BI Spektrum
tdwi WISSEN
ONLINETHEMENSPECIAL

Fachartikel

Methoden und Techniken für die agile Integration komplexer Daten

von Jens Albrecht

Die Bereitstellung einer integrierten Sicht auf Daten aus verschiedenen Quellen ist ein Prozess, der meist hohen manuellen Aufwand für die Sicherung der Datenqualität erfordert. Jede Quelle muss neu angebunden, jedes Feld mit dem bestehenden Datenbestand abgeglichen werden. Im Big-Data-Umfeld ist ein solcher Aufwand sowohl aus finanziellen als auch aus zeitlichen Gründen meist nicht leistbar. Dabei ist weniger das Datenvolumen das Problem, sondern vielmehr die Anzahl und Heterogenität der Datenquellen. Darüber hinaus müssen Daten für die Überprüfung von Hypothesen und die Erstellung von Modellen (Data Science / Machine Learning) nicht zwangsweise periodisch bereitgestellt werden. Stattdessen ist die Integrationsaufgabe häufig nur einmal dediziert für eine bestimmte Analyse durchzuführen. Der Aufwand in fachlicher und technischer Hinsicht sollte daher der Aufgabe angemessen sein. Es geht dabei vor allem um eine effektive Balance zwischen Automatisierung und Datenqualität.

Dieser Artikel gibt einen Überblick über die Herausforderungen und Lösungsansätze für die anwendungsbezogene Integration großer heterogener Datenbestände.

Datenintegration

Unter Datenintegration versteht man den einheitlichen, standardisierten Zugriff auf verschiedene autonome und heterogene Datenquellen [DoH12]. Die klassische Inkarnation eines Datenintegrationssystems ist ein Data Warehouse (DWH). Über einen aufwendigen ETL-Prozess werden Daten aus unterschiedlichsten Quellen in eine einzige Datenbank integriert, damit sie dann einheitlich zusammen analysierbar sind.

Datenintegration ist grundsätzlich anspruchsvoll, da die Quellen meist unabhängig voneinander entwickelt werden und sich zusätzlich auch strukturell über die Zeit verändern. Je stärker Unternehmen digitalisiert und datengetrieben arbeiten, desto mehr Datenquellen entstehen und müssen angebunden werden. Die durch die Digitalisierung getriebene Dynamik bei der Entstehung und die damit erforderliche Anbindung von Quellen macht Big-Data-Integration besonders herausfordernd. Folgende Faktoren wirken dabei als Treiber:

Mehr Datenquellen: Neben den klassischen ERP-Systemen, die üblicherweise ein DWH speisen, kommen eine Vielzahl interner und externer Datenquellen hinzu. Das Spektrum reicht von Sensoren an Maschinen über Logfiles bis hin zu Dokumenten jeglicher Art aus dem Intra- und Internet.

Verschiedenste, komplexe Datenformate: Nachdem die Daten nicht nur aus Datenbanken stammen, liegen sie in einer Vielzahl unterschiedlichster Formate und Strukturen vor. Erklärende Metadaten sind meist nicht vorhanden.

Hohe Änderungsgeschwindigkeit: Die Datenquellen selbst, die Auswahl der Datenquellen und die Verwendung bzw. der

Bedarf für die Daten sind häufigen Änderungen unterworfen.

Agilität: Ein potenzieller Vorteil der Digitalisierung ist eine Steigerung der Flexibilität, das heißt der Möglichkeiten eines Unternehmens, auf veränderte Rahmenbedingungen zu reagieren. Dafür müssen Daten zügig erschlossen und verfügbar gemacht werden.

