Bild: Shutterstock

FastChangeCo und seine schnelle Veränderung mit einem anpassungsfähigen Cloud-Data-Warehouse

Model-Driven Decision Making

Ein Beitrag von Dirk Lerner, André Dörr und Mathias Brink

Das fiktive Unternehmen FastChangeCo hat eine Möglichkeit entwickelt, nicht nur Smart Devices herzustellen, sondern auch die Smart Devices als Wearables in Form von Bio-Sensoren auf Kleidung und Lebewesen auszudehnen. Bei jedem dieser Geräte entsteht eine große Menge an (sensiblen) Daten, genauer gesagt: durch die Aufzeichnung und Aufbereitung sowie die Auswertung personen- und umweltbezogener Daten. Das Geschäftsmodell von FastChangeCo hat nicht zum primären Ziel, die Technologie zu lizenzieren und gewinnbringend zu verkaufen. Im Gegenteil – das Ziel besteht darin, über niedrige Preise die Technologie möglichst breit im Markt zu positionieren, um später mit einer großen Datenbasis Geld zu verdienen.

Anhand der gesammelten Daten will FastChangeCo zukunftsgerichtete Entscheidungen treffen und innovative Services aufbauen. Sie sollen zum einen Kunden gezielt zum Kauf animieren, aber auch die Qualität der vorhandenen Produkte signifikant verbessern und als Datenbasis für die Entwicklung zukünftiger Produkte dienen – (Data) Model-Driven Decision Making.

Die Sensordaten, welche die Smart Devices erzeugen, summieren sich über die Zeit auf ein sehr großes Datenvolumen. Dies hat die Entscheidung des BI-Teams von FastChangeCo für eine hybride Datenhaltung, vollständig in der Cloud, wesentlich beeinflusst [BrL16].

Die Umsätze mit den Produkten von FastChangeCo, wie Sportbekleidung, Lifestyle-Produkten sowie Ernährungs-Supplementen, ist während des Wochenendes am höchsten. Daher ist der ideale Zeitpunkt für das regelmäßige Training der Predictive-Modelle frühmorgens zum Wochenbeginn. Das ist ein weiterer Grund, die Cloud einzusetzen, da der hohe Bedarf an Rechenkapazitäten für das wöchentliche Training der Predictive-Modelle nicht mit bestehenden Reporting-Workloads konkurrieren soll. Ansonsten müsste das BI-Team das Data-Warehouse-System um hohe Rechenkapazitäten erweitern, die den Rest der Zeit ungenutzt blieben. Dazu ist FastChangeCo nicht bereit.

Datenmodellierung

Um die Daten zu einer logischen Gesamtsicht zu kombinieren, hat sich FastChangeCo einer konsequenten Umsetzung der Informationslandschaft mit Methoden der Datenmodellierung verschrieben. Ein logisches Datenmodell bildet dabei die notwendigen Geschäftsobjekte sowie deren Attribute und Eigenschaften unabhängig von der eingesetzten Technologie eines Unternehmens ab.

Wo, wie und mit welcher Technologie die anfallenden Daten des Data Warehouse persistiert werden, ist im logischen Datenmodell nicht relevant. Erst mit der Entwicklung der physischen Datenmodelle wird die eingesetzte Technik wichtig und somit die physische Modellierungsmethode

[Hob09]. Hier bietet sich in dieser Konstellation die Modellierungsmethode Data Vault an, in der technologieübergreifend die Daten modelliert werden können, sowie eine Virtualisierungstechnik, die als zentrale Instanz alle beteiligten Technologien vereint [Lin14]. Der zentrale Aspekt der Virtualisierung ist, dass es für den Anwender unerheblich ist, wo die Daten tatsächlich liegen, der Anwender jedoch eine vollständig integrierte Datenlandschaft vorfindet.

Ein weiterer Aspekt ist, dass die Daten für das Predictive Modelling bereits gut strukturiert und qualitativ hochwertig aufbereitet sind. So ist es FastChangeCo möglich, die daraus neu gewonnenen Erkenntnisse für die Umsetzung ihrer Ziele, nämlich der zukunftsgerichteten Entscheidungen und neuer innovativer Services, zu verwenden.

Um die dafür notwendigen Recommendations aus dem Predictive Modelling zu erhalten, sind große Mengen an Rohdaten der zugrunde liegenden Geschäftsobjekte erforderlich. In Gesprächen mit den Fachbereichen stellte sich heraus, dass die Geschäftsobjekte Kunde (Customer, „C“), Produkt (Part, „P“) und zugehöriger Lieferant (Supplier, „S“) sowie Bestellung (Order, „O“) und Sensordaten (Measure, „M“) dafür notwendig sind. Zusätzlich ist das neue Geschäftsobjekt der Recommendations („R“) im logischen Datenmodell zu entwerfen.

Abbildung 1 zeigt, dass die Sensordaten („M“) abhängig von den von Kunden bestellten („O“) Produkten („P“) sind. Ein Kunde erzeugt somit Sensordaten durch mehrere ihm zugeordnete Produkte. Der Lieferant („S“) ist als Geschäftsobjekt notwendig, da ein und dasselbe Produkt von unterschiedlichen Lieferanten produziert wird. Zum einen ist so eine eindeutige Zuordnung eines Stücks dieses Produkts zum Hersteller möglich. Zum anderen – für diesen Use-Case viel wichtiger – generiert dasselbe Produkt aufgrund unterschiedlicher verbauter Sensoren durch die Lieferanten leicht unterschiedliche Daten.

Das logische Datenmodell überführen die Datenmodellierer von FastChangeCo im nächsten Schritt in ein physisches Data-Vault-Datenmodell, das schon konkret die spätere Implementierung auf der Zieltechnologie darstellt. Wie Abbildung 2 zeigt, liegen die Sensordaten aufgrund der Datenmenge und der hohen Datenrate auf einem Hadoop-Cluster in der Cloud. Die besondere Kritikalität der personen- und umweltbezogenen Daten bedeutet für die Datenmodellierer, dass sie diese in mehrere Schutzklassen einordnen müssen, die auch den Anforderungen der Datenschutz-Grundverordnung (DSGVO) gerecht werden. Da Kunden Anspruch auf die Löschung ihrer Daten haben, achten die Datenmodellierer bereits beim Design der Datenmodelle auf die Möglichkeit einer einfachen Löschung anhand der Schutzklassen, zum Beispiel durch eine zusätzliche Separation der Daten sowohl im Datenmodell als auch in der physischen Implementierung. Die Schutzklassen sind vor einem unberechtigten Zugriff durch ein Nutzer- und Rollenkonzept geschützt sowie nach Bedarf verschlüsselt.

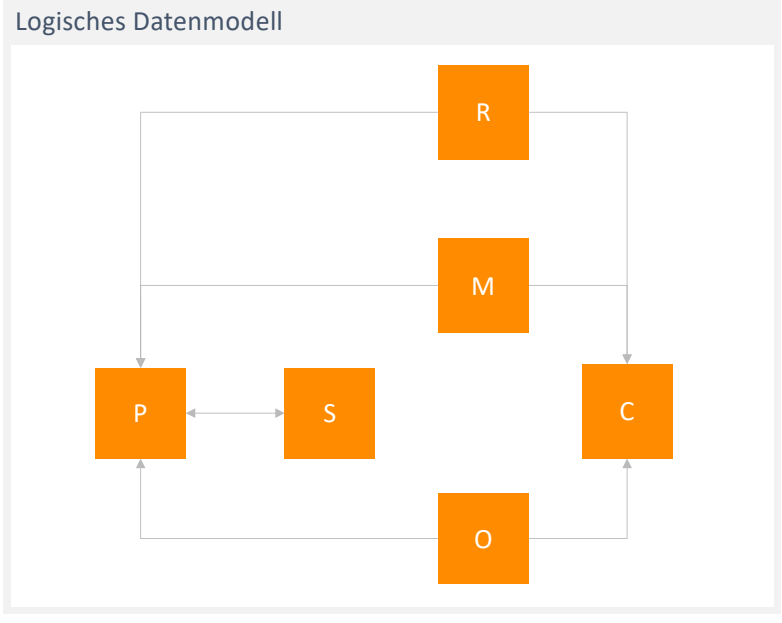


DIRK LERNER ist ein unabhängiger, erfahrener Berater und Gründer sowie Geschäftsführer der TEDAMOH GmbH. Seit über 18 Jahren leitet er BI-Projekte und gilt als umfassender Experte für BI-Architekturen und Datenmodellierung. Als Pionier für Data Vault und FCO-IM in Deutschland veröffentlichte er verschiedene Publikationen, ist internationaler Redner auf Konferenzen und Autor des Blogs https://tedamoh.com/blog.
E-Mail: Dirk.Lerner@tedamoh.com

ANDRÉ DÖRR ist Data Scientist bei Exasol. Mit mehr als 12 Jahren Erfahrung im BI-Umfeld und in unterschiedlichsten Branchen ist er der führende Data Warehouse Architect bei Exasol. Er ist Sprecher auf Konferenzen und Autor des auf Sportwetten fokussierten Data-Science-Blogs https://beatthebookie.blog.
E-Mail: Andre.Doerr@exasol.com

MATHIAS BRINK betreut seit 9 Jahren Kunden im Aufbau und Betrieb von analytischen Plattformen. Die Integration unterschiedlichster Technologien in komplexen Data-Warehouse-Lösungen ist seine Leidenschaft und Vision als Solution Manager bei Exasol.
E-Mail: Mathias.Brink@exasol.com

Abb. 1: Logisches Datenmodell für das Model-Driven Decision Making, vereinfacht ohne Attribute



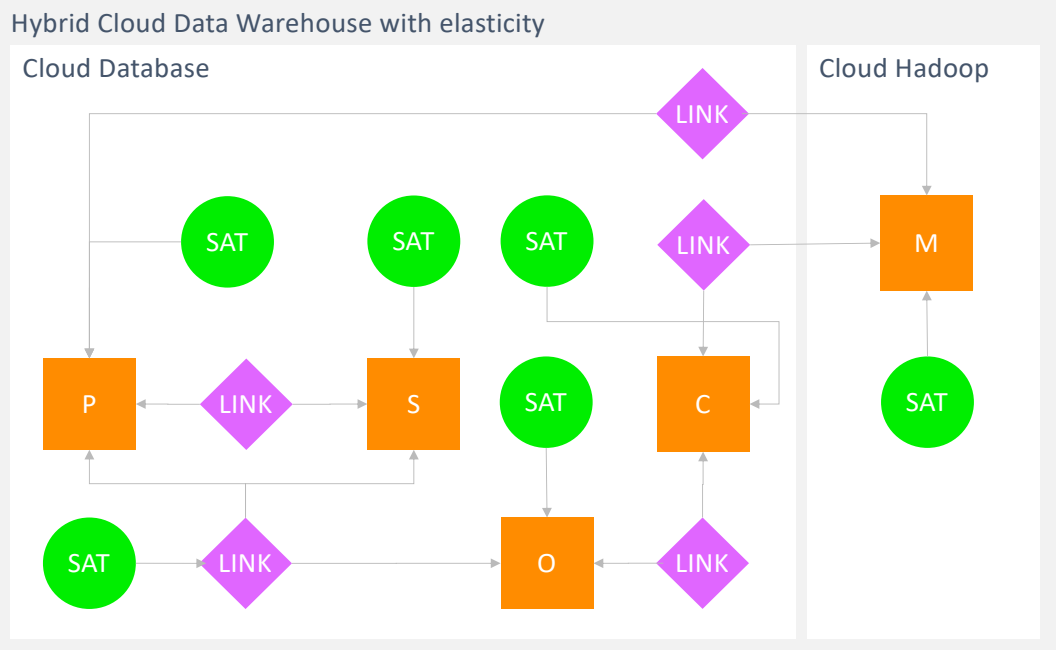


Abb. 2: Physisches Data-Vault-Datenmodell in der Cloud

Die eingesetzten Technologien ermöglichen einen transparenten Zugriff auf die Daten. Ohne Kenntnisse der zugrunde liegenden Systeme kann jeder Nutzer – Personen oder weitere nachgelagerte Systeme – mit einfachem SQL die Daten abfragen. So kann auch die Recommendation Engine, wie in Abbildung 4, Schritt 2 gezeigt, die Rohdaten für Kunden („C"), Produkte („P") und Sensordaten („M") ohne Weiteres abholen und verarbeiten. Die schlussendlich generierten Empfehlungen modelliert das BI-Team als virtuelle Data-Vault-Objekte (Abbildung 4, Schritt 4), die direkt auf das Predictive Model im Cloud-Filesystem zugreifen.

Predictive Modelling

Descriptive Analytics wird im Bereich der Business Intelligence schon seit geraumer Zeit eingesetzt, um Information und Wissen aus historischen Daten zu gewinnen. Predictive Analytics geht dabei noch einen Schritt weiter und nutzt diese historischen Daten, um daraus Vorhersagen für die Zukunft zu erstellen. Große Datenmengen, wie die Sensordaten, über das Verhalten der Nutzer ermöglichen es, Vorhersagevariablen für einfache wie auch komplexe Problemstellungen (Abbildung 3, Schritt 1) zu definieren, zum Beispiel: Wann sollte sich ein Läufer neue Schuhe kaufen? Vorhersagemethoden ermitteln mit Hilfe dieser Variablen die Wahrscheinlichkeit zukünftiger Ereignisse oder in dem hier gezeigten Use-Case Empfehlungen für die Kunden.

Diese Methoden kann man grob in drei Gruppen einteilen:

- Statistische Modelle
- Klassische Machine-Learning-Methoden
- Deep Learning

Deep-Learning-Methoden bzw. neuronale Netze bieten hierbei das größte Potenzial, da sie in der Regel eine höhere Genauigkeit erreichen und vielseitiger einsetzbar sind [Sho18; Hal18].

Für die Predictive Analytics kommt bei FastChangeCo das von Google entwickelte Tensorflow zum Einsatz. Es ist aktuell eines der bekanntesten und meist genutzten Deep-Learning-Frameworks für den Aufbau neuronaler Netze. Da zu Beginn des Projekts die Data Scientists bei FastChangeCo eher unerfahren im Umgang mit Deep Learning waren, verwendeten sie zusätzlich das Paket tflearn, das als High-Level-API für Tensorflow die Erstellung neuronaler Netze leichter macht. Zusammen mit dem Adam-Optimizer für den Trainingsprozess vereinfachte dies aufgrund der automatischen Learning-Rate-Anpassung die für das Verbessern des Deep Learning notwendige Optimierung der Parameter.

Bevor ein neuronales Netz jedoch ermitteln kann, wann ein Läufer sich neue Schuhe kaufen sollte, muss dieses trainiert werden. Das BI-Team von FastChangeCo verwendet hier das sogenannte Supervised Learning.

Wie bereits erwähnt, müssen zunächst passende Vorhersagevariablen bestimmt werden, die

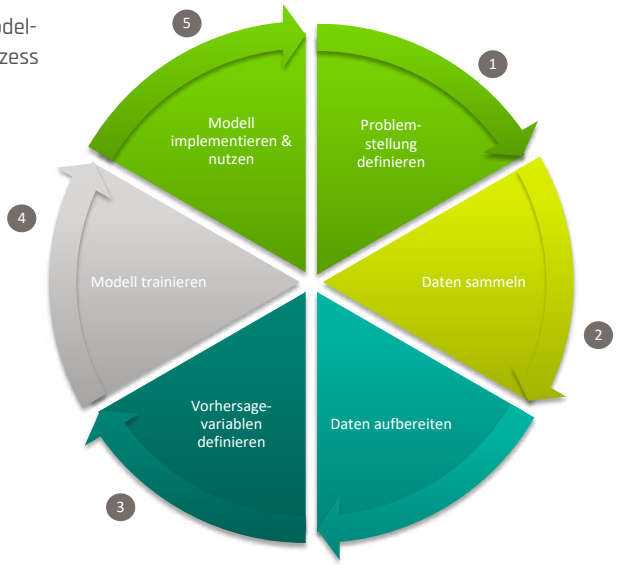


Abb. 3: Predictive-Modelling-Entwicklungsprozess

mit der Haltbarkeit von Laufschuhen korrelieren (Abbildung 3, Schritt 3). Die gelaufene Strecke und das Gewicht des Läufers sind Beispiele für Vorhersagevariablen, geliefert von den verkauften Smart Devices.

Rohdaten aus dem Data Warehouse über weitere Bestellungen der den Smart Devices zugeordneten Kunden liefern dem Deep-Learning-Framework zusätzliche Informationen, zum Beispiel wann und in welchen Abständen Kunden in der Vergangenheit neue Schuhe gekauft haben (Abbildung 4, Schritt 2 und Abbildung 3, Schritt 2). Dies entspricht den Ausgabeklassen für die Vorhersage [Doe17].

Während der Trainingsphase ermittelt das neuronale Netz den Zusammenhang zwischen den verschiedenen Vorhersagevariablen und den Vorhersageklassen (Abbildung 3, Schritt 4). Das trainierte Netz ist danach in der Lage, auf Basis der aktuellen Laufdaten zu bestimmen, wann ein Kunde wieder über ein neues Paar Schuhe nachdenken beziehungsweise aus Sicht von FastChangeCo eines kaufen sollte (Abbildung 3, Schritt 5).

Diese Vorgehensweise kann das BI-Team auf weitere Problemstellungen anwenden. So ist es FastChangeCo möglich, Kunden mit Hilfe von deren Sensordaten und Sportprofilen gezielt mit Empfehlungen und Hinweisen zu versorgen sowie ihnen weitere Dienstleistungen anzubieten.

Umsetzung

Für die technologische Umsetzung des physischen Datenmodells (siehe Abbildung 2) im Zusammenhang mit Deep-Learning-Algorithmen musste das BI-Team mehrere Anforderungen meistern. Neben einer kosteneffizienten Speicherung der enormen Menge an Sensordaten ist eine performante Auswertung der Abverkaufsdaten in Standard-Reports oder Self-Service-Analysen eine wichtige Anforderung. Während die Auswertung der Daten einen relativ festen Workload darstellt und keine dynamischen Ressourcen benötigt, wird für das Trainieren der Modelle nur zu ausgewählten Zeiten eine hohe Rechenleistung benötigt.

Da FastChangeCo sich einer kompletten Cloud-Strategie verschrieben hat und Amazon Web Services (AWS) zurzeit am weitesten am Markt verbreitet ist, hat sich das BI-Team für die AWS Cloud entschieden [AWS19]. Dazu trug ebenfalls bei, dass innerhalb der FastChangeCo bereits die AWS Cloud verwendet wird und so das Know-how aus anderen Bereichen nutzbar ist. Die mit allen führenden Anbietern (Amazon, Google, Microsoft) durchgeführten Proofs of Concept zeigten ebenso, dass für diesen hier vorgestellten Use-Case AWS ideal ist. Nichtsdestotrotz ist es für jeden Use-Case zu empfehlen, eine dedizierte, erneute kritische Prüfung der Technologien durchzuführen, insbesondere in Bezug auf die potenziell anfallenden Kosten.

Die von FastChangeCo in der Cloud eingesetzten Technologien sollen aber auch auf andere Cloud-Anbieter portierbar sein, um zukünftigen Technologietrends folgen zu können oder Kostenvorteile nutzbar zu machen.

Für die kostengünstige Ablage der Sensordaten (Abbildung 4, „Cloud Hadoop“) hat sich das BI-Team hier für ein Amazon-EMR-Cluster entschieden [EMR19]. Weitere Hadoop-Distributionen wurden betrachtet und hatten sowohl in Funktionsumfang als auch in Wartbarkeit ein gleiches Scoring wie EMR. Allerdings konnte EMR durch ein besseres Preis-Leistungs-Verhältnis punkten. Um sowohl die Verkaufszahlen (Abbildung 4, „Cloud Database“) performant auswerten zu können, aber auch mit den Sensordaten transparent vereinigen zu können, fiel die Entscheidung auf die analytische Datenbank von Exasol. Diese bietet mittels virtueller Schemen die Möglichkeit, Daten aus verschiedenen Quellen zu vereinigen ohne den zwingenden Bedarf, diese persistent in die Datenbank zu laden. Daneben hat FastChangeCo mit Exasol bereits On-Premise gute Erfahrungen

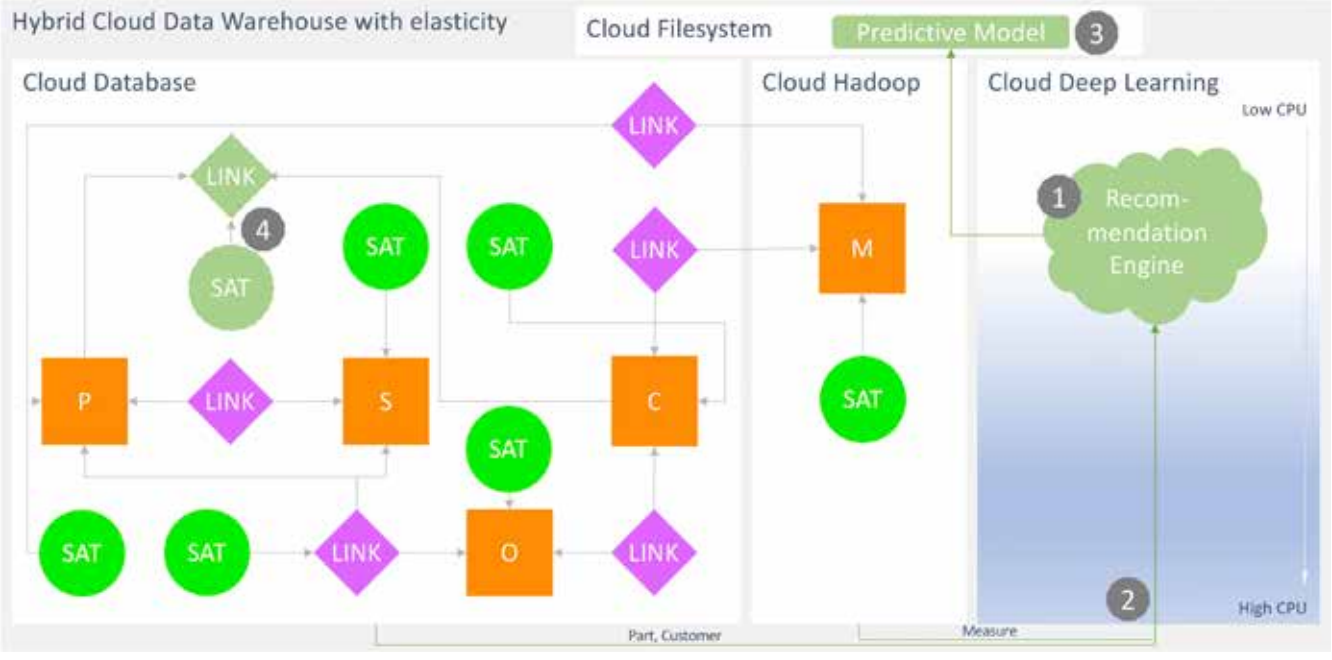


Abb. 4: Recommendation Engine, Predictive Model und virtualisierte Integration-Empfehlungen

gesammelt und eine Migration auf eine andere Datenbanktechnologie konnte vermieden werden.

Ein weiterer Vorteil der vom BI-Team gewählten technischen Architektur ist die Möglichkeit, über die Programmiersprache Python das Modelltraining direkt aus der Datenbank anzustoßen und die trainierten Modelle direkt in der Datenbank für Vorhersagen zu nutzen.

Für das Training werden nutzerdefinierte Funktionen verwendet, die dynamisch einen Python-Server in AWS starten (Abbildung 4, Schritt 1) und Daten zwischen der Datenbank und dem Trainingsserver transferieren (Abbildung 4, Schritt 2). So sind eine minimale Belastung der Datenbank und dynamische Ressourcen für das Training der Predictive-Modelle sichergestellt.

Die trainierten Modelle schreibt die Recommendation Engine zurück in das Cloud- (Exasol-Bucket-) Filesystem (Abbildung 4, Schritt 3). Die aus den trainierten Modellen abgeleiteten Vorhersagen bindet das BI-Team über einen virtuellen Link und Satelliten, denen nutzerdefinierte Funktionen zugrunde liegen, in die Datenstrukturen des Data Warehouse ein (Abbildung 4, Schritt 4).

Fazit

Mit der Kombination einer logischen und physischen Datenmodellierung, einem agilen Ansatz und Best-of-Breed-Komponenten in einer öffentlichen Cloud konnte FastChangeCo die gestellten Anforderungen kosteneffizient und schnell umsetzen. Das gewählte Konzept kombiniert die Vorteile der einzelnen Technologien ideal.

Eine kosteneffiziente Speicherung der Sensordaten bei gleichzeitiger Möglichkeit der Auswertung in einem Amazon-EMR-Cluster wird um die performante In-Memory-Datenbank ergänzt, um schnelle Analysen der Abverkaufsdaten und eine transparente Integration der anderen Technologien zu gewährleisten.

FastChangeCo ist davon überzeugt, dass der gewählte Weg zukunftssicher ist, da aufgrund der logischen Modellierung die fachlichen Anforderungen und das Datenmodell technologieunabhängig sind und somit Technologien zukünftig ausgetauscht oder auch um weitere Konzepte oder Technologien erweitert werden können.

Literatur

[AWS19] Amazon Web Services, https://aws.amazon.com/, abgerufen am 22.7.2019

[BrL16] Brink, M. / Lerner, D.: Hybrides Data Warehouse am Beispiel von Data Vault – Shared Data in einer virtuellen Architektur. In: BI-Spektrum, 5-2016

[Doe17] Doerr, A.: How To: Develop predictive models. https://beatthebookie.blog/2017/04/11/how-to-develop-predictive-models/, abgerufen am 22.7.2019

[EMR19] Amazon EMR. https://aws.amazon.com/emr/, abgerufen am 22.7.2019

[Hal18] Hale, J.: Deep Learning Framework Power Scores 2018. 20.9.2018, https://towardsdatascience.com/deep-learning-framework-power-scores-2018-23607ddf297a, abgerufen am 22.7.2019

[Hob09] Hoberman, S. et al.: Data Modeling Made Simple. Technic Publications 2009

[Lin14] Linstedt, D.: Dan Linstedt.com, http://danlinstedt.com/allposts/datavaultcat/nosql-platforms-and-datavault-curiosity-bigdata-datamodeling/, abgerufen am 22.7.2019

[Sho18] Shorten, C.: Machine Learning vs. Deep Learning. 7.9.2018, https://towardsdatascience.com/machine-learning-vs-deep-learning-62137a1c9842, abgerufen am 22.7.2019

Die nächste Ausgabe von BI-Spektrum erscheint am 17.10.2019

Impressum

Verlag
SIGS DATACOM

SIGS DATACOM GmbH
Lindlaustr. 2c, D-53842 Troisdorf
Tel: +49 2241 2341-100
Fax: +49 2241 2341-199
E-Mail: bi-spektrum@sigs-datacom.de
www.bi-spektrum.de

Verlagsleitung
Günter Fuhrmeister
Emanuel Rosenauer
E-Mail: emanuel.rosenauer@sigs-datacom.de

Herausgeber
Prof. Dr. Carsten Felden, TU Bergakademie Freiberg
Assistentin: Dr. Tatiana Arzhakova

Schlussredaktion
Kirsten Skacel, Lektorat Rotstift
E-Mail: bi-spektrum@lektorat-rotstift.de

Fachbeirat
Frank Beier, msg systems AG; Dr. Klaus Detemple, SEVEN PRINCIPLES AG; Dr. Carsten Dittmar, Alexander Thamm GmbH; Daniel Eiduzzis, DXC Technology; Markus Enderlein, INFOMOTION GmbH; Martin Engler, Stadtwerke Leipzig GmbH; Dr. Ralf Finger, Information Works GmbH; Tom Gansor, OPITZ CONSULTING GmbH; Prof. Dr. Uwe Haneke, Hochschule Karlsruhe; Leif Hitzschke; Dr. Joachim Philippi, SEVEN PRINCIPLES AG; Dr. Torsten Priebe, Simplity; Christian Weinberger, meta Informationssysteme GmbH; Ingo Weishaupt, Viridium Gruppe; Dr. Michael Zimmer, Deloitte Consulting GmbH

Chefredaktion
Christoph Witte (V.i.S.d.P.)
E-Mail: christoph.witte@sigs-datacom.de

Redaktions- und Herstellungsleitung Zeitschriften
Emanuel Rosenauer

Herstellung
Bonifatius GmbH, Druck · Buch · Verlag
Karl-Schurz-Str. 26, D-33100 Paderborn
Tel: +49 5251 153-0, Fax: +49 5251 153-104
Grafik: Elke Brosch
E-Mail: anzeigendaten@bi-spektrum.de

Anzeigen
Andreas Dietz, Tel: +49 2241 2341-577
E-Mail: andreas.dietz@sigs-datacom.de
Martin Bena, Tel: +49 2241 2341-588
E-Mail: martin.bena@sigs-datacom.de

Datenpflege Digitalausgabe/App
Klaus Gebert

Anzeigenpreise
Es gilt die Anzeigenpreisliste Nr. 15 vom 01.11.2018

Illustrationen Titelbild: Shutterstock

Abonnentenservice
IPS Datenservice GmbH,
Postfach 1331, D-53335 Meckenheim
Tel: +49 2225 70 85-374
Fax: +49 2225 70 85-376
E-Mail: sigsdatacom@aboteam.de

Erscheinungsweise
5 Mal jährlich plus Sonderhefte

Bezugspreise
Einzelverkaufspreis:
D € 18,00; A € 21,50; SFR 33,25
Jahresabonnement: D € 72,00, Ausland: € 84,00,
Studentenabonnement: € 36,00

Sie möchten in BI-SPEKTRUM veröffentlichen? Schicken Sie uns Ihren Artikel!

Senden Sie Ihre Vorschläge an christoph.witte@sigs-datacom.de

Senden Sie uns gern Ihre Artikelvorschläge zu den genannten Titelthemen.

Selbstverständlich sind wir darüber hinaus aber auch an weiteren spannenden Themen und Artikeln interessiert. Das können zum Beispiel Anwenderberichte, Beiträge zu technologischen Themen und Entwicklungen oder Kommentare zu Fachthemen sein.

Bitte senden Sie Ihre Vorschläge mit einer Kurzbeschreibung des geplanten Beitrags und einer Autorenkurzbiografie an die Redaktion unter **christoph.witte@sigs-datacom.de**

BI-SPEKTRUM plant für die übernächste Ausgabe das folgende Titelthema:

Ausgabe 05/2019

Analytical

Face to the Customer

Einreichungen bis 30.08.2019

Wir freuen uns auf Ihre Ideen!