ORACLEs IT-Architektur für die Verarbeitung von „Big Data“

Oliver Zandner

System-Berater im Bereich ORACLE Datenbank, Region Nord
ORACLE Deutschland B.V. & Co. KG
Thurnithistraße 2
30519 Hannover
oliver.zandner@oracle.com

Abstract: „Big Data“ sind große Datenmengen, wie sie im Internet in sozialen Netwerken oder durch Sensoren entstehen. „Big Data“ stellt neue Anforderungen an Speicherung und Verarbeitung, so dass sich neue Technologien herausgebildet haben, u.a. sog. NoSQL-Datenbanken. Die Firma ORACLE bietet ein Lösungs-Portfolio von Soft- und Hardware an, das „Big Data“ in ORACLE Datenbanken integriert. Durch diese Verbindung entfaltet „Big Data“ seinen eigentlichen Informations-Mehrwert für Organisationen.

1 Was ist „Big Data“?

Bei "Big Data" handelt es sich um große Datenmengen, die u.a. erzeugt werden

- von Sensoren (Mess-Geräte, RFID-Lesegeräte usw)
- im Internet in sozialen Netzwerken (XING, Facebook, LinkedIn etc.) oder in den Zugriffs-Protokollen von Web-Servern
- durch mobile vernetzte digitale End-Geräte (Smart Phones, Tablet PCs)
- in der Wissenschaft wie etwa der Geologie, der Klimaforschugn oder der Genetik

Diese großen Datenmengen zeichnen sich dadurch aus, dass sie

- unstrukturiert bzw. nur schwach strukturiert sind
- nicht selbst-beschreibend sind bzw. keine einheitliche Bedeutung aufweisen - d. h. keinem Schema unterliegen

- erst bereinigt, aggregiert und in vorhandene „traditionelle“ Datenbestände integriert werden müssen, damit sie Aussagekraft gewinnen für die Organisation, die sie nutzen möchte

Bislang basierte die Speicherung und Auswertung von Daten in Organisationen in der Regel auf relationalen Datenbanken in relationalen Datenbankmanagement-Systemen (RDBMS). Diese Daten zeichnen sich dadurch aus, dass sie

- wohlstrukturiert und von einer selbst-beschreibenden Semantik sind - d.h. einem definierten Schema unterliegen
- mittels der Standard-Sprache SQL beschrieben, abgefragt und verändert werden können.
- aus bekannten Quellen meist aus der betreibenden Organisation selbst stammen
- von hoher Informations-Dichte sind – d.h. jeder Datensatz gleichermassen relevant und aussagekräftig für die betreibende Organisation ist

2 Verarbeitungs-Ansätze für „Big Data“

Zur Verarbeitung von „Big Data“ haben sich zwei neue Paradigmen herausgebildet:

Erstens: Der sog. „Key-Value-Store“, d.h. die Speicherung und Verarbeitung von Daten als Paare aus Schlüssel und zugeordnetem Wert. Die Bedeutung des Schlüssels und des Wertes geht nicht aus der Datenstruktur selber hervor (wie sie in einem RDBMS es aus den Metadaten tut) und sie ist nicht festgelegt, denn ein nachfolgendes Schlüssel-Wert-Paar kann eine andere Bedeutung haben. Gespeichert werden die Schlüssel-Wert-Paare in sog. NoSQL-Datenbanken (wie z.B. von ORACLE).

Zweitens: Die Daten werden in Dateien in verteilten Datei-Systemen gespeichert wie z.B. dem Hadoop Distributed Files System (HDFS). Dies ermöglicht, die Daten datei-orientiert zu verarbeiten.

Beiden Ansätzen ist gemeinsam, dass sie dasselbe Verfahren nutzen, um die gespeicherten Daten auszuwerten. Um „Big Data“ auszuwerten – z.B. mit der Fragestellung „Wie häufig wird Produkt X in den Kommentaren der Nutzer eines sozialen Netzwerkes erwähnt?“ – müssen die Daten aggregiert werden, so dass am Ende die relevante Untermenge verbleibt. Dies geschieht in sog. „MapReduce-Jobs“ in verteilten Rechner-Architekturen:

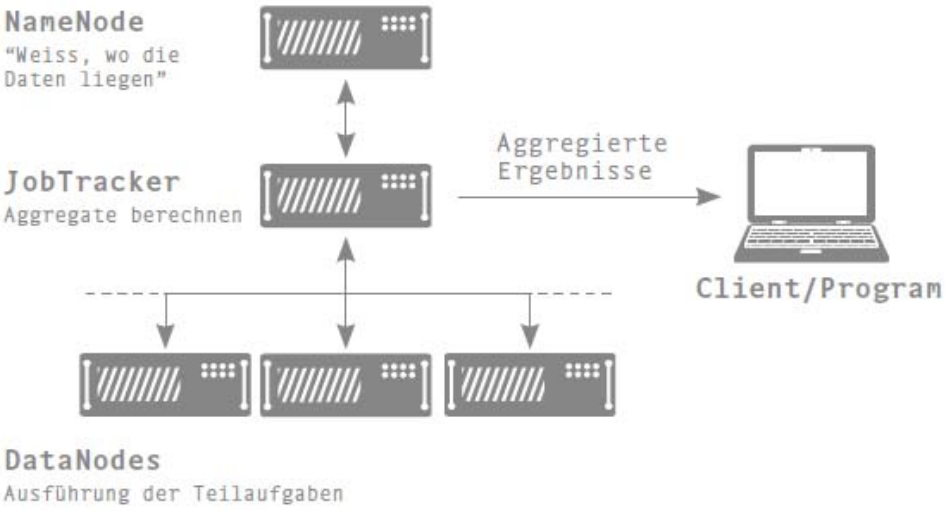

Abbildung 1: „Map-Reduce-Job“ in verteilter Architektur

3 Integrierte Architektur von ORACLE

Die Firma ORACLE bietet eine Architektur an, die den NoSQL-Ansatz mit dem traditionellen relationalen Ansatz verbindet:

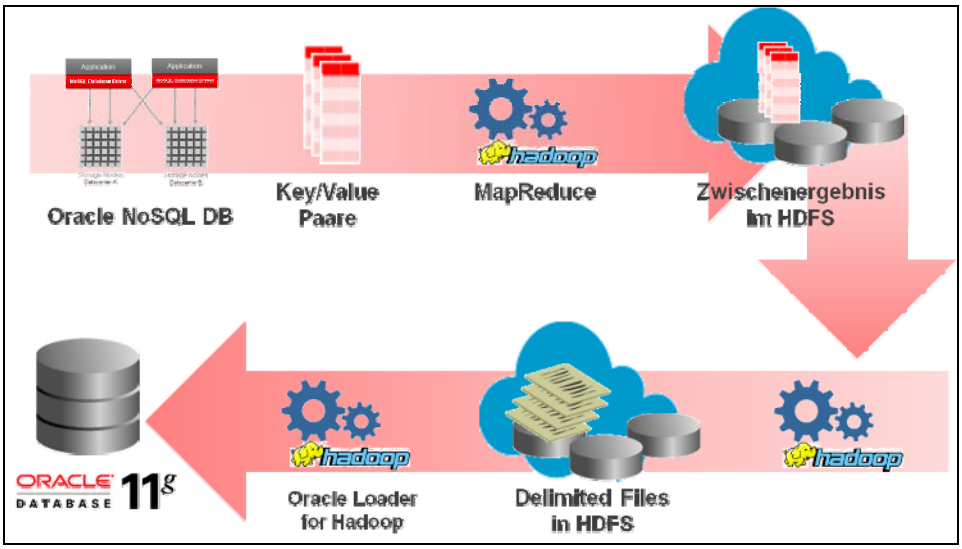

Abbildung 2: Integrierte Architektur von ORACLE zur Verarbeitung von „Big Data“

In dieser Architektur werden:

- Schlüssel-Wert-Paare von der ORACLE NoSQL-Datenbank gespeichert oder aus verteilten Datei-Systemen ausgelesen
- Schlüssel-Wert-Paare mittels des quelloffenen Hadoop-Frameworks in „MapReduce-Jobs“ aggregiert
- Zwischenergebnisse dieser Aggregation im verteilten Datei-System von Hadoop (HDFS) abgelegt und ggfs. anschließend weiter aggregiert
- die Ergebnisse der Aggregation mittels verschiedener Programm-Schnittstellen von ORACLE (sog. APIs) in das RDBMS von ORACLE geladen

Dieser integrierte Ansatz bietet folgende Vorteile:

- Die Organisation, die das RDBMS betreibt, ist in der Lage, „Big Data“ in Beziehung zu bereits vorhandenen Informationen zu setzen, wie zum Beispiel Kunden-Stammdaten oder Bestell-Historien. Aus dieser Möglichkeit des Korrelierens ergibt sich für die Organisation erst der eigentliche Mehrwert von „Big Data“.
- Durch die Integration in das ORACLE RDBMS kann „Big Data“ mit der kompletten Bandbreite an Analyse-Werkzeugen ausgewertet werden, die das ORACLE RDBMS bietet. Hierher gehören u.a.: Data-Mining-Verfahren mittels „R“, Online Analytical Processing (OLAP), Verfahren zur Auswertung geographischer Bezüge (ORACLE Spatial) usw.

Schließlich bietet ORACLE zur Verarbeitung von „Big Data“ auch ein integriertes Gesamt-System aus Soft- und Hardware: Die ORACLE Big Data Appliance. Sie umfasst spezialisierter Hardware, die erforderlichen Software-Komponenten sowie die erforderlichen Konfigurationen. Damit werden Inbetriebnahme und Aministration drastisch vereinfacht.

Weiterführende Informationen

Carsten Czarski (2012), Big Data: Eine Einführung Oracle NoSQL Database, Hadoop MapReduce, Oracle Big Data Connectors: http://www.oracle.com/webfolder/technetwork/de/community/dojo/index.html --> Oracle Dojo #2)

Oracle White Paper „Big Data Strategy“: http://www.oracle.com/us/technologies/big-data/big-data-strategy-guide-1536569.pdf

Oracle White Paper “Oracle: Big Data for the Enterprise” (Januar 2012): http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

Oracle White Paper “Oracle Information Architecture: An Architect's Guide to Big Data” (August 2012): http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf?ssSourceSiteId=ocomen