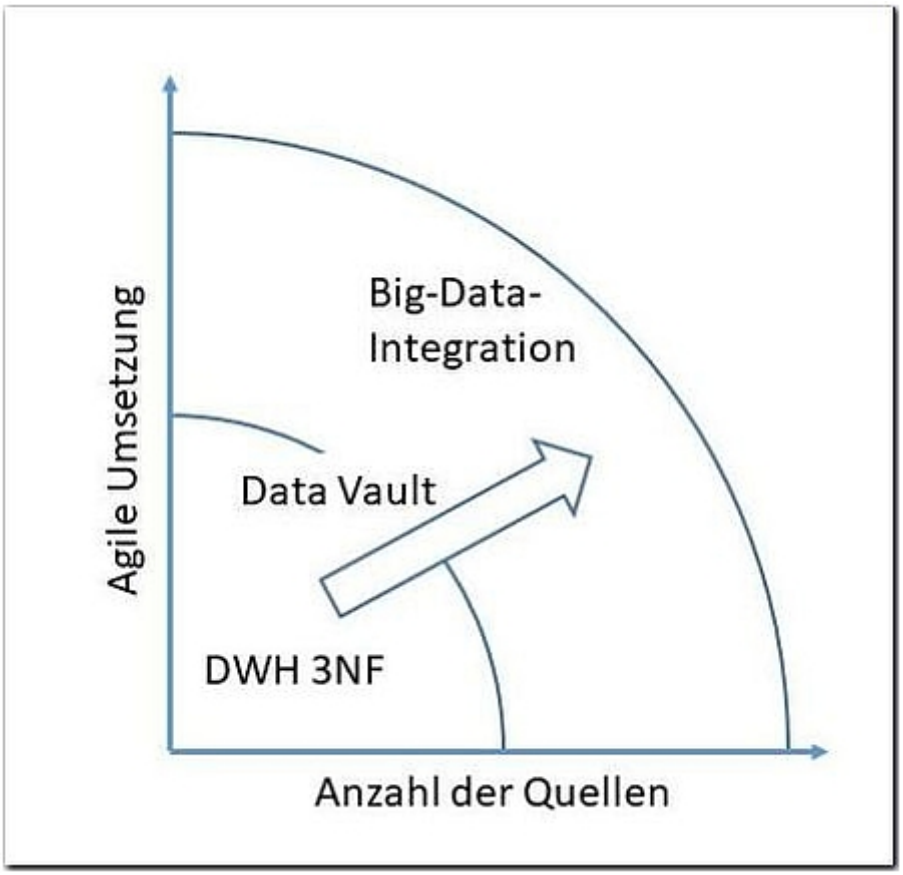

Abb. 1: Big-Data-Integration

Eine wachsende Zahl von Datenquellen steht also der Anforderung nach schnellerer Datenbereitstellung gegenüber. Das geht nur durch Automation. Dabei sollte die Datenqualität einerseits möglichst hoch sein, andererseits aber möglichst wenig menschliche Eingriffe erfordern. Wer die gleiche Datenqualität wie im DWH haben möchte, muss in der Regel auch den gleichen Aufwand treiben. Vielfach geht es aber bei Big-Data-Analysen nicht um präzise Aussagen („Wie hoch war der Umsatz letzte Woche?“), sondern um Stimmungen, Meinungen und Trends („Was bewegt unsere Zielgruppe?“). Insofern müssen die Daten nur hinreichend gut integriert werden, um eine valide und relevante Aussage zu treffen.

Quellen identifizieren und bewerten

Bevor überhaupt integriert werden kann, muss zunächst einmal bekannt sein, welche Daten wo in welcher Form verfügbar und relevant für die Fragestellung sind. Was sich trivial anhört, ist angesichts der schier unendlichen Informationsflut im Web oder Tausender unterschiedlichster Dateien im Data Lake gar nicht so einfach zu klären. Rein rechtlich ist gegebenenfalls zu prüfen, ob die Daten überhaupt verwendet werden dürfen. Datenschutz- und Urheberrechtsfragen sowie ethische Aspekte könnten dem entgegenstehen.

Bei externen Datenquellen, wie sie im Big-Data-Umfeld häufig zum Einsatz kommen, ist darüber hinaus zu klären, welche Informationen überhaupt aus den Daten abgeleitet werden können. Welche Kennzahlen können beispielsweise einem Jahresbericht oder einer Pressemeldung entnommen werden? Worauf bezieht sich eine Umsatzzahl? Wesentlich sind auch Aussagen zur Vertrauenswürdigkeit (Konfidenz). Sind die Daten aussagekräftig genug für den Anwendungsfall?

Werden beispielsweise Kommentare in einem Forum analysiert, so spiegeln sich hier nur die Meinungen derjenigen, die in dem Forum Beiträge schreiben. Das sind in der Regel deutlich weniger Personen, als Beiträge lesen, und sie sind nur für eine sehr spezielle Domäne repräsentativ, wenn überhaupt. Das ist besonders wichtig, wenn maschinelle Lernverfahren in den integrierten

Daten nach Mustern suchen oder basierend darauf Vorhersagen machen sollen, denn sie sind besonders anfällig für Daten-Schieflagen. Die Ergebnisse sind zudem schwer nachvollziehbar, sodass schlechte Modelle aufgrund schlechter Trainingsdaten leicht für gute gehalten werden können.

