O'REILLY®

Mit zahlreichen Illustrationen

Daten-architekturen

Modern Data Warehouse, Data Fabric, Data Lakehouse und Data Mesh richtig einsetzen

James Serra

Übersetzung von Frank Langenau

Stimmen zum Buch *Datenarchitekturen*

In *Datenarchitekturen* leistet James Serra einen tollen Job, indem er die Entwicklung der führenden Datenarchitekturen und die Kompromisse zwischen ihnen erläutert. Dieses Buch sollte zur Pflichtlektüre für derzeitige und angehende Datenarchitektinnen und -architekten werden.

– Bill Anton, Data Geek, Opifex Solutions

James hat über 30 Jahre Wissen und Weisheit über Datenarchitekturen in diesem umfassenden und sehr lesenswerten Buch zusammengefasst. Für diejenigen, die die eigentliche Arbeit bei der Bereitstellung von Analysen leisten müssen, anstatt Lobeshymnen darüber anzustimmen, ist dieses Buch ein Muss.

– Dr. Barry Devlin, Gründer und Direktor, 9sight Consulting

Diese Referenz sollte im Bücherregal eines jeden Datenarchitekten stehen. Mit klaren und aufschlussreichen Beschreibungen der aktuellen und geplanten Technologien werden die Leserinnen und Leser ein gutes Gefühl dafür bekommen, wie sie ihre Unternehmen steuern, um die Herausforderungen der sich ständig weiterentwickelnden Datenlandschaft zu meistern. Dies ist eine für Neueinsteiger und erfahrene Datenarchitekten gleichermaßen unschätzbare Referenz.

– Mike Fung, Master Principal Cloud Solution Architect, Oracle

Das Marketing-Getöse und das Gerede von Vordenkern der Branche haben viel Verwirrung über Datenarchitekturmuster gesät. Mit seiner umfassenden Erfahrung und seinen Kommunikationsfähigkeiten durchbricht James Serra die Störgeräusche und verschafft Klarheit sowohl über altbewährte Datenarchitekturmuster als auch über die neuesten Methoden in der Branche, die sowohl Datenpraktikern als auch Datenverantwortlichen helfen werden. Legen Sie es auf Ihren Schreibtisch – Sie werden es oft zurate ziehen.

– Sawyer Nyquist, Inhaber, Autor und Consultant, The Data Shop

Die Welt der Datenarchitekturen ist komplex und voller Störgeräusche. Dieses Buch bietet eine frische, praktische Perspektive, die auf jahrzehntelanger Erfahrung aufbaut. Egal ob Sie Einsteiger oder Expertin sind, jeder mit einem Interesse an Daten muss dieses Buch lesen!

– Piethein Strengholt, Autor von »Data Management at Scale«

Ein lehrreiches Juwel! *Datenarchitekturen* schafft ein perfektes Gleichgewicht zwischen Einfachheit und Tiefe, sodass Technologieexpertinnen und -experten aller Ebenen die entscheidenden Datenkonzepte erfassen und die wesentlichen Kompromissentscheidungen verstehen, die bei der Planung einer Datenreise wirklich wichtig sind.

– Ben Reyes, Mitbegründer und Managing Partner, ZetaMinusOne LLC

Ich empfehle *Datenarchitekturen* als Quelle, die das Wissen liefert, um die verfügbaren Optionen zu verstehen und sich darin zurechtzufinden, wenn es darum geht, eine Datenarchitektur zu entwickeln.

– Mike Shelton, Cloud Solution Architect, Microsoft

Datenmanagement ist entscheidend für den Erfolg eines jeden Unternehmens. Das Buch *Datenarchitekturen* zerlegt die Schlagwörter in einfache und verständliche Konzepte sowie praktische Lösungen, um Ihnen zu helfen, die richtige Architektur für Ihren Datensatz zu finden.

– Matt Usher, Director, Pure Storage

Als Berater und Community-Leiter verweise ich oft auf das Blog von James Serra, um aktuelle und ausführliche Informationen über moderne Datenarchitekturen zu erhalten. Dieses Buch ist eine großartige Sammlung, die Serras Reichtum an herstellerneutralem Wissen verdichtet. Mein Favorit ist Teil III, in dem James die Vor- und Nachteile der einzelnen Architekturen diskutiert. Ich glaube, dass dieses Buch jedem Unternehmen, das seinen Datenbestand modernisieren will, ungemein nützen wird.

