

Jesse Anderson

Daten-Teams

Ein einheitliches Managementmodell für
erfolgreiche, datenorientierte Teams

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.

© Der/die Herausgeber bzw. der/die Autor(en), exklusiv lizenziert an APress Media, LLC, ein Teil von Springer Nature 2024

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

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	xix
Über den technischen Gutachter	xxi
Danksagungen	xxiii
Einführung	xxv
Teil I: Einführung in Daten-Teams	1
Kapitel 1: Data Teams	3
Big Data und Datenprodukte	3
Die schrecklichen 3er, 4er, 5er	4
Die „Kann-nicht“-Definition	4
Warum das Management die Definition von Big Data kennen muss	5
Warum ist Big Data so kompliziert?	6
Data Pipelines und Datenprodukte	7
Allgemeine Missverständnisse	8
„Es sind nur Daten“	8
„Ist das nicht nur etwas anders als ...?“	9
Warum werden Data Teams für Big Data benötigt?	11
Warum einige Teams scheitern und andere erfolgreich sind	12
Die drei Teams	13
Data Science	13
Data Engineering	15
Betrieb	16

- Warum werden drei Teams benötigt? 19
 - Drei Teams für kleine Organisationen 19
- Was passiert, wenn eine Organisation nicht richtig verwaltet?..... 20
- Kapitel 2: Die guten, die schlechten und die hässlichen Data Teams 21**
 - Erfolgreiche Data Teams..... 21
 - Wie Erfolg mit Big Data aussieht..... 22
 - Wie ein Scheitern von Big Data aussieht 23
 - Wie unterdurchschnittliche Projekte aussehen..... 24
 - Was passiert, wenn ein Team fehlt 25
 - Den erzeugten Wert herausfinden..... 26
 - Probleme mit der Skalierung in Data Science 27
 - Automatisieren Sie so viel wie möglich 28
- Teil II: Ihr Data-Team aufbauen 31**
 - Kapitel 3: Das Data-Science-Team 33**
 - Welche Fähigkeiten werden benötigt? 34
 - Mathematik 35
 - Programmierung 36
 - Verteilte Systeme 37
 - Kommunikation 37
 - Domänenwissen..... 38
 - Technische Schulden in Data-Science-Teams..... 39
 - Einstellung und Ausbildung von Data Scientists..... 41
 - Die Hindernisse bei der Umschulung 42
 - Verbesserung der Fähigkeiten von Data Scientists..... 43
 - Finden von Data Scientists..... 44
 - Die Bedürfnisse von Data Scientists erfüllen 45
 - Einführung von Software-Engineering-Praktiken 46
 - Zu viel Prozess hemmt den Fortschritt 47

Kapitel 4: Das Data-Engineering-Team	49
Welche Fähigkeiten werden benötigt?	50
Verteilte Systeme	52
Programmierung	53
Analyse.....	54
Visuelle Kommunikation.....	54
Mündliche Kommunikation	55
SQL.....	55
Schema	56
Domänenwissen.....	58
Andere wichtige Fähigkeiten	58
Ebenen der Expertise.....	60
Neue Data Engineers.....	60
Qualifizierter Data Engineer	60
Veteran.....	61
Weitere Spezialisierung	61
Sollte sich das Data-Engineering-Team nur auf Big Data konzentrieren?.....	62
Häufiges Missverständnis	63
Warum Data Engineering nicht nur Datenverwaltung ist	64
Warum ein Data Engineer kein Data Scientist ist.....	64
Warum Data Engineering mehr als Data Wrangling ist	66
Die Beziehung zwischen einem Data-Engineering-Team und einem bestehenden Data-Science-Team	66
Umschulung bestehender Mitarbeiter	67
Software Engineers	68
SQL-fokussierte Positionen.....	69
Die Rolle der Architekten	70
Platzierung in der Organisation.....	72

Kapitel 5: Das Operations Team.....	75
Die besondere Herausforderung des Betriebs von verteilten Systemen	75
Stellenbezeichnungen für Operations Teams	77
Welche Fähigkeiten werden benötigt?	78
Hardware.....	79
Software/Betriebssysteme.....	79
Verteilte Systeme	80
Fehlerbehebung	80
Sicherheit.....	81
Datenstrukturen und Formate.....	82
Skripterstellung/Programmierung	82
Best Practices für die Operationalisierung.....	83
Überwachung und Instrumentierung	83
Katastrophenwiederherstellung.....	84
Einrichten von Service-Level-Vereinbarungen	85
Batch-SLAs	85
Echtzeit-SLAs	86
Spezifische Service-/Technologie-SLAs.....	86
Organisationscode und Bereitstellungs-SLAs	87
Typische Probleme.....	87
Ihr Organisationscode	88
Daten und Datenqualität	88
Framework-Software	89
Hardware.....	90
Das Operations Team besetzen	90
Die Notwendigkeit spezialisierter Schulungen für Big Data	90
Umschulung bestehender Mitarbeiter.....	91
Warum ein Data Engineer kein guter Operations Engineer ist.....	92

