Data Warehouse und Big Data

DR-ING- DIRK ORTLOFF, CAMLINE GMBH

Das Data Warehouse stellt ein zentrales Datenbanksystem dar, das zu Analysezwecken im Unternehmen einsetzbar ist. Das System extrahiert, sammelt und sichert relevante Daten aus verschiedenen heterogenen Datenquellen und versorgt nachgelagerte Systeme.

Data Warehouse beschreibt eine Plattform, die Daten aus verschiedenen Datenquellen sammelt, verdichtet, sie langfristig sichert und nachgelagerte Analysesysteme versorgt. Oft wird das Data Warehouse auch als Datenlager bezeichnet. Vorteil des Datenlagers ist, dass eine globale Sicht auf Daten aus unterschiedlichen Datenbeständen entsteht. Gleichzeitig vereinfacht sich der Zugriff auf die Daten für Anwender, da sie in einer zentralen Datenbank konsistent und strukturiert bereitgestellt sind. Außerdem werden durch die Einführung eines Data Warehouses die teilweise rechenintensiven Abfragen und Auswertungen von den Online-Systemen für die Datenaufnahme aus z. B. der Produktion entkoppelt. Dies hat den Vorteil, dass die produktiven Systeme, die teils in sehr kurzen Zeiten reagieren müssen, nicht durch die Abfrage, Auswertung und Reporting-Aufgaben belastet werden.

Den nachgelagerten Anwendungen bietet das Data Warehouse spezifisch erstellte Auszüge, die sogenannten Data Marts. Die bereitgestellten Daten lassen sich nach bestimmten Mustern analysieren und beispielsweise zur Ermittlung von betrieblichen Kennzahlen einsetzen. Oft stellt das Datenlager die Ausgangsbasis für das Data Mining dar [1]. Die Gesamtheit aller Prozesse zur Datenbeschaffung, Verwaltung, Sicherung und Bereitstellung der Daten nennt sich Data Warehousing.

Das **Data Warehousing** ist in vier Teilprozesse aufteilbar:

- Datenbeschaffung: Beschaffung und Extraktion der Daten aus verschiedenen Datenbeständen
- Datenhaltung: Speicherung der Daten im Datenlager inklusive Langzeitarchivierung
- Datenversorgung: Versorgung der nachgelagerten Systeme mit den benötigten Daten, Bereitstellung von Data Marts
- Datenauswertung: Analysen und Auswertungen der Datenbestände.

Architektur und Prozesse des Data Warehouse.

Die Prozesse des Data Warehouse lassen sich in einem Architekturschaubild vier verschiedenen Bereichen zuordnen:

- Quellsysteme
- Data Staging Area
- Data Presentation Area
- Data Access Tools.

Die Daten für das Datenlager werden von verschiedenen Quellsystemen bereitgestellt. Die Staging Area des Data Warehouse extrahiert, strukturiert, transformiert und lädt die Daten aus den unterschiedlichen Systemen. Über die Staging Area gelangen die Daten in die eigentliche Datenbank des Datenlagers. Diese Datenbank stellt eine parallele Speicherplattform, die Data Presentation Area, zu den eigentlichen Quellsystemen dar und ermöglicht einen separaten Datenzugriff für Anwendungen und nachgelagerte Systeme.

Der Datenzugriff erfolgt über diverse Data Access Tools auf verschiedenen Ebenen, den Data Marts. In der Regel basiert das Data Warehouse auf relationalen Datenbanken, die sich mittels SQL-Abfragen (Structured Query Language) auslesen lassen [2]. Bei besonders großen Datenmengen kommen oft OLAP-Datenbanken (OLAP: Online Analytical Processing) für eine hierarchische Strukturierung der Daten zum Einsatz.

Das Data Warehouse wird meist in regelmäßigen Abständen mit neuen Daten beladen. Mehr und mehr setzen sich Systeme durch, bei der die Versorgung des Datenlagers in Echtzeit erfolgt. Das Data Warehouse sorgt für die saubere Trennung von operativen und auswertenden Systemen und ermöglicht Analysen in Echtzeit. Diese sind wiederum dafür nutzbar, operative Systeme zu steuern.

Data Warehouse im Unternehmensumfeld:

Im Unternehmensumfeld kommt das Data Warehouse in vielen Bereichen zum Einsatz. Es soll als unternehmensweit nutzbares Instrument verschiedene Abteilungen und die Entscheider flexibel unterstützen. Das Datenlager stellt die benötigten Daten für die Anwender zur Analyse von Unternehmensprozessen und -kennzahlen bereit. Für folgenden Aufgaben ist das Datenlager nutzbar:

- Kosten- und Ressourcenermittlung
- Analyse von Geschäfts- und Produktionsprozessen
- Bereitstellung von Reports und Statistiken
- Ermittlung von Unternehmenskennzahlen
- Bereitstellung von Daten für weitergehende Analysen und Data Mining
- Strukturierung und Harmonisierung von Datenbeständen für eine globale Unternehmenssicht.

Big Data – der Data Lake als Ergänzung zum Data Warehouse.

Im Big-Data-Umfeld ist es notwendig, auf eine Vielzahl an Informationen zuzugreifen, die oft nur in unstrukturierter Form zur Verfügung stehen [3]. Zudem sind deutlich größere Datenmengen zu beschaffen und bereitzustellen.

Um diese Herausforderungen zu meistern, ist das ergänzende Konzept des Data Lakes entstanden [4]. Das Data Warehouse kann mithilfe des Data Lakes zu einer Big-Data-Analyseplattform ausgebaut werden. Der Data Lake bietet hohe Speicherkapazität und ermöglicht es, große Datenmengen abzulegen. Gleichzeitig ist er in der Lage, verschiedene Datenformate, auch unstrukturierte, zu verarbeiten. Die im Data Lake gespeicherten Daten können bei Bedarf für Analysen herangezogen werden.

Allerdings sind die heterogenen Data-Lake-Informationen in einem Zwischenschritt aufzubereiten, damit Anwender mit den passenden Tools darauf zugreifen können. Durch geeignete Transformationen entstehen aus den unstrukturierten Rohdaten des Data Lakes strukturierte Datenbestände, die sich mit den Data Access Tools des Data Warehouse darstellen und analysieren lassen.

Die systematische Auswertung von immer größer werdenden Datensammlungen stellt Unternehmen vor immer größeren Herausforderungen. Dabei fehlt aber teilweise einfach das Know-how, um Big-Data-Projekte erfolgreich durchführen zu können. Man folgt einfach nur den gerade aktuellen Trends und Buzz-Words. Dadurch kommt es häufig zu wenig erfolgreichen Herangehensweisen und es treten regelhaft ähnliche Muster auf:

Suche ohne konkretes Ziel. Es wird ein Datentopf ausgewählt und der Auftrag lautet, mit Hilfe von neuronalen Netzen nach „Interessantem“ zu suchen, was typischerweise kostengünstig von einem Praktikanten erledigt werden soll. Das führt allerdings dazu, dass ohne klares Ziel meist wenige Erkenntnisse gewonnen werden, keine adäquate Methodenauswahl stattfindet und somit die Ergebnisse weit hinter den Erwartungen zurückbleiben.

„Wir machen erst einmal einen Data Lake.“ Hier wird alles umfänglich gesammelt, aber meist ohne Vorüberlegungen oder -strukturierungen. Dies basiert auf dem Irrglauben, dass man so viele Daten wie möglich in das System packen sollte, um maximal viele und maximal flexible Auswertungen durchführen zu können. Dies führt allerdings zu hohen Anfangsinvestitionen mit sehr hohen Ergebniserwartungen, aber ohne klare Abschätzung des späteren Nutzens und ohne klare Use Cases. Daher ist die Gefahr hoch, evtl. auch „aufs falsche Pferd zu setzen“ und es ist mit Performance- Problemen zu rechnen.

Sammeln und systematisieren. Beim gegenteiligen Extrem zum Data Lake werden zwar auch möglichst alle Datentöpfe zusammengeführt, allerdings mit deutlicher Systematisierungsabsicht. Analysen stehen erst nach Erledigung der „Hausaufgabe“ an. Bei diesem Vorgehen fallen viele Organisationen dann „auf der anderen Seite vom Pferd“. Das Systematisieren nimmt kein Ende, Datenanalysen und Quick Wins bleiben aus und es entstehen wiederum hohe Kosten ohne absehbaren, konkreten Nutzen. Oftmals fehlen dann bei ersten Analysen auch immer noch die richtigen Daten in der richtigen Form.

Deshalb müssen sich Unternehmen auch bei Big-Data-Projekten im Vorfeld Gedanken machen über sinnvolle Use Cases und passende Analysemethodiken, die unter Berücksichtigung der Kosten-Nutzen-Relation einen Mehrwert bieten. Ein blindes Vertrauen auf den neuesten Technologiehype ist vielfach nicht zielführend. Eine vorgeschaltete Potenzialanalyse ist zu empfehlen und kann hier Aufschluss geben. Für weitere Details im Umgang mit und der Anwendung von Big Data sei auf die in Kürze erscheinende VDI-Richtlinie 3714 [5] verwiesen.

Literatur

[1] www.bigdata-insider.de/was-ist-data-mining-a-593421

[2] www.bigdata-insider.de/was-ist-eine-relationale-datenbank-a-643028

[3] www.bigdata-insider.de/was-ist-big-data-a-562440

[4] www.bigdata-insider.de/was-ist-ein-data-lake-a-686778

[5] www.vdi.de/nc/richtlinie/?tx_wmdbvdirilisearch_pi1%5Brpro_id%5D=7309&cHash=2a55ca4b5e31bd0557e6e262f32d0c82