– Teo Lachev, Consultant, Prologika

Das Blog von James ist meine erste Anlaufstelle, wenn es darum geht, Architekturkonzepte zu entmystifizieren, technische Fachbegriffe zu verstehen und sich im Leben eines Lösungsarchitekten oder Dateningenieurs zurechtzufinden. Seine Fähigkeit, komplexe technische Konzepte in klare, leicht verständliche Erklärungen zu verwandeln, ist wirklich bemerkenswert. Dieses Buch ist eine unschätzbare Sammlung seiner Arbeit und dient als umfassendes Nachschlagewerk, um Architekturen zu entwerfen und nachzuvollziehen.

– Annie Xu, Senior Data Customer Engineer, Google

James hatte schon immer die Superkraft, komplexe Themen aufzugreifen und sie auf einfache Weise zu erklären. In diesem Buch trifft er alle wichtigen Punkte, um Ihnen zu helfen, die richtige Datenarchitektur auszuwählen und häufige (und teure!) Fehler zu vermeiden.

– Rod Colledge, Senior Technical Specialist (Daten und KI), Microsoft

Dieses Buch ist ein großer Meilenstein in der Entwicklung unseres Umgangs mit Daten in der Technologiebranche, und zwar über mehrere Jahrzehnte hinweg, was für die meisten wohl einer ganzen Karriere entspricht. Der Inhalt bietet großartige Einblicke für die nächste Generation von Datenexperten in Bezug darauf, was sie bei der Entwicklung zukünftiger Lösungen bedenken müssen.

– Paul Andrew, CTO, Cloud Formations Consulting

Ein fantastischer Leitfaden für Datenarchitekten, dieses Buch ist vollgepackt mit Erfahrungen und Einsichten. Die umfassende Abdeckung sich entwickelnder Trends und verschiedene Ansätze machen es zu einem unverzichtbaren Nachschlagewerk für alle, die ihr Verständnis des Fachgebiets erweitern möchten.

– Simon Whiteley, CTO, Advancing Analytics Limited

Es gibt niemanden, dem ich mit seinem Wissen über Datenarchitekturen und Datenprozesse mehr vertraue als James Serra. Dieses Buch bietet nicht nur eine umfassende und klare Beschreibung der wichtigsten architektonischen Prinzipien, Ansätze und Fallstricke, sondern befasst sich auch mit den überaus wichtigen menschlichen, kulturellen und organisatorischen Problemen, die Datenprojekte allzu oft gefährden, bevor sie in Gang kommen. Dieses Buch ist dazu prädestiniert, ein Grundlagenwerk für die Branche zu werden, und zwar sowohl für Hochschulstudentinnen und -studenten als auch für Geschäftsleute, die zum ersten Mal mit Daten in Berührung kommen (und vielleicht auch zum zweiten und dritten Mal)!

– Wayne Eckerson, Präsident der Eckerson Group

Das Buch *Datenarchitekturen* ist ein unverzichtbarer herstellerneutraler Leitfaden für die Datenexperten von heute. Es vergleicht aufschlussreich historische und moderne Architekturen. Hervorgehoben werden dabei die wichtigsten Kompromisse und Nuancen der Entscheidungsfindung bei der Auswahl einer geeigneten Architektur für die sich entwickelnde datengesteuerte Landschaft.

– Stacia Varga, Autorin und Data Analytics Consultant, Data Inspirations

Mit tiefgehender Praxiserfahrung ausgestattet, haben die neuesten Szenarien auf dem heutigen Markt eine anbieterspezifische Ausrichtung, inklusive Terminologie und Verkaufsoptionen. James nutzt seine jahrelange Expertise, um anbieterunabhängige, Cloud-übergreifende und branchenübergreifende Ansätze für kleine bis große Unternehmen zu zeigen.

– Jordan Martz, Senior Sales Engineer, Fivetran

Data Lake, Data Lakehouse, Data Fabric, Data Mesh ... Es ist nicht einfach, die Spreu vom Weizen zu trennen. Das Wissen und die Erfahrung von James Serra sind eine großartige Ressource für alle, denen Verantwortung für die Datenarchitektur übertragen wurde.

– Dave Wells, Industry Analyst, eLearningcurve

Zu oft findet man in Büchern Anleitungen ohne Hintergrund oder Logik – dieses Buch löst dies. Mit einem umfassenden Überblick darüber, warum Daten auf eine bestimmte Art und Weise angeordnet sind, erfahren Sie mehr über den richtigen Weg, das »Wie« zu implementieren.

