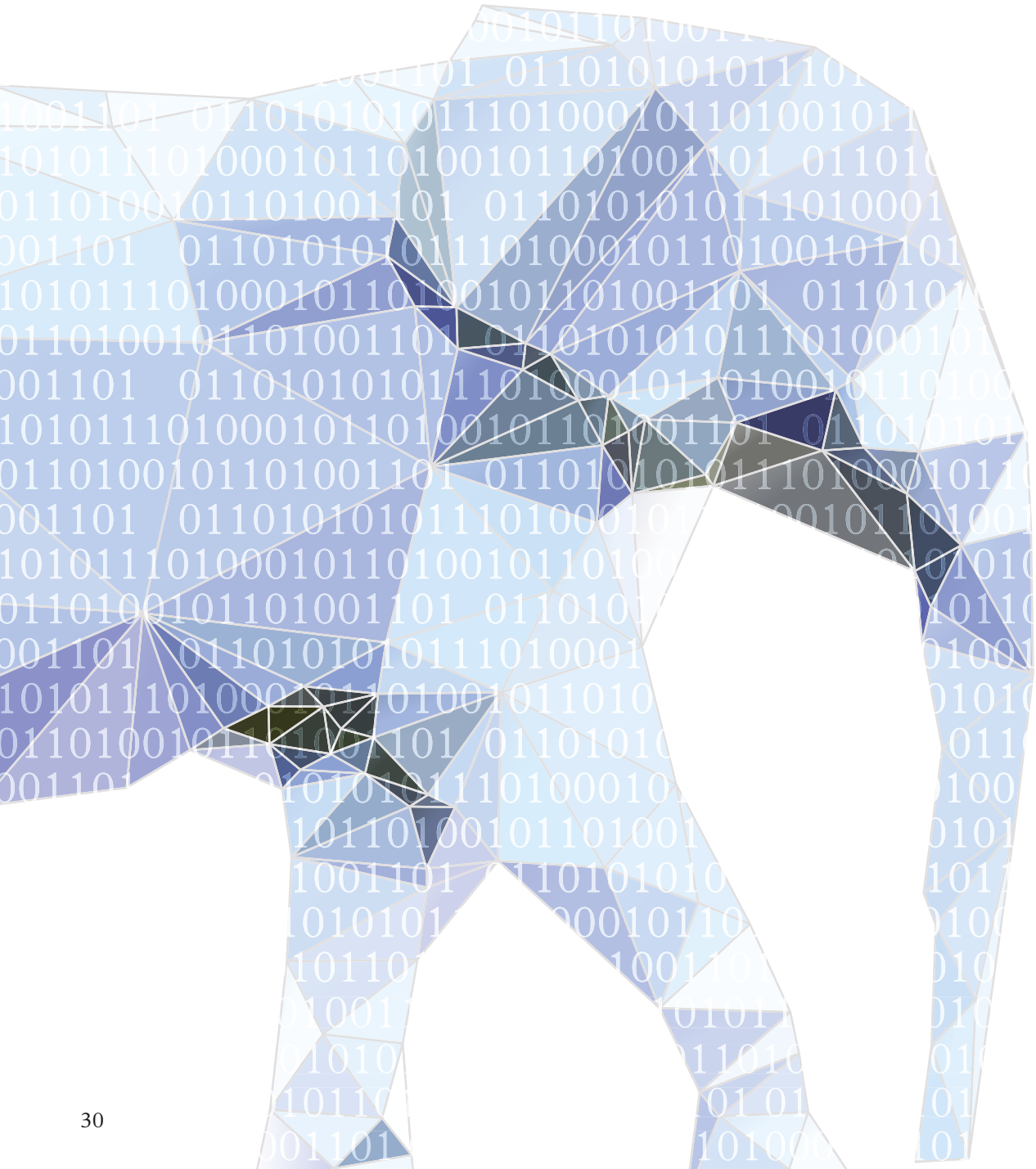
Big Data, Hadoop und Data Vault

Ein evolutionärer Ansatz für Big Data



Big Data liefert neue Möglichkeiten mit schwach strukturierten beziehungsweise unstrukturierten Daten umzugehen. Der »Data Lake« soll alle Daten sammeln; die Analysten von Gartner haben das Konstrukt in einen »Data Swamp« umbenannt. Die Entwicklungsgeschwindigkeit für neue Werkzeuge rund um Hadoop ist sehr hoch, es entstehen immer wieder neue Möglichkeiten der Datenanalyse. Es wird Zeit mit einem evolutionären Vorgehen die Vorteile zu nutzen, ohne gleich die komplette BI-Struktur neu aufzusetzen.

Hadoop bietet viele neue Möglichkeiten mit schwach strukturierten Daten umzugehen. Vor allem beschreibende Daten – wie Sensordaten, Umfragen, Verhaltensdaten (Weblogs) – sind nur schwer in einem relationalen Datenbanksystem zu halten. Nicht weil die Strukturen fehlen, sondern weil die Normalisierung der Daten sehr umfangreich ist und dabei unter Umständen sogar wichtige Informationen verloren gehen.

Daten und Metadaten gemeinsam ablegen. Schwach strukturierte Daten haben nicht nur eine lange Reihe von Attributen, sondern haben Unterstrukturen, sind untergliedert. JSON ist ein Format, in dem sich solche Daten gut darstellen und speichern lassen (siehe Abbildung 1).

Bei der Übertragung auf ein relationales System müssen solche Sätze auf mehrere Tabellen aufgegliedert werden. Wenn sich nun die Struktur auch noch laufend verändert, weil – wie bei Twitter – kontextbezogen jeweils andere Daten gesammelt werden, entsteht viel Arbeit in der Normalisierung dieser Daten, ohne jedoch einen Nutzen zu liefern.

Formate wie JSON speichern in dieser polystrukturierten Form neben den Daten auch die Namen und Formate der einzelnen Attribute. Jetzt kann beim Lesen der Daten anhand dieser Metainformationen entschieden werden, mit welchen Attributen weiter gearbeitet wird.

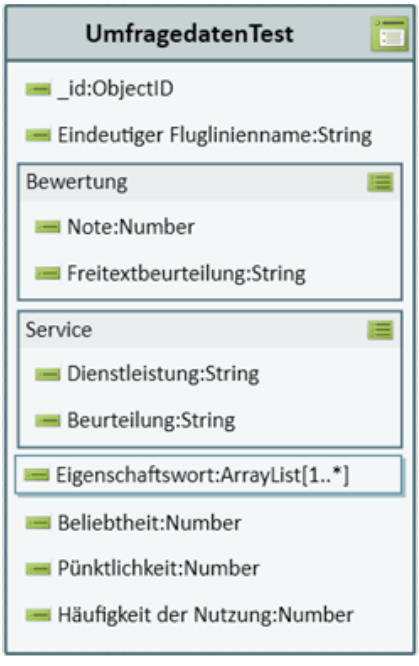


Abbildung 1: JSON ist ein Format, in dem sich schwach strukturierte Daten mit Attributen und Unterstrukturen gut darstellen und speichern lassen.

Entspannt auswerten. Neben JSON stehen mit AVRO und Parquet zwei weitere Formate für die Verarbeitung zur Verfügung. Bei Parquet handelt es sich sogar um ein spaltenbasiertes Speicherformat und ist damit ideal für viele Auswertungen. Der Zugriff auf diese Daten kann dank der enthaltenen Metadaten dann über SQL erfolgen. Hierzu stehen mit Hive und Apache Drill entsprechende Werkzeuge zur Verfügung. Etliche endnutzerfähige Werkzeuge für die Berichterstattung können via SQL auch auf diese Daten zugreifen.

Paradigmenwechsel in der Datenspeicherung. Hadoop ist ein billiger Speicher. Zusammen mit der Philosophie die notwendigen Metadaten mit abzulegen, ergibt sich ein Paradigmenwechsel. Das lässt sich gut am Beispiel der Sensordaten für eine Produktionsstrecke betrachten: bisher hat man nur die wichtigsten Daten gespeichert und diese aufwendig normalisiert. Mit billigem Speicher und ohne die Notwendigkeit zur Normalisierung können nun alle Daten übernommen werden. Jetzt stehen viel mehr Daten über den Produktionsprozess bereit und erlauben aufwändigere Analysen mit noch wertvolleren Erkenntnissen. Zudem sind sofort auch historische Vergleichswerte zur Validierung der Ergebnisse vorhanden.

Integration in die bisherige BI-Landschaft. Mit dem neuen Ansatz stehen die Daten nun schneller und billiger bereit. Um den vollen Nutzen zu erreichen, müssen diese Daten nun mit der bestehenden BI-Landschaft verknüpft

werden. Jede dieser schwach strukturierten Daten bezieht sich auf ein Geschäftsobjekt. Sensordaten beziehen sich auf das Werkstück und die Maschine, Umfragedaten auf den Kunden. Für dieses Geschäftsobjekt müssen die Schlüsselbegriffe sowie die Schlüssel für Referenzen auf andere Geschäftsobjekte identifiziert werden. Diese Schlüssel und ihre Beziehungen sind dann in das bisherige DWH zu übertragen. So entsteht ein Brückenkopf, an dem bei der Auswertung weitere beschreibende Attribute hinzugeschlüsselt werden können.

Hashkeys vereinheitlichen Schlüssel. Die Schlüssel in den schwach strukturierten Daten sind fachliche Schlüssel, setzen sich mitunter aus mehreren Attributen zusammen. Die Information über die Verknüpfung ist somit nur schwer verständlich und muss jeweils dokumentiert und an die Nutzer weitergegeben werden. Hierzu gibt es bei Data Vault einen interessanten Ansatz. Data Vault ist eine Methode für BI, die Standards für Vorgehen, Modellierung und Architektur eines Data Warehouse setzt. Diese Standards bieten viele neue Möglichkeiten zur Automatisierung des DWH. Zudem werden agile Ansätze auch im Core Warehouse möglich, da das Datenmodell flexibel änderbar wird.

Im Data Vault sind auch verteilte Datenarchitekturen möglich. Hierzu müssen Schlüssel in mehreren Systemen gepflegt und dennoch verknüpfbar gehalten werden. Deshalb werden bei Data Vault 2.0 die fachlichen Schlüssel nicht mehr als Surrogat-ID, sondern als Hashkey gepflegt. Dabei werden die Schlüsselinformationen mit Standardhashverfahren wie MD5 oder SHA1 verschlüsselt und als Hex-Codes gespeichert. Nun haben wir einheitliche, deutlich erkennbare Schlüssel, die auf mehreren Plattformen gleich sind, ohne dass auf einem Mastersystem alle Schlüssel generiert werden müssen.

Dieser Ansatz kann auch in ein klassisches Data Warehouse integriert werden, in dem die relevanten Geschäftsobjekte einen alternativen Schlüssel erhalten beziehungsweise der bestehende Schlüssel ersetzt wird. Die Verknüpfung der Daten kann nun an der Stelle erfolgen, an der es am meisten Nutzen stiftet. Das kann sogar erst im Self-Service-BI-Tool erfolgen. Dennoch empfiehlt es sich immer, die Schlüsselinformationen und deren Beziehungen ins Core Warehouse (in Abbildung 2 in den Data Vault) zu übernehmen. So ist die Integration der Daten sichergestellt und Abweichungen in den Schlüsselinformationen können frühzeitig festgestellt und beseitigt werden.

Data Vault

Quelle: MID

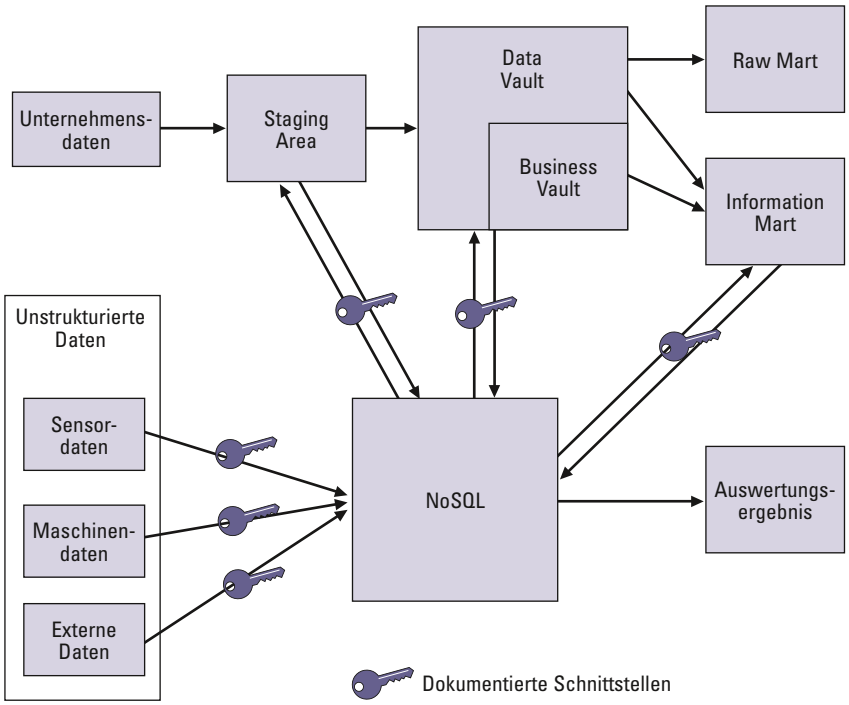


Abbildung 2: Durch die verteilten Datenarchitekturen bei Data Vault müssen Schlüssel in mehreren Systemen gepflegt und dennoch verknüpfbar gehalten werden. Dieser Ansatz kann auch in ein klassisches Data Warehouse integriert werden, in dem die relevanten Geschäftsobjekte einen alternativen Schlüssel erhalten beziehungsweise der bestehende Schlüssel ersetzt wird. Es empfiehlt sich die Schlüsselinformationen und deren Beziehungen ins Core Warehouse (Data Vault) zu übernehmen.

Neue Wege gehen und an die bisherigen anbinden. Big Data und Hadoop bieten neue Lösungsmöglichkeiten. Darum muss nicht alles verworfen und neu erstellt werden. Im Gegenteil durch die Konzentration auf die neuen Möglichkeiten gewinnt die bestehende Lösung an Attraktivität und bleibt dabei stabil. Daten, die bisher nicht effizient geladen werden konnten, sind nun schnell und billig verfügbar. Der Data Lake oder Data Swamp wird zum Bewässerungssystem für die vorhandene BI. Mögen die Daten blühen.

Michael Müller

Michael Müller, Dipl.-Inf. (FH), ist Principal Consultant bei der MID GmbH und beschäftigt sich seit 2000 mit Business Intelligence, Data Warehousing und Data Vault. Seine Schwerpunktthemen sind Architekturen, Modellierung und modellgetriebene Automation für Business Intelligence.