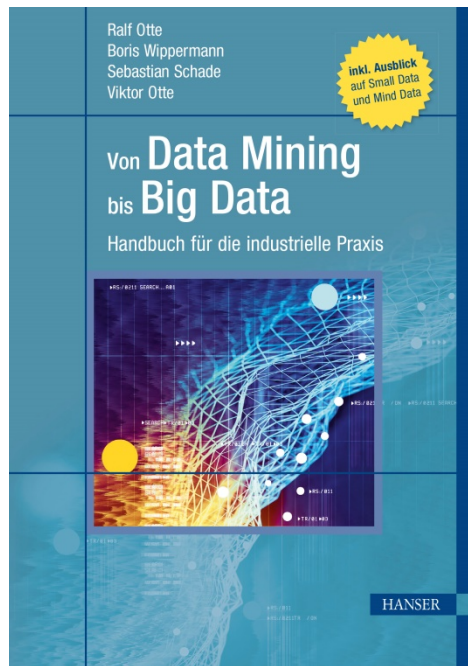


# HANSER



## Leseprobe

zu

## „Von Data Mining bis Big Data“

von Ralf Otte et al.

Print-ISBN: 978-3-446-45550-4  
E-Book-ISBN: 978-3-446-45717-1

Weitere Informationen und Bestellungen unter  
<http://www.hanser-fachbuch.de/978-3-446-45550-4>

sowie im Buchhandel

© Carl Hanser Verlag, München

# Vorwort

Das vorliegende Buch entstand im Ergebnis einer über 25-jährigen Beschäftigung der Autoren mit Data-Mining-Projekten im industriellen Umfeld und einer mehrjährigen Tätigkeit im Big-Data-Business im kommerziellen Bereich.

Als 2004, also vor über 15 Jahren, das Buch „Data Mining für die industrielle Praxis“ im HANSER Verlag von zwei der heutigen Autoren erschien [OTT04], reagierte es auf einen zunehmenden Trend in Wirtschaft, Verwaltung, Forschung und Industrie, die Informationen vorliegender Prozessdaten für die Prozessoptimierung besser zu nutzen. „Graben in Daten“ war damals bereits angesagt. Die Effizienz der vorhandenen Rechnersysteme war gestiegen, neue mathematische Verfahren zur statistischen Datenanalyse waren hinreichend erprobt und der Bedarf der potenziellen Anwender dieser neuen Technologie war groß. Schon bis zum Erscheinen des damaligen Buches, aber insbesondere in der darauffolgenden Zeit, entstanden vielfältige Data-Mining-Projekte, die nahezu alle Facetten dieser relativ neuartigen Technologie tangierten. Es musste auch viel gelernt werden. Bei der Projektrealisierung wurden viele handwerkliche Fehler gemacht, insbesondere die Phase der Datenbereitstellung umfangs- und zeitmäßig unzureichend geplant, die Kosten unterschätzt und die Projekteinführung in den industriellen Prozess zu optimistisch prognostiziert. Jeder, der ein derartiges Projekt real begleitet oder sogar selbst verantwortlich geführt hat, weiß davon ein Lied zu singen. Letztlich hat das dem Siegeszug von Data Mining oder Data Science aber nicht geschadet.

Die Aufgaben zur Optimierung vorhandener Prozesse sind heute noch genauso aktuell wie vor 15 Jahren und deshalb haben sich die Autoren des damaligen Buches in Abstimmung mit dem Verlag entschlossen, trotz neuer Entwicklungen und Trends in der Datenverarbeitung wie Big Data oder Small Data dem klassischen Data Mining in diesem Buch erneut hinreichend Platz einzuräumen. Will man Data Mining anwenden, so muss man die dazugehörigen Verfahren bzw. Methoden beherrschen oder zumindest theoretisch verstehen. Obwohl auch dieses Buch kein Lehrbuch sein will, so erschien es uns zweckmäßig, die im Buch von 2004 beschriebenen mathematischen Grundlagen für das vorliegende Buch komplett zu übernehmen, ergänzt durch einige zweckmäßig ausgewählte Beispiele. Der Leser

unseres Vorgängerbuches wird also in den Kapiteln 2 bis 3 nichts wesentlich Neues finden. Es lohnt sich aber auf jeden Fall, sich mit den verschiedenen, problemangepassten Möglichkeiten der Datenselektion und -zusammenführung, der Datenvorverarbeitung, der Datentransformation und der statistischen Datenanalyse wie z.B. multivariaten Verfahren, Künstlichen Neuronalen Netzen (KNN), Support-Vektor-Maschinen, Selbstorganisierenden Merkmalskarten (SOM) oder Clusterverfahren und den daraus ziehbaren Schlussfolgerungen, wie sie z. B. durch verschiedene Regelgenerierungsverfahren oder Visualisierungen hochdimensionaler Datenräume möglich werden, nochmals zu beschäftigen. Zur Klassifizierung von Daten im Sinne von Lernprozessen hat sich die Informatik in der letzten Zeit sehr stark auf Künstliche Neuronale Netze konzentriert, vor allem angeregt durch die neuen Möglichkeiten des Deep Learning mit tiefen Neuronalen Netzen. Dazu wird jedoch Spezialliteratur empfohlen, die Grundlagen der neuronalen Lernprozesse muss man aber verstanden haben. Auch deshalb ist es zweckmäßig, sich nochmals mit der Theorie zu beschäftigen. In Kapitel 4 werden einige Auswertungsmöglichkeiten für praktische Anwendungsfälle dargestellt, die heute wie vor 15 Jahren noch hochaktuell sind.

Viele der hier beschriebenen mathematischen Verfahren findet die Leserin bzw. der Leser auch im KI-Buch „Künstliche Intelligenz für Dummies“ aus dem Jahre 2019 von einem unserer Autoren [OTT19]. Da die Künstliche Intelligenz heute zu einem Großteil auf maschinelles Lernen setzt, sind viele Verfahren und auch Anwendungsbeispiele in beiden Büchern nahezu deckungsgleich. Manche Algorithmen, wie die zu Unrecht fast nirgends erwähnten Selbstorganisierenden Merkmalskarten (SOM), sind in dem hier vorliegenden Buch allerdings detaillierter erläutert, andere, eher übliche KI-Verfahren, wie Entscheidungsbäume, Backpropagation-Netze oder Deep-Learning-Faltungsnetze (CNN) sind in dem KI-Buch näher beschrieben. Der Schwerpunkt des hier vorliegenden Buches liegt eindeutig in der Datenanalyse und nicht so sehr in der Beschreibung der Künstlichen Intelligenz selbst.

In den letzten Jahren hat sich ein gravierender Wandel in den Datenverarbeitungstechnologien ergeben. Während klassische Data-Mining-Projekte immer mit einem gewissen Zeitpolster bearbeitet werden konnten und weiterhin auch können, gibt es zunehmend Aufgaben in allen Bereichen moderner Gesellschaften, bei denen riesige, temporär anfallende Daten (Volume) unterschiedlichster Formate (Variety) in extrem kurzer Zeit (Velocity) verarbeitet werden müssen. Die drei „Vs“ sind zum Charakteristikum einer neuen Form der Datenverarbeitung geworden, die als Big Data bezeichnet wird. Alle Datenverarbeitung bleibt aber nur sinnvoll, wenn den ermittelten Informationen ein inhaltlicher Wert (Value), ein Nutzen, zugeordnet werden kann. Betrachtet man die täglich anfallenden riesigen Datenmengen in sogenannten Sozialen Netzen oder, um ein diametrales Beispiel zu nennen, im Gesundheitswesen, so wird schnell klar, dass die abgeleiteten Informationen bzgl.

ihre Korrektheit oder Zuverlässigkeit sehr unterschiedlich bewertet werden müssen. Hinweise zum Kaufverhalten des Nutzers eines Online-Shops haben durchaus einen anderen Qualitätsanspruch als eine medizinische Diagnose, die aus Vergleichsdaten ähnlicher Fälle computergestützt erstellt wird. Ansprüche an Prädiktion und Wertvorstellungen sind also kontextabhängig. Wir wollen sie hier nicht betrachten, sondern uns stattdessen auf die technologischen Herausforderungen der drei Vs konzentrieren.

Kapitel 5 beschreibt zunächst Architekturkonzepte zum Umgang mit vielen Daten, geht dann auf die Verarbeitungsmöglichkeiten in Form eines deskriptiven oder präskriptiven Datenmodells ein, benennt Möglichkeiten der Hardware- und Datenvirtualisierung und den weiteren Trend zur Entkoppelung von Speicherung und Verarbeitung von Daten. Nach diesen grundsätzlichen Bemerkungen werden in Kapitel 6 die Komponenten der Big-Data-Pionierlösung Hadoop und ihre Weiterentwicklungen dargestellt, auf Apache Spark, eine weitere Evolutionsstufe, eingegangen (die sich gegenwärtig als Standard herausbildet) und anschließend die vielfältigen Modifikationen und Neuentwicklungen beschrieben, die verschiedenen Nutzern den Zugang zu Big-Data-Diensten erleichtern sollen. Dieses Kapitel informiert auch über solche Schlüsseltechnologien wie Apache Kafka oder Kappa, die insbesondere für eine Streamingverarbeitung genutzt werden. Von besonderem Interesse könnte für Leser auch der zunehmende Einsatz von NoSQL-Datenbanken sein, die das übliche relationale Speicherkonzept auch zugunsten von massiv verteilten WEB-Anwendungen aufgegeben haben. Es werden verschiedene Einsatzfälle beschrieben. Hier sind hunderte Open-Source-Anbieter auf dem Markt. Abschließend informiert das Kapitel über die Möglichkeit, für wiederkehrende Aufgaben klar definierte Lösungsmuster, sogenannte Stacks, einzusetzen. Damit sind die Voraussetzungen genannt, um die in den folgenden Abschnitten spezialisierten Ausführungen zu Datenarchitekturen, zum Aufbau von Data Lakes und zum Cloud Computing einordnen zu können. Die Frage, wie Big Data und Künstliche Intelligenz (KI) zusammenwirken, wird abschließend diskutiert. Big Data wird zu einer Quelle für KI und für den Data Scientist, der letztlich für den Mehrwert der Datenanalyse zuständig ist.

Kapitel 7 und 8 schildern Anwendungen. Die gesamten Ausführungen dieser Kapitel sind sehr kompakt gehalten und man bekommt viele Detailinformationen, die auch als Kompass zum Navigieren in diesem äußerst vielseitigen und sich gegenwärtig noch ausdehnenden Terrain dienen. Die Autoren wissen natürlich, dass die dargestellten Informationen nur eine Übersicht sein können und kein Lehrbuch zum Studium einzelner Facetten dieses riesigen Gebietes der Datenverarbeitung. Die angegebenen Quellen, im Schwerpunkt Internetadressen, helfen aber weiter, wenn man die Absicht hat, sich tiefer in die Problematik von Big Data einzuarbeiten. In vielen, in den vergangenen Jahren realisierten Beispielen hat sich gezeigt, dass die ausreichende Vorbereitung und Planung von Data Mining oder Big-Data-

Projekten oft sträflich unterschätzt wird, oftmals sogar bereits die Aufgabenstellung nicht hinreichend an den unternehmerischen Belangen, sondern zu sehr aus einer IT-Perspektive heraus ausgerichtet ist. Hierbei sind also nicht nur die Informatiker und Kenner der inhaltlichen Probleme gefragt, sondern vor allem die Geschäfts- und Prozessverantwortlichen sowie Entscheider. Hier wird auch das Urteil darüber gefällt, ob das Data-Projekt erfolgreich durchgeführt werden wird, ob es beginnt, wie es realisiert oder ob es bereits in einer frühen Phase gestoppt wird. Wir haben uns deshalb entschlossen, diese Problematik sehr ausführlich darzustellen. Die dafür zuständigen Autoren kennen die zu behandelnden Aufgaben genau, da sie sich in ihrem beruflichen Alltag mit allen diesen Fragen der Projektplanung, Projektrealisierung und Projekteinführung täglich beschäftigen müssen. Die gegebenen methodischen Hinweise sind deshalb äußerst praxisnah und oft auch wie ein Kochrezept abzuarbeiten. Dem Leser, der ein Data-Projekt von der Planung bis zur Einführung bearbeiten muss, sind die in diesen Kapiteln gegebenen Hinweise deshalb nachdrücklich zu empfehlen, insbesondere auch deshalb, weil hier verschiedene betriebliche Anwendungsfelder behandelt werden. Einige der Beispiele zu Anwendungsfällen von Data Mining sind wieder dem älteren Buch „Data Mining für die industrielle Praxis“ entnommen, da sie ihre didaktische Funktion, die in Kapitel 2 und 3 dargestellte Theorie am praktischen Beispiel zu erläutern, noch immer gut erfüllen können. Weitere, kleine Ausführungsbeispiele findet man in dem bereits genannten Fachbuch „Künstliche Intelligenz für Dummies“ von einem unserer Autoren.

Verfolgt man die Entwicklung zu Datenverarbeitungstechnologien, so taucht seit einiger Zeit ein neuer Begriff auf, das sogenannte Small Data. Hier wird das Paradigma geändert: Nicht noch immer mehr und immer schneller, sondern zurück zu den Anfängen. Small Data erklärt sich aus der Absicht, nur wenige Daten zu nutzen oder auch nur zur Verfügung zu haben, aber auch aus ihnen wesentliche Informationen zur Lösung eines Problems ableiten zu müssen. In der Philosophie des Ansatzes versteckt sich auch die Sorge, dass extrem viele Daten möglicherweise den Problemerkern verschleiern könnten, man also besser erst einmal versucht, die Dinge zu vereinfachen und mit dem zu arbeiten, was ohnehin zur Verfügung steht. Kapitel 9 widmet sich diesem Thema, wobei schnell klar wird, dass es sich eigentlich nicht um eine neue Technologie, sondern tatsächlich nur um einen Ansatz handelt, der etwas in Vergessenheit geraten ist. Neue Erkenntnisse über die Welt gewinnt man durch Induktion und Deduktion, also durch Extraktion versteckter Inhalte aus riesigen Datenbergen mittels statistischer Verfahren (induktives Lernen), aber auch durch logische (oder kreative) Schlussfolgerung aus oft nur wenigen (oder gar keinen) Daten (deduktives Schließen). Der Mensch kann beides, die modernen Datenverarbeitungstechnologien haben den Akzent seit Data Mining und Big Data aber mehr auf die Statistikvariante gelegt. Es wird also Zeit, sich wieder der Deduktion zu erinnern und zu hinterfragen, wie eigentlich der Mensch mit

seinem Gehirn Small Data und letztlich Deduktion praktiziert. Wie also wird die Zukunft eines technisch optimierten Small Data aussehen?

Hier nun haben die Verfasser dieses Kapitels den bisher beschrittenen Weg der Analyse und Zusammenfassung vorliegender, bewährter Fachinhalte verlassen und – mit dem Risiko „zweifelnder Akzeptanz“ beim Leser – einen vollkommen neuen, hypothetischen Ansatz gewählt. Sie bezeichnen die im menschlichen Gehirn ablaufenden Datenverarbeitungsprozesse als Mind Data und führen aus, dass sich Mind Data nur unter Nutzung von Bewusstsein realisieren lässt. Damit entsteht die grundsätzliche Frage, wie maschinelles, technisches Bewusstsein erzeugt werden kann. Es wird versucht, Antworten darauf aus einem Konzept abzuleiten, welches Bewusstsein als imaginären physikalischen Prozess beschreibt. Dieses Problem tangiert die aktuellen Forschungsarbeiten zur Künstlichen Intelligenz und es wird sofort klar, dass hier auch die neuesten Ansätze der KI hineinspielen, Künstliche Neuronale Netze mit neuromorpher Struktur oder Quantentechnologien aufzubauen.

Damit schließt sich der Kreis dieses Buches, angefangen mit Erläuterungen zu statistischen Verfahren zur Informationsgewinnung im Rahmen von Data Mining, über den Einsatz modernster Datenverarbeitungstechnologien zur Verarbeitung extrem großer Datenmengen in extrem kurzer Zeit durch Big Data bis zur Wiedergeburt der Erkenntnis des Small Data, dass auch in geringen Datenbeständen viel Information enthalten sein kann, die sich mit stark reduzierten Data-Technologien und auch alten, konventionellen Methoden gewinnen lässt. Daraus folgt zwangsläufig der Ansatz eines Mind Data, also einer Datenverarbeitung, die von den zukünftigen Entwicklungen der KI zum Aufbau eines rudimentären, technischen Bewusstseins geprägt sein wird.

Die Autoren hoffen, Ihnen, liebe Leser, damit einen verständlichen und interessanten Überblick über Technologien gegeben zu haben, die im „Zeitalter der Digitalisierung“ schon heute und künftig noch nachhaltiger unser Leben bestimmen werden.

Das vorliegende Buch ist das gemeinschaftliche Werk von vier Autoren. Selbstverständlich gibt es bei den Fachkapiteln dedizierte Verantwortlichkeiten. Insbesondere verantwortlich für die Fachkapitel 2, 3, 4 und 9 und für zahlreiche Industriebeispiele in diesem Buch, unter anderem in den Kapiteln 7 und 8, sind Ralf Otte und Viktor Otte. Das Kapitel 4.1 basiert auf Ausarbeitungen von Christian Rohdanz aus Konstanz, wofür wir uns an dieser Stelle nochmals ganz herzlich bedanken wollen. Unser Autor Sebastian Schade ist als Big-Data-Experte bei den Kapiteln 5 und 6 und bei zahlreichen anderen Praxisbeispielen in verschiedenen Kapiteln federführend gewesen, und Boris Wippermann, als Experte in der Verknüpfung von Digitalisierung und Unternehmensoptimierung, zeichnet hauptverantwortlich für die Anwendungskapitel 7 und 8.

In der Drucklegung des Buches werden leider keine Farben wiedergegeben. Die Autoren weisen extra darauf hin, da bei verschiedenen Bildern die dargestellten Grauwerte nicht immer als unterschiedlich identifiziert werden können.

Wenn Sie Fragen, Anregungen oder auch Verbesserungsvorschläge an einen unserer Autoren haben, freuen wir uns über eine Mail an den Hauptverantwortlichen des Buches unter *ralf.otte@email.de*.

Am Ende des Vorwortes möchten wir allen danken, die zur Erstellung des Buches beigetragen haben, aber ganz besonders auch unseren Familienangehörigen, die durch Verzicht, aber auch Aufmunterung am Gelingen des Projektes beteiligt waren.

Die Autoren

Weinheim, München, Frankfurt/Main und Magdeburg im Frühjahr 2020

# Inhalt

<b>Vorwort</b> .....	<b>V</b>
<b>1 Einführung</b> .....	<b>1</b>
<b>2 Warum Data Mining? Wozu Big Data?</b> .....	<b>3</b>
2.1 Definition und Einordnung der Begriffe .....	6
2.1.1 Was ist Data Mining? .....	6
2.1.2 Was ist Big Data? .....	14
2.1.3 Data Mining im Kontext anderer Datenanalyseverfahren	15
2.2 Spezielle Anforderungen der Industrie an die Datenanalyse . . . .	22
2.3 Gibt es einen Handlungsbedarf für die Industrie? .....	29
<b>3 Das theoretische und mathematische Konzept der technischen Datenauswertung</b> .....	<b>33</b>
3.1 Einführung .....	33
3.2 Datenselektion und Datenzusammenführung .....	35
3.2.1 Aufbau einer Datentabelle .....	35
3.2.2 Denormalisierung von Datentabellen .....	36
3.2.3 Synchronisierung von Datentabellen .....	37
3.3 Datenvorverarbeitung .....	39
3.3.1 Festlegung der Datentypen .....	39
3.3.2 Diskretisierung von metrischen Daten .....	41
3.3.3 Statistiken und Tests für metrische Daten .....	43
3.3.4 Das Problem ungenauer Messungen .....	48
3.3.5 Behandlung von Datenlücken .....	51
3.3.6 Behandlung von Ausreißern .....	53
3.3.7 Behandlung von Mehrdeutigkeiten .....	55
3.4 Datentransformation .....	60



3.5	Datenanalyse .....	64
3.5.1	Visuelle explorative Analysen .....	64
3.5.2	Überblick über multivariate Verfahren zur Datenanalyse .....	67
3.5.2.1	Regressionsanalysen .....	67
3.5.2.2	Varianzanalyse .....	73
3.5.2.3	Diskriminanzanalyse .....	76
3.5.2.4	Korrelationsanalyse .....	79
3.5.2.5	Faktoranalyse .....	83
3.5.2.6	Clusteranalyse .....	86
3.5.3	Einführung in Data-Mining-Methoden .....	93
3.5.4	Data Mining zum Auffinden von Zusammenhängen ....	97
3.5.4.1	Neuronale Netze .....	99
3.5.4.2	Support-Vektor-Maschinen .....	114
3.5.4.3	Gütemaße für Modelle und Klassifikatoren ...	119
3.5.5	Data Mining zum Auffinden von Strukturen .....	127
3.5.5.1	Fuzzy-Clusterverfahren .....	128
3.5.5.2	Demographisches Clustern .....	130
3.5.5.3	Selbstorganisierende Merkmalskarten .....	131
3.5.5.4	Gütemaße für Clusterverfahren .....	143
3.5.6	Data Mining zum Generieren von Regeln .....	145
3.5.6.1	Bayessche Netze .....	146
3.5.6.2	Entscheidungsbäume .....	152
3.5.6.3	Assoziationsregeln .....	162
3.5.6.4	Gütemaße für Regeln .....	165
3.5.7	Data Mining zum Visualisieren hochdimensionaler Datenräume .....	166
3.5.7.1	Selbstorganisierende Merkmalskarten für topologieerhaltende Projektionen .....	166
3.5.7.2	Gütemaße für Projektionen .....	173
3.5.8	Zusammenfassung der Data-Mining-Verfahren .....	177
3.6	Interpretation der Ergebnisse .....	180
3.6.1	Fehlinterpretationen .....	181
3.6.2	Strittige Interpretationen .....	187
3.6.3	Konsequenzen .....	189
<b>4</b>	<b>Hilfreiche Auswertemöglichkeiten für praktische Anwendungsfälle .....</b>	<b>191</b>
4.1	Text Mining – das Auswerten unstrukturierter Daten .....	191
4.2	Versuchsplanungen zur Erzeugung von Prozessdaten .....	197
4.3	Automatische Diskretisierungen .....	202

4.4	Güte und Sicherheit von Regressionsschätzungen .....	204
4.5	Auffinden der sensitiven Einflussgrößen .....	208
4.6	Ausschluss von zufälligen Zusammenhängen .....	212
4.7	Datenbasierte Optimierungen .....	216
<b>5</b>	<b>Big Data – die Datenhaltungs- und Verarbeitungskonzepte der Gegenwart .....</b>	<b>229</b>
5.1	Digitale Transformation und Big Data .....	230
5.2	Grundprinzipien eines Paradigmenwandels .....	232
5.2.1	Die drei Vs – und der Wert .....	232
5.2.2	Scale-up und Scale-out .....	232
5.2.3	Unabhängige Verarbeitung direkt auf den Daten .....	233
5.2.4	Schema on Read versus Schema on Write .....	234
5.2.5	Hardwarevirtualisierung und Containermanagement ..	234
5.2.6	Datenvirtualisierung .....	235
5.2.7	Entkoppelte Systeme .....	236
<b>6</b>	<b>Technische Big-Data-Lösungen zur industriellen und kommerziellen Datenanalyse .....</b>	<b>237</b>
6.1	Datenmanagement im Big-Data-Umfeld .....	237
6.1.1	Hadoop machte den Anfang .....	237
6.1.2	Apache Spark – die nächste Evolutionsstufe .....	240
6.1.3	Abstrahierte Datenverarbeitung und -speicherung .....	241
6.1.4	Komplexe Eventverarbeitung mit Kafka & Co. ....	245
6.1.5	Das beste beider Welten – von Lambda und Kappa .....	246
6.1.6	Big-Data-Plattformen .....	247
6.1.7	NoSQL-Datenbanken .....	248
6.1.8	Anwendungsfälle für NoSQL-Datenbanken .....	249
6.1.9	Technologiestacks .....	250
6.2	Datenzentrische Architekturen .....	251
6.2.1	AI-basierte Systeme brauchen IA-basierte Plattformen ..	251
6.2.2	Die logische Architektur .....	252
6.2.3	Die Softwarearchitektur .....	252
6.2.4	Die technische Architektur .....	252
6.3	Der Supervised Data Lake (SDL) .....	253
6.3.1	Ein Data Lake braucht ein Konzept, damit der See nicht zum Sumpf wird .....	253
6.3.2	Die unterschiedlichen Bereiche eines SDL .....	255
6.3.3	Quellen und Ladearten .....	255

6.3.4	Raw Zone	256
6.3.5	Ingestion Zone	256
6.3.6	Discovery und Sandbox	256
6.3.7	Integration	257
6.3.8	Serving	258
6.3.9	Associated Processes	258
6.3.10	Access und Application	259
6.4	Aufbau eines Data Lakes	259
6.4.1	Think Big – Start Small – Act Now	259
6.4.2	Vision, Ziele und Standortbestimmung	260
6.4.3	Konzeption des Data Lakes	260
6.4.4	Implementierung der Basisumgebung	261
6.4.5	Data Lake Ramp-up – Use Case Driven	261
6.4.6	Industrialisierung – die betriebsfokussierte Datenfabrik	262
6.5	Cloud-Computing und Services	263
6.5.1	Die Cloud-Ausbaustufen – Everything as a Service	264
6.5.2	Offene Ökosysteme	265
6.5.3	Der Data Lake in der Cloud	266
6.6	Big Data, Data Mining und Artificial Intelligence	268
6.6.1	Analytic Data Hub	269
6.6.2	Data-Science- und Data-Mining-Plattformen	270

<b>7</b>	<b>Die Anwendersicht – Systematik für industrielle Anwendungen</b>	<b>279</b>
7.1	Aufgabenstellung und Zielsetzung	279
7.1.1	Datengetriebene Identifikation von Aufgabenstellungen	279
7.1.2	„Produktgetriebene“ Identifikation	280
7.1.3	Geschäftsorientierte Identifikation von Aufgabenstellungen	280
7.1.3.1	Reduktion von Kosten, Verlusten, Verschwendungen	283
7.1.3.2	Erhöhung operativer Performance	284
7.1.3.3	Ergebnisverbesserung funktionaler Prozesse	285
7.2	Vorgehensmethodik	286
7.2.1	Workshop zur Ideenfindung und Datenanalyse	289
7.2.1.1	Design-Thinking-Workshop	289
7.2.1.2	Wertschöpfungsschritte	290
7.2.1.3	Perspektiven	291
7.2.1.4	Schmerzpunkte und Mehrwerte	292
7.2.1.5	Erzeugen des Mehrwertes	292

7.2.1.6	Geschäftsmodell .....	294
7.2.1.7	Anwendungen und Lösungsansätze identifizieren .....	296
7.2.2	Hackathons als alternative Möglichkeit der Lösungsfindung und Pilotierung .....	297
7.2.3	Aufsetzen konkreter Aufgabenstellungen .....	299
7.2.3.1	Definition der Aufgabenstellung .....	299
7.2.3.2	Modellauswahl .....	300
7.2.3.3	Beauftragung von Dienstleistern .....	301
7.2.4	Explorations- und Umsetzungsphase eines Use Case ...	302
7.2.4.1	Sichtung der Daten .....	302
7.2.4.2	Bestimmung der sensitiven Eingangsgrößen ..	308
7.2.4.3	Modellierung und Ergebnisbewertung .....	315
7.2.4.4	Die Königsklasse: Vektorielle Optimierung eines Use Case .....	316
7.2.5	Auswertung und Detailkonzept, Applikationserstellung und Implementierung .....	321
<b>8</b>	<b>Die Anwendersicht – typische Anwendungsfelder am konkreten Beispiel .....</b>	<b>327</b>
8.1	Anwendungen in den Geschäftsfunktionen .....	330
8.1.1	Forschung und Entwicklung .....	330
8.1.2	Engineering .....	333
8.1.3	Produktmanagement .....	334
8.1.4	Einkauf, Supply Chain Management, Logistik .....	336
8.1.5	Fertigung und Produktion .....	338
8.1.6	Qualitätsmanagement .....	340
8.1.7	Service und Instandhaltung .....	342
8.1.8	Service und After Market .....	344
8.1.9	Marketing und Vertrieb .....	347
8.2	Ausgewählte Data-Mining- und Big-Data-Beispiele .....	350
8.2.1	Forschung, Entwicklung und Engineering .....	351
8.2.1.1	Beschleunigung einer Produktentwicklung ...	351
8.2.2	Einkauf .....	358
8.2.2.1	Spend Cube .....	360
8.2.2.2	Bündelung .....	363
8.2.2.3	Spezifikations- und Kostenhebel .....	366
8.2.3	Produktion, Fertigung und Service .....	370
8.2.3.1	Störungsanalysen .....	370
8.2.3.2	Instabilitätsanalysen in einem Klärwerk .....	372
8.2.3.3	Fehlerdetektion in einem Kraftwerk .....	381

8.2.3.4	Analyse der Dynamik von chemischen Batchprozessen .....	390
8.2.4	Instandhaltung und Service .....	394
8.2.4.1	Aufbau einer Datenbasis für erweiterte Analysen und Monitoring von Industrieanlagen .....	394
8.2.4.2	Erweiterung eines digitalen Zwillings um Maschinendaten und Strompreisdaten im Bereich Windenergie .....	396
8.2.5	Marketing und Vertrieb .....	398
8.2.5.1	Cross-Selling-Effekte mit Data Mining finden ..	398
8.2.5.2	Cross-Selling-Analysen mit Big-Data- Technologien beschleunigen .....	405
8.2.5.3	Optimale Preisschwellen mit Data Mining aufspüren .....	407
8.2.6	Data Mining für die strategische Unternehmensführung	412
<b>9</b>	<b>Small Data gehört die Zukunft .....</b>	<b>421</b>
9.1	Einführung in die Thematik .....	421
9.2	Charakteristik von Small Data .....	423
9.3	Machine Learning versus menschlicher Geist – die Mind-Data- Hypothese .....	428
9.4	Bewusstsein als übergeordnete Ordnungsstruktur neuronaler Systeme .....	431
9.5	Mind-Data-Auswertungen mit maschinellem Bewusstsein .....	442
<b>10</b>	<b>Ausblick und mögliche Weiterentwicklungen von Data Mining und Big Data .....</b>	<b>451</b>
<b>11</b>	<b>Liste der häufig verwendeten Formelzeichen und Symbole ..</b>	<b>457</b>
<b>12</b>	<b>Literaturverzeichnis .....</b>	<b>461</b>
<b>13</b>	<b>Autoren .....</b>	<b>471</b>
	<b>Index .....</b>	<b>473</b>

# 5

## Big Data – die Datenhaltungs- und Verarbeitungskonzepte der Gegenwart

Wenn wir uns in den folgenden Kapiteln 5 und 6 über Big Data unterhalten wollen, müssen zwangsläufig die zugeordneten Fachbegriffe genannt werden. Für den Experten sind sie sofort klar, der Laie wird aber wohl sehr oft einen in der Big-Data-Branche wie selbstverständlich gebrauchten Terminus nicht verstehen und (im Internet) nachschlagen müssen. Denn im Unterschied zu den mittlerweile auch außerhalb der Community bekannten maschinellen Lernverfahren der Kapitel 3 und 4, wie etwa neuronale Netze, Entscheidungsbäume, Assoziationsregeln oder eventuell auch selbstorganisierende Merkmalskarten, sind wichtige Big-Data-Begriffe für sehr viele Nutzer noch echtes Neuland. Das vertiefende Nachlesen einiger Fachbegriffe erscheint daher unumgänglich.

Gleichzeitig können die nachfolgenden Ausführungen aber auch wie ein „What’s what“-Fachlexikon benutzt werden, da alle relevanten Fachbegriffe in einer dem Thema angepassten Struktur angeboten werden. Manchen Lesern werden die Ausführungen deshalb als eine sehr nützliche, logisch strukturierte Zusammenfassung ihnen schon bekannter Fachinhalte dienen können, für die anderen, noch themenfremden, aber interessierten Leser werden sie zu der Erkenntnis führen, dass sie hier noch viel investieren müssen, wenn sie diese moderne Technologie durchdringen wollen. Manche aber, die relativ leichtfertig von Big Data, künstlicher Intelligenz und digitaler Transformation sprechen, sollten verinnerlichen, dass wir mit Big Data nicht nur einen revolutionären Technologiewandel benennen, sondern vor allem äußerst komplexe technologische Vorgänge intellektuell erfassen müssen. Der Big-Data-Zug fährt schon eine Weile, aber noch kann man aufspringen. Wir hoffen, dass es Ihnen gelingt.

## ■ 5.1 Digitale Transformation und Big Data

Nachdem Sie bisher wichtige Data-Mining-Verfahren kennengelernt haben, teilweise auch im mathematischen Detail, soll in den Folgekapiteln ein Überblick über Big-Data-Konzepte und ihre Technologien gegeben werden, denn Big Data stellt eine der wesentlichen Grundlagen für die digitale Transformation dar.

Die Digitalisierung und die damit einhergehende digitale Transformation sind allgegenwärtig und bringen tief greifende Veränderungen mit sich. Dabei ist die Digitalisierung, wie wir sie gegenwärtig erleben, das Zusammenspiel ganz unterschiedlicher Entwicklungen: Sie resultiert aus einer Beschleunigung und Integration eines ganzen Bündels technologischer Innovationen, die, jede für sich, heute bereits einen bestimmten Reifegrad erreicht haben. Für Firmen bedeutet dies, dass sie – weit über die bislang bestehenden Grenzen klassischer IT-gestützter Prozesse hinaus – in umfassender Weise neue Informationstechnologien einsetzen. Oder, anders betrachtet: Unternehmen nutzen IT-Innovationen intensiver als je zuvor und verwenden dabei unter anderem neben eigenen auch unterschiedlichste externe Daten.

Dabei lassen sich verschiedene Varianten der Digitalisierung unterscheiden. In manchen Fällen liegt der Fokus auf der Granularität, mit der die reale Welt in einem „digitalen Zwilling“ abgebildet wird. In anderen Fällen geht es eher um Automatisierung, Optimierung, Individualisierung sowie Geschwindigkeit, und in wieder anderen Fällen liegt der Akzent auf der Innovation von veränderten Geschäftsprozessen. In allen Fällen sind Daten der eigentliche Treibstoff der Entwicklung oder, wie es oft heißt, neben Material, Arbeitskraft und Kapital der „vierte Produktionsfaktor“ einer modernen Ökonomie.

Ob im privaten, öffentlichen oder im geschäftlichen Umfeld – die Nutzung und Produktion von Informationen sowie die daraus entstehenden Datenmengen nehmen ständig zu. All diese Einflüsse führen zu einem Wandel innerhalb der Unternehmen, der sich in verschiedenen Handlungsfeldern niederschlägt. Die Digitalisierung wird in Unternehmen strategisch verankert, neue Digitalisierungsabteilungen entstehen oder sind entstanden; Methoden, Prozesse und Technologien müssen so angepasst werden, dass sie mit der stetigen Zunahme der Datenvielfalt und -menge umgehen können.

Und so wird die IT-Landschaft, die uns begegnet, immer komplexer. Die entstehenden Daten strömen auf Unternehmen ein und sollen verarbeitet und gespeichert werden. Allerdings gibt es nicht die *eine* Technologie oder die *eine* Plattform, die alle Daten in ihrer Vielfalt speicher- und verarbeitbar macht. Vielmehr entstehen unterschiedliche Verarbeitungen, Speichermedien oder Plattformen, die auf den jeweiligen Anwendungsfall optimiert sind.



**Bild 5.1** Anforderungen an moderne Datenplattformen

Die Anforderungen an moderne Datenplattformen werden somit immer komplexer (siehe Bild 5.1). Sie müssen bei Bedarf schnell und effizient erweiterbar sein und eine Datenhaltung bieten, die agil und flexibel den sich ändernden Anforderungen des Unternehmens angepasst werden kann. Neue Technologien und Methoden fördern hierbei die Fähigkeit, via Self-Service explorativ auf Datenbestände zuzugreifen und den Zeitraum von der Idee bis zur Analyse und Marktreife eines datengetriebenen Produkts oder Prozesses zu verkürzen. Neue komplexe Datenquellen müssen schnell und in Echtzeit angebunden werden. Des Weiteren verschmelzen hierbei operative, dispositive sowie analytische Systeme und agieren in virtuellen Datenmodellen miteinander. Für erweiterte Analysen und Anwendungen im Kontext von künstlicher Intelligenz (KI, AI) oder dem in Kapitel 3 beschriebenen Data Mining bzw. maschinellem Lernen (Machine Learning, ML) ist dieser (virtuell) zusammenhängende Datenspeicher die Basis für die Erstellung von Modellen und erweiterter Analyse. Die in vorherigen Kapiteln verwendeten Flat Tables sind eine einfache Ausführungsvariante dieser virtuellen Datenspeicher.

Im Folgenden sollen Grundprinzipien des Paradigmenwandels, der Digitalisierung und Big Data antreibt, dargestellt, auf Kerntechnologien eingegangen und Konzepte zum Aufbau von Big-Data-Systemen sowie deren Zusammenspiel mit Data-Science-Plattformen aufgezeigt werden.



## ■ 5.2 Grundprinzipien eines Paradigmenwandels

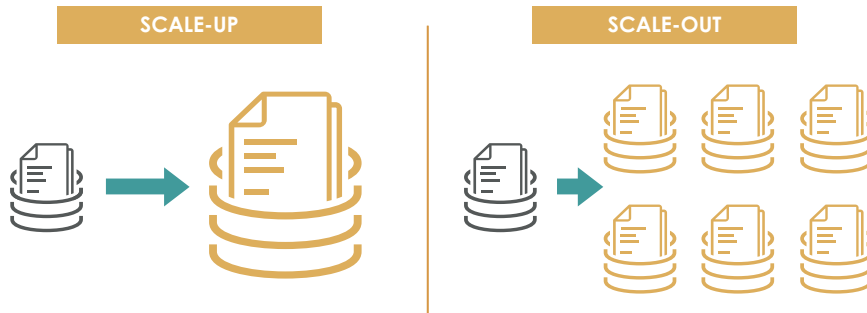
In einer zunehmend digitalisierten Welt fallen Daten in großer Menge, mit hoher Geschwindigkeit und/oder in unterschiedlicher Form an. Hinzu kommen innovative und zugleich kostengünstige Arten der Verarbeitung, die Analysemöglichkeiten, Entscheidungsfindung und Prozessautomatisierung verbessern und erweitern. Nachstehend werden die Grundprinzipien von Big Data vorgestellt und deren Einfluss auf die Infrastruktur und die Abläufe in Projekten erläutert.

### 5.2.1 Die drei Vs – und der Wert

Zu Beginn des Big-Data-Zeitalters sind die drei sogenannten V-Begriffe entstanden, die das Thema genauer umschreiben sollen – *Volume*, *Variety* und *Velocity*. *Volume* (Menge) bezeichnet die schiere Masse an Daten, die mit herkömmlichen Methoden der Datenverarbeitung nicht mehr gespeichert, verarbeitet oder gar analysiert werden können. *Variety* (Vielfalt) steht für die Anforderung, wenig oder gänzlich unstrukturierte Daten durch Algorithmen in eine auswertbare Struktur umzuformen, in der Zusammenhänge erkannt und mit bekannten Informationen verknüpft werden können. *Velocity* (Geschwindigkeit) steht für den Umstand, Daten in kleinsten Zeitabständen zu verarbeiten und auswertbar zu machen. Im Laufe der Jahre sind weitere „Vs“ als Begriffsunterstützung hinzugekommen [UM19]. Hier sei lediglich noch der wichtige Begriff des Wertes (*Value*) zu nennen, denn eine zentrale Frage sollte bei der gesamten Datenverarbeitung immer im Vordergrund stehen: Welcher Nutzen lässt sich aus den Daten ziehen?

### 5.2.2 Scale-up und Scale-out

Ein klassischer Ansatz zur Skalierung, um Datenwachstum und Ressourcenanforderungen gewachsen zu sein, ist das *Scale-up-Prinzip* (siehe Bild 5.2). Ganz nach dem Motto „*The Bigger the Better*“ entstehen immer stärkere und größere monolithische Serversysteme. Dies führt neben Vorteilen zu steigenden Hardware- und Lizenzkosten – denn nicht selten koppeln Softwareanbieter ihre Lizenzen an die Hardwaregegebenheiten. Die Datenzentralität und damit der „Single point of failure“, also ein neuralgischer Fehlerpunkt im System, ist als weiteres Problem aufzuführen; in diesen Fällen kann nur durch kostenintensive Backup- und K-Fall-Szenarien die Ausfallwahrscheinlichkeit gemindert werden. Die Datenverarbeitung ist zudem mit einer hohen Netzwerkauslastung verbunden, da die Daten oft für die Clients kopiert werden müssen.



**Bild 5.2** Scale-up und Scale-out

Vorteile hat dieses Vorgehen in der Architektur, weil es die Verwaltung des Systems vereinfacht. Doch bei ständig wachsenden Datenmengen ist in der Regel irgendwann die maximale Ausbaustufe erreicht. Diese Limitierung überwindet der sogenannte *Scale-out-Ansatz*, der als Basisarchitektur für eine verteilte Verarbeitung im Rechnerverbund anzusehen ist. Hierbei werden mehrere kleine Systeme als Cluster zusammengeschlossen; sowohl Daten als auch Verarbeitung befinden sich gleichzeitig auf allen Rechnern (Nodes) des Clusters. Bei einer Erweiterung um Hardwarekomponenten ist eine lineare Skalierung der Leistung und Speicherkapazität möglich. Durch die Verwendung von vielen kleinen, in der Regel kostengünstigeren Rechnern wird hierbei zusätzlich eine Kostenersparnis erzielt.

### 5.2.3 Unabhängige Verarbeitung direkt auf den Daten

Die verteilte Rechnerarchitektur arbeitet nach einem Ansatz, bei dem jeder Rechner die Speicherung und Verarbeitung unabhängig von anderen Systemen vornimmt (*Shared-Nothing-Prinzip*). Anders als in anderen Serverarchitekturen werden hier keine großen Kopiervorgänge vorgenommen, um die Daten einer Verarbeitung zuzuführen; stattdessen wird hier die Verarbeitung der Daten den einzelnen Nodes im Cluster zugeführt. Durch den hohen Grad an Parallelität wird so die Verarbeitungsperformance enorm gesteigert (siehe Bild 5.3).



#### LESEZEIT FÜR 3 TB DATEI



**Bild 5.3** Verteilte Verarbeitung

### 5.2.4 Schema on Read versus Schema on Write

In einem Data Warehouse oder darauf aufsetzenden Data Marts, aber auch in operativen Systemen im relationalen Umfeld kennt man seit Langem den Ansatz *Schema on Write*. Besonders bei der deskriptiven Datenmodellierung werden Annahmen über Datenbeziehungen getroffen, die in ein Modell münden, das ein statisches Datenbankschema abbildet. Dieses Schema wird vor der Datenbeladung definiert, neue Attribute oder Spalten müssen explizit hinzugefügt werden.

Bei der sogenannten präskriptiven Datenmodellierung, dem Ansatz *Schema on Read*, werden die Daten im Rohformat abgelegt, und die Schemata werden erst für den Lesevorgang erzeugt (Parser). Abfragen über den Urzustand der Daten sind hiermit möglich, Beziehungen zwischen Datenobjekten werden erst bei der Abfrage hergestellt. Auf denselben Datenbestand können also unterschiedliche Schemadefinitionen angewendet werden. Dieses Vorgehen findet man in verteilten Frameworks, wie beispielsweise Hadoop, und es eignet sich besonders für explorative Ansätze [ML19].

### 5.2.5 Hardwarevirtualisierung und Containermanagement

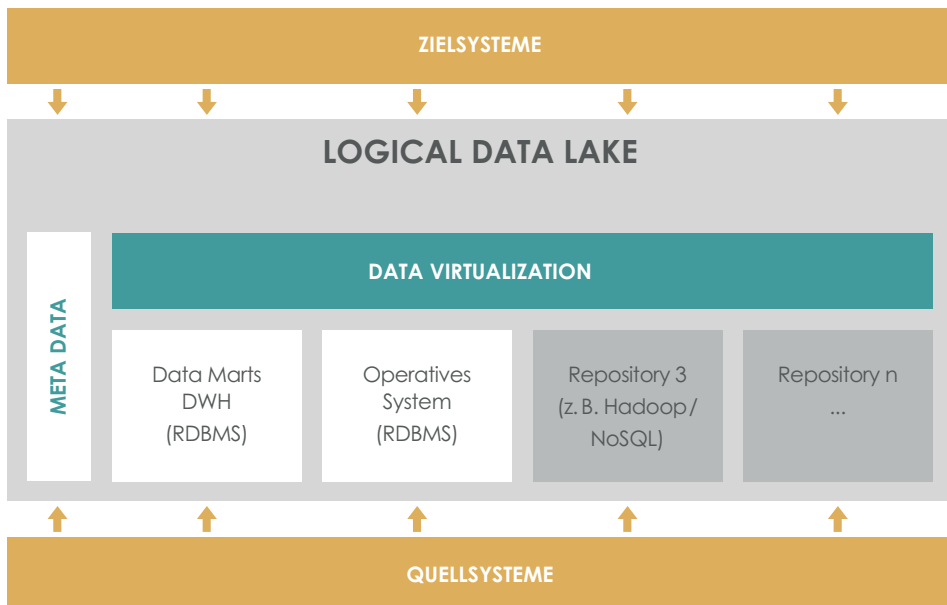
Durch die Möglichkeit, virtuelle Hardware in gekapselten Objekten zur Verfügung zu stellen, verringert man zum einen Abhängigkeiten, zum anderen wird die Kompatibilität mit beliebigen Betriebssystemen erhöht. Mit Containern in der IT verhält es sich wie mit Containern in der Logistik: Sie erhöhen die Flexibilität, Portabilität, Effizienz und Wiederverwendbarkeit und senken damit erheblich die Kosten. Bei Docker handelt es sich um eine Open-Source-Software, mit der sich Anwendungen ähnlich wie bei einer Betriebssystemvirtualisierung in Containern

isolieren lassen. Docker baut und verwaltet Container. Jede abhängige Datei, die für den ordnungsgemäßen Betrieb einer Applikation benötigt wird, ist in einer sehr kleinen Datei namens Container verpackt. Solange ein Container auf einen ähnlichen Docker-Host geladen wird, läuft die Applikation ohne größere Konfigurationsanpassungen weiter. Container beinhalten isoliert nur die Abhängigkeiten einer einzelnen Applikation und lassen sich aufgrund ihrer geringen Größe schnell installieren und transportieren.

### 5.2.6 Datenvirtualisierung

Techniken zur Datenvirtualisierung erlauben es, die unterschiedlichen physikalischen Datenquellen eines Unternehmens virtuell zusammenzuführen und zu verknüpfen, ohne die Daten dabei von ihrem Ursprungsort zu entfernen. Datenvirtualisierung kombiniert einen integrierten Echtzeitzugriff auf heterogene Quellsysteme mit *In-Memory-Zugriffsverfahren* und bietet Übersetzungsmechanismen für Abfragesprachen, um auf strukturierte und unstrukturierte Datenquellen zuzugreifen. Durch Datenvirtualisierung werden losgelöste Datenbestände verschiedener Unternehmensteile aus unterschiedlichen Blickwinkeln und Detaillierungsgraden auf Nutzdaten und Metaebene miteinander verbunden, und es entsteht eine übergreifende virtuelle Nutzungsschicht.

Mit einem Bruchteil der Integrationsaufwände herkömmlicher Verfahren können auswertbare Datenmodelle für eine Analyse zur Verfügung gestellt werden.



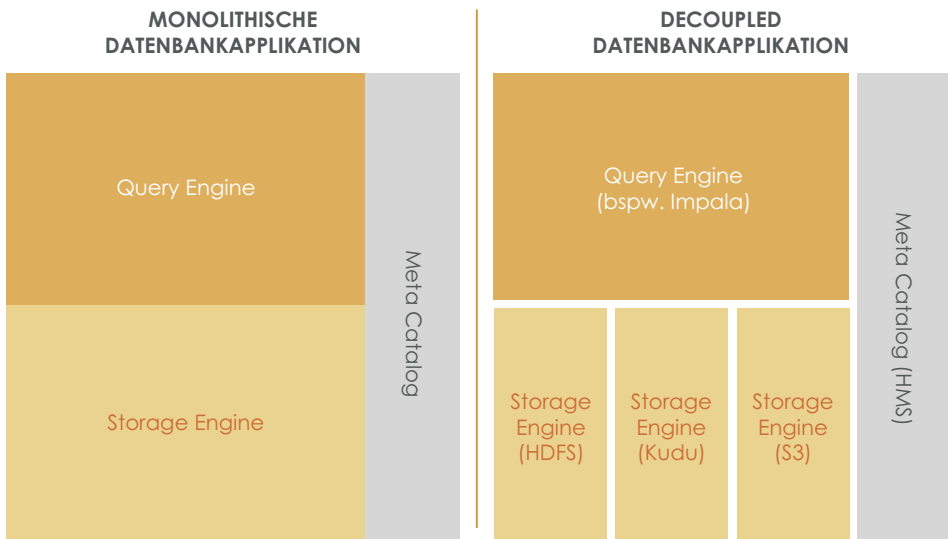
**Bild 5.4** Logical Data Lake

Die Konzepte eines Logical Data Warehouses oder eines sogenannten *Logical Data Lakes* nutzen Datenvirtualisierung zur Integration von Systemen sowie zur Abstraktion zwischen Backend- und Frontendsystemen (siehe Bild 5.4). Für alle Zielsysteme kann die technische Komplexität des Kerns vollständig verborgen werden – durch die Datenvirtualisierung wirkt das System wie eine fachlich und technisch vollständig integrierte zentrale Datenebene.

## 5.2.7 Entkoppelte Systeme

Daten, Verarbeitung und Metadaten sind in monolithischen Systemen fest miteinander verbunden. Der Trend, in Systemen die Zuständigkeiten in unterschiedliche Bereiche zu entkoppeln – sogenannte *Decoupled Systems* aufzubauen –, ist ein weiterer Treiber im Big-Data-Umfeld (siehe Bild 5.5).

Die hierbei vorgenommene Trennung der Daten von der Verarbeitungs-Engine ermöglicht es, optimierte Werkzeuge für unterschiedliche Anwendungen zu verwenden.



**Bild 5.5** Entkoppelte Systeme

Prof.Dr.-Ing. Ralf Otte ist seit 2015 Hochschullehrer für Prozessautomatisierung und Künstliche Intelligenz an der Technischen Hochschule Ulm. Von 1992 bis 2004 war er Leiter des Data Mining Centers und der BI-Chef der ABB Process Industries und in dieser Funktion verantwortlich für die Durchführung zahlreicher Data-Mining-Projekte weltweit. Im Jahr 2004 wechselte Otte in die Schweiz und führte 10 Jahre lang als Geschäftsführer die tecData AG, ein Unternehmen im Bereich der unternehmensweiten Datenanalyse.

Boris Wippermann ist Principal bei der h&z Unternehmensberatung aus München. Dort beschäftigt er sich schwerpunktmäßig mit der Operational Excellence von Unternehmen, insbesondere durch Digitalisierungsanwendungen. Bereits Ende der 1990er Jahre baute er bei der ABB den Bereich des technischen Consultings für Optimierungslösungen durch Digitalisierung in der Prozessindustrie auf.

Sebastian Schade ist Lead Consultant der INFOMOTION GmbH Frankfurt und arbeitet dort seit Jahren als leitender Entwickler und Architekt für datenbasierte Systeme. Die INFOMOTION GmbH ist das führende Beratungsunternehmen für Business Intelligence, Big Data und Digital Solutions im deutschsprachigen Raum.

Prof.Dr.-Ing. habil. Viktor Otte ist emeritierter Hochschullehrer der Universität Wuppertal im Fachgebiet Maschinenbau und langjähriger Berater für Data-Mining-Projekte auf Kundenseite in den Bereichen Fertigungstechnik und Automobilzulieferer.

# Index

## A

After Sales 281  
Agglomerative Verfahren 89  
Ähnlichkeitsmaß 87  
Aika 273  
Aktivierungsfunktion 103  
Alteryx 276  
Amazon 243, 264  
Angoss 273  
Anlagenfahrer 4  
Apache 244, 266  
Apache Cassandra 251  
Apache Drill 258  
Apache Foundation 240  
Apache HBase 238  
Apache Spark 240, 251  
Aphelion 273  
Arbeitspunkte 312  
Artificial Intelligence 268  
Assoziationsregeln 20, 145, 162  
Ausreißer 45, 53  
Autokorrelation 378  
Automatische Diskretisierungsverfahren 451  
Autonomes Fahren 15  
Azure 241  
Azure Data Lake 395

## B

B2C 347  
Backpropagation-Lernregel 108, 111  
BANF 360

Batchlauf 393  
Batchprozesse 390  
Bayessche Netze 145, 151  
Bedeutung einer Information 8  
Benfordsches Gesetz 61  
Bestimmtheitsmaß 70 f., 119  
Bestrafungsterme 317  
Beweisbarkeit 430  
Bewusstseinsobjekte 435  
Bewusstseinszustände 434  
Big Data 3, 5, 14, 229, 243  
Big-Data-Anbieter 237  
Big Data Mining 279  
Big-Data-Projekte 299  
Box-Plots 55  
Bremsverhalten 351  
Business Intelligence 268  
Business Understanding 34

## C

Caffe 273  
CART 273  
Cascading 242  
Cash Flow Return on Investment 413  
Chatbots 196  
Claim Management 281  
Cloud-Computing 263  
Cloudera 247, 258, 265  
Clusteranalyse 12 f., 19 f., 86, 139, 311  
CNN 445  
Confident-Modelle 207  
Confident-RBF-Netz 201

CRISP-Phasen 34  
 CRISP-Standard 33  
 Cross-Selling 403  
 Cross-Selling-Effekte 398  
 Crunch 242  
 Cybenko-Theorem 106

## D

Data Cleansing 269  
 Data Lake 227, 253, 256, 259  
 Data Marts 22  
 Data Mining 3, 8f., 15, 356  
 Data-Mining-Applikationen 331  
 Data-Mining-Ingenieur 263  
 Data Preparation 34  
 Data-Science-Plattformen 231  
 Data Transformation 34  
 Data Understanding 34  
 Data Virtualisation 270  
 Data Warehouses 22  
 Daten 6f.  
 Datenauswertungen mit Bewusstsein 431  
 Datenlücken 51f.  
 Datenmangel 356  
 Datenplattformen 231  
 Datenreduktion 23  
 Datenschutz 328  
 Datenschutz-Grundverordnung 254, 329  
 Datentransformation 60  
 Datenvielfalt 426  
 Datenvirtualisierung 235  
 Decoupled Systems 236  
 Delta-Lernregel 108, 111  
 Demographisches Clustern 130  
 Dendrogramm 91  
 Denormalisierung 36  
 Deployment 34  
 Design-Thinking-Workshop 288f.  
 Design-to-Cost 366  
 Digitalcomputer 442  
 Digitale Transformation 230  
 Digitalisierung 230  
 Diskretisierung 41, 202

Diskriminanzachse 78  
 Diskriminanzanalyse 18, 76  
 Diskriminanzmaß 77  
 Divisive Verfahren 89  
 D-U-N-S-Nummer 362  
 Dynamische Prozesse 376  
 Dynamische Sensitivitätsanalyse 218

## E

Echtzeitzugriff 235  
 e-Log-Modelle 68  
 Empirische Inferenz 216  
 Empirische Modellbildung 25  
 Engineeringaufwände 28  
 Enterprise Map 413  
 Entscheidungsbäume 20, 98, 145, 219, 355, 401  
 Entscheidungsbaumverfahren 203  
 ETL 242, 252, 266  
 EU-DSGVO 330  
 Evaluation 34  
 Evidenzbasierte Medizin 216  
 Evolution der schwachen KI 449  
 Evolutionsstrategien 21, 216  
 Experimentelle Modellbildung 25  
 Exponentialmodelle 68

## F

Faktoranalyse 19, 83, 139  
 Faktorielle Versuchspläne 198  
 Faktorstufen 74  
 Fehler 48  
 Fehlerdetektion 370, 381  
 Fehlerdiagnose 387  
 Fehler erster Art 180  
 Fehlerfortpflanzungsrechnung 50  
 Fehlerlokalisierung 371  
 Fehlerquadrat 68  
 Fehlerursachenbestimmung 371  
 Fehler zweiter Art 180  
 Fehlinterpretationen 204  
 Flat Table 35, 37  
 Flugwespe 437



Flume 243  
F-Test 46  
Fuzzy-Techniken 21  
F-Wert 75

## G

Gain 203  
Genetische Algorithmen 21, 216  
Gesamtstreuung 70  
Gewinnerneuron 208, 385  
GIS-Karte 330, 405  
Google 264  
Gütefunktion 222  
Gütemaße 119, 143  
Gütemaße für Projektionen 173  
Gütemaße für Regeln 165

## H

H2O 243, 273  
H2O.ai 276  
Hackathons 288, 297  
Hackfest 298  
Hadoop 247  
Hadoop Distributed File System  
237  
Hardwarevirtualisierung 234  
Hauptachsen 86  
Hauptachsentransformation 93  
Hauptkomponentenanalyse 304  
HBase 239, 249, 258  
HDFS 238 f., 258  
Hebbsche Lernregel 108  
Heuristischer Algorithmus 422  
Hidden-Neuronen 123  
hierarchische Clusteranalyse 89  
Histogramm 42, 65, 202, 399  
Hive 241, 257  
Hortonworks 247  
Hyperebene 117  
Hyperkugeln 92

## I

IBM 264  
IBM SPSS Modeler 276  
Identifikation von Ausreißern 451  
Imaginäre Struktur 439  
Imaginärprozesse 444  
Information 7, 153  
Informationsbedarf 154  
Informationsgewinn 155, 201  
Instabilitätsanalysen 372  
Instandhaltung 342  
Intervallskala 39  
Inverse Modelle 68  
Irrtumswahrscheinlichkeit 46 f., 180

## J

JAR-Skripts 238  
Java 237

## K

Kafka 245  
Kafka-Cluster 245  
Kaiser-Kriterium 85  
Kappa 246  
Key Performance Indicator 192  
Klassifikation der KI 450  
k-means Verfahren 90, 92  
KNIME 192, 243, 271, 273, 276  
KNN 99  
Kohonen-Karte 134  
Kommunalitäten 84  
Komponentenkarten 171  
Konfidenz 164 f.  
Konfidenzintervall 205, 315  
Konnektionistische Systeme 6  
Kontingenzanalyse 18  
Korrelationsanalyse 12, 19, 79, 140  
Korrelationskoeffizient 79  
Korrelationsmatrix 83  
Korrelationswerte 84  
KPIs 412  
Kraftwerksvorgänge 386

Kreuztabelle 358  
Kreuzvalidierung 125  
Kudu 239, 258  
Künstliche Intelligenz 8, 231, 251, 455

## L

Lambda 246  
Latent Dirichlet Allocation 193  
Lean-Prinzipien 283  
Least Median Error 54  
Lernende Vektor-Quantisierung 115  
Levenberg-Marquardt 216  
Lift 165  
Lineare Modelle 68  
Lineare Regression 355  
LinkedIn 245  
Link-Graph 400  
Link-Nodes 67  
Logarithmierung 61  
Logical Data Lake 236  
Logistische Modelle 68  
Long Short Term Memory 195

## M

Machine Vision 432  
Mahalanobis-Distanz 89  
Mahout 243  
Manhattan-Distanz 87  
MapR 247, 265  
MapReduce 238, 241  
MARS-Regression 409  
Maschinelles Lernen 6, 20, 231  
Maschinelles Sehen 429  
Massendatendisplays 24  
MATLAB 274  
Maximaler Informationsgewinn 203  
Median 44  
Mehrdeutigkeit 55, 57  
Mehrdeutigkeitsanalyse 59  
Merkmalskarte 131  
Messfehler 48  
Metagütefunktion 318  
Metrische Skalen 39

Microsoft Cognitive Toolkit 274  
MIDOS 212, 214, 218, 319  
Mind Data 428, 436  
Mind-Data-Applikationen 448  
Minkowski-Metrik 87  
Mischung von Experten 205  
Mittelwert 43 f.  
Modeling 34  
Modus 43  
MS Azure 264  
Multivariate Statistik 18

## N

Netzhaut 432  
Neuromodell 402  
Neuromorphe Computer 443  
Neuromorphe Datenauswertungen 452  
Neuromorphe Hardware 443  
Neuromorphes System 8, 447  
Neuronale Lernverfahren 100  
Neuronale Netze 21, 98 f.  
Neuronale Sensitivitätsanalyse 416  
Neuronales Korrelat 432, 440  
Neuronales Vorhersagemodell 13  
Neuronenmodell 100  
Nichtlokalität 441  
Nicht-metrische Skalen 39  
Nominalskalen 39  
Normierung 60  
NoSQL 248  
NoSQL-Datenbanken 238  
Nullhypothese 46

## O

ODBC 259  
OEE 338, 394  
Offlineoptimierung 323  
Online Analytical Processing 17  
Onlineoptimierung 323  
Optimierungskriterium 286  
Optimierungsvorgaben 221  
Oracle Data Mining 274  
Orange 274

Ordinalskala 39  
Ordnungsstruktur 439  
Overfitting 122

## P

Patentanalysen 331  
Physikalisches Symbolsystem 442  
Pig 242  
Piranha 274  
Pivotal 265  
PL2 430  
Plateauauswahl 225  
Polyoptimum 319  
Potenzmodelle 68  
Prädikatenlogik 430, 433  
Preis-Absatz-Daten 408  
Preisschwellen 407  
Proximitätsmaß 86  
Prozessanalyse 23, 26, 322  
Prozessmodellierung 23, 25  
Prozessoptimierung 23, 28  
Prozessprognose 23, 28  
Prozessvisualisierung 23  
PSSH-Theorem 442  
p-Wert 48, 213  
Python 257  
Pytorch 274

## Q

Q-Korrelationskoeffizient 88  
Qualitätskosten 340  
Qualitätsmängel 337  
Qualitätsschwankungen 390  
Qualitätssicherung 281  
Quantencomputer 443  
Quantenprozesse 442  
Quantensprung 427

## R

R 274  
Rapid Miner 271, 274, 276  
Rationalskala 40

Raw Zone 256  
RBF-Netz 221  
Referenzkanal 393  
Regelbäume 20  
Regressionsanalyse 12, 18, 67, 83, 98,  
140, 308  
Regressionsfunktion 68 f.  
Regressionskoeffizient 68 f.  
Reifenproduktion 352  
Relief-Algorithmus 211  
Repräsentative Daten 303  
Robuster Prozessentwurf 320  
ROI 415  
Rückverfolgbare Prozesse 304

## S

Sandbox 256  
SAS 274  
Scale-up 233  
Scatter-Plot 11  
Scatter-Plots 306  
Scheinkorrelation 82  
Schema on Read 234  
Schema on Write 234  
Schwachstellenanalyse 414  
Schwellenwertevent 53  
Selbstorganisierende Merkmalskarten  
131  
Semantische Information 197  
Sensitivitätsanalyse 72, 208, 211, 310,  
313, 387, 402  
Separierbarkeit 102  
Shared-Nothing-Prinzip 233  
Shogun 275  
Signalorientierte Prozessanalyse 29  
Single Linkage-Verfahren 90  
Skalenniveaus 39  
Small Data 421, 427  
Social Media 226, 249  
Soft Computing 6, 21  
Solr 258  
SOM 131, 380  
SOM-Karte 132, 200, 223, 304, 384  
SOM-Komponentenkarte 220

SOM-Modelle 219  
SOM-Networks 108  
SOM-Verfahren 199  
Spannweite 45  
Spark 247  
Spark Core 240  
Spark ML 243, 268, 271  
Spark Streaming 243  
Spend Cube 360  
Stabilitätsanalyse 225  
Stakeholder 286  
Standardabweichung 45, 119, 206  
Standardisierung 60  
Steilheitsanalysen 223  
Storm 243  
Störungen 371  
Störungsanalyse 373  
Streaming 245  
Streuung 70  
Strukturanalyse 10  
Superposition 222  
Supply Chain 336  
Supply Chain Management 281  
Support 165  
Support-Vektor-Maschinen 20, 98, 114

## T

Tanimoto-Koeffizient 87  
TensorFlow 262, 271, 275  
Teradata 276  
Text Mining 191  
Tez 241  
Theoretische Modellbildung 25  
TICK 251  
Toleranzbänder 27  
Topic Modeling 193  
Topologie von neuronalen Netzwerken 104  
Torch 275  
Trajektoriendarstellung 86  
Trendanalyse 10  
t-Test 46, 323

## U

UIMA 275  
U-Matrix-Methode 141, 168  
Unternehmensführung 412  
Unternehmensfunktionen 327  
Use Case 302  
USP 349

## V

Varianz 43, 119  
Varianzanalyse 18, 73  
Variety 232  
Vektordiagramm 357  
Vektorielle Prozessoptimierung 453  
Vektorraum 384  
Velocity 232  
Verarbeitungsgeschwindigkeit 427  
Versuchsplanungen 197  
Versuchsplanungsmethoden 197  
Vertrauensgrenzen 126  
Vertrauensschutz 328  
VIGRA 275  
Visualisierung der Prozesstrajektorie 384  
Volume 232  
Volumen der Daten 426  
Vorgehensmethodik 287  
Vowpal 275  
VUCA 286

## W

Wahrscheinlichkeit 154, 213  
Warenkorb 406  
Warenkorbbepreisung 367  
Wärmetauscher 343  
Web Mining 454  
Weka 275  
WEKA 353  
What-if-Analyse 26, 71  
What-if-Simulationen 217  
Windkraftanlagen 397  
Wissensextraktion 192

Wolfram Mathematica 275  
Workshops 287  
WRM-Verfahren 200

## Y

YARN 238  
Yooreeka 275  
Yottabytes 14

## Z

Zeroth 275  
Zettabyte 14  
Zielgrößengebirge 224  
Zusammenhangsanalyse 10