– Buck Woody, Principal Data Scientist, Microsoft

Datenarchitekturen ist nicht nur gründlich und detailliert, sondern bietet auch eine kritische Perspektive auf das, was funktioniert, und – was vielleicht noch wichtiger ist – auf das, was vielleicht nicht gut funktioniert. Egal ob ältere Datenansätze oder neuere wie Data Mesh diskutiert werden, das Buch bietet Weisheiten und Erkenntnisse, die jedem Datenpraktiker helfen, seine Datenreise zu beschleunigen.

– Eric Broda, Unternehmer, Data Consultant, Autor von »Implementing Data Mesh« (O'Reilly)

Kein anderes Buch, das ich kenne, erklärt so umfassend Data Lake, Warehouse, Mesh, Fabric und Lakehouse! Es ist Pflichtlektüre für alle Datenarchitektinnen und -ingenieure.

– Vincent Rainardi, Datenarchitekt und Autor

Zum liebevollen Andenken an meine Großeltern – Dolly, Bill, Martha und Bert

Datenarchitekturen

Modern Data Warehouse, Data Fabric, Data Lakehouse und Data Mesh richtig einsetzen

James Serra

Übersetzung von Frank Langenau

O'REILLY®

James Serra

Lektorat: Alexandra Follenius
Übersetzung: Frank Langenau
Copy-Editing: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner
Umschlaggestaltung: Karen Montgomery, Michael Oréal, *www.oreal.de*

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über *http://dnb.d-nb.de* abrufbar.

ISBN:
Print 978-3-96009-254-4
PDF 978-3-96010-874-0
ePub 978-3-96010-875-7

1. Auflage 2025

Wieblinger Weg 17
69123 Heidelberg

Authorized German translation of the English edition of *Deciphering Data Architectures*, ISBN 9781098150761 © 2024 James Serra. This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«.
O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wissen: *kommentar@oreilly.de*.

Inhalt

Teil II: Allgemeine Datenarchitekturkonzepte

Teil IV: Menschen, Prozesse und Technologien

Vorwort

Noch nie in der Geschichte des modernen, technologiegestützten Unternehmens hat sich die Datenlandschaft so schnell entwickelt. Da sich das Tempo des Wandels weiter beschleunigt, wird das Datenökosystem immer komplexer – insbesondere die Wertschöpfungsketten, die Zulieferer und Kunden mit dem Unternehmen verbinden. Daten scheinen überall zu fließen. Sie sind zu einem der strategischsten Vermögenswerte eines jeden Unternehmens geworden und bilden die Grundlage für die digitale Transformation, Automatisierung, künstliche Intelligenz, Innovation und vieles mehr.

Dieses zunehmende Tempo des Wandels macht es umso wichtiger, die Datenarchitektur Ihres Unternehmens zu optimieren, um kontinuierliche Anpassungsfähigkeit, Interoperabilität und Wartungsfreundlichkeit zu gewährleisten. In diesem Buch stellt James Serra eine Reihe klarer Entscheidungen für den Datenarchitekten vor, unabhängig davon, ob Sie ein belastbares Design erstellen oder einfach die technische Schuld reduzieren möchten.

Scheinbar jeder Wissensarbeiter kennt eine Geschichte eines Meetings mit Dateningenieuren und -architekten, das sich wie ein Dilbert-Cartoon anfühlte, in dem niemand die gleiche Sprache zu sprechen schien und die Entscheidungen zu kompliziert und undurchsichtig waren. Indem er Konzepte definiert, Bedenken ausräumt, Mythen aufklärt und Workarounds für Fallstricke vorschlägt, vermittelt James den Leserinnen und Lesern brauchbare Kenntnisse über Datenarchitekturen und das nötige Vertrauen, um fundierte Entscheidungen zu treffen. Als Führungskraft bin ich sehr zufrieden damit, wie gut die Inhalte des Buchs dazu beitragen, die Ausrichtung meines Datenteams zu stärken, indem sie ihnen ein gemeinsames Vokabular und Referenzen an die Hand geben.

Als ich James vor etwa 15 Jahren kennenlernte, überlegte er, sein Fachwissen über die Datenbankadministration hinaus auf Business Intelligence und Analytik auszuweiten. Sein unstillbares Verlangen, Neues zu lernen und das Gelernte mit anderen zu teilen, um dem Allgemeinwohl zu dienen, hat mich stark beein-

druckt. Das treibt ihn auch heute noch an. Die unzähligen Blogposts, Präsentationen und Vorträge, in denen er seine tiefgreifenden Erfahrungen weitergibt, kulminieren nun in dieser umfassenden Ressource. Dieses Buch wird uns allen, die wir mit Daten umgehen, auf dem Weg in eine ungewisse Zukunft von Nutzen sein.

Wenn Sie die Grundprinzipien der Datenarchitektur verstehen, können Sie auf den Wellen des Wandels reiten, wenn neue Datenplattformen, Technologieprovider und Innovationen auftauchen. Dankenswerterweise hat James eine grundlegende Ressource für uns alle geschaffen. Dieses Buch wird Ihnen helfen, das große Ganze zu sehen und eine strahlende Zukunft zu gestalten, in der Ihre Daten den Wettbewerbsvorteil schaffen, den Sie suchen.

– Sean McCall, Chief Data Officer, Oceaneering International

Houston, Dezember 2023

Einführung

Seit fast 40 Jahren bin ich in der Informationstechnologie (IT) unterwegs. Ich habe in Unternehmen aller Größenordnungen gearbeitet, war als Berater tätig und habe mein eigenes Unternehmen gegründet. Ich arbeite seit neun Jahren als Datenarchitekt bei Microsoft, und in den letzten 15 Jahren habe ich mich mit Data Warehousing beschäftigt. Ich habe schon Tausende Male vor Kunden und Gruppen über Daten gesprochen.

Im Laufe meiner Karriere habe ich viele Datenarchitektinnen und -architekten kommen und gehen sehen. Ich habe zu viele Unternehmen gesehen, die sich über den besten Ansatz streiten und am Ende die falsche Datenarchitektur aufbauen – ein Fehler, der sie Millionen von Dollar und Monate an Zeit kosten kann und sie weit hinter ihre Konkurrenten zurückwirft.

Hinzu kommt, dass Datenarchitekturen sehr komplex sind. Ich habe aus erster Hand erfahren, dass den meisten Menschen die damit verbundenen Konzepte nicht klar sind, sofern sie sie überhaupt kennen. Jeder scheint mit Begriffen wie *Data Mesh*, *Data Warehouse* und *Data Lakehouse* um sich zu werfen – aber wenn Sie zehn Menschen fragen, was ein Data Mesh ist, werden Sie elf verschiedene Antworten erhalten.

Wo soll man da überhaupt anfangen? Handelt es sich dabei nur um Schlagwörter mit viel Hype, aber wenig Substanz, oder sind es praktikable Ansätze? In der Theorie mögen sie toll klingen, aber wie praktisch sind sie? Was sind die Vor- und Nachteile der einzelnen Architekturen?

Keine der in diesem Buch besprochenen Architekturen ist »falsch«. Alle haben ihre Berechtigung, aber nur in bestimmten Anwendungsfällen. Es gibt keine Architektur, die für jede Situation geeignet ist. Daher geht es in diesem Buch nicht darum, Sie davon zu überzeugen, eine bestimmte Architektur den anderen vorzuziehen. Stattdessen erhalten Sie ehrliche Meinungen zu den Vor- und Nachteilen der einzelnen Architekturen. Bei allen gibt es Kompromisse, und es ist wichtig, diese zu verstehen und sich nicht einfach für eine Architektur zu

entscheiden, die mehr als die anderen angepriesen wird. Außerdem kann man von jeder Architektur viel lernen, auch wenn man sie nicht nutzt. Wenn Sie zum Beispiel verstehen, wie ein Data Mesh funktioniert, werden Sie über Datenbesitz nachdenken, ein Konzept, das sich auf jede Architektur anwenden lässt.

Dieses Buch bietet eine grundlegende Einführung in gängige Datenarchitekturkonzepte. Es gibt *so* viele Konzepte, dass es einschüchternd sein kann, sich für eines zu entscheiden und herauszufinden, wie es zu implementieren ist. Ich möchte Ihnen helfen, all diese Konzepte und Architekturen auf einem hohen Niveau zu verstehen, damit Sie ein Gefühl für die Optionen bekommen und erkennen können, welche für Ihre Situation am besten geeignet ist. Das Ziel des Buchs ist es, Ihnen zu ermöglichen, auf intelligente Weise über Datenkonzepte und -architekturen zu sprechen und sich dann mit denjenigen zu befassen, die für die Lösung, die Sie aufbauen, relevant sind.

Es gibt keine Standarddefinitionen von Datenkonzepten und Architekturen. Andernfalls wäre dieses Buch überflüssig. Meine Hoffnung ist, Standarddefinitionen zur Verfügung zu stellen, damit alle auf Augenhöhe miteinander diskutieren können. Allerdings mache ich mir keine Illusionen darüber, dass meine Definitionen allgemein akzeptiert werden, aber ich möchte uns allen einen Ausgangspunkt für Gespräche darüber geben, wie wir diese Definitionen anpassen können.

Ich habe dieses Buch für jeden geschrieben, der daran interessiert ist, Nutzen aus Daten zu ziehen, egal ob Sie Datenbankentwickler oder -administratorin, Datenarchitekt, CTO oder CIO oder sogar jemand in einer Rolle außerhalb der IT sind. Sie können am Anfang Ihrer Karriere stehen oder ein erfahrener Veteran sein. Die einzigen Fähigkeiten, die Sie benötigen, sind ein wenig Vertrautheit mit Daten aus Ihrer Arbeit und ein gewisses Maß an Neugierde.

Für Leserinnen und Leser, die weniger Erfahrung mit diesen Themen haben, biete ich einen Überblick über Big Data (Kapitel 1) und Datenarchitekturen (Kapitel 2) sowie über grundlegende Datenkonzepte (Teil II). Wenn Sie schon eine Weile mit Daten zu tun haben, aber neue Architekturen verstehen müssen, könnte Ihnen Teil III von großem Nutzen sein, in dem ich auf die Details bestimmter Datenarchitekturen eingehe und einige der Grundlagen wiederhole. Diesen Teil können Sie auch überfliegen – überspringen Sie einfach die Abschnitte mit dem Material, das Sie bereits gut kennen. Der Schwerpunkt liegt zwar auf Big Data, doch gelten die Konzepte und Architekturen auch für »kleine« Daten.

Dies ist ein anbieterneutrales Buch. Die Architekturen und Konzepte, die Sie hier lernen, sollten Sie also auch bei jedem Cloud-Provider anwenden können.

An dieser Stelle möchte ich erwähnen, dass ich bei Microsoft beschäftigt bin. Die hier geäußerten Meinungen sind jedoch allein meine und spiegeln nicht die Ansichten meines Arbeitgebers wider.

Ich habe dieses Buch geschrieben, weil ich eine angeborene Neugier habe, die mich dazu antreibt, Dingen auf den Grund zu gehen und sie dann so weiterzugeben, dass sie jeder verstehen kann. Dieses Buch ist die Summe meines bisherigen Schaffens. Ich hoffe, es wird ein wertvolles Arbeitsmittel für Sie.

Konventionen in diesem Buch

In diesem Buch werden die folgende typografische Konventionen verwendet:

Kursiv
: Kennzeichnet neue Begriffe, URLs, E-Mail-Adressen, Dateinamen und Dateierweiterungen.

`Schreibmaschinenschrift`
: Wird in Programmlistings verwendet und im Fließtext für Programmelemente wie zum Beispiel Variablen- oder Funktionsnamen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter.

Dieses Element kennzeichnet einen Tipp oder Vorschlag.

Danksagungen

Dieses Buch wäre ohne die unermüdliche Unterstützung und Geduld meiner Frau Mary nicht möglich gewesen. Ihre Ermutigung war in den langen Nächten, die ich mit dem Schreiben verbracht habe, ein entscheidender Antrieb, selbst wenn das bedeutet hat, dass ich auf Kartenspiele mit Familie und Freunden verzichten musste. Ihre Anwesenheit war eine ständige Quelle des Trostes und der Motivation.

Meine Reise wurde durch die Unterstützung meiner Familie bereichert: meine Eltern Jim und Lorraine, meine Schwestern Denise, Michele und Nicole und meine inzwischen erwachsenen Kinder Lauren, RaeAnn und James, die eine Quelle der Inspiration waren, obwohl sie keine Ahnung hatten, worum es in dem Buch geht!

Ein herzliches Dankeschön geht an meine Mentoren und Kollegen aus meinen Jahren bei Microsoft. Menschen wie Steve Busby, Eduardo Kassner und Martin Lee haben meine Karriere geprägt und mir Ratschläge mit auf den Weg gegeben, die an vielen Stellen in dieses Buch eingeflossen sind.

Mein Dank gilt auch denjenigen, die mir ihr kritisches Auge und konstruktives Feedback geschenkt haben, insbesondere Piethein Strengholt, Barry Devlin, Bill Inmon und Mike Shelton. Ihre Einblicke waren unschätzbar.

Besonders dankbar bin ich Sean McCall, der mich vor vielen Jahren in die Welt des Data Warehousing eingeführt hat, aber auch ein treuer Freund ist und sich bereit erklärt hat, das Vorwort zu diesem Buch zu schreiben.

Schließlich möchte ich all den großartigen Menschen bei O'Reilly danken, die dieses Buch möglich gemacht haben: Sarah Grey, deren großartiges Lektorat und deren Vorschläge dieses Buch so viel besser gemacht haben, als wenn ich nur auf mich selbst gestellt gewesen wäre, Aaron Black, der mir geholfen hat, den Abstract zum Buch zu erstellen und ihn bestätigen zu lassen, Paula Fleming für ihr außergewöhnliches Copyediting, Katie Tozer, die die Herstellung des Buchs geleitet hat, Kristen Brown, die dafür gesorgt hat, dass alles reibungslos gelaufen ist, und Suzanne Huston für ihre wunderbare Vermarktung des Buchs.

Ich möchte Ihnen, den Leserinnen und Lesern, meine tiefe Dankbarkeit aussprechen, denn Ihr Interesse und Ihr Engagement für dieses Werk machen die unzähligen Stunden, die ich mit dem Schreiben zugebracht habe, nicht nur lohnenswert, sondern zutiefst erfüllend.

Während ich dieses Kapitel meines Lebens abschließe und mich auf die neuen Horizonte freue, die vor mir liegen, bin ich zutiefst dankbar für die Reise, auf die mich dieses Buch geführt hat, und für die unglaublichen Menschen, die daran teilgenommen haben.

TEIL I
Grundlagen

In Teil I dieses Buchs geht es um die Grundlagen für die Entschlüsselung von Datenarchitekturen. Zunächst beschreibt Kapitel 1, was man unter Big Data versteht, während Kapitel 2 einen Überblick über die Arten von Datenarchitekturen und deren Entwicklung gibt. Kapitel 3 zeigt, wie Sie eine Architektur-Design-Sitzung (*Architecture Design Session*) durchführen können, um die beste Datenarchitektur für Ihr Projekt zu bestimmen.

Dieser Teil des Buchs bietet Ihnen einen guten Ausgangspunkt, um den Wert von Big Data und die Geschichte der Architekturen, die diese Daten erfasst haben, zu verstehen. In den späteren Kapiteln befassen wir uns dann ausgiebiger mit den Details.

KAPITEL 1

Big Data

Die Anzahl der Firmen, die Datenarchitekturen erstellen, ist in den 2020er-Jahren sprunghaft gestiegen. Es ist unwahrscheinlich, dass sich dieses Wachstum in absehbarer Zeit verlangsamt, vor allem weil mehr Daten als je zuvor zur Verfügung stehen: angefangen bei sozialen Medien über IoT-Geräte (Internet der Dinge) bis hin zu selbst entwickelten Anwendungen und Software von Drittanbietern, um nur einige Quellen zu nennen. Laut einer BCG-Studie aus dem Jahr 2023 (*https://oreil.ly/hpOPt*) »hat sich der Umfang der generierten Daten von 2019 bis 2021 auf etwa 84 ZB ungefähr verdoppelt, wobei zu erwarten ist, dass es mit dieser Wachstumsrate weitergeht«. Die Forscher »schätzen, dass der Umfang der generierten Daten mit einer jährlichen Wachstumsrate (*Compound Annual Growth Rate*, CAGR) bei 21 % von 2021 bis 2024 auf 149 ZB ansteigen wird. Die Unternehmen wissen, dass sie Millionen Dollar sparen und den Umsatz erhöhen können, indem sie diese Daten sammeln und anhand der Vergangenheits- und Gegenwartsdaten Vorhersagen über die Zukunft treffen – doch um das zu tun, brauchen sie eine Möglichkeit, um alle diese Daten zu speichern.

Überall in der Geschäftswelt wird versucht, so schnell wie möglich Datenarchitekturen aufzubauen. Diese Architekturen müssen auch in der Lage sein, zukünftig zu erfassende Daten – unabhängig von ihrer Größe, Geschwindigkeit oder Art – zu verarbeiten und ihre Genauigkeit zu gewährleisten. Und diejenigen von uns, die mit Datenarchitekturen arbeiten, müssen genau wissen, wie sie funktionieren und welche Möglichkeiten sie bieten. Genau hier setzt dieses Buch an. Ich habe aus erster Hand erfahren, was passiert, wenn man die Konzepte der Datenarchitektur nicht richtig versteht. Ein mir bekanntes Unternehmen hat in zwei Jahren eine Datenarchitektur für 100 Millionen Dollar aufgebaut, nur um dann festzustellen, dass die Architektur die falsche Technologie verwendet hat, zu schwierig in der Anwendung und nicht flexibel genug war, um bestimmte Datentypen zu verarbeiten. Sie musste verworfen und von Grund auf neu aufgebaut werden. Lassen Sie nicht zu, dass Ihnen das passiert!

Es geht darum, die richtigen Informationen zur richtigen Zeit und im richtigen Format an die richtigen Personen weiterzugeben. Dazu benötigen Sie eine Datenstruktur, mit der Sie die Daten erfassen, speichern, umwandeln und modellieren können (Big-Data-Verarbeitung), damit sie präzise und einfach genutzt werden können. Sie benötigen eine Architektur, die es jedem Endbenutzer, selbst einem mit sehr geringem technischem Wissen, ermöglicht, die Daten zu analysieren und Berichte und Dashboards zu erstellen, anstatt sich darauf zu verlassen, dass IT-Mitarbeiter mit profundem technischem Wissen dies für sie tun.

Kapitel 1 führt in Big Data und einige seiner grundlegenden Ideen ein. Anschließend erörtere ich, wie Unternehmen ihre Daten nutzen, wobei der Schwerpunkt auf Business Intelligence liegt, und wie diese Nutzung zunimmt, wenn die Datenstruktur eines Unternehmens reift.

Was ist Big Data, und wie kann Big Data Ihnen helfen?

Auch wenn *Big Data* das Adjektiv *big* (groß) enthält, geht es nicht nur um die Größe der Daten. Vor allem geht es um alle Daten, egal ob groß oder klein, die in Ihrem Unternehmen existieren, sowie alle Daten außerhalb Ihres Unternehmens, die für Sie hilfreich sein könnten. Die Daten können in jedem Format vorliegen und mit beliebiger Regelmäßigkeit gesammelt werden. Um Big Data zu definieren, betrachtet man sie am besten als die Daten *in ihrer Gesamtheit*, unabhängig von ihrer Größe (*Volume*), Geschwindigkeit (*Velocity*) oder Vielfalt (*Variety*). Neben diesen Kriterien gibt es drei weitere Faktoren, mit denen Sie Daten beschreiben können: Wahrhaftigkeit (*Veracity*), Variabilität (*Variability*) und Wert (*Value*). Nach den Anfangsbuchstaben der englischen Bezeichnungen sind sie allgemein als »die sechs Vs« von Big Data bekannt, wie Abbildung 1-1 zeigt.

Sehen wir uns jedes einzelne V genauer an:

Volume (Datenvolumen)
: Das Datenvolumen ist die schiere Menge der erzeugten und gespeicherten Daten. Das Volumen kann von Terabyte bis Petabyte reichen, und die Daten können aus einer Vielzahl von Quellen stammen, darunter soziale Medien, E-Commerce-Transaktionen, wissenschaftliche Experimente, Sensordaten von IoT-Geräten und viele mehr. Beispielsweise können die Daten von einem Auftragseingabesystem pro Tag mehrere Terabyte ausmachen, während IoT-Geräte Millionen von Ereignissen pro Minute streamen und Hunderte von Terabytes an Daten pro Tag erzeugen können.

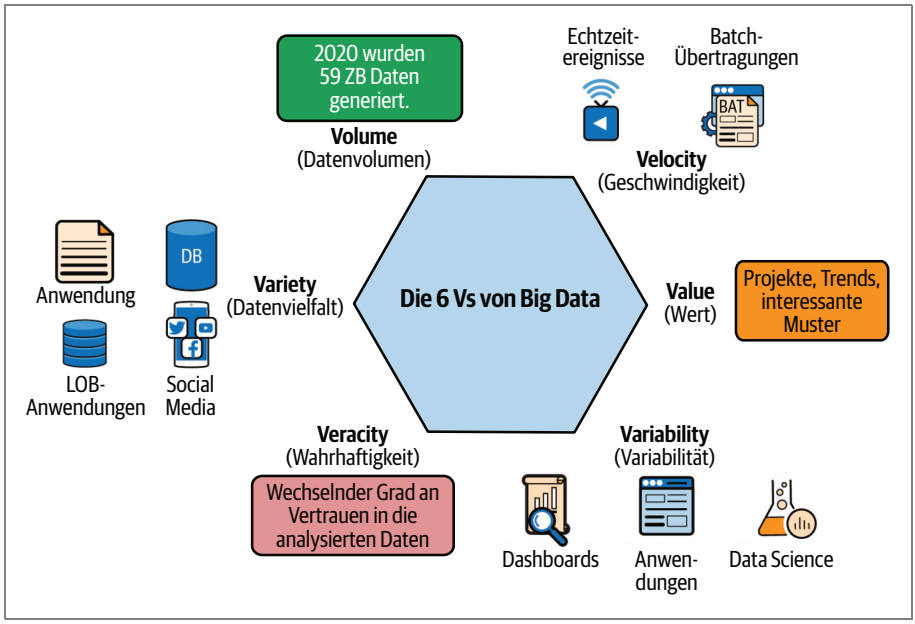

Abbildung 1-1: Die sechs Vs von Big Data (Quelle: The Cloud Data Lake von Rukmani Gopalan [O'Reilly, 2023])

Variety (Datenvielfalt)

Die *Datenvielfalt* bezieht sich auf das breite Spektrum an Datenquellen und -formaten. Diese lassen sich weiter unterteilen in *strukturierte Daten* (aus relationalen Datenbanken), *teilstrukturierte Daten* (wie zum Beispiel Protokolle und Daten in den Formaten CSV, XML und JSON), *unstrukturierte Daten* (wie E-Mails, Dokumente und PDFs) und *binäre Daten* (Bilder, Audio, Video). Zum Beispiel wären Daten aus einem Auftragseingabesystem strukturierte Daten, da sie aus einer relationalen Datenbank stammen, während Daten von einem IoT-Gerät wahrscheinlich im JSON-Format vorliegen.

Velocity (Geschwindigkeit)

Die *Geschwindigkeit* gibt an, wie schnell Daten erzeugt und verarbeitet werden. Wenn Daten eher selten erfasst werden, spricht man oft von *Stapelverarbeitung* (*Batch Processing*). Zum Beispiel könnten die tagsüber eingegangenen Bestellungen jede Nacht zusammengefasst und verarbeitet werden. Es ist aber auch üblich, dass Daten sehr häufig oder sogar in Echtzeit erfasst werden, insbesondere wenn sie mit hoher Geschwindigkeit entstehen, wie es beispielsweise bei Daten von sozialen Medien, IoT-Geräten und mobilen Anwendungen der Fall ist.

Veracity (Wahrhaftigkeit)

Mit *Wahrhaftigkeit* sind Genauigkeit und Zuverlässigkeit der Daten gemeint. Die Quellen für Big Data könnten unterschiedlicher nicht sein. Unzuverlässige oder unvollständige Daten beeinträchtigen gegebenenfalls die Qualität der Daten. Wenn die Daten zum Beispiel von einem IoT-Gerät kommen, etwa von einer Sicherheitskamera vor Ihrem Haus, die auf die Einfahrt gerichtet ist, und die Ihnen eine Textnachricht sendet, wenn eine Person erkannt wird, ist es durchaus möglich, dass Umgebungseinflüsse wie zum Beispiel das Wetter dazu führen, dass eine Person statt einer Katze erkannt wird, und das Überwachungsgerät somit verfälschte Daten sendet. Daher ist es unumgänglich, die Daten zu validieren, sobald sie empfangen werden.

Variability (Variabilität)

Variabilität meint die Konsistenz (oder Inkonsistenz) von Daten hinsichtlich ihres Formats, ihrer Qualität und ihrer Bedeutung. Strukturierte, teilstrukturierte und unstrukturierte Datenformate zu verarbeiten, verlangt verschiedene Tools und Techniken. So können beispielsweise Art, Häufigkeit und Qualität der Sensordaten von IoT-Geräten sehr unterschiedlich sein. Temperatur- und Luftfeuchtigkeitssensoren können Datenpunkte in regelmäßigen Intervallen erzeugen, während Bewegungssensoren möglicherweise nur dann Daten liefern, wenn sie eine Bewegung erkennen.

Value (Wert)

Das wichtigste V steht für *Value*, d.h. den Wert, der sich auf die Nützlichkeit und Relevanz der Daten bezieht. Unternehmen nutzen Big Data, um Erkenntnisse zu gewinnen und Entscheidungen zu treffen, die zu einem geschäftlichen Nutzen führen können, zum Beispiel zu höherer Effizienz, zu Kosteneinsparungen oder zu neuen Einnahmequellen. So können Unternehmen das Verhalten, die Vorlieben und die Bedürfnisse ihrer Kunden besser verstehen, indem sie die Kundendaten analysieren. Anhand dieser Informationen sind sie in der Lage, zielgerichtete Marketingkampagnen zu entwickeln, die Kundenzufriedenheit zu verbessern und den Umsatz zu steigern.

Mithilfe von Big Data können Unternehmen Erkenntnisse gewinnen, die ihnen helfen, bessere Geschäftsentscheidungen zu treffen. Die *prädiktive Analyse* ist eine Art der Datenanalyse, die statistische Algorithmen und Machine Learning einbezieht, um historische Daten zu analysieren und Vorhersagen über zukünftige Ereignisse und Trends zu treffen. Dadurch können Unternehmen proaktiv und nicht nur reaktiv handeln.