Datenkatalogisierung

Insbesondere im Big-Data-Umfeld wurde schon vor einiger Zeit der Bedarf erkannt, diese Prozesse zu unterstützen. Data Catalogs sind Werkzeuge zur Verwaltung heterogener Datenbestände, wie es zum Beispiel in Data Lakes erforderlich ist. Dabei werden neue Datenquellen automatisch auf Datensatz-Ebene durchsucht, Beziehungen zu bekannten Datenquellen werden aufgezeigt, alles wird verschlagwortet und durchsuchbar gemacht. Werteverteilungen, Trends und Beziehungen werden weitgehend automatisch ermittelt. Die Vertrauenswürdigkeit der Daten kann über Team Collaboration ermittelt und dokumentiert werden. Der gesamte Prozess zur Aufnahme von Daten in den Bestand und für das Wiederfinden wird so stark vereinfacht und unterstützt.

Als Beispiele für kommerzielle Lösungen können beispielsweise IBM (Watson Knowledge Catalog) und Oracle (Oracle Enterprise Metadata Management) genannt werden, Newcomer im Hadoop-Umfeld sind Waterline oder Cloudera (Navigator). Eine aktuelle Marktübersicht wurde kürzlich von Forrester erstellt [For18].

Technische Integration

Sind die Rahmenbedingungen geklärt, kann die technische Integration der Systeme erfolgen. Dafür muss ein Zugriff auf verschiedene Datenbestände, die in unterschiedlichen Formaten auf unterschiedlichen Systemen vorliegen, realisiert werden.

Physische Datenintegration
Klassischerweise werden Daten für die gemeinsame Analyse aus Quellsystemen extrahiert und in ein integriertes System (DWH) kopiert. Die gängigen ETL-Plattformen bieten dafür eine Vielzahl von Konnektoren zu all diesen Plattformen, einschließlich Hadoop und NoSQL. Der Zeitaufwand sowohl für die Implementierung des Prozesses als auch für das periodische Laden der Daten selbst ist allerdings sehr hoch. Dafür setzen nachgelagerte Analysen auf einem sauberen, dedizierten Datenbestand auf, der unabhängig von den Quellsystemen gepflegt und optimiert werden kann.

Daten-Virtualisierung
Alternativ kann eine virtuelle Integrationsschicht eingesetzt werden, die den transparenten Zugriff, zumeist mit SQL, auf die Datenquellen ermöglicht und eine technische Integration on-the-fly vornimmt. Haupttreiber für den Einsatz von Daten-Virtualisierungslösungen ist die Möglichkeit, schneller auf neue Use Cases zu reagieren und Anforderungen zeitnah umzusetzen.

