Handbuch Data Engineering

Robuste Datensysteme planen und erstellen

DAS VORWORT

Vorwort

Wie ist dieses Buch entstanden? Der Ursprung ist tief verwurzelt in unserer eigenen Entwicklung von Data Science zu Data Engineering. Wir bezeichnen uns oft scherzhaft als *genesende Data Scientists*. Wir haben beide die Erfahrung gemacht, dass wir mit Data-Science-Projekten betraut wurden und dann Schwierigkeiten hatten, diese Projekte auszuführen, weil uns die notwendigen Grundlagen fehlten. Unsere Entwicklung in Richtung Data Engineering begann, als wir Aufgaben aus dem Bereich Data Engineering übernahmen, um die Grundlagen und die Infrastruktur aufzubauen.

Mit dem Vormarsch von Data Science gaben Unternehmen viel Geld für Data Scientists aus in der Hoffnung, davon reichlich profitieren zu können. Sehr oft jedoch hatten Data Scientists mit Problemen zu kämpfen, für deren Lösung ihr Vorwissen und ihre Ausbildung nicht ausreichten – Datenerfassung, Datenbereinigung, Datenzugriff, Datentransformation und Dateninfrastruktur. Dies sind Probleme, die mit Data Engineering gelöst werden sollen.

Was dieses Buch nicht ist

Bevor wir darauf eingehen, worum es in diesem Buch geht und wie Sie vom Inhalt profitieren werden, lassen Sie uns kurz darauf eingehen, was dieses Buch *nicht ist*. Dieses Buch handelt nicht von Data Engineering unter Verwendung eines bestimmten Tools, einer Technologie oder einer Plattform. Es gibt zwar viele hervorragende Bücher, die sich aus dieser Perspektive mit Technologien des Data Engineering befassen, diese Bücher haben jedoch oft eine kurze Lebensdauer. Wir fokussieren uns stattdessen auf die grundlegenden Konzepte des Data Engineering.

Worum es in diesem Buch geht

Dieses Buch soll eine bestehende Wissenslücke in den vorhandenen Inhalten und Materialien zum Thema Data Engineering schließen. Obwohl es keinen Mangel an

technischen Informationsquellen gibt, die sich mit spezifischen Tools und Technologien des Data Engineering befassen, ist es schwierig zu verstehen, wie diese Komponenten zu einem schlüssigen Ganzen zusammengefügt werden können, das in der Praxis funktioniert. Dieses Buch beschreibt die einzelnen Schritte des gesamten Datenprozesses. Es zeigt Ihnen, wie Sie verschiedene Technologien kombinieren können, um die Anforderungen der nachgelagerten Datennutzer wie Analysten, Data Scientists und Machine-Learning-Engineers zu erfüllen. Dieses Buch dient als Ergänzung zu den Büchern von O'Reilly, die sich mit Details bestimmter Technologien, Plattformen und Programmiersprachen befassen.

Die Grundidee dieses Buchs ist der *Data Engineering Lifecycle*: Datengenerierung (*Data Generation*), Datenspeicherung (*Data Storage*), Dateningestion (*Data Ingestion*), Datentransformation (*Data Transformation*) und Datenbereitstellung (*Data Serving*). Seit den Anfängen der Datenverarbeitung haben wir den Aufstieg und Fall unzähliger spezifischer Technologien und Anbieterprodukte erlebt, aber die Phasen des Data Engineering Lifecycle sind im Wesentlichen unverändert geblieben. Mit unserem Rahmenwerk erhalten die Leserinnen und Leser ein fundiertes Grundwissen über die Anwendung von Technologien auf praktische Probleme.

Unser Ziel ist es, Richtlinien zu formulieren, die sich über zwei Achsen erstrecken. Erstens wollen wir Data Engineering in allgemeine Grundsätze fassen, die *alle relevanten Technologien* umfassen. Zweitens wollen wir Prinzipien vorstellen, die *langfristig* Bestand haben werden. Wir hoffen, dass unsere Überlegungen die Lektionen widerspiegeln, die wir in den letzten 20 Jahren im Bereich der Informationstechnologie gelernt haben, und dass unser konzeptueller Rahmen auch noch in zehn oder mehr Jahren nützlich sein wird.

Noch eine Anmerkung vorweg: Wir verfolgen ganz klar die Cloud-first-Strategie. Wir sehen die Cloud als eine revolutionäre Entwicklung, die jahrzehntelang Bestand haben wird; die meisten lokalen Datensysteme und Arbeitslasten werden letztendlich in die Cloud verlagert. Wir gehen davon aus, dass Infrastrukturen und Systeme *kurzlebig* und *skalierbar* sind und dass Data Engineers dazu übergehen werden, verwaltete Dienste in der Cloud bereitzustellen. Dennoch lassen sich die meisten Konzepte in diesem Buch auch auf nicht cloudbasierte Umgebungen übertragen.

Für wen ist dieses Buch gedacht?

Unser primäres Zielpublikum sind technische Anwender, Softwareingenieurinnen und -ingenieure auf mittlerer bis höherer Ebene, Data Scientists sowie Analystinnen und Analysten, die sich für Data Engineering interessieren, ebenso Data Engineers, die sich mit bestimmten Technologien auskennen, aber eine umfassendere Fachkompetenz entwickeln möchten. Unsere sekundäre Zielgruppe sind Stakeholder aus dem Datenbereich, die neben technischen Fachleuten arbeiten, z.B. die

Managerin eines Datenteams mit technischem Hintergrund, die ein Team von Data Engineers leitet, oder der Direktor für Data Warehousing, der von einer lokalen Technologie zu einer cloudbasierten Lösung migrieren möchte.

Im Idealfall sind Sie neugierig und wollen lernen – weshalb sonst würden Sie dieses Buch lesen? Sie halten sich über die neuesten Technologien und Trends im Umgang mit Daten auf dem Laufenden, indem Sie Bücher und Artikel über Data Warehousing/Data Lakes, Batch- und Streaming-Systeme, Orchestrierung, Modellierung, Management, Analyse, Entwicklungen bei Cloud-Technologien usw. lesen. Dieses Buch wird Ihnen helfen, das Gelesene zu einem vollständigen Bild des Data Engineering über Technologien und Paradigmen hinweg zu verweben.

Voraussetzungen

Wir gehen davon aus, dass die Leserschaft mit den gängigen Datensystemen in Unternehmen vertraut ist. Darüber hinaus setzen wir voraus, dass sie mit SQL und Python (oder einer anderen Programmiersprache) einigermaßen vertraut ist und Erfahrung mit Cloud-Diensten hat.

Für angehende Data Engineers gibt es zahlreiche Ressourcen zum Erlernen von Python und SQL. Kostenlose Onlineressourcen gibt es im Überfluss (Blogbeiträge, Tutorials, YouTube-Videos), und jedes Jahr werden viele neue Python-Bücher veröffentlicht.

Die Cloud bietet beispiellose Möglichkeiten, praktische Erfahrungen mit Datentools zu sammeln. Wir empfehlen angehenden Data Engineers, Konten bei Cloud-Diensten wie AWS, Azure, Google Cloud Platform, Snowflake, Databricks, usw. einzurichten. Zwar bieten viele dieser Plattformen *kostenlose Tier-Optionen* an, es ist jedoch ratsam, sich die Kosten genau anzusehen und zu Beginn mit kleinen Datenmengen und Einzelknotenclustern zu arbeiten.

Es ist nach wie vor schwierig, sich außerhalb einer betrieblichen Umgebung mit den Datensystemen von Unternehmen vertraut zu machen, was für angehende Data Engineers, die ihren ersten Job im Bereich Datenverarbeitung noch nicht gefunden haben, gewisse Hindernisse darstellt. Hier kann dieses Buch weiterhelfen. Wir schlagen vor, dass Einsteiger in die Datenverarbeitung das Buch lesen, um sich einen Überblick über die wichtigsten Ideen zu verschaffen, und sich dann die Materialien in den Abschnitten »Weitere Quellen« jeweils am Schluss eines Kapitels ansehen. Notieren Sie sich beim erneuten Durchlesen alle unbekannten Begriffe und Technologien. Außerdem könnten Sie Google, Wikipedia, Blogbeiträge, YouTube-Videos und Webseiten von Anbietern nutzen, um sich mit neuen Begriffen vertraut zu machen und Wissenslücken zu schließen.

Was Sie lernen werden und wie Sie Ihre Kenntnisse erweitern können

Dieses Buch soll Ihnen helfen, eine fundierte Grundlage für die Lösung praxisrelevanter Probleme beim Data Engineering zu erlangen.

Nach der Lektüre dieses Buchs werden Sie Folgendes gelernt haben:

- wie sich Data Engineering auf Ihre aktuelle Position auswirkt (Data Scientist, Softwareentwickler oder Teamleiter),
- wie man den Marketing-Hype durchschaut und die richtigen Technologien, Datenstrukturen und Prozesse auswählt,
- wie man den Data Engineering Lifecycle nutzt, um eine stabile Infrastruktur zu entwerfen und aufzubauen, sowie
- bewährte Verfahren für jede Phase des Datenlebenszyklus.

Außerdem werden Sie in der Lage sein:

- Prinzipien des Data Engineering in Ihre gegenwärtige Tätigkeit zu integrieren (Data Scientist, Analyst, Softwareentwickler, Teamleiter usw.),
- eine Vielzahl von Cloud-Technologien zu kombinieren, um die Bedürfnisse der Datenkonsumenten zu erfüllen,
- Probleme des Data Engineering mit bewährten Verfahren zu bewerten und
- Datenschutz und -sicherheit in den gesamten Data Engineering Lifecycle einzubeziehen.

Wegweiser durch dieses Buch

Dieses Buch besteht aus vier Abschnitten:

- Teil I: »Grundlagen und Bausteine«
- Teil II: »Der Data Engineering Lifecycle im Detail«
- Teil III: »Sicherheit, Datenschutz und die Zukunft des Data Engineering«
- Anhang A und B: Serialisierung und Kompression bzw. Cloud-Networking

In Teil I definieren wir in Kapitel 1 Data Engineering und skizzieren in Kapitel 2 den Data Engineering Lifecycle. In Kapitel 3 stellen wir *gute Architektur* vor. Abschließend stellen wir in Kapitel 4 eine Orientierungshilfe für die Auswahl der richtigen Technologie vor – obwohl wir häufig sehen, dass Technologie und Infrastruktur in einen Topf geworfen werden, handelt es sich in Wirklichkeit um sehr unterschiedliche Themen.

Teil II baut auf Kapitel 2 auf, um den Data Engineering Lifecycle eingehender zu beleuchten; jede Phase des Lebenszyklus – Datengenerierung, Speicherung, Ingestion, Transformation und Bereitstellung – wird in einem eigenen Kapitel beschrie-

ben. Teil II ist zweifellos das Herzstück des Buchs, und die weiteren Kapitel dienen der Unterstützung der hier vorgestellten zentralen Themen.

Teil III beinhaltet ergänzende Themen. In Kapitel 10 befassen wir uns mit *Sicherheit und Datenschutz*. Sicherheit war zwar schon immer ein wichtiger Bestandteil im Data Engineering, ist aber mit dem Aufkommen kommerzieller Hackerangriffe und staatlich geförderter Cyberattacken nur noch wichtiger geworden. Und was können wir zum Datenschutz sagen? Die Zeit des Datenschutz-Nihilismus in Unternehmen ist vorbei – kein Unternehmen möchte seinen Namen als Schlagzeile über nachlässige Datenschutzpraktiken sehen. Der rücksichtslose Umgang mit personenbezogenen Daten kann seit der Einführung von GDPR, CCPA und anderen Vorschriften auch erhebliche rechtliche Folgen haben. Zusammenfassend lässt sich sagen, dass Sicherheit und Datenschutz bei jeder Tätigkeit im Data Engineering oberste Priorität haben müssen.

Im Laufe unserer Arbeit im Bereich Data Engineering, der Recherche für dieses Buch und der Befragung zahlreicher Experten haben wir uns viele Gedanken darüber gemacht, wie sich das Thema in den nächsten Jahren entwickeln wird. Kapitel 11 umreißt unsere äußerst spekulativen Vorstellungen von der Zukunft des Data Engineering. Die Zukunft ist nur schwer vorhersehbar, sie wird zeigen, ob einige unserer Ansichten zutreffend sind. Wir würden uns freuen, von unseren Lesern zu erfahren, inwieweit ihre Zukunftsvisionen mit unseren Vorstellungen übereinstimmen bzw. sich von ihnen unterscheiden.

In den Anhängen decken wir einige technische Themen ab, die für die tägliche Arbeit im Data Engineering äußerst wichtig sind, aber nicht in den Hauptteil des Buchs passten. Insbesondere müssen Data Engineers Serialisierung und Kompression verstehen (siehe Anhang A), sowohl um mit Dateien arbeiten zu können als auch um die Leistungsfähigkeit von Datensystemen beurteilen zu können. Die Vernetzung in der Cloud (siehe Anhang B) ist ein wichtiges Thema, da sich Data Engineering in die Cloud verlagert.

In diesem Buch verwenden wir folgende Konventionen

In diesem Buch werden die folgenden typografischen Konventionen angewandt:

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

`Feste Zeichenbreite`
: Wird für Programmlistings sowie innerhalb von Absätzen verwendet, um auf Programmelemente wie Variablen- oder Funktionsnamen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter zu verweisen.

Dieses Symbol kennzeichnet einen Tipp oder Vorschlag.

Dieses Symbol zeigt einen allgemeinen Hinweis an.

Dieses Symbol weist auf eine Warnung oder einen Warnhinweis hin.

Danksagung

Als wir mit der Arbeit an diesem Buch begannen, warnten uns viele Menschen, dass wir vor einer schwierigen Aufgabe stünden. Solch ein Buch besteht aus vielen einzelnen Teilen, und da es einen umfassenden Überblick über das Thema Data Engineering bieten will, waren eine Menge Recherchen, Interviews, Diskussionen und ein tiefes Eintauchen in die Materie erforderlich. Wir können nicht garantieren, dass wir alle Nuancen des Data Engineering erfasst haben, aber wir hoffen, dass die Ergebnisse bei Ihnen Anklang finden. Zahlreiche Personen haben uns geholfen, und wir sind dankbar für die Unterstützung, die wir von vielen Expertinnen und Experten erhalten haben.

Zunächst einmal möchten wir uns bei unseren wunderbaren technischen Reviewern bedanken. Sie haben sich durch viele Versionen gekämpft und unschätzbares (und oft schonungslos direktes) Feedback gegeben. Ohne ihre Bemühungen wäre dieses Buch nur ein Bruchteil dessen, was es ist. In keiner bestimmten Reihenfolge gebührt unser Dank Bill Inmon, Andy Petrella, Matt Sharp, Tod Hansmann, Chris Tabb, Danny Lebzyon, Martin Kleppman, Scott Lorimor, Nick Schrock, Lisa Steckman, Veronika Durgin und Alex Woolford.

Des Weiteren hatten wir die einmalige Gelegenheit, in unseren Liveshows, Podcasts, Treffen und endlosen Telefonaten mit den führenden Experten aus dem Bereich der Daten zu sprechen. Ihre Ideen prägen unser Buch. Es sind zu viele, um sie einzeln zu nennen, daher gilt stellvertretend unser Dank Jordan Tigani, Zhamak Dehghani, Ananth Packkildurai, Shruti Bhat, Eric Tschetter, Benn Stancil, Kevin Hu, Michael Rogove, Ryan Wright, Adi Polak, Shinji Kim, Andreas Kretz, Egor Gryaznov, Chad Sanderson, Julie Price, Matt Turck, Monica Rogati, Mars Lan, Pardhu Gunnam, Brian Suk, Barr Moses, Lior Gavish, Bruno Aziza, Gian Merlino, DeVaris Brown, Todd Beauchene, Tudor Girba, Scott Taylor, Ori Rafael, Lee Edwards, Bryan Offutt, Ollie Hughes, Gilbert Eijkelenboom, Chris Bergh, Fabiana

Clemente, Ori Reshef, Nick Singh, Mark Balkenende, Kenten Danas, Brian Olsen, Rhaghu Murthy, Greg Coquillo, David Aponte, Demetrios Brinkmann, Sarah Catanzaro, Michel Tricot, Levi Davis, Ted Walker, Carlos Kemeny, Josh Benamram, Chanin Nantasenamat, George Firican, Jordan Goldmeir, Minhaaj Rehmam, Luigi Patruno, Vin Vashista, Danny Ma, Jesse Anderson, Alessya Visnjic, Vishal Singh, Dave Langer, Roy Hasson, Todd Odess, Che Sharma, Scott Breitenother, Ben Taylor, Thom Ives, John Thompson, Brent Dykes, Josh Tobin, Mark Kosiba, Tyler Pugliese, Douwe Maan, Martin Traverso, Curtis Kowalski, Bob Davis, Koo Ping Shung, Ed Chenard, Matt Sciorma, Tyler Folkman, Jeff Baird, Tejas Manohar, Paul Singman, Kevin Stumpf, Willem Pineaar und Michael Del Balso von Tecton, Emma Dahl, Harpreet Sahota, Ken Jee, Scott Taylor, Kate Strachnyi, Kristen Kehrer, Taylor Miller, Abe Gong, Ben Castleton, Ben Rogojan, David Mertz, Emmanuel Raj, Andrew Jones, Avery Smith, Brock Cooper, Jeff Larson, Jon King, Holden Ackerman, Miriah Peterson, Felipe Hoffa, David Gonzalez, Richard Wellman, Susan Walsh, Ravit Jain, Lauren Balik, Mikiko Bazeley, Mark Freeman, Mike Wimmer, Alexey Shchedrin, Mary Clair Thompson, Julie Burroughs, Jason Pedley, Freddy Drennan, Jason Pedley, Kelly und Matt Phillipps, Brian Campbell, Faris Chebib, Dylan Gregerson, Ken Myers, Jake Carter, Seth Paul, Ethan Aaron und vielen mehr.

Wenn du nicht ausdrücklich erwähnt wurdest, nimm es nicht persönlich. Du weißt, wer du bist. Lass es uns wissen, und wir erwähnen dich in der nächsten Ausgabe.

Außerdem möchten wir uns beim Ternary Data Team (Colleen McAuley, Maike Wells, Patrick Dahl, Aaron Hunsaker und anderen), unseren Studierenden und den unzähligen Menschen weltweit bedanken, die uns unterstützt haben.

Die Zusammenarbeit mit O'Reilly war fantastisch! Besonderer Dank gilt Jess Haberman für sein Vertrauen, als wir unsere Idee für dieses Buchs vorstellten, sowie unseren großartigen und äußerst geduldigen Lektorinnen Nicole Taché und Michele Cronin für ihr unschätzbares Lektorat, ihr Feedback und ihre Unterstützung. Außerdem danken wir dem exzellenten Produktionsteam von O'Reilly (Greg und sein Team).

Joe ist seiner Familie – Cassie, Milo und Ethan – dankbar, dass sie ihm erlaubt hat, ein Buch zu schreiben. Sie mussten eine Menge durchmachen, und Joe verspricht, nie wieder ein Buch zu schreiben. ;)

Matt ist seinen Freunden und seiner Familie für ihre anhaltende Geduld und Unterstützung dankbar. Er hofft, dass Seneca nach so viel Mühe und trotz verpasster Zeit mit der Familie über die Feiertage eine Fünf-Sterne-Bewertung abgeben wird.