Big Data und Big Data Management

Die vier Charakteristika von Big Data sind Volume, Velocity, Variety und der steigende Bedarf an Analysen. Wir haben gesehen, dass klassische BI-Werkzeuge bei Big Data oft scheitern und neue Verfahren wie Textanalytik, Big Data-Extraktionswerkzeuge und analytische Datenbanken notwendig werden.

Traditionelles Information Management stößt jetzt ebenfalls an seine Grenzen. Es hat sich zu „Big Data Management" weiterentwickelt. Dabei setzen sich die drei Hauptkomponenten von traditionellem Information Management entsprechend fort.

Big Data Integration. Hier werden zunächst einmal die traditionellen Datenintegrations-Technologien wie ETL- und ELT-Prozesse[1] und Echtzeit-Verarbeitung (change data capture, event triggering, Web Services) weiter genutzt. Neu dazu kommen MapReduce-basierte Flat-File-Verarbeitung zum Sortieren, Filtern, Mischen und Aggregieren von Daten inklusive einiger Basisarithmetischer Funktionen. Beispiel hierzu ist das FileScale-Verfahren von Talend, das auch von Anbietern wie Uniserv genutzt wird. Alternativ kann man hier aber auch auf alte und sehr bewährte Technologien wie DMExpress von Syncsort zurückgreifen, die im Zuge von Big Data „wiederentdeckt" werden. Schließlich braucht man jetzt auch – das ist neu – Unterstützung für Hadoop, also Import, Export und interne Verarbeitung.

Big Data Lineage. Data Lineage basiert auf einem Repository zur Metadatenverwaltung und der Verwaltung aller Transformationsregeln: Alle Objekte der Datenintegrations-Plattform werden im Endeffekt hier abgebildet, damit sie vom Ursprung bis zum Ziel über den gesamten Informationslebenszyklus verfolgt werden können und bei Änderungen so weit wie möglich automatisiert auch alle betroffenen Objekte gleich mit geändert werden. Das bleibt grundsätzlich so im Big Data, außer dass jetzt auch alle Metadaten zu Big-Data-Objekten und Entitäten zu verwalten sind. Das Repository wird also wichtiger, und es kommt jetzt besonders auf die Performanz an, die mit der Repository-Technologie erreicht werden kann, denn sonst haben wir hier einen Engpass.

Big Data Quality. Datenqualität spielt auch im Big Data eine wichtige Rolle, vor allem dann, wenn Unternehmensdaten mit Information aus dem Big Data angereichert werden sollen, also beispielsweise Kundendaten durch Daten aus den sozialen Medien ergänzt werden sollen oder Patientendaten mit therapeutischen Daten im Gesundheitswesen. Die Grundaufgaben von Data Quality Management bleiben die gleichen. Es geht wie immer um das Profiling, das Cleansing und das Anreichern und Abgleichen mit Referenzdaten. Aber auch hier steigen im Big Data sowohl die Bedeutung von Datenqualität – das Schaffen des „single point of truth" ist beim gegebenen Datenvolumen schwieriger geworden – als auch die Anforderungen an die Performanz der Datenqualitätslösungen, damit man auch in (nahezu) Echtzeit arbeiten kann.

Auf der technologischen Seite muss man also im Big Data Management sicherstellen, dass die Performance stimmt: Alle Werkzeuge, Services und Plattformen müssen entsprechend skalierbar sein. Das wird in der Regel durch Parallelverarbeitung erreicht. Dazu kommen die Anforderungen der neuen Methoden wie Hadoop. Ein weiteres „Muss" ist die Service-Orientierung der Plattform und der Werkzeuge. Dann lassen sich auch hybride Cloud-Lösungen betreiben, beispielsweise ein Datenqualitäts-Management as a Service in ETL-Prozessen, um Social Media-Daten über Referenzdaten auf korrekte Adressen zu prüfen. Einer der ersten Anbieter hierzu ist die deutsche Uniserv. Das ist entscheidend, um auch mit Big Data im Unternehmen den „single point of truth" zu bewahren.

Big Data Management fordert nicht nur die Technologie, sondern auch die Menschen: Neue Skills insbesondere in der IT werden gebraucht. In einigen Unternehmen wie Amazon, eBay, Facebook, Google u.a., die sich schon einige Zeit mit Big Data beschäftigen, haben sich neue Rollen wie „Data Scientists" gebildet. Das sind Mitarbeiter mit folgendem Profil:

- Technische Expertise: Tiefe Kenntnisse in einer Natur- oder Ingenieurs-Wissenschaft sind notwendig.
- Problembewusstsein: die Fähigkeit, ein Problem in testbare Hypothesen aufzubrechen.
- Kommunikation: die Fähigkeit, komplexe Dinge per Anekdoten durch einfach verständliche und gut kommunizierbare Sachverhalte darzustellen.
- Kreativität: die Fähigkeit, Probleme mit anderen Augen zu sehen und anzugehen („thinking out oft he box").

Im Endeffekt wird so Datenmanagement wieder zur eigentlichen und Hauptaufgabe der IT[2] , während das Beherrschen der Prozesse und der Analytik die Hauptaufgabe der Fachbereiche ist.

Fazit

Die Nutzung von Big Data reichert das traditionelle Wissen über Markt, Kunden und Produkte an und schafft insbesondere ein erweitertes Marktwissen und neue Einsichten, denn in den Social Media und im Web findet man ja nicht nur seine Kunden, sondern auch seine Mitbewerber und deren Kunden ebenso wie auch Presse, Marktmultiplikatoren und alle anderen Marktteilnehmer. Die kritischen Erfolgsfaktoren – neben dem Beherrschen der Big-Data-Technologien und Methoden – sind aber die gleichen wie in der Vergangenheit:

- Ohne Sponsor auf der Geschäftsführungsebene geht es nicht.
- Ohne eine Strategie für Big Data im Rahmen von Supply Chain, CRM und BI geht es nicht: Analytik ist kein Selbstzweck, sondern macht die Geschäftsprozesse intelligent.
- Ohne Governance (Organisation per Kompetenzzentrum und entsprechende Prozesse) geht es nicht. Die Konsequenz ist ein Alignment von Business und IT (gemeinsame Begriffe und Sprache, gemeinsames Verständnis).
- Ohne die Menschen geht es nicht: Die Mitarbeiter müssen motiviert und mitgenommen werden, auch ins Big Data.

Dr. Wolfgang Martin
Wolfgang Martin Team
E-Mail: info@wolfgang-martin-team.net

[1] ETL = extract, transform, load; ELT = extract, load, transform. Beide Verfahren unterscheiden sich durch die Reihenfolge der drei Schritte. Bei ETL wird der transform-Schritt in der Datenintegrations-Plattform ausgeführt, bei ELT in der Datenbank.

[2] Das unterstreichen einige neuere Marktstudien, siehe den Beitrag bei InformationAge http://www.information-age.com/channels/information-management/features/1687078/its-focus-shifts-to-data-management.html