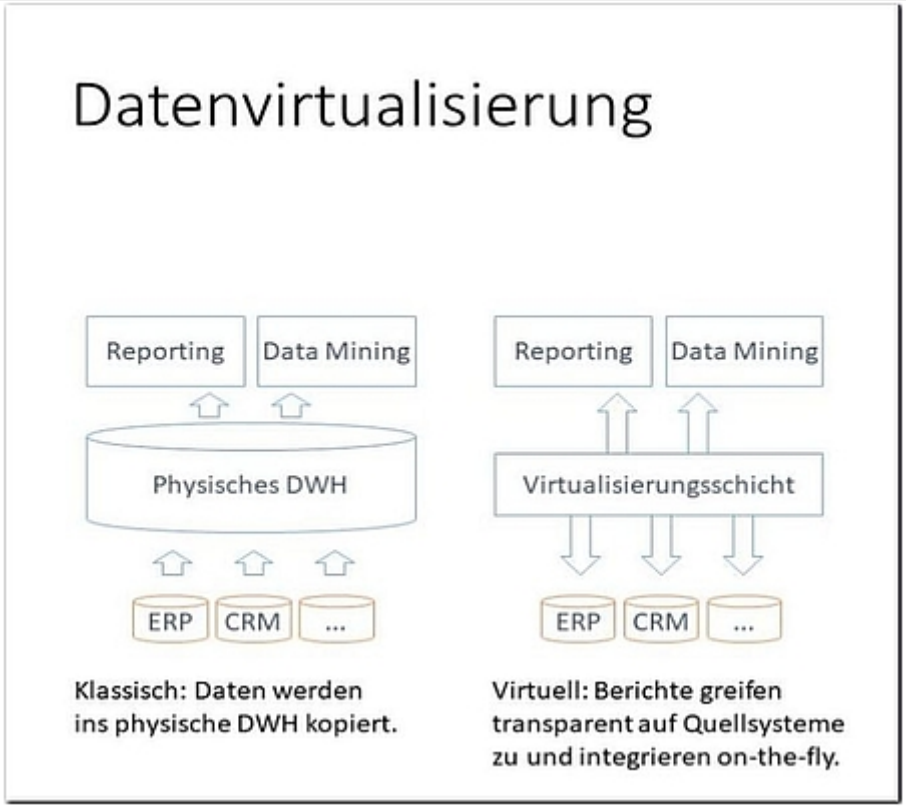

Abb. 2: Physische Datenintegration vs. Daten-Virtualisierung

Einige BI-Systeme ermöglichen direkt den leichtgewichtigen Zugriff auf heterogene Quellen, entweder direkt im Bericht oder im BI-Server-Layer. Die Nutzung der integrierten Daten ist allerdings auf diese Weise eingeschränkt. Für komplexe analytische Plattformen ist es daher sinnvoller, eine Daten-Virtualisierungsschicht auf Ebene (wenn es nach den Anbietern ginge sogar anstatt) des Enterprise Data Warehouse einzuziehen (siehe Abbildung 2).

Ein solches virtuelles Data Warehouse stellt eine logisch integrierte Sicht auf die Daten bereit, physisch verbleiben sie aber in den Quellsystemen. Erst beim Zugriff werden die Daten geladen. Um die Quellsysteme zu entlasten und die Performance zu steigern, können Daten auch gecachet werden. Anders als im Data Warehouse werden die Daten aber nicht dauerhaft materialisiert, sondern bei Platzbedarf wieder verworfen.

Es gibt hier eine ganze Reihe von Anbietern, die inzwischen ausgereifte Lösungen für Enterprise Data Virtualisation anbieten [For17]. Darunter finden sich die großen ETL-Anbieter ebenso wie einige interessante Newcomer. Ist bereits eine Big-Data-Plattform im Einsatz, kann Apache Spark eine kostengünstige Alternative zu einer teuren Virtualisierungssoftware sein. Spark bietet von Hause aus die Möglichkeit, auf verschiedenste Datenquellen mit SQL zuzugreifen. Dadurch können mit minimalem Aufwand Daten aus beispielsweise einer relationalen und einer NoSQL-Datenbank sowie einer Text-Datei auf SQL-Ebene verknüpft werden. Die Bereitstellung dieser integrierten Sicht nach außen hin, das heißt über die Anwendung innerhalb eines einzelnen Ladeprozesses zur Datenaufbereitung hinaus, ist allerdings noch nicht out-of-the-box möglich und erfordert einigen Konfigurationsaufwand ohne die gängige visuelle Unterstützung.

Inhaltliche Integration

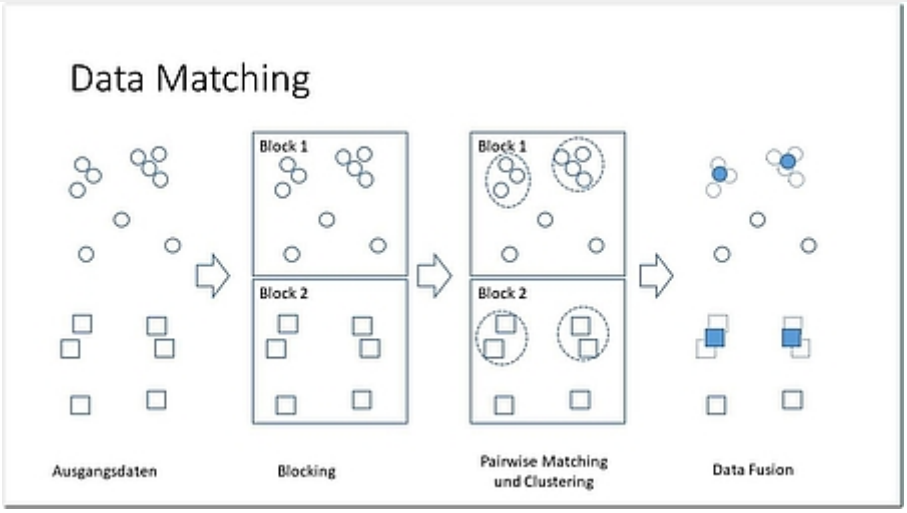

Abb. 3: Phasen der Daten-Integration: Schema Matching, Data Matching und Data Fusion

Ist die technische Integration vollzogen, beginnt die schwierige Aufgabe der inhaltlichen Integration. Dabei können drei Phasen unterschieden werden [DoS13]: Schema Matching, Data Matching und Data Fusion (siehe Abbildung 3).

Schema Matching

Unter der Schema-Angleichung (Schema Alignment, Schema Matching, Schema Integration) versteht man die Überführung der heterogenen Quellschemata in eine einheitliche Darstellung. Es geht hierbei also um eine Harmonisierung der Datenstrukturen als Voraussetzung für Abfragen und Analysen auf dem integrierten Datenbestand.

Im klassischen DWH-Ansatz wird ein integriertes Schema explizit durch den Menschen modelliert. Hier wird eher ein Top-down-Ansatz gefahren, das heißt, ausgehend vom Informationsbedarf (Berichte) wird ein DWH mit einem einheitlichen Schema konzipiert und modelliert. Dabei steht von vornherein ein ganzheitlicher Ansatz im Vordergrund: Ziel ist eine weitgehende Integration aller Daten.

Im Big-Data-Szenario wird hingegen eher Bottom-up, das heißt datengetrieben, nach einem Pay-as-you-go-Ansatz vorgegangen. Es wird nur so weit integriert, wie es gerade benötigt wird. Ein Data Lake unterstützt dieses Vorgehen, da Daten relativ problemlos im Data Lake abgelegt werden können, ohne dass zwangsläufig eine Integration erforderlich wird.

Die Erstellung eines integrierten („mediated“) Schemas kann in komplexen Umgebungen algorithmisch unterstützt werden. Die Attribute der Quelldaten werden dabei in Bezug auf Namen und Wertebereiche verglichen, um Ähnlichkeiten zu ermitteln. Ein „Exact Match“ ist dabei aber eher selten. Meist stimmen die Wertebereiche nicht ganz überein oder Attribute müssen gar zerlegt bzw. zusammengefasst werden (zum Beispiel PLZ und Ort). Auf wissenschaftlicher Seite gibt es eine Vielzahl von Veröffentlichungen in diesem Bereich [ShE05], die allerdings kaum den Weg in kommerzielle Produkte gefunden haben.

Data Matching bzw. Entity Resolution

Der wichtigste und schwierigste Schritt bei der Datenintegration ist das Data Matching, das heißt das Auffinden zusammengehöriger Datensätze aus verschiedenen Datenbanken, die sich auf die gleichen Objekte bzw. Entitäten beziehen [Chr12; EIV07; GeM13] (siehe Abbildung 4). In diesem Zusammenhang tauchen verschiedene Begriffe auf, die zum Teil auch synonym verwendet werden:

Entity Resolution thematisiert die Zuordnung von Daten aus unterschiedlichen Quellen zu einem Objekt der realen Welt.

Deduplikation, das heißt die Erkennung und Eliminierung von Kopien des gleichen realen Objekts (zum Beispiel „Bundeskanzlerin Merkel“, „Angela Merkel“), behandelt Entity Resolution innerhalb einer Datenbank.

Record Linkage bezeichnet die Verknüpfung von Datensätzen aus verschiedenen Quellen, die sich auf das gleiche reale Objekt beziehen (zum Beispiel Befragungsdaten aus einem Online-Survey mit Kundenstammdaten).

Disambiguierung ist ein Prozess, der ähnliche Repräsentationen, die zu unterschiedlichen Objekten der realen Welt gehören, voneinander abgrenzt (Beispiel: „Samsung Galaxy Note“ vs. „Samsung Galaxy Note 9“ oder „Alexander Schmidt“ vs. „Alexandra Schmidt“).

Name	Straße	Ort	Tel.	Versicherungsnummer
Peter Meier	Schulstr. 1b	80796 München	089-4624-4342	KV-2415633265

Datensatz 1: Kundenstamm bei Versicherung

Name	Ort	Tel.	Art des Anrufs
Peter F. Meier	80796 München	+498946244342	Anfrage

Datensatz 2: Anruf beim Customer Service

Abb. 4: Probabilistisches Data Matching – eine exakte Zuordnung der Datensätze ist nicht möglich, daher muss anhand der Ähnlichkeit der Feldinhalte ein Score für die Wahrscheinlichkeit der Gleichheit ermittelt werden

Algorithmische Verfahren zum Data Matching nutzen Techniken des maschinellen Lernens, um die Ähnlichkeiten von Objekten zu erkennen. Dabei werden die Schritte Blocking, Pairwise Matching und Clustering unterschieden (siehe Abbildung 3) [DoS13].

