tdwi Ausgabe 2 | 14. Jahrgang 2019 | D € 18,00 A € 21,50 SFR 33,25 ISSN 1862-5789

BI Spektrum

EINE PUBLIKATION DES TDWI E.V.

BI ist tot! Es lebe das neue Miteinander von BI, Analytics und Big Data ab Seite 8

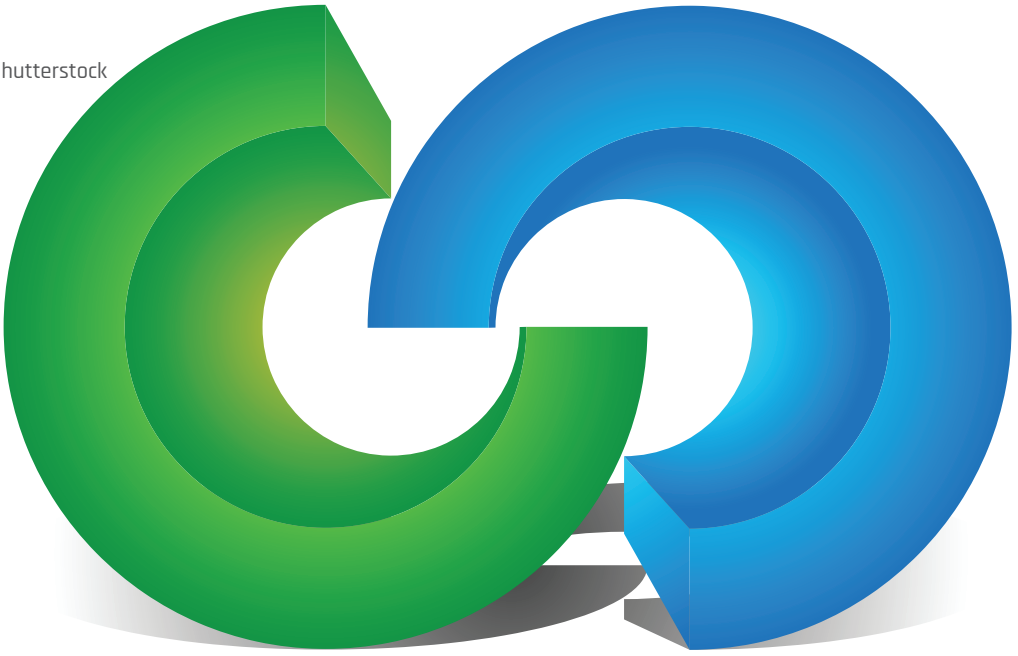
Digitale Transformation
Cloud als Schlüssel zu smarten Produkten
Seite 34

Ende des IT-Bottleneck
BI & Analytics im Self-Service
Seite 44

Interview
„AI muss nicht immer das große Ding sein“
Seite 30

Dr. Jürgen Wirtgen, Microsoft

Grafik: Shutterstock



Symbiose von Data-Warehouse- & Big-Data-Lösungen

Koexistenz in BI-Landschaften

Ein Beitrag von
Boris Vogt

Daten sind in Unternehmen mittlerweile ein anerkannter Vermögenswert. Um diesen Vermögenswert intelligent zu nutzen, haben sich in der Praxis konventionelle Data-Warehouse- und moderne Big-Data-Ansätze entwickelt. Da Daten die Grundlage beider Ansätze bilden, werden diese aus einer Makroperspektive häufig gleichgesetzt. Tatsächlich nehmen sie jedoch unterschiedliche technische, fachliche und rechtliche Parameter in den Blick: Ein klassisches Data Warehouse (DWH) und Big-Data-Technologien bieten Lösungen für ganz unterschiedliche Aufgabenstellungen und Zielrichtungen. Für eine optimale Ausschöpfung der Daten ist daher eine zielgerichtete Symbiose beider Ansätze sinnvoll.

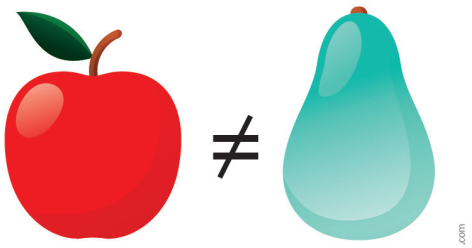


Abb. 1: Data Warehouse vs. Big Data

In diesem Artikel werden beide Lösungsansätze diskutiert. Hierzu wird im ersten Schritt eine Definition der Lösungen vorgenommen, um sie anschließend einander gegenüberzustellen. Ziel des Artikels ist es, die Frage „Warum und wofür brauche ich welche Ansätze“ zu beantworten (vgl. Abbildung 1).

Data Warehouse

In einem Data Warehouse werden Daten aus gleich- und verschiedenartigen Quellsystemen extrahiert, transformiert und geladen (ETL). Die so im DWH historisch gespeicherten Daten bilden den Single Point of Data und unterstützen die Berichts- und Analysesysteme eines Unternehmens. Es ist ein zentraler Bestandteil heutiger Business-Intelligence- und Business-Analytics-Umgebungen.

Folgende charakteristischen Eigenschaften muss ein DWH erfüllen [Inm05] (siehe Abbildung 2):

- **Subject-Oriented** (Themenorientiert): Während sich in operativen Systemen die Daten an den Prozessen des Unternehmens ausrichten, orientieren sich die Daten im DWH an fachlichen Themen – sogenannten Subjects. Die Konzentration auf solche Themen (Kunde, Produkt, ...) ist für eine Entscheidungsfindung sehr hilfreich. Die Themenorientierung hilft auch allen Stakeholdern, Daten schnell zu finden und auszuwerten sowie interdisziplinär mittels eines einheitlichen Datenglossars zu diskutieren. Grundlage hierfür ist der themenorientierte, sachlogische Bezug der Daten.
- **Integrated** (Vereinheitlicht): Die Daten in einem DWH sind integriert. Das bedeutet, dass die Datenbestände aus den vielen unterschiedlichen Quellsystemen zusammengeführt und Inkonsistenzen entfernt beziehungsweise korrigiert werden. Auch einheitliche Namenskonventionen, Referenzdaten und physische Datentypen sowie klare Definitionen von Regeln zur Berechnung von KPIs erhöhen den Grad der Integration. Der Datenhaushalt wird somit harmonisiert und ermöglicht zusätzlich zu

den obigen Punkten einen höheren Grad der Automatisierung.

- **Time-Variant** (Zeitorientiert): In einem DWH werden Daten über größere Zeiträume gespeichert als in operativen Systemen. Es ist möglich, für fachliche Themen den historischen Verlauf nachzuvollziehen. Hiermit werden Analysen bezüglich Entwicklungen, Mustern und Vorhersagen unterstützt. Abfragen sind somit über aktuelle Daten (as-is) und über jeden beliebigen Zeitraum in der Vergangenheit (as-was) problemlos möglich.
- **Non-Volatile** (Beständig): Daten, die einmal ihren Weg ins DWH gefunden haben, werden nicht mehr (zum Beispiel vom Anwender) verändert oder gelöscht. In einem operativen System ist dieses möglich, wenn nicht sogar notwendig. Somit sind in einem DWH per Definition alle Änderungen von Daten zu jeder Zeit nachvollziehbar und reproduzierbar.

BORIS VOGT verantwortet den Bereich Business Intelligence & Data Analytics bei der MID GmbH. Er befasst sich seit fast 20 Jahren mit IT-Projekten im Allgemeinen und seit ca. 15 Jahren im Speziellen mit den Schwerpunkten Datenintegration, -migration und -transformation sowie der Modellierung und Architektur von Data-Warehouse-Lösungen und deren Visualisierung. Unterstützt wurde Boris Vogt bei der Erstellung des Artikels von **KAY DÖHLA**, der seit vielen Jahren in IT-Projekten im Data-Warehouse- und BI-Umfeld tätig ist.
E-Mail: b.vogt@mid.de

Big Data

Big Data bezeichnet eine Lösung zur Verarbeitung von sehr großen, komplexen und teilweise semi- beziehungsweise unstrukturierten sowie schnelllebigen Datenmengen. Die gesammelten Daten können dabei aus verschiedensten Quellen stammen, werden größtenteils in Rohform gespeichert und zur Visualisierung, Analyse und Data Mining beziehungsweise Machine Learning verwendet.

In der ursprünglichen Definition von Big Data bezieht sich das „Big“ auf die drei Vs [Lan01] (siehe Abbildung 3):

- **Volume** (Datenvolumen): Sehr große Datenmengen können verarbeitet und gespeichert werden.
- **Velocity** (Geschwindigkeit): Daten werden in kürzester Zeit und mit hoher Geschwindigkeit, nahezu in Echtzeit, verarbeitet.
- **Variety** (Bandbreite der Datentypen und -quellen): Daten unterschiedlichster Typen und Herkunft werden verarbeitet. Sie können strukturiert (relationale Datenbanken), semistrukturiert (CSV, Logs, XML, JSON), unstrukturiert (zum Beispiel Mails, Dokumente wie PDF, DOCX) und binär (Bilder, Video, Audio) sein. Somit ist auch die Anbindung von Systemen wie Social Media, Web Tracker und Web-Suche möglich.

Unterschiede beider Welten

Beide Lösungen haben durchaus Gemeinsamkeiten wie zum Beispiel die Verwaltung großer Datenmengen und deren Verwendung im Reporting. Dennoch gibt es einige wichtige Unterschiede, auf die im Folgenden eingegangen wird (siehe dazu auch den Kasten auf der nächsten Seite).

Datenspeicherung

Beim klassischen DWH werden die Daten in einem relationalen Datenbanksystem gespeichert. Sie werden strukturiert und gruppiert in Tabellen und Spalten gleichen Typs abgelegt. Eine entsprechende Gleichheit der Daten muss vorliegen. Geschwindigkeit und maximale Datenmenge sind vom verwendeten Datenbankhersteller abhängig. Performance-Optimierungen können mittels unterschiedlicher Technologien, zum Beispiel In-Memory

Abb. 2: Grundprinzipien des Data Warehouse

Abb. 3: Grundprinzipien von Big Data

und Massive Parallel Processing (MPP), erzielt werden. Die Erweiterbarkeit solcher Systeme ist jedoch an physische Grenzen gebunden.

Bei Big Data hingegen werden verteilte Dateisysteme (DFS – Distributed File System) verwendet. Diese verarbeiten die Daten, indem sie als kleine Dateien verteilt in einem Rechner-Cluster gespeichert und verwaltet werden. Solche Cluster lassen sich in puncto Geschwindigkeit und Datenmenge beliebig erweitern.

Datenherkunft

Da beim klassischen DWH die Daten strukturiert in einem relationalen Datenbanksystem gespeichert werden, können nur Daten aus Datenquellen verarbeitet werden, die die Daten strukturiert bereitstellen. Dies sind zum Beispiel operative Systeme, die ebenfalls relationale Datenbanksysteme verwenden. Semi- beziehungsweise unstrukturierte Daten können teilweise in strukturierte Daten umgewandelt werden, um sie zu laden. Dies ist jedoch mit erhöhtem Aufwand verbunden. Moderne relationale Datenbanksysteme bieten auch die Möglichkeit, semi- und unstrukturierte Daten ohne Umwandlung in strukturierte Daten zu speichern. Dies ist für manche Anwendungsfälle zwar sinnvoll, widerspricht aber dem Grundsatz der Integrität.

Big Data und die Verwendung der Dateisystemspeicherung bieten die Möglichkeit, alle Formen von Daten zu verarbeiten und zu speichern (strukturiert, semistrukturiert, unstrukturiert, binär).

Orientierung

Im DWH werden die Daten nach fachlichen Themen (Kunde, Produkt, ...) organisiert, gespeichert und integriert.

Bei Big Data orientiert sich die Organisation der Daten an der Datenherkunft beziehungsweise den Quellsystemen.

Integration der Daten

Die Integration der Daten, also die Verknüpfung von Daten unterschiedlicher Herkunft, spielt im klassischen DWH eine wichtige Rolle. Hierzu werden ETL-Prozesse verwendet.

In Big-Data-Systemen ist die Integration der Daten zweitrangig und wird meist nicht durchgeführt. Die Datenbeladung erfolgt mittels ELT- beziehungsweise Data-Ingest-Lösungen.

Datenqualität

Beim DWH sind die Anforderungen an die Datenqualität hoch. Insbesondere wenn es sich um rechtssichere, buchhalterische Berichte handelt, müssen die Daten sehr genau sein. Die Verwendung des DWH als „Single Point of Truth“ setzt hundertprozentige Korrektheit der Daten voraus.

Bei Big Data ist die Anforderung an die Datenqualität deutlich niedriger. Hier arbeitet man mit Werten, die sich wegen der hohen Datenmengen an die Realität annähern. Unschärfen werden in Kauf genommen beziehungsweise statistisch entschärft. Der Businessnutzen der Daten wird über die Qualität gestellt.

Historische Daten

Die großen Datenmengen in einem DWH entstehen vor allem durch das Speichern über große Zeiträume sowie deren Veränderungen. Diese Veränderungen bilden einen hohen Mehrwert für den klassischen DWH-Ansatz sowie deren Analyse.

Bei Big Data wird die Datenmenge durch die Quellsysteme bestimmt. Die Speicherung von historischen Veränderungen ist je nach Datenbasis möglich, aber kein vorrangiges Ziel.

Beständigkeit

In einem DWH sollen keine Daten gelöscht werden. Dies ist unter anderem notwendig, um die Historie der Daten aufrechtzuerhalten.

Bei Big Data besteht diese Anforderung genauso. Allerdings wird sie wegen geringerer Qualität und unstrukturierter Sammlung von Daten sowie meist kürzerer Auswertungszeiträume weniger strikt gehandhabt. Gelegentlich macht ein Löschen der „alten/schlechten“ Daten sogar Sinn, um die Ergebnisse bei Data-Mining- oder Machine-Learning-Prozessen zu verbessern. Über die letzten Jahre haben sich viele Data Lakes in einen wahren Datenfriedhof verwandelt, der aufgrund hoher Datenmengen und damit verbundener steigender Kosten für Speicherplatz zu Datenlöschungen führt.

Datenmodelle

Im DWH werden die Daten strukturiert in einem Datenbanksystem abgelegt. Diese Struktur unterliegt klaren Regeln und ermöglicht es, einheitliche Datenmodelle zu erstellen. Diese Datenmodelle, zum Beispiel Data-Vault- [Lin15] und Star-/Snowflake-Schemata [Kim13], haben sich über die

Gegenüberstellung: Data Warehouse und Big Data

Unterschiede	Data Warehouse	Big Data
Datenspeicherung	Relationales Datenbanksystem oder In-Memory	Distributed File System
Datenherkunft	vorrangig strukturiert; eingeschränkt auch semi- und unstrukturiert sowie binär	strukturiert, semistrukturiert, unstrukturiert, binär
Orientierung	nach fachlichen Themen	nach Datenherkunft
Integration der Daten	per Definition ein Muss-Kriterium	per Definition ein Kann-Kriterium
Datenqualität	per Definition sehr hoch, Single Point of Truth	per Definition nicht relevant, von der Konzeption abhängig und flexibel gestaltbar
Historische Daten	lückenlose Umsetzung per ETL, fehlende Zeiträume werden mit Dummies gefüllt	abhängig von der Konzeption, aufgrund der Flexibilität und Performance im Normalfall nicht lückenlos
Volatilität	niedrig bzw. keine	hoch
Datenmodelle	vollständig umsetzbar	Umsetzbarkeit abhängig von der Konzeption, im Normalfall aufgrund der Flexibilität nicht notwendig
Data Lineage	vollständig umsetzbar	abhängig von der Konzeption, aufgrund der Flexibilität nicht im Fokus

letzten Jahrzehnte etabliert und bieten ausgereifte Lösungen. Weiterhin sind diese Modelle normiert und somit allgemein leicht nachzuvollziehen und zu verstehen.

Da Big Data neben strukturierten auch semistrukturierte, unstrukturierte und binäre Daten beinhaltet, ist eine Modellierung der Strukturen sehr schwierig und mit hohem Aufwand verbunden. Eine umfängliche Datenmodellierung steht der Flexibilität, die Big Data in puncto Quelldaten bietet, entgegen. Ansätze und Lösungen stecken hier noch in den Kinderschuhen, sind aus fachlicher Sicht im Normallfall aber auch nicht notwendig.

Data Lineage

Die klaren Regeln bei der Modellierung im klassischen DWH erlauben es, über die Metadaten die Abhängigkeiten der Objekte und den Datenfluss nachzuvollziehen. Dies kann für Anpassungen am System sowie Audit-Fähigkeiten hilfreich sein. Die Data Lineage kann hierbei auch den Informationsraum der Daten verlassen und zum Beispiel Prozesse inkludieren. Sie ist für einige regulatorische Anforderungen, wie BCBS239 (Basel Committee on Banking Supervision's standard number 239 [Bas13]), unumgänglich.

Da eine Modellierung in Big Data eher hinderlich ist, fehlt die Möglichkeit, solche Abhängigkeiten und Datenflüsse in der Gesamtheit der Daten zu dokumentieren und nachzuvollziehen. Mit speziellen Methoden (Machine Learning etc.) kann man näherungsweise Rückschlüsse auf Zusammenhänge und Abhängigkeiten ermitteln.

Nutzung und Zielrichtung

Die Entscheidung zur Nutzung einer klassischen DWH-Welt oder einer modernen Big-Data-Welt hängt maßgeblich von der Zielrichtung ab (vgl. Abbildung 4). Im Folgenden werden unterschiedliche Zielrichtungen von beiden Welten aus dem praktischen Erfahrungsschatz vieler Kundenprojekte der Autoren diskutiert.

Motivation

Sowohl im DWH als auch in der Big-Data-Welt können dem Endbenutzer Daten zur Verfügung gestellt werden.

Im DWH sind folgende Zielrichtungen optimal umsetzbar:

- Standardisierte Auswertungen über Zeitreihen, das heißt, es können x-beliebige Historien abgebildet werden
- Drillfunktionalitäten, um mittels Slice & Dice in Daten hinein beziehungsweise heraus zu drillen, das heißt, die Granularität der Daten kann per Knopfdruck geändert werden
- Harmonisierte Datenbestände mit sehr hoher Datenqualität, das heißt, mittels eines sauberen Daten-Glossars sind alle Attribute und Metriken unternehmensweit eindeutig und die Kommunikation der Stakeholder wird verbessert
- Planungen können auf verlässlichen Datenbeständen durchgeführt werden

Abb. 4: Zielrichtung als Entscheidungsgrundlage

Im Gegenzug ist das Big-Data-Umfeld auf folgende Zielrichtungen ausgerichtet:

- Anforderungen der Künstlichen Intelligenz und des Machine Learning sind einfacher zu produzieren und abzubilden
- x-beliebige Daten können flexibel und einfach im Data Lake ergänzt werden
- Data Mining und Mustererkennung sind nicht nur auf harmonisierte Daten, sondern flexibel auf jeglichen zugänglichen Daten möglich
- Data Science wird optimal unterstützt

Gesetzliche Motivation

Im DWH sind grundsätzlich regulatorische und gesetzliche Anforderungen und Vorschriften umsetzbar, zum Beispiel die Vorschrift BCBS239 im Bankensektor. Darin geforderte Stresstests im Risikobereich mit unterschiedlichen Parametern für Szenarien sind über Zeitreihen darstell- und vergleichbar und zu beliebigen Zeitpunkten in der Zukunft mit gleichen Ergebnissen reproduzierbar. Ebenso ist die allgemein gültige Datenschutz-Grundverordnung (DSGVO) umsetzbar. So können die Anonymisierung von Daten automatisiert gewährleistet sowie die entsprechenden Verpflichtungen zur Datenkorrektur beziehungsweise Löschung sichergestellt werden.

Diese Anforderungen können derzeit in einer reinen Big-Data-Lösung nicht vollständig abgebildet werden.

Technische Nutzung

Klassische DWHs haben jahrelang technische Lösungen über die gesamte Wertschöpfungskette von der allgemeinen Datenhaltung über die Datentransformation bis hin zur Datenvisualisierung perfektioniert und standardisiert. Bei der Datenhaltung stehen vor allem relationale Datenbanken und In-Memory-Datenbanken im Fokus. Die Datentransformation läuft mittels ausgereifter ETL-Tools. Die Visualisierung wird mittels Frontend-Tools spezialisiert auf Abteilungs- beziehungsweise Enterprise-Anforderungen erfüllt.

Big-Data-Lösungen unterliegen derzeit noch einem starken Wandel und diversen Innovationen. Aktuell sind in der Datenhaltung Lösungen von NoSQL-Datenbanken, zum Beispiel Graph-Datenbanken oder dokumentenbasierte Datenbanken,

im Fokus. Der Datentransfer wird mittels ELT- beziehungsweise Streaming-Diensten vorgenommen. Eine Visualisierung erfolgt entweder mit Frontend-Werkzeugen des DWH-Bereichs beziehungsweise mit speziellen Werkzeugen für die genannten fachlichen Zielrichtungen. Diese Werkzeuge sind auf Programmiersprachen wie R, PMML beziehungsweise Python optimiert. Hierdurch wird eine maximale Flexibilität erzielt.

Sicherheit und Berechtigungen

DWHs sind aufgrund ihrer jahrelangen Erprobung auf Stabilität, Ausfallsicherheit und Benutzerberechtigungen mittlerweile sehr ausgefeilt. Es können nahezu alle Anforderungen abgebildet und sichergestellt werden.

Moderne Big-Data-Lösungen unterliegen derzeit hinsichtlich Infrastruktur, Architektur und Software einem rapiden Wandel. Lösungsansätze sind zu vielen Punkten vorhanden, aber bisher nicht ausgereift. Sie liegen noch hinter dem Stand bewährter DWH-Lösungen, bieten aber in puncto Flexibilität und Veränderungsmöglichkeit wesentlich mehr Spielraum.

Maximaler Mehrwert durch DWH und Big Data

Im Vergleich der beiden Lösungsansätze lassen sich folgende Punkte zusammenfassen: Alle Anforderungen eines Geschäftsbetriebs, die harmonisierte, standardisierte, automatisierte und revisionssichere Lösungen erfordern, sind mittels DWH umzusetzen. Dies sind zum Beispiel Controlling-, Planungs- und Steuerungsanforderungen. Anforderungen, die einen empirischen Charakter aufweisen, sind eher mittels Big Data zu realisieren. Use Cases sind hier verstärkt im Marketing und der Produktqualität vorzufinden.

Nach aktuellem Stand sind DWH- und Big-Data-Lösungen nicht in der Lage, sich gegenseitig zu ersetzen. Die naheliegende Frage, ob ein Unternehmen lieber auf DWH oder Big Data setzen sollte, greift allerdings zu kurz. Beide Systeme bieten, wie skizziert, je nach vorliegenden Voraussetzungen und konkreter Zielrichtung ihre Vor- beziehungsweise Nachteile. Ein Unternehmen wird deshalb seine Daten nur dann optimal für alle geschäftlichen Belange nutzen können, wenn es sowohl auf DWH- als auch auf Big-Data-Lösungen setzt. Im Zuge der Digitalisierung sind somit alle Unternehmen gefordert, eine Koexistenz beider Lösungen aufzubauen, um weiterhin am Markt konkurrenzfähig und erfolgreich agieren zu können (vgl. Abbildung 5). Die Kombination der beiden Welten schafft aus Daten und dem daraus zu ziehenden Wissen den maximalen Mehrwert, der durch ein Entweder-Oder nicht erreichbar ist.

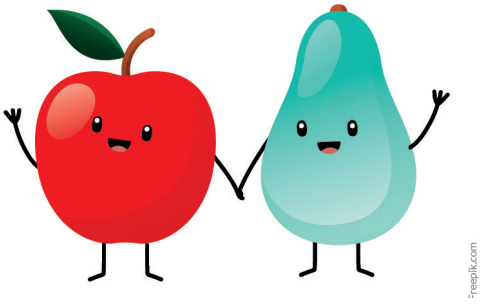


Abb. 5: Mehrwert durch Koexistenz

Fazit

Im Jahr 2014 schrieb William Inmon einen Blog-Eintrag: „Big Data or Data Warehouse? Turbocharge your Porsche – buy an Elephant." Hier bemängelte er unter anderem die Vermischung von Architektur (Data Warehouse) und Technologie (Big Data). Ralph Kimball hingegen sprach sich für den Einsatz von Hadoop als Data-Warehouse-Plattform aus [Wel15]. Die unterschiedliche Haltung der beiden Data-Warehouse-Päpste sowie ihre dogmatische Haltung zu ihren Ansätzen ist aus Sicht der Autoren hier zu kurz gesprungen.

In der heutigen Welt geht es im Grunde schon gar nicht mehr um Dogmen wie Data Warehouse, Big Data oder Data Science. Es geht vielmehr um eine Enterprise Data Intelligence! Damit ist gemeint, dass sämtliche verfügbaren Daten innerhalb eines Unternehmens sowie externe Daten, ja selbst Prozessmanagement- und Anforderungsmanagementdaten einen Mehrwert darstellen. Somit sind zwar alle Ansätze der „Päpste" eine Hilfestellung für Unternehmen – wie diese Daten allerdings in Wissen verwandelt werden können, das „Warum" und „Wie", ist immer auf die speziellen Bedürfnisse eines jeden Unternehmens anzupassen. Je nach Use-Case ist somit der eine oder der andere Ansatz beziehungsweise eine Mischung die passende Lösung. Unter innovativen Gesichtspunkten kann somit nur festgehalten werden: Es lebe die Koexistenz, und man darf gespannt sein, unter welchem Begriff dieser disruptive Umgang mit Daten zukünftig Einzug in die Unternehmen und die Literatur hält.

Literatur

[Bas13] Basler Ausschuss für Bankenaufsicht: BCBS239 – Grundsätze für die effektive Aggregation von Risikodaten und die Risikoberichterstattung. 2013, https://www.bis.org/publ/bcbs239_de.pdf, abgerufen am 12.4.2019

[Inm05] Inmon, W. H.: Building the Data Warehouse. 4. Aufl., Wiley 2005

[Kim13] Kimball, R.: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3. Aufl., Wiley 2013

[Lan01] Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety. In: Application Delivery Strategies, published by META Group Inc. File 949, 2001

[Lin15] Linstedt, D.: Building a Scalable Data Warehouse with Data Vault 2.0. Morgan Kaufmann 2015

[Wel15] Welker, P.: Big Data oder Data Warehouse. 24.6.2015, https://www.computerwoche.de/a/big-data-oder-data-warehouse,3092517, abgerufen am 12.4.2019