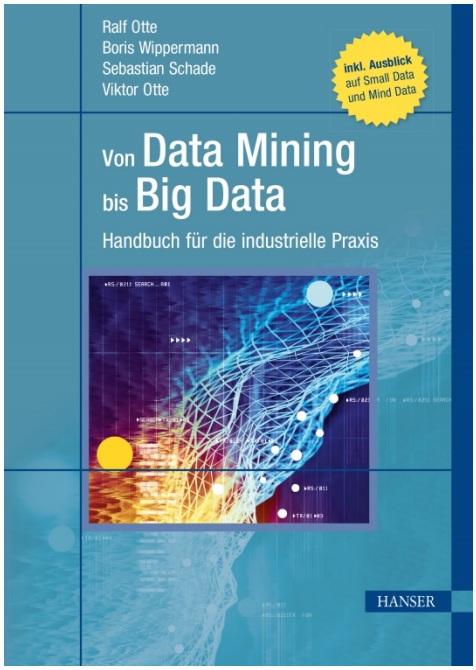
HANSER



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

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 Auswertmöglichkeiten 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 Problemkern 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 Rohrdantz 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

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.

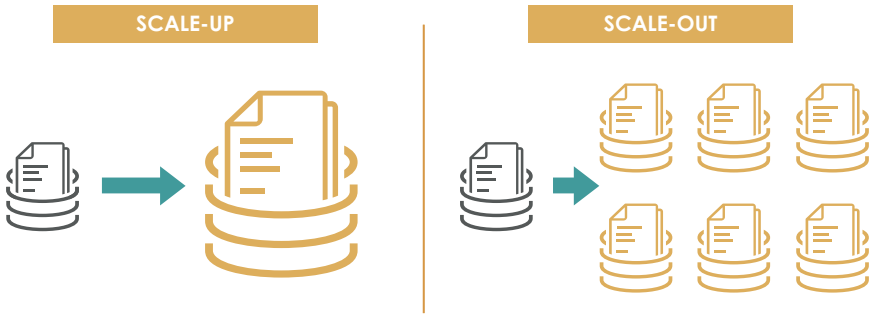


Bild 5.2 Scale-up und Scale-out

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).

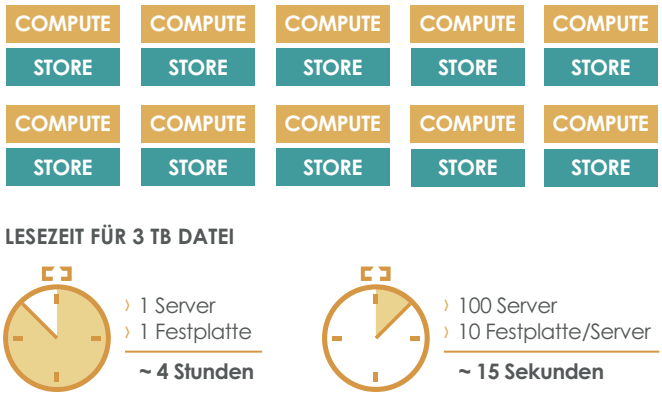


Bild 5.3 Verteilte Verarbeitung

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.

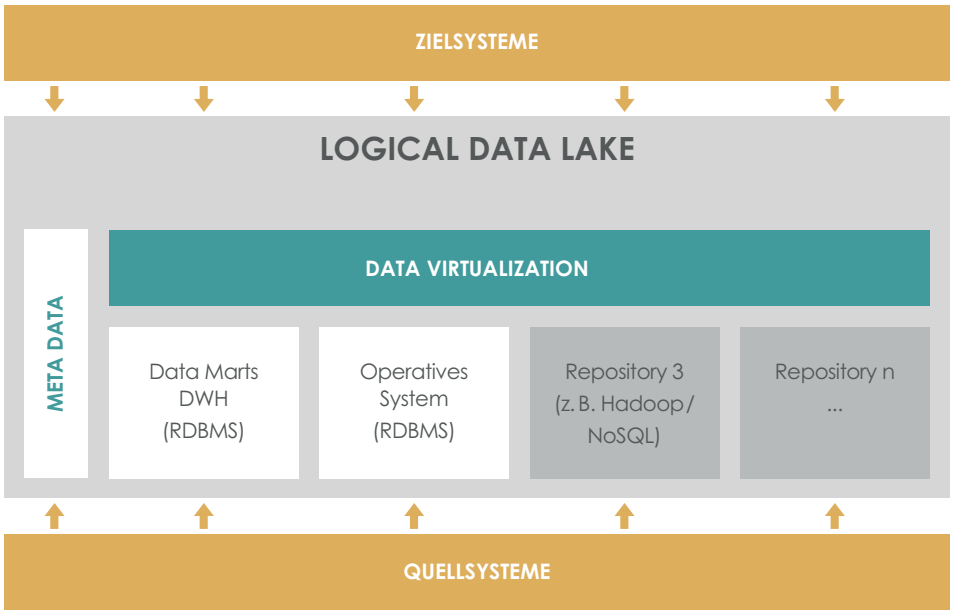


Bild 5.4 Logical Data Lake

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.

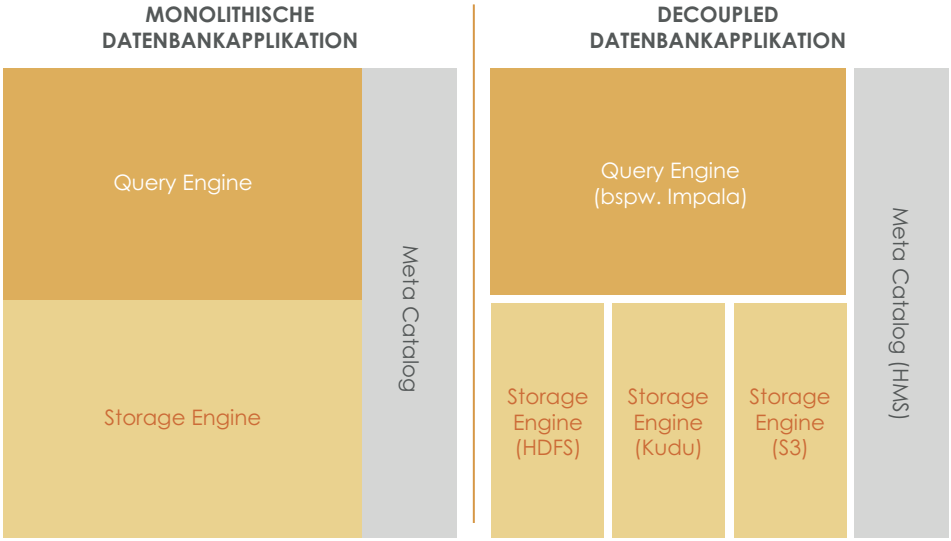


Bild 5.5 Entkoppelte Systeme

13 Autoren

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

D

E

F

T

U

V

W

Y

Z