Blocking

Algorithmen für Duplicate Detection oder Record Linkage haben eine quadratische Komplexität, da jedes Objekt mit jedem anderen abgeglichen werden muss (kartesisches Produkt). Eine häufig eingesetzte Methode, um die Effizienz der Verfahren zu steigern, ist Blocking. In einem ersten Schritt werden die Daten in disjunkte Blöcke aufgeteilt, danach werden die Duplikate nur noch innerhalb eines Blocks gesucht [EIV07] (siehe Abbildung 3). Allerdings muss sichergestellt werden, dass keine zwei Duplikate in unterschiedlichen Blöcken liegen. Häufig lassen sich solche Blöcke aber leicht identifizieren, zum Beispiel über Regionen, Sprachen, Schlüsselkreise, Quellsysteme.

Pairwise Matching

Im zweiten Schritt wird jetzt innerhalb jedes Blocks (das heißt jeder Datenpartition) der Abstand bzw. die Ähnlichkeit zwischen allen Objektpaaren berechnet. Mögliche Ähnlichkeitsmaße sind aus dem Information Retrieval und dem Data Mining bekannt. Häufig wird der Grad der Übereinstimmung auf Text- oder Attribut-Ebene gemessen (zum Beispiel Levenshtein- oder Jaccard-Distanz, TF-IDF mit Cosinus-Ähnlichkeit etc.) [EIV07].

Neue Verfahren basieren auf Deep Learning [Ebr18]. Dabei wird die Methode der Word-Embeddings auf Datensätze übertragen. Verfahren für Word-Embeddings wie word2vec oder Glove transformieren Worte oder Wortgruppen in einen numerischen Vektor, sodass ähnliche Worte auch durch ähnliche Vektoren abgebildet werden (siehe [Col16] für einen Überblick). Ähnlichkeit wird dabei über den Kontext bestimmt, das heißt zwei Wörter sind ähnlich, wenn sie in ähnlichen Kontexten verwendet werden. Dadurch können sogar Aussagen wie „king – queen + woman = man“ aus den Daten abgeleitet werden. Der Vorteil dieser Verfahren ist, dass eine semantische Ähnlichkeit ermittelt werden kann, ohne dass ein Mensch dafür die Regeln definieren muss.

Clustering

Im letzten Schritt werden die Objekte, die einen sehr geringen Abstand haben, zu Äquivalenzklassen zusammengefasst. Wenn Trainingsdaten vorhanden sind, ist das Auffinden von „matching records“ ein Klassifikationsproblem (ein Paar von Datensätzen ist ein Match oder nicht), das mit einem überwachten Lernverfahren gelöst werden kann. Nachdem Trainingsdaten aber im Big-Data-Umfeld meist fehlen, kommen häufig Varianten unüberwachter Lernverfahren zum Einsatz, wobei die Anzahl der Cluster in der Regel sehr hoch ist (ein Cluster pro Real-World-Entity) und sehr viele Cluster nur aus einem Datensatz (Singleton) bestehen. Hinzu kommt, dass ein Cluster-Repräsentant ermittelt werden muss, das heißt aus den Informationen aus unterschiedlichen Quellen wird eine kanonische Darstellung der Entität mit maximaler Information synthetisiert.

Data Fusion

Nach dem Data Matching müssen im letzten Schritt der inhaltlichen Integration die (lückenhaften) Datensätze aus den verschiedenen Quellen in eine einheitliche Repräsentation, das heißt eine möglichst detailreiche kanonische Form der Entität überführt werden (siehe Abbildung 3). Sofern die gleiche Information (zum Beispiel Adresse) mehrfach und gegebenenfalls inkonsistent vorliegt, ist eine Bewertung der Qualität und Vertrauenswürdigkeit der Datenquellen erforderlich.

Zusammenfassung

Datenintegration im Big-Data-Umfeld ist eine komplexe Aufgabe, die aufgrund der Anzahl und Heterogenität der Quellen und der

erforderlichen Agilität bei der Umsetzung kurzfristiger Datenanforderungen weitgehend automatisiert und werkzeugunterstützt ablaufen muss. Die Anbieter von Datenintegrationslösungen bieten hier bereits für viele Teilaufgaben Unterstützung an. Neue Werkzeuge zur Datenkatalogisierung erleichtern das Auffinden und Matching von bereits katalogisierten Datenbeständen im Data Lake. Der manuelle Aufwand ist dennoch weiterhin hoch. Und insbesondere die bei Big Data besonders wichtige Bewertung der Validität und Vertrauenswürdigkeit der Daten bleibt auf absehbare Zeit dem menschlichen Experten vorbehalten.

Literatur

[Chr12] Christen, P.: Data Matching. Springer 2012, ISBN 978-3-642-31163-5, http://link.springer.com/10.1007/978-3-642-31164-2

[CHT18] Chen, C. / Halevy, A. / Tan, W.-C.: BigGorilla : An Open-Source Ecosystem for Data Preparation and Integration. In: Data Engineering Bulletin, 2018

[Cle18] Clearpeaks.com: Data Quality with Informatica – Part 3: Data Deduplication. https://www.clearpeaks.com/data-deduplication-with-informatica/

[Col16] Colyer, A.: The amazing power of word vectors. the morning paper (Blog) 2016, blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

[DoS13] Dong, X. L. / Srivastava, D.: Big data integration. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), 2013, ISBN 978-1-4673-4910-9

[EIV07] Elmagarmid, A. K. / Ipeirotis, P. G. / Verykios, V. S.: Duplicate record detection: A survey. In: IEEE Transactions on Knowledge and Data Engineering, 2007, ISBN 1041-4347

[Ebr18] Ebraheem, M. / Thirumuruganathan, S. / Joty, S. / Ouzzani, M. / Tang, N.: Distributed Representations of Tuples for Entity Resolution. In: Proceedings of the VLDB Endowment (PVLDB), Bd. 11 (2018), Nr. 11

[For17] Forrester Research: The Forrester WaveTM: Enterprise Data Virtualization, Q4 2017. www.forrester.com/report/The+Forrester+Wave+Enterprise+Data+Virtualization+Q4+2017/-/E-RES133042, abgerufen am 14.10.2018

[For18] Forrester Research: Now Tech: Machine Learning Data Catalogs, Q1 2018. www.forrester.com/report/Now+Tech+Machine+Learning+Data+Catalogs+Q1+2018/-/E-RES137445

[Fra17] Franczuk, S.: Data Matching 101: What Tools Does Talend Have? Talend Blog, März 2017, www.talend.com/blog/2017/03/13/data-matching-101-tools-talend/, abgerufen am 14.10.2018

[GeM13] Getoor, L. / Machanavajjhala, A.: Entity resolution for big data. In: KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, ISBN 9781450321747

[Kon16] Konda, P. / Das, S. / Suganthan, P. G. C. / Doan, A. / Ardalan, A. / Ballard, J. R. / LI, H. / Panahi, F. u. a.: Magellan: Toward Building Entity Matching Management Systems. In: Proceedings of the Vldb Endowment, 2016

[Mud18] Mungal, S. / Li, H. / Rekatsinas, T. / Doan, A. / Park, Y. / Krishnan, G. / Deep, R. / Arcaute, E. u. a.: Deep Learning for Entity Matching. In: Proceedings of the 2018 International Conference on Management of Data – SIGMOD '18, 2018 — ISBN 9781450347037

[ShE05] Shvaiko, P. / Euzenat, J.: A Survey of Schema-Based Matching Approaches. In: Spaccapietra, S. (Hrsg.): Journal on Data

Semantics IV., Berlin, Heidelberg 2005, ISBN 978-3-540-31447-9, S. 146–171

Prof. Dr. Jens Albrecht

ist an der Fakultät Informatik der Technischen Hochschule Nürnberg für das Lehrgebiet Datenbanken und Big Data verantwortlich. Seine Interessensgebiete umfassen Big-Data-Technologien und maschinelles Lernen, insbesondere im Kontext Natural Language Processing. Bevor Prof. Albrecht 2012 an die Hochschule wechselte, arbeitete er als DWH-Architekt und IT-Manager bei Oracle und der GfK in Nürnberg.

E-Mail: jens.albrecht(at)th-nuernberg.de

Bildnachweis: