Hadoop Eine Erweiterung für die Oracle DB?

Matthias Fuchs
Capgemini Deutschland
Nürnberg

Schlüsselworte
Hadoop, Exadata, BigDataSQL, Big Data Appliance, HDFS, R

Einleitung
Oracle hat, neben der Exadata, seit längerem die Big Data Appliance im Programm, ein Engineered System um Daten schnell und sicher abzulegen. Die Big Data Appliance kann mit Cloudera, einer Apache Hadoop Distribution, betrieben werden. Neben strukturierten Daten, wie sie in einer RDBMS vorherrschen, können somit auch unstrukturierte Daten kostengünstig abgelegt werden.
Wie aber können die Daten auf dem Hadoop Cluster verwendet werden? Muss es eine Big Data Appliance sein, oder funktioniert auch jedes andere Hadoop Cluster? Was ändert sich für den DBA, wenn nicht mehr alle Daten eines Data Warehouse in der Oracle DB liegen? Ergeben sich neue Möglichkeiten in der Datennutzung? Wie verhält es sich mit der Sicherheit?
Im Vortrag werden die möglichen Verbindungen zwischen Oracle DB und Hadoop vorgestellt. Im Detail werden Erfahrungen aus Proof of Concepts, die in einem Testcenter durchgeführt wurden, vorgetragen. Hier wird auf Features, Performance und Einsatzszenarien einer Exadata - Big Data Kombination eingegangen. Ebenso werden die Unterschiede der Lösung mit Engineered Systems zu Standard Implementationen aufgezeigt.

Hadoop Kurzübersicht
Im Vortrag wird als Basis eine Hadoop Installation angenommen, die auf der Cloudera Distribution beruht. Weiterhin wird auf Funktionen eingegangen, die gerade im Zusammenspiel von Exadata und Big Data Appliance möglich sind. Als Übersicht dient die folgende Abbildung:

