Editorial

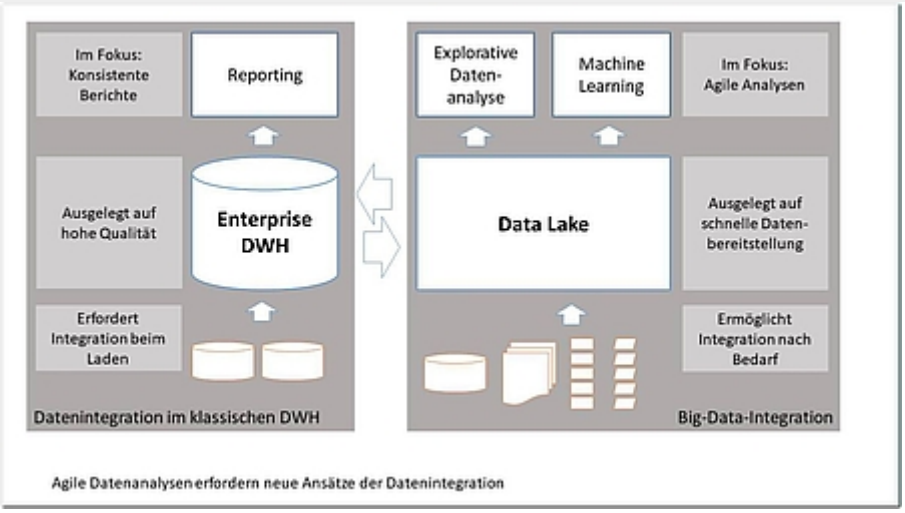
Big Data Integration

von Jens Albrecht

Auch wenn der Begriff „Big Data“ marketingtechnisch schon etwas abgegriffen ist: Ohne Daten geht heutzutage nichts mehr, sie sind der vierte Produktionsfaktor. Um die Nase vorn zu haben, müssen sich Unternehmen daher immer schneller neue Datenquellen erschließen. Insbesondere maschinelle Lernverfahren, die Basis für intelligente, autonome Systeme, sind sehr datenhungrig. Je umfangreicher die Informationen sind, desto genauer lassen sich datenbasierte Prognosen erstellen und Muster erkennen. Die angestrebte ganzheitliche Sicht auf einen Prozess wird jedoch meist erst durch die Verknüpfung verschiedener Daten, also deren Integration, ermöglicht.

Die Notwendigkeit zur Datenintegration war der Grund für die Entstehung für Data Warehouses in den 90er-Jahren. Aber die klassischen Unternehmensdaten aus den Bereichen Finanzen, Vertrieb, Produktion etc. allein reichen heutzutage nicht mehr aus. So müssen für den 360-Grad-Blick auf Kunden Daten aus verschiedensten Quellen, wie beispielsweise Call-Centern, Weblogs oder User-Foren, miteinander verwoben werden.

Das Data Warehouse mit dem „Single Point of Truth“-Ansatz ist hierfür nur bedingt geeignet. Denn Datenintegration ist aufwendig, und zwar nicht nur rein technisch, sondern vor allem auch fachlich-organisatorisch. Die Aufnahme neuer Informationen ins Data Warehouse ist deshalb meist ein langwieriger Prozess, der viel Abstimmung erfordert. Das führt zu langen Projektlaufzeiten, die dem Ziel der Agilität in Unternehmen, also der Fähigkeit, schnell auf Veränderungen zu reagieren, im Wege stehen. Hinzu kommt, dass gerade im Big-Data-Umfeld viel stärker projektbezogen mit Daten gearbeitet wird. Für eine bestimmte Fragestellung müssen die Daten beschafft und analysiert werden. Dabei kann sich herausstellen, dass mehr oder andere Daten benötigt werden, oder auch, dass das Ziel gar nicht erreicht werden kann. Viele Analysen haben daher einmaligen Charakter. All das erfordert mehr Agilität in Bezug auf die Erschließung neuer Datenquellen.



Agile Datenanalysen erfordern neue Ansätze der Datenintegration

Die schnellste und flexibelste Möglichkeit, eine skalierbare Big-Data-Plattform aufzubauen, ist die Nutzung einer cloudbasierten Lösung. Dimitri Groß von Opitz Consulting diskutiert in seinem Artikel die Vor- und Nachteile von Cloud-Lösungen für Big Data sowie verschiedene Architekturvarianten.

Der zweite Artikel bietet einen Überblick über aktuelle Lösungsansätze bei der Integration von Big Data. Fachlich-organisatorisch unterstützen heutzutage Data Catalogs den Prozess der Ordnung und Verwaltung der Daten, während Virtualisierungstechniken und Data Lakes die technische Integration erleichtern. Die schwierige inhaltliche Integration, das „Matching“ von Schema und Daten, geht noch nicht automatisch, aber Werkzeuge helfen auch hier.

Der praxisorientierte Beitrag von IT-Novum beinhaltet mehrere Anwender-Interviews zum Thema Datenintegration. Die Interview-Partner stammen unter anderem von der Bundespolizei, dem Kernforschungszentrum CERN, Bosch und dem Gesundheitsministerium von Mosambik.

Dr. Maik Thiele von der TU Dresden beschreibt in seinem Fachartikel den Ansatz der Ad-hoc-Datenintegration mittels Entity-Augmentation-Systemen. Verfahren zur Ad-hoc-Datenintegration erlauben die weitgehend automatisierte, intelligente Integration von Daten on-the-fly, das heißt zur Laufzeit einer entsprechenden Anfrage. Der vorgestellte Ansatz illustriert anschaulich, in welche Richtung sich das Thema Datenintegration entwickelt und welche Funktionalität die Integrationswerkzeuge von morgen haben werden.

Die Rolle des Chief Data Officer im Rahmen der digitalen Transformation wird von Prof. Dr. Kristin Weber von der Hochschule Würzburg-Schweinfurt thematisiert. Dabei wird insbesondere darauf eingegangen, welche Teilaufgaben im Rahmen der Datenplanung, -beschaffung, -organisation und -integration anfallen.

Alexander Thume und Jörg Stephan von Oraylis erläutern im letzten Fachbeitrag die Bedeutung von Data Lakes bei der Modernisierung einer vorhandenen Data-Warehouse-Lösung. Dabei wird erläutert, welche Aufgaben ein Data Lake übernehmen sollte und wie er sich vom Aufbau her von einem DWH unterscheidet. Dieser Beitrag erschien bereits in BI-SPEKTRUM 3/2017.

Wir hoffen, Ihnen mit diesen Beiträgen hilfreiche Anregungen zu diesem herausfordernden Themenbereich geben zu können, und wünschen Ihnen viel Spaß beim Lesen!

Ihr Jens Albrecht

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:

© Jens Albrecht, TH Nürnberg