Cloud vs. On-Premises-Systeme	93
Verwaltete Cloud-Dienste und Betrieb	94
Kapitel 6: Spezialisiertes Personal	95
DataOps	95
Die Kompromisse in DataOps.....	96
Personal für DataOps finden	99
Der Wert von DataOps	99
Beziehung zwischen den DataOps- und Data-Engineering-Teams.....	100
Wann mit DataOps beginnen?.....	101
Ingenieure für maschinelles Lernen	103
Finden von Ingenieuren für maschinelles Lernen	105
Wo man Ingenieure für maschinelles Lernen findet	106
Teil III: Zusammenarbeiten und Verwalten von Data Teams	107
Kapitel 7: Arbeiten als Data Team	109
Menschen dazu bringen, zusammenzuarbeiten	109
Personalverhältnisse.....	110
Sollten Data Teams getrennt oder zusammen sein?	110
Hochbandbreitenverbindungen	112
Ein iterativer Prozess	113
Politik, Erwartungen und Egos.....	114
Datenprojektgeschwindigkeit	114
Erstellen Sie einen Kriechen-, Gehen-, Laufen-Plan	115
Die Liebe verbreiten	116
Den Wert von Data Engineering, Data Science und Betrieb kommunizieren	116
Datenspezialisten benötigen die Beiträge anderer Teams	117
Datenhortung	119

INHALTSVERZEICHNIS

Tod durch tausend Schnitte	119
Der Fluch der Komplexität.....	119
Abseits des ausgetretenen Pfades gehen.....	120
Technische Schulden.....	121
Kapitel 8: Wie das Geschäft mit Data Teams interagiert	123
Wie Veränderung erreicht werden kann	124
Von oben nach unten drücken	124
Druck von der Mitte nach oben und unten.....	125
Von unten nach oben	126
Wie sollte das Geschäft mit den Data Teams interagieren?	126
Fallstudie: Fachwissen im Bereich der Krankenversicherung	128
Umstellung von Software als Produkt auf Daten als Produkt	132
Symptome unzureichender oder ineffektiver Interaktion.....	134
Zusammenarbeit mit dem QA-Team	135
Zusammenarbeit mit Projektmanagern	137
Finanzierung und Ressourcen	137
Personalbeschaffung	137
Software und Hardware	138
Cloud	140
Themen für die Interaktion	140
Datenstrategien.....	141
Risiken und Belohnungen	142
Verwaltung und Erstellung realistischer Ziele.....	143
Anwendungsfälle und Technologieauswahl	143
Sich der Herausforderung stellen	144
Aufmerksamkeit auf Führungsebene.....	145
Umgang mit Unsicherheit	146
Datenquellen ändern sich	147
Ausgabe von Modellen ist ein Grad an Sicherheit.....	147

Fürchten Sie nicht die Datenprodukte des Sensenmanns.....	148
Vergessen Sie nicht die menschliche Seite	149
Data Warehousing/DBA-Mitarbeiter	149
SQL-Entwickler/ETL-Entwickler	150
Betrieb.....	151
Business Intelligence und Data-Analysten-Teams.....	151
Einrichten von Key Performance Indicators (KPIs)	152
Data-Science-Team.....	152
Data-Engineering-Team	152
Betrieb.....	153
Kapitel 9: Management von Big-Data-Projekten	155
Planung der Erstellung und Nutzung von Datenprodukten.....	156
Einmalige und Ad-hoc-Einblicke	156
Ausstellen von Datenprodukten	157
Zuweisung von Aufgaben an Teams	158
Data Engineering, Data Science oder Betrieb?	158
Zusammenarbeit.....	159
Probleme mit bestehenden Teams beheben	160
Auswirkungen der Verwendung des falschen Teams für eine Aufgabe	161
Langfristiges Projektmanagement.....	162
Technologieauswahl	163
Ein mentaler Rahmen für das Hinzufügen oder Auswählen von Technologien	163
Technologieauswahl einschränken	164
Unterstützung von Programmiersprachen	166
Projektmanagement.....	169
Wenn alles schiefgeht	171
Achten Sie auf N-te Ordnungsfolgen	172

Kapitel 10: Ein Team gründen	175
Neue Teams gründen.....	175
Der erste Data Engineer	176
Der erste Data Scientist	177
Betrieb.....	177
Die kritische erste Einstellung	178
Standort und Status des Data Teams	180
Einbettung des Data Teams in die Geschäftseinheit	180
Nabe und Speiche	181
Kompetenzzentrum	182
Datendemokratisierung.....	182
Das Team von Anfang an richtig aufstellen.....	184
Data-Science-Start-ups	184
Unternehmen ohne Datenfokus	185
Kleine Organisationen und Start-ups	185
Mittelgroße Organisationen.....	188
Große Organisationen.....	190
Wie sollte die Berichtsstruktur aussehen?	191
VP of Engineering.....	191
Zusammenarbeit mit Personalabteilungen.....	195
Es geht nicht nur um Titel	195
Neue Titel hinzufügen	196
Fokus auf Technologiekenntnisse	198
Kapitel 11: Die Schritte für erfolgreiche Big-Data-Projekte	201
Vorschritte	201
Betrachten Sie Daten und Datenprobleme.....	202
Geschäftsbedürfnisse identifizieren.....	203
Ist das Problem ein Big-Data-Problem?.....	204
Entscheiden Sie sich für eine Ausführungsstrategie.....	205

Zwischen Planung und Start.....	206
Finden Sie einige leicht zu erreichende Ziele	206
Warten auf Daten	207
Die Anwendungsfälle grundlegend verstehen	208
Start des Projekts	208
Externe Hilfe in Anspruch nehmen.....	208
Wählen Sie die richtigen Technologien	211
Schreiben Sie den Code	213
Erstellen Sie das Computer-Cluster	213
Erstellen Sie Erfolgsmetriken.....	214
Iterative Schritte	215
Bewerten.....	215
Wann zu wiederholen.....	216
Kapitel 12: Organisatorische Veränderungen.....	217
Zusammenarbeit mit der alten Garde.....	218
Was tun mit den bisherigen Teams?.....	219
Innovation und Fortschritt in der Datenverarbeitung und Data Science.....	220
Bewertung von Kompetenzlücken	222
Bewertung der Expertise.....	222
Wie man eine Lückenanalyse durchführt.....	223
Ihre Gap-Analyse interpretieren	224
Fähigkeitslücken	225
Hardwareänderungen	226
Cloud-Nutzung	226
Cluster kaufen.....	226
Systeme für Big-Data-Teams	227
Daten für Data Scientists bereitstellen	229

Kapitel 13: Diagnose und Behebung von Problemen 233

- Festgefahrene Teams 234
 - Das Team sagt, es wird die Dinge erledigen, aber es hat das Gleiche schon vor einem Monat gesagt 234
 - Immer wenn ich dem Team eine Aufgabe gebe, kommt es zurück und sagt, es sei nicht möglich 235
 - Ich kann den Unterschied zwischen Fortschritt und Stillstand des Teams nicht erkennen 236
 - Wir haben Probleme mit kleinen Datensystemen, und wir haben noch mehr Probleme mit Big Data 236
- Unterperformende Teams 237
 - Immer wenn ich dem Data Team eine Aufgabe gebe, erstellt das Team etwas, das nicht funktioniert 237
 - Immer wenn ich dem Data Team eine Aufgabe gebe, erstellt es etwas, das nicht wirklich funktioniert 238
 - Das Team kann grundlegende Dinge tun, aber es kann nie etwas Komplizierteres tun 239
- Fähigkeits- und Kompetenzlücken 240
 - Wie kann ich den Unterschied zwischen einer Kompetenzlücke und einer Fähigkeitslücke erkennen? 240
 - Hat das Team Schwierigkeiten, das Programmieren zu erlernen? 241
 - Das Team kommt aus einem kleinen Datenhintergrund und hat Schwierigkeiten, Big Data und verteilte Systeme zu erlernen 242
 - Mein Team sagt, dass Sie falsch liegen und dass die Aufgabe einfach ist 243
 - Warum kann ich nicht einfach einige Anfänger einstellen und von diesen das Projekt erstellen lassen? 244
 - Wir haben ein Cluster für verteilte Systeme aufgebaut, und niemand benutzt es 245
- Verzerrte Verhältnisse 245
 - Meine Data Scientists beschwerten sich ständig, dass sie alles selbst machen müssen 246
 - Meine Data Scientists brechen ständig Projekte ab oder stoppen sie 247
 - Die Analyse der Data Engineers ist nicht sehr gut 248

Projektfehlschläge und schnelle Lösungen	248
Wir haben mehrere Projekte ausprobiert und keines von ihnen hat irgendwohin geführt.....	249
Wir haben uns von Big Data verabschiedet, weil wir keinen ROI erzielen können	250
Wir haben die Cloud ausprobiert und sind gescheitert; jetzt scheitern wir mit Big Data	251
Das ist wirklich schwer; gibt es einen einfacheren Weg, das zu erledigen?	253
Gibt es einfache Wege oder Abkürzungen zu diesen Problemen?	254
Wir haben eine Beratungsfirma eingestellt, um uns zu helfen, aber sie schafft es nicht.....	254
Wir haben alles befolgt, was unser Anbieter uns gesagt hat, und wir sind immer noch nicht erfolgreich.....	255
Wir haben unsere Datenstrategie definiert, aber es wird nichts erstellt.....	256
Unsere Datenmodelle scheitern ständig in der Produktion.....	256
Der Heilige Gral.....	257
Wir haben jemandes Architektur kopiert und wir erzielen nicht den gleichen Wert	257
Wir haben einen wirklich ehrgeizigen Plan und wir haben Schwierigkeiten, ihn zu erreichen	259
Die Software oder Data Pipeline versagt ständig in der Produktion.....	260
Wir haben ständig Produktionsausfälle	261
Die Daten bringen unser System ständig zum Absturz, und wir können es nicht stoppen.....	261
Es dauert viel zu lange, Probleme im Code und in der Produktion zu finden und zu beheben.....	263
Teil IV: Fallstudien und Interviews.....	265
Kapitel 14: Interview mit Eric Colson und Brad Klingenberg	267
Über dieses Interview	267
Hintergrund	267
Ausgangspunkt	268
Wachstum und Einstellung.....	269

INHALTSVERZEICHNIS

Die primäre Aufteilung in Data-Science- und Plattformteams.....	270
Bottom-up-Ansatz.....	273
Projektmanagement.....	275
Der Wettbewerbsvorteil von Daten.....	276
Ratschläge an andere Unternehmen.....	278
Erkenntnisse im Nachhinein	280
Kapitel 15: Interview mit Dmitriy Ryaboy	283
Über dieses Interview	283
Hintergrund.....	283
Einstellung bei Twitter.....	283
Herausforderungen von Daten und Analysen.....	286
Aufgabenbesitzstruktur.....	287
Die Schwierigkeit der Technologieauswahl und der Erstellung verteilter Systeme	288
Data Engineers, Data Scientists und Operations Engineers.....	289
Projektmanagement-Framework	291
Interaktionen zwischen Geschäfts- und Data Teams	291
Schlüssel zum Erfolg mit Data Teams	292
Kapitel 16: Interview mit Bas Geerdink.....	293
Über dieses Interview	293
Hintergrund.....	293
Die Data Teams der ING.....	294
ING Organisationsstruktur	296
Projektmanagement-Frameworks	298
Data Science im Bankwesen einsetzen	299
KPIs für Data Teams	299
Ratschläge für andere.....	300

Kapitel 17: Interview mit Harvinder Atwal	303
Über dieses Interview	303
Hintergrund	303
Teamstruktur	304
Barrieren und Reibungspunkte entfernen	305
Projektmanagement-Frameworks	307
Team-KPIs und Ziele	308
Veränderungen in den Data Teams	309
Ratschläge an andere	310
Kapitel 18: Interview mit einem großen britischen Telekommunikationsunternehmen	313
Über dieses Interview	313
Hintergrund	313
Die Initiative starten	314
Arbeiten mit dem Geschäft	314
Data Scientists und Data Engineers	315
Modelle für das Unternehmen unterstützen	315
Vom Proof of Concept zur Produktion	316
Erstellung von Unternehmensinfrastruktur und -betrieb	318
Projektmanagement-Frameworks	319
Ratschläge für andere	319
Kapitel 19: Interview mit Mikio Braun	321
Über dieses Interview	321
Hintergrund	321
Organisationsstruktur	321
Etablierung des Wertes des maschinellen Lernens	323
Definitionen von Jobtiteln	325
Reibung reduzieren	326

INHALTSVERZEICHNIS

Projektmanagement-Frameworks 327

KPIs 328

Verbesserung der Engineering-Fähigkeiten von Data Scientists 328

Integration von Operations Teams und Data Teams 329

Die Unterschiede zwischen europäischen und US-Unternehmen..... 329

Ratschläge an andere 330

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

EINFÜHRUNG

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.

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.

¹Wenn 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 *verteilt*es 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

²Für 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

³Es 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/).