Jesse Anderson

Daten-Teams

Ein einheitliches Managementmodell für erfolgreiche, datenorientierte Teams

Springer

Daten-Teams

Ein einheitliches Managementmodell für erfolgreiche, datenorientierte Teams

Jesse Anderson

Springer

Daten-Teams: Ein einheitliches Managementmodell für erfolgreiche, datenorientierte Teams

Jesse Anderson
Sintra, Portugal

ISBN 979-8-8688-0071-9 ISBN 979-8-8688-0072-6 (eBook)
https://doi.org/10.1007/979-8-8688-0072-6

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://portal.dnb.de abrufbar.

Übersetzung der englischen Ausgabe: „Data Teams“ von Jesse Anderson, © Jesse Anderson 2020. Veröffentlicht durch Apress. Alle Rechte vorbehalten.

Dieses Buch ist eine Übersetzung des Originals in Englisch „Data Teams“ von Anderson, Jesse, publiziert durch APress Media, LLC im Jahr 2020. Die Übersetzung erfolgte mit Hilfe von künstlicher Intelligenz (maschinelle Übersetzung). Eine anschließende Überarbeitung im Satzbetrieb erfolgte vor allem in inhaltlicher Hinsicht, so dass sich das Buch stilistisch anders lesen wird als eine herkömmliche Übersetzung. Springer Nature arbeitet kontinuierlich an der Weiterentwicklung von Werkzeugen für die Produktion von Büchern und an den damit verbundenen Technologien zur Unterstützung der Autoren.

Planung/Lektorat: Susan McDermott
Springer ist ein Imprint der eingetragenen Gesellschaft APress Media, LLC und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: 1 New York Plaza, New York, NY 10004, U.S.A.

Das Papier dieses Produkts ist recyclebar.

Dieses Buch ist Sara, Ashley und Grace gewidmet.

Inhaltsverzeichnis

Über den Autor

Jesse Anderson dient in drei Rollen am Big Data Institute: Data Engineer, Kreativingenieur und Geschäftsführer. Er arbeitet mit Unternehmen, die von Start-ups bis zu 100 von Fortune gelisteten Firmen reichen, an Big Data. Seine Arbeit umfasst Schulungen zu Spitzentechnologien wie Apache's Kafka, Hadoop und Spark. Er hat über 30.000 Menschen die Fähigkeiten vermittelt, die sie benötigen, um Data Engineer zu werden.

Jesse wird allgemein als Experte auf seinem Gebiet und für seine innovativen Lehrmethoden anerkannt. Er hat für O'Reilly und Pragmatic Programmers veröffentlicht. Er wurde in renommierten Publikationen wie *The Wall Street Journal,* CNN, BBC, NPR, Engadget und *Wired* vorgestellt. Er hat mindestens die letzten sechs Jahre damit verbracht, Data Teams zu beobachten, zu betreuen und mit ihnen zu arbeiten. Er hat dieses Wissen, warum Teams erfolgreich sind oder scheitern, in diesem Buch zusammengefasst.

Über den technischen Gutachter

Harvinder Atwal ist ein Datenprofi mit einer umfangreichen Karriere, in der er Analysen verwendet hat, um das Kundenerlebnis zu verbessern und die Geschäftsleistung zu steigern. Er ist nicht nur von Algorithmen begeistert, sondern auch von den Menschen, Prozessen und technologischen Veränderungen, die notwendig sind, um Wert aus Daten zu liefern.

Er genießt den Austausch von Ideen und hat auf der O'Reilly-Strata-Data-Konferenz in London, der Open-Data-Science-Konferenz (ODSC) in London und dem Data Leaders Summit in Barcelona gesprochen.

Harvinder leitet derzeit die Gruppendatenfunktion, die für den gesamten Datenlebenszyklus verantwortlich ist, einschließlich Datenakquisition, Datenmanagement, Daten-Governance, Cloud- und On-Premises-Datenplattformmanagement, Data Engineering, Business Intelligence, Produktanalytik und Data Science bei der Moneysupermarket Group. Zuvor leitete er Analyseteams bei Dunnhumby, der Lloyds Banking Group und British Airways. Seine Ausbildung umfasst einen Bachelor-Abschluss vom University College London und einen Master-Abschluss in Betriebsforschung von der Birmingham University School of Engineering.

Danksagungen

Danke an alle, die an mich geglaubt haben. Dieses Buch war eine schwierige Geburt. An meine Frau und Kinder. An meinen ersten Redakteur Jared Richardson, möge er in Frieden ruhen. Ein großes Dankeschön an Andy Oram.

Vielen Dank an alle Mitwirkenden: Ted Malaska, Paco Nathan, Lars Albertsson, Dean Wampler und Ben Lorica. Vielen Dank an die Personen, die sich für Fallstudieninterviews zur Verfügung gestellt haben: Eric Colson, Brad Klingenberg, Dmitriy Ryaboy, Harvinder Atwal, Bas Geerdink und ein anonymer Mensch in Großbritannien.

Einführung

Willkommen zu meinem Buch. Ich hoffe, Sie sind hier, um Ihr aktuelles Team zu sortieren, oder Sie stehen kurz davor, ein großes Datenprojekt zu starten. Vor Ihnen liegt ein langer Prozess, um ein Team zu reparieren oder zu erstellen – eher wie drei Teams.

Wenn es Ihnen wie mir geht, lesen Sie als Erstes die Einleitung, um zu sehen, worum es geht. Diese Einleitung dient als kurze Zusammenfassung davon, was mir im Kopf durchgeht, während ich dieses Buch schreibe.

Dies ist nicht der Ort, um über die neuesten Technologien oder sogar alte Technologien zu lernen. Tatsächlich werde ich Technologien und ihre Diskussion darüber bewusst meiden. Sie ändern sich ständig, und das sollten Sie wissen. Wahrscheinlich werden Sie keine 10- bis 20-jährigen Lebenszyklen aus diesen verteilten Systemtechnologien erreichen. Sie ändern sich oft, und es gibt eine neue „beste“ Technologie gleich um die Ecke, die versucht, den aktuellen Champion zu entthronen.

Im Zuge meiner Beratungstätigkeit konnte ich mit vielen Organisationen und Branchen auf der ganzen Welt zusammenarbeiten. Ich konnte die Muster und Gemeinsamkeiten sehen, weil ich Zugang zu einer größeren Datenmenge hatte und experimentieren konnte, was die besten Praktiken sein sollten. Wenn Sie das Buch lesen, bedenken Sie, dass jeder Gedanke und jede Idee keine akademische Theorie ist. Dies waren alles persönliche Erfahrungen und Wahrheiten, die ich manchmal auf die harte Tour gelernt habe. Ich werde nicht nur meine Geschichten und die Geschichte meines Unternehmens teilen, sondern die Geschichte aller Unternehmen, mit denen ich gearbeitet habe.

Ich schreibe dieses Buch hauptsächlich, um die besten Praktiken für die Erstellung von Data Teams zu etablieren und zu dokumentieren. Ich gehe seit vielen Jahren auf dieses quixotische Abenteuer. Ich war es leid, dass Organisationen keine Best Practices anwenden und bei ihren Datenprojekten scheitern. Ich begann zu erforschen, warum diese Misserfolge passierten. Während dieser Forschung stellte ich fest, dass nur wenige Menschen sich die Zeit oder Mühe genommen hatten, über ihre besten Praktiken zu sprechen. Es war nicht die Schuld des Managements, dass sie keine Best Practices befolgten, weil sie Stammeswissen waren! Um dies zu beheben, habe ich das Werk rund

um Technologie und Management für Big Data erstellt. Ich hoffe, dass ich Menschen dabei helfen kann, einige der Misserfolge zu vermeiden.

Wenn Sie ein Buch oder einen Artikel schreiben, kommt manchmal der aufschlussreichste Dialog während der Feedbackphase von den Gutachtern. Die Gutachter werden ein Stück durchlesen und großartiges Feedback zu einem Konzept geben. Dieses Feedback kann sich zu einem Thread entwickeln, der genauso interessant ist wie das Stück, aber nicht veröffentlicht wird. Ich habe versucht, die Kommentare der Gutachter in das Buch und die Unterabschnitte einzubringen. Auf diese Weise können Sie sehen, was sie gesagt haben und ihre Standpunkte nachvollziehen. Einige dieser Abschnitte im Buch werden dem zustimmen und das, was ich gesagt habe, erweitern. Andere Abschnitte werden meinen Ideen widersprechen, und ich respektiere die Standpunkte der Gutachter. Es gibt nicht nur einen Weg zum Erfolg mit Data Teams.

Im Einklang mit meinem Wunsch nach vielen Stimmen und Meinungen habe ich Fallstudien in das Buch aufgenommen. Anstatt nur Fallstudien über Unternehmen zu machen, habe ich Fallstudien über Menschen gemacht. Ich habe Menschen durch den Jobwechsel begleitet und gesehen, was sie beibehielten oder änderten, als sie zu einem neuen Unternehmen wechselten. Dies ermöglichte es mir, Fragen darüber zu stellen, wie ihre einzigartigen Erfahrungen ihre Strategie veränderten, als sie einem neuen Unternehmen beitraten.

Für einige wird dieses Buch eine Bestätigung dessen sein, was sie in ihrer eigenen Organisation erleben, sehen und zu ändern versuchen. Sie werden nicht gehört oder ihre Meinungen werden nicht beachtet. Dieses Buch wird als externe Validierung ihrer Ansichten dienen. Wenn ich auf einer Konferenz spreche oder einen Beitrag schreibe, sagen die Leute oft, dass ich die Dinge schreibe, die sie nicht sagen können oder nicht die Worte haben, um sie zu sagen. Es ist mehr eine Gruppentherapie als ein Konferenzgespräch.

Gleichzeitig fragen sich andere Leute, warum der Rest der Welt die Wahrheiten in diesem Buch nicht kennt, weil sie offensichtlich scheinen. Der Grund ist, dass das Management von Data Teams relativ neu ist und die Unterschiede nicht offensichtlich sind. Das ist ein weiterer Grund, warum ich dieses Buch schreibe. Ich möchte das weitergeben, was für einige von uns offensichtlich ist, und einige unserer Meinungen, damit Sie sich Ihr eigenes Urteil bilden können.

In den letzten Jahren habe ich ausführlich über Data Teams geschrieben. Manchmal habe ich weit mehr über ein Thema geschrieben, als ich geplant hatte. Achten Sie auf

diese Fußnoten, dort verweise ich Sie auf Links, um mehr über die Themen zu erfahren, die Sie interessieren.

Ich habe dies nicht geschrieben, um den endgültigen Leitfaden für die Erstellung von Data Teams zu erstellen. Dies ist der Anfang des Lernens, und ich versuche, so viele Unbekannte wie möglich zu beseitigen. Ich erwarte, dass Sie die Informationen nutzen und wissen, was als Nächstes kommt und einige der Überlegungen zu treffen sind, wenn Sie diese Entscheidungen treffen. Ich möchte die Fragen teilen, die Sie stellen sollten und die Antworten, die Sie suchen sollten.

Während ich mich auf verteilte Systeme und Big Data konzentriere, gelten die gleichen Arten von Prinzipien auch für Small Data. Bei kleinen Datengrößen wird die Komplexität sinken. Diese geringe Komplexität wird die technische Erfolgsschwelle auf ein leichter erreichbares Niveau verschieben.

Also lehnen Sie sich zurück, entspannen Sie sich und genießen Sie. Viel Glück auf Ihrer Reise!

Für weitere Informationen und Extras zum Buch gehen Sie bitte zu `www.datateams.io`.

TEIL I

Einführung in Daten-Teams

Bevor wir uns ausführlich mit den Details jedes Teils des Data Teams und der Interaktion mit diesen befassen, muss ich Ihnen eine allgemeine Einführung in Data Teams geben. Sobald Sie die Grundlagen jedes Teams verstehen, können wir anfangen, uns auf die Details zu konzentrieren.

KAPITEL 1

Data Teams

Oh, I get by with a little help of my friends

—„With a Little Help from My Friends“ von The Beatles

Die Nutzung von Big Data ist eine Teamsportart. Es braucht verschiedene Arten von Mitarbeitern, um Dinge zu erledigen, und in allen, außer in den kleinsten Organisationen, sollten diese in mehrere Teams organisiert sein. Wenn Sie die Hilfe von diesen Freunden erhalten, können Sie einige großartige Dinge tun. Wenn Ihnen Ihre Freunde fehlen, scheitern Sie und bleiben unter Ihren Möglichkeiten.

Wer sind diese Freunde, was sollten sie tun und wie machen sie es? Dieses Buch beantwortet diese Fragen. Das Buch behandelt viele Aspekte der Bildung von Data Teams: Welche Arten von Fähigkeiten Sie bei Mitarbeitern suchen sollten, wie Sie Mitarbeiter einstellen oder befördern sollten, wie die Teams miteinander und mit der größeren Organisation interagieren sollten und wie Sie Probleme erkennen und abwenden können.

Big Data und Datenprodukte

Um sicherzustellen, dass dieses Buch für Sie geeignet ist – dass meine Themen dem entsprechen, woran Ihre Organisation arbeitet – werde ich mir etwas Zeit nehmen, um die Arten von Projekten zu erklären, die in diesen Seiten behandelt werden.

Wie könnten wir ein Buch über Big-Data-Management ohne eine Definition von Big Data beginnen? Was ich hier erreichen möchte, ist über Schlagworte hinauszugehen und zu einer Definition zu gelangen, die dem Management wirklich hilft.

J. Anderson, *Daten-Teams*, https://doi.org/10.1007/979-8-8688-0072-6_1

Die schrecklichen 3er, 4er, 5er …

Jeder akzeptiert, dass Big Data ein ziemlich abstraktes Konzept ist: Sie können nicht einfach sagen, dass Sie Big Data haben, weil die Größen Ihrer Datensätze bestimmte Metriken erreichen. Sie müssen qualitative Unterschiede zwischen kleinen und großen Daten finden. Das wird schwierig.

Einer der ursprünglichen Versuche von Gartner, Big Data zu definieren, führte zur Schaffung der 3 Vs. Ursprünglich waren die Vs variety, velocity und volume. Es ist schwierig für das Management, diese Definition zu verstehen. Sie war zu breit. Als Ergebnis sagte jedes Unternehmen, dass sein Produkt Big Data sei, und das Management verstand die Definition immer noch nicht.

Dies führte dazu, dass die Menschen ihre eigene Definition wählten. Wählen Sie eine Zahl zwischen 3 und 20. Das ist die Anzahl der Vs, die definiert wurde.

Anstatt Klarheit zu schaffen, haben diese Definitionen zur Verwirrung geführt. Die Leute suchten einfach im Wörterbuch nach Vs, die so klangen, als sollten sie passen. Manager lernten nichts, was ihnen half, moderne Datenprojekte zu managen.

Die „Kann-nicht“-Definition

Für das Management bevorzuge ich die *Kann-nicht*-Definition. Wenn man gebeten wird, eine Aufgabe mit Daten zu erledigen, sagt die Person oder das Team, dass sie/es es nicht tun kann, normalerweise aufgrund einer technischen Einschränkung. Zum Beispiel, wenn Sie Ihr Analyseteam um einen Bericht bitten und dieses daraufhin meint, dass dies nicht möglich ist, haben Sie wahrscheinlich ein Big-Data-Problem.

Es ist zwingend notwendig, dass das *Kann-nicht* aufgrund eines technischen Grundes und nicht aufgrund der Fähigkeiten des Personals erfolgt. Hören Sie auf die Gründe, warum das Team sagt, dass es die Aufgabe nicht erledigen kann. Dies sind einige Beispiele für technische Gründe für ein *Kann-nicht*:

- Die Aufgabe wird zu lange dauern.
- Die Aufgabe wird unsere Produktionsdatenbank herunterfahren oder verlangsamen.
- Die Aufgabe erfordert zu viele Schritte zur Fertigstellung.
- Die Daten sind an zu vielen Orten verstreut, um die Aufgabe auszuführen.

Offensichtlich werden Ihre technisch besser ausgebildeten Leute eine präzisere und technischere Definition anbieten. Ich schlage dringend vor, dass Sie überprüfen, ob Ihre Data Teams wirklich verstehen, was Big Data ist und was nicht. Wenn das nicht der Fall ist, verlassen Sie sich auf vielleicht Leute, die die Anforderungen nicht wirklich erfüllen.

Einige Organisationen sind kleiner oder Start-ups. Was sollten sie tun, da sie noch nicht „kann nicht" sagen. Die Frage sollte dann lauten: Wird die Organisation in der Zukunft Big Data haben? Auf dieses zukünftige Big Data konzentrieren sich viele Unternehmen. Es ist schwierig, viele Pipelines und Codes zurückzugehen und neu zu entwickeln, um einen anderen Technologie-Stack zu verwenden. Einige Organisationen ziehen es vor, diese Probleme von Anfang an zu lösen, anstatt zu warten.

Warum das Management die Definition von Big Data kennen muss

Es ist entscheidend für das Management, zu verstehen, was Big Data ausmacht, und sie sollten von technisch qualifizierten Personen geleitet werden. Dies liegt daran, dass die Diskrepanz zwischen großen und kleinen Datenproblemen die Produktivität und Wertschöpfung wirklich zunichtemachen kann.

Die Verwendung von Small-Data-Technologien für Big-Data-Probleme führt zu *Kann-nicht*. Die Verwendung von Big-Data-Technologien für Small-Data-Probleme ist auch ein Problem und nicht nur, weil es zu Überengineering führt – es verursacht große Kosten und Probleme.[1]

Nur weil viele Big-Data-Technologien Open Source sind, bedeutet das nicht, dass sie billig sind. Ihre Kosten werden für Infrastruktur und Gehälter steigen. Big-Data-Technologien neigen dazu, spitz und voller Dornen zu sein, während Small-Data-Technologien weniger Nuancen haben, mit denen Sie kämpfen müssen. Während Small Data es Ihnen ermöglicht, alle Ihre Prozesse auf ein paar Computern auszuführen, explodiert die Anzahl der Computer mit Big Data. Plötzlich sind Ihre Infrastrukturkosten viel höher. Mit Big Data könnten Ihre Antwortzeiten, Verarbeitungszeiten und End-to-End-Zeiten steigen. Big Data ist nicht unbedingt schneller; es ist nur schneller und effizienter als die Verwendung von Small-Data-Technologien für zu große Daten.

[1]Wenn Sie Small Data haben, nehmen Sie dies nicht als Herabsetzung Ihres Anwendungsfalls, Ihres Unternehmens oder Ihres Projekts. Vielmehr sollten Sie erleichtert sein, dass Sie dem Thema Big Data entkommen sind (siehe www.jesse-anderson.com/2018/07/saying-you-have-small-data-isnt-belittling-your-use-case/).

Das Management muss wissen, wann es gegen den Hype um Big Data vorgehen muss. Das Management sollte nicht zulassen, dass es von Ingenieuren überzeugt wird, etwas zu verwenden, das nicht das richtige Werkzeug für den Job ist. Dies könnte eine Aufpolierung des Lebenslaufs von der Ingenieursseite sein. Aber im Falle von „Kann-nicht“ könnte Big Data der richtige Ansatz sein.

Wenn Sie es schließlich mit Small-Data-Technologien gerade so schaffen, werden Big-Data-Technologien noch schwieriger sein. Ich habe festgestellt, dass, wenn eine Organisation kaum in der Lage ist, Small-Data-Technologien zu nutzen oder zu produzieren, der signifikante Anstieg der Komplexität zu Misserfolgen oder unterdurchschnittlichen Projekten führt.

Warum ist Big Data so kompliziert?

Big Data ist 10- bis 15-mal komplizierter zu nutzen als kleine Daten.[2] Diese Komplexität erstreckt sich von technischen Problemen bis hin zu Managementproblemen. Ein Missverständnis, eine Unterschätzung oder Ignoranz dieser signifikanten Zunahme an Komplexität führt dazu, dass Organisationen scheitern.

Technisch gesehen beruht diese Komplexität auf der Notwendigkeit verteilter Systeme. Anstatt alles auf einem einzigen Computer zu erledigen, müssen Sie einen verteilten Code schreiben. Die verteilten Systeme selbst sind oft schwierig zu nutzen und müssen sorgfältig ausgewählt werden, da jedes spezifische Kompromisse hat.

Ein *verteiltes System* ist eine Aufgabe, die aufgeteilt und gleichzeitig auf mehreren Computern ausgeführt wird. Dies könnte auch bedeuten, dass Daten aufgeteilt und auf mehreren Computern gespeichert werden. Big-Data-Frameworks und -Technologien sind Beispiele für verteilte Systeme. Ich verwende diesen Begriff anstelle eines spezifischen Big-Data-Frameworks oder -Programms. Ehrlich gesagt kommen und gehen diese Big-Data-Frameworks leider.

Das Management wird auch komplexer, weil das Personal auf einer Ebene und mit einer Konsistenz über die Organisation hinweg erreichen muss, die es zuvor nie tun musste: verschiedene Abteilungen, Gruppen und Geschäftseinheiten. Zum Beispiel mussten Analyse- und Business-Intelligence-Teams nie die schieren Interaktionslevel

[2]Für eine vollständige Erklärung dieser Zunahme an Komplexität empfehle ich Ihnen, www.oreilly.com/learning/on-complexity-in-big-data/ zu lesen.

mit IT oder Engineering haben. Die IT-Organisation musste nie das Datenformat dem Operations Team erklären.

Sowohl aus technischer als auch aus Managementperspektive mussten Teams zuvor nicht mit einer so hohen Bandbreitenverbindung zusammenarbeiten. Es mag vorher ein gewisses Maß an Koordination gegeben haben, aber nicht so hoch.

Andere Organisationen stehen vor der Komplexität von Daten als Produkt anstelle von Software oder APIs als Produkt. Sie mussten noch nie für die in der Organisation verfügbaren Daten werben oder diese propagieren. Mit Data Pipelines wissen die Data Teams möglicherweise nicht einmal oder kontrollieren, wer Zugang zu den Datenprodukten hat.

Einige Teams sind sehr abgeschottet. Mit kleinen Daten konnten sie zurechtkommen. Es gab nie die Notwendigkeit, sich auszutauschen, zu koordinieren oder zu kooperieren. Die Zusammenarbeit mit diesen Maverick-Teams kann eine Herausforderung für sich sein. Hier ist das Management wirklich komplizierter.

Data Pipelines und Datenprodukte

Die Teams, über die wir in diesem Buch sprechen, befassen sich mit *Data Pipelines* und *Datenprodukten*. Kurz gesagt, eine Data Pipeline ist eine Möglichkeit, Daten verfügbar zu machen: Sie in eine Organisation zu bringen, sie an ein anderes Team zu übertragen und so weiter – aber in der Regel die Daten auf dem Weg zu transformieren, um sie nützlicher zu machen. Ein Datenprodukt nimmt einen Datensatz auf, organisiert die Daten auf eine Weise, die für andere konsumierbar ist, und stellt sie in einer Form zur Verfügung, die von anderen nutzbar ist.

Genauer gesagt ist eine Data Pipeline ein Prozess, um Rohdaten zu nehmen und sie so zu transformieren, dass sie vom nächsten Empfänger in der Organisation nutzbar sind. Um erfolgreich zu sein, müssen diese Daten von Technologien bereitgestellt werden, die die richtigen Werkzeuge für die Aufgabe sind, und die für die Anwendungsfälle korrekt sind. Die Daten selbst sind in Formaten verfügbar, die die sich verändernde Natur der Daten und die Unternehmensnachfrage danach widerspiegeln.[3]

[3]Es ist auch erwähnenswert, wie andere Leute Data Pipelines definieren und verwandte Fragen stellen (siehe www.jesse-anderson.com/2018/08/what-is-a-data-pipeline/).