tat ist, herstellt, sich oder einem anderen verschafft, verkauft, einem anderen überlässt, verbreitet oder sonst zugänglich macht.*“

Die Strafbarkeit nach § 202c StGB setzt nicht notwendigerweise voraus, dass der Täter selbst eine Straftat nach § 202a StGB oder § 202b StGB plant. Auch wer einem Dritten durch die in § 202c StGB beschriebenen Handlungen eine solche Straftat ermöglicht (d. h. dies zumindest billigend in Kauf nimmt), macht sich strafbar.

Die Strafbarkeit der Herstellung und Verbreitung von Hacking-Tools bereitet in der Praxis einige Schwierigkeiten, da solche Tools gerade auch dazu entwickelt werden, Sicherheitslücken aufzudecken und so Einfallsstore für Hacking-Angriffe zu schließen. Der Gesetzgeber wollte einer unbeabsichtigten Ausüferung der Strafbarkeit im Bereich der IT-Sicherheit dadurch begegnen, dass er den Tatbestand auf solche Tools beschränkte, de-

ren objektive Zweckbestimmung die Begehung einer Straftat ist. Allgemeine Techniken, Anwendungen und Werkzeuge erfüllen den Tatbestand daher nicht. Problematisch sind freilich Tools, die sowohl zur Überprüfung der IT-Sicherheit als auch für rechtswidrige Angriffe genutzt werden können (Dual Use). Da es nach der überwiegenden Ansicht der Juristen in Deutschland ausreicht, dass ein Tool zumindest auch für die Begehung von Straftaten ist (vgl. Weidemann 2014, § 202c Rn. 7) setzen sich Programmierer solcher Dual-Use-Tools stets einem gewissen juristischen Risiko aus.

3.3.1.7 Ausblick

Rechtswissenschaft und Gesetzgeber diskutieren schon seit mehreren Jahren, inwieweit die Bedeutung von Daten als Wirtschaftsgut Anpassungen der bestehenden Gesetzeslage notwendig macht. Vergangene Gesetzesänderungen betrafen vor allem das Immaterialgüterrecht, etwa in Gestalt eines neuen Leistungsschutzrechts für Presseverleger (vgl. hierzu Abschn. 3.2.4.1).

Nach einer Initiative des Bundesrates ist der Gesetzgeber derzeit damit befasst, die sogenannte Datenhehlerei unter Strafe zu stellen (vgl. hierzu BT-Drucksache 18/1288). Der Gesetzesvorschlag ähnelt in seinem Wortlaut dem bestehenden Hehlerei-Tatbestand des Strafgesetzbuchs. Strafbar soll sein, wer mit Daten, die ein anderer ausgespäht oder sonst rechtswidrig erlangt hat, Handel treibt oder diese in anderer Weise in Umlauf bringt oder hält, um dabei sich oder einen anderen wirtschaftlich zu bereichern. Nach der Gesetzesbegründung hat der Gesetzgeber hierbei vor allem den illegalen Handel mit Zugangsdaten, E-Mail-Adressen und vergleichbaren Informationen im Auge. Nach dem Wortlaut des Entwurfs greift die Vorschrift aber auch ein, wenn andere Daten von wirtschaftlichem Interesse betroffen sind. Ob das Gesetz so, wie vom Bundesrat vorgeschlagen, in Kraft treten wird, war bei Fertigstellung dieses Handbuchs noch offen.

3.3.2 Zivilrechtlicher Schutz: Daten als absolut geschützte Rechtsgüter

Das Zivilrecht trifft für den Schutz von Rechtsgütern eine prinzipielle Unterscheidung:

- Die besonders bedeutsamen Rechtsgüter wie beispielsweise die körperliche Integrität und das Sacheigentum sind durch jedermann zu respektieren („absolut geschützte Rechtsgüter“, § 823 BGB).
- Sonderbeziehungen zwischen Personen (z. B. aufgrund von Verträgen oder Nachbarschaft oder familiärer Beziehung) führen zu weitergehenden Pflichten; § 241 Abs. 2 spricht von einer „Rücksicht auf die Rechte, Rechtsgüter und Interessen des anderen Teils“. In Vertragsbeziehungen ist auch das Interesse geschützt, keine Vermögensschäden zu erleiden.

In einem Schadensfall ist zunächst zu prüfen, ob eine solche Sonderbeziehung besteht, denn die Sonderbeziehung bewirkt typischerweise weitergehende Schadensersatzansprüche. Sodann ist zu prüfen, ob eines der absolut geschützten Rechtsgüter verletzt ist.

§ 823 Abs. 1 BGB definiert diese Rechtsgüter so:

... das Leben, der Körper, die Gesundheit, die Freiheit, das Eigentum oder ein sonstiges Recht.

Unter Eigentum versteht das Gesetz ausschließlich das Recht an körperlichen Gegenständen. Bei anderen Wirtschaftsgütern, z. B. Urheberrechten und Forderungen, spricht man von Inhaberschaft.

Als „sonstiges Recht“ wird durch die Rechtsprechung beispielsweise der eingerichtete und ausgeübte Gewerbebetrieb verstanden. Eingriffe in die Daten eines Unternehmens können zur erheblichen Störung des Gewerbebetriebs führen.

3.3.2.1 Daten auf eigenen Datenspeichern

Grundsatz

Wo Daten auf Datenspeichern verkörpert sind (z. B. im Buch, auf der Festplatte), scheint der Rechtsschutz unproblematisch. Der Datenträger ist zweifelsfrei ein Eigentumsgegenstand nach § 823 Abs. 1 BGB. Der Eigentumsschutz erfasst nicht nur die Substanzverletzung (z. B. Zerreißen, Verbrennen), sondern jede nicht nur kurzfristige Beeinträchtigung des bestimmungsgemäßen Gebrauchs der Sache, die nachteilige Beeinflussung ihrer Beschaffenheit (Sprau 2014, § 823 Rn. 7).

Bei Datenspeichern ist die Integrität in Bezug auf die Daten selbstverständlich eine zentrale und gesetzlich geschützte Eigenschaft.

Schutzlücken

Dieser Schutz der Daten über den Schutz des Eigentums am Datenträger funktioniert aber nur dann problemlos, wenn der Eigentümer des Datenträgers und der Inhaber der Daten identisch sind. Das war die früher ganz typische Situation. Durch moderne Formen von Dienstleistungen im Bereich der Datenverarbeitung, z. B. durch Hosting und Cloud-Dienste, liegen heute vielfach Daten auf Datenträgern, die dem Dateninhaber nicht eigentumsrechtlich zugeordnet sind. Auch schon bisher musste der Nutzer des Datenspeichers nicht sein Eigentümer sein; IT-Ausstattung konnte geleast werden oder von der Bank gegen Übertragung des Sicherungseigentums finanziert werden.

Beispiel: A arbeitet auf einer Reise am PC seines Gastgebers B und speichert auf diesem Gerät seine Arbeitsergebnisse. C beschädigt das Gerät; die Daten sind gelöscht. Bei B ist ein Schaden in Bezug auf die Daten nicht entstanden, denn sie gehören nicht ihm. A kann seinen Datenverlust nicht über ein Eigentumsrecht am Gerät geltend machen, weil er nicht der Eigentümer ist.

3.3.2.2 Daten als absolut geschützte Rechtsgüter

Grundsatz

Wie oben zitiert, erfasst § 823 Abs. 1 BGB auch „sonstige Rechte“. Die Formulierung ist offen. Gehören Daten zu den „sonstigen Rechten“?

Wert und Bedeutung von Daten stehen außer Zweifel. Der Bundesgerichtshof hat einmal entschieden, dass ein Datenbestand ein selbständiges vermögenswertes Gut sei; dies werde daran deutlich, „*dass er für sich von der Klägerin gegen Entgelt veräußert werden könnte.*“ (Bundesgerichtshof, Urteil v. 02.07.1996 – X ZR 64/94).

Diese Formulierung öffnet allerdings noch nicht die Einstufung als absolut geschütztes Rechtsgut nach § 823. Denn auch Forderungen sind vermögenswerte Güter, die gegen Entgelt veräußert werden können; sie sind aber nicht über § 823 Abs. 1 BGB geschützt.

In der juristischen Fachliteratur wird seit Jahren diskutiert, ob man Daten als solche „sonstigen Rechten“ des § 823 BGB erklärt (vgl. Zech 2012, S. 386–387 m. w. N.; Spindler 2011, S. 261).

Rechtsprechung und Rechtslehre sind seit jeher sehr zurückhaltend darin, Rechtsgüter und rechtliche Gegebenheiten als „sonstiges Recht“ nach § 823 Abs. 1 BGB einzustufen. Der Wertungshintergrund ist regelmäßig die Wertung der Verfassung.

Vom Bundesverfassungsgericht kommt Unterstützung zugunsten des Arguments, Daten als solche durch § 823 Abs. 1 BGB zu schützen. In seinem Urteil zur Vertraulichkeit und Integrität informationstechnischer Systeme betonte das Bundesverfassungsgericht, dass Daten von herausragender Bedeutung sind und dass aus grundrechtlicher Sicht ihr Schutz nicht vom eher zufälligen Eigentum am Datenträger abhängig gemacht werden darf (Bundesverfassungsgericht, Urteil vom 27.02.2008 – 1 BvR 595/07).

Das Urteil erging zwar zu personenbezogenen Daten. Aber auf der Grundlage des seitens des Bundesverfassungsgerichts sehr ausgedehnten grundrechtlichen Eigentums-schutzes (Art. 14 Grundgesetz) liegt die Ausdehnung des Arguments auf Daten, die in Unternehmen genutzt werden, sehr nahe. Wirtschaftsgüter, die bislang als Sachen auf den Markt kamen (insbesondere als Datenträger, z. B. als Buch), werden heute virtualisiert als Daten gehandelt; wer ein Buch lesen will, lädt sich die Daten auf sein Lesegerät. Das funktionale Ersatzstück verdient grundsätzlich denselben Schutz wie das Ersetzte. Nach Wertungsgesichtspunkten ist Gleichbehandlung geboten.

In dieser Diskussion wurde auch darauf hingewiesen, dass die noch aus dem römischen Recht stammende prinzipielle Unterscheidung zwischen Sachen (körperlichen Rechtsgütern) und anderen Gegenständen (z. B. Rechten, Forderungen, Daten) den heutigen Sachverhalt nicht mehr sinnvoll abbildet. So besteht bei den meisten Rechtsfragen, die es in Bezug auf Software gibt, kein Differenzierungsgrund, ob die Software dem Nutzer auf einem Datenträger oder online zukam (Schneider und Spindler 2014, S. 213 mit Hinweisen zur umfangreichen Diskussion).

Daten sind also Rechtsgüter, die gegenüber jedermann absolut geschützt sind, gleich wer der Eigentümer des jeweiligen Datenträgers ist. Der Schutz bedeutet, dass die Löschung, die Veränderung, übrigens auch die Vermischung mit weiteren Daten Sachverhalte sind, die nach § 823 Abs. 1 BGB zum Ersatz des daraus entstehenden Schadens verpflichten.

Einschränkungen

Das Postulat, technisch moderne Sachverhalt parallel einzustufen zu den Sachverhalten, die durch moderne Technik ersetzt oder verändert sind, führt allerdings zu einer Einschränkung dieses Grundsatzes.

Auch bisher war die Änderung oder Löschung eines einzelnen Datums nicht immer ein Sachverhalt, der über § 823 Abs. 1 BGB zum Schadensersatz führte. Der Schutz jedes einzelnen Bits ginge deshalb über eine Analogie hinaus.

Das oben zitierte Urteil des Bundverfassungsgerichts zur Vertraulichkeit und Integrität informationstechnischer Systeme benennt eine hohe Schwelle für Schadensersatzansprüche; es müsse um „wesentliche Teile der Lebensgestaltung“ gehen, die von dem Eingriff betroffen seien.

Dies liegt auf der Linie der Rechtsprechung in Bezug auf den Schutz des eingerichteten und ausgeübten Gewerbebetriebs (vgl. oben bei Abschn. 3.3.2.1; Sprau 2014, § 823 Rn. 126 ff.). Auch hier stellt der Bundesgerichtshof Schadensersatzansprüche aus § 823 Abs. 1 BGB unter hohe Anforderungen. Voraussetzung ist eine unmittelbare Beeinträchtigung des Betriebes als solchen oder eine Bedrohung seiner Grundlagen. Eingriffe in einzelne Rechtsgüter des Betriebes genügen typischerweise nicht. Die Integrität des Gewerbebetriebs muss tangiert sein. Dies kann aber auch durch Angriffe auf betriebsbezogene Daten geschehen (Sprau 2014, § 823 Rn. 127). Die schadensrechtliche Behandlung von Daten außerhalb von Betrieben sollte parallel zu diesen Fällen vorgenommen werden.

Ein weiteres Argument zur Beschränkung des Schutzes von Daten auf wesentliche Eingriffe ist dem Datenbankschutz zu entnehmen. Der im Urheberrechtsgesetz verankerte Schutz von Datensammlungen (§ 87 a UrhG, im Gesetz „Datenbank“ genannt) setzt in Bezug auf die technische Qualität und in Bezug auf den Erstellungsaufwand ein recht hohes Niveau voraus (Dreier und Schulze 2013, § 87 a Rn. 11 ff.).

Es liegt auf der Ebene dieser Wertungen, auch den Schutz von Daten durch § 823 BGB an ein vergleichbares Niveau zu binden. Schutzgegenstand nach § 823 Abs. 1 BGB wird deshalb nur eine Datensammlung von erheblicher Bedeutung sein. Das Abgrenzungskriterium sollte die Frage sein, ob der Nachteil aus dem Eingriff in das Rechtsgut mit den Schwellen vergleichbar ist, die die Gerichte in den zitierten Fällen und die das Gesetz beim Datenbankschutz installiert haben.

Diese Beschränkung ist insbesondere im Bereich Big Data und Smart Data leicht nachvollziehbar. Bei dieser Technik geht es gerade nicht darum, der Richtigkeit einer einzelnen kleinen Information zu vertrauen, nicht um die absolut korrekte Weitergabe und Speicherung von Daten, wie dies beispielsweise bei Telefonnummern erforderlich ist. Hier geht es darum, aus einer möglichst großen Fülle scheinbar nicht zusammenhängender Daten Ergebnisse zu erzielen, die über den Erkenntniswert der einzelnen Informationen hinausgehen. Ein schadensrechtlich relevanter Eingriff in eine solche Datenmenge setzt deshalb einen Eingriff von solcher Stärke voraus, dass das aus der Datenmenge erzielte Ergebnis verfälscht wird.

3.3.2.3 Ansprüche aus Schutzgesetzen

Zu diesem unmittelbaren Schutz des absolut geschützten Rechtsgut tritt ein Schutz über Schutzgesetze hinzu. Wer ein Schutzgesetz verletzt, muss die hieraus entstehenden Schäden begleichen (§ 823 Abs. 2 BGB).

Schutzgesetze in diesem Sinne gibt es in großer Zahl (Sprau 2014, § 823 Rn. 61–72). Für den Schutz von Daten sind mehrere Vorschriften des Strafgesetzbuches relevant, beispielsweise:

- das Ausspähen und Abfangen von Daten (§ 202 a, § 202 b, § 202 c StGB),
- die Verletzung oder Verwertung fremder Geheimnisse (§ 203, § 204 StGB),
- die Fälschung technischer Aufzeichnungen (§ 268 StGB),
- die Täuschung im Rechtsverkehr bei Datenverarbeitung (§ 270 StGB),
- die Fälschung beweiserheblicher Daten (§ 269 StGB),
- die Datenveränderung und die Computersabotage (§ 303 a, § 303 b StGB).

Dieser strafrechtliche Schutz, der durch zahlreiche Normen in Sondergesetzen ergänzt ist, wirkt über § 823 Abs. 2 BGB auch als Grundlage für Schadensersatzansprüche.

§ 823 BGB lässt als Verschuldensvorwurf prinzipiell Vorsatz und Fahrlässigkeit genügen.

3.3.2.4 Rechtsfolgen

Unterlassung

Wer ein absolut geschütztes Rechtsgut beeinträchtigt, muss die Beeinträchtigung unterlassen. Das folgt aus § 1004 BGB. Die Vorschrift spricht zwar nur vom Eigentum (womit Sacheigentum gemeint ist), wird aber nach einheitlicher Auffassung auf alle absolut geschützten Rechte angewandt, also nach der vorliegenden Darstellung auch auf geschützte Daten. Eine Beeinträchtigung liegt nicht nur in der unmittelbaren Änderung der Daten (durch Fälschung oder Löschung) und nicht nur in der Entziehung oder Vorenthalten des Zugangs zu den Daten, sondern auch im Kopieren und Entnehmen der Daten.

Wer Daten kopiert hat, ohne hierzu berechtigt zu sein (vgl. § 1004 Abs. 2 BGB) muss also die Beeinträchtigung beseitigen. Dies bedeutet, dass er die übernommenen Daten löschen muss. Zusätzlich muss er sich zur Unterlassung verpflichten. Dies geschieht dadurch, dass er eine mit einer Vertragsstrafe beschwerte Unterlassungsverpflichtung abgeben muss.

Diese Ansprüche sind nicht von einem Verschulden des Datenübernehmers abhängig. Schuldner des Anspruchs ist jeder, der durch seine Handlung oder pflichtwidrige Unterlassung die Beeinträchtigung bewirkt hat. Schädiger ist auch derjenige, der unerlaubt Inhaber der kopierten Daten ist, wenn die Beeinträchtigung wenigstens mittelbar auf seinen Willen zurückgeht (vgl. Bassenge 2014, § 1004 Rn. 15 ff.).

Schadensersatz

Unerlaubte Handlungen in Bezug auf Daten sind jede Beeinträchtigung des Inhalts der Daten und des Zugangs zu den Daten und zusätzlich die unerlaubte Kopie und alle denkbaren Verwendungen, die mit einer solchen unerlaubten Kopie geschehen. Wer solche Vorgänge durchführt oder veranlasst, schuldet Schadensersatz.

Nach § 249 Abs. 1 BGB gilt zunächst das Prinzip der Naturalrestitution: Der Schädiger hat den Zustand herzustellen, der ohne die unerlaubte Handlung bestände. Wo er illegale Daten hat, muss er sie also löschen. Er muss alles tun, damit eventuelle Dritte, die nun ebenfalls die Daten haben, ebenfalls die Daten löschen.

Wo der unerlaubte Eingriff zu Vermögenseinbußen führt, besteht ein Ersatzanspruch in Geld. Da Eingriffe in Daten als solche nicht Gegenstand der Rechtsprechung waren und auch in der Literatur kaum behandelt sind, sind insofern noch praktische Fragen offen. So kommt in Frage, die unerlaubte Inhaberschaft und insbesondere die unerlaubte Benutzung von Daten schadensrechtlich so zu behandeln, wie die entsprechende Situation beim Eingriff in Urheberrechte behandelt wird. § 97 UrhG bietet drei Berechnungsmöglichkeiten für den Schadensersatz, nämlich die Abführung des Verletzergewinns, die Analogie zur marktüblichen Vergütung, die für die Handlung bei Rechtmäßigkeit (also bei Gestaltung) verlangt worden wäre, und die Erstattung des tatsächlich eingetretenen Schadens.

Wo Persönlichkeitsrechte verletzt sind, kommt zusätzlich Schmerzensgeld in Frage.

Das Gericht ist berechtigt, den Schadensersatzbetrag gegebenenfalls durch Schätzung festzulegen (§ 287 ZPO).

Mit der Schadensbemessung im Fall einer Vernichtung des Datenstandes hat sich der Bundesgerichtshof in einem Urteil vom 09.12.2008 beschäftigt (VI ZR 173/07). Das Urteil skizziert folgende Grundsätze:

- Wenn die Daten wiederhergestellt werden können, ist hierfür der Aufwand zu erstatten.
- Wenn die Herstellung nur mit unverhältnismäßigem Aufwand möglich ist, ist eine angemessene Entschädigung in Geld zu leisten (§ 251 Abs. 2 BGB).
- Wenn die Daten eine qualifizierte geistige oder schöpferische Leistung („Unikat“) bilden, kommt die Erstattung der Wiederherstellungskosten nicht in Betracht.
- Bei der Bemessung der schadensrechtlichen Kompensationszahlung ist auch der Nachteil zu berücksichtigen, den das Fehlen der Daten in den Betriebsabläufen verursacht hat.

Insgesamt bietet dieses Urteil eine gute Ausgangsposition für die Erörterung der Schadenshöhe.

3.4 Reglementierung der Erhebung von Big Data

Carsten Ulbricht

Ein weit verbreiteter Weg große Datenmengen zu erheben, ist die Entnahme öffentlicher zugänglicher Informationen aus dem Internet. Dabei bestehen verschiedene technische Möglichkeiten die erheblichen Datenmengen aus fremden Internetplattformen auszulesen (sog. Screen-Scraping) und durch Aggregation oder Auswertung dieser Daten und Informationen neue Erkenntnisse zu gewinnen. Um genau das zu verhindern, untersagen zahlreiche Plattformen über die jeweiligen Nutzungsbedingungen eine systematische Entnahme von Daten und Informationen der Webseite in Form des Screen-Scraping.

Nachdem sich die vorgehenden Ausführungen vor allem mit der Frage beschäftigt haben, wie und welche gesetzlichen Vorgaben bei Big Data Projekten zu beachten sind, soll nachfolgend untersucht werden, ob beziehungsweise unter welchen Voraussetzungen Nutzungsbedingungen, deliktsrechtliche Regelungen oder Rechte Dritter einem Parsen oder Screen-Scraping fremder Daten rechtlich entgegenstehen.

3.4.1 Rechtliche Bewertung des Screen-Scraping

Nach der oben bereits erläuterten Vorschrift des § 87 b Abs. 1 Satz 2 UrhG ist ein Plattformbetreiber als Datenbankhersteller grundsätzlich gegen die wiederholte und systematische Vervielfältigung, Verbreitung und öffentliche Wiedergabe von nach Art und Umfang wesentlichen Teilen der Datenbank geschützt, sofern diese Handlung einer normalen Auswertung der Datenbank zuwiderläuft oder die berechtigten Interessen des Datenbankherstellers unzumutbar beeinträchtigt sind.

Vor dem Jahr 2009 wurde deshalb in mehreren Urteilen entschieden, dass ein solches Screen-Scraping aus einer fremden Datenbank in aller Regel als rechtwidrig anzusehen ist (vgl. LG München I, Urteil vom 18.09.2001 – 7 O 6910/01; LG Berlin Beschluss vom 27.10.2005 – 16 O 743/05; OLG Köln, Urteil vom 15.12.2006 – 6 U 229/05).

Seit 2009 sind dann jedoch unterschiedliche Entscheidungen ergangen, in denen Klagen der Datenbankhersteller gegen ungefragte Datenübernahme zurückgewiesen worden sind. So wurde in den Entscheidungen des OLG Frankfurt am Main (Urteil vom 05.03.2009 – 6 U 221/08) beziehungsweise des OLG Hamburg (Urteil vom 24.10.2012 – 5 U 38/10) eine unzulässige Nutzung nach § 87 b Abs. 1 Satz 2 UrhG verneint, wenn sich die Nutzung von Datensätzen der dort streitgegenständlichen einzelnen Flugverbindungen im Rahmen einer normalen Auswertung der Datenbank halte. Die berechtigten Interessen der betroffenen Flugunternehmen würden dann nach Auffassung der Gerichte insofern nicht unzumutbar beeinträchtigt. Eine ähnliche Entscheidung ist bezüglich der Entnahme von Daten aus diversen Automobilbörsen ergangen. Das OLG Hamburg hat in seiner Entscheidung (Urteil vom 16.04.2009 – 5 U 101/08) eine Rechtswidrigkeit für ein automatisiertes Verfahren, in dem in sehr kurzen Zeitabständen Suchanfragen bei mehreren

Internetautomobilbörsen durchgeführt worden sind, um die gewonnenen Informationen unmittelbar den Nutzern anzugeben, entsprechend verneint.

In einem aktuellen Urteil hat der BGH (Urteil vom 30.04.2014 – I ZR 224/12 – Flugvermittlung im Internet) eine Rechtsverletzung ebenfalls verneint und die wesentlichen rechtlichen Grundlagen zur Bewertung entsprechender Rechtsfragen ausgeführt.

Für eine Rechtswidrigkeit des Screen-Scraping ist es nach Auffassung des BGH erforderlich, dass eine Beeinträchtigung der wettbewerblichen Entfaltungsmöglichkeit festgestellt werden könnte, die über die mit jedem Wettbewerb verbundene Beeinträchtigung hinaus geht und bestimmte Unlauterkeitsmomente aufweist. Allein der Umstand, dass sich die dortige Beklagte, die die Daten entnommen hatte, über den von der Klägerin in ihren Geschäftsbedingungen geäußerten Willen hinwegsetze, keine Vermittlung von Flügen im Wege des sogenannten Screen-Scrapings zuzulassen, führte nach Auffassung des Bundesgerichtshofes nicht zu einer wettbewerbswidrigen Behinderung. Ein Unlauterkeitsmoment sei hingegen dann anzunehmen, wenn eine technische Schutzvorrichtung überwunden wird, mit der ein Unternehmen verhindern möchte, dass sein Internetangebot durch übliche Such- oder Analysedienste genutzt werden kann. Einer solchen technischen Schutzmaßnahme steht es nach Auffassung des Bundesgerichtshofes jedoch nicht gleich, wenn ein Unternehmen die Nutzung seiner Internetseite „nur“ von der Akzeptanz der entsprechenden Nutzungsbedingungen durch notwendiges Ankreuzen eines Kästchens abhängig macht.

3.4.2 Technische Schutzmaßnahmen

Wenn und soweit eine Internetplattform also keine hinreichend wirksamen technischen Schutzmaßnahmen gegen Screen-Scraping vorhält, können – jedenfalls nach deutschem Recht – Nutzungsbedingungen einer Plattform die Entnahme von Daten nicht rechtlich wirksam verbieten.

Technische Maßnahmen sind gemäß § 59 a Abs. 2 Satz 1 UrhG Technologien oder Vorrichtungen, die im normalen Betrieb dazu bestimmt sind, geschützte Werke betreffenden Handlungen, die vom Rechteinhaber nicht genehmigt sind, zu verhindern und einzuschränken.

Als hinreichend wirksame technische Schutzmechanismen, die im Zusammenwirken mit entsprechenden Nutzungsbedingungen auch zu einem wirksamen Verbot von Screen-Scraping führen können, werden derzeit gemeinhin zum Beispiel IP-Sperren, Captchas oder die Notwendigkeit eines Logins angesehen.

3.4.2.1 IP-Sperren

IP-Sperren funktionieren in der Regel in der Form, dass eine regelmäßige, sich wiederholende Abfrage von einer bestimmten IP-Adresse dazu führt, dass der Zugriff für diese IP-Adresse entsprechend gesperrt wird.

Wie oben erläutert, kann die Verwendung einer IP-Sperre nach der Rechtsprechung dazu führen, dass das Umgehen als Überwindung einer technischen Schutzvorkehrung interpretiert wird, was zur Rechtswidrigkeit des Datenzugriffs durch den Screen Scraper führt.

Da IP-Sperren objektiv dazu bestimmt sind, geschützte Werke betreffender Handlung, die vom Rechteinhaber nicht genehmigt sind, zu verhindern oder einzuschränken, ist im Grundsatz auch von einer entsprechenden technischen Schutzmaßnahme auszugehen. Auch wenn die Frage, ob die Umgehung von IP-Sperren auch tatsächlich zu einer Unzulässigkeit des Screen-Scraping führt, noch nicht gerichtlich entschieden worden ist, spricht einiges dafür, dass das Vorhalten einer entsprechenden IP-Sperre im Zusammenwirken mit einer in den Nutzungsbedingungen aufgenommenen entsprechenden Verbotsklausel der Datenentnahme grundsätzlich entgegensteht.

3.4.2.2 Captcha

Die Funktion des Captcha (engl. Completely Automated Public Turing test to tell Computers and Humans Apart) wird verwendet, um zu entscheiden, ob das Gegenüber ein Mensch oder eine Maschine ist. In der Regel wird durch die Abfrage einer als Bild dargestellten Zahlen- oder Buchstabenkombination über die Eingabe in das Internetformular sichergestellt, dass ein Mensch die Daten abfragt und nicht eine Maschine (Roboter, kurz „Bot“). Captchas dienen also der Verhinderung von voll automatischen Zugriffen und sind insoweit wohl als hinreichend wirksame technische Schutzmaßnahme gegen Missbrauch anzusehen.

3.4.3 Zusammenfassung

Angesichts der aktuellen Entscheidung des Bundesgerichtshofes kann davon ausgegangen werden, dass die bloße Regelung eines Verbotes von Parsing oder Screen-Scraping in Nutzungsbedingungen der Plattform der Entnahme von Daten nicht grundsätzlich entgegensteht. In vielen Fällen dienen entsprechende Nutzungsbedingungen also vor allem der Abschreckung.

Wenn und soweit die Erhebung und Verarbeitung von Daten also den oben stehenden gesetzlichen Restriktionen nicht entgegensteht, ist auch die Entnahme von Daten aus fremden Webpräsenzen grundsätzlich zulässig. Gerade bei Big Data Projekten, die in vielen Fällen die Entnahme wesentlicher Teile einer bestehenden Datenbank entnehmen, können allerdings vor allem die genannten datenbankrechtlichen Vorschriften zur Unzulässigkeit eines entsprechenden Big Data Projektes führen.

Rechtlich problematisch ist die Entnahme großer Datenmengen, wenn zur Verhinderung einer einfachen Abfrage technische Schutzmaßnahmen integriert worden sind, die den obenstehenden Vorgaben entsprechen. In diesen Fällen wird schon die Entnahme der so geschützten Daten regelmäßig unzulässig sein.

3.5 Anwendungsszenarien

Nachdem die rechtlichen Fragen im Rahmen etwaiger Big Data Projekten oben stehend allgemein dargestellt worden sind, sollen nun in diesem Kapitel die rechtlichen Gestaltungsmöglichkeiten für einige spezifische Anwendungsszenarien betrachtet werden.

3.5.1 Auswertung des Nutzungsverhaltens im Internet

Soll das Verhalten von Nutzern im Internet im Rahmen von Big Data Projekten gespeichert und verarbeitet werden (vgl. etwa die Analyse des Nutzungsverhaltens im Rahmen der Customer Journey im E-Commerce), so stellen sich vor allem datenschutzrechtliche Fragen.

Für die rechtliche Bewertung ist es dabei entscheidend, ob tatsächlich personenbezogene, oder „nur“ pseudonyme oder anonyme Daten erhoben und verarbeitet werden.

Da die Daten in der Regel über das jeweilige Telemedium erhoben werden, hat sich die rechtliche Zulässigkeit an den Vorgaben des Telemediengesetzes zu orientieren.

Die im Rahmen der Nutzung von Telemedien entstehenden Daten dürfen verarbeitet werden, wenn dies erforderlich ist, um die Inanspruchnahme von Telemedien zu ermöglichen und abzurechnen (§ 15 Abs. 1 TMG). Als Beispiele werden Merkmale zur Identifikation des Nutzers, Angaben über Beginn und Ende sowie des Umfangs der jeweiligen Nutzung und Angaben über die von Nutzern in Anspruch genommenen Telemedien aufgeführt. Nutzungsdaten dürfen etwa zu Abrechnungszwecken verwendet werden.

Da sich die Auswertung des Nutzungsverhaltens nur in den seltensten Fällen als (zwingend) erforderlich darstellen lassen wird, ist eine personenbezogene Auswertung entsprechender Nutzungsdaten wohl nur auf Grundlage einer Einwilligung des jeweiligen Nutzers wird darstellen lassen. Da sich die Nutzer in zahlreichen Fällen zu irgendeinem Zeitpunkt auf der jeweiligen Internetpräsenz oder einer anderen Anwendung (z. B. einer mobilen Applikation (App) anmelden, lässt sich die Auswertung des weiteren Nutzungsverhaltens dadurch legitimieren, dass der Nutzer im Rahmen der Anmeldung der jeweiligen Datenschutzerklärung ausdrücklich zustimmt. Wenn und soweit diese Datenschutzerklärung vor Beginn der Speicherung der jeweiligen umfassend und transparent über Art, Umfang und Zwecke der Erhebung und Verwendung der personenbezogener Daten aufklärt, lässt sich auch die weitere Datenerhebung und -verarbeitung rechtskonform gestalten.

Wenn und soweit eine pseudonyme Auswertung für die jeweiligen Zwecke ausreicht, so bildet § 15 Abs. 3 TMG, der heute schon im Rahmen unterschiedlicher Analyse- und Auswertungswerkzeuge herangezogen wird, einen weiteren wichtigen Erlaubnistatbestand. Nach dieser Vorschrift darf der Anbieter des Dienstes zu Werbe – und Forschungszwecken oder zur bedarfsgerechten Gestaltung der Telemedien pseudonyme Nutzungsprofile erstellen. Solche pseudonymen Nutzungsprofile dürfen erhoben werden, wenn eine Datenschutzerklärung den Nutzer auf dem jeweiligen Telemedium über diese Auswertung und sein Recht zum Widerspruch informiert. Wenn der betroffene Nutzer darüber hinaus

die Möglichkeit hat, die Profilierung im Rahmen seines Widerspruchs „auszuschalten“ (Opt-Out) beziehungsweise die Nutzungsprofile nicht mit Daten über den Träger des Pseudonyms zusammengeführt werden (§ 15 Abs. 3 Satz 3 TMG), so stehen der Erhebung solch pseudonymer Nutzungsprofile auch im Rahmen von Big Data Projekten keine rechtlichen Einwände entgegen.

3.5.2 Social Media Analysen

Bei Social Media Analysen (z. B. Social Media Monitoring) geht es in der Regel um die Aus- und Verwertung der enormen Datenmengen, die täglich von den Nutzern der Sozialen Medien über eigene Inhalte und das Nutzungsverhalten im Internet eingestellt werden. So stellen die Nutzer von Plattformen wie XING, Facebook, Twitter oder YouTube zahlreiche eigene Texte, Bilder, Musik und Videoinhalte (nutzergenerierte Inhalte oder User Generated Content) auf den Plattformen ein, deren Aus- und Bewertung den Unternehmen neue Erkenntnisse im Hinblick auf Marktforschung, Produktgestaltung oder Vertrieb liefern können.

Bei der Durchführung entsprechender Projekte sind diverse rechtliche Implikationen zu berücksichtigen, die oben auf eher abstrakter Ebene dargestellt worden sind.

So werden viele dieser Informationen und Inhalte als personenbezogene Daten zu werten sein. Die Erhebung, Speicherung oder Verarbeitung bedarf damit einer datenschutzrechtlichen Legitimation.

Neben der Einwilligung der Betroffenen ist – soweit diese nicht zu erlangen sind – vor allem der Legitimationstatbestand des §§ 28 Abs. 1 Satz 1 Nr. 3 BDSG relevant. Danach dürfen allgemein zugängliche Daten als Mittel für die Erfüllung eigener Geschäftszwecke verarbeitet werden, es sei denn, dass das schutzwürdige des Betroffenen an dem Ausschluss der Verarbeitung oder Nutzung gegenüber den berechtigten Interessen der verantwortlichen Stellen offensichtlich überwiegt (Ohrtmann und Schwiering 2014, S. 2986).

Da Aktivitäten in vielen Sozialen Netzwerken eine Anmeldung und Authentifizierung erfordern, ist weiter rechtlich umstritten, ob die Daten und Informationen in solchen Netzwerken tatsächlich als allgemein zugänglich angesehen werden können (Taeger 2014, § 28 Rn. 82, 83).

Nach der hier vertretenen Ansicht sind die Daten, solange der Betroffene den Zugriff eben nicht weitergehend eingeschränkt, jedenfalls dann als öffentlich zugänglich anzusehen, wenn der Anbieter des Dienstes die dafür vorgesehenen Daten über eine offene Schnittstelle (Application Programming Interface, API) öffentlich zur Verfügung stellt. Solange die Daten „allgemein zugänglich“ sind, spielt es im Übrigen auch keine Rolle, woher die verantwortliche Stelle die Daten hat. Sie kann sie der Primärquelle unmittelbar entnommen haben, ist aber genauso berechtigt, auf Sekundär- oder Tertiärquellen zurückzugreifen, selbst wenn diese nicht allgemein zugänglich sind. Ebenso wenig interessiert

sich das Gesetz dafür, wie die Daten der konkreten Informationsquelle jeweils entnommen werden (Simitis 2014, § 28 Rn. 158 f.).

Soweit die Informationen und Inhalte in und aus Sozialen Netzwerken als allgemein zugänglich anzusehen sind, ist deren Speicherung und Verarbeitung zulässig, wenn Interessen des Betroffenen, dem nicht offensichtlich entgegenstehen. Bei dieser Vorschrift kommt die gesetzgeberische Wertung zum Ausdruck, dass Daten die vom Betroffenen oder einem berechtigten Dritten öffentlich zugänglich gemacht werden, als weniger schutzbedürftig anzusehen sind. Bei entsprechenden Daten wird, gerade weil sie öffentlich gemacht worden sind, in der Regel nicht davon ausgegangen werden können, dass offensichtliche Interessen der Speicherung und Verarbeitung entgegenstehen.

Tendenziell können Big Data Analysen aus Sozialen Medien datenschutzrechtlich gerechtfertigt werden, wenn die Datenverarbeitung legitimen Interessen des Unternehmens dient. Denkbar sind Zwecke der Konzeption von Produkten, Marktanalysen und strategischem Management.

Neben der datenschutzrechtlichen Bewertung sind – je nach Konzeption der jeweiligen Big Data Analyse – auch urheberrechtliche Fragen zu berücksichtigen.

Da die nutzergenerierten Inhalte, je nach urheberrechtlicher Werkart (Text, Bilder, Audio- und Videoinhalte) nach § 2 Abs. 1 UrhG rechtlich geschützt sein können, bedarf auch eine Vervielfältigung (z. B. durch Speicherung auf eigenen Servern) oder eine anderweitige Veröffentlichung einer entsprechenden rechtlichen Prüfung.

Wenn in diesen Fällen keine Schrankenregelung, wie etwa das Zitatrecht (§ 51 UrhG) oder das Recht auf vorübergehende Vervielfältigungshandlung (§ 44a UrhG) oder andere urheberrechtliche Gestaltungsmöglichkeiten (z. B. Embedding) die jeweilige Nutzung legitimiert, wird sich der Verwender bei urheberrechtlich relevanten Nutzungshandlungen (§§ 15 ff. UrhG) ein entsprechendes Nutzungsrecht vom jeweiligen Rechteinhaber einräumen lassen müssen.

Social Media Analysen sind also datenschutzrechtlich zulässig, wenn die Betroffenen entweder eingewilligt haben oder die Legitimation des § 28 Abs. 1 Nr. 3 BDSG eingreift. Wenn und soweit urheberrechtlich geschützte Inhalte vervielfältigt oder veröffentlicht werden sollen, sind die vorgenannten urheberrechtlichen Fragen abzuklären.

3.5.3 Big Data in der Industrie (Industrie 4.0)

Neben der Nutzung verschiedener Daten, die von den Nutzern im oder über das Internet entstehen, rücken zunehmend auch Daten aus Industrieanlagen in den Fokus. Wer kritische Informationen, wie Schwingungen, Vibrationen oder andere relevante Werte aus- und bewerten kann, erhält wichtige Hinweise zur Optimierung von Abläufen und der damit einhergehenden Effizienzsteigerung.

Diese Entwicklung, die unter dem Stichwort Industrie 4.0 ausdrücklich von der Bundesregierung vorangetrieben wird, wirft ebenfalls einige zentrale Rechtsfragen auf.

In den meisten Fällen wird der Schwerpunkt mangels Speicherung und Verarbeitung personenbezogener Daten allerdings nicht so sehr auf datenschutzrechtlichen Fragen liegen. Anders könnte der Fall zu bewerten sein, wenn im Rahmen der Datennutzung auch IP-Adressen, die teilweise in der Literatur als personenbezogene Daten angesehen werden, verarbeitet werden. Im betrieblichen Kontext wird zu prüfen sein, ob der insoweit häufig heranzuhaltende Erlaubnistratbestand des § 28 Abs. 1 Satz 2 BDSG im Hinblick auf die Erforderlichkeit der jeweiligen Datenerhebung zur Erfüllung eigener Geschäftszwecke, die jeweilige Datennutzung legitimieren kann. Ansonsten wird die Einwilligung des jeweils Betroffenen einzuholen sein.

Bei Industrie 4.0 Projekten stellt als zentrale und insofern auch zwingend vertraglich zu regelnde Frage, wem „gehören“ die anfallenden Daten und Auswertungsergebnisse und wer darf diese wie verwerten.

In vielen Fällen werden die Industriedaten automatisiert (verschlüsselt oder unverschlüsselt) über entsprechende Mess- und Überwachungssysteme (z. B. Sensoren und Industrie PCs) gesammelt (Peschel und Rockstroh 2014, S. 571).

In zahlreichen Fällen wird sowohl die Datenerhebung und -auswertung, als auch der Abruf der ausgewerteten Informationen (z. B. über ein entsprechendes Internetportal) von einem Dienstleister und eben nicht dem Industrieunternehmen selbst vorgenommen werden. Mangels hinreichend konkreter und in vielen Fällen nicht sachgerechter gesetzlicher Regelungen, sollten die Parteien die gegenseitigen Rechte und Pflichten vertraglich regeln.

Dabei ist zu berücksichtigen, dass das Zivilrecht Eigentumsschutz für Daten nicht kennt. So erfasst § 903 BGB ausdrücklich nur das Eigentum an Sachen.

Beschränkungen der Nutzungsbefugnis an den Daten können sich hingegen aus anderen Rechtsvorschriften, insbesondere dem Urheber- und Datenbankrecht, möglicherweise auch dem Strafrecht und dem Wettbewerbsrecht ergeben.

Während Urheberrechtsschutz mangels Vorliegen einer hinreichenden „persönlichen Schöpfung“ im Sinne des § 2 Abs. 2 UrhG an Industriedaten in der Regel nicht angenommen werden können wird, können entsprechende systematischen Sammlungen und Auswertungen von Industriedaten die in Abschn. 3.2.2 beschriebenen Datenbankrechte (§ 87a ff. UrhG) begründen.

Wenn und soweit die erhobenen Daten – im Gegensatz zu einem reinen ungeordneten „Datenhaufen“ – unter Einsatz einer wesentlichen Investition auf eine systematisch und methodische Art und Weise aufbereitet werden, so wird in den meisten Fällen von einer geschützten Datenbank auszugehen sein.

Hersteller und damit Inhaber dieser Datenbank ist nach § 87a Abs. 2 UrhG, derjenige der die wesentliche Investition getätigt hat, die zur Beschaffung des Inhalts der Datenbank erforderlich ist und daher der Ermittlung von vorhandenen Elementen und deren Zusammensetzung in einer Datenbank dient.

In Rahmen einer gerichtlichen Auseinandersetzung hat der Bundesgerichtshof in seinem Urteil vom 25. März 2010 (Aktenzeichen: I ZR 47/08, „Autobahnmaut“) entschieden, dass nicht die Bundesrepublik Deutschland als Auftraggeber des landesweiten Mautsys-

tems Inhaber der datenbankrechtlich geschützten Mautdaten ist, sondern die Toll Collect GmbH, als Träger der organisatorischen Verantwortung und der Investitionen für die Geräte und das Rechenzentrum.

Ohne entsprechend anderslautende vertragliche Vereinbarung wird also derjenige Hersteller und damit Inhaber der Datenbank sein, der die Kosten für die Datenerhebung, -verarbeitung und -darstellung beziehungsweise die Kosten für den Betrieb und Wartung der Anlagen trägt.

Da dieses Ergebnis in vielen Fällen nicht sachgerecht erscheint, sollte die Inhaberschaft an den Daten und den Auswertungsergebnissen vertraglich geregelt werden.

Weitere Verfügungsbeschränkungen können sich in strafrechtlicher Hinsicht aus den §§ 303 a StGB ergeben, der eine unbefugte Datenveränderung unter Strafe stellt. Danach wird bestraft, wer rechtswidrig Daten im Sinne des § 202a Abs. 2 StGB löscht, unterdrückt, unbrauchbar macht oder verändert (s. hierzu näher in Abschn. 3.3). Durch den jeweiligen Auftrag an den Dienstleister wird eine strafrechtliche Relevanz regelmäßig ausgeschlossen sein. Denkbar ist diese nur dann, wenn der Dienstleister Daten in einer Art und Weise löscht, unbrauchbar macht oder verändert, die deutlich über das vertraglich Vereinbarte hinausgeht. Umso wichtiger ist es also für die Beteiligten, den Umfang der Datenerhebung und -verarbeitung genau zu definieren.

Da Industriedaten in vielen Fällen aufgrund ihrer Rückbeziehbarkeit auf Herstellungsverfahren, Eigenschaften und Wirkweisen von Maschinen auch als Betriebs- und Geschäftsgeheimnisse im Sinne des § 17 UWG interpretiert werden können, sollte auch insoweit die Speicher-, Nutzungs- und Weitergabebefugnis des Dienstleisters genau definiert werden. Sollte es sich um entsprechend sensible Daten handeln, erscheinen auch Vorgaben zur Datensicherheit bei der Übertragung, Speicherung und Verarbeitung im Sinne des § 9 BDSG sinnvoll.

Schließlich sollten Verträge über Projekte im Bereich Industrie 4.0 bei mehreren Beteiligten mögliche Gewährleistungs- und Haftungsfragen regeln. Die Qualität der Daten und der Auswertung kann für das Ergebnis ebenso relevant sein, wie die Frage, wer für etwaige Mängel, wie einzustehen hat (s. hierzu im Folgenden Abschn. 3.6.2).

Zusammenfassend werden bei Industrie 4.0 Projekten wohl weniger datenschutzrechtliche, umso mehr aber vertragsrechtliche Fragen eine Rolle spielen. Mangels klarer Vorgaben im Gesetz sind sowohl die Auftraggeber, als auch die Auftragnehmer bei entsprechend datenbasierten Diensten gut beraten, die Fragen nach der Inhaberschaft und der (Reichweite) der Nutzungsbefugnis, aber auch Gewährleistungs- und Haftungsfragen, wie auch Löschpflichten nach Abschluss des Auftrages klar festzulegen.

3.5.4 Zusammenfassung

Die oben stehenden Ausführungen zeigen, dass sich Big Data-Projekte bei rechtzeitiger Einbeziehung der entsprechenden rechtlichen Problemstellungen auch insofern in zulässiger Art und Weise darstellen lassen. Dabei zeigt die Erfahrung, dass eine frühzeitige

Einbindung eines kundigen Juristen oder Datenschutzbeauftragten die Projektverantwortlichen in die Lage versetzt, bereits in einem frühen Stadium die „richtigen Weichen zu stellen“. Es stellt sich regelmäßig als schwierig dar, die Rechtskonformität eines Big Data-Projektes sicherzustellen oder für diese zu sorgen, wenn der prüfende Jurist das Projektergebnis – wie häufig – erst im Rahmen einer finalen Abnahme zur Beurteilung vorgelegt bekommt.

Je nach Projekt, Art der Daten und der entsprechenden Verarbeitung sind unterschiedliche rechtliche Implikationen zu berücksichtigen. Bei Beachtung der ausgeführten datenschutz-, urheber- und vertragsrechtlichen Grundlagen sind entsprechende Big Data-Ansätze aber in vielen Fällen in zulässiger Form gestaltbar.

3.6 Verträge über Daten und Datenanalysen

Joachim Dorschel

Daten sind ein Wirtschaftsgut und eine Ware, mit der Handel getrieben werden kann. In vielen Fällen handelt es sich bei dem Erfasser, dem Analysten und dem Nutzer von Daten um unterschiedliche Unternehmen. Für einen rechtssicheren Umgang mit Daten und Datenanalysen ist es daher wichtig, die (geschriebenen oder ungeschriebenen) vertraglichen Beziehungen zwischen den Akteuren der jeweiligen Wertschöpfungsketten zu klären.

Für die vertragliche Einordnung von Geschäften über Daten sind unter anderem die folgenden Fragen relevant:

- Handelt es sich um eigene Datenbestände des Unternehmens oder sollen fremde Daten erworben werden?
- Werden die Daten dauerhaft oder auf Zeit überlassen?
- Geht es um die Überlassung fertiger Datenbestände, um deren Analyse oder die Erfassung neuer Daten?
- Erfolgt die Datenspeicherung auf eigenen Servern des Unternehmens oder in der Cloud?

Big Data-Projekte können je nach ihrer Zielrichtung unterschiedliche Rechtsgeschäfte nach sich ziehen. Denkbar ist, dass ein Unternehmen Marktdaten erwirbt, mit eigenen Vertriebsdaten zusammenführt, den Datenbestand in der Cloud speichert und ein Marktforschungsinstitut mit der Durchführung von Analysen beauftragt. Rechtlich kommt es in diesem Beispiel zu einem Kauf- oder Mietvertrag über die Marktdaten, einem Mietvertrag über den Cloud-Speicher und einem Dienst- oder Werkvertrag über die Datenanalyse.

Das vorliegende Kapitel betrachtet Rechtsgeschäfte, bei denen Daten unmittelbarer Leistungsgegenstand sind. Verträge über begleitende Leistungen, die typischerweise Bestandteil von Big Data-Projekten sind, etwa bei Einkauf von Beratung, der Erwerb von Software oder Hardware etc., sind in einschlägigen IT-rechtlichen Werken ausführlich dargestellt und sollen hier nicht näher betrachtet werden.

3.6.1 Wichtige Vertragstypen

3.6.1.1 Kaufverträge über Daten

Werden Daten dauerhaft gegen Vergütung überlassen, liegt rechtlich ein Kaufvertrag vor. Je nachdem, ob die Daten auf einem Datenträger oder online übergeben werden, handelt es sich um einen Sachauf gemäß § 433 BGB oder um einen Rechtskauf gemäß § 453 BGB. Da in beiden Fällen die kaufrechtlichen Regeln des BGB gelten, ist diese Unterscheidung für die Praxis selten relevant.

Wesentliche Vertragspflicht des Verkäufers ist es, dem Käufer die Daten in der vereinbarten Form zu überlassen. Im Big Data-Kontext werden Daten maschinell weiterverarbeitet. Das Format der überlassenen Daten ist daher für den Käufer wesentlich. Hier ist eine vertragliche Regelung zwingend geboten.

Ist der überlassene Datenbestand rechtlich geschützt (vgl. hierzu oben Abschn. 3.2) ist der Verkäufer verpflichtet, dem Käufer ein dauerhaftes Nutzungsrecht, also eine Lizenz, einzuräumen. Diese Lizenz wird den Käufer typischerweise berechtigen, die Daten selbst und für eigene Zwecke zu nutzen. Weitergehende Nutzungsbefugnisse, etwa das Recht des Käufers, die Daten Dritten im Internet zugänglich zu machen, bedürfen einer gesonderten vertraglichen Vereinbarung.

3.6.1.2 Zeitlich begrenzte Datennutzung

Werden dem Kunden die Daten nur für einen bestimmten Zeitraum zur Verfügung gestellt, wobei der Kunde ein zeitbezogenes Entgelt zu bezahlen hat, handelt es sich rechtlich um Miete. Eine entsprechende Einordnung hat der Bundesgerichtshof für das sogenannte Application-Service-Providing vorgenommen (vgl. BGH, Urteil v. 15.11.2006 – XII ZR 120/04). Die dort aufgestellten Grundsätze kann man auf eine zeitlich begrenzte Datenüberlassung ohne Weiteres übertragen. In der Praxis handelt es sich hier meist um Fälle, in denen der Kunde einen Online-Zugang zu Datenbeständen erhält, die durch oder im Auftrag des Datenlieferanten gehostet werden. Wirtschaftlich bedeutsame Beispiele sind etwa Marktdatenprovider und Anbieter wissenschaftlicher Recherchedatenbanken.

Wird der Zugang zu den Daten unentgeltlich gewährt, liegt juristisch ein Leihvertrag vor.

Auch bei der zeitlich begrenzten Überlassung ist zwischen der Verpflichtung zur Überlassung oder Zugangsgewährung und der Einräumung entsprechender Nutzungsrechte zu unterscheiden. Zu beachten ist, dass die bloße Onlinenutzung von Daten, bei der die Daten nicht, auch nicht vorübergehend, auf den Client des Nutzers kopiert werden, aus Sicht des Nutzers urheberrechtlich nicht relevant ist. Die bloße Zugangsgewährung erfordert also nicht die Einräumung einer Lizenz an den Nutzer. Sind die Daten rechtlich geschützt, muss der Anbieter freilich über das Recht verfügen, Daten zur Nutzung öffentlich zugänglich zu machen.

3.6.1.3 Aufträge zur Datenanalyse

Unternehmen können Big Data-Analysen selbst vornehmen oder Dritte hiermit beauftragen. Es hängt vom konkreten Leistungsinhalt ab, wie entsprechende Verträge juristisch einzuordnen sind.

Praktisch bedeutsam sind zunächst Angebote, bei denen Software-Provider dem Kunden Analysetools als Cloud-Service zur Verfügung stellen. Die Leistung des Providers beschränkt sich hier auf die Bereitstellung der Software und entsprechender IT-Ressourcen in der Cloud. Die Analyse selbst führt der Kunde unter Nutzung der ihm bereitgestellten Tools durch.

Bei diesem Geschäftsmodell handelt es sich letztlich um Software as a Service, welches vertragsrechtlich als Miete einzuordnen ist (vgl. hierzu die bereits oben zitierte Entscheidung zum Application-Service-Providing). Dass die bereitgestellte Software der Datenanalyse dient, ändert an dieser Einordnung nichts.

Anders verhält es sich, wenn das Kundenunternehmen den Anbieter mit der Durchführung der Datenanalyse beauftragt. Dies kann beispielsweise im Rahmen von Beratungsprojekten zu konkreten Problemstellungen geschehen. Auswahl und Aufbereitung der Daten und die Durchführung der eigentlichen Analyse obliegen hier dem Anbieter allein, ggf. unter Mitwirkung des Kunden.

Vertragsrechtlich kann es sich hier um einen Dienst- oder um einen Werkvertrag handeln. Ein Werkvertrag liegt vor, wenn der Anbieter ein konkretes Ergebnis schuldet, der Vertrag also nur erfüllt ist, wenn bestimmte vorab definierte Ziele (z. B. ein Gutachten, eine Präsentation etc.) erreicht sind. Schuldet der Anbieter dagegen bloße Beratung, wird er also für einen bestimmten Zeitraum tätig und hierfür bezahlt, handelt es sich rechtlich um einen Dienstvertrag. Der wesentliche Unterschied zwischen diesen beiden Vertragstypen ist, dass der Anbieter beim Werkvertrag einen definierten Erfolg schuldet, wobei das Leistungsergebnis frei von Mängeln sein muss. Erreicht der Anbieter diesen Erfolg nicht, haftet er dem Kunden wegen Vertragsverletzung. Beim Dienstvertrag dagegen erfüllt der Anbieter den Vertrag bereits dadurch, dass er mit den zugesagten Ressourcen (z. B. unter Einsatz von Consultants mit einem vereinbarten Skill Level) tätig wird. Dass diese Tätigkeit zu dem erwarteten Erfolg führt, ist vertragsrechtlich nicht entscheidend.

Verträge über Datenanalysen im Auftrag sollten eine Regelung zu den Rechten an den Arbeitsergebnissen enthalten. Dies gilt unabhängig davon, ob die Leistungsbeziehungen dienst- oder werkvertraglich ausgestaltet sind. Sind in das Projekt Betriebsgeheimnisse oder Know-how des Auftraggebers eingeschlossen, sollte sich das Kundenunternehmen die ausschließlichen Rechte an allen Arbeitsergebnissen ausbedingen und den Anbieter verpflichten, über jedwede Inhalte Stillschweigen zu wahren (sogenanntes Non-Disclosure Agreement).

3.6.1.4 Datenerhebung im Auftrag

Insbesondere im Bereich der Markt und Meinungsforschung kommt es vor, dass Unternehmen mit der Erhebung von Primärdaten beauftragt werden (hierzu eingehend Abschn. 2.6). Die Datenerhebung kann manuell etwa in Form von Befragungen oder Inter-

views oder maschinell in Form von Online-Umfragen oder der Auswertung vorhandener Datenquellen (z. B. des öffentlich zugänglichen Internet) erfolgen.

Für die vertragsrechtliche Einordnung entsprechender Aufträge gilt das zur Datenanalyse Gesagte (vgl. Abschn. 3.6.1.3). Wird ein konkretes Ergebnis geschuldet, z. B. die Durchführung einer anhand konkreter Merkmale festgelegten Stichprobe, handelt es sich um einen Werkvertrag. Wird der Auftragnehmer für einen bestimmten Zeitraum mit qualifiziertem Personal tätig, liegt ein Dienstvertrag vor.

Wie bei der Datenanalyse sollte man auch bei der Datenerhebung im Vertrag festlegen, in welchem Format die Ergebnisse übergeben werden und welche Rechte der Auftraggeber hieran erwirbt. Regelungsbedürftig ist auch die Frage, welche Qualität die zu erhebenden Daten haben müssen, ob diese z. B. eine bestimmte statistische Signifikanz aufweisen müssen (zu den spezifischen Problemen in diesem Zusammenhang beim Einsatz von Big Data vgl. Abschn. 2.3.4.1).

Besonderes Augenmerk verdient bei der Datenerhebung im Auftrag die rechtliche Bewertung der Erhebungsmethode. Das europäische und deutsche Recht halten ein engmaschiges Netz aus Vorschriften und Anforderungen vor, die zu beachten sind, wenn Unternehmen oder Verbraucher zu kommerziellen Zwecken, und hierzu zählen auch die Markt- und Meinungsforschung, kontaktiert werden. Hervorzuheben sind hier insbesondere die einschlägigen Bestimmungen im Gesetzes gegen den unlauteren Wettbewerb (UWG), des Datenschutzrechts und des Telemediengesetzes. Rechtsverstöße in diesem Bereich können regelmäßig zu kostenpflichtigen Abmahnungen durch Wettbewerber sowie Wettbewerbs- und Verbraucherverbände führen.

Bedient sich der Auftragnehmer rechtswidriger Mittel (z. B. durch den Versand von E-Mails ohne vorherige Zustimmung des Empfängers), kann der Auftraggeber Gefahr laufen, hierfür mit in Haftung genommen zu werden. Um dem vorzubeugen, ist der Auftraggeber gut beraten, den Auftragnehmer vertraglich auf eine Einhaltung der einschlägigen rechtlichen Bestimmungen zu verpflichten. Je enger die Grenzen hier gezogen werden, desto weniger wird man dem Auftraggeber eine eigene Einstandspflicht zuschreiben können, wenn der Auftragnehmer den Boden des rechtlich Zulässigen verlässt. Zugleich versetzt sich der Auftraggeber so in die Lage, im Schadensfall beim Auftragnehmer Regress zu nehmen.

3.6.1.5 Datenspeicherung im Auftrag

Viele Unternehmen lagern ihre Daten ganz oder teilweise in Rechenzentren aus. Im Big Data-Kontext liegt dies insbesondere nahe, wenn auch bei der Analyse auf Cloud-Lösungen zurückgegriffen wird.

Die Bereitstellung von Speicherplatz als Service ist in der Regel mietvertragsrechtlich einzurordnen.

Die inhaltliche Verantwortung für die Daten trägt in dieser Leistungsbeziehung der Auftraggeber. Der Auftragnehmer wird sich daher vertraglich absichern wollen, nicht für rechtswidrige Dateninhalte zur Verantwortung gezogen zu werden. In der Praxis verpflichtet sich der Auftraggeber daher regelmäßig, den Auftragnehmer bei einer Inanspruchnahme durch Dritte freizustellen.

3.6.2 Leistungsstörungen

Bei Störungen, insbesondere einer mangelhaften Leistungserbringung durch den Anbieter, gelten die allgemeinen Regeln des BGB, die im Rahmen des gesetzlich Zulässigen durch spezifische Vertragsklauseln angepasst oder ergänzt werden können.

Bei Kauf-, Werk- oder Mietverträgen muss dem Kundenunternehmen der Leistungsgegenstand frei von Sach- und Rechtsmängeln übergeben werden. Werden Daten überlassen, müssen diese die vereinbarten Eigenschaften haben, d. h., das vereinbarte Format, die vereinbarte Qualität usw. Fehlen entsprechende Regelungen im Vertrag, ist das Geschuldete durch Auslegung zu ermitteln.

Besondere Bedeutung hat bei der Überlassung von Daten die Freiheit von Rechtsmängeln. Ein Rechtsmangel liegt vor, wenn Dritte an den vertragsgegenständlichen Daten Rechte haben, die einer vertragsgemäßen Nutzung entgegenstehen. Dies ist z. B. der Fall, wenn die einer Datenbank entnommen wurden, ohne dass der Datenbankhersteller dem zugestimmt hätte (zum rechtlichen Schutz von Datenbanken vgl. Abschn. 3.2.2). Ein Rechtsmangel ist auch dann gegeben, wenn personenbezogene Daten entgegen den Bestimmungen des Datenschutzrechts gespeichert, verarbeitet oder übermittelt werden, die Betroffenen also der vertragsgemäßen Nutzung durch das Kundenunternehmen entgegentreten können. Gleiches gilt, wenn die Daten fremde Geschäfts- oder Betriebsgeheimnisse enthalten. Inwieweit Daten, die unter Verletzung strafrechtlicher Normen beschafft wurden (hierzu näher Abschn. 3.3.1), allein aus diesem Grund rechtsmängelbehaftet sind, ist eine Frage des Einzelfalls. In der Regel wird man davon ausgehen können, dass Berechtigte oder Geschädigte einer weiteren Nutzung der Daten durch den Auftraggeber entgegentreten kann, auch wenn dieser mit der rechtswidrigen Datenbeschaffung nichts zu tun hat.

Es ist die grundsätzliche Pflicht des Datenlieferanten, die Herkunft der Daten und ihre rechtliche Legitimation sorgfältig zu überprüfen.

3.6.3 Auftragsdatenverarbeitung

Handelt es sich bei den Daten, die Gegenstand der Leistungserbringung sind, um solche mit Personenbezug, so muss klar sein, welche Beteiligten der einzelnen Leistungsbeziehungen verantwortliche Stelle im Sinne des Datenschutzrechts sind. Jede verantwortliche Stelle benötigt für die rechtmäßige Verarbeitung personenbezogener Daten eine gesetzliche Erlaubnis oder eine Einwilligung der Betroffenen (vgl. Abschn. 3.1.1).

Wird im Zuge der Leistungserbringung einem anderen Unternehmen Zugriff auf Datenbestände gewährt, handelt es sich hierbei datenschutzrechtlich um eine Datenübermittlung, die als solche wiederum rechtfertigungsbedürftig ist. Dies gilt unabhängig davon, ob der Empfänger Kopien der Daten erhält oder ihm lediglich die Möglichkeit gegeben wird, auf Daten zuzugreifen. In der Praxis werden solche Fälle häufig über eine Auftragsdatenverarbeitung gemäß § 11 BDSG geregelt. Hiernach bedarf die Datenübermittlung

vom Auftraggeber an den Auftragnehmer keiner Einwilligung oder gesetzlichen Erlaubnis, wenn

- der Auftraggeber den Auftragnehmer unter besonderer Berücksichtigung der Eignung der von ihm getroffenen technischen und organisatorischen Maßnahmen (vgl. § 9 BDSG) ausgewählt hat,
- zwischen dem Auftraggeber und dem Auftragnehmer eine schriftliche Vereinbarung (sogenannte Auftragsdatenverarbeitungsvereinbarung, ADV) geschlossen wurde, in der bestimmte in § 11 Abs. 2 BDSG geforderte Punkte ausdrücklich geregelt sind,
- der Auftraggeber sich vor Beginn der Datenverarbeitung und sodann regelmäßig von der Einhaltung der beim Auftragnehmer getroffenen technischen und organisatorischen Maßnahmen überzeugt und das Ergebnis dieser Überprüfungen dokumentiert und
- der Auftragnehmer sich in Bezug auf die Verarbeitung personenbezogener Daten vollständig den Weisungen des Auftraggebers unterwirft und nur gemäß diesen Weisungen handelt.

Die Auftragsdatenvereinbarung ist ein in der Praxis des IT-Outsourcing weitverbreitetes Instrument der datenschutzrechtlichen Absicherung. Es existieren zahlreiche Muster entsprechender Vereinbarungen¹. Beim Einsatz solcher Musterformulierungen ist freilich Vorsicht geboten. Die vom Gesetz verlangten Elemente der Vereinbarung, etwa zum Gegenstand und zur Dauer des Auftrags, zum Umfang, die Art und den Zweck der Datenverarbeitung, die Art der Daten und den Kreis der Betroffenen, müssen sich auf das konkrete Auftragsverhältnis beziehen. Allgemeine Formulierungen und Umschreibungen genügen diesen Anforderungen nicht.

Das Instrument der Auftragsdatenverarbeitung gemäß § 11 BDSG steht nur zur Verfügung, wenn die Datenverarbeitung innerhalb der Europäischen Union stattfindet. Der Auftragnehmer darf also keine Rechenzentren außerhalb der Europäischen Union betreiben (vgl. hierzu Funke und Wittmann 2013). Dies ist insbesondere zu beachten, wenn für die Datenvereinbarung oder Speicherung auf Cloud-Infrastrukturen zurückgegriffen werden soll.

Für eine Datenverarbeitung außerhalb der Europäischen Union kommt eine Auftragsdatenverarbeitung auf Grundlage der sogenannten EU-Standardvertragsklauseln (Kommissionsbeschluss 2010/87/EU vom 5. Februar 2010) in Betracht.

Die EU-Standardvertragsklauseln stellen jedoch spezifische Anforderungen an die beteiligten Unternehmen, die im Einzelnen nicht immer leicht zu erfüllen sind. Auch bieten die Standardvertragsklauseln nicht in allen Fällen die gleiche Rechtssicherheit wie eine EU-interne Auftragsdatenverarbeitung.

¹ Hervorzuheben ist hier die „Mustervereinbarung zum Datenschutz und zur Datensicherheit in Auftragsverhältnissen“ nach § 11 BDSG vom 28.09.2010, abrufbar unter <https://www.datenschutz-hessen.de/ft-auftragsdatenverarbeit.htm>, die zwischen den Aufsichtsbehörden der Länder abgestimmt ist und von diesen allgemein anerkannt wird.

Die Vereinbarung über die Auftragsdatenverarbeitung muss schriftlich geschlossen werden. Diese Anforderung ist nur durch einen von Auftraggeber und Auftragnehmer auf der gleichen Urkunde unterschriebenen Vertrag oder durch ein mit einer qualifizierten elektronischen Signatur versehenes Dokument erfüllt (Funke und Wittmann 2013, S. 225). Andere Formen der Auftragsverteilung, etwa ein Austausch von E-Mails oder Telefaxen, ist nicht ausreichend.

Literatur

Literatur zu 3.1

- Dammann, U. (2014), Kommentierung zu § 3 BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Dorschel, J./Nauerth, Ph. (2013), Big Data und Datenschutz – ein Überblick über die rechtlichen und technischen Herausforderungen, *Wirtschaftsinformatik und Management* (S. 32–38). Springer: Heidelberg.
- Finsterbusch, St./Knop, C. (2012): Im Meer der Daten, *Frankfurter Allgemeine Zeitung*, 26.10.2012, <http://www.faz.net/-gqe-73wle>, zuletzt abgerufen am 25.06.2014.
- Gola, P/Schomerus, R. (2015): *BDSG – Kommentar*. C.H. Beck: München; 12. Aufl.
- Grüneberg, Ch. (2014), Kommentierung zu § 276 BGB, in: Palandt: *Bürgerliches Gesetzbuch – Kommentar*. C.H. Beck: München; 73. Aufl.
- Jüngling, Th. (2013): Wie die Sammler von Big Data uns durchleuchten, *Die Welt*, 04.03.2013, <http://www.welt.de/114121023>, zuletzt abgerufen am 18.03.2015.
- Katko, P/Babaei-Beigi, A. (2014): Accountability statt Einwilligung? Führt Big Data zum Paradigmenwechsel im Datenschutz?, *MultiMedia und Recht – MMR* (S. 360–364). C.H.Beck: München.
- Poppenhäger, H. (2003), Datenschutz in der amtlichen Statistik, in: Roßnagel, A. (Ed.): *Handbuch Datenschutzrecht* (S. 1622–1643). C.H. Beck: München.
- Roßnagel, A. (2013): Big Data – Small Privacy? Konzeptionelle Herausforderungen für das Datenschutzrecht, *Zeitschrift für Datenschutz – ZD* (S. 562–567). C.H.Beck: München.
- Roßnagel, A./Pfitzmann, A./Garstka, H. (2001): *Modernisierung des Datenschutzrechts*. Bundesministerium des Innern: Berlin.
- Schirrmacher, F. (2013): *Ego – Das Spiel des Lebens*. Karls Blessing Verlag: München.
- Scholz, Ph. (2011), Kommentierung zu § 3 a BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 7. Aufl.
- Schulz, S. (2013), Kommentierung zu § 3 a BDSG, in: Wolff, H. A./Brink, St. (Eds.): *Beck'scher Online-Kommentar Datenschutzrecht*. C.H.Beck: München.
- Simitis, S. (2014): *BDSG - Kommentar*. Nomos: Baden-Baden; 8. Aufl.
- Weichert, T. (2008) Datenschutz. in: Kilian, W/Heussen, B. (Eds.): *Computerrechts-Handbuch*, C.H. Beck, München.
- Weichert, Th. (2013): Big Data und Datenschutz – Chancen und Risiken einer neuen Form der Datenanalyse, *Zeitschrift für Datenschutz – ZD* (S. 251–259). C.H. Beck: München.
- Witt, B. (2010): *Datenschutz kompakt und verständlich*. Vieweg + Teubner: Wiesbaden; 2. Aufl.

Literatur zu 3.1.2–3.1.5

- Arning M./Forgó N./Krügel T. (2008), Datenschutzrechtliche Aspekte der Forschung mit genetischen Daten, Datensicherheit und Datenschutz (S. 700–705). Vieweg: Heidelberg.
- Dammann, U. (2014), Kommentierung zu § 3 BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Ehmann, E. (2014), Kommentierung zu § 30a BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Eichler, C./Kamp M. (2014), Kommentierung zu § 3 a BDSG, in: Wolff, H. A./Brink, St. (Eds.) *Beck'scher Online-Kommentar Datenschutzrecht*. C.H.Beck: München
- Ernestus, W. (2014), Kommentierung zu § 9 BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Gola, P./Schomerus, R. (2015): *BDSG – Kommentar*. C.H. Beck: München; 12. Aufl.
- Roßnagel, A. (2013): Big Data – Small Privacy? Konzeptionelle Herausforderungen für das Datenschutzrecht, *Zeitschrift für Datenschutz – ZD* (S. 562–567). C.H.Beck: München.
- Schaar, P. (2002): *Datenschutz im Internet*, C.H.Beck: München.
- Simitis, S. (2014): Kommentierung zu § 28 BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Weichert, Th. (2008): Datenschutz. in: Kilian, W./Heussen, B. (Eds.): *Computerrechts-Handbuch*, C.H. Beck, München.
- Weichert, Th. (2013): Big Data und Datenschutz – Chancen und Risiken einer neuen Form der Datenanalyse, *Zeitschrift für Datenschutz – ZD* (S. 251–259). C.H. Beck: München.
- Zieger, C./Smirra, N. (2013): Fallstricke für Big-Data Anwendungen – Rechtliche Gesichtspunkte bei der Analyse fremder Datenbestände, *Multimedia und Recht* (S. 418–421). C.H.Beck: München.

Literatur zu 3.1.6

- Dix, A. (2014): Kommentierung zu § 34 BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Gola, P./Schomerus, R. (2015): *BDSG – Kommentar*. C.H. Beck: München; 12. Aufl.
- Forgó, S. (2015), Kommentierung zu § 33 BDSG, in: Wolff, H. A./Brink, St. (Eds.): *Beck'scher Online-Kommentar Datenschutzrecht*. C.H.Beck: München.
- Wolff/Brink *Beck'scher Online-Kommentar Datenschutzrecht*, Stand: 01.08.2014

Literatur zu 3.1.8

- Bäcker, M. (2015): Kommentierung zu § 4 BDSG, in: Wolff, H. A./Brink, St. (Eds.): *Beck'scher Online-Kommentar Datenschutzrecht*. C.H.Beck: München.
- Gola, P./Schomerus, R. (2015): *BDSG – Kommentar*. C.H. Beck: München; 12. Aufl.
- Kania, T. (2015): Kommentierung zu § 94 BetrVG, in: Müller-Glöge, R./Preis, U./Schmidt, I. (Eds.): *Erfurter Kommentar zum Arbeitsrecht*. C.H. Beck: München 15. Aufl.
- Linck, R. (2013): Beschäftigtendatenschutz, in: Schaub, G. (Begr.) *Arbeitsrechts-Handbuch*, C.H. Beck: München, 15. Aufl.
- Schiedmair, S. (2014): Kommentierung zu § 25 BDSG, in: Wolff, H. A./Brink, St. (Eds.): *Beck'scher Online-Kommentar Datenschutzrecht*. C.H. Beck: München.

- Seifert, A. (2014): Kommentierung zu § 32 BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Simitis, S. (2014): Kommentierung zu § 4a BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Spindler, G., Nink, J., (2015): Kommentierung zu § 28 BDSG, in: Spindler, G./Schuster, F. (Eds.), *Recht der elektronischen Medien*. C.H. Beck: München; 3. Aufl.
- Thüsing, G. (2014): Kommentierung zu § 92 BetrVG, in: Richardi, R. (Ed.), *Betriebsverfassungsgesetz mit Wahlordnung*, C.H. Beck: München; 14. Aufl.
- Thüsing, G./Forst, G. (2014): Informationserhebung bei der Einstellung und beim beruflichen Aufstieg, in: Thüsing, G. (Ed.) *Beschäftigungsdatenschutz und Compliance*, C.H. Beck: München. 2. Aufl.

Literatur zu 3.1.9

- Gola, P./Schomerus, R. (2015): *BDSG – Kommentar*. C.H. Beck: München. 12. Aufl.
- von Lewinski, K. (2015): Kommentierung zu § 6a BDSG, in: Wolff, A./Brink, S. (Eds.), *Beck'scher Online-Kommentar Datenschutzrecht*. C.H. BECK: München; 11. Aufl.

Literatur zu 3.2

- Dreier, T./Schulze, G. (2013): Kommentar zum Urheberrechtsgesetz. C.H. Beck: München.
- Kahler, J./Helbig K. (2012): Umfang und Grenzen des Datenbankschutzes bei dem Screen Scraping von Onlinedatenbanken durch Online-Reiseportale, *Wettbewerb in Recht und Praxis* (S. 48–55), dfv: Frankfurt a.M.
- Wiebe (2013): Grundzüge des Immaterialgüterrechts im Bereich der Informationstechnologie, in: Leupold, A./Glossner, S. (Eds.) *Münchener Anwaltshandbuch IT-Recht*, C. H. Beck: München.
- Zieger, C./Smirra, N. (2013): Fallstricke für Big-Data Anwendungen – Rechtliche Gesichtspunkte bei der Analyse fremder Datenbestände, *Multimedia und Recht* (S. 418–421). C.H.Bek: München.

Literatur zu 3.3.1

- Hoeren, T./ Völkel, J. (2014): Eigentum an Daten, in: Hoeren, T. (Ed.) *Big Data und Recht*. C.H. Beck: München.
- Weidemann, M. (2014): Kommentierung zu § 202a StGB und § 303 StGB. in: Heintschel-Heinegg, B. (Ed.) *Beck'scher Online-Kommentar StGB*, C.H. Beck: München; 25 Aufl.
- Wieck-Noodt, B. (2014): Kommentierung zu § 303a StGB. in: Heintschel-Heinegg, B. (Ed.) *Beck'scher Online-Kommentar StGB*, C.H. Beck: München.

Literatur zu 3.3.2

- Bartsch, M. (2008): Die Vertraulichkeit und Integrität Informationstechnischer Systeme als sondiges Recht nach § 823 Abs. 1. Computer und Recht (S. 613–617). Verlag Dr. Otto Schmidt: Köln.
- Bassenge, P. (2014): Kommentierung von § 1004 BGB, in: Palandt, *Bürgerliches Gesetzbuch*. C. H. Beck: München. 73. Aufl.
- Dreier, T./Schulze, G. (2013): Kommentar zum Urheberrechtsgesetz. C.H. Beck: München.

- Schneider, J./Spindler, G. (2014): Der Erschöpfungsgrundsatz bei „gebrauchter“ Software im Praxistest. *Computer und Recht*, (S. 213–223) Verlag Dr. Otto Schmidt: Köln.
- Spindler, G. (2011): Der Schutz virtueller Gegenstände. Leible, Lehmann, H. Zech: *Unkörperliche Güter im Zivilrecht*. Mohr Siebeck: Tübingen.
- Sprau, H. (2014): Kommentierung von § 823 BGB, in: Palandt, *Bürgerliches Gesetzbuch*. C. H. Beck: München. 73. Aufl.
- Zech, H. (2012): *Information als Schutzgegenstand*. Mohr Siebeck: Tübingen.

Literatur zu 3.5

- Ohrtmann, P./Schwiering, S. (2014): Big Data und Datenschutz – Rechtliche Herausforderungen und Lösungsansätze, *Neue Juristische Wochenschrift* (S. 2984–2989). C.H. Beck: München.
- Peschel C./Rockstroh S. (2014), Big Data in der Industrie – Chancen und Risiken neuer datenbasierter Dienste, *Multimedia und Recht* (S. 571–576). C.H. Beck: München.
- Simitis, S. (2014), Kommentierung zu § 28 BDSG, in: Simitis, S. (Ed.): *Bundesdatenschutzgesetz*. Nomos: Baden-Baden; 8. Aufl.
- Taeger, J. (2014), Kommentierung zu § 28 BDSG, in: Taeger J./Gabel D. (Eds.): *Bundesdatenschutzgesetz*. Deutscher Fachverlag: Frankfurt a.M. 2. Aufl.

Literatur zu 3.6

- Funke, M./Wittmann, J. (2013): Cloud Computing – ein klassischer Fall der Auftragsdatenverarbeitung? *Zeitschrift für Datenschutz*, 2013, (S. 221–228). C.H. Beck: München.

Gernot Fels, Carsten Lanquillon, Hauke Mallow, Fritz Schinkel und Christian Schulmeyer

4.1 Grenzen konventioneller Business-Intelligence-Lösungen

Carsten Lanquillon und Hauke Mallow

4.1.1 Business Intelligence: Ein Überblick

Der Begriff *Business Intelligence* (kurz BI) ist seit mehr als zwei Jahrzehnten stark verbreitet und BI-Lösungen sind inzwischen ein fester Bestandteil in vielen Unternehmen. Dennoch herrscht immer noch Unstimmigkeit darüber, was BI konkret umfasst.

4.1.1.1 Verwendung und Definitionen des Begriffs

Die Verwendung des Begriffs BI geht bis weit ins 19. Jahrhundert zurück. In der „Cyclopaedia of commercial and business anecdotes“ aus dem Jahre 1865 wird beispielsweise beschrieben, wie ein Banker sich bemüht, systematisch an relevante Informationen zu gelangen, um aus diesem Wettbewerbsvorteil wirtschaftlichen Nutzen zu ziehen. Mit zunehmender Digitalisierung historischer Buchbestände und angemessenen Suchfunktionen mögen gar noch frühere Quellen ans Licht kommen.

Gernot Fels 

Walzbachtal, Deutschland

Prof. Dr. Carsten Lanquillon

Heilbronn, Deutschland

Hauke Mallow

Leinfelden-Echterdingen, Deutschland

Dr. Fritz Schinkel

Unterhaching, Deutschland

Dr. Christian Schulmeyer

Seeheim-Jugenheim, Deutschland

Im IT-Kontext wird BI vermutlich erstmals im Jahr 1958 vom IBM-Forscher Hans Peter Luhn verwendet. In einem Artikel mit dem Titel „A Business Intelligence System“ beschreibt Luhn, wie automatisch auf die Nutzer abgestimmte Informationen an die richtigen Stellen in einer Organisation verteilt werden sollen (Luhn; 1958, S. 314). Dabei fasst er *Business* als Zusammenfassung jeglicher zielgerichteter Aktivitäten auf. Es geht also nicht nur um Unternehmen im betriebswirtschaftlichen Sinne, sondern z. B. auch um Behörden oder Forschungseinrichtungen. Der englische Begriff *Intelligence* wird als die Fähigkeit verstanden, in vorliegenden Daten Zusammenhänge zu erkennen, die sodann zielführendes Handeln ermöglichen. Es geht folglich nicht um menschliche oder künstliche Intelligenz in einer Organisation, sondern um die Gewinnung von Einsichten oder Erkenntnissen in Analogie zur Verwendung des Begriffs Intelligence im Kontext nachrichtendienstlicher Tätigkeit (vgl. CIA).

Übergeordnetes Ziel von BI ist es, das fundierte und rechtzeitige Treffen von strategischen, taktischen oder operativen Entscheidungen in einem Unternehmen zu unterstützen. Dazu werden unternehmensinterne und -externe Daten zu Erkenntnissen veredelt, die insbesondere im Rahmen der Leistungskontrolle, der Optimierung von Geschäftsprozessen und der Planung benötigt werden.

Trotz dieser ursprünglich klaren Ausrichtung und Zielsetzung gibt es keine einheitliche BI-Definition. Waren die ersten Definitionen in den 1990er Jahren noch sehr technikorientiert, so kommen bei aktuellen Definitionen eine fachliche und eine organisatorische Perspektive hinzu.

Vergleicht man unterschiedliche Definitionen, so gilt BI typischerweise als begriffliche Klammer für Konzepte, Prozesse und Technologien zur systematischen Sammlung, Vereinheitlichung, Speicherung, Auswertung und Darstellung von Daten.

Hervorzuheben ist die ganzheitliche Sichtweise, bei der BI als unternehmensspezifischer Gesamtansatz zur betrieblichen Entscheidungsunterstützung aufgefasst wird, wobei die Integration verschiedener analytischer Anwendungen und die Ausrichtung von BI an der Unternehmensstrategie entscheidende Erfolgsfaktoren bei der Einführung und Umsetzung von BI-Vorhaben in Unternehmen sind (Kemper, Baars, and Mehanna; 2010, S. 8).

4.1.1.2 Evolution entscheidungsunterstützender Systeme

Die Idee einer computerbasierten Entscheidungsunterstützung ist nicht neu. Seit den 1960er Jahren wurden unter unterschiedlichen Begriffen entscheidungsunterstützende Systeme entwickelt und auf den Markt gebracht: Management Information Systems (MIS), Decision Support Systems (DSS), Executive Information Systems (EIS), Data-Warehouse-Systeme mit einem Fokus auf Online Analytical Processing (OLAP) und schließlich BI-Lösungen sind, wie in Abb. 4.1 dargestellt, Teil dieser Evolution (Humm und Wietek 2005).

Mit fortschreitendem Stand der Technik und Forschung wurde insbesondere der Umfang der integrierten Datenquellen sowie deren Detaillierungsgrad und Aktualität, die Bandbreite der Analysemöglichkeiten und der Anwenderkreise mit zunehmender Akzeptanz gewachsen.

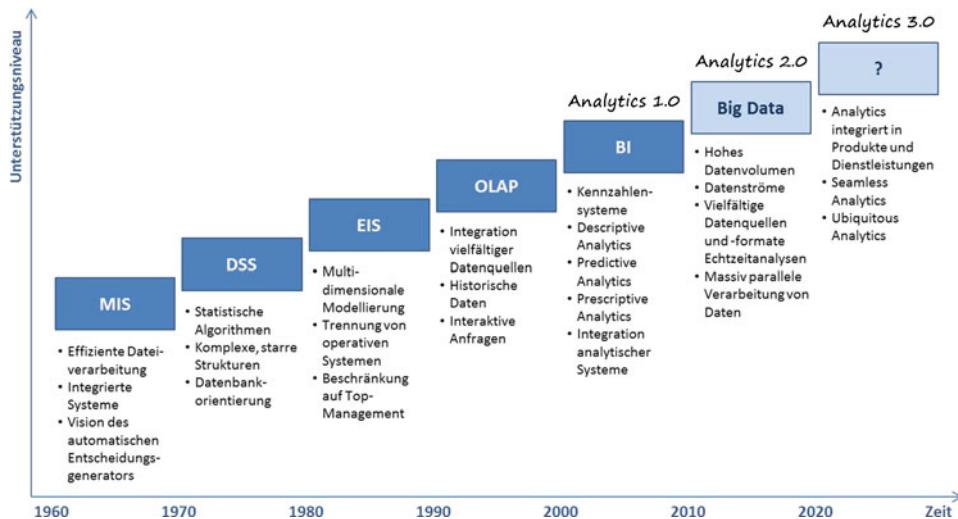


Abb. 4.1 BI-Stammbaum: Evolution entscheidungsunterstützender Systeme in Anlehnung an Humm und Wietek (2005) und ergänzt um Aspekte aus Davenport (2013)

tanz ausgeweitet. Weiterhin wurde die Integration mit Geschäftsprozessen und angrenzenden betrieblichen Anwendungssystemen vorangetrieben.

Führt man diese Entwicklung fort, lässt sich Big Data oder im Prinzip genauer Big Data Analytics als Business Intelligence der nächsten Generation auffassen. Diese Einschätzung setzt allerdings voraus, dass man BI im weitesten Sinne auslegt und ihr jegliche Form der Datennutzung zuschreibt, wie im nächsten Abschnitt über den Funktionsumfang dargelegt. Außerdem gilt es zu berücksichtigen, dass neben der Gewinnung von Erkenntnissen zur Entscheidungsunterstützung auch das Betreiben von Geschäftsmodellen hinzugekommen ist, die im Kern auf Big Data aufbauen.

Als nächste Stufe der Entwicklung lässt sich eine zunehmende Einbettung von Analysefunktionalität in Produkte und Dienstleistungen ausmachen (Davenport; 2013). Themen wie das *Internet der Dinge* und *Industrie 4.0* stehen damit direkt in Verbindung. Mit Blick auf die einzubettende Analysefähigkeit kann der Integrationsgrad variieren, beispielsweise mag lediglich die Einbettung einer Modellanwendung oder aber auch eine Modellanpassung oder gar die gesamte Modellerstellung erforderlich sein (vgl. auch Abschn. 2.3.4.1).

4.1.1.3 Diskussion um das Analysespektrum

Insbesondere über den tatsächlichen Funktionsumfang von BI gibt es unterschiedliche Auffassungen. Die Kenntnis darüber ist hilfreich, um Diskussionen über eine Abgrenzung von BI und Big Data bzw. Kritik an BI-Systemen im Kontext von Big Data einordnen zu können.

In der Praxis wird BI oft ausschließlich mit statischem Berichtswesen (Reporting) und dynamischen Berichten und Abfragemöglichkeiten (OLAP) assoziiert. Dies liegt zum einen an der verfügbaren Funktionalität bei gängiger BI-Software und einer Fokussierung auf das statische und dynamische Berichtswesen in der Anfangsphase vieler BI-Programme. Eine weitere Ursache sind zum anderen sicherlich die Marketingstrategien der Softwareanbieter.

Diese eingeschränkte Sichtweise führt dazu, dass BI oft als lediglich vergangenheitsorientiert bemängelt wird. Big Data sei dagegen zukunftsorientiert, denn mithilfe von Predictive Analytics oder gar Prescriptive Analytics würden Vorhersagen und die Angabe konkreter Handlungsempfehlungen ermöglicht. Dadurch liefere Big Data gerade im Vergleich zu BI einen erheblichen Mehrwert für Unternehmen.

Eine Reduzierung auf eine vergangenheitsorientierte Nutzung von Daten im Kontext von BI erscheint allerdings künstlich und wenig sinnvoll. Trotz der Vielfalt an BI-Definitionen lässt sich kaum ein Hinweis darauf finden, dass eine tiefere Analyse der bereitgestellten Daten zur Beantwortung weiterführender Fragestellungen auszuschließen sei. Dies ist auch nicht zu erwarten, wenn es der Gewinnung von Erkenntnissen und letztlich der Unterstützung von Entscheidungen und somit dem primären Ziel von BI dient.

Daher legen wir BI im weitesten Sinn aus und zählen jede sinnvolle Form der Datennutzung zur Gewinnung von Erkenntnissen für die Beantwortung oder Lösung fachlicher Fragestellungen dazu. Somit umfasst BI alle Nutzungsarten der bereitgestellten Daten von einfachen SQL-Abfragen und statischen Berichtswesen bis hin zu Vorhersagen und dem Ableiten von Handlungsempfehlungen. Die Abb. 4.2 zeigt gängige analyseorientierte, datengetriebene BI-Anwendungen, die inzwischen oft unter dem Schlagwort *Analytics* zusammengefasst werden, und benennt die Fragestellungen, die damit typischerweise adressiert werden. Das Thema *Advanced Analytics mit Big Data* wurde bereits ausführlich in Abschn. 2.3 behandelt. Neben den genannten gibt es auch noch weitere Anwendungen, wie etwa Planung, Optimierung und Simulation.

4.1.1.4 BI-Referenzarchitektur

In einem BI-System lassen sich die Bereiche Datenbereitstellung, oft auch als Data Warehousing bezeichnet, und Datennutzung identifizieren. Werkzeuge zur Datennutzung werden oft als BI-Anwendungen (Applikationen) bezeichnet. Beide Bereiche lassen sich noch weiter funktional untergliedern. Dies führt zu einer in Schichten organisierten Referenzarchitektur, wie in Abb. 4.3 dargestellt.

Wie im vorangegangenen Abschnitt erläutert, zählt im weitesten Sinne jede Art der Datennutzung als potenzielle BI-Anwendung. Sinnvoll ist eine Trennung zwischen der tatsächlichen Berechnung oder Umsetzung von Abfragen mit einer gewissen Semantik bzw. der Anwendung von Analysemethoden und der Präsentation der Ergebnisse, in welcher Art und Weise diese auch immer gefordert wird, denn einige BI-Anwendungen unterscheiden sich primär in der Aufbereitung der Ergebnisse und nicht in den Ergebnissen selbst.

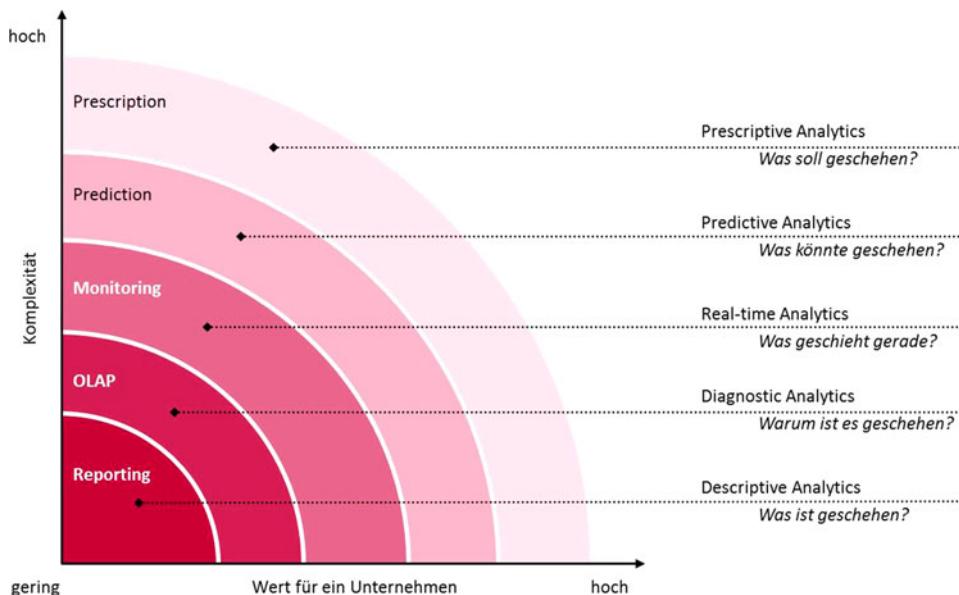


Abb. 4.2 BI-Analysespektrum: Fragestellungen im Kontext von BI in Anlehnung an Eckerson (2007)

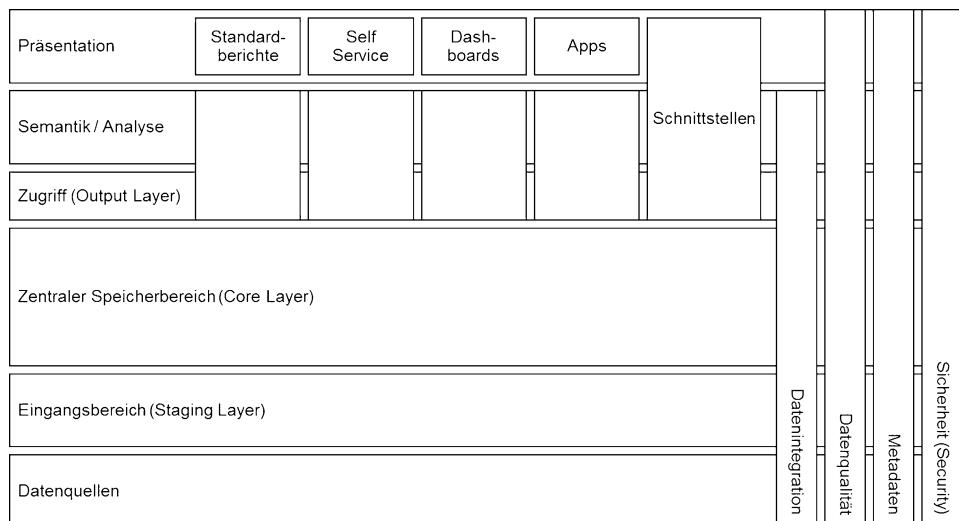


Abb. 4.3 BI-Referenzarchitektur: Funktionale Schichten eines BI-Systems

Den Kern klassischer BI-Systeme bildet das sogenannte Data Warehouse. Meist kommen dabei relationale Datenbankmanagementsysteme (RDBMS) zum Einsatz, die für die Speicherung und Analyse strukturierter Daten optimiert wurden.

Entscheidungsrelevante Daten aus überwiegend unternehmensinternen, aber auch einigen externen Quellen laufen über einen sogenannten Eingangsbereich (Staging Layer) in den zentralen Speicherbereich (Core Layer). Datenintegrationsprozesse stellen sicher, dass die Daten dort bereinigt und qualitätsgeprüft gespeichert werden. Die Verarbeitung geschieht mit sogenannten ETL-Tools, wobei ETL für die Extraktion, die Transformation und das Laden der ausgewählten Daten steht.

Die Datenintegration ist von zentraler Bedeutung für das Data Warehouse, denn Daten können nur geladen werden, sofern sie in ein vordefiniertes relationales Datenbankschema passen. Was nicht passt, wird passend gemacht oder geht verloren, sofern kein separater Speicherort für fehlerhafte Daten vorgesehen ist. Dieser Ansatz wird als *Schema-on-Write* bezeichnet und erfordert oft sehr aufwendige Transformationsschritte.¹

In der Schicht für den Datenzugriff (Output Layer) erfolgt die Aufbereitung der Daten für die bereits genannten Nutzungsvarianten der verschiedenen BI-Anwendungen. Zur Beschleunigung oder Vereinfachung des Datenzugriffs werden dafür meist sogenannte Data Marts erstellt, die einen domänen spezifischen Teilausschnitt der im zentralen Speicherbereich verfügbaren Daten repräsentieren. Berichtsorientierte Anwendungen, insbesondere das Standard-Reporting, Ad-hoc-Reporting und OLAP, decken oft schon einen großen Teil der Benutzeranforderungen ab. Um die dafür notwendigen Datenzugriffe zu beschleunigen, erfolgt die Speicherung der Data Marts oft in sogenannten OLAP-Würfeln, die teilweise sogar im Hauptspeicher gehalten werden.

4.1.2 Grenzen von BI-Lösungen im Kontext von Big Data

Big Data bezeichnet sehr große Datenmengen (Volume) mit großer Vielfalt (Variety), die zum Teil aus Quellen mit zweifelhafter Glaubwürdigkeit (Veracity) stammen sowie in hoher Geschwindigkeit (Velocity) erzeugt werden und verarbeitet werden sollen, um daraus einen wirtschaftlichen Nutzen zu generieren. Diese Eigenschaften stellen große Herausforderungen sowohl an das Datenmanagement als auch an die Analyse. Die Grenzen klassischer BI-Lösungen insbesondere mit einem Fokus auf Data-Warehouse-Architekturen lassen sich gut anhand dieser Eigenschaften erläutern.

4.1.2.1 Volume

Klassische relationale Datenbanken bieten kaum horizontale Skalierungsmöglichkeiten, das sogenannte Scaling-out wird meist nicht unterstützt. Database-Appliances, die für Data Warehousing optimiert sind, ermöglichen zwar diese Skalierungsmöglichkeit und

¹ Vergleiche Tab. 4.3 in Abschn. 4.2.2.1 prüfen für eine Gegenüberstellung der Konzepte *Schema-on-Write* und *Schema-on-Read*.

können damit auch Zugriffe auf größere Datenmengen performant verarbeiten, jedoch meistens zu nicht unerheblichen Kosten. Außerdem setzen auch Database-Appliances wie alle anderen relationalen Datenbanken ein vordefiniertes Schema voraus.

Für diese Datenbanken bieten einige ETL-Tools die Möglichkeit, Verarbeitungen in die Datenbank auszulagern. Es handelt sich dabei um einen sogenannten Push-Down, bei dem Datenbank-Anweisungen (Statements) in die Datenbank gedrückt werden. Allerdings kann man nicht von einer engen Verzahnung zwischen Verarbeitung der Anweisungen und den im Rahmen der Datenvorverarbeitung erforderlichen Transformationen sprechen. Eine horizontale Skalierung ist im Sinne einer Gesamtarchitektur daher nicht möglich, und so sind die verarbeitbaren Datenmengen hinsichtlich der Größe begrenzt.

Gleches gilt für den Analyseprozess, sofern fortschrittliche Analysemethoden (Advanced Analytics) nicht über SQL auf dem RDBMS laufen können. Dann nämlich werden derartige Analysemethoden, also z. B. Data-Mining-Methoden, in der Regel auf einem Datenextrakt durchgeführt, und zwar entweder auf lokalen Rechnern oder auf dedizierten Analyseservern. Das Bewegen sehr großer Datenmengen ist allerdings sehr zeitintensiv und sprengt ab einer gewissen Grenze auch die Kapazitäten der für die Analyse vorgesehenen Rechner. Daher werden bei Big Data-Lösungen die Daten dort verarbeitet und analysiert, wo sie liegen. Die Analysemethoden werden zu den Daten gebracht, nicht umgekehrt.

4.1.2.2 Velocity

Klassische BI-Systeme sind in der Regel auf gespeicherte Daten (*Data-at-Rest*) ausgelegt, d. h. es wird sowohl für die Datenintegration als auch für die Analyse davon ausgegangen, dass auf gespeicherte (statische) Daten zugegriffen wird. Dies hat einen erheblichen Einfluss auf die Art und Weise, wie diese Daten verwendet werden können. Lässt man Laufzeitaspekte außen vor, so kann beliebig oft auf alle verfügbaren Daten zugegriffen werden.

Datenströme (*Data-in-Motion*), die in der Regel in Echtzeit verarbeitet werden müssen, können nicht direkt angebunden werden, ein direkter Zugriff zumindest auf einen Teil der Daten ohne vorherige Speicherung ist nicht möglich. Sollen Daten aus Datenströmen verwendet werden, sind diese vorher zu speichern. Bei größeren Mengen erforderte dies in der Regel eine Auswahl relevanter Teile, was die späteren Analysemöglichkeiten einschränken kann. Außerdem stünden die Daten nur mit einer gewissen Latenz zur Verfügung, was im Rahmen von Echtzeit-Anwendungen äußerst kritisch wäre.

Sollen größere Mengen an Daten aus einem Datenstrom in Echtzeit verarbeitet werden, dann bleibt nicht viel Zeit für die in einem Data-Warehouse üblichen Datenintegrations- und Bereinigungsprozesse. Daher geht die Datenbereitstellung in diesem Fall stets zu Lasten der Konsistenz. Folglich wird zum einen der Schema-on-Write-Ansatz der kritische Faktor bei der technischen Umsetzung und zum anderen widerspricht ein Speichern ohne angemessene Integration und Bereinigung den Ansprüchen an die Datenqualität, die in der Regel an die Daten in einem Data Warehouse gestellt werden.

4.1.2.3 Variety

Relationale Datenbanken eignen sich hervorragend zur Speicherung strukturierter Daten. Nach allgemeinen Schätzungen sind heutzutage jedoch 80–85 % der Daten unstrukturiert (z. B. Textdokumente oder andere Multi-Media-Daten) oder semi-strukturiert (z. B. maschinengenerierte Logfiles). Folglich kann ein Großteil der verfügbaren Daten gar nicht oder erst nach sehr aufwendigen Transformationsschritten in einem Data Warehouse gespeichert werden.

Da aber immer mehr Anwendungsszenarien auch die Verarbeitung un- und semi-strukturierter Daten verlangen, steigt die Notwendigkeit, auch diese Daten zu speichern. Gefragt sind daher Lösungen, die sowohl strukturierte als auch semi- und unstrukturierte Daten speichern können, um Anwendungen in diesen Fällen einen einfachen und transparenten Zugriff auf alle benötigten Daten bieten zu können.

Grenzen bei der Verarbeitung strukturierter Daten, insbesondere im Zusammenhang mit dem Schema-on-Write-Konzept, äußern sich auch im Scheitern des in den 1990er und 2000er Jahren häufig propagierten Enterprise-Data-Warehouse-Ansatzes. Ziel dieses Ansatzes ist die Speicherung sämtlicher Unternehmensdaten in einem einzigen Data Warehouse, um dieses dann als einzige Datenquelle für alle dispositiven und analytischen Anwendungen in einem Unternehmen zu nutzen. Insbesondere in größeren Konzernen setzte sich dieser Ansatz nicht durch, da es nicht möglich war, sich unternehmensweit auf ein Datenmodell zu einigen. Ein vereinfachter Ansatz beschränkte sich nur auf die Stammdaten und fand sich in sogenannten Master-Data-Management-Systemen wieder. Auch hier war häufig die fehlende Akzeptanz in einem Unternehmen der erfolgsverhindernde Faktor.

4.1.2.4 Veracity

Die strukturierten Daten aus den operativen Vorsystemen sind in der Regel mit Bedacht modelliert und weisen zumindest im Rahmen ihrer ursprünglichen Verwendungsabsicht ein vertretbar hohes Maß an Datenqualität auf. Bei „zweckentfremdetem“ Einsatz wird die Datenqualität jedoch schon deutlich schlechter eingestuft. Bei internen Datenquellen gibt es zumindest direkte Ansatzpunkte zur Steigerung der Datenqualität. Im Big-Data-Umfeld sieht die Situation oft anders aus. Die un- oder semi-strukturierten Daten stammen zum Teil aus externen Datenquellen und haben oft eine zweifelhafte Datenqualität. Daher muss bei Analysen mit einem gewissen Grad an Unsicherheit und fehlender Vertrauenswürdigkeit umgegangen werden. Mechanismen, die dies bei der Verarbeitung berücksichtigen oder bei der Bereitstellung von Ergebnissen entsprechend kennzeichnen, sind in klassischen BI-Systemen nicht vorgesehen, da dort von einer hohen Datenqualität ausgegangen wird, obgleich diese in der Praxis oft weit hinter den Erwartungen zurück bleibt.

4.1.3 Zusammenfassung

Business Intelligence (BI) und Big Data haben eine vergleichbare Zielsetzung, nämlich die Gewinnung von Erkenntnissen, die einen wirtschaftlichen Nutzen für ein Unternehmen haben. Während BI-Lösungen primär interne strukturierte Daten dafür nutzen, umfasst Big Data letztlich alle Arten und Formen von Daten, die für ein Unternehmen relevant sein könnten. Wäre es nur das Datenvolumen, könnten BI-Lösungen sicherlich entsprechend aufgerüstet werden. In Kombination mit der Vielfalt und Geschwindigkeit, mit der die Daten verarbeitet werden müssen, sind klassische BI-Lösungen überfordert. Stattdessen müssen neuartige Technologien zum Einsatz kommen, wie in folgenden Kapiteln ausführlich dargelegt.

4.2 Big Data-Lösungen

Carsten Lanquillon und Hauke Mallow

4.2.1 Anforderungen an Big Data-Lösungen

Die Anforderungen an ein Big Data-System werden ganz wesentlich durch die Anforderungen bestimmt, die sich aus den sogenannten vier Vs – also Volume, Variety, Velocity und Veracity – und den angestrebten Analysen (Analytics) ergeben. Aber auch bewährte Anforderungen an ein BI-System müssen berücksichtigt werden, denn ein Big Data-System muss ebenso gut auch Eigenschaften einer BI-Lösung wie etwa die Verarbeitung und Ad-hoc-Analyse strukturierter Daten bewältigen. Tabelle 4.1 stellt die High-Level-Anforderungen für ein Big Data-System dar (vgl. hierzu auch Kimball; 2011, S. 9f.).

4.2.2 Big Data-Referenzarchitekturen

Eine Big Data-Referenzarchitektur soll die im Abschnitt 4.2.1 dargestellten Anforderungen abdecken. In vielen Veröffentlichungen finden sich überwiegend technisch getriebene Architektur-Muster (Patterns). Diese stellen meist eine Komponentensicht dar, in der beispielsweise erläutert wird, welche Arten von Lösungen für welche Anwendungsfälle zur Speicherung der benötigten Daten geeignet sind, wie etwa Graph-Datenbanken für die Speicherung von sozialen Netzwerken und Key-Value-Stores für definierte Abfragen.

4.2.2.1 Funktionale Big Data-Referenzarchitektur

Abbildung 4.4 skizziert eine allgemeine Referenzarchitektur unter funktionalen Aspekten, die im Folgenden detaillierter erläutert wird.

Die Aufgaben, die eine solche Architektur lösen muss, werden logischen Schichten zugeordnet. Zur besseren Einordnung werden zu den jeweiligen Architekturkomponen-

Tab. 4.1 High-Level-Anforderungen für ein Big Data-System

Volume	Sehr große Datenvolumen, die aktuell im Petabyte-Bereich liegen, müssen gespeichert und verarbeitet werden können.
	Ein Big Data-System muss gut skalieren und fehlertolerant auf den Ausfall einzelner Komponenten reagieren (Robustheit).
	Sowohl kleine Datenpakete (Einzelsätze, Ereignisse) als auch große Datenmengen müssen vom System verarbeitet und gespeichert werden können.
Variety	Beliebige Datenquellen und Formate (z. B. Videos, Bilder, Texte, Web-Inhalte) müssen unterstützt werden.
	Die Schemalosigkeit muss unterstützt werden, d. h. Daten müssen geladen werden können, ohne dass die Struktur bekannt ist. Weiterhin muss es möglich sein, für die Daten unterschiedliche Schemata definieren zu können.
	Einer hohen Änderungshäufigkeit von Metadaten in Datenquellen muss begegnet werden können (Agilität).
Velocity	Sehr große Datenströme (<i>Data Streams</i> oder <i>Data-in-Motion</i>) müssen verarbeitet werden können.
	Auf Ereignisse oder Ereigniskonstellationen in Datenströmen muss in Echtzeit reagiert werden können.
Veracity	Für die Integration unterschiedlichster Datenquellen muss eine Metadatenintegration möglich sein.
	Metadaten für die Analyse von unstrukturierten Daten müssen zur Verfügung stehen.
	Eine Datenqualitätskomponente muss enthalten sein, z. B. mit Funktionalitäten zur Datenbereinigung und -harmonisierung, Messung der Datenqualität.
	Daten sollten immer in ihrem Ursprungsformat vorliegen, ohne inhaltlich verändert worden zu sein.
Analytics	Strukturierte mengenorientierte Abfragen (SQL) müssen mit niedrigen Latenzzeiten möglich sein.
	Sowohl der Zugriff auf Einzelsätze durch feststehende Abfragen als auch nicht vorhersehbare Abfragemuster (Ad-hoc-Abfragen) müssen mit niedrigen Latenzzeiten unterstützt werden.
	Analysen auf Datenströmen müssen in Echtzeit möglich sein.
	Unstrukturierte Daten, wie z. B. Textdokumente oder Videos, müssen analysiert werden können. Hierfür sollten auch Suchfunktionen, Indizierung, semantische Anreicherung und das Annotieren von Inhalten (Tagging) unterstützt werden.
	Fortschrittliche Analysemethoden (Advanced Analytics) insbesondere für prädiktive und präskriptive Fragestellungen müssen zur Verfügung stehen.
	Neue Datenquellen und Analyseansätze müssen einfach in Sandboxes evaluiert werden können.
	Abhängig von den Inhalten, müssen die Daten gegen Missbrauch und vor unberechtigten Zugriff geschützt werden können.
	Der Lebenszyklus von Daten muss berücksichtigt werden, wie z. B. das Löschen personenbezogener Daten nach bestimmten Zeiträumen aufgrund rechtlicher Vorgaben.
	Die Visualisierung muss neuen Anforderungen gerecht werden, wie z. B. die Darstellungen von Knoten und Kanten in großen Graphen oder Echtzeitaktualisierungen.

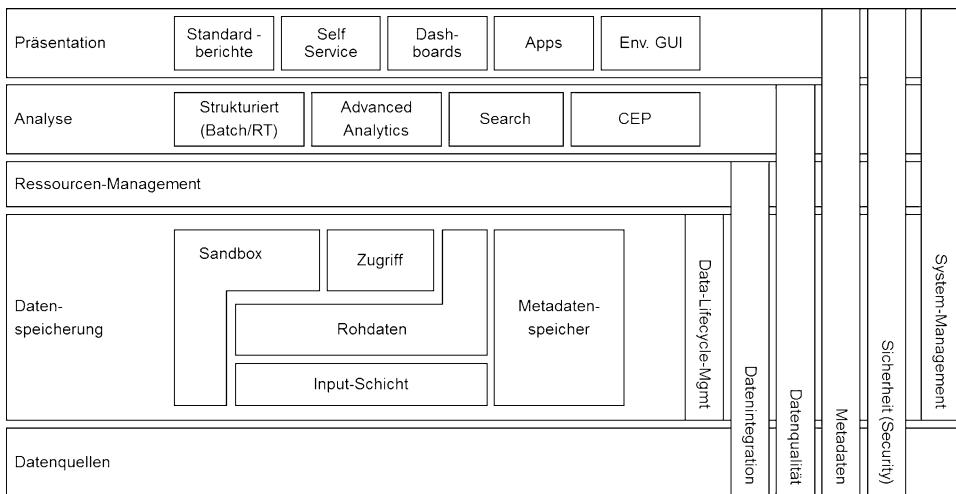


Abb. 4.4 Referenzarchitektur für Big Data-Lösungen aus funktionaler Sicht

ten konkrete Realisierungen aus dem Hadoop-Ökosystem genannt, das sich zu einer der Kerntechnologien im Kontext von Big Data entwickelt hat. Details zum technischen Hintergrund der verschiedenen Komponenten finden sich im entsprechenden Kapitel dieses Handbuchs.

Die Eigenschaften Variety und Velocity beeinflussen die Architekturmerkmale am stärksten. Einerseits müssen nicht nur strukturierte Daten, sondern auch unstrukturierte und semi-strukturierte Daten gespeichert werden können. Datenbestände mit Daten unterschiedlicher Struktur werden in diesem Zusammenhang auch als polystrukturiert bezeichnet. Andererseits sollen nicht nur gespeicherte Daten (*Data-at-Rest*), sondern auch Datenströme (*Data-in-Motion*) verarbeitet werden.

Aus diesen primären Einflussfaktoren lassen sich weitere sekundäre ableiten. Werden unstrukturierte Daten wie z. B. Textdokumente gespeichert, so müssen diese auch analysiert werden können, denn eine Speicherung ohne spätere Nutzung ist sinnlos. Die Verarbeitung von Datenströmen erfordert Möglichkeiten, diese in Echtzeit oder wenigstens zeitnah zu verarbeiten, um auf eventuelle Ereignisse reagieren zu können.

Datenquellen

Gegenüber klassischen Data-Warehouse-Architekturen, die in der Regel fokussiert sind auf unternehmensinterne strukturierte Daten, wird eine Big Data-Architektur mit einer Vielzahl weiterer Datenquellen konfrontiert.

Bei den Datenquellen werden bestimmte Eigenschaften unterschieden, die wesentlich die Anforderungen beeinflussen, insbesondere hinsichtlich der Datenintegration und Speicherung. Tabelle 4.2 gibt Beispiele für die im Folgenden beschriebenen Eigenschaften.

Tab. 4.2 Eigenschaften verschiedener Datenquellen

Datenquelle	Herkunft	Vielfalt	Geschwindigkeit
Operative Daten (ERP etc.)	Intern	Strukturiert	Data-at-Rest
Dokumenten-Management	Intern	Unstrukturiert	Data-at-Rest
Web-Server-Logs	Intern	Semi-strukturiert	Data-in-Motion
Netzwerk-Router-Logs	Intern	Semi-strukturiert	Data-in-Motion
Sensor-Zeitreihen	Intern	Semi-strukturiert	Data-in-Motion
Intelligente Stromzähler	Intern	Semi-strukturiert	Data-in-Motion
Nielsen-Daten	Extern	Strukturiert	Data-at-Rest
Internet-Foren	Extern	Unstrukturiert	Data-at-Rest
Blogs	Extern	Unstrukturiert	Data-at-Rest
Micro-Blogs	Extern	Unstrukturiert	Data-in-Motion

Datenquelle Welche Art von Anwendung hat die Daten erzeugt? War es z. B. eine interne Geschäftsanwendung (z. B. ERP-System, CRM-System oder ein Dokumenten-Management-System), eine Datenquelle aus dem Web (z. B. soziale Netzwerke oder Video-Plattformen), Kaufdaten (Wetterdaten des Deutschen Wetterdienstes), etc.?

Herkunft Handelt es sich um interne oder externe Datenquellen? Diese Einordnung der Datenquellen hat eine große Bedeutung, da beispielsweise auf interne Datenquellen mehr Einfluss etwa im Rahmen eines Datenqualitätsmanagements ausgeübt werden kann.

Vielfalt Liegen die Daten strukturiert in einer relationalen Datenbank vor? Oder handelt es sich um semi-strukturierte Daten wie etwa Log-Dateien im JSON²-Format oder gar unstrukturierte Daten beispielsweise in Textdokumenten oder Videos?

Geschwindigkeit Wurden die Daten persistiert (*Data-at-Rest*) und können im Stapel (Batch) verarbeitet werden oder handelt es sich um Datenströme (*Data-in-Motion*) wie beispielsweise die Zeitreihen, die von Sensoren entlang einer Fertigungsstraße geliefert werden?

Datenspeicherung

Die Datenspeicherung muss unterschiedlichen Anforderungen an die Struktur der Daten, die Geschwindigkeit, der Form der Bereitstellung (Einzelsatz vs. viele Daten) und dem Zugriff gerecht werden.

Nachfolgend sind die einzelnen Schichten aufgeführt. Allerdings ist gegenüber klassischen Data-Warehouse-Architekturen hervorzuheben, dass Daten nicht zwingend alle Schichten durchlaufen müssen, da das verarbeitete Datenvolumen bzw. die Anforderungen an die Geschwindigkeit Persistierungen nicht mehrfach oder erst sehr spät erlauben

² Abkürzung für JavaScript Object Notation, Datenformat zum Datenaustausch vergleichbar mit XML.

und zum Teil schon vor der Speicherung erste Analysen auf Datenströmen stattfinden müssen.

Eingangsschicht Diese Schicht ist der sogenannte *Landing Space*, in dem Daten in der Form aufgenommen werden, in der sie ankommen. Daten müssen diese Schicht nicht zwingend durchlaufen, wenn sie in Echtzeit verarbeitet werden sollen oder sie keiner Transformation für die Speicherung in der Rohdaten-Schicht mehr unterzogen werden müssen. In der Regel sind jedoch auch in rein Datei-basierten Speicherformen Verarbeitungen notwendig, wie beispielsweise die Durchführung einer Datenkompression. Im Falle der Echtzeitverarbeitung durchlaufen die Daten stattdessen die Datenstrom-Verarbeitung.

Rohdaten-Schicht In der Rohdaten-Schicht – es wird hier häufig auch vom sogenannten *Data Lake* gesprochen – werden die Daten für die Big Data-Architektur optimiert gespeichert. Wichtig ist, dass es hier keinen Informationsverlust geben darf. Daten, die in einem XML-Format geliefert werden, bleiben in diesem Format. Es werden hier nur Eigenschaften in der Speicherung angepasst, z. B. die Transformation des Kompressionsformats oder aber das Zusammenführen vieler kleiner Input-Dateien. Eine inhaltliche Prüfung der Daten findet nicht statt. Die Rohdaten-Schicht ist am ehesten vergleichbar mit dem Eingangsbereich (Staging-Layer) im klassischen Data Warehouse, der jedoch meist nicht persistent ist. Die Rohdaten-Schicht ist die Quelle für alle weiteren Verarbeitungsschritte. Daten werden dort dauerhaft gespeichert, sofern dies aufgrund rechtlicher Vorgaben möglich ist (vgl. hierzu Abschn. 3.1.5.2, Speichern von Big Data).

Zugriffsschicht In der Zugriffsschicht liegen die Daten transformiert für den Zugriff mit Analysewerkzeugen vor. Dies entspricht beispielsweise den vorausberechneten Sichten (Views) in der Lambda-Architektur. Eine Zugriffsschicht (Output-Layer) findet sich in der Regel auch in klassischen Data-Warehouse-Architekturen wieder.

Sandboxing Sollen neue Datenquellen in eine Big Data-Systemlandschaft integriert werden, treten häufig Schwierigkeiten auf. Da neue Datenquellen meist unbekannt sind, können Fachbereiche nur schwer einschätzen, welche Attribute für Auswertungen relevant sind. Schon für interne Datenquellen liegen meist nur unzureichende Informationen vor und Dokumentationen sind häufig lückenhaft. Die Qualität der neuen Datenquellen lässt sich a priori schwer bewerten. Erst über Analysen kann die Qualität der Daten im Rahmen eines sogenannten Data Profilings ermittelt werden. Hierbei sind dann auch Analysen notwendig, inwiefern die neuen Daten zu den bisherigen passen. Lässt sich also beispielsweise eine eindeutige Beziehung von in einem Internetforum diskutierten Produkten zu den eigenen herstellen? Eine explorative Untersuchung der Daten gibt auch erste Aufschlüsse darüber, welche Transformationsschritte im Einzelnen genau durchgeführt werden müssen, um zu bestimmten Analyseergebnissen zu kommen.

Tab. 4.3 Unterschiede zwischen Schema-on-Read und Schema-on-Write in Anlehnung an White (2012), Abschn. 12.4

	Schema-on-Read	Schema-on-Write
Schemaprüfung	Keine Überprüfung zum Zeitpunkt des Ladens. Die Schemaprüfung findet bei der Abfrage statt.	Zum Zeitpunkt der Beladung. Daten, die nicht schemakonform sind, werden abgewiesen.
Datenbeladung	Geht sehr schnell, da nur ein Kopieren der Daten stattfindet.	Die notwendigen Typprüfungen, die beim Laden stattfinden, benötigen Zeit.
Flexibilität	Hoch, mehrere Schemata für die gleichen Daten sind möglich.	Niedrig, das Schema ist immer einzuhalten.

An dieser Stelle bietet sich der Einsatz von Sandboxing-Funktionalitäten an. Eine Sandbox ist ein abgetrennter Bereich, in dem mit bestehenden oder neuen Daten experimentiert werden kann. Neue Daten können gespeichert und bestehende Daten können beliebig transformiert und mit neuen Daten versuchsweise verknüpft werden. Außerdem können neue Analysemöglichkeiten evaluiert werden. In diesem Bereich können Anwender letztlich neue Szenarien experimentell entwickeln und ggf. verwerfen oder aber später der Allgemeinheit zugänglich machen (vgl. hierzu auch Kimball; 2011, S. 21).

Klassische Data-Warehouse-Ansätze mit komplexen Ladeprozessen und einer strukturierten Speicherung sind für die Verwendung als Sandbox zu unflexibel, denn konventionelle relationale Datenbanken folgen dem *Schema-on-Write-Ansatz*, bei dem das Schema schon beim Schreiben feststeht. Daher muss beim oder vielmehr vor dem Schreiben eine Prüfung der Daten auf Typkonformität stattfinden. Dagegen ist das im Big Data-Kontext weit verbreitete *Schema-on-Read-Konzept* ideal auch für das Sandboxing geeignet. Bei diesem Ansatz werden keine feste Struktur und keine Typdefinitionen vorausgesetzt, die bei der Speicherung berücksichtigt werden müssen. Stattdessen wird eine bestimmte Struktur erst beim Lesen angewendet. Die Unterschiede zwischen den beiden Ansätzen sind in Tab. 4.3 zusammengefasst.

Metadaten-Speicherung Für die Metadaten gibt es einen abgetrennten Speicherbereich. Projekte im Hadoop-Ökosystems, wie Hive oder Oozie, aber auch Werkzeuge aus dem BI-Umfeld, wie ETL- oder Analyse-Werkzeuge, benötigen in der Regel relationale Datenbanksysteme für die Speicherung ihrer Metadaten (Meta-Stores, Repositories). Diese bilden die Basis für das Metadaten-Management.

Die dargestellten Anforderungen für die Speicherung lassen sich in der Praxis nicht in einer Technologie vereinen. Meist kommen mehrere Technologien parallel zum Einsatz, die sich dann in Architektur-Mustern wie der Lambda-Architektur wiederfinden. Zu betonen ist, dass diese Muster darin begründet sind, dass es keine universelle Technologie gibt, die alle Anforderungen hinreichend abdeckt. So bieten Key-Values-Stores, wie beispielsweise *HBase*, extrem schnelle Zugriffsmöglichkeiten auf Daten bei bekannten Zugriffsmustern, die über klar definierte Schlüssel auf definierte Zellen oder Bereiche von

Zellen geht (vgl. George; 2011, Kap. 1). Sie sind jedoch schlecht geeignet für Ad-hoc-Abfragen. Diese können über Echtzeit-Query-Engines (z. B. *Impala*) direkt auf Daten im Hadoop-Dateisystem (*HDFS*) zugreifen. Für die Verarbeitung von Metadaten eignen sich relationale Datenbanksysteme hervorragend, das *HDFS* ist hierfür komplett ungeeignet.

Ressourcenmanagement

Auf einem Big Data-System laufen unterschiedlichste Anwendungen, Prozesse und Applikationen. Das Ressourcenmanagement hat die Aufgabe, die vorhandenen Hardware-Ressourcen nach definierten Regeln zu verteilen. Es soll beispielsweise verhindert werden, dass bestimmte Applikationen, beispielsweise ein komplexer Data-Mining-Prozess, alle Ressourcen bindet und andere Prozesse, z. B. ein Datenintegrationsprozess, nicht mehr ausgeführt werden können. Hierfür muss es möglich sein Gruppen zu definieren, denen Applikationen oder Benutzer zugeordnet werden können. Für diese Gruppen werden Prioritäten und Gewichtungen für die Systemnutzung vergeben. So kann die Auslastung einer Big Data-Umgebung dediziert gesteuert werden (vgl. hierzu Abschn. 4.3.2.1, *YARN*).

Analyse

In der Analyseschicht lassen sich insbesondere folgende Nutzungsarten unterscheiden:

Strukturierte Analysemöglichkeiten Die Speicherung der Daten erfolgt zum Teil oder in Gänze schemalos. Geeignete Schemata für strukturierte Analysen finden sich in der Analyseschicht. Advanced-Analytics-Komponenten können sich dieser Schemata bedienen. Die strukturierten Analysen können im Batch durchgeführt werden, z. B. über *Apache Hive* oder *Pig*, oder über Echtzeit-Query-Engines wie *Impala* oder *Shark*.

Advanced Analytics Zu Advanced Analytics zählen insbesondere automatisierte Analysemethoden etwa aus der Statistik oder dem Data Mining, aber auch semantische Methoden, die z. B. für die Analyse von Texten benötigt werden. Es stehen dafür Statistikbibliotheken, wie *MADlib*, oder Machine-Learning-Werkzeuge, wie *Apache Mahout*, zur Verfügung. Der Bereich Advanced Analytics mit Big Data wird im Abschn. 2.3, Advanced Analytics mit Big Data, ausführlicher behandelt. Auch für semantische Technologien zur Analyse von Texten sei auf das Abschn. 4.4, Big Data-Analyse auf Basis technischer Methoden und Systeme, verwiesen.

Search Insbesondere für unstrukturierte Daten, wie etwa Textdokumente, stehen Komponenten für die Volltextsuche zur Verfügung. Beispiele für Search-Komponenten sind *Solr* und *Elastic Search*, die auf der Programmabibliothek *Apache Lucene* zur Volltextsuche basieren. Lucene erstellt einen Index auf Dokumenten und bietet unterschiedlichste Suchalgorithmen an. Weitere Details zum Thema Suchmaschinen werden in einem eigenen Kapitel dieses Handbuchs erläutert.

Complex Event Processing (CEP) Über CEP werden Analysen auf einer Serie von Ereignissen – also einer speziellen Form eines Datenstroms – ausgeführt. Im Unterschied zu den anderen dargestellten Analysemöglichkeiten müssen die Daten noch nicht gespeichert sein. Wichtig ist, dass Analysen in Echtzeit stattfinden, also während die Ereignisse passieren. In der Regel wird eine überschaubare Anzahl von Ereigniskombinationen analysiert, die kontinuierlich erzeugt werden. Siehe Eckert and Bry (2009) für eine detailliertere Beschreibung.

Präsentation

Die Präsentationsschicht übernimmt primär die Visualisierung der Ergebnisse der analytischen Schicht. Sie kann über klassische BI-Tools umgesetzt werden. Für bestimmte Big Data-Anwendungsfälle sind die klassischen Visualisierungen jedoch häufig nicht ausreichend und müssen mit neuen Visualisierungsmöglichkeiten ergänzt werden. Geht es beispielsweise darum, Beziehungen in sozialen Netzwerken darzustellen, so eignen sich hierfür Graphen besonders gut. Häufig verwendete Begriffe in Texten lassen sich gut über sogenannte *Word Clouds* visualisieren. Werden Datenströme analysiert, so sollten auch Visualisierungen die Ergebnisse in Echtzeit darstellen können.

Data-Lifecycle-Management

Im Rahmen eines Data-Lifecycle-Managements (DLM), das Daten aktiv über ihre Lebensdauer verwaltet, werden folgende Aspekte im thematisiert.

Vernichtung von Daten Aufgrund von Datenschutzvorgaben kann es zum Beispiel notwendig sein, Daten zu löschen. Insbesondere bei Daten mit Personenbezug gibt es hier in vielen Ländern klare rechtliche Vorgaben.

Komprimierung von Daten Aus Performance- und Platzgründen werden Daten in einer Big Data-Umgebung in der Regel mit bestimmten Kompressionsverfahren gespeichert. Es gibt unterschiedliche Kompressionsverfahren, die sich im Wesentlichen im Komprimierungsfaktor, also dem Faktor, um den die Daten verkleinert werden, und der Geschwindigkeit, mit der die Daten komprimiert bzw. dekomprimiert werden können, unterscheiden. Da sich dies in der CPU-Belastung widerspiegelt, wird man für Daten, auf die weniger häufig zugegriffen wird, Verfahren mit einer hohen Kompression (aber eher langsamer Kompressions- bzw. Dekompressionsgeschwindigkeit) und für Daten, die sich häufiger im Zugriff befinden, ein schnelleres Verfahren mit schlechterer Kompressionsrate wählen. Da die Zugriffshäufigkeit häufig mit zunehmendem Alter der Daten abnimmt, ist es sinnvoll, Kompressionsverfahren über die Lebensdauer der Daten zu variieren.

Archivierung Die Archivierung von Daten, also das Auslagern von Daten z. B. auf Bänder, war bisher primär notwendig, um Teile eines Datenbestandes auf die seltener zugegriffen werden, zu reduzieren. Diese Anforderung ist im Big Data-Umfeld nicht mehr notwendig und auf die herkömmliche Weise auch kaum umzusetzen. Ein anderer Grund

kann der Schutz von Daten für den Katastrophenfall sein. Diese Anforderung besteht auch weiterhin. In der Regel erfolgt die Lösung über ein zweites Big Data-System, welches stärker auf Speicherung und weniger auf Rechenleistung ausgelegt ist.

Datenintegration

Im Rahmen klassischer Data-Warehouse-Architekturen wird der Begriff der Datenintegration häufig synonym zu ETL verwendet, wie im Abschn. 4.1.2 über die Grenzen klassischer BI-Lösungen dargestellt. Für Big Data-Architekturen wird die Datenintegration weiter gefasst und geht über die klassischen ETL-Funktionalitäten hinaus. Insbesondere steht die Datenbeschaffung als Verallgemeinerung der Extraktionsphase viel stärker im Fokus.

Extraktion aus strukturierten Datenquellen Auch wenn bei Big Data häufig die Varianz und Geschwindigkeit im Vordergrund stehen, müssen dennoch auch „klassische“ Datenquellen angebunden werden, also Daten aus z. B. relationalen Datenbanksystemen integriert werden. Hierfür stehen im Hadoop-Umfeld Tools wie *Apache Sqoop* zur Verfügung, können aber durchaus auch Einsatzgebiet für klassische ETL-Tools sein, die hier auch Funktionalitäten mitbringen, um eine Delta-Datenversorgung zu unterstützen oder alte Host-Systeme an eine Big Data-Architektur anzubinden.

Beschaffung externer Daten Abbildung 4.5 zeigt verschiedene Möglichkeiten der Datenbeschaffung aus dem Web.

Bei der Integration von Daten aus Internet-Foren werden die Daten mit einem Web-Crawler abgegriffen, müssen aber in der Regel mit einem Pre-Parsing vorverarbeitet werden, da das Design verschiedener Webseiten oft stark variiert.

Für andere, meist kommerzielle Webseiten oder auch Mikro-Blogging-Anwendungen wie Twitter, gibt es vordefinierte Programmierschnittstellen (Application Programming Interface, API), die anzubinden sind. Da diese Daten permanent generiert werden, wird hier häufig eine Streaming-Komponente zwischengeschaltet, wie z. B. *Apache Flume*.

Bestimmte Daten sind nicht frei verfügbar und müssen über Drittanbieter gekauft werden.

Sind verteilte Quelldaten, wie z. B. Log-Daten von verschiedenen Web-Servern, zu analysieren, so müssen diese verteilten Quellendaten transportiert und zusammengefügt werden. Sie werden laufend generiert, sodass hierfür eine Unterstützung von Data Streams notwendig wird.

Integration von Streams (Echtzeitfähigkeit) Je nach Anwendungsfall können die Anforderungen in Richtung Stream-Verarbeitung steigen und sind durch spezielle massiv-parallele Streaming-Architekturen zu unterstützen, insbesondere wenn eine Realtime-Reaktion auf einzelne Ereignisse oder komplexe Ereignis-Kombinationen notwendig ist (vgl. hierzu CEP). Beispiele für Streaming-Komponenten sind *Storm* und *Apache Spark Streaming*.

Eine weitere wichtige Funktion im Kontext der Datenintegration ist die Transformation der Daten. Diese kann analog zu den Transformationsprozessen in einem klassischen Data-Warehouse stattfinden. Zu den Aufgaben gehört zum einen die Transformation von Daten in ein abfrageorientiertes Schema (z. B. Star-Schema) in der Zugriffsschicht. Zum anderen sind Datenqualitätsüberprüfungen und Aufgaben im Rahmen eines Datenqualitätsmanagements durchzuführen. Grundsätzlich ist hervorzuheben, dass eine Big Data-Architektur im Vergleich zu einem klassischen Data Warehouse zunächst mit weniger Datentransformationen auskommt. Die Rohdaten liegen als solche vor und werden keinen inhaltlichen Verarbeitungsschritten unterworfen.

System-Management

In den Bereich System-Management fallen insbesondere folgende Aufgaben:

- Überwachung (Monitoring) der Systemhardware,
- Analyse der Ressourcennutzung und -auslastung,
- proaktive Health-Checks,
- Überwachung (Monitoring) von Prozessen und Abfragen,
- Konfiguration unterschiedlicher Dienste wie das Log-Management, Event-Triggering, Softwarekonfiguration und -verteilung sowie Backup und Recovery.

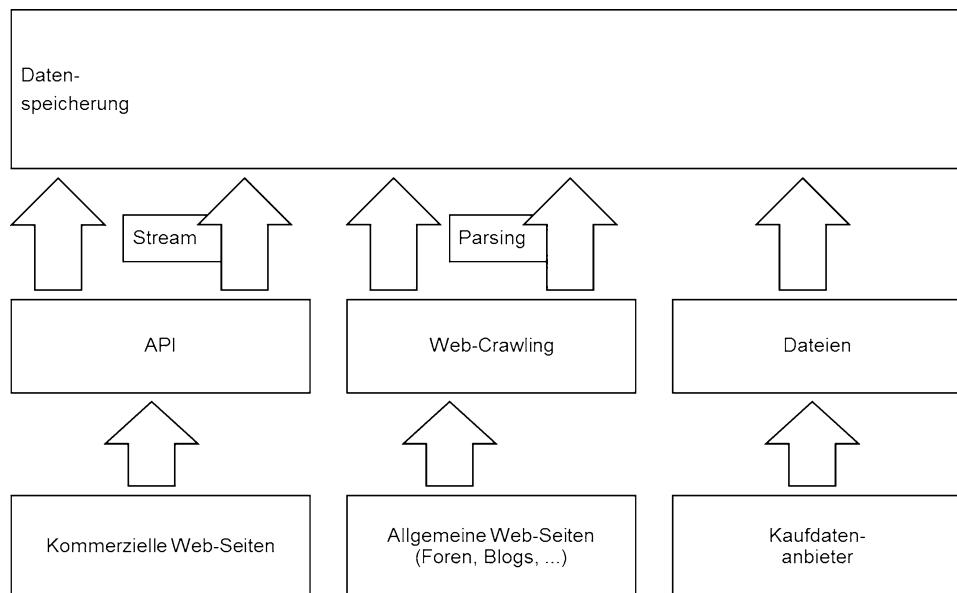


Abb. 4.5 Möglichkeiten der Datenbeschaffung aus dem Web

Metadaten

Big Data-Projekte sind analog zu BI-Projekten Metadaten-getrieben. Da die Vielfalt der Quellen im Big Data-Kontext steigt, ist die Bedeutung eines Metadatenmanagements noch wichtiger geworden. Metadaten lassen sich in die drei Hauptkategorien technische Metadaten, Prozess-Metadaten und Geschäfts-Metadaten unterteilen (Kimball et al.; 2008, S. 116 ff.).

Technische Metadaten Technische Metadaten beschreiben die technische Sicht einer Big Data-Lösung. Beispiele hierfür sind:

- Tabellendefinitionen in einem Datenbank-Repository (Metastore),
- Hive-Schemadefinitionen für HDFS-Dateien,
- Datenprofile,
- in ETL-Tools entwickelte Transformationsschritte,
- Zugriffsberechtigungen eines BI-Tools.

Prozess-Metadaten Hierbei handelt es sich um Daten, die während der Ausführung von Prozessen in einem Big Data-System gespeichert werden, wie etwa die Ausführungszeiten von Map-Reduce-Jobs oder die dynamisch von Berichten generierten HiveSQL-Statements.

Geschäfts-Metadaten Geschäfts-Metadaten reichern die geschäftsorientierte Sicht auf die Big Data-Lösung an. Dies können Geschäftsregeln oder auch Sichten sein. Hilfreich sind hier auch Taxonomien, die Datenobjekte fachlich klassifizieren, z. B., ob es sich um demografische oder geografische Daten handelt (vgl. hierzu auch Mohanty, Jagadeesh, and Srivatsa; 2013, S. 141). Weiterhin können hierunter auch Metadaten gefasst werden, die für eine datenschutzrechtskonforme Verarbeitung notwendig sind wie etwa eine Kennzeichnung personenbezogener Daten, die Angabe besonderer Arten personenbezogener Daten, Speicherungszwecke, Löschungspflichten oder auch Daten, die zur Erfüllung von Betroffenenrechten erforderlich sind wie z. B. bei einem Auskunftsersuchen (vgl. hierzu Abschn. 3.1.6 Betroffenenrechte Datenschutzrechtliche Rahmenbedingungen von Big Data).

Gegenüber klassischen Data-Warehouse-Architekturen gibt es bei einer Big Data-Architektur Unterschiede. Die Schemalosigkeit (siehe hierzu auch Schema-on-Read vs. Schema-on-Write in Tab. 4.3), führt dazu, dass sich Metadaten über die Zeit ändern können (z. B. unterschiedliche Avro-Definitionen) oder aber sogar zum gleichen Zeitpunkt unterschiedliche Metadaten auf den gleichen physischen Daten modelliert sind. Die Logik, um aus unstrukturierten Daten strukturierte Informationen zu erhalten, wird insbesondere für unstrukturierte Quelldaten komplexer. Wichtige Metadaten sind verstärkt in den oberen Systemschichten zu finden, insbesondere der Analyseschicht. Zu beobachten ist auch, dass die Vielfalt der Metadaten-Speicher (engl. Meta Stores) mit der Vielfalt der Tools und Projekte eher zunimmt und es kaum Vereinheitlichungen und Standards gibt.

Datenqualität

Für die strukturierten Datenquellen lassen sich Datenqualitätseigenschaften definieren, wie etwa Vollständigkeit, Korrektheit, Konsistenz (siehe Pipino, Lee, and Wang (2002) für weitere Dimensionen der Datenqualität). Die Eigenschaften sind messbar und können innerhalb eines Datenqualitätsmonitorings dargestellt werden. Innerhalb der Datenintegrationsprozesse kann die Datenqualität durch Cleansing- und Enrichment-Prozesse (deutsch: Bereinigung und Anreicherungsprozesse) erhöht werden. Für strukturierte Daten innerhalb einer Big Data-Lösung greifen diese Ansätze auch analog. Wobei hier wie im Data Warehousing auch, die Datenqualität vorzugsweise in den anliefernden Systemen verbessert werden sollte. Wie verhält es sich jedoch für semi- oder unstrukturierte Daten aus Texten, Web-Einträgen, Sensordaten? Sind die Ansätze für strukturierte Daten innerhalb Big Data eins zu eins anwendbar?

Um diese Frage zu beantworten, ist es wichtig zu berücksichtigen, dass Big Data-Lösungen fokussiert sind auf die Sammlung von vielen Daten und Geschwindigkeit in der Beschaffung. Das höhere Volumen und die Varianz der Daten bedeuten auch mehr fehlerhafte Daten. Wie oben dargestellt, werden in einer Big Data-Lösung immer die Originaldaten unverändert gespeichert, d. h. hier greifen keine Cleansing-Prozesse, die zum Verwerfen von Informationen führen könnten. Qualitätsprüfungen greifen später, zum Teil erst in der Analyse. Für bestimmte analytische Fragestellungen ist ein gewisser Prozentsatz von fehlerhaften Daten sogar tolerierbar, für andere nicht, es sollte hier also abhängig von der Art der Daten ein Vorgehen gewählt werden.

Auch für Aspekte der Datenqualität ist ein durchgängiges Metadaten-Konzept sehr wichtig. Für die Anwender muss transparent sein, woher welche Daten kommen und wie sie von der Quelle bis zur Analyse transformiert worden sind (Data Lineage). Bestimmte Datenqualitätsprobleme werden auch erst durch Analysen und durch die Kombination vieler Quellinformationen sichtbar. Beispielsweise wäre eine Situation zu hinterfragen, in der Kundenäußerungen ein negatives Sentiment haben, dieser Kunde jedoch einen verhältnismäßig hohen Umsatz generiert. Bei den Metadaten sind auch Informationen wichtig, die letztendlich zu qualitativ hochwertigen strukturierten Informationen führen, z. B. Lexika für die Analyse von Texten (Mohanty, Jagadeesh, and Srivatsa; 2013, S. 142 ff.).

Sicherheit

Das Hauptziel einer Big Data-Umgebung besteht darin, große Datenmengen zugreifbar und analysierbar zu machen. Sicherheitsaspekte laufen diesem Ziel oftmals entgegen, da Zugriffe auf bestimmte Daten beschränkt werden oder Daten aufgrund von Datenschutzanforderungen vergröbert oder maskiert werden müssen (Anahory and Murray; 1997, S. 188 ff.).

Für ein umfassendes Sicherheitskonzept sind verschiedene Aspekte zu betrachten:

Zugriffsbeschränkung der Umgebung Die Zugriffsbeschränkung wird über eine Authentifizierung gelöst. Hierfür werden Authentifizierungsdienste genutzt, wie z. B. Kerberos. Bei massiv parallelen Umgebungen, die aus vielen Rechnern bestehen (Hadoop-

Clustern), kann es sinnvoll sein, die Authentifizierung zu zentralisieren und über ein Gateway einen einzigen Authentifizierungspunkt anzubieten. Hierüber wird dann auch die interne Netzwerktopologie eines Hadoop-Clusters nicht transparent nach außen sichtbar. Ein Beispiel für ein derartiges Gateway ist *Apache Knox*.

Verschlüsselung Die Daten lassen sich durch Verschlüsselung vor unberechtigter Sichtbarkeit schützen. Ein weiteres Verfahren ist die Maskierung von besonders sensiblen Daten, beispielsweise von Kreditkartennummern. Hierbei werden beispielsweise Teile der Nummern ausgeixt.

Anonymisierung Bei der Anonymisierung ist das Ziel, Daten so zu verändern, dass kein Personenbezug mehr hergestellt werden kann. Praktisch geschieht dies dadurch, dass beispielsweise Kundennummer und Kundenname aus den Daten entfernt werden. Zu beachten ist hierbei, dass Personen auch über Merkmalskombinationen wie Geburtsdatum, Postleitzahl und Geschlecht identifizierbar sein können. Falls solche Merkmale vorhanden sind, sind diese zu gegebenenfalls zu vergröbern, sofern sich auf einzelne Personen mit hoher Wahrscheinlichkeit rückschließen lässt (vgl. hierzu auch entsprechende Ausführungen im Abschn. 3.1.3.1, Anonymisierung).

Autorisierung Über die Authentifizierung wird geregelt, wer Zugriff auf das Big Data-System haben darf, über die Autorisierung wird geregelt, was ein authentifizierter Benutzer machen darf. So kann über die Autorisierung gesteuert werden, welche Dateien, Tabellen oder Datenbanken ein Benutzer sehen und was er mit diesen machen darf, also z. B. ob er Daten nur lesen oder auch löschen kann. Für die Klassifikation der Daten sind unterschiedlichste Ansätze möglich, in der Regel orientieren sie sich an der Sensitivität der Daten. Die Autorisierung kann etwa über *Apache Sentry* implementiert werden.

Auditing Auditing kann notwendig sein aufgrund von Vorgaben in der Organisation oder aufgrund rechtlicher Rahmenbedingungen. Über die sogenannte Auditing-Funktionalität kann nachvollzogen werden, wer auf welche Daten zugegriffen hat. Diese Funktionalität dokumentiert die Nutzung von Daten und kann helfen, Lücken im aufgesetzten Sicherheitskonzept zu entdecken.

Datenübermittlung Bei der Datenübermittlung ist zu berücksichtigen, wie die Verschlüsselung der Daten stattfindet, dies ist insbesondere bei einem Transfer über öffentliche Netze wichtig.

4.2.2.2 Erweiterung einer Data-Warehouse-Architektur mit Big Data-Technologien

Nachfolgend wird aufgezeigt, wie sich neue Architekturen in bestehende Landschaften eingliedern und welche Aufgaben sie dort übernehmen können als Ergänzung im Rahmen bestehender Data Warehouses.

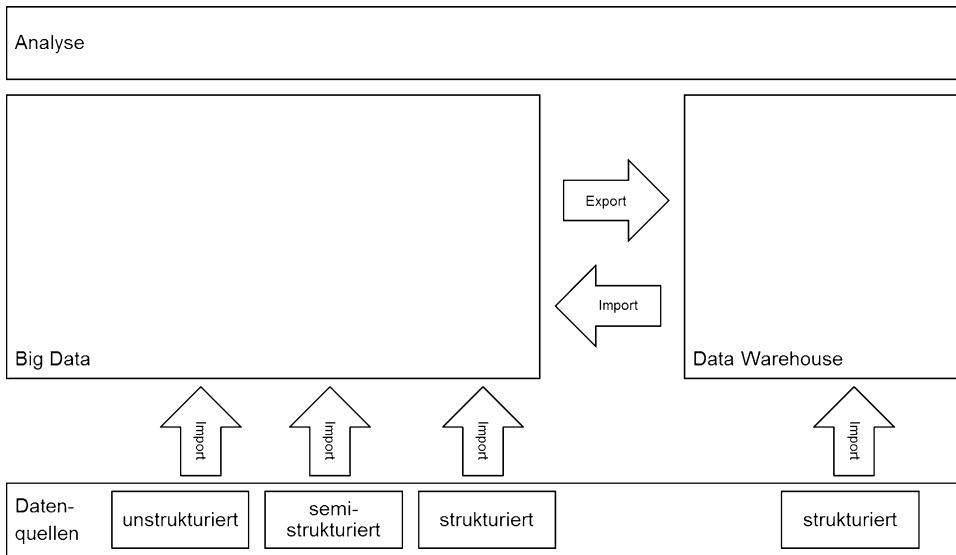


Abb. 4.6 Integration von BI-Systemen und Big Data-Technologie

Die Ergänzung einer bestehenden Data-Warehouse-Landschaft mit Big Data-Komponenten, insbesondere dem Hadoop-Ökosystem oder kurz Hadoop, bietet sich für folgende Einsatzgebiete an (Russom; 2013, S. 9):

Data Staging Hadoop ist eine ideale Staging-Plattform insbesondere für polystrukturierte Daten. Verarbeitungsschritte, die hier stattfinden, können den Workload auf dem relationalen Datenbank-Management-System (RDBMS) reduzieren. Es bietet sich als schnellere Option für die Aufnahme von Daten an, insbesondere von Dateien in unterschiedlichsten Formaten, wie etwa Log-Dateien, JSON-Objekte, XML-Dokumente. Das Staging in Hadoop erfolgt persistent, womit Originaldaten in ihren originären Strukturen erhalten bleiben. Der Vorteil eines solchen Konzeptes ist, dass somit auch zukünftige Analysetechnologien und Fragestellungen bedient werden können.

Datenarchivierung Hadoop kann auch für die Archivierung genutzt werden. Erfolgt das Staging auch über Hadoop, so muss innerhalb des Data Warehouses nur ein Housekeeping-Prozess implementiert werden, der Daten löscht.

Polystrukturierte Daten Im Hadoop-System können polystrukturierte Daten abgelegt werden, ohne dass sie vorher in komplexen Transformationsprozessen aufbereitet werden müssen. Hadoop kann hier also als ergänzende Persistenzschicht zur Speicherung dieser Daten dienen.

Datentransformationen Es werden komplexe Verarbeitungsprozesse auf eine Big Data-Plattform ausgelagert. Man spricht hierbei auch vom sogenannten Prozess-Off-Loading. Daten müssen hierzu aus dem Data Warehouse entladen werden, falls diese nicht über ein Big Data-Plattform vorverarbeitet werden (Staging). Die Auslagerung hat den Vorteil, dass die Verarbeitung auf Hadoop günstiger und oft effizienter durchgeführt werden kann als auf traditionellen Data-Warehouse-Plattformen.

Advanced Analytics Im Bereich Advanced Analytics ist SQL nicht der einzige Ansatz und viele Werkzeuge, wie beispielsweise *Apache Mahout*, sind nicht für RDBMS verfügbar. Weiterhin werden Advanced-Analytics-Methoden häufig direkt auf Rohdaten ausgeführt (vgl. hierzu auch Abschn. 2.3).

4.2.3 Zusammenfassung und Ausblick

Die Datenvielfalt und Geschwindigkeit der Daten sind wesentliche Herausforderungen für Big Data-Lösungen, die maßgeblich die Architekturanforderungen gegenüber BI-Lösungen beeinflussen.

Der *Schema-on-Read-Ansatz* bringt eine viel höhere Agilität und Flexibilität mit sich, verglichen mit dem Korsett der starren Modellierung in relationalen Datenbankmanagementsystemen. Mit dem Ansatz gehen jedoch gleichzeitig größere Herausforderungen an das Metadaten- und Datenqualitätsmanagement einher.

Big Data Analytics muss sowohl den Anforderungen aus BI-Lösungen mit der Analyse strukturierter Informationen als auch den unstrukturierten Daten und den Analysen auf Datenströmen (*Data-in-Motion*) gerecht werden.

Der Stand der Technik, macht es inzwischen möglich, alle für ein Unternehmen relevanten externen und internen Daten dauerhaft zu speichern. Dies birgt auch Risiken, insbesondere in Richtung Datenschutz und Datensicherheit. Gerade deshalb müssen in Big Data-Lösungen diese Aspekte hinreichend Berücksichtigung finden.

Es kann davon ausgegangen werden, dass Big Data-Lösungen zunächst häufig parallel zu bestehenden IT-Landschaften aufgebaut werden. Die Rolle als zentraler Unternehmensdatenspeicher wird sicher zunehmen und mit weiteren Software-Evolutionen werden auch immer mehr BI-Anwendungen in Big Data-Lösungen aufgehen.

Technologisch greifen immer mehr Konzepte die Real-Time-Anforderungen besser unterstützen und vom klassischen Map-Reduce-Programmiermodell weggehen. Zu nennen sind hier z. B. *Apache Spark*. Auch andere Projekte, wie die Machine-Learning-Bibliothek *Apache Mahout* wollen zukünftig die Spark-Infrastruktur nutzen.

4.3 IT-Infrastrukturen für Big Data

Gernot Fels und Fritz Schinkel

4.3.1 Herausforderungen an die Infrastruktur

Eine der wesentlichen Herausforderungen bei Big Data ist die Speicherung und Verarbeitung immer größer werdender Datenmengen zu überschaubaren Kosten. Das exponentielle Wachstum der Speicherkapazität von Plattspeichersystemen ermöglicht zwar die Speicherung großer Datenmengen als Basis für analytische Fragestellungen. Allerdings vergrößern sich bei zunehmenden Datenmengen auch die erforderlichen Zeiten zur Bearbeitung dieser Fragestellungen, insbesondere wenn man traditionelle Infrastrukturkonzepte in Erwägung zieht. Die gewünschten Antwortzeiten oder sogar eventuelle Echtzeitanforderungen können dann nicht mehr erfüllt werden.

Um Rechenleistung zu steigern, kommen auf der Serverseite generell die vertikale und die horizontale Skalierung in Betracht.

Bei der vertikalen Skalierung (Scale-up) geht es darum, einen Server mit Prozessoren, Hauptspeicherkapazität und Plattspeicher-Konnektivität aufzurüsten, um ihn leistungsfähiger zu machen. Doch egal wie leistungsstark ein Server auch sein mag, für jede seiner Komponenten gibt es eine Obergrenze, die nicht überschritten werden kann. Angesichts des wachsenden Datenvolumens werden diese Obergrenzen früher oder später zum Problem. Sicher werden diese Obergrenzen im Laufe der Zeit weiter nach oben verschoben, aber das Gesamtvolumen der Daten wird in derselben Zeit noch schneller ansteigen.

Bei der horizontalen Skalierung (Scale-Out) wird die Verarbeitung auf mehrere Rechner verteilt, welche sich die Daten vom Speichersystem holen und die Ergebnisse in der Regel auch wieder dort ablegen müssen. Da mehrere Server gleichzeitig auf den Datenbestand zugreifen, werden die Speicherverbindungen zur entscheidenden Schwachstelle. Gleichzeitig steigt der Koordinationsaufwand für den Zugriff auf gemeinsam genutzte Daten mit der Anzahl der verwendeten Server.

Folglich steigen die Zugriffs- und Bearbeitungszeiten mit dem Wachstum der Daten. Ergebnisse werden je nach Einsatzfall erst dann geliefert, wenn sie nicht mehr relevant sind. Das ist im praktischen Einsatz natürlich inakzeptabel.

Neben ruhenden Datenbeständen (Data at Rest), die nur darauf warten, verarbeitet zu werden, spielen bei Big Data auch im Fluss befindliche Daten (Data in Motion) eine wichtige Rolle. Es handelt sich dabei um große Datenströme, die gegebenenfalls kontinuierlich und in hoher Frequenz erzeugt werden, und auch in Echtzeit verarbeitet werden müssen. Dabei gilt es, diese Datenströme geschickt auf mehrere Verarbeitungseinheiten verteilen, damit die Ergebnisse auch in Echtzeit zur Verfügung stehen.

Da wir also schnell an die Grenzen der klassischen Lösungskonzepte stoßen, sei es bei der Bearbeitung großer Datenbestände sowie auch bei der Bearbeitung großer Datenströme, sind neue Ansätze erforderlich, die es ermöglichen, auch bei steigendem Datenvolumen mit den bestehenden Anforderungen Schritt zu halten.

4.3.2 Verteilte Parallelverarbeitung großer Datenbestände

Konstante Verarbeitungszeit bei steigendem Datenvolumen lässt sich über verteilte Parallelverarbeitung erreichen. Dabei werden die Datenmengen auf die lokalen Platten vieler Knoten eines Rechner-Clusters verteilt. Die Verarbeitung der Daten wird jeweils auf diejenigen Rechnerknoten verlagert, auf denen die Daten liegen. Somit sind auch gleichzeitige Datenzugriffe der einzelnen Rechnerknoten völlig unabhängig voneinander, was folglich zu keinerlei gegenseitiger Beeinträchtigung führt. Man spricht in diesem Zusammenhang auch von einer „Shared Nothing“ Architektur.

Bei steigendem Datenvolumen werden bei Bedarf weitere Rechnerknoten hinzugefügt (horizontale Skalierung) und bei der Verteilung der Daten mit berücksichtigt. Die Unabhängigkeit der einzelnen Rechnerknoten garantiert eine lineare Skalierbarkeit, die praktisch keine Grenzen kennt.

Um die Gesamtkonfiguration fehlertolerant zu machen, werden die Daten auf mehrere Rechnerknoten repliziert und damit redundant gehalten. Fällt ein Server aus, kann die entsprechende Aufgabe auf einem Server mit einer Datenkopie fortgeführt werden. Bei Datenänderungen müssen alle Replikate berücksichtigt werden; ansonsten gilt das System als inkonsistent.

War diese Art der Parallelisierung bislang dem High Performance Computing mit spezialisierter Hard- und Software und hohen Kosten vorbehalten, so stehen inzwischen Softwaresysteme zur Verfügung, welche eine solche Parallelverarbeitung auf einen Cluster gewöhnlicher Standardserver mit lokalem Plattspeicher umsetzen können.

4.3.2.1 Apache Hadoop

Als Standard für Big Data und verteilte Parallelverarbeitung hat sich im Markt Hadoop etabliert. Hadoop ist ein Projekt der Apache Software Foundation und ein in der Sprache Java programmiertes Open-Source-Framework für Batchbetrieb. Es lässt sich horizontal von wenigen auf mehrere tausend Serverknoten skalieren, toleriert Serverausfälle, die in großen Server-Farmen als „Normalzustand“ anzusehen sind, und sorgt so für stabile Speicher- und Analyseprozesse.

Hadoop ist frei verfügbar unter der Apache-Lizenz. Darüber hinaus gibt es verschiedene Hadoop-Distributionen mit zusätzlichen Softwareprodukten und Serviceleistungen.

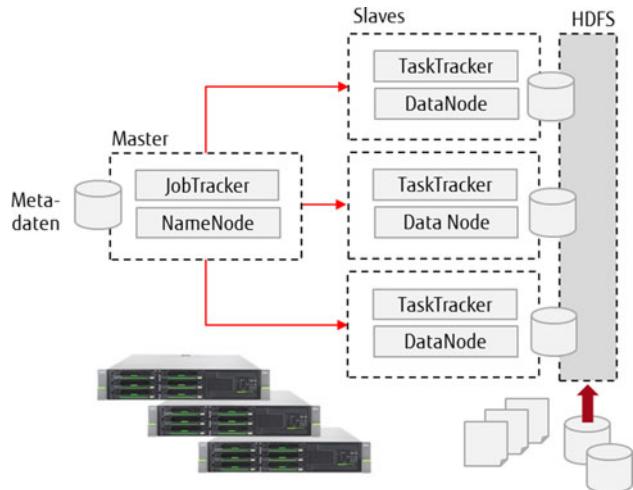
Die Hauptbestandteile von Hadoop sind das Hadoop MapReduce-Framework und das Hadoop Distributed File System (HDFS), die nachfolgend etwas näher beleuchtet werden.

Hadoop Distributed File System (HDFS)

HDFS ist ein verteiltes Dateisystem, welches insbesondere für die Verarbeitung hoher Volumina und hohe Verfügbarkeit konzipiert und optimiert ist.

Die Daten sind über die sogenannten DataNodes verteilt. Dabei handelt es sich typischerweise um Industrie-Standard-Server, die persistente Daten auf ihren lokalen Platten ablegen. Typischerweise ist die Datenhaltung redundant ausgelegt; standardmäßig ist jeder Datenblock dreimal vorhanden. Neben dem Primärblock existieren eine Kopie typischer-

Abb. 4.7 Hadoop Distributed File System und MapReduce
(Quelle: Fujitsu)



weise auf einem zweiten Server innerhalb desselben Server-Racks und eine zusätzliche Kopie in einem anderen (entfernten) Rack. Um die Verfügbarkeit zu erhöhen, lassen sich die Daten auch auf unterschiedliche Lokationen verteilen. Gegen logische Fehler beim Kopieren oder Löschen von Daten hilft das natürlich nicht; hier sind zusätzliche Datensicherungsverfahren anzuwenden.

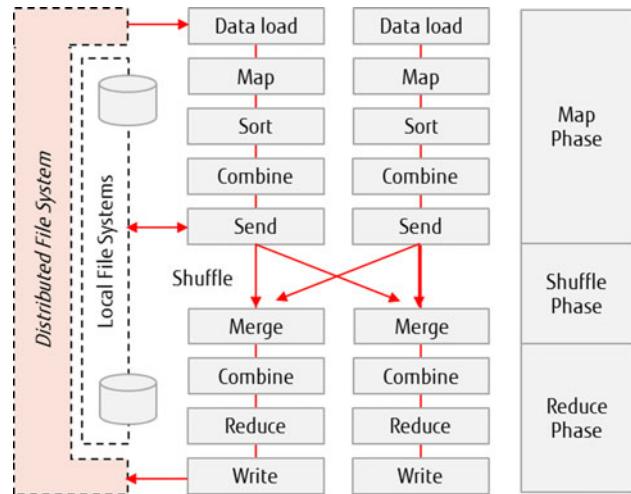
Die zentrale Komponente im HDFS ist der NameNode. Er verwaltet die Metadaten, weiß also jederzeit Bescheid, welche Datenblöcke zu welchen Dateien gehören, wo die Datenblöcke liegen und wo welche Speicherkapazitäten belegt sind. Über periodisch übermittelte Signale kann der NameNode feststellen, ob die DataNodes noch funktionsfähig sind. Anhand ausbleibender Signale erkennt der NameNode den Ausfall von DataNodes, entfernt ausgefallene DataNodes aus dem Hadoop-Cluster und versucht jederzeit, die Datenlast gleichmäßig über die zur Verfügung stehenden DataNodes zu verteilen. Und er sorgt dafür, dass stets die festgelegte Anzahl von Datenblockkopien zur Verfügung steht.

Um sich gegen einen Ausfall des NameNode abzusichern, werden zwei NameNodes auf unterschiedlichen Maschinen im Cluster aufgesetzt. Dabei ist zu jeder Zeit einer aktiv und der andere passiv (Hot Stand-by). Änderungen der Daten, sowie ggf. der Metadaten, werden auf den NameNodes protokolliert. Die Synchronisation der Journaldaten kann über einen Shared-Storage (NAS) oder Quorum-basiert über mehrere Rechnerknoten erfolgen. Da immer beide NameNodes von den DataNodes direkt mit aktuellen Blockinformationen und Heartbeat-Signalen versorgt werden, kann bei Ausfall des aktiven NameNode sehr schnell automatisch umgeschaltet werden.

Hadoop MapReduce

MapReduce ist ein Framework, welches ein definiertes Problem in eine Vielzahl sinnvoller Teilaufgaben, sogenannte Map-Tasks zerlegt, diese über das Netz auf eine Reihe von Rechnerknoten zur parallelen Bearbeitung verteilt, Zwischenergebnisse zwischen den

Abb. 4.8 MapReduce-Ablaufschema (Quelle: Fujitsu)



Knoten austauscht (Shuffling) und danach die Ergebnisse der einzelnen Bearbeitungsschritte wieder zusammenfasst, um ein Endergebnis zurückzuliefern (Reduce-Tasks). Da die Bearbeitungsprozesse zu den zu verarbeitenden Daten bewegt werden und nicht umgekehrt, ist es möglich, in der Map-Phase die I/O-Aktivität zu parallelisieren und die Netzbelastrung fast vollständig zu vermeiden. Um auch in der Shuffling-Phase Skalierungsengpässe zu vermeiden, ist ein sich nicht blockierendes, leistungsfähiges Switched Network zwischen den Rechnerknoten erforderlich.

Optional können Zwischenergebnisse aus der Map-Phase und der Shuffle-Phase aggregiert werden (Combine), um die Datenmenge, die von den Map-Tasks an die Reduce-Tasks übergeben werden müssen, gering zu halten.

Während sich die Eingangsdaten für MapReduce wie auch die Endergebnisse im HDFS befinden, werden Zwischenergebnisse in den lokalen Dateisystemen der DataNodes hinterlegt.

Das MapReduce-Framework übernimmt dabei die gesamte Ablaufsteuerung. Wie HDFS arbeitet auch MapReduce nach dem Master-Slave-Prinzip. Der Master, der sogenannte JobTracker, teilt einen Job in Teilaufgaben auf, verteilt diese als Tasks an die sogenannten TaskTracker, die die Slaves darstellen. Die TaskTracker laufen normalerweise auf den Rechnerknoten, auf denen sich die Daten für die Map-Tasks befinden.

Sollten manche Rechnerknoten bereits stark überlastet sein, wird ein Knoten ausgewählt, der möglichst kurze Wege zu den Daten hat, also bevorzugt ein Knoten im selben Server-Rack. Der JobTracker steht ständig mit den Slaves in Kontakt, überwacht die Slaves und sorgt dafür, dass unterbrochene bzw. abgebrochene Tasks erneut ausgeführt werden.

Meldet eine Task lange Zeit keinen Fortschritt oder fällt ein Knoten ganz aus, so werden die noch nicht beendeten Tasks auf einem anderen Server neu gestartet, in der Regel auf

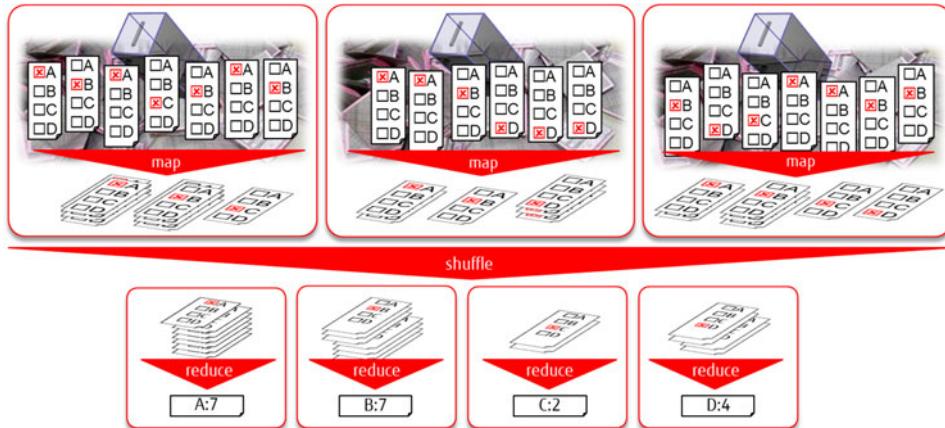


Abb. 4.9 Auszählen von Stimmen mittels MapReduce (Quelle: Fujitsu)

einem Rechnerknoten, auf dem eine Kopie der betreffenden Daten vorhanden ist. Wenn eine Task sehr langsam läuft, kann der JobTracker den Auftrag auf einem anderen Server noch einmal starten (spekulativer Ausführung), um den Gesamtauftrag sicher zu erfüllen.

Die einzige Schwachstelle ist der JobTracker selbst, der einen Single Point of Failure darstellt. Dieser Rechner sollte deshalb in sich möglichst viele redundante Komponenten enthalten, um die Ausfallwahrscheinlichkeit so gering wie möglich zu halten.

Die fachliche Aufgabenstellung einschließlich der Interpretation der Daten wird in Form der Map- und Reduce-Funktionen eingebracht. Die problembezogene Programmierung ist äußerst flexibel; sie kann sich auf die beiden Funktionen Map und Reduce beschränken und setzt damit kein tiefes Cluster-Know-How voraus. Zur Optimierung gibt es weitere Schnittstellen, um bspw. die Zwischenergebnisse auf einem Knoten zu verdichten oder das Shuffling zu beeinflussen.

MapReduce wird zur Ausführung von Business Analytics-Abfragen angewendet, wird aber auch genutzt, um Daten erst in eine für Analyseverfahren optimierte Form zu bringen.

Beispiel: Auszählen von Stimmen nach einer Wahl mittels MapReduce

An einer Wahl nehmen die Parteien A, B, C und D teil. Die Stimmzettel befinden sich in 3 Wahlurnen. Jede Map Task verarbeitet eine Urne und erzeugt ein Key-Value-Paar (P,1) für jeden Stimmzettel mit einer Stimme für Partei P. Alle Paare mit gleichem Schlüssel werden in das für die Partei P vorgesehene Fach, eine sogenannte Partition gelegt. Das Shuffling übergibt den Inhalt dieser Fächer an die Stimmenzähler für jede Partei, in unserem Falle die Reduce Tasks. Diese zählen die Key-Value-Paare aus den einzelnen Fächern und liefern jeweils als Endresultat die Partei P und die Anzahl der gezählten Stimmen.

Optional kann das Zählen der in einem Fach abgelegten Stimmzettel bereits lokal erfolgen (Combine), sodass in der Shuffle-Phase nur die Ergebnisse zu den einzelnen Fächern an die Stimmenzähler weitergereicht werden müssen. Die Teilsummen werden von den

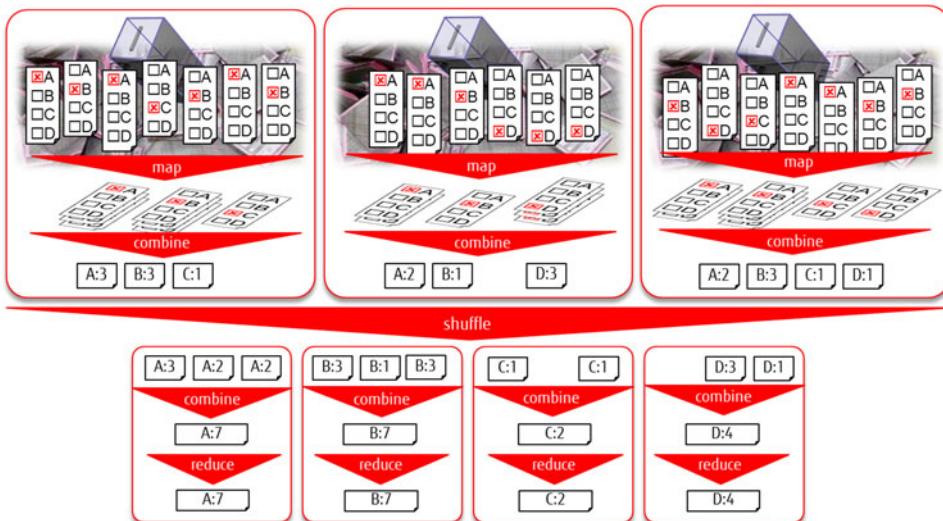


Abb. 4.10 Auszählen von Stimmen mittels MapReduce und Combine (Quelle: Fujitsu)

Reduce-Tasks übernommen und zu einer Gesamtsumme der jeweiligen Partei verdichtet (Combine). Somit beschränkt sich die verbleibende Aufgabe der Reduce Tasks auf das Übernehmen dieser Ergebnisse.

YARN (Yet Another Resource Negotiator)

YARN und MapReduce Version 2 sind Weiterentwicklungen des MapReduce-Frameworks, welche gegenüber der ursprünglichen MapReduce-Implementierung eine Reihe attraktiver Fortschritte mit sich bringen.

Bei YARN werden die beiden Aufgaben des JobTracker, Ressourcenverwaltung und Applikationssteuerung, durch zwei voneinander getrennte Instanzen, den ResourceManager und den ApplicationMaster erledigt. Nur der sehr schlanke Ressourcen-Manager bleibt als zentrale Instanz vorhanden. Die gesamte Job-Planung wird dagegen an die ApplicationMaster delegiert, die auf jedem Clusterknoten ablaufen können. Pro Job wird ein ApplicationMaster dynamisch auf einem der Slave-Knoten erzeugt und steht exklusiv für diesen Jobauftrag zur Verfügung. Das klassische MapReduce-Muster ist dann nur noch eine Ausprägung der parallelen Ablaufsteuerung; andere Strategien können durch andere ApplicationMaster festgelegt werden und gleichzeitig in demselben Cluster angewandt werden. Der Ressourcen-Manager sichert den jeweiligen Zustand seiner Aufträge und ermöglicht damit eine Wiederherstellung. Die verteilte Applikationssteuerung erhöht den Grad der Parallelisierung, beseitigt Engpässe und ermöglicht Cluster mit zehntausenden von Rechnerknoten.

Natürlich gehen die Entwicklungen weiter. Kaum hat YARN Einzug in den Markt gehalten, stehen bereits einige ApplicationMaster für dedizierte Einsatzfälle (Typen von

YARN-Applikationen) zur Verfügung. Neben dem ApplicationMaster für MapReduce ist Apache Tez zu nennen. Apache Tez bietet gegenüber dem starren MapReduce reichhaltigere Möglichkeiten, den Fluss der Datenverarbeitung festzulegen. Für komplexe Abfragen kann damit die Effizienz gegenüber einer Abbildung auf mehrere MapReduce-Tasks erheblich gesteigert werden. Einen ähnlichen Ansatz verfolgt das Stratosphere PACT Projekt.

MapReduce und Shared Storage

Wie bereits ausführlich erläutert, erstreckt sich in einem Hadoop-Cluster das verteilte Dateisystem üblicherweise über die lokalen Platten der beteiligten Rechnerknoten. Es gibt aber auch Einsatzszenarien, bei denen es sich anbietet, das Dateisystem auf ein von allen Rechnerknoten gemeinsam benutztes Speichersystem (Shared Storage) zu verlagern.

Eine derartige Lösung hat durchaus einige Vorteile. Die verfügbare Kapazität eines Shared Storage kann insgesamt besser genutzt werden als die Kapazitäten der einzelnen Rechnerknoten. Durch Nutzung der in solchen Speichersystemen von Hause aus implementierten Verfügbarkeitskonzepten, wie bspw. unterschiedliche RAID-Levels, muss weniger Speicherkapazität für Datenreplikate reserviert werden. Präventive Wartungsmaßnahmen für den Hadoop-Cluster können besser geplant werden, und das oft doch etwas schwierige Ausbalancieren des Verhältnisses aus Rechenleistung und Plattenkapazität bei den einzelnen Knoten des Clusters entfällt.

Trotz aller genannten Vorteile, darf man nicht verschweigen, dass beim parallelen Zugriff auf ein gemeinsam benutztes Speichersystem durch mehrere Rechnerknoten die Skalierbarkeit äußerst begrenzt ist, und sich daher dieser Lösungsansatz nur für kleinere, überschaubare Hadoop-Cluster (mit einer Knotenzahl im maximal zweistelligen Bereich) eignen kann.

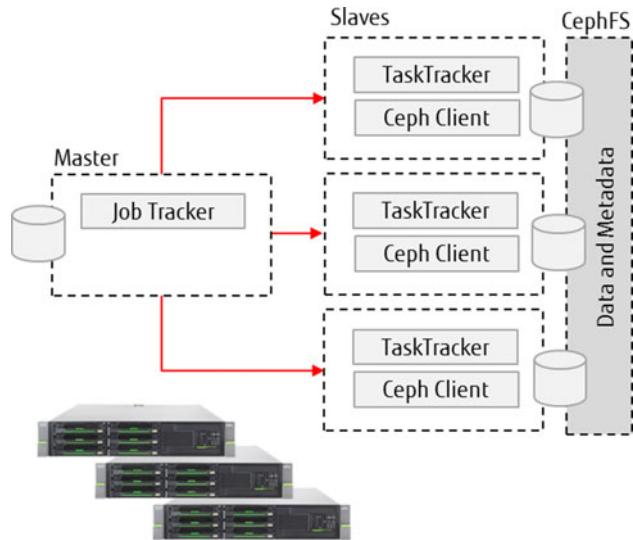
MapReduce und CephFS

HDFS ist ein verteiltes Dateisystem, das für serielle Verarbeitung optimiert ist. Es handelt sich um kein Standard-POSIX-Dateisystem, welches für jede Applikation und jeden Einsatzzweck verwendbar ist.

Darüber hinaus werden im HDFS sämtliche Metadaten von dem im Prinzip nur einmal für den gesamten Hadoop-Cluster vorhandenen (aktiven) NameNode verwaltet. Viele gleichzeitige Veränderungen der Metadaten, bspw. wenn viele Dateien geöffnet oder geschlossen werden, können daher zu Performance-Engpässen führen. Da man die Metadaten ständig benötigt, werden sie resident im Hauptspeicher gehalten, was die Anzahl der Dateien im Dateisystem von vornherein begrenzt. Ob diese Begrenzung nun tatsächlich zum Problem werden kann, hängt vom jeweiligen Anwendungsfall ab. Fakt ist aber, dass die Zugriffszeiten zu den Metadaten entscheidend für die Leistung des Gesamtsystems sind.

MapReduce lässt sich nicht nur mit dem HDFS kombinieren, sondern auch mit anderen Dateisystemen wie bspw. mit dem verteilten Dateisystem CephFS (Ceph File System). Dieses Dateisystem verwaltet seine Metadaten nicht zentral, sondern verteilt diese auf be-

Abb. 4.11 MapReduce und CephFS (Quelle: Fujitsu)



liebig viele Rechnerknoten im Cluster. Damit wird eine hohe Verfügbarkeit der Metadaten erreicht, und man vermeidet Engpässe, die durch viele Zugriffe auf ein zentrales System entstehen können. Bevor die steigende Anzahl von Dateien oder Operationen auf Metadaten zu einem Engpass führt, kann der Metadaten-Cluster horizontal skaliert werden. Eine Begrenzung nach oben gibt es praktisch nicht. Dadurch können die bei HDFS zumindest theoretisch vorhandenen Grenzen der Skalierung umgangen werden.

Da CephFS universell einsetzbar ist, können durchaus manche Datenquellen ebenfalls über CephFS organisiert sein, was die Ladevorgänge erheblich vereinfacht. Lasttests zeigen zwar, dass MapReduce im Zusammenspiel mit CephFS im Vergleich zu HDFS leichte Geschwindigkeitsnachteile hat; die Tatsache, dass man CephFS als Plattform für einige Datenquellen nutzen kann, wird aber je nach Einsatzfall diese Nachteile mehr als nur kompensieren.

Wie bei HDFS werden auch im Falle CephFS die MapReduce Tasks auf denselben Rechnerknoten zum Ablauf gebracht, auf denen sich die Daten befinden.

Neben CephFS gibt es weitere kommerzielle Angebote an für MapReduce einsetzbaren Dateisystemen.

MapReduce und NoSQL-Datenbanken

Durch Kombination von MapReduce mit NoSQL-Datenbanken kann der gezielte Zugriff auf Daten verbessert werden. Die speziell für Big Data entwickelten NoSQL-Datenbanken, werden in Abschn. 4.4.3 ausführlich behandelt.

Apache Hadoop-Unterprojekte und Aufsatzprodukte

Neben den beiden Basisfunktionen, dem Hadoop Distributed File System (HDFS) und MapReduce bzw. YARN gibt es eine Reihe zusätzlicher Hadoop-Unterprojekte, die in Kombination mit den beiden Kernkomponenten, Hadoop zu einem umfassenden Ökosystem und einer universell einsetzbaren Plattform für Analyseanwendungen machen. Nachfolgend werden die wichtigsten Unterprojekte zusammengefasst.

Pig mit der Skriptsprache „Pig Latin“ ermöglicht die Erstellung von Scripts, die in MapReduce-Jobs kompiliert werden.

Mittels **Hive** und der deklarativen Abfragesprache HiveQL können Ad-hoc-Abfragen und die Erstellung von Reports ohne großen Aufwand durchgeführt werden. Die Kompilierung in die entsprechenden MapReduce-Jobs findet automatisch statt. Als SQL-ähnliche Sprache ermöglicht HiveQL zwar einen schnellen Einstieg für Analytiker und Programmierer mit Kenntnissen in relationalen Datenbanken; gegenüber SQL hat die Sprache jedoch einen etwas eingeschränkten Funktionsumfang. Hive wird auch verwendet, um Daten aus HDFS in ein bestehendes Data Warehouse oder in die NoSQL-Datenbank HBase zu laden.

Sqoop wird für den Import und Export von Massendaten zwischen HDFS und relationalen Datenbanken herangezogen. Es existieren Konnektoren zu Oracle-Datenbanken, MySQL und Microsoft SQL Server. Andere Datenbanken können über Standard-SQL-Schnittstellen, bspw. JDBC integriert werden.

Flume eignet sich insbesondere für den Import von Datenströmen, wie bspw. Web-Logs oder andere Protokolldaten, in das HDFS.

Avro dient der Serialisierung strukturierter Daten. Die strukturierten Daten werden in Bitketten konvertiert und in einem kompakten Format effizient im HDFS abgelegt. Die serialisierten Daten enthalten auch Informationen über das ursprüngliche Datenschema.

Die beiden NoSQL Datenbanken **HBase** und **Cassandra** erlauben eine sehr effiziente Speicherung großer Tabellen und einen effizienten Zugriff darauf. NoSQL-Datenbanken werden im nächsten Abschnitt ausführlicher behandelt.

Mahout ist eine Bibliothek von Modulen für maschinelles Lernen und Data Mining auf Basis unterschiedlichster Verfahren. Die Bausteine dienen dazu, aus vorhandenen, historischen Daten oder in der Anwendung gemachten Erfahrungen zu lernen und dieses Wissen auf neue Situationen übertragbar zu machen. Damit können für Probleme, die auf herkömmliche Weise gar nicht oder nur schwer mit hohem Aufwand lösbar sind, Lösungsansätze entwickelt bzw. weiterentwickelt werden.

ZooKeeper ist eine Bibliothek von Modulen zur Implementierung von Koordinations- und Synchronisationsdiensten in einem Hadoop-Cluster. Die Module werden von anderen Hadoop-Unterprojekten, wie Hive, HBase, Flume und Chukwa verwendet.

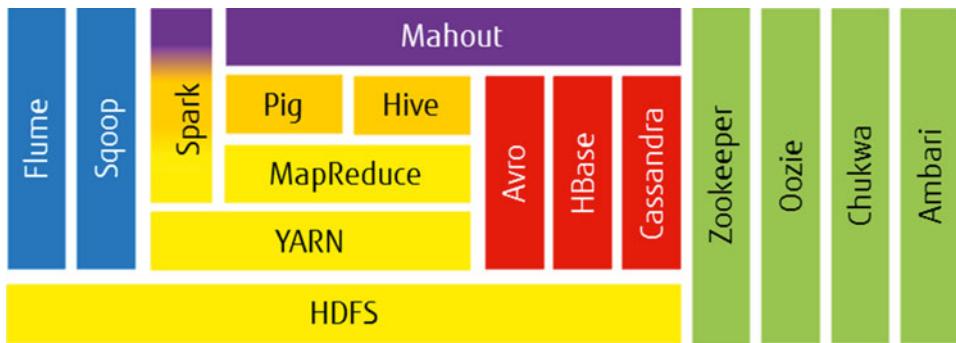


Abb. 4.12 Hadoop-Kernkomponenten und Unterprojekte (Quelle: Fujitsu)

Abläufe lassen sich mit **Oozie** beschreiben und automatisieren. Dies geschieht unter Berücksichtigung von Abhängigkeiten zwischen einzelnen Jobs. Bspw. wird eine Abfrage erst dann abgesetzt, wenn bestimmte andere Jobs abgeschlossen sind. Ebenso können die Abläufe protokolliert und die Zustandsinformationen visualisiert werden.

Chukwa überwacht große Hadoop-Umgebungen mit entsprechender Visualisierung. Protokolldaten werden von den Clusterknoten gesammelt und verarbeitet und die Ergebnisse visualisiert.

Ambari erlaubt eine einfache Verwaltung von Hadoop-Umgebungen über ein intuitives Web-Interface. DataNodes werden hinsichtlich ihres Ressourcenverbrauchs und ihres Gesundheitszustandes überwacht. Cluster-Installation und Upgrades können automatisiert ohne jegliche Serviceunterbrechung durchgeführt werden. Dies erhöht die Verfügbarkeit und beschleunigt die Wiederherstellung im Notfall. Ambari nutzt Apache Nagios, welches Alarne mittels E-Mail, SMS und SNMP erzeugt, sowie Ganglia zur Performance-Überwachung.

Spark ist ein Rahmenwerk für verteilte parallele Verarbeitung. Unter YARN kann Spark gemeinsam mit anderen Parallelisierungsstrategien auf einem Cluster betrieben werden. Die zu verarbeitenden Daten können in HDFS, aber auch in HBase oder Cassandra verwaltet werden. Spark bietet auch die Möglichkeit, Daten direkt in den Hauptspeicher der Cluster-Knoten zu laden. Dies ist insbesondere dann von Vorteil, wenn mehrfache Abfragen auf dieselben Datenbestände abgesetzt werden, was z. B. bei Machine Learning der Fall ist. So kann für bestimmte Applikationen eine erhebliche Beschleunigung im Vergleich zu Hadoop MapReduce erreicht werden; Spark bietet dafür als höherwertige Schnittstellen mit Shark einen SQL-Prozessor, die Stream-Verarbeitung Spark Streaming, die Machine Learning Bibliothek MLlib und die Graph-Verarbeitung GraphX.

Die Programmierung komplexerer Abläufe durch Map- und Reduce-Funktionen wird schnell sehr aufwändig. Hier helfen eine Vielzahl von Tools und Sprachen, die das Pro-

blem auf einer höheren Abstraktionsebene prozedural oder deklarativ beschreiben lassen und daraus entsprechende MapReduce-Abläufe generieren. Manche dieser Tools, die in Hadoop-Umgebungen für Analysen und Reporting, sowie zum Suchen nach Informationen und zur Visualisierung eingesetzt werden, sind als Open Source Software frei verfügbar, andere dagegen sind kommerzialisiert.

4.3.2.2 Reale oder virtuelle Server?

Ob man für verteilte Parallelverarbeitung eher reale oder virtuelle Server verwendet, lässt sich nicht allgemein beantworten. Es kommt wie so oft auf die jeweilige Situation an. Daher versuchen wir, die Vor- und Nachteile der beiden Ansätze gegenüberzustellen.

Virtuelle Server sind sicherlich einfacher und schneller bereitzustellen als reale Server. Insbesondere wenn sich neue Lastanforderungen dynamisch und spontan ergeben, vielleicht auch sogar von unterschiedlichen Mandanten, ist eine virtuelle Lösung von Vorteil. Ebenso können Produktivsysteme und Entwicklungs- bzw. Testsysteme elegant voneinander getrennt werden, was auch einen problemlosen Parallelbetrieb unterschiedlicher Versionen des Betriebssystems oder des Hadoop-Frameworks (möglicherweise auch von verschiedenen Hadoop-Distributionen) ermöglicht.

Beim Einsatz virtueller Server ist nur mit geringen Performance-Einbußen durch den Hypervisor, kaum nennenswerter Verringerung des Netzdurchsatzes und leicht höheren Kosten zu rechnen. Die wesentliche Einschränkung ergibt sich allerdings durch den persistenten Speicher. In virtualisierten Umgebungen wird typischerweise eine physikalische Serverfarm mit einem Netzwerkspeicher zusammengeschaltet. Dadurch geht die Lokalität der Daten verloren, was die Skalierbarkeit unter Umständen deutlich herabsetzen kann.

4.3.3 NoSQL-Datenbanken

Verteilte Dateisysteme, wie das HDFS, können problemlos für große Datenmengen und Daten unterschiedlichen Typs verwendet werden. Im Vergleich zu einem Dateisystem erlaubt eine Datenbank jedoch eine weitaus effizientere Bereitstellung und einfache Bearbeitung der Daten.

Die heute am weitesten verbreitete Form der Datenbank ist die relationale Datenbank. Relationale Datenbanken eignen sich hervorragend für die Transaktionsverarbeitung strukturierter Daten von begrenztem Umfang. Sie sind optimiert für den satzweisen parallelen Zugriff vieler Benutzer, sowie für Operationen zum Einfügen, Aktualisieren und Löschen von Datensätzen.

Zur Vermeidung von Redundanzen wird der Datenbestand als Verknüpfung mehrerer Tabellen dargestellt (Normalform). Dieses steigert die Effizienz etwa bei Änderungen und sowie bei der persistenten Datenspeicherung. Das Einlesen und Zusammenstellen der Verknüpfungen macht die Abfragen dagegen aufwändiger. Klassische relationale Datenbanken realisieren zur reibungslosen Abwicklung vieler Transaktionen die ACID-Eigenschaften, d. h. jede Transaktion wird ganz oder gar nicht ausgeführt (Atomicity),

der Datenbestand ist immer in einem konsistenten Zustand (Consistency), die Transaktionen wickelt das System isoliert voneinander ab, ohne dass sich der Programmierer darum kümmern muss (Isolation), abgeschlossene Transaktionen und ihre verändernde Wirkung bleiben dauerhaft bestehen (Durability). Diese Eigenschaften erzeugen einen gewissen Overhead und beschränken den Durchsatz, sodass es bei großen Datenmengen nicht mehr gelingt, Abfragen innerhalb akzeptabler Zugriffszeiten abzuwickeln. Wie im einführenden Abschnitt zu IT-Infrastrukturen für Big Data beschrieben, helfen hierbei auch die vertikale und die horizontale Skalierung nicht. Außerdem liegen relationalen Datenbanken starre Schemata zu Grunde, deren Änderung meist mit erheblichen Reorganisationsauswänden der Daten verbunden ist.

Im Gegensatz zu den sehr allgemein gehaltenen relationalen Datenbanken mit ihrem schwergewichtigen Transaktionskonzept adressieren NoSQL-Datenbanken (Not only SQL) jeweils nur eine spezialisierte Aufgabenstellung und erreichen durch ihre Einfachheit sehr hohe Durchsatzraten in den für Big Data typischen immensen Datenmengen. Da sie auf keinem festen Schema aufsetzen, können sie recht einfach beliebige Datentypen aufnehmen. Ebenso sind Datentypen einfach veränderbar. Die NoSQL-Datenbank sollte in ihrer Datenhaltung passend zum fachlichen Problem bzw. zur Anwendung gewählt werden, sodass für eine Transaktion nur ein Objekt eingelesen werden muss, anstatt viele Speicherobjekte zu lesen und zu kombinieren. Daten und Abfragen lassen sich dann gut auf die Rechnerknoten eines Clusters verteilen und ermöglichen fast lineare und unbegrenzte Skalierbarkeit, sowie hohe Fehlertoleranz durch Datenrepplikation und automatische Wiederherstellung im Fehlerfall.

Da hohe Geschwindigkeit bei Datenzugriffen und Datenverarbeitung im Vordergrund steht, ist in vielen NoSQL-Implementierungen eine Caching-Funktion integriert, bei der häufig benutzte Daten resident im Hauptspeicher der Rechnerknoten gehalten werden.

Statt der großen Vielfalt möglicher relationale Abfragen eines traditionellen Systems soll ein Big Data-System relativ wenige Anfragen mit jeweils enorm großen Datenvolumina verarbeiten. Hier muss vor allem sichergestellt sein, dass die Verarbeitung vollständig durchgeführt wird. Fehler in einer Verarbeitung (z. B. Absturz eines Knoten) dürfen nicht zum Abbruch der ganzen Query führen. Stattdessen muss das System Gegenmaßnahmen einleiten und eventuell Teilaufgaben erneut starten, bzw. sogar Ungenauigkeiten akzeptieren, bis die gesamte Query abgearbeitet ist. Anstelle der ACID-Konformität setzt man hier auf die BASE-Eigenschaften (Basic Availability, Soft State, Eventual Consistency).

Sehr deutlich bringt das CAP Theorem die unterschiedlichen Ziele und ihre Unvereinbarkeit zum Ausdruck. Von den drei Zielen Consistency, Availability und Partition Tolerance (daher CAP), kann ein System immer nur zwei Eigenschaften erfüllen. Konsistenz bedeutet hier, dass alle Datenreplikate identisch sind. Verfügbarkeit heißt, alle Anfragen an das System können stets beantwortet werden; und Partitionstoleranz besagt, das System arbeitet auch bei Trennungen des Netzes und Aufteilung des Clusters in mehrere getrennte Partitionen weiter.

Klassische ACID-konforme Datenbanken setzen auf Consistency und Availability (CA), während Big Data Systeme auf Availability und Partition Tolerance (AP) oder auf

Consistency und Partition Tolerance (CP) setzen (wobei nicht jede Störung die Konsistenz oder die Verfügbarkeit beschädigt, sondern nur die 100 %-Garantie nicht mehr gegeben werden kann).

Je nach Anwendungsfall sind die Anforderungen an Konsistenz und Verfügbarkeit abzuwegen, und eine bestmögliche Kombination für die jeweilige Anwendung zu finden. Meist wird die gewählte Kombination von Eigenschaften auf das gesamte NoSQL-System bezogen. Man könnte aber auch für jede Transaktion festlegen, auf welche zwei der drei Eigenschaften zu achten ist. Die NoSQL-Datenbank Cassandra realisiert genau das und steigert auf diese Weise insgesamt sowohl die Verfügbarkeit wie auch die Konsistenz.

Zu den kommerziellen NoSQL-Datenbankprodukten, die es teilweise schon seit langerem gibt, kommen im Zeitalter von Big Data immer mehr Open Source-Produkte hinzu. Die heute verfügbaren Produkte lassen sich in unterschiedliche Kategorien untergliedern, je nachdem für welche Problemstellung sie konzipiert und optimiert wurden.

4.3.3.1 Key-Value Stores

Die erste Variante, die wir uns anschauen, sind die Key-Value-Stores, in denen Paare aus Schlüssel und Wert in großen Mengen gespeichert werden. Der Zugriff auf den Wert erfolgt über den Schlüssel, über den der Wert eindeutig referenziert werden kann. Die Struktur der Values wird nicht von der Datenbank interpretiert. Das ist allein Sache der Applikation.

Für den Key-Value-Store gibt es nur wenige Operationen: Hinzufügen und Löschen eines Key-Value-Paares, Auslesen und Ändern eines Wertes. Diese Operationen werden sehr effizient und parallelisierbar implementiert.

Key-Value Stores eignen sich insbesondere für die schnelle Verarbeitung von Daten aus dem Internet, wie bspw. Clickstreams oder Suchmaschinen. Daher sind sie für Online Shopping-Anwendungen und Suchfunktionen, z. B. in Mail-Systemen äußerst interessant.

4.3.3.2 Beispiel: Key-Value Store mit Produktinformationen

In Abb. 4.13 haben alle Produkte eine Identifikation, einen Preis und eine Beschreibung. Unterschiedliche Produkttypen können durchaus unterschiedliche Attribute haben; bei einem Buch kann dies der Autor sein, bei Jeans ist es stattdessen die Länge. Attribute können unterschiedlich lang sein, bspw. die Tracks in Musikalben.

Key	Value
24989109	ID:024989109; Price:1.08; Description: Milk; Quantity: 1L
44455137	ID:044455137; Price:18.00; Description: Novel; Author: Pete Swan; Title: Beyond Big Data; Pages: 436;
69487316	ID:069487316; Price:99.00, Description: Blue Jeans; Width:32; Length: 36;
93543198	ID:093543198; Price:39.99; Description: Green Jeans; Size: S;
849836501	ID:849836501; Price:9.75, Description: CD;Artist: Fat Daddy; Track1: Drunken in Heaven; Track2: Under the hood; Track3: ...;
903983733	ID:903983733; Price:0.75; Description: Tomatos (canned);

Abb. 4.13 Key-Value Store (Quelle: Fujitsu)

4.3.3.3 Dokument-orientierte Datenbanken (Document Stores)

Bei Dokument-orientierten Datenbanken werden die Daten in Form von Dokumenten gespeichert. Ähnlich wie bei den Key-Value-Stores werden die Dokumente (Values) durch eindeutige Namen (Keys) referenziert.

Jedes Dokument ist vollkommen frei bezüglich seiner Struktur oder seines Schemas; das heißt, es kann das Schema verwendet werden, das für die Applikation benötigt wird. Sollten neue Anforderungen auftreten, so ist leicht eine Anpassung möglich. Neue Felder können hinzugefügt und bereits verwendete Felder können weggelassen werden. Völlig andersartige Schemata oder auch schemalose Daten (z. B. Text- oder Binärdaten) können in derselben Dokumenten-orientierten Datenbank gespeichert sein.

Da dokumentorientierte Datenbanken über keine Mittel zur Verarbeitung der Dateiinhalte verfügen, ist der Datenzugriff vollständig durch die Applikation zu leisten und die Programmierung etwas aufwändiger.

Dokumentorientierte Datenbanken eignen sich besonders zum Speichern zusammenhängender Daten in einem Dokument (bspw. HTML-Seiten), oder serialisierter Objektstrukturen (bspw. im JSON-Format).

Beispielsweise setzt Twitter eine dokumentorientierte Datenbank zur Verwaltung der Nutzerprofile unter Einbeziehung der Follower und Kurznachrichten (Tweets) ein.

4.3.3.4 Spaltenorientierte Datenbanken (Columnar Stores)

Die geläufigste und wahrscheinlich am häufigsten genutzte Variante der NoSQL-Datenbank ist die spaltenorientierte Datenbank. Ihr Hauptanwendungsgebiet liegt in der Verarbeitung großer Mengen strukturierter Daten, die sich mit relationalen Datenbanksystemen nicht angemessen bewältigen lassen.

Man stelle sich eine Tabelle mit Milliarden von Zeilen vor, wobei jede einzelne Zeile einen Datensatz darstellt. Die Anzahl der benötigten Spalten pro Datensatz ist dagegen vergleichsweise gering. Abfragen bei Analyseaufgaben beziehen sich in aller Regel nur auf wenige Spalten. Auf Grund der zeilenweisen Ablage müssen jedoch bei jeder Abfrage alle Spalten gelesen werden. Dies ist äußerst ineffizient.

Eine spaltenweise Speicherung der Daten in einer spaltenorientierten Datenbank erhöht die Effizienz. Der Zugriff wird auf die für die Abfrage relevanten Spalten beschränkt, wodurch das Lesen irrelevanter Daten vermieden wird. Auf diese Weise lässt sich die zu lesende Datenmenge erheblich reduzieren. Sehr häufig kommen auch dünnbesetzte

Id	Name	Vorname	Geschlecht	Geburtsdatum	Familienstand	Land	Stadt	PLZ	Straße
1	Schultze	Hans	männlich						
2	Müller	Eva	weiblich	1985-03-12	ledig	D	München	80807	Domagkstraße
3	Ihrgau	Horst	männlich			F			
4	Schmidt	Udo							
5	Meier	Fritz	männlich						

Abb. 4.14 Zeilenorientierte Datenspeicherung (Quelle: Fujitsu)

Tabellen oder Matrizen vor, die nicht nur Milliarden von Zeilen sondern ähnlich viele Spalten besitzen, von denen aber nur sehr wenige mit Werten besetzt sind.

Spaltenorientierte Datenbanken sind selbstindizierend. Obwohl sie dieselben Vorteile für die Abfrageleistung bieten wie Indizes, ist kein zusätzlicher Indexbereich erforderlich.

Da Spaltendaten einen einheitlichen Datentyp haben und es oft nur wenige verschiedene Werte pro Spalte gibt, lässt sich eine hohe Datenkomprimierung erzielen. Typische Komprimierungsfaktoren liegen im Bereich von 10 bis 50. Somit lassen sich die benötigten Speicherkapazitäten und folglich auch die Speicherkosten reduzieren.

Die spaltenorientierte Struktur ermöglicht eine einfache Verteilung der Daten auf viele Server und somit einer extrem gute Skalierbarkeit. Dagegen erfordert Schreiben in die Datenbank, insbesondere wenn mehrere unterschiedliche Spalten betroffen sind, vergleichsweise längere Bearbeitungszeiten.

Einige Columnar Stores unterstützen Spaltenfamilien. Das Datenbankschema enthält nur die grobe Struktur dieser Familien. Jede Spaltenfamilie steht typischerweise für einen inhaltlich zusammenhängenden Bereich von Daten mit ähnlichen Zugriffsmustern. Eine Unterteilung der Familie in mehrere Spalten, auch das Hinzufügen neuer Spalten, kann dynamisch erfolgen, ohne dass das Schema geändert und die Datenbank reorganisiert werden muss. Dies macht den Einsatz der Datenbank sehr flexibel. Neue oder geänderte Einträge werden mit einem Zeitstempel versehen, sodass nicht nur der aktuelle Stand, sondern auch eine gewisse Historie der Daten gespeichert wird. Dadurch wird das dynamische Verhalten nachvollziehbar, und es ist möglich, ggf. auftretende Inkonsistenzen im Nachhinein zu beheben.

Columnar Stores eignen sich besonders, um sehr große Datenmengen in der Struktur dünn besetzter Tabellen performant für Ad-hoc-Abfragen zur Verfügung zu stellen. Als ein Beispiel sei hier die BigTable von Google genannt, die den Inhalt gefundener Webseiten und ihre Historie speichert, und für die Aufbereitung von Suchergebnissen bereitstellt. Open Source-Implementierungen dieses Ansatzes sind HBase und Cassandra.

4.3.3.5 Graph-Datenbanken (Graph Databases)

Bei Graph-Datenbanken werden Informationen in Graphen mit Knoten und Kanten, sowie deren Eigenschaften dargestellt. Eine der häufigsten Berechnungen in einer Graphen-Datenbank ist die Suche nach dem einfachsten und effizientesten Pfad durch den gesamten Graphen. Anwendungsgebiete sind die Fahrplanoptimierung, geografische Informations-

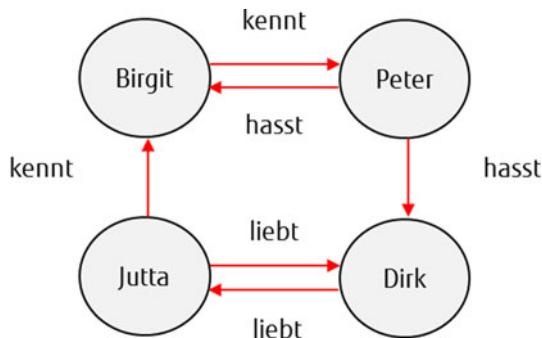
Id	Name	Id	Vorname	Id	Geschlecht	Id	Geburtsdatum	Id	Familienstand	Id	Land	Id	Stadt	Id	PLZ	Id	Straße
1	Schultze	1	Hans	1	männlich	2	12.03.1985	2	ledig	2	D	2	München	2	80807	2	Domagkstraße
2	Müller	2	Eva	2	weiblich					3	F						
3	Ihrgau	3	Horst	3	männlich												
4	Schmidt	4	Udo	5	männlich												
5	Meier	5	Fritz														

Abb. 4.15 Spaltenorientierte Datenspeicherung (Quelle: Fujitsu)

Id	Spaltenfamilie: Person			Spaltenfamilie: Adresse		
1	2011-05-14 17:45:22:648	Name	Schultze			
		Vorname	Hans			
		Geschlecht	männlich			
2	2010-02-13 12:57:16:145	Name	Müller	2010-02-13 12:57:16:145	Land	D
		Vorname	Eva	Stadt	München	
		Geschlecht	weiblich	PLZ	80807	
		Geburtsdatum	1985-03-12	Straße	Domagkstraße	
2	2013-12-03 10:21:10:760	Name	Thurgau	2013-12-03 10:21:10:760	Land	F
		Familienstand	verheiratet			
3	2013-12-03 10:21:10:760	Name	Thurgau	2013-12-03 10:21:10:760	Land	F
		Vorname	Horst			
		Familienstand	verheiratet			

Abb. 4.16 Columnar Store mit Spaltenfamilien (Quelle: Fujitsu)

Abb. 4.17 Beziehungen zwischen Personen als Graph dargestellt (Quelle: Fujitsu)



systeme (GIS), Hyperlinkstrukturen sowie die Darstellung von Nutzerbeziehungen innerhalb sozialer Netzwerke.

Wenn die Anzahl der Knoten und Kanten für einen Server zu groß wird, muss der Graph partitioniert werden, was im Prinzip einer horizontalen Skalierung entspricht. Dabei wird der Gesamtgraph in Teilgraphen zerlegt, was sich als durchaus schwierige, manchmal sogar unlösbare Aufgabe erweisen kann. In letzterem Falle müssten manche Knoten mehreren Teilgraphen zugeordnet werden. Man spricht dann auch von einer überlappenden Partitionierung.

Bei großen Datenbanken mit hoher Lese-Intensität und relativ geringer Schreiblast, kann eine Replikation der Daten auf mehrere Server zu einer beschleunigten Bearbeitung beitragen.

Im Gegensatz zu den zuvor behandelten Kategorien von NoSQL-Datenbanken erfüllen viele der Graph-Datenbanken die von relationalen Datenbanken bekannten ACID-Eigenschaften.

4.3.4 In-Memory-Technologien

Verteilte Parallelverarbeitung auf Basis verteilter Dateisysteme oder verteilter NoSQL-Datenbanken kann sehr wohl für umfangreiche Analyseaufgaben herangezogen werden. Verwendet man verteilte Dateisysteme, steht jedoch häufig die Vorverarbeitung großer Datenmengen im Vordergrund. Dabei werden die Daten bereinigt; Redundanzen und Widersprüche werden entfernt; und letztlich werden die Daten in eine Form gebracht, die sich besser für spontane (ad-hoc) Abfragen eignet, deren Antworten in Echtzeit erwartet werden. Das Ergebnis dieser Transformation ist in der Regel sehr viel kleiner als die Eingangsdatenmenge, enthält aber alle für den Anwendungsfall wichtigen Informationen und lässt sich somit als destillierte Essenz der Gesamtdaten auffassen.

Es steht außer Frage, dass der Zugriff zu Daten auf Festplatte oder auch schnelleren Speichermedien wie bspw. SSD (Solid State Disks) nie so schnell sein kann wie auf Daten, die sich resident im Hauptspeicher eines Servers und somit näher bei den Applikationen befinden. Um die destillierte Essenz für Echtzeitaufgaben schnell zugänglich zu machen, bietet sich eine Datenhaltung im Hauptspeicher eines oder mehrerer Rechner an. In der Praxis werden mit der Einführung derartiger In-Memory-Plattformen die Datenzugriffe um das 1000- bis 10.000-fache beschleunigt. Die Analyse von Geschäftsdaten kann so in Echtzeit erfolgen und nimmt nicht mehr Stunden, Tage oder sogar Wochen in Anspruch. Wichtige Entscheidungen lassen sich somit viel schneller treffen.

Um die Datenverfügbarkeit zu erhöhen, können Hauptspeicherinhalte zwischen verschiedenen Rechnerknoten gespiegelt und damit synchron gehalten werden. Natürlich reduziert sich dadurch die insgesamt verfügbare Netto-Hauptspeicherkapazität entsprechend.

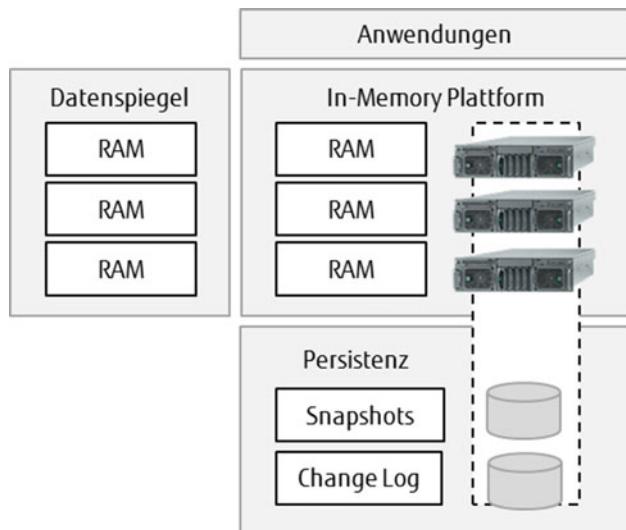
Für Analysezwecke kann auf Plattspeicher komplett verzichtet werden. Zu beachten ist allerdings, dass Daten, die sich nur im flüchtigen Hauptspeicher eines Rechners befinden, bei Systemabsturz verloren gehen können, insbesondere natürlich nach einem Stromausfall. Der Einsatz von Notfallbatterien für den Hauptspeicher hilft hier sicher nur temporär.

Um Datenverlustes zu verhindern, müssen die Dateninhalte, die sich im Hauptspeicher befinden, zusätzlich auf ein nichtflüchtiges, persistentes Speichermedium gebracht werden. Hierfür sind mehrere Lösungsansätze denkbar.

Eine gängige Methode ist die kontinuierliche Replikation der Hauptspeicherdaten auf Festplatte oder SSD. Ein identischer Status von Hauptspeicher und Platte ist damit jederzeit garantiert.

Zur Reduzierung der I/O-Last können alternativ in regelmäßigen Abständen oder beim kontrollierten Abschalten des Gesamtsystems Snapshot-Dateien erzeugt werden, die den jeweils aktuellen Zustand des Hauptspeicherinhalts festhalten. Bei einem Absturz können hierbei allerdings Datenänderungen, die sich seit dem letzten Snapshot ergeben haben, verloren gehen. Durch Mitprotokollieren sämtlicher Datenänderungen ist auch diese Lücke zu schließen. Aus dem letzten Snapshot und dem Protokoll der inzwischen getä-

Abb. 4.18 In-Memory-Plattform – Applikationen, Daten im Hauptspeicher, Persistenzschicht (Quelle: Fujitsu)



tigten Änderungen kann der zuletzt gültige Hauptspeicherinhalt automatisch wiederhergestellt werden.

Als Speicher für die Datenreplikate, die Snapshots bzw. die Änderungsprotokolle kommen sowohl lokale Platten der beteiligten Rechnerknoten, wie auch Speichersysteme im Netz in Betracht. Die Verwendung von Solid State Disks (SSD) beschleunigt natürlich die Wiederherstellung der Daten nach einem Absturz.

Infolge der ständig sinkenden Preise für Hauptspeicher und der steigenden Leistung von Netzkomponenten, welche dazu beitragen, die Hauptspeicherinhalte mehrerer Rechner zu einer logischen Einheit zu formen, werden In-Memory-Plattformen mit immer größer werdenden Datenkapazitäten zu einem wichtigen Infrastrukturbaustein für Big Data.

Nachfolgend werden verschiedene Implementierungsmöglichkeiten von In-Memory-Plattformen näher unter die Lupe genommen.

4.3.4.1 In-Memory-Datenbanken (IMDB)

Bei einer IMDB befindet sich der komplette Datenbestand inklusive Verwaltungsdaten komplett im Hauptspeicher. Auch die Datenoperationen werden nur noch im Hauptspeicher durchgeführt. Lesevorgänge und Schreibvorgänge von und zur Festplatte entfallen bei Verwendung von In-Memory-Datenbanken komplett und ohne Ausnahme. Hierdurch können Daten extrem schnell gespeichert, abgerufen und verarbeitet werden. So lassen sich komplett Analysen der Daten vollständig im Hauptspeicher abwickeln, was zu enormen Leistungsgewinnen führt. Man spricht hier von In-Memory Analytics.

Der Einsatz von IMDB liegt nahe, wenn voneinander unabhängige Applikationen aus vollkommen unterschiedlichen Blickwinkeln auf Datenbestände zugreifen, darauf dyna-

misch im Vorfeld noch nicht bekannte ad-hoc Abfragen absetzen, und die Ergebnisse in Echtzeit zu bereitzustellen sind.

Die ersten im Markt verfügbaren IMDB waren für vertikale Skalierung ausgelegt. Mittlerweile gibt es aber auch IMDB, die horizontal skalieren. Die Skalierbarkeit ist hier jedoch bei Weitem nicht so ausgeprägt wie bei Hadoop-Konfigurationen. Der Trend geht derzeit in die Richtung, horizontale und vertikale Skalierung gleichermaßen zu unterstützen.

Ob eine In-Memory-Datenbank für die destillierte Essenz in Frage kommt, hängt von verschiedenen Faktoren ab. Zum einen ist dies die Datengröße, welche durch die im gesamten Servercluster verfügbare Hauptspeicherkapazität begrenzt wird. Bei spaltenorientierten Architekturen ist hier auch der Komprimierungsfaktor ausschlaggebend, der je nach IMDB und verwendetem Komprimierungsverfahren bei 10 bis 50 liegt und demzufolge die Kapazitätsgrenze um das 10- bis 50-fache nach oben verschieben lässt. Wie viele Rechnerknoten insgesamt verwendet werden dürfen, ist in der Regel auch von IMDB zu IMDB verschieden. Außerdem spielt natürlich das verfügbare Budget eine Rolle; obwohl Arbeitsspeicher ständig preiswerter wird, sind die Kosten im Vergleich zu anderen Speichermedien immer noch vergleichsweise hoch.

Einige IMDBs, wie bspw. SAP HANA oder Oracle DB mit IMDB-Option, wickeln OLTP und den OLTP-Betrieb nicht beeinträchtigende OLAP-Abfragen in derselben Datenbankinstanz ab, und ermöglichen so eine Analyse von Produktionsdaten in Echtzeit. In solchen Implementierungen verwendet man häufig hybride Tabellen, bestehend aus einem für OLAP optimierten Spaltenspeicher und einem Zeilenspeicher. Dabei werden aktuelle Datenänderungen während der Transaktionsverarbeitung im Zeilenspeicher hinterlegt, und nach bestimmten Kriterien in den Spaltenspeicher eingesortiert, um Konsistenz zu gewährleisten.

4.3.4.2 In-Memory Data Grids (IMDG)

Kommt eine IMDB nicht in Frage, verdient ein In-Memory Data Grid (IMDG) zwischen dem Plattenspeicher und den Applikationen Beachtung.

Ein IMDG ist ein in der Regel über mehrere Server verteilter residenter Hauptspeicherbereich, in den große Datenbestände von externen Speichersystemen eingelagert werden. Jede Art von Datenobjekt kann in einem IMDG verwaltet werden, also auch vollkommen unstrukturierte Daten, wie man sie bei Big Data mehrheitlich vorfindet. Im Gegensatz zu einer IMDB muss hier nicht zwingend der gesamte Datenbestand in den Hauptspeicher passen. Dennoch wird die I/O-Last drastisch reduziert, Applikationen werden beschleunigt, und Analysen können in Echtzeit durchgeführt werden.

Während sich IMDB für oftmals im Vorfeld noch nicht bekannte ad-hoc Abfragen von unterschiedlichen Anwendungen eignen, welche nach Belieben dynamisch auf den Datenbestand zugreifen, liegt der Haupteinsatz von IMDG eher bei exklusiven Nutzern und vordefinierten, somit also im Vorfeld schon bekannten Abfragen. Die Applikation bestimmt und steuert die Verarbeitung innerhalb der einzelnen Datenobjekte, das IMDG kümmert sich um den Zugriff auf die Daten, ohne dass die Anwendungen wissen müssen,

wo sich die Daten befinden. Zusätzliche Funktionen zu Suche und Indexierung der im IMDG gespeicherten Daten lassen die Grenzen zwischen einer IMDG-Lösung und einer NoSQL-Datenbank fließend erscheinen.

Ein IMDG kann eine kosteneffiziente Alternative zu einer IMDB sein, vor allem wenn die Dateninhalte nicht häufig aktualisiert werden. Die Anwendungen, die auf die Daten zugreifen, können in der Regel unverändert genutzt werden; bei entsprechend vorhandener Expertise kann aber durch eine Anpassung der Anwendungen der Nutzen eines IMDG optimiert werden.

Wo letztlich die Daten in einem IMDG herkommen, spielt für das IMDG keine Rolle. Die Daten können von Speichersystemen im Netz kommen, oder es kann sich um Daten aus dem HDFS handeln. Weniger denkbar sind NoSQL-Datenbanken, da diese oftmals eine integrierte Caching-Funktion besitzen, die bereits manche der Vorteile eines IMDG bietet.

Java Heap und das Garbage Collector-Problem

Viele der heute am Markt verfügbaren IMDG-Lösungen sind in der Sprache Java programmiert und somit auch den üblichen Herausforderungen einer Java-Umgebung ausgesetzt. Datenobjekte werden in der Java Heap erzeugt.

Datenobjekte, die nicht mehr benötigt werden, können von der Java Heap entfernt werden, um Speicherengpässe zu vermeiden. Diese Aufgabe übernimmt der Garbage Collector (sprichwörtlich die Müllabfuhr). Sein hoher Ressourcenbedarf verlangsamt jedoch den Betrieb und führt sogar zum Stillstand der Anwendung, wenn der Heap sehr groß ist. Es werden deshalb Heap-Größen von unter 1 GB empfohlen. Will man den physikalischen Hauptspeicher moderner Server ausnutzen, kann man sehr viele Java Virtual Machines mit kleinem Heap auf einer Server-Hardware einrichten, was aber einen nicht unerheblichen administrativen Aufwand darstellt.

Daher ist es ratsam, IMDG-Lösungen zu berücksichtigen, bei denen Datenobjekte nicht auf der Halde erzeugt werden, und somit für den Garbage Collector auch nicht sichtbar sind. Dadurch kann das Garbage Collector-Problem umgangen werden, ohne jegliche Auswirkungen auf den Betrieb. Die Halde kann minimal dimensioniert werden, und eine einzige JVM kann quasi den kompletten verfügbaren Hauptspeicher nutzen, was den Verwaltungsaufwand erheblich verringert. Ein Beispiel für eine derartige IMDG-Lösung ist Terracotta BigMemory von der Software AG.

4.3.5 Verarbeitung großer Ereignisströme

In den bisherigen Abschnitten wurde stets davon ausgegangen, dass Daten auf lokalen oder zentralen Plattenspeichern oder resident im Hauptspeicher von Rechnersystemen ruhen (Data at Rest) und auf Verarbeitung warten. Die Regeln, wie die Verarbeitung erfolgen soll, konnte in Form von Abfragen dynamisch vorgegeben werden.

Anders verhält es sich bei Datenströmen (Data in Motion). Hier werden Ereignisse kontinuierlich und mit hoher Frequenz bspw. von Sensoren erzeugt. Ein einzelnes Ereignis ist noch relativ einfach zu bewältigen. Anders dagegen sieht es aus, wenn ganze Ereignisströme nicht nur in Echtzeit von verschiedenen Quellen erfasst, sondern auch gefiltert und miteinander korreliert werden müssen, sowohl in logischer wie auch in zeitlicher Hinsicht. Dies genau ist die Aufgabe von CEP (Complex Event Processing).

Die Analyse von Ereignisströmen und die entsprechende Reaktion auf die Ereignisströme wird typischerweise durch Regeln definiert. Die Regeln bestehen aus einer Bedingung und einer Aktion. Die Bedingung kann durch bestimmte Konstellationen der Ereignisse im Eingabestrom erfüllt sein, was dann zur Auslösung der Aktion führt. Die einfachste Bedingung ist das Auftreten eines beliebigen Ereignisses. Das Auftreten eines bestimmten Ereignistyps wird von einer Filterbedingung geprüft. Mehrfaches Auftreten eines bestimmten Typs oder einer bestimmten Folge von Ereignistypen stellen eine einfache Korrelationsbedingung dar. Das Auftreten oder Nichtauftreten eines oder mehrerer Ereignistypen in einem bestimmten Zeitintervall sind weitere Möglichkeiten der Korrelation. Die Aktion beim Zutreffen der Bedingung ist die Erzeugung eines neuen Ereignisses. Dieses aus mehreren Ereignissen hervorgegangene komplexe Ereignis kann durch weitere Regeln analysiert werden oder zu einer externen Reaktion wie z. B. Schicken einer E-Mail oder Eintrag in einer Datenbank umgewandelt werden. Je nach CEP-Engine können Regeln z. B. deskriptiv in einer Query-Sprache oder prozedural in einer Script-Notation formuliert werden. Ereignisse können neben dem Typ eine innere Struktur haben, die ebenfalls durch die Regeln verarbeitet werden kann.

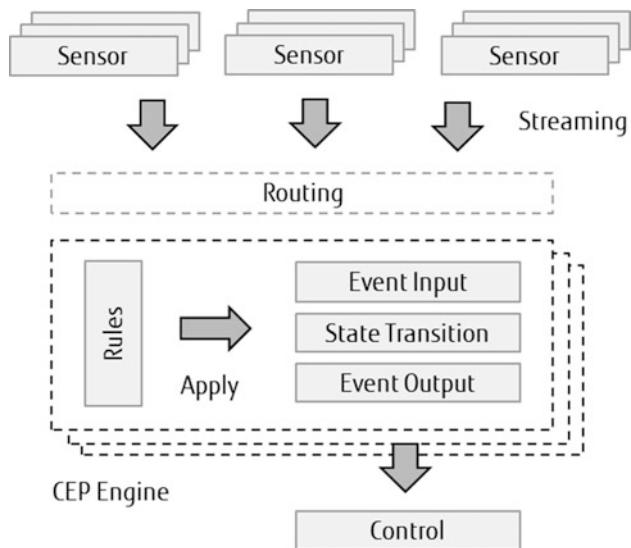
Neu eintreffende Daten warten nicht, bis sie an der Reihe sind, verarbeitet zu werden; die Verarbeitung muss sofort erfolgen. Typischerweise sind Hunderttausende oder gar Millionen von Ereignissen pro Sekunde bewältigen. Ein weiterer kritischer Parameter ist die Latenz, d. h. in diesem Fall die Zeit, die zwischen Ereigniseingabe- und Ereignisausgabe verstreichen darf. Typische Anforderungen bezüglich der Latenzwerte liegen im Mikro- bzw. Millisekunden-Bereich.

Eine CEP-Lösung muss demzufolge vor allem schnell und skalierbar sein. Um zeitraubende Plattenzugriffe zu vermeiden, werden die Datenströme hauptspeicherresident organisiert. Lediglich Betriebsprotokolldaten werden in eine Datei auf Platte geschrieben, sofern von dieser Option überhaupt Gebrauch gemacht wird. In diesem Protokoll werden unter anderem Engpässe bei Betriebsmitteln festgehalten, oder auch Anomalitäten, um im Nachgang die Ursache dafür zu ermitteln.

Ist die auftretende Last nicht von einem Server zu bewältigen, ist eine Verteilung auf mehrere Server erforderlich. Große Datenströme werden auf mehrere Server mit entsprechenden CEP Engines verteilt. Eingehende Ereignisse werden verarbeitet oder an andere Server weiter geroutet.

Sind sehr viele Regeln und Abfragen abzuarbeiten, findet eine Verteilung auf mehrere Server mit voneinander unabhängigen CEP Engines statt. Dabei wird versucht, Abfragen die in einer Beziehung miteinander stehen, soweit möglich, auf denselben Rechnerknoten zu dirigieren.

Abb. 4.19 Architektur einer CEP-Plattform (Quelle: Fujitsu)



Bei Regeln mit hohem Hauptspeicherbedarf, bspw. auf Grund größerer zu betrachtender Zeitfenster, kann ein IMDG helfen, welches sich über mehrere Server erstreckt. Datenverluste können über die in IMDG-Lösungen realisierten Hochverfügbarkeitsfunktionen vermieden werden (bspw. automatische Datenreplikation auf unterschiedlichen Rechnerknoten, bzw. eine temporäre Zwischenlagerung auf persistenten Speichermedien).

Komplexe Abfragen, die ein einzelner Server nicht alleine bearbeiten kann, können in Teilaufgaben zerlegt und auf mehrere Server verteilt werden.

Je nach Anwendungsfall können von einer CEP-Engine erarbeitete Ergebnisse auch in HDFS (bspw. die Betriebsprotokolldaten), NoSQL-Datenbanken, In-Memory-Datenbanken oder IMDG hinterlegt und für weitere Analysezwecke oder zur Visualisierung, bzw. für das Berichtswesen genutzt werden.

Theoretisch ist es auch denkbar, von Sensoren erzeugte Daten in anderen Datenspeichern zu konservieren. Allerdings ist dies in vielen Fällen nach ihrer Auswertung gar nicht erforderlich. Es gibt aber Fälle, wo alle Vorgänge nachvollziehbar sein müssen und deshalb sämtliche eintreffenden Daten protokolliert werden.

Neben Open Source-Produkten, wie bspw. ESPER gibt es heute auch einige etablierte kommerzielle CEP-Produkte, die sich teilweise auch mancher Funktionen aus den Open Source-Produkten bedienen. Insbesondere das Internet of Things wird zu einem drastischen Anstieg der Einsatzfälle für CEP führen.

4.3.6 Referenzarchitektur für Big Data-Infrastrukturen

Nachdem nun die wesentlichen Infrastrukturkonzepte für Big Data behandelt wurden, soll im nächsten Schritt eine Referenzarchitektur betrachtet werden, in der diese Konzepte berücksichtigt sind. Je nach Anwendungsfall kommen manche Technologien nur alternativ zum Einsatz, oder sie sind möglicherweise gar nicht erforderlich. Die Referenzarchitektur stellt letztlich einen Technologiebaukasten dar, aus dem für jede kundenspezifische Situation eine optimale Lösung als Kombination verschiedener Technologien konzipiert werden kann.

Daten werden aus verschiedenen Datenquellen extrahiert. Bei diesen Datenquellen kann es sich handeln um Transaktionssysteme und Data Warehouses, aber auch um IT Logs, Click Streams, Internetseiten, die auch miteinander verlinkt sein können, E-Mails, Textdateien und Multimedia-Dateien. Und falls bereits eine IMDB im Einsatz ist, kann diese auch als Datenquelle herangezogen werden. Zur Beschleunigung der Extraktion von Daten auf Plattenspeichern kann dem Plattenspeichersystem optional ein IMDG vorgeschaltet werden. Dies ist nicht nur für die Big Data-Anwendung von Vorteil, sondern auch für jede andere Applikation, die diese Daten benötigt.

Die Daten aus den genannten Quellen werden in ein verteiltes Datenhaltungssystem importiert, bspw. das HDFS, welches als Sammelbecken für große Datenmengen dient und sich über die lokalen Plattenspeicher von Rechnerknoten erstreckt. Über verteilte Parallelverarbeitung werden die Daten bereinigt und vorbearbeitet.

Die Analyse und die Visualisierung der Analyseergebnisse kann unmittelbar auf das verteilte Datenhaltungssystem angewandt werden. Alternativ können durch verteilte Parallelverarbeitung transformierte Daten als destillierte Essenz in ein anderes Datenhaltungssystem exportiert werden. Dies ist in aller Regel entweder eine SQL- oder eine

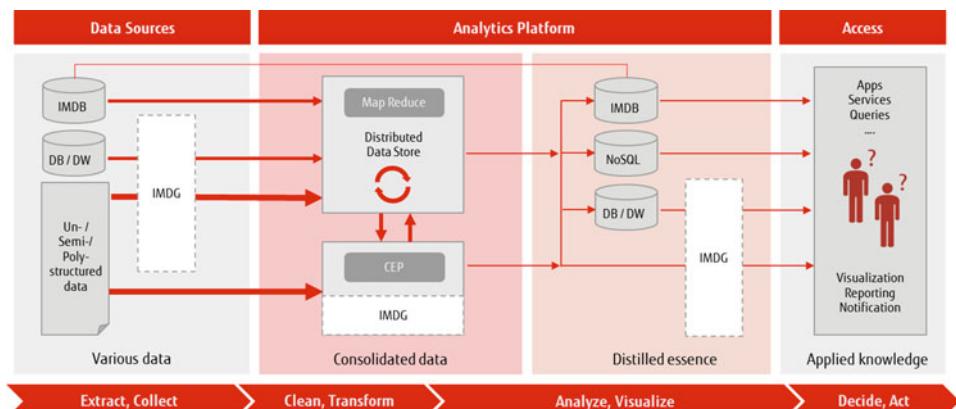


Abb. 4.20 Referenzarchitektur für Big Data-Infrastrukturen (Quelle: Fujitsu)

NoSQL-Datenbank. Auf diese Datenbank finden dann entsprechend auch die Analysen statt.

Bei weniger zeitkritischen Anwendungsfällen, kann sich die SQL-Datenbank durchaus auf einem zentralen Plattspeichersystem befinden, das zweckmäßigerweise über eine Hochgeschwindigkeitsverbindung mit dem entsprechenden Datenbank-Serversystem verfügt. Die Verwendung schneller Speichersysteme, wie bspw. All-Flash-Arrays (AFA), die von Hause aus allumfassend für Solid State Disks (SSD) konzipiert sind, und sich durch extrem hohe I/O-Leistung und geringe Latenzzeiten auszeichnen, kann hier weitere Vorteile bringen.

Sind jedoch Analyseergebnisse in Echtzeit gefragt, bietet sich bspw. eine IMDB an. Alternativ kann ein IMDG etabliert werden, um die destillierten Essenz entweder komplett oder zumindest zu großen Teilen hauptspeicherresident zu halten, und die Analyse-Applikationen drastisch zu beschleunigen.

Während Daten aus den oben erwähnten Quellen typischerweise mit Hilfe von Hadoop und MapReduce im Batchbetrieb verarbeitet werden, müssen z. B. Sensordaten, Eindringversuche, Kreditkartentransaktionen, oder andere mit hoher Frequenz generierte Datenströme mittels einer CEP-Plattform in Echtzeit erfasst und analysiert werden, damit auch die entsprechenden Maßnahmen in Echtzeit eingeleitet werden können. Als Datenspeicher für Complex Event Processing wird der Hauptspeicher genutzt, je nach Komplexität auch ein IMDG. An das Ergebnis der CEP-Engine ist ein Alarm gekoppelt mit entsprechenden Maßnahmen; je nach Anwendungsfall können aber auch Ergebnisse aus der CEP-Engine nach Hadoop, bzw. an die destillierte Essenz außerhalb Hadoop zwecks Weiterverarbeitung oder Visualisierung weitergeleitet werden. Gleichermaßen werden Hadoop-Daten manchmal auch für CEP genutzt.

Wie jeder hybride Lösungsansatz erweckt dies auf den ersten Blick den Anschein einer hohen Komplexität. Umso wichtiger ist es, den Endanwender nicht mit dieser hohen Komplexität zu konfrontieren. Dies ist über den Einsatz entsprechender Konnektoren zwischen den Teilsystemen, bzw. durch Datenadapter zu erreichen, die einen reibungslosen Übergang ermöglichen. Letztlich entscheidend für den Endanwender ist jedoch, dass die für ihn bereitgestellten Analysefunktionen sich – für ihn vollkommen transparent – der richtigen Daten aus den richtigen Datenspeichern bedienen. Für den Anwender ist nicht die zu Grunde liegende Technologie entscheidend, sondern die Einsicht in die verfügbaren Daten, und das Gewinnen von Erkenntnissen, die ihn und seinen Geschäftsauftrag nach vorne bringen.

4.3.7 Lambda-Architektur

Alternativ zu der eben erläuterten Referenzarchitektur, betrachten wir nun die Lambda-Architektur, die zumindest ansatzweise (Near Real-Time) Echtzeitanforderungen bei der Analyse großer Datenbestände erfüllen kann.

Unser Ausgangspunkt sind Rohdaten, die aus unterschiedlichen Datenquellen stammen können und im HDFS abgelegt wurden. Abfragen werden im Batchbetrieb bearbeitet und erfordern viel Zeit. Sollen neue Daten in die Betrachtung einbezogen werden, sind sie demzufolge erst nach erneuten langwierigen Batch-Jobs im Ergebnis zu finden. Ergebnisse werden zumindest im Falle großer Datenmengen nicht zeitnah nach dem Eintreffen der neuen Daten geliefert.

Um Abfragen auch ansatzweise in Echtzeit zu erledigen, führen wir nun 3 Bearbeitungsebenen ein: die Batch Ebene, die Serving-Ebene und die Speed-Ebene.

Die Batch-Ebene hat im Grunde genommen die Aufgabe, die bis zu einem bestimmten Zeitpunkt gesammelten Rohdaten zu verarbeiten (dies geschieht z. B. mit MapReduce im HDFS), und die Ergebnisse der Datenvorbearbeitung in die sogenannten Batch Views zu überführen. Die Aufbereitung der Batch Views geschieht über MapReduce. Neue inkrementelle Daten, die während der Vorbearbeitung eintreffen, werden dem aktuellen Datenbestand hinzugefügt, allerdings zunächst in den Batch Views nicht berücksichtigt. Die Batch Views werden zwar kontinuierlich überarbeitet, wegen der langen Laufzeit ist das Ergebnis zum Teil erheblich zeitlich verzögert.

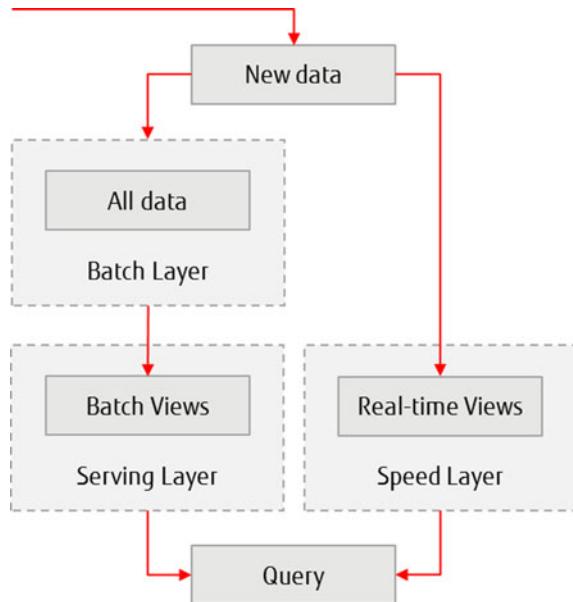
Die Serving-Ebene nimmt die Indizierung der Batch Views vor, um einen schnellen wahlfreien Zugriff zu ermöglichen. Ad-hoc Abfragen können demzufolge in Echtzeit beantwortet werden. Zur Implementierung der Serving-Ebene wird in der Regel eine NoSQL-Datenbank, wie bspw. Apache HBase verwendet. Es ist allerdings zu beachten, dass die Ergebnisse auf Grund inzwischen neu eingetroffener Daten, die in den Batch Views noch nicht berücksichtigt sind, nicht aktuell sind.

Die Speed-Ebene schließlich dient dazu, diese Aktualitätslücke zu schließen. Alle neuen Daten werden nicht nur an die Batch-Ebene, sondern auch an die Speed-Ebene übergeben. Die Speed-Ebene erzeugt aus den neuen Daten in extrem hoher Geschwindigkeit sogenannte Echtzeit-Views, die somit immer aktuell sind. Allerdings werden hier wegen der hohen Geschwindigkeitsanforderungen geringere Anforderungen an die Genauigkeit der Bearbeitung gestellt. Die Speed-Ebene überbrückt die Batch-Laufzeiten und die damit verbundenen Verzögerungen zwischen dem Eintreffen von Daten und deren Berücksichtigung im Batch View. Die Echtzeit-Views werden wieder verworfen, sobald sie in der Serving-Ebene berücksichtigt wurden.

Für die Bearbeitung von Abfragen werden die Ergebnisse aus den Batch Views und Echtzeit-Views kombiniert; und das aktuelle Gesamtergebnis kann in Echtzeit zurückübergeben werden.

Zur Implementierung der Lambda-Architektur werden verschiedene Technologien herangezogen. Für die Batch-Ebene ist sicherlich Hadoop MapReduce prädestiniert, wobei die Datenbestände zum Beispiel im HDFS oder in der NosQL-Datenbank HBase liegen. Für die die Serving-Ebene und die Speed-Ebene kommen additiv die beiden Open Source Software-Produkt Impala und Storm hinzu, die wir nachfolgend etwas näher betrachten.

Abb. 4.21 Lambda-Architektur (Quelle: Fujitsu)



4.3.7.1 Impala

MapReduce ist dafür konzipiert, Datenbestände im Petabyte-Bereich zu speichern und im Batchbetrieb zu verarbeiten. Das robuste Management von MapReduce erzeugt dabei einen gewissen Overhead, wodurch auch bei kleineren Datenmengen die Ablaufzeiten nicht beliebig klein werden. Mit Hive erstellte Abfragen werden üblicherweise in mehrere MapReduce-Läufe aufgeteilt. Damit entstehen Laufzeiten, die Hive für Ad-Hoc-Analysen im Dialogbetrieb ungeeignet machen.

Hier bietet Impala eine deutliche Verbesserung. Impala interpretiert mit kleinen Einschränkungen die von Hive bekannte Hive QL. Impala verwendet zur Leistungssteigerung eine in C++ implementierte In-Memory Query-Engine. Die komplette Query wird ausgeführt, ohne dass Zwischenergebnisse auf Platte geschrieben werden. Die Query-Ausführung und Verteilung auf den Cluster wird direkt von Impala ohne Festlegung auf eine Folge von MapReduce-Schritten geplant, was zusätzliche Optimierungsmöglichkeit bietet. Der Umfang der Eingaben, Ergebnisse und Zwischenergebnisse einer Query ist natürlich durch die konsolidierte Hauptspeichergröße der Serverfarm limitiert, was Impala auf den Einsatz für kleinere und mittlere Datenmengen im Terabytebereich einschränkt.

Impala nutzt als Schnittstellen HDFS oder HBase und ist damit hervorragend mit Hadoop integriert. Ohne Umkopieren von Daten steht das Ergebnis von MapReduce als destillierte Essenz zur Verfügung. Impala kann diese Daten nun sehr schnell verarbeiten, sodass ad-hoc-Analysen im Dialog möglich werden. Impala eignet sich damit sehr gut für die Serving-Ebene der Lambda-Architektur.

4.3.7.2 Storm

MapReduce verarbeitet eine große Datenmenge und liefert nach einer gewissen Verarbeitungszeit ein Ergebnis. Wenn sich der in den Eingabedaten abgebildete Teil der Welt inzwischen ändert, ist das Ergebnis zwar richtig in Bezug auf die früheren Eingabedaten, aber nicht mehr aktuell. Mit der Speed-Ebene versucht man, dieses Problem zu lösen oder zumindest zu verringern.

Storm wurde eigens für diese Aufgabe konzipiert. Storm ist darauf ausgerichtet, ein-treffende Daten, bzw. Datenänderungen schnell und zuverlässig in Echtzeit zu verarbeiten. Das System soll auch einem Wachstum der Daten bzw. einer Steigerung der Änderungsrate standhalten; daher spielen Fehlertoleranz, Parallelität und Skalierbarkeit eine entscheidende Rolle. Anders als MapReduce wird Storm nicht periodisch immer wieder ausgeführt, sondern wartet als Dienst hinter einer Eingangswarteschlange und beginnt beim Eintreffen von Daten sofort mit der Verarbeitung. Je nach Intensität der eingehenden Datenströme muss die Verarbeitungslast auf mehrere Server verteilt werden. Die Tatsache, dass jeder Server zustandslos ist, vereinfacht diese Vorgehensweise.

Ein weiterer Unterschied zu MapReduce besteht darin, wie Abhängigkeiten im gesamten Datenbestand berücksichtigt werden können. MapReduce kann hier beliebige Daten in der Map-Funktion mit einem gemeinsamen Key kennzeichnen und damit demselben Reducer für eine gemeinsame Auswertung zustellen. Storm dagegen verarbeitet eingehende Daten immer sofort und hat kein langes Gedächtnis, mit dem früher eingegangene Daten zu den aktuellen Daten in Beziehung gebracht werden können. Damit ist klar, dass nicht jedes mit MapReduce lösbare Problem mit Storm vollständig lösbar ist. Die exakte Lösbarkeit hängt von der Vollständigkeit der inkrementellen Eingabedaten in Bezug zur jeweiligen Fragestellung ab. Falls kein exaktes Ergebnis möglich ist, versucht man ein Ergebnis zu schätzen. Bei Storm wird deshalb auch von „möglicher Korrektheit“ (Eventual Accuracy) gesprochen.

Die Verarbeitung der Daten kann durch beliebig viele Callback-Funktionen spezifiziert werden. Wie diese auf die Daten angewandt werden, folgt nicht einem starren zweistufigen Schema wie bei MapReduce, sondern kann durch einen gerichteten Graphen („Topologie“) vom Anwender festgelegt werden. Die Eingangsknoten werden als Spouts bezeichnet, Knoten mit Vorgänger als Bolts, und die Datenkanäle zwischen diesen Knoten als Streams. Hinter jedem Knoten liegt eine Callback-Funktion, die aus Eingabedaten ein oder mehrere Werte-Tupel als Ergebnis berechnet und diese in bestimmte Ausgabestreams schreibt, um sie anderen Knoten zur Weiterverarbeitung zur Verfügung zu stellen. Andere Bolt-Knoten lesen wiederum aus diesen Streams und verarbeiten die Tupel weiter. Der Transport von Ergebnisdaten eines Knoten als Eingabe zu einem anderen wird durch vorhandene Kanten des gerichteten Graphen festgelegt. Mit diesen Mitteln hat der Programmierer viele Möglichkeiten, die Verarbeitung der Datenströme zu spezifizieren.

Das Storm Framework stellt sicher, dass die Berechnungen entsprechend der festgelegten Topologie und Callback-Funktionen ausgeführt wird. Dazu werden die die Callback-

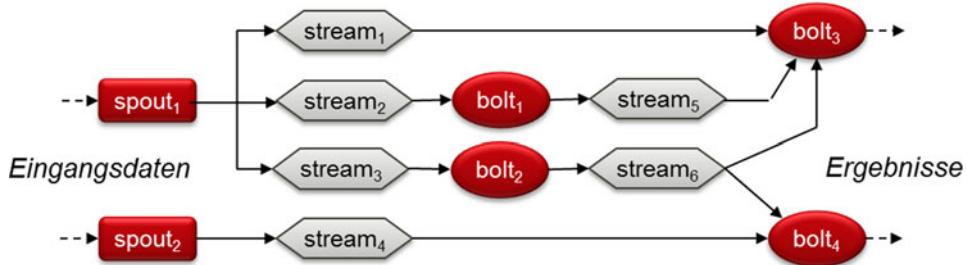


Abb. 4.22 Datenfluss als Storm-Topologie (Quelle: Fujitsu)

Funktionen auf Server im Cluster verteilt, die Daten-Tupel entsprechend der Topologie zugestellt und die Callback-Funktionen gestartet.

Ein Storm-Cluster besteht aus einem „Master Node“ und den „Worker Nodes“. Auf dem Master Node kommt der sogenannte „Nimbus“-Dämon zum Einsatz, der den für die Verarbeitung relevanten Code im Cluster verteilt, die Topologie in Teilaufgaben zerlegt, diese den Rechnerknoten zuteilt und die Überwachung des Clusters übernimmt. Auf den Worker Nodes läuft jeweils der „Supervisor“-Dämon, der den Ablauf der Arbeitsprozesse sicherstellt, die sich auf Grund der von Nimbus zugeteilten Teilaufgaben ergeben.

Die Koordinierung zwischen Nimbus und den Supervisors übernimmt Hadoop Zookeeper. Zookeeper ist jederzeit über den Gesamtzustand des Storm-Clusters im Bilde. Selbst wenn ein Nimbus-Dämon abbricht, kann ein solcher sofort wieder an anderer Stelle aktiviert werden. Dies macht die Gesamtkonfiguration äußerst stabil und gewährleistet somit, dass neu eintreffende, inkrementelle Daten in einer Lambda-Architektur stets in Echtzeit bearbeitet werden können.

4.3.8 Betrieb von Big Data-Infrastrukturen

Nachdem wir nun recht ausführlich über zusammengesetzte Konzepte und Referenzarchitekturen für Big Data mit ihren verschiedenen Technologiebausteinen gesprochen haben, bleibt die Frage, wie die IT-Infrastruktur für Big Data zweckmäßigerweise betrieben werden sollte.

Eine Option ist sicherlich der Eigenbetrieb im eigenen Rechenzentrum. Hier können vordefinierte, vorgetestete und vorintegrierte Konfigurationen, Appliances und optimal aufeinander abgestimmte Hardware- und Softwarekomponenten dazu beitragen, die Einführung und auch die Verwaltung entscheidend zu vereinfachen.

Der Aufbau, die Integration und der Betrieb einer Big Data-Infrastruktur ist jedoch aufwändig, erfordert Personal und entsprechendes Spezialwissen. Insbesondere bei Personalknappheit oder nicht vorhandenem Spezialwissen, kann es sinnvoll sein, sich auf seine Kernkompetenzen zu konzentrieren und den Betrieb der Infrastruktur einem Dienst-

leister zu überlassen, der diesen Service schneller, besser und auf Grund der erzielbaren Skaleneffekte auch kostengünstiger bereitstellen kann.

Immer mehr Unternehmen sind nicht mehr daran interessiert, komplexe Infrastrukturen im eigenen Rechenzentrum zu halten. In diesem Falle ist Big Data aus der Cloud eine interessante Alternative, die den Anwender schließlich von Vorabinvestitionen und allen Betriebsaufgaben befreit. Der Aufwand für Installation, Konfiguration und Wartung entfällt vollständig. Auch eine Kapazitätsplanung ist nicht mehr erforderlich. Die benötigten Kapazitäten lassen sich flexibel an den sich verändernden Bedarf anpassen, insbesondere erübriggt sich dann auch die häufig diskutierte Frage, ob die riesigen Mengen externer Daten tatsächlich durch die Unternehmens-Firewall geschleust werden müssen. Abgerechnet wird Big Data aus der Cloud nach Nutzung der angebotenen Services. Welche Parameter bei der Kostenfindung eine Rolle spielen, hängt vom Cloud-Dienst und vom Cloud-Dienstleister ab.

Es gibt aber auch gute Gründe, die Daten im eigenen Rechenzentrum zu belassen. So zum Beispiel, wenn es um interne Daten geht, bspw. Sensordaten von Fertigungsstraßen oder Videodaten von Sicherheitssystemen, die einen erheblichen Datenverkehr im Netz verursachen würden. Ebenso bei sensiblen Daten, bei denen der Verlust ein enorm hohes Risiko für das Unternehmen mit sich bringt, bzw. die auf Grund von Compliance-Anforderungen gar nicht in ein Public Cloud verlagert werden dürfen. Oder aber riesige Mengen flüchtiger Daten, die in einer derart hohen Frequenz anfallen, dass sie nicht schnell genug in die Cloud weitergeleitet werden können.

4.3.8.1 IaaS, PaaS, SaaS oder sogar Data Science als Service?

Wenn man erst mit Big Data aus der Cloud liebäugelt, stellt sich die Frage, was nun genau aus der Cloud geliefert werden soll. Am einfachsten ist es sicherlich, wenn man sich weder um die Analyse-Applikationen und die Middleware, noch die Infrastruktur inklusive Serversystemen, Speichersystemen und Netzen kümmern muss.

Unternehmen können die Infrastruktur eines Cloud-Anbieters inklusive Serversystemen, Speichersystemen und Netzkomponenten nutzen, kümmern sich aber selbst um Middleware und Analysetools. In diesem Fall spricht man von Infrastructure as a Service (IaaS). Wird zusätzlich zur Infrastruktur auch die Big Data Middleware, bspw. das Hadoop-Ökosystem vom Cloud-Anbieter bezogen, spricht man von Platform as a Service (PaaS). Bei Software as a Service (SaaS) nutzen Unternehmen auch die Analysetools des Cloud-Anbieters.

Somit bleibt für das Unternehmen „nur“ noch die Aufgabe, die so oft zitierte Nadel im Heuhaufen zu finden, also die Data Science-Fragen zu klären, für die man angeblich die händeringend gesuchten Data Scientists benötigt. Da diese sehr rar sind, gehen manche Cloud-Anbieter noch einen Schritt weiter und bieten Data Science as a Service an, ein Service, der dem Kunden quasi die Nadel im Heuhaufen sucht und findet.

Hinweis zum Hintergrund

- Apache Hadoop, Hadoop, HDFS, Flume, Sqoop, Ambari, Avro, Cassandra, Chukwa, HBase, Hive, Mahout, Pig, Spark und Zookeeper sind Warenzeichen der Apache Software Foundation.
- Oozie ist ein Warenzeichen von Yahoo! Inc.
- Cloudera Impala ist ein Warenzeichen von Cloudera.
- Ceph und CephFS sind Warenzeichen von Inktank Storage Inc.
- Esper ist ein Warenzeichen von EsperTech Inc.

4.4 Big Data-Analyse auf Basis technischer Methoden und Systeme

Christian Schulmeyer

Im folgenden Kapitel werden die Systemwelten und die Methoden dargestellt, die es erlauben, mittels komplexer Analyseverfahren zielgerichtet Informationen aus großen semi- und unstrukturierten Datenmengen zu extrahieren.

4.4.1 Herausforderungen an Big Data-Analyse

4.4.1.1 Was sind Big Data aus technischer Sicht?

Unter dem Begriff Big Data versteht man aus technischer Sicht sehr große un- oder semistrukturierte Datenmengen, die sehr inhomogen sind, also verschiedenartige Formate und Strukturen aufweisen und im Gegensatz zu Daten in Datenbanken kein einheitliches Schema aufweisen und aus vielen verschiedenen Quellen stammen können. Es gibt für diese Datenansammlungen keine einheitliche Beschreibung in Form von Metadaten und keinerlei Einschränkung in Bezug auf die Datentypen und die thematische Herkunft der Daten. Es können sowohl Daten aus den Bereichen Medizin, Gentechnik, Kernphysik oder anderen wissenschaftlichen Bereichen als auch Geschäftsdaten, Maschinen-, Audio- und Videodaten, Daten aus RFID-Lesern sowie E-Mails, SMS oder alle anderen Daten aus dem frei zugänglichen Internet und dem Bereich der sozialen Medien (Facebook, Twitter u. a.) sein. Durch den Mangel einer einheitlichen Beschreibung, eines Schemas oder von Metadaten, muss der Informationsinhalt dieser Daten anhand der Daten selber erkannt werden, der Inhalt der Daten muss *verstanden* werden, um sinnvoll Informationen extrahieren zu können.

4.4.1.2 Abgrenzung zu BI

Ein wichtiger Schritt hin zum Verständnis der Besonderheiten der Analyse un- bzw. semistrukturierter Daten ist die Abgrenzung zu den klassischen Verfahren der Business Intelligence (BI). Der Begriff Business Intelligence beschreibt die systematische Analyse

von *strukturiert* vorliegenden Geschäftsdaten, also Daten aus Datenbanken mit Buchhaltungsdaten, Produktionsdaten, Lagerhaltungsdaten usw. Diese basieren größtenteils auf zahlenbasierten, seltener auf textbasierten Daten, die mit klar definierten Schemata in meist relationalen Datenbanken vorliegen. Die BI-Analyseverfahren werden genutzt, um auf Basis der Unternehmensziele operative und strategische Entscheidungen durch Informationsgewinnung und -qualifikation zu unterstützen und Vorhersagen z. B. auf Basis von Trendanalysen durchzuführen. Die analytischen Methoden orientieren sich meist an Verfahren der Statistik, auch wenn es um die Analyse von in Textform vorliegenden Daten geht. Texte werden bei BI-Analysen aber von den analysierenden Systemen nicht *gelesen*, geschweige denn *verstanden*, sondern als abstraktes Element betrachtet und mittels statistischer Methoden verarbeitet. Begriffe sind hier Data-Mining, Text-Mining oder Online-Analytical Processing (OLAP). Die Analyseergebnisse werden in Einzelreports auf Papier oder sogenannten Dashboards dargestellt. Dashboards sind computergestützte grafische Oberflächen, die sehr übersichtlich die wichtigsten Analyseergebnisse leicht erfassbar darstellen. Klassisches BI basiert also in der Hauptsache auf strukturierten, in relationalen Datenbanken vorliegenden Zahlenmaterialien, wohingegen Big Data-Analysen auf unstrukturierten eher textbasierten, format- und beschreibungsfreien Datenformaten bestehen. In der Literatur werden Big Data-Analysen des Öfteren als *BI auf Texten* beschrieben.

4.4.1.3 Datenmengen

Auf das konstituierende Merkmal von Big Data – den immer weiter wachsenden Datenmengen – ist in Abschn. 2.1.1.1 bereits ausführlich eingegangen worden. Von *echten* Big Data beginnt man heute (2014) ab einer untersten Datenmenge von Terabytes (10^{12} Bytes) zu sprechen, die folgenden Stufen sind Petabytes (10^{15} Bytes), Exabytes (10^{18} Bytes) und Zettabytes (10^{21} Bytes). Um hier einen fassbaren Vergleich zu haben, drei Beispiele für Datenmengen in dieser Größenordnung.

Die aktuell umfangreichste Repräsentation des Weltwissens in (fast) allen Sprachen ist zurzeit die Online-Enzyklopädie Wikipedia. Die Gesamtzahl aller Artikel in allen Sprachen beträgt zurzeit (Februar 2014) 30 Millionen (Jüngling 2013, online)³, was bei einer angenommenen Durchschnittsgröße von 4 kB je Artikel einem Speicherbedarf von ca. 115 GB, also einer Größe, die auf jede heute handelsübliche Festplatte passt. Die Speicherkapazität des menschlichen Gehirns beträgt in etwa 2,5 Petabyte, also ein Datenbereich, den zum Beispiel das Online-Spiel World of Warcraft im Jahre 2013 spielend meisteerte, um alle Spieler auf allen Servern am Laufen zu halten. Nimmt man die Prognose für das Jahr 2020 mit 40 Zettabyte an, so sind das laut Expertenschätzungen ca. 57-mal die Anzahl aller Sandkörner an allen Stränden unserer Erde.

³ Vgl. <http://stats.wikimedia.org/DE/TablesArticlesTotal.htm> (zugegriffen am 20.05.2014).

4.4.1.4 Heterogenität der Datenquellen und der Datenformate sowie fehlende Beschreibung

Ein wichtiger Aspekt, der unter den Begriff Big Data subsumierten sehr großen Datenmengen und deren Analyse ist das Faktum, dass es sich um Daten handelt, die sowohl in ihrer Herkunft (Datenquelle) als auch in ihrer Art (Datenformat) extrem heterogen sein können. Wie bereits weiter oben beschrieben, können diese Daten aus nahezu beliebigen Quellen mit vielen unterschiedlichen Formaten stammen. Dies können Dokumente aus Textverarbeitungssystemen, Tabellenkalkulationssystemen, PDF, Präsentationen, HTML-Seiten, XML-Dateien, reine Textformate (ASCII oder UTF), Transskripte von Audio- und Video-Dateien, Maschinendaten im Binärformat, Transaktionsdaten von Web-Shops usw. sein. Ebenfalls sind die Datenquellen nicht eindeutig definiert. Dies kann der eigene Rechner unter dem Schreibtisch sein, ein zentraler Server, ein Speicherelement, eine Produktionsmaschine, eine zentrale Datenbank, ein Filesystem usw. Ein Ziel der Big Data-Analyse ist es, die Daten ganz verschiedener Datenquellen zu analysieren und die Analyseergebnisse *miteinander in Beziehung zu setzen*. Dieses Ziel bedingt, dass quasi alle geeigneten Datenquellen für Big Data-Analysen zum Einsatz kommen müssen.

Ein weiterer Grad von Komplexität kommt hinzu, da die Daten nicht beschrieben sind. Es gibt keine Instanz, die einer Big Data-Analysesoftware mitteilt, was die gerade zu analysierenden Daten darstellen oder beinhalten. In einer Datenbank gibt es eine formale Beschreibung der Daten. Diese wird Datenbankschema genannt. So weiß ein analysierendes System, dass sich z. B. in einer bestimmten Spalte ein Zahlenwert im Integer-Format befindet, der eine Geldsumme darstellt. Bei abgespeicherten Texten in einer Datenbank sind ggf. Metadaten vorhanden, die aussagen, dass diese hier abgespeicherten Textdateien wöchentliche Umsatzreports sind, die von einem bestimmten Mitarbeiter erstellt wurden. Die Entität *Mitarbeiter* ist in der Metadatenbeschreibung immer im gleichen Format und an der gleichen Stelle zu finden, also für ein System eindeutig erkennbar. Wesentliches Merkmal von Big Data-Analysen ist das weitgehende Fehlen solcher Informationen und die dadurch notwendig gewordene Fähigkeit, den Inhalt der zu analysierenden Daten *inhärent zu erkennen* und diese Erkenntnis bezüglich des Analysezwecks hin anzuwenden.

Aus dem Vorgenannten ergibt sich, dass die Big Data-Analyse keine singuläre Technologie sein kann, sondern immer eine Mischform aus verschiedenen Verfahren sein muss, die sich je nach Erkenntnisziel (rein informativ zur Entscheidungsunterstützung), Aktion (z. B. Prozesssteuerung) und zeitlicher Kritikalität (Echtzeitanalyse) ausrichtet.

4.4.2 Daten

4.4.2.1 Unstrukturierte und semistrukturierte Daten

Ein wesentliches Merkmal von Big Data-Analysen ist die *Unstrukturiertheit* der zu Grunde liegenden Daten. Das Merkmal „unstrukturiert“ ist dann gegeben, wenn die Daten keinerlei bzw. nur rudimentären formalen Ordnungskriterien unterliegen. Ein formales Ordnungskriterium wäre eine Feldbezeichnung innerhalb eines Dokuments, eine bezeich-

nete Spalte in einer Datenbank oder eine Metabeschreibung über Form und Inhalt einer Datei. Hochstrukturierte Daten findet man in relationalen Datenbanken, die einem einheitlichen Datenmodell und einem Datenbankschema unterliegen. Jedes Feld wird in Inhalt und Format beschrieben, ist eindeutig bestimmbar und die Verknüpfung zu anderen Datenbankelementen ist klar definiert. Ein Beispiel für völlig unstrukturierte Daten ist eine reine Textdatei mit einem fortlaufenden Text, ohne Metadaten und sonstigen Beschreibungen.

Die weitaus meisten Daten, mit denen man es in der Big Data-Analyse zu tun hat, sind rein formal gesprochen *semistrukturierte Daten*. Dies sind zum Beispiel Daten aus Textverarbeitungsprogrammen, PDF-Dateien, Internetseiten (HTML-Dateien), E-Mails oder Präsentationsdateien. Diese enthalten einige wenige beschreibende Elemente (in E-Mails zum Beispiel das „an:“-Feld, den Betreff, bei einer Datei aus dem Textverarbeitungsprogramm, den Autor, die Zeichenanzahl usw.). Da es bei der Big Data-Analyse jedoch hauptsächlich auf die inhaltlichen Daten ankommt und diese unstrukturiert sind, werden die semistrukturierten Daten im allgemeinen Sprachgebrauch ebenfalls unter die unstrukturierten Daten gezählt. Echte semistrukturierte Daten sind zum Beispiel XML-Dateien und Daten aus Tabellenverarbeitungsprogrammen. Hier können lange unstrukturierte Textblöcke vom Markups⁴ (XML) eingeschlossen oder in Tabellenspalten (Tabellenverarbeitungsprogramm) abgelegt sein.

4.4.2.2 Text und nicht-Text-Formate (Audio, Video, Grafik, Bilder)

Die Unterscheidung der unstrukturierten Daten in Text und nicht-Text-Formate ist ein wichtiger Bestandteil für das Verständnis der Komplexität der Big Data-Analyse. Ein nicht unerheblicher Teil der Daten, die täglich produziert werden, sind Audio- und Video-Dateien und nicht zu vergessen die Bilderflut, die im Internet vorhanden ist bzw. jeden Tag von Millionen von Nutzern hochgeladen wird. Ein Ziel der Analyse sehr großer Datenmengen ist die Analyse von Audiostreams. Ein Audiostream ist ein fortlaufender Datenstrom mit digitalisierten Tonaufnahmen, hier speziell Sprachaufnahmen. Für eine Analyse dieser Audiodaten müssen diese nach dem aktuellen Stand der Technik zuerst in eine Textform transkribiert werden. Der Datenstrom wird mittels des technischen Verfahrens der Spracherkennung in Text umgewandelt und dieser Text dann entweder mittels In-Memory-Technologien⁵ in nahe-Echtzeit analysiert oder in Dateien abgelegt und dann unter weniger zeitlich kritischen Bedingungen analysiert. Ähnlich geschieht dies bei Videoformaten (z. B. Nachrichtensendungen, die als Videostreams über das Internet erhältlich sind). Bei diesen Formaten ist die Bild- und Tonspur getrennt und man transkribiert die Tonspur. Noch in den Anfängen steckt die Technologie zur Analyse echter Bilddaten in Fotos, Grafiken oder Videos, doch gibt es hier auch erste vielversprechende Ansätze, sinnvolle Analyseergebnisse aus diesen Daten zu erzeugen. Als Beispiel möge hier die Analyse chemischer Strukturformeln dienen, die als Grafik in Texten vor-

⁴ Markups sind Textauszeichnungen innerhalb von Auszeichnungssprachen um Strukturmerkmale darzustellen.

⁵ Hierzu näher in Abschn. 4.3.4.

liegend bereits sehr zufriedenstellend erkannt und analysiert werden können. Auch die Erkennung bestimmter Bildelemente, wie Menschen, Gesichter, Fahrzeuge, Gebäude ist schon weit gediehen. Schwierigkeiten bereitet hier noch das Erkennen von komplexen Handlungsabläufen. Besonderheit der Verarbeitung von Bilddaten ist die benötigte große Rechenleistung der beteiligten Computer. Insbesondere wenn es in Richtung einer Echtzeitverarbeitung geht, sind hier noch weitere Entwicklungen nötig, bis eine Praxisreife erreicht werden kann. Doch scheint dies in Betracht der aktuellen technischen Entwicklung nur eine Frage der Zeit zu sein.

4.4.2.3 Multilinguale Daten

Dadurch, dass die Analysen im Big Data-Bereich vorwiegend auf Texten basieren, ist die Fragestellung der Multilingualität von wesentlicher Bedeutung. Jeden Tag werden Hunderttausende von Dokumenten in allen beliebigen Sprachen dieser Erde erstellt. Das Online-Lexikon Wikipedia beinhaltet Artikel in 236 Sprachen (Stand Februar 2014)⁶. Da Big Data-Analysen vorwiegend textbasiert sind und diese Texte von den analysierenden Systemen auch verstanden werden müssen, ist das Beherrschen mehrerer Sprachen von großer Wichtigkeit. Ziel der Analysen ist es ja, über eine große Varietät von Datenquellen – also auch eine große Varietät von Sprachen – übergreifende Analysen zu machen und diese Datenquellen miteinander zu verknüpfen. Dies funktioniert nur, wenn Textinformationen in den wichtigsten Sprachen maschinell gelesen und verstanden werden können, was sehr hohe Ansprüche an die Qualität der linguistischen und semantischen Fähigkeiten der Systeme bedingt, die Big Data-Analysen erstellen sollen. Ein elementarer Schritt ist es, die in einem Dokument vorgefundene Sprache zu erkennen, um den nachfolgenden Analyseschritten die Grundlage des Textverständnisses geben zu können.

4.4.2.4 Datenzugriff

Wie der Name Big Data schon sagt, geht es hier um sehr große Mengen von Daten. Ein Verschieben oder Kopieren von Daten vom ursprünglichen Ort verbietet sich daher, sowohl aus Sicht der dafür einzusetzenden IT-Ressourcen, der benötigten Bandbreiten für die Datenübertragung und aus ökonomischen Gründen. Dies begründet sich in der hohen Bandbreite, die eine Bewegung dieser Daten verursachen würde und der dann gegebenen Redundanz der Daten, die komplexe Abgleichprozesse notwendig machen, die selber wiederum Bandbreite benötigen. Weiterhin beobachtet man oft, dass – falls diese Daten manuell erstellt werden – die Ersteller dieser Daten oft mit Widerstand reagieren, wenn ihre Arbeitsprozesse verändert werden. Ziel muss es daher sein, die Daten an ihrem ursprünglichen Ort zu belassen und über geeignete Mechanismen auf sie zuzugreifen, ohne sie selbst bewegen zu müssen.

Hierfür gibt es in der IT mehrere geeignete Verfahren. Grundlage aller Verfahren ist ein Zugang zu den Daten über eine Netzwerkverbindung jedweder Art, die aber als gegeben angenommen werden kann. Der einfachste Prozess ist das Crawling-Verfahren. Hier

⁶ Vgl. <http://stats.wikimedia.org/DE/TablesArticlesTotal.htm> (zugegriffen am 20.05.2014).

wird ein spezielles Softwareprogramm eingesetzt, welches die Daten über das Netzwerk an dem Ort voranalysiert, an dem sie liegen, die notwendigen Daten überträgt und dem Big Data-Analysesystem zur Verfügung stellt. Diese Programme werden auch Spider oder Searchbot genannt. Dieses Verfahren wird in den meisten Fällen für Internet- (Intra- und Extranet-)Quellen angewendet. In gewissem Umfang können Crawler auch für Dateien verwendet werden. Weitere Verfahren werden unter dem ETL-Prozess subsumiert. ETL steht für: Extract Transform Load. Hier werden nur die notwendigen Daten aus den Originaldaten extrahiert, für die Big Data-Analyse entsprechend transformiert und zentral geladen. Der ETL-Prozess kann Daten aus den verschiedensten Quellen und in verschiedenen Formaten verarbeiten und konsolidieren.

4.4.3 Systemische Grundlagen

4.4.3.1 Indexerstellung

Grundlegend für das Verständnis von Big Data-Analysen ist die Tatsache, dass diese Analysen fast nie auf den Originaldaten stattfindet. Ähnlich wie Google nicht direkt „im Internet“ sucht, sondern in einem auf Basis der Originaldaten erstellten Index, finden Analyseschritte bei Big Data-Analysen meistens ebenfalls auf einem Index statt. Dieser Index kann eine sehr große Datei sein, die man sich als mit einer speziellen Struktur versehene Tabelle vorstellen kann. In dieser Tabelle stehen alle Informationen, die der Crawler vorher gefunden oder der ETL-Prozess extrahiert hat. Diese Informationen werden mit vielen weiteren Informationen angereichert, um aussagekräftiger zu werden. Der Index enthält des Weiteren noch eine Art Wegweiser zur Originalquelle der Information. Der Index wird in einem zu definierenden zeitlichen Abstand immer wieder aktualisiert, um immer die aktuellen Informationen zu haben. Da wir über Big Data sprechen, ist die initiale Erstellung eines Index sehr zeitintensiv, je nach den für die Big Data-Analyse benötigten Zusatzdaten kann die Erstellung eines Index mehrere Stunden bis mehrere Tage dauern. Dieser Index überführt un- oder semistrukturierte Daten in eine für die Big Data-Analyse geeignete Strukturform (Berman 2013, S. 1–13).

4.4.3.2 In Memory Computing

Da ein zu erstellender Index selbst sehr viele Gigabyte groß sein kann, sind hier Zugriffsmethoden notwendig, die extrem schnelle Zugriffszeiten ermöglichen. Diese schnellen Zugriffszeiten kann man im Allgemeinen nicht mittels einer Datenspeicherung auf magnetische oder optische Speichermedien nicht erreichen. Insbesondere weil eine Big Data-Analyse aus sehr vielen kleinen Leseoperationen besteht, sind die Zugriffszeiten auch auf modernen Festplattenlaufwerken zu lang.

Es müssen adäquate Methoden zur Beschleunigung dieser Prozesse bereitgestellt werden. Einer davon ist das In-Memory Computing, welches ausführlich in Abschn. 4.3.4 dargestellt ist.

4.4.3.3 MapReduce

Das MapReduce-Verfahren wird in Abschn. 4.3.2 ausführlich beschrieben, hier soll daher nur der grundlegende Mechanismus anhand eines Beispiels erläutert werden.

Das MapReduce-Verfahren ist eine Funktion, in der sehr große Datenmengen in einzelne in sich konsistente Blöcke aufgeteilt werden und damit eine Parallelisierung der Verarbeitung auf mehrere Einzelrechner in einem Cluster (miteinander vernetzte Rechnergruppen) ermöglichen (bzgl. Clustering vgl. Agrawal et. al. 2013, S. 192–211). Das Verfahren besteht aus einem *Map-Prozess*, der den vorgefundenen Daten eine eindeutige Position innerhalb des Datenblocks zuordnet (damit einzelne identische Datenwerte in jedem Block eindeutig identifiziert werden können), jedem Datenwert einen Schlüssel (z. B. Datenwert = Temperatur, Schlüssel = Jahr) zuweist und diese sinnvoll auf Basis des Schlüssels gruppiert (alle Temperaturwerte des Jahres 1980 werden dem Schlüssel 1980 zugewiesen). Wenn es für jeden Tag des Jahres 1980 einen Eintrag gab, ist somit in den Daten aus 365 Mal der Jahreszahl „1980“ mit jeweils einem zugeordneten Temperaturwert nur einmal die Jahreszahl „1980“ mit 365 zugeordneten Temperaturwerten erzeugt worden, also 1/365 der ursprünglichen Datenmenge der Zeichenkette „1980“. Dies wird in allen Blöcken durchgeführt.

Im *Reduce-Prozess* werden die wie oben bearbeiteten Blöcke wieder zusammengeführt. Wurde die Ursprungsmenge der Daten in z. B. 1000 Blöcke aufgeteilt, so befindet sich nach dem Map-Prozess in jedem Block nur noch einmal die Jahreszahl 1980 mit allen 365 dazugehörigen Temperaturwerten, während der Zusammenführung werden jetzt die Datenwerte nochmal genau einmal dem gemeinsamen Schlüssel (hier der Jahreszahl 1980) zugewiesen, sodass nun in den wieder zusammengeföhrten Daten statt 1000 Mal die Jahreszahl 1980 diese nur noch ein Mal vorhanden ist. Für die Jahreszahl 1980 ist die Datenmenge in der Form verkleinert worden, dass von ursprünglich 365.000 Mal der Zeichenkette „1980“ genau einmal die Zeichenkette „1980“ übrig geblieben ist, eine Reduzierung auf 1/365.000 der ursprünglichen Datenmenge dieser Zeichenkette. Dieser Prozess wird für alle anderen Schlüsselbegriffe im Datenkorpus ebenfalls durchgeführt. Das Ergebnis der Zusammenführung aller Blöcke wird dann in eine neue Datei geschrieben, die alle für eine Analyse relevanten Informationen enthält, aber um ein vielfaches kleiner ist als die des ursprüngliche Datenkorpus.

4.4.3.4 Skalierbarkeit

Bei der Analyse von sehr großen Datenmengen hat die Skalierbarkeit der zu Grunde liegenden Software als auch der Hardware eine besondere Relevanz, da hier hohe Verarbeitungsgeschwindigkeiten Grundlage einer ökonomischen und effektiven Datenverarbeitung sind. Um mit großen Datenmengen umgehen zu können, müssen die Systeme so ausgelegt sein, dass sie a) mit wechselnden Lastanforderungen zureckkommen und b) auch sehr große Datenmengen ohne Ausfälle und in ausreichender Geschwindigkeit bewältigen können. Insbesondere den Antagonismus zwischen Datenmenge und Verarbeitungsdauer gilt es zu meistern. Nimmt die Datenmenge zu, steigt auch die Verarbeitungsdauer. Die Reaktion darauf ist, schnellere Rechner zu verwenden. So skaliert gute

Tab. 4.4 Skalierung von Systemen bei der Indizierung großer Datenmengen

Testdaten	Menge	Größe	Cores	Dauer
Ein Tag Twitter weltweit (2010)	200.000.000 Tweets	0,6 KB/tweet	16	12:41 h
			640	19 sec
Deutsches Wikipedia (2010)	1.6 Mio. Artikel	4 KB/doc	16	1:59 h
			640	3 sec

Big Data-Analyse-Software linear proportional zu der Anzahl der Prozessorkerne (Cores), im Idealfall ist eine 16-Core-Maschine achtmal so schnell wie eine Dualcore- (zwei-Core) Maschine. Tabelle 4.4. zeigt die Skalierungseffekte bei der Indizierung eines Datenkorpus.

Eine weiterführende Beschreibung von Skalierungsmethoden befindet sich in Abschn. 4.3.

4.4.4 Methoden

4.4.4.1 Suche ist nicht gleich Suche

Jede Datenanalyse basiert auf dem Verarbeiten von Daten verschiedener Art und Provenienz. Bei strukturierten Daten ist dies recht einfach, weil man genau weiß, wo in einer Datenbank z. B. die Tage und Monate abgespeichert sind und wo die dazugehörigen Umsatzzahlen liegen. In Kenntnis dieser Zahlen kann man den ganzen Methodenbaukasten der BI nutzen. Schwieriger wird es, wenn nicht klar definierte Informationen (wie Worte, Begriffe, ganze Sätze oder Verknüpfungen) erkannt und in die Analyse miteinbezogen werden sollen. Höchst komplex wird es aber nun, wenn statt strukturierter Daten unstrukturierte Daten vorliegen, es sich also in erster Linie um textuelle Daten handelt und es nicht bekannt ist, welche Informationen wo vorliegen. Es müssen vor jeglicher Analyse erst die zu analysierenden Informationen und Zusammenhänge in den Daten erkannt bzw. gefunden werden.

Daraus folgt, dass ein initialer Bestandteil einer Datenanalyse bei un- bzw. semistrukturierten Daten die Suche nach den zu analysierenden Datenelementen innerhalb eines Datenkorpus ist. Einer Suche geht – insbesondere bei sehr großen Datenmengen – immer die Indizierung der zu durchsuchenden Daten voran. Sucht man bestimmte Informationen im Internet, z. B. bei Google, wird hier natürlich nicht „im Internet“ danach gesucht, sondern in dem riesigen Index, den Google auf Basis der gefundenen Daten aus dem Internet aufbaut. Ein Index ist eine Datei, die nach einem festgelegten Schema aufgebaut ist und es einer Suchmaschine erleichtert, bestimmte Worte oder Phrasen mit einer bestimmten Lokation (im Internet wäre das eine Internetseite, im Big Data-Bereich vielleicht ein einzelnes Dokument) zu verknüpfen. In den Index können weitere zusätzliche Informationen einfließen wie der Kontext, in dem diese Information gefunden wurde, Stammformen von Verben, zeitliche Angaben, Synonyme usw. Wichtig bei Google ist in diesem Zusammenhang zum Beispiel die Verlinkung der zu findenden Seite mit anderen Seiten. Je mehr

andere Seiten im Suchzusammenhang auf diese Seite verlinken, desto höher wird sie im Suchranking stehen.

Jedoch ist gerade im Kontext der Big Data-Analyse *Suche nicht gleich Suche*. Google ist zwar sicher die bekannteste und von den im Index vorhandenen Volumina sicher größte Suchmaschine, doch für die speziellen Aufgaben einer Big Data-Analyse nur sehr bedingt geeignet. Dies liegt daran, dass ein wichtiger Bestandteil von Google für die Qualität eines Suchtreffers die oben beschriebene Verlinkung von anderen Seiten ist (was für Google ein Merkmal ihrer Qualität darstellt). Im Kontext einer Big Data-Analyse ist dies jedoch nicht maßgeblich, da es meist um die durch die Verlinkung ausgedrückte „Beliebtheit“ von Internetseiten geht und Querverlinkungen eine geringe Relevanz haben. Ziel bei einer Suche nach Big Data ist es, genau das Richtige sicher zu finden und nicht Vieles oder Beliebtes. Das heißt, der Suchalgorithmus muss *verstehen* was gesucht wird. Ein weiteres wichtiges Merkmal für eine qualitativ hochwertige Suche im Big Data Bereich ist es, dass die Maschine Inhalte und Bedeutungen erfassen kann. Man kann einen identischen Sachzusammenhang mit völlig verschiedenen Worten beschreiben; ein Mensch erkennt, dass das Gleiche gemeint ist – im Kontext der Big Data-Analyse muss eine Suchsoftware dies auch können und die relevanten Angaben, die für die Analyse wichtig sind, trotz verschiedener Syntax und Grammatik eindeutig erkennen, entsprechend auszeichnen und in den Index übernehmen.

So ist für die Suche im Big Data-Bereich die Fähigkeit, Texte zu „lesen“ und zu verstehen, Entitäten in Daten zu erkennen und diese in den richtigen Sinnzusammenhang zu bringen, von elementarer Bedeutung. Hierfür sind eine linguistische und morphosyntaktische⁷ Analyse sowie ein semantisches Verständnis von Texten die Grundlage.

4.4.4.2 Keywordbasierte Suche

Die keywordbasierte Suche ist die einfachste Form der Suche in großen Datenmengen. Diese Suche basiert auf der Nutzung des oder der eingegebenen Suchworte als Schlüsselwörter nach denen in einem Datenkorpus gesucht wird, ohne auf Kontexte, Synonyme, Hyperonyme oder Homonyme usw. zu achten. Die keywordbasierte Suche kann optimiert werden, in dem sie mit linguistischen Methoden angereichert wird, sodass Worte auf Grundformen zurückgeführt oder Synonym- und Homonymlisten beigefügt und angewendet werden, um mit dem Keyword „Auto“ auch „KFZ“ zu finden. Weitere Optimierungen sind statistische Methoden, die zum Beispiel Häufungen von Keywords in Dokumenten erkennen oder bei der Eingabe von mehreren Keywords deren räumliche Nähe zueinander im Text eines Dokuments berücksichtigen. Keywords können weiterhin mittels boolescher Operatoren⁸ miteinander verknüpft werden. So findet die Sucheingabe *Auto AND Zeitschrift* Dokumente, in denen die Worte Auto und Zeitschrift zusammen vorkommen.

⁷ Die morphosyntaktische Analyse beschreibt die grammatisch basierte Zerlegung von Sätzen in ihre Grundbestandteile bzw. -formen.

⁸ Boolesche Operatoren sind Ausdrücke der booleschen Algebra wie AND, OR, EXOR, NOT, NEAR usw.

Der Ausdruck *Porsche NEAR 914* wird Dokumente liefern, in denen im Text das Wort Porsche und die Zahl 914 sehr eng zusammenstehen. Da die Keywordsuche nur *Zeichenketten* vergleicht, ohne ihre linguistische Bedeutung oder ihre Semantik zu kennen, kann sie in keinem Fall Inhalte und Bedeutungen bzw. Relationen, Kontexte und Ähnlichkeiten erkennen. Dies ist jedoch bei textbasierten und unstrukturierten Daten für eine hohe Qualität der Suchergebnisse unerlässlich.

4.4.4.3 Linguistik und Semantik

Die als Texte vorliegenden un- oder teilstrukturierten Daten müssen in einem ersten Schritt *normalisiert* werden. Dies geschieht einmal durch eine Beseitigung aller nicht rein textuellen Elemente in den Daten (z. B. Markups) und einer Transformation in ein einheitliches Format (z. B. bei unterschiedlichen Zeichencodierungen). Danach folgt eine morphosyntaktische und linguistische Bearbeitung der Texte, der eine Spracherkennung vorgeschaltet ist. Hier werden die Texte tokenisiert (z. B. Erkennung der Wortgrenzen mittels Leerzeichen, Erkennung von Satzzeichen usw.), die Sätze segmentiert (z. B. Erkennen von Satzgrenzen, Haupt- und Nebensätzen) und auf ihre Normalformen zurückgeführt. Dies ist die Wortstammreduktion (z. B. „Flugzeuges“ und „Flugzeugen“ auf „Flugzeug“) sowie die Lemmatisierung, die z. B. ein Verb auf seine Grundform zurückführt (z. B. „flog“ auf „fliegen“). Weitere höherwertige Aufgaben dieser Teilanalyse sind das *Part-of-Speech Tagging*, welches den erkannten Wörtern die Wortarten (Verb, Adjektiv, Nomen usw.) zuweist und das Erkennen von Bezügen im Text auch komplexerer Art (z. B. ein Ereignis wird mit einer Person in Verbindung gebracht). Hierfür ist noch eine *Entity-Extraction* notwendig, die im Text bestimmte Elemente (Entities) erkennt. Dies können Eigennamen, Städte, Länder, Daten, Ereignisse usw. sein.

Wenn nun aus den Daten ein Index erstellt wird, werden alle Informationen, die aus den oben beschriebenen Analysen gewonnen wurden, mit in den Index eingebracht und stehen als weitere Informationen für eine Suche und/oder Analyse zur Verfügung.

Im nächsten Schritt werden nun semantische Verfahren eingesetzt. Die *Semantik* ist die Lehre von der Bedeutung der Zeichen⁹, also auch von Worten und Sprache. Man weiß, dass ein Wort viele verschiedene Bedeutungen haben kann. Diese kann vom Kontext, von der Stellung im Satz, vom Sprecher, vom kulturellen Hintergrund oder von vielen anderen Einflussfaktoren abhängen. Man möchte das System für eine Big Data-Analyse in die Lage versetzen, das „Gelesene“ zu „verstehen“. Ein treffendes Beispiel sind die Bezeichnungen für Geld. Dieses kann auch Kohle, Moos, Penunze, Mäuse oder Kies (und vieles anderes mehr) genannt werden. Jedes dieser Worte hat eigentlich eine völlig andere Semantik, bezeichnet also völlig andere Dinge als Geld. Doch abhängig vom Kontext oder z. B. dem sozialen Hintergrund des Sprechers kann eines dieser Worte auch die Bedeutung Geld haben.

⁹ Aus dem altgriechischen *σημαίνειν* (sēmaínein) = *bezeichnen* oder *zum Zeichen gehörig*.

Man kann ein und denselben Sachverhalt in völlig verschiedenen Worten beschreiben oder etwas völlig Verschiedenes in sehr ähnlichen Worten. Eine Analyse ohne semantische Bearbeitung wird dies nicht erkennen können und wichtige gesuchte Inhalte eines großen Datenkorpus übersehen oder falsche Zusammenhänge liefern.

Ziel einer Analyse großer unstrukturierter Datenmengen ist es nun, neben den oben beschriebenen rein textuellen Merkmalen auch die Bedeutung des Geschriebenen richtig zu erkennen und als Annotation mit in den Index aufzunehmen, sodass diese zusätzliche Information für eine Suche auch zur Verfügung steht. Als Beispiel möge folgende Suchanfrage dienen:

Frage: „Wo trifft Angela Merkel Barack Obama?“

Gefunden werden müssen Dokumente, die Folgendes beinhalten:

„Frau Bundeskanzlerin Merkel plant während des Weltwirtschaftsgipfels in Davos ein Treffen mit den Vertretern der G8-Staaten. Der Präsident der USA, Barack Obama, wird persönlich vor Ort sein.“

„Die Deutsche Bundeskanzlerin wird sich im Januar mit dem wiedergewählten amerikanischen Präsidenten treffen.“

„Wir werden eine Demo starten, wenn sich die Angie mit dem Barack in der Schwyz trifft.“

„The German Chancellor Angela Merkel will meet the President of the United States during the Summit in Davos.“

Weitere sehr wichtige Elemente der semantisch erweiterten Analyse sind Domänenwissen und Wissensmodelle, also das Wissen um sehr spezielle Wissensbereiche und die dort inhärenten Relationen und Verknüpfungen. Hierfür nutzt die semantische Analyse von Texten spezielle Wissensmodelle, Taxonomien und Ontologien.

4.4.4.4 Wissensmodelle, Taxonomien und Ontologien

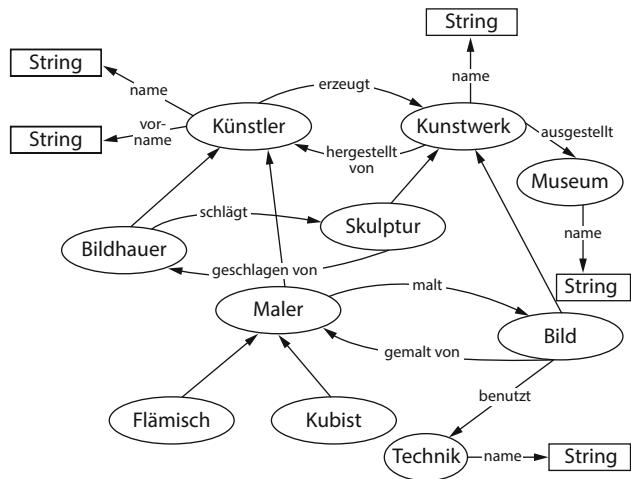
Ein Wissensmodell fasst in geeigneter Weise Begriffe und Kategorien eines bestimmten Wissensbereiches zusammen. Allgemein werden Wissensmodelle neben den oben beschrieben syntaktischen Analysen mittels Taxonomien, Wissensmodellen (auch Wissenslandkarten genannt) oder Ontologien beschrieben (vgl. Sathi 2012, S. 31–46).

Eine Taxonomie ist ein gerichtetes Modell, das Begriffe hierarchisch von einer hohen zu einer niedrigeren Ebene ordnet und somit die Relation zwischen Ober- und Unterbegriffen darstellt. Bekanntestes Beispiel sind hier die taxonomischen Modelle in der Biologie, in denen Lebewesen in Reiche, Stämme, Klassen, Ordnungen usw. dargestellt werden¹⁰.

Für die semantische Erfassung von Informationen in Daten ist eine Taxonomie aber nicht ausreichend. Hier werden andere Darstellungsformen benötigt. Diese findet man in Form von Wissenslandkarten oder – übertragen in die Informatik – in Ontologien. Diese Ontologien werden als ungerichtete Graphen dargestellt, die die Beziehungen und

¹⁰ Eine der ersten modernen taxonomischen Ordnungen wurde von Carl von Linné entworfen, bekannt als das Linnésche System, welches in seinen Grundlagen heute noch in der Biologie Gültigkeit hat.

Abb. 4.23 Beispiel einer einfachen Ontologie (übernommen aus Wikipedia)



Zusammenhänge zwischen Begriffen eines bestimmten Wissensgebietes darstellen. Sie stellen eine formale Spezifikation eines in einem Wissensgebiet verwendeten Vokabulars bzw. dessen Bedeutungen dar und erlauben es, Bedeutungen und Begriffe, Relationen und bestimmte Objekte miteinander zu verknüpfen. In diesen Wissensmodellen werden z. B. Homonyme, Hyperonyme und Synonyme aufgenommen, sodass eine kontextuelle Erschließung von Wort- und Satzbedeutung in Texten möglich wird (Berman 2013, S. 35–48). Für eine semantische Analyse von Texten sind Ontologien daher den reinen Taxonomien überlegen.

Die Knoten stellen die Begriffe (Konzepte) dar, die Kanten geben die Relation zueinander an. Während einer semantischen Analyse von Texten nutzt das Analysesystem die Ontologie, blickt quasi „durch“ die Ontologie auf die Informationen im Text und kann diese somit kontextuell und in ihrer spezifischen Bedeutung erfassen und verstehen.

Technisch gesehen wird beim Erstellen des Index dieser semantisch annotiert, d. h. um weitergehende, aus dem Wissensmodell (z. B. Synonym-, Homonym-, Hyperonym- und Antonymlisten, Taxonomien und Ontologien) entnommene und im direkten Kontext des gerade analysierten Textes stehenden Informationen angereichert, sodass bei einer Suche innerhalb eines solchen Index alle durch die genutzte Ontologie vorhandenen Zusatzinformationen (als Annotationen im Index) für das erfolgreiche Finden von Informationen genutzt werden können.

So würden – um bei dem in obiger Abbildung gezeigten Beispiel zu bleiben – bei der Suche nach dem Konzept „Kubismus“ auch Dokumente über den Maler Georges Braque oder das Metropolitan Museum of Art in New York, welches seit 2013 eine sehr große kubistische Sammlung beherbergt, gefunden werden.

4.4.4.5 Assoziative Methoden der Suche

Die assoziative Suche ahmt rudimentär den Prozess nach, der im menschlichen Gehirn abläuft, wenn wir versuchen, etwas aktuell über unsere Sinnesorgane wahrgenommenes mit einer unserer Erinnerungen – also unseren gemachten Erfahrungen – in Verbindung zu bringen. Hier geht es nicht darum – wie in der keywordbasierten Suche – einen exakten Treffer zu finden, sondern wir nutzen Assoziationen; also Ähnlichkeiten zwischen dem eben Erfahrenen und dem bereits in den Erinnerungen gespeicherten. Schon Philosophen wie Arthur Schopenhauer erkannten, dass konkretes unmittelbares Wissen zu einer konkreten Begebenheit erst dann praktisch nutzbar wird, wenn es abstrahiert, also auf eine höhere Ebene gehoben wird, und somit auch auf andere, ähnliche bzw. vergleichbare Erfahrungen anwendbar ist. Dieses eher heuristische Vorgehen erlaubt es, die „best matches“ durch Assoziation, also dem Verknüpfen von ähnlichen im Speicher (beim Menschen das Gehirn) liegenden Informationen, mit der gerade in die Suche eingegebenen (vom Menschen gesehenen oder gehörteten) Information in Relation zu bringen.

Technisch gesprochen basieren die assoziativen Suchmethoden auf den Beziehungen zwischen Informationen auf Basis ihrer linguistischen, semantischen oder inhaltlichen Ähnlichkeiten. Für die Suchmethode heißt das, dass sie den Text der Bedeutung nach „verstehen“ muss. Sie muss erkennen, dass der Satz „dies ist ein Buch von Hermann Hesse“ eine völlig andere Bedeutung hat als der Satz „dies ist ein Buch über Hermann Hesse“. Google würde hier keine Unterscheidung machen. Weiterhin ist es für eine assoziative Suche notwendig, Beziehungen in argumentativen Ketten zu erkennen und richtig zu deuten. Als Beispiel diene folgender Text: „Der Airbus A380 ist das größte Passagierflugzeug der Welt. Die Endmontage dieses Airbus-Typs findet in Toulouse statt. Es gilt als eines der leisten in Betrieb befindlichen Verkehrsflugzeuge.“ Hier muss erkannt werden, dass sich das Verb *leise* auf den Flugzeugtyp *Airbus A380* bezieht, obwohl die Phrase *Airbus A380* im eigentlichen Satz und im mittleren Satz nicht genannt wird. Eine typische Assoziation wäre nun, wenn man die obigen drei Sätze in eine entsprechende Suchmaschine eingibt und u. a. ein Dokument findet, das folgenden Inhalt haben könnte: „Die Boeing 747 ist inzwischen seit 46 Jahren in Betrieb und kann in Ihrer neuesten Version 747-8 durchaus noch mit der europäischen Konkurrenz von Airbus mithalten, viele Passagiere empfinden jedoch die Geräuschkulisse im Inneren der Kabine als zu laut.“

Anhand dieses Beispiels kann man auch sehr gut die Nutzung der im obigen Abschnitt beschriebenen Ontologien beschreiben, die das Suchsystem für seine Assoziationen nutzt. Sowohl der Airbus A 380 als auch die Boeing 747 sind beides Großraumflugzeuge für den Passagierbetrieb. Hier ist schnell eine Bedeutungsübereinstimmung gefunden. Weiterhin bestehen Beziehungen zwischen den Begriffen „Airbus“ und „Boeing“ als Wettbewerb und „Geräuschkulisse“ sowie „laut“ und „leise“, sodass auch hier schnell durch Nutzung einer Ontologie eine passende Assoziation gefunden ist. Das Suchsystem nutzt also für das Herstellen von Assoziationen die weiter oben beschriebenen Wissensmodelle, -netzwerke und Ontologien sowie Synonyme (Passagierflugzeug und Verkehrsflugzeug), Homonyme (Airbus die Firma und Airbus das Flugzeug) und Antonyme (laut, leise). Es ist offen-

sichtlich, dass dann die besten Suchergebnisse erzielt werden, wenn man als Sucheingabe möglichst ganze Sätze oder sogar größere Texte nimmt, also die geschriebene natürliche Sprache als Eingabe nutzt. Ein häufiger Nutzungsfall ist die Suche nach Dokumenten ähnlichen Inhalts, in dem man ein mehrseitiges Dokument, z. B. eine Patentschrift, in ein Suchsystem eingibt und auf der Datenbasis vieler Millionen bereits be- und anerkannter Patente sucht, ob es hier bereits ein Patent gibt, welches dem als Sucheingabe genutzten entspricht oder Ähnlichkeiten aufweist.

4.4.4.6 Case Based Reasoning (CBR)

Eng verwandt mit der assoziativen Suche ist das fallbasierte Schließen oder Case Based Reasoning (CBR). Wo bei der assoziativen Suche oft mit völlig unstrukturierten Daten gearbeitet wird, wird beim fallbasierten Schließen eine bestimmte Struktur notwendig. Dies erschließt sich aus der Natur der zugrundeliegenden Wissensbasis. Diese wird hier „Fall“ genannt. Ein Fall ist ein Vorgang in der Vergangenheit, der eine bestimmte Problemstellung darstellt und deren Lösung einschließt. Dieser Fall liegt neben sehr vielen anderen auf einem Datenträger gespeichert in einer bestimmten Struktur vor. Der Nutzungsfall des CBR ist, dass man von einem aktuell vorliegenden neuen Fall, den man in dieser speziellen inhaltlichen Ausprägung nicht kennt und daher auch keine Lösung für das hinter diesem Fall liegende Problem hat, auf einen in der Vergangenheit bereits vorgekommenen Fall und dessen Lösung schließt und diese Lösung auf den neuen Fall bzw. das zugrundeliegende Problem anwendet und das Problem damit löst. Ist das Problem erfolgreich gelöst, wird der neue Fall und die (ggf. etwas adaptierte Lösung) in die Fallbasis mit aufgenommen, und die dem CBR-System zugrundeliegende Wissensbasis wächst um einen weiteren Fall. Auch dieses Vorgehen ähnelt den Heuristiken, die wir Menschen zur Problemlösung einsetzen. Als Beispiel möge folgender Vorgang dienen: Ein junger Mechatroniker steht vor dem Problem, einen Vergaser eines 50 Jahre alten Jaguar E-Type zu reparieren. Er hat davon keine Ahnung, weil in seiner Ausbildung der Vergaser keine Rolle mehr spielte (aktueller Fall mit Problemstellung). In der Werkstatt gibt es nun einen alten Meister, der noch mit Vergasern umgehen kann (die Fallbasis), wenn auch nicht direkt mit dem im Jaguar E-Type verbauten SU-Gleichdruckvergasern. Dieser alte Meister nutzt nun sein Wissen um die Arbeitsweise von Vergasern im Allgemeinen und das über Gleichdruckvergaser im Speziellen (Ähnlichkeit eines bekannten Falles mit dem aktuellen Fall) und löst das Problem des jungen Mechatronikers. Natürlich merkt sich der alte Meister den Lösungsweg und hat damit seine Fallbasis wieder erweitert.

Wie bereits weiter oben erwähnt, ist eine gewisse Struktur notwendig, um die Vergleichbarkeit von Fällen und die Genauigkeit des Vergleiches zu erhöhen. Ein gutes Beispiel sind Berichte von Servicetechnikern, die Haushaltsgeräte zu reparieren haben. Diese Berichte sind meist insoweit strukturiert, dass Maschinenart (Waschmaschine), Maschinentyp (Profigerät), Problembeschreibung und Lösungsansatz klar unterschieden und entsprechend in einer strukturierten Form digital gespeichert vorliegen. Im CBR-System kann man nun den einzelnen Elementen der Struktur eine höhere oder niedrigere Wichtigkeit zumessen. Während des Vergleichs berechnet die Maschine

anhand weiterer vorgegebener Parameter die Ähnlichkeit (Similarity) zwischen den Strukturelementen und gewichtet diese mit der gegebenen Gewichtung. Das Suchergebnis gibt nun den Fall samt Lösung mit der höchsten gewichteten Ähnlichkeit zum aktuellen Fall als am höchsten geranktes Suchergebnis wieder. Das CBR-Verfahren kommt insbesondere dort zum Einsatz, wo sehr große, in einer gewissen Struktur vorliegende Problembeschreibungen samt Lösung vorliegen. Beispiele sind hier medizinische Diagnosedatenbanken, Service-, Diagnose- und Reparaturberichte (z. B. KFZ-Diagnosen) und der Kundeservice (Call Center). Der Nutzen von CBR-Systemen ist klar erkennbar: Bestehendes Wissen um Problemstellungen und deren Lösungen wird systematisiert und steht auch Personen ohne umfangreiches Fachwissen und großem Erfahrungshintergrund zur Verfügung.

4.4.4.7 Mischformen/Kombinationen

Speziell bei Analysen sehr großer Datenmengen ist es von Vorteil, die oben beschriebenen Methoden zu kombinieren, um bestmögliche Ergebnisse in Kontext zum Suchziel zu gewinnen. Die benannten Methoden können parallel oder nacheinander angewendet werden, um optimale Ergebnisse zu erzielen. Ein Beispiel für eine sinnvolle Kombination von Methoden ist eine Suche in einem sehr großen Datenkorpus zu bestimmten Begriffen, bei denen die oben beschriebenen linguistischen und semantischen Verfahren eingesetzt werden. Auf dem dadurch reduzierten Korpus, in dem die gesuchten Begriffe enthalten sind (der reduzierte Korpus besteht nur noch aus den Treffern der ersten Suche), können nun höherwertige Analysemethoden eingesetzt werden, um Bedeutungsähnlichkeiten und Relationen in bestimmten Kontexten zu finden. Würden diese höherwertigen Methoden auf den Gesamtkorpus angewendet, wären die Rechenzeiten zu lang bzw. der IT-Ressourceneinsatz zu groß. Eine Beispiel für eine solche Kombination von Suchmethoden ist die Suche nach stark kontextorientierten Informationen. Wenn ein Versicherungsunternehmen Informationen zu ganz bestimmten Schadensfällen sucht, z. B. Wetterschäden ab einer bestimmten Schadenhöhe, wird die Suchzeit sehr verkürzt, wenn in einem ersten Schritt eine IT-Ressourcenschonende allgemeine Suche auf Basis von Keywords und Basis-Wissensmodellen nach Informationsquellen, die Informationen über Wetterschäden allgemein durchgeführt wird. Der zweite Schritt ist dann die detaillierte Suche genau auf den vorher gefundenen inhaltlich passenden Informationsquellen mit den komplexeren und IT-Ressourcen stark beanspruchenden Suchmethoden (wie komplexen Ontologien, assoziativen und semantischen Methoden). Nutzt man hier Systeme, die alle vorher genannten Suchfunktionalitäten beinhalten, können solche komplexen zwei oder auch dreistufigen Suchvorgänge sehr effizient und in kurzer Zeit durchgeführt werden. Moderne Systeme können solche Suchvorgänge teilweise schon in Echtzeit, z. B. mit strömenden Medien (wie RSS-Newsfeeds oder Audiostreams) durchführen.

4.4.5 Zeitlicher Aspekt

4.4.5.1 Retrospektive Analysen

In der retrospektiven Analyse von sehr großen Datenmengen wird auf Daten zurückgegriffen, die in der Vergangenheit erhoben und in geeigneter Art gespeichert worden sind (Datenbanken, Dokumente in Dateisystemen usw.). Eine Auswahl typischer Anwendungsfälle zeigt Tab. 4.5.

Einer der oben ersichtlichen Vorteile von retrospektiven Analysen ist die einfache Kombination verschiedenster Datenquellen unterschiedlichster Struktur. Da die oben beschriebenen Analysemethoden mit unstrukturierten Daten arbeiten und aus diesen Inhalte, Bedeutungen und Ähnlichkeiten extrahieren, also eine Abstraktion der eigentlichen Information erstellen, können auf dieser Basis verschiedenste Quellen, Datentypen und Formate miteinander verknüpft und eine Gesamtaussage über alle Datenquellen hinweg getroffen werden.

4.4.5.2 Echtzeitanalysen

Eine besondere Herausforderung stellen Echtzeitanalysen bei sehr großen Datenmengen dar. In diesen müssen die Analyseergebnisse sehr zeitnah nach dem Eintreten des zu analysierenden Ereignisses vorliegen, da diese zumeist zum Auslösen von weiteren Aktionen genutzt werden (Barlow 2013, S. 6 ff.). Die Analyse basiert also nicht auf bestehenden, in

Tab. 4.5 Typische retrospektive Big Data-Analysen

Analyse	Datenkorpus
Trendanalysen	Offene Quellen im Internet, in denen sich Personen austauschen
Kaufverhalten in Online-Shops	Logdaten von Onlineshops, Foren und Blogs
Marktforschung	Foren, Blogs, Facebook usw. kombiniert mit qualitativen Umfragen
Auswertungen in der Kriminalistik und der Terrorismusbekämpfung	Alle offenen Internetquellen kombiniert mit internen polizeilichen Informationen
Erstellen von Bewegungsprofilen bei geheimdienstlicher Arbeit	Alle offenen Quellen kombiniert mit internen geheimdienstlichen Informationen wie Grenzübertritte, Beobachtungsprotokolle usw.
Verbrauchsprognosen für die Energieversorger um zukünftig Spitzen abdecken zu können	Verbrauchsdaten aller Verbraucher in verschiedenen Formaten
Nachträgliche Überprüfung von Finanztransaktionen	Börsendaten, Tagesnachrichten, Fachzeitschriften
Analysen von Hackerangriffen in Log-Files von Rechenzentren	Log-Files aller beteiligten Rechner in verschiedenen Formaten
Analysen von Windverhältnissen in der Vergangenheit um optimale Plätze für das Aufstellen von Windkrafträder zu finden	Meteorologische Daten, Wetterberichte, einzelne Beobachtungen

der Vergangenheit erhobenen Daten, sondern auf Daten, die im Moment entstehen und sofort zur Analyse zur Verfügung stehen. Die Echtzeitanalyse soll anhand zweier Beispiele erläutert werden.

Die Triebwerke eines Verkehrsflugzeuges sind mit umfangreicher Sensorik ausgestattet. Diese Sensoren senden laufend ihre Daten über eine Funkstrecke an eine Zentrale, parallel hierzu werden weitere Daten aus dem Flugzeug erfasst, wie Vibrationen, Temperaturen, Flughöhe, Geschwindigkeit usw. und mit weiteren Daten wie z. B. mit den meteorologischen Daten im aktuellen geographischen Kontext kombiniert. Die Daten an sich haben verschiedene Strukturen und einen unterschiedlichen Grad der Strukturierung, die Datenmenge kann einige MB pro Sekunde betragen. Ziel einer Big Data-Echtzeitanalyse ist es nun, diesen kombinierten Datenstrom sehr zeitnah nach der Entstehung auf Auffälligkeiten zu untersuchen. Hierbei werden nicht nur die Triebwerksdaten analysiert, sondern immer auch in die Kombination mit allen anderen Daten, um auch kritische Zustände erkennen zu können, die bei Überwachung nur eines Elements (wie der Triebwerke) nicht in ihrer wirklichen Kritikalität erkannt werden könnten. Ziel ist es, bei kritischen Situationen sofort eine (ggf. automatische) Warnung entsprechend der Kritikalität und sogleich detaillierte Handlungsanweisungen mit allen bekannten Hintergrundinformationen an die Piloten geben zu können. Hierbei kommt unter anderem oft das CBR-Verfahren zu Einsatz.

Weiteres Beispiel ist ein großer Konzern, der überwachen möchte, welche Nachrichten über ihn weltweit in textbasierten Internet-Nachrichtenmeldungen (z. B. über RSS-Feeds) oder in das Internet gestreamten Nachrichtensendungen gemeldet werden, um bei nicht adäquaten Nachrichten schnellstmöglich reagieren zu können. Hierbei werden die Tonspuren der ausgewählten digitalen Nachrichtenkanäle in Echtzeit transkribiert (also in Text umgewandelt) und dieser Text mittels der oben beschrieben Methoden analysiert. Werden Inhalte detektiert, die eine bestimmte Aussage oder eine bestimmte Bedeutung haben, erkennt das Analysesystem dies und kann eine Meldung generieren, sodass das Unternehmen quasi im Moment der Sendung dieser Nachricht reagieren kann.

Echtzeitsysteme stellen sehr hohe Anforderungen an die dem Analysesystem zu Grunde liegenden IT-Infrastrukturen und an die Analysesysteme selber, denn diese müssen sowohl auf Hochvolumen als auch auf schnellste Verarbeitung ausgelegt sein, da die Latenz zwischen dem Eingang der Information und der Analyse mit Ergebnisbewertung und dem Anstoßen von Prozessen je nach Anwendungsfall sehr klein sein muss.

4.4.6 Erkenntnisziele der Big Data-Analyse

4.4.6.1 Datengold

Oft liest man in Zeitschriftenartikeln oder in der Fachliteratur über das Thema Big Data den Begriff „Datengold“. Die damit verbundene Assoziation, das wichtige und richtige Informationsnugget aus einer riesigen Menge von Informationssediment zu finden, ist bei der Big Data-Analyse durchaus angebracht. Siehe auch Abschn. 2.3.

Wie man am vorher Ausgeführten gesehen hat, ist es für den gewinnbringenden Einsatz semantischer Technologien sehr wichtig, bereits ein ausgefeiltes und auf die vorliegende Domäne angepasstes Wissensmodell (z. B. Ontologie) einzusetzen. Weiterhin sollte man unter dem gesamten vorliegenden Datenkorpus nur diejenigen Datenquellen benutzen, die zum Erreichen des Erkenntnisziels beitragen. Nutzt man z. B. das gesamte Internet als Datenkorpus, ist es sinnvoll, sich auf die Domänen zu beschränken, die auch wirklich Inhalte in Bezug auf das Erkenntnisziel haben. Diese Vorauswahl spart Ressourcen und erhöht die mögliche Qualität der Ergebnisse. Dafür sind Überlegungen notwendig, was genau das Erkenntnisziel sein soll. Je genauer man sein Erkenntnisziel eingrenzt, desto besser sind die Ergebnisse qualitativ. Der diesbezüglich oft gehörte Einwand, dass man gerade die unbekannten Informationen oder Relationen in einem Datenkorpus aufspüren möchte, ist deswegen unbegründet, da ohne jegliche Vorgabe zumindest einer groben Richtung, die Maschine quasi blind sucht, ohne zu wissen, was jetzt wichtig für das Erkenntnisziel ist und was nicht. Das richtige Vorgehen ist hier ein iteratives, in dem man im ersten Schritt erst eine sehr grobe Vorgabe macht und dann in den nächsten Schritten anhand der durch die vorhergehende Iteration gewonnenen Erkenntnisse das Modell schärft, erweitert und verbessert.

4.4.6.2 Vorhersagen

Eines der typischen Erkenntnisziele der klassischen BI sind Trendaussagen, in denen der Verlauf der Vergangenheitswerte in die Zukunft extrapoliert wird. Diese Trendberechnungen basieren bisher auf im Wesentlichen *strukturiert* vorliegenden Zahlenreihen, die in der Vergangenheit aufgenommen wurden und mittels komplexer mathematischer Methoden in die Zukunft verlängert werden (vgl. Shroff 2013, S. 187–234).

Die semantische Analyse von Daten aus frei zugänglichen Datenquellen im Internet, in denen Personen sich in den sozialen Medien austauschen, ihre Meinung äußern, Erlebnisberichte schreiben und Produkte bewerten, erlaubt nun auch Trendaussagen auf Basis von un- und semistrukturerter textbasierter Daten. Da das Internet ein Datenspeicher ist, der (fast) nie gelöscht wird, können auch Aussagen in der Vergangenheit und die Veränderungen von Meinungen, Ideen und Wünschen im Laufe der Zeit erkannt und analysiert werden. Erstmals können auch Aussagen beliebiger Personen durch die Äußerungen in sozialen Medien über zukünftige Wünsche, Vorhaben und Ideen analysiert werden, *ohne* diese Personen in eine persönliche Befragungssituation (mit einem Interviewer oder mit einem Fragebogen in einer Gruppe) zu bringen (Stichworte: soziale Erwünschtheit, Befangenheit usw.), die die Ergebnisse verfälschen kann. Weiterhin kann der immer fehlerbehaftete Versuch, qualitative Aussagen aus Interviews in eine quantitative Form und damit in eine analysierbare Form zu bringen, vermieden werden.

Durch die semantische Analyse von Äußerungen in den sozialen Medien ist es möglich, auf Basis von Aussagen sehr vieler Personen in vielen Ländern und Sprachbereichen, zukünftige Entwicklungen (Modetrends, Konsumverhalten, Imageveränderungen, Wahlentscheidungen, Änderungen der öffentlichen Meinung usw.) schneller, genauer und zielgruppen- bzw. milieugenauer vorherzusagen, als dies früher der Fall war. Dies begrün-

det sich darin, dass die Personen diese Aussagen ohne jegliche Beeinflussung durch eine Befragungssituation und in einem (von den Personen vermuteten) freien und anonymen Raum ohne Korrektive tätigen – also der „Wahrheitsgehalt“ einer Einzelaussage sehr hoch ist. Durch die oben beschriebenen Analysemethoden können nun Art und Inhalt der Aussage sowie Zeitbezug (wann wurde die Aussage gemacht und welchen Zeitbezug hat sie: Vergangenheit, Gegenwart oder Zukunft?) und Kontext (wo wurde die Information gefunden, welche Wortwahl und wurde getroffen und welche Milieusprache wurde benutzt) erkannt und in eine Trendaussage mit einbezogen werden.

Weiteres Einsatzfeld von Vorhersagen ist der Bereich des technischen Service von Investitionsgütern. In der Investitionsgüterindustrie hat in den letzten Jahren ein Paradigmenwechsel stattgefunden. War es vor Jahren noch üblich, einen Großteil der Marge beim Verkauf komplexer und teurer Investitionsgüter über langlaufende und mit hohen Fixkosten versehenen Serviceverträgen zu generieren, so hat sich die Situation heute grundlegend gewandelt. Investitionsgüter werden heute seltener als Ganzes verkauft, sondern es werden Betreibermodelle gewählt, in denen nur die Leistung aber nicht die Maschine als solche verkauft wird. So erwirbt ein Automobilproduzent nicht mehr die ganze Karosseriepieresse, sondern er erwirbt eine bestimmte Anzahl von Pressvorgängen pro Jahr. Und nur diese bezahlt er auch. War vorher das Interesse des Herstellers, die Services so oft als möglich und für den Hersteller so margenträchtig wie möglich zu gestalten, sehen sich die Hersteller von Investitionsgütern nun vor der Herausforderung, die Kosten für Betrieb und Service einer Maschine so gering wie möglich zu halten. Denn nur durch geringe Kosten während des Betriebs kann die Marge gesteigert werden. Aus diesem Grund werden Methoden benötigt, die es ermöglichen, den Betriebszustand der Maschine auf Basis ausgewählter Sensorwerte laufend zu überwachen und zu analysieren, um etwaige Service- oder Schadenfälle möglichst vor Eintritt kostenintensiver Reparaturfälle erkennen und warten zu können. Weiterhin soll das Servicepersonal nur genau dann vor Ort sein, wenn es absolut nötig ist und schon vorher mit einer genauen Problembeschreibung und der notwendigen Literatur versehen sein, um den Einsatz so effizient und effektiv wie möglich zu gestalten. Dies wird erreicht durch eine Problemerkennung mit dem Echtzeitvergleich von bekannten oder historischen Sensorwertkonstellationen sowie der relevanten Umfeldparameter, bei denen ein Problem auftrat oder eine Servicenotwendigkeit vorlag und den aktuell von der Maschine kommenden Sensorwerten. Im Allgemeinen wird hier das Case Base Reasoning-Verfahren eingesetzt. Wird eine kritische Wertekonstellation erkannt, werden automatisch semantisch orientierte Suchen auf den entsprechenden Datensammlungen abgesetzt, um das Problem näher beschreiben zu können, sowie alle kontextbezogenen Informationen gesammelt, um einen Servicetechniker mit genau den passenden Informationen auszustatten.

4.4.6.3 Schwache Signale

Eine sehr interessante Analysemöglichkeit auf Basis sehr großer Datenmengen ist das Finden sogenannter *schwacher Signale*. Ein schwaches Signal ist ein Begriff aus der Nachrichtentechnik, der Signale beschreibt, die noch im Grundrauschen versteckt sind und die

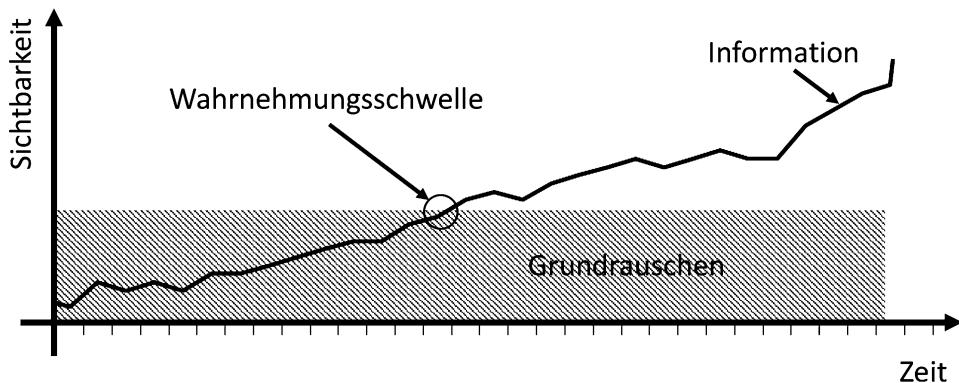


Abb. 4.24 Schwache Signale unterhalb der Wahrnehmungsschwelle

signaltechnische Wahrnehmungsschwelle noch nicht überschritten haben. Übertragen auf die Big Data-Analyse sind dies z. B. Informationen zu bestimmten Themen, die von den klassischen Suchmaschinen nicht wahrnehmbar sind, weil diese Informationen noch nicht in ausreichender Menge im Internet vorhanden sind, noch nicht oder nur sehr gering mit anderen Seiten verknüpft (verlinkt) sind oder in einem bestimmten Kontext noch nicht aufgetreten sind.

Für Firmen kann es nun sehr gewinnbringend sein, solche sehr speziellen Informationen vor allen anderen zu erhalten, um einen Wettbewerbsvorteil daraus zu ziehen. Beispielhaft könnte man annehmen, dass eine Gruppe Studenten herausgefunden hat, dass ein Material mit bestimmten Eigenschaften für sehr spezielle Anwendungsfälle geeignet ist. Diese Information ist aber nur in sehr wenigen offenen Quellen zu finden und in diesem Kontext noch nicht bekannt. Für eine Firma, die einen Geschäftsbereich hat, in dem diese Anwendungsfälle wichtig sind, kann es einen großen wirtschaftlichen Vorteil haben, diese Information als Erste und bevor sie die Internet-Wahrnehmungsschwelle überschritten hat, zu kennen und zu sichern. Dieses Vorgehen wird branchentypisch „Competitive Intelligence“ genannt.

Oft gibt es auch Situationen, in denen große Konzerne selbst keinen Überblick über das Wissen („Wenn ich wüsste, was ich alles weiß ...“, Fachterminus „unbekanntes Wissen“) haben, das in der eigenen Organisation vorhanden ist. Dieses Wissen ist meistens sehr granular und heterogen (Intranet, Dateien, Rechner von Mitarbeitern usw.) verteilt und anderen Organisationsbereichen nicht bekannt bzw. für diese anderen Organisationsbereiche nicht auffindbar, da es zu „versteckt“ ist, um mit den üblichen Methoden (z. B. Suche über das Intranet) hier Erfolg zu haben. Für die Suche auf internen Quellen können ähnliche Methoden wie beim Competitive Intelligence eingesetzt werden. Dies wird dann „Corporate Intelligence“ genannt.

4.4.6.4 Neue Erkenntnisse (knowing the unknown unknown)

Die Begrifflichkeit der „unknown unknowns“ geht auf eine Sentenz von Donald Rumsfeld, dem ehemaligen Verteidigungsminister unter George Bush zurück. In ganzer Länge sagte er:

There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.¹¹

Er drückte damit das „unbekannte Unwissen“ aus, welches es zu erkennen gilt. Gesamthaft kann man die Wissenszustände wie in Tab. 4.6 beschreiben.

Gerade das unbekannte Unwissen stellt eine Gefahr dar, weil man sich diesen Dingen nicht bewusst ist, im Gegensatz zu dem bekannten Unwissen – nach dem man gezielt forschen kann – ist das Gefährdungspotential des unbekannten Unwissen sehr hoch, da nichts dafür getan wird, diese Wissenslücke zu schließen. Die Methoden der Big Data-Analyse bieten nun aber Möglichkeiten auf Basis aller verfügbaren Daten nach solchem unbekannten Unwissen zu suchen. Spezielle Methodensets bieten nicht nur die Möglichkeit, auf Basis vorher konzipierter Wissensmodelle, Ontologien, Verbindungen, Relationen und Zusammenhänge zu finden sondern sie weisen auch auf im Datenkorpus gefundene signifikante in den Gesamtkontext der Analyse passende Relationen und Informationen hin, die man aufgreifen und weiter analysieren kann. Ein Beispiel ist eine Firma, die Batterien für Elektrofahrzeuge herstellt und nach Konkurrenten im Markt und die verwendeten Technologien sucht (z. B. unter Nutzung von Competitive Intelligence). Neben den Treffern in Bezug auf die modellierten Konzepte (Wissensmodelle und Ontologien) schlägt das Analysesystem nun weitere im Datenkorpus gefundene Phrasen vor, die in irgendeiner dem Analysesystem erkenntlichen Beziehung zum Gesuchten stehen. Dies kann in diesem Beispiel der Name einer Person sein. Dem Analysten ist dieser Name völlig unbekannt, insbesondere in Bezug auf seinen Analysezweck. Nach einem Blick auf die Originalquelle aus der das Analysesystem diesen Namen hat, stellt sich heraus, dass diese Person ein chinesischer Investor ist, der eine chinesische Batteriefirma übernommen hat und eine Expansion auf den europäischen Markt plant. Deren Namen wird nicht genannt, steht aber mit der Firma des Analysten bald in starkem Wettbewerb, wenn sie nach Europa expandiert. Nach dem Namen dieser Firma kann nun gezielt gesucht werden. Dies war vorher eine völlig unbekannte Information, die bei weiterem Nichtwissen zu einer großen Gefahr

Tab. 4.6 Wissenszustände

	Wissen	Unwissen
Bekannt	Bekanntes Wissen	Bekanntes Unwissen
Unbekannt	Unbekanntes Wissen	Unbekanntes Unwissen

¹¹ <http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636> (zugegriffen am 18.05.2014).

hätten werden können. Da nun aus dem unbekannten Unwissen ein bekanntes Unwissen geworden ist, kann nun konkret nach Informationen gesucht werden, die das Unwissen in Wissen wandeln.

4.4.6.5 Relationen/Verknüpfung von Daten

Die spezielle Fähigkeit semantischer Methoden in Verbindung mit assoziativen Methoden, die Bedeutung von Texten auf Basis spezieller Wissensmodelle und Ontologien festzustellen, eröffnet im Bereich des Findens offener und versteckter Relationen in sehr umfangreichen Datenmengen neue Möglichkeiten. Zum ersten Mal ist es möglich, auch komplexe Inhalte und Zusammenhänge systemisch prüfen zu lassen, wo vorher ein menschlicher Analyst schon aufgrund der schieren Datenmenge keine Chance gehabt hätte, wirklich alle vorhandenen Relationen, Ähnlichkeiten und Verknüpfungen zu finden. Eine ideale Beispianwendung ist hier die Patentrecherche.

Bei der Patentrecherche muss ein meist sehr umfangreicher im meist komplexer Sprache gehaltenen Patentantrag dahingehend geprüft werden, ob ein gleiches oder ähnliches Verfahren bereits zum Patent angemeldet worden ist – und dies über eine sehr lange Zeitspanne zurück in die Vergangenheit. Dieser oft an die 100 Seiten lange Patentantrag wird mittels der vorher beschrieben Methoden vom System „gelesen“ und in seinen Grundzügen „verstanden“, dann erfolgt eine Suche auf den Korpus aller bereits eingereichter Patente. Dass man es in diesem Falle mit einer sehr großen Datenmenge zu tun hat ist leicht zu erkennen; auch die Komplexität von Text und Inhalt der Patentschriften ist evident – trotzdem muss mit sehr hoher Sicherheit und in ausreichend kurzer Zeit das System gleiche oder ähnliche Patente finden und diese einem Rechercher aufzeigen, der dann die eigentlich Detailprüfung vornimmt.

4.4.7 Zusammenfassung

Die hier beschriebenen Methoden der Big Data-Analyse eröffnen neue Wege, mit den vorhandenen unstrukturierten großen Datenmengen neue Erkenntnisse erzielen zu können. Die Tatsache, dass ca. 80 % aller im unternehmerischen, öffentlichen und staatlichen Sektor vorliegenden Daten unstrukturierte Daten sind, also in Form von weitgehend textbasierten Dokumenten vorliegen, und die Tatsache, dass diese ungeheuren Datenmengen bisher völlig ungenutzt waren, macht es geradezu zwingend, dass man neue Methoden einsetzt, die die in diesen Daten enthaltenen Informationen nutzbar machen. Bisherige Verfahren waren sowohl in technischer Hinsicht (zu viele Ressourcen, zu teuer und zu langsam) als auch in Hinsicht auf die Analysemethoden nicht in der Lage, mit großen und sehr großen Datenmengen umzugehen und diese zielorientiert und mit guter Qualität zu analysieren. Jetzt ist die Situation gegeben, dass sowohl die Technologie als auch die Methoden hierzu geeignet sind; gepaart mit der immer evidenter werdenden Einsicht, dass in diesen Daten wichtige und bisher völlig ungenutzte Informationspotenziale liegen, die einen Erkenntnis- oder Wettbewerbsvorteil bieten können. Die am Anfang

dieser Erkenntnis liegende Goldgräberstimmung, mit der damit verbundenen ungerichteten Analyse der vorhandenen Datenmengen, ohne genau zu wissen, was der eigentliche Analysezweck ist, ist nun – nach einer Ernüchterungsphase – der konkreten Suche nach sinnvollen Nutzungsszenarien gewichen, die den Erkenntniszweck und (zumindest in ökonomisch getriebenen Unternehmungen) einen konkreten ökonomischen Mehrwert bietet. In vielen Bereichen der Wirtschaft und des öffentlichen Lebens sind Big Data-Analysen bereits gewinnbringend im Einsatz: Sei es in der Marktforschung, dem Kaufverhalten von Kunden auf Shopping-Portalen, dem Kundenservice bzw. dem technischen Service oder bei der Echtzeitanalyse von Nachrichtenquellen. Überall laufen Big Data-Analysen bereits sehr erfolgreich und bieten einen evidenten Erkenntnisgewinn.

Literatur

Literatur zu 4.1

- Davenport T (2013) Analytics 3.0. Harvard Business Review S. 64–72
- Eckerson W (2007) Predictive Analytics: Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report, TDWI Research
- Humm B, Wietek F (2005) Architektur von Data Warehouses und Business Intelligence Systemen. Informatik Spektrum 23:3–14
- Kemper HG, Baars H, Mehanna W (2010) Business Intelligence – Grundlagen und praktische Anwendungen. Vieweg+Teubner
- Luhn H (1958) A Business Intelligence System. IBM Journal of Research and Development 2(4):314–319

Literatur zu 4.2

- Anahory S, Murray D (1997) Data Warehousing in the Real World: A practical guide for building decision support systems, 1st edn. Addison-Wesley
- Eckert M, Bry F (2009) Complex Event Processing (CEP). URL <http://www.gi.de/service/informatiklexikon/detailansicht/article/complex-event-processing-cep.html>, zuletzt abgerufen am 31.03.2014
- George L (2011) HBase: The Definitive Guide, 1st edn. O'Reilly
- Kimball R (2011) The Evolving Role of the Enterprise Data Warehouse in the Era of Big-Data Analytics. Whitepaper, Kimball Group, URL <http://www.kimballgroup.com/2011/04/29/the-evolving-role-of-the-enterprise-data-warehouse-in-the-era-of-big-data-analytics/>
- Kimball R, Ross M, Thornthwaite W, Mundy J, Becker B (2008) The Data Warehouse Lifecycle Toolkit, 2nd edn. Wiley
- Mohanty S, Jagadeesh M, Srivatsa H (2013) Big Data Imperatives. Apress
- Pipino L, Lee Y, Wang R (2002) Data Quality Assessment. Communications of the ACM 45(4):211–218
- Russom P (2013) Integrating Hadoop into Business Intelligence and Data Warehousing. TDWI Best Practices Report, TDWI Research
- White T (2012) Hadoop: The Definitive Guide, 3rd edn. O'Reilly

Literatur zu 4.4

- Agrawal, A., Patwary, M., Hendrix, W., Liao, W., Choudhary, A. (2013): High Performance Big Data Clustering), 2013, S. 192–211
- Barlow, M. (2013): Real Time Big Data Analytics: Emerging Architecture, 2013, S. 6 ff.
- Berman, J. (2013): Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information, 2013, S. 1–13
- Cukier, K. (2010): The Data Deluge, in: The Economist: Onlineausgabe vom 25.02.2010. <http://www.economist.com/node/15579717> (zugegriffen am 24.05.2014)
- Jüngling, T. (2013): Die Zeit: Datenvolumen verdoppelt sich alle zwei Jahre. In: Die Welt, Onlineausgabe vom 16.07.2013. <http://www.welt.de/wirtschaft/webwelt/article118099520/Daten-volumen-verdoppelt-sich-alle-zwei-Jahre.html> (zugegriffen am 26.05.2014)
- Moore, G. E. (1965): *Cramming more components onto integrated circuits*. In: *Electronics*, 38, Nr. 8, 1965, S. 114–117
- Sathi, A. (2012): Big Data Analytics: Disruptive Technologies for Changing the Game, 2012, S. 31–46
- Shroff, G. (2013): The intelligent Web: Search, smart Algorithms, and Big Data, 2013, S. 187–234

Sachverzeichnis

A

Abhängigkeitsanalyse, 67
Abmahnung, 248
Abweichungsanalyse, 68
ACID, 288, 289, 293
Advanced Analytics, 55, 62, 269
Aggregation der Daten, 103
Algorithmushandel, 141
Alignment, 31
allgemein zugänglichen Quelle, 195
Ambari, 287
Analyse, 55
 Prozessmodell, 68
Analyseaufgaben, 63
Analyseprozess, 68
Analysespektrum, 56, 257
Analytics, 55, 56, 257, 264
 Advanced, 62, 269
 Big Data, 63, 74
 Descriptive, 56, 66
 Diagnostic, 57
 Embedded, 60
 Exploratory, 58
 In-Database, 61
 In-Memory, 61
 Location, 59
 Predictive, 64
 Predictive, 57, 258
 Prescriptive, 57, 258
 Real-time, 59
 Social Media, 59
 Stream, 61
 Text, 58
 Visual, 58
 Web, 59
Analytics 2.0, 257

Analytik, 55
Änderung, 235
angemessenes Datenschutzniveau, 202, 203
Anonymisierung, 172, 185–190, 275
Anwendungsarchitektur, 36
Apache Hadoop, 279, 286
API, 241
Apple Pay, 137
Application Programming Interface, 241
ApplicationMaster, 283
Application-Service-Providing, 246
Arbeitnehmerdatenschutz, 206
Arbeitnehmerdatenschutzrecht, 208
arbeitsplatzbezogene Daten, 205
Architekturrahmen, 42
Assoziationsanalyse, 67
Assoziative Methoden der Suche, 319
Atomicity, 288
Audiostreams, 310
Auditing, 275
Auftragsdatenverarbeitung, 201, 203, 204, 249
Auftragsdatenverarbeitungsvereinbarung, 250
Aufzeichnung, technische, 235
Auskunft, 193, 194, 196, 197
Auswertung, 217, 219–222
Auswertung der Nutzungsverhaltens, 240
Avro, 286
AWS Big Data, 147

B

Backoffice, 135
Bank, 134
Banken-Stresstest, 138
Bargeld, 146
BASE, 289
Basel III, 138, 145

- Batch, 134
Batch Ebene, 302
Batch View, 302
BCBS 239, 138
BCR, *siehe* verbindliche Unternehmensregelungen
BDSG, 174–176, 178, 179, 181–189, 191, 192
Bedien- und Auswertelogik, 96
Benachrichtigung, 194, 195, 197
best match, 319
Beteiligungsrecht des Betriebsrates, 209
Betriebsvereinbarung, 210
Betriebsverfassungsrecht, 209
Betroffener, 193–199
Bewegungsprofile, 322
Big Data, 234, 307
 Analytics, 269
 Anforderungen, 263
 Funktionale Referenzarchitektur, 263
 Lösungen, 263
Big Data Analytics, 55, 63, 74
Big Data Einführung, 41, 51
Big Data Reifegrad, 22
Big Data Strategie, 30
Big Data System, 29
Big Data-Analyse, 307
BigTable, 292
binding corporate rules, *siehe* verbindliche Unternehmensregelungen
Bolt, 304
Bonitätsrating, 143, 212
Bot, 239
Bundesdatenschutzgesetz, 174, 175, 206
Bundesverfassungsgericht, 233
Business Analytics, 56
Business Capability Map, 40
Business Intelligence, 126, 135, 255, 307
 Analysespektrum, 56, 257
 Grenzen, 260
 Referenzarchitektur, 258
Business-Architektur, 18
- C
Callback-Funktion, 304
CAP, 289
Capabilities, 40
Captcha, 239
Case Based Reasoning, 320
Cassandra, 286, 287, 290, 292
- CEP, 298, 299, 301
Ceph, 284
Chukwa, 286, 287
Click Streams, 300
Client-Server-Architektur, 160
Cloud, 248
Cloud Computing, 146, 203, 204
Cloud-Service, 247
Cluster, 279, 280, 282–287, 303, 305
Clusteranalyse, 67
Columnar Store, 293
Competitive Intelligence, 326, 327
Complex Event Processing, 10, 61, 270, 298, 301
Computersabotage, 228, 235
Core-Banking-System, 135
Crawling, 311
CRISP-DM, 68
Crowd Funding, 144
Customer Journey, 113, 128, 130, 131
- D
Data at Rest, 261, 278, 297
Data Governance, 23
Data in Motion, 261, 278, 298
Data Lake, 267
Data Mining, 286
Data Science, 87
Data Scientist, 306
Data-Hub, 78
Data-Lifecycle-Management, 270
Data-Mining-Prozess, 68
Data-Warehouse, 136, 159, 260
Data-Warehouse-Architektur, 275
Data-Warehouse-System, 256
Daten, 232
Daten, Abfangen, 235
Daten, Abfangen von, 229
Daten, Ausspähen, 235
Daten, Ausspähen von, 228
Daten, Fälschung, 235
Daten, Löschung, 234
Daten, personenbezogene, 233
Datenanalyse, 55, 247
Datenarchitektur, 37
Datenbank, 213, 216–223, 234, 243
Datenbankhersteller, 219, 221
Datenbankschutz, 234
Datenbankwerk, 217, 218, 220

- datenbasierten Capabilities, 33
Datendesintegration, 146
Datenerhebung, 175, 176, 179, 181–184, 191, 193, 237, 247
Datenhehlerei, 231
Datenintegration, 80, 260, 271
Daten-Kidnapping, 227
Datenqualität, 76, 78, 274
Datenschutz, 274
Datenschutz-Grundverordnung, 201
Datenschutzrecht, 206
datenschutzrechtliche
 Abwägungsentscheidung, 169
datenschutzrechtliche Einwilligung, 168
Datensicherheit, 158, 274
Datensparsamkeit, 171
Datenspeicherung, 248
Datenträger, 232, 233
Datenveränderung, 226, 235
Datenverarbeitung, 176–178, 184, 187, 189, 190, 193
Datenverarbeitung in der EU, 200, 201
Datenverarbeitung in Drittstaaten, 200, 201
Datenverarbeitung zu statistischen Zwecken, 173
DB2, 135
Depots, 135
Descriptive Analytics, 56, 66
destillierte Essenz, 294, 296, 300, 301, 303
Diagnostic Analytics, 57
Dienst, 232
Dienstvertrag, 247, 248
Dodd–Frank Act, 138
Dokument-orientiert, 291
DTA, 140
Dual Use, 231
Durability, 289
dynamisches Ablaufverhalten, 99
- E**
Echtzeitanalyse, 322
Echtzeitdaten, 155
Echtzeit-Monitoring, 140
Echtzeitplanungs-System, 153
Echtzeitverarbeitung, 79
Echtzeit-View, 302
E-Commerce, 123
Eigentum, 232
Eigentumsschutz, 232
- Einwilligung, 175, 176, 178, 183, 193, 201, 207
Einzelentscheidungen, automatisierte, 211
Embedded Analytics, 60
EMIR, 138
Ensemble-Methode, 73, 81
Enterprise Architecture, 33
Enterprise Architecture Management, 28, 33
Enterprise Architecture Management
 Einführung, 41
Entity-Extraction, 316
Entscheidungs- und Umsetzungskultur, 17
Entscheidungsmodell, 20
Entscheidungsunterstützung, 256
Entscheidungsunterstützung in der
 Produktionsplanung, 103
ereignis-diskrete Simulation, 97
Erlaubnisnorm, 206
Erlaubnistratbestand, 174, 183, 190
ERP-System, 135
erzwingbares Mitbestimmungsrecht, 210
ESPER, 299
ETL, 8, 260
ETL-Prozess, 312
EU-Standardvertragsklausel, 202–204
Eventual Accuracy, 304
Eventual Consistency, 289
Expertensystem, 156
Exploratory Analytics, 58
- F**
FATCA, 138
Fehlerfreundlichkeit, 25
Fernmeldegeheimnis, 230
FIDUCIA, 134
Finanz Informatik, 134
Finanzdienstleister, 134
Finanztransaktion, 322
Fintechs, 137
Flume, 286
Forum Shopping, 201
Freiraum, 25
Frontoffice, 135
Führung im Big Data, 32
Fusion, 136
- G**
GAD, 134
Ganglia, 287
Garbage Collector, 297

- Gesamtbanksteuerung, 144
Geschäftsarchitektur, 36
Geschäftsfähigkeitslandkarte, 40
gesetzliche Erlaubnis, 201
Gewerbebetrieb, 232, 234
gläserne Gesellschaft, 173
Google, 292
Google Wallet, 137
Graph-Datenbank, 292, 293
GraphX, 287
grenzüberschreitende Datenverarbeitung, 199
Großrechner, 134
Gütekriterium, 118
- H**
Hacking-Paragraf, 228
Hacking-Tools, 230
Hadoop Distributed File System, 279, 280, 286
Hadoop-Distribution, 279, 288
Handel, 141
HBase, 286, 287, 292, 302, 303
HDFS, 279–281, 284–288, 297, 299, 300, 302, 303
Hive, 286, 303
Hochfrequenzhandel, 138, 141
Homon, 315
Hosting, 232
- I**
IaaS, 306
IDV, 135
IMDB, 295–297, 300, 301
IMDG, 296, 297, 299–301
Impala, 302, 303
IMS-DB, 135
In Memory Computing, 312
Incidentmanagement, 27
In-Database Analytics, 61
Index, 312
individuelle Datenverwaltung, 136
Industrialisierung, 135
Industrie 4.0, 242, 257
Information Governance, 27
informationelle Selbstbestimmung, 168, 193
Informationsarchitektur, 36
In-Memory Analytics, 61, 295
In-Memory Data Grid, 296
In-Memory Query-Engine, 303
In-Memory-Datenbank, 296
- In-Memory-Technik, 159
In-Memory-Technologie, 10, 310
integratives Planungssystem, 101
integrierte Datenhaltung, 155
Integritätsschutz, 225
intelligentes Produktionssystem, 153
interaktive Austaktung einer Produktionslinie, 101
Interessenabwägung, 174, 180–182, 185, 192
Internet der Dinge, 7, 257
Internet of the things (IOT), 16
Internet of Things, 299
Investition, 218, 219, 221, 222
IP-Sperre, 238
- J**
JSON-Format, 291
JVM, 297
- K**
Kaufvertrag, 246
Kernbankensystem, 135
Key-Value Store, 290
Key-Value-Paar, 282
keywordbasierte Suche, 315
Klassifikation, 64
kollaborative Entscheidungsmodell-Entwicklung, 26
Kommunikationskompetenz, 26
Kompetenz, 23
Komplexität der Lieferkette, 152
Komplexität der Wirkzusammenhänge, 156
Konto, 135
Konzeptbeschreibung, 66
Korrektur, 194, 197, 198
Kreditgeschäft, 142
Kreditkartenabrechnung, 141
Kreditprodukte, strukturierte, 142
Kreditwirtschaft, 134
kundenindividuelle Produktgestaltung, 154
- L**
Lambda-Architektur, 301–303, 305
Landesbank, 134
Landing Space, 267
Lastschrift, 139
Legacy-System, 135, 136
Leihvertrag, 246
Leistungsschutzrechte, 217, 224
Leistungssteuerung in Echtzeit, 150

- Leistungsstörung, 249
Linguistik, 316
Lizenz, 246
Location Analytics, 59
logistischen Simulation, 103
Lösung, 193, 194, 197, 198
- M**
Machine Learning, 287
Mahout, 286
Mainframe, 134
Malware, 228
Managed Services, 147
Management des Daten-Lebenszyklus, 28
Map-Phase, 281
MapReduce, 83, 279–288, 301–304, 313
Marktdatenversorgung, 142
Marktforschung, 112, 322
Meldewesen, 135
Menschenwürde, 211
Metadaten, 273, 307
Miete, 246
MiFID II, 138
MiFIR, 138
Mitbestimmungsrecht des Betriebsrates, 210
MLib, 287
Moore'sches Gesetz, 7
morphosyntaktische Analyse, 315
Multichannel, 134, 145
Multikanal, 134
Multilingualität, 311
- N**
Nachahmung, 220, 222
Nagios, 287
NameNode, 280, 284
natürliche Sprache, 8
Near Real-Time, 301
nichtvalidierte Daten, 22
Niederlassung, 200
Nimbus, 305
Non-Disclosure Agreement, 247
NoSQL, 10, 285, 286, 288–291, 293, 294, 297, 299, 301, 302
NSA-Skandal, 9
Numerische Vorhersage, 65
nutzergenerierter Inhalt, 241
- O**
öffentliche Wiedergabe, 219–221
OLAP, 10, 256
OLTP, 10
Online-Analytical Processing, 308
Online-Handel, 124
Online-Leistungssteuerung, 102
Online-Produktkonfiguration, 149
Ontologie, 317
Oozie, 287
Open Government Data, 170
Opt-Out, 241
Ordnungswidrigkeit, 195, 196
Outsourcing, 136, 146, 227, 250
Overfitting, 80
- P**
PaaS, 306
PACT, 284
Partition Tolerance, 289
Part-of-Speech Tagging, 316
PayPal, 137
Personaleinsatzplanung, 100
Personalflexibilität, 93
Personalfragebogen, 209
Personalplanung, 209
Personalvertretungsrecht, 209
personenbezogene Daten der Mitarbeiter, 206
personenbezogene Mitarbeiterdaten, 208
Persönlichkeitsbild, 169
Persönlichkeitsmerkmal, 211
Phasen der EAM-Einführung, 45
Pig, 286
Planungsgenauigkeit, 93
Planungskaskade in der Automobilindustrie, 91
Plausibilitätsprüfungen von Eingangs-,
Simulations- und Ergebnisdaten, 98
Portfoliosimulation, 142
Praxisbeispiel, 19
Predictive Modeling, 64
Predictive Analytics, 57, 64, 258
Prescriptive Analytics, 57, 258
proaktive Planung und Steuerung in Echtzeit,
160
Produktempfehlung, 128
Produktion in Echtzeit, 159
Programm- bzw. Sequenzplanung, 90
Prozess-Off-Loading, 277
Pseudonymisierung, 185, 187–189

Q

qualifizierte Beobachtung, 22

R

Raiffeisenbank, 134

Ranking, 65

Realtime Analytics, 59

Realtime-Informationsmanagement-Konzepte, 149

Realtime-Monitoring, 140

Recht auf vergessen werden, 198, 199

Rechtsgeschäfte, 245

Rechtsgüter, geschützte, 231

Rechtskauf, 246

Rechtsmangel, 249

Reduce, 281–283, 287

Referenzarchitektur, 300, 301

Regression, 65

regulatorische Anforderung, 136

Reihenfolgeplanung in Echtzeit, 152

relevantes Zielsystem, 31

ResourceManager, 283

Resourcenmanagement, 269

Risikocontrolling, 138

Risikomanagement, 135, 145

Roboter, 239

Rohdaten-Schicht, 267

S

Sachbeschädigung, 226

Sachkauf, 246

Safe Harbor Principles, 203

Sandboxing, 267

SAP Banking Services, 135

Scale-Out, 278

Scale-up, 278

Schema-on-Read, 79, 268

Schema-on-Write, 260, 268

Schufa-Auskunft, 143

Schutz des Presseverlegers, 224

Schutzgesetz, 235

schutzwürdiges Interesse, 202

schwaches Signal, 325

Scoring, 143, 212

Screen-Scraping, 196, 220, 221, 237

Search, 269

Semantik, 316

semantische Technologie, 10

semistrukturierte Datenmenge, 307

SEPA, 139

sequenzierte Produktionslinie, 92, 99

Serving-Ebene, 302, 303

Shared Nothing, 279

Shark, 287

Shuffle-Phase, 281, 282

simulationsbasierte Planung und Steuerung in Echtzeit, 95

simulationsgestützte Planung, 90

Simulationsstudie, 95

Sitzlandprinzip, 200, 204

Skalierbarkeit, 313

Slave-Knoten, 283

Smart Data, 234

Social Media, 112, 119, 127, 129

Social Media Analyse, 241

Social Media Analytics, 59

Social Media Monitoring, 241

Software as a Service, 247

Solid State Disk, 294, 295, 301

soziales Netzwerk, 195

Spaltenfamilie, 292

Spaltenorientierte Datenbank, 291, 292

Spark, 287

Spark Streaming, 287

Sparkasse, 134

Speed-Ebene, 302, 304

spekulative Ausführung, 282

Sperrung, 193, 194, 198

Spout, 304

SQL, 286, 287, 289, 300, 301

Sqoop, 286

Stamm- und Planungsdaten, 102

Stammformen von Verben, 314

Standardapplikation, 135

Standardsoftware, 136

Standardvertragsklausel, 250

Storm, 302, 304, 305

Strafrecht, 225

Stratosphere, 284

Stream Analytics, 61

Stream-Verarbeitung, 271

Subgroup Discovery, 66

Subgruppenerkennung, 66

Suche, 314

Suchmaschine, 198, 199

Supervised Learning, 64

Synonym, 315

Systemdatenschutz, 172

- System-Management, 272
Szenarien-Technik, 157
szenario-basierte Steuerung, 160
- T**
TaskTracker, 281
Taxonomie, 317
technische Schutzmaßnahme, 238
technische Schutzvorkehrung, 223
technische und organisatorische Maßnahmen, 250
Technologiearchitektur, 37
Technologie-Roadmap, 154
Telekommunikationsgesetz, 174, 175, 177, 230
Telemediengesetz, 174, 175
Territorialprinzip, 200
Text Analytics, 58
Text Mining, 58, 119, 308
Tez, 284
Time-to-Market, 135
Tipps für Entscheider, 30
TOGAF, 31
TOGAF Architecture Development Method, 42
Transaktionsdaten, 141
transparentes Zielsystem, 17
Transparenz, 154, 170
Trendanalyse, 322
- U**
Überwachtes Lernen, 64
Überweisung, 139
Universalbank, 134
unknown unknowns, 327
Unmittelbarkeit der Datenerhebung, 171
unstrukturierte Daten, 8
unstrukturierte Datenmenge, 307
unstrukturierte und semistrukturierte Daten, 309
Unsupervised Learning, 66
Unternehmensarchitektur, 33
Unternehmenskultur, 23
Unternehmensmodellierung, 50
unternehmerische Entscheidung, 15
Unüberwachtes Lernen, 66
Urheberrecht, 232
User Generated Content, 241
- V**
variantenreiche Produkte, 90
- variantenreiche Serienfertigung, 98
Variety, 8, 262, 264
Velocity, 7, 261, 264
Veracity, 8, 262, 264
verbindliche Unternehmensregelung, 202
Verbot mit Erlaubnisvorbehalt, 168
Verbrauchsprognose, 322
Verbreitung, 216, 217, 219, 221
Verbundorganisation, 134
Verknüpfung von Daten, 328
Verschlüsselung, 185, 186, 188, 189
verteilte Parallelverarbeitung, 10
Vertrag, 245
Vertraulichkeit und Integrität informationstechnischer Systeme, 233
Vervielfältigung, 216, 217, 219–221
verwandtes Schutzrecht, 217, 224
Videostreams, 310
Virtualisierung, 147
Visual Analytics, 58
Visualisierung, 287, 288, 299–301
Volksbank, 134
Volume, 7, 260, 264
Vorabkontrolle, 171
Vorhersage, 324
V's, 6
- W**
Wahrscheinlichkeitswert, 212
Web Analytics, 59
Webanalyse, 113, 123, 125
Web-Crawling, 195, 196
Werkvertrag, 247, 248
Wert von Informationen, 19
Wissensmodell, 317
Worker Node, 305
- Y**
YARN, 283, 286, 287
Yet Another Resource Negotiator, 283
- Z**
Zahlungsverkehr, 139
Zahlungsverkehrsdaten, 141
Zeitreihenanalyse, 65
Zookeeper, 305
Zweckbindung, 169
Zweikreis-Modell der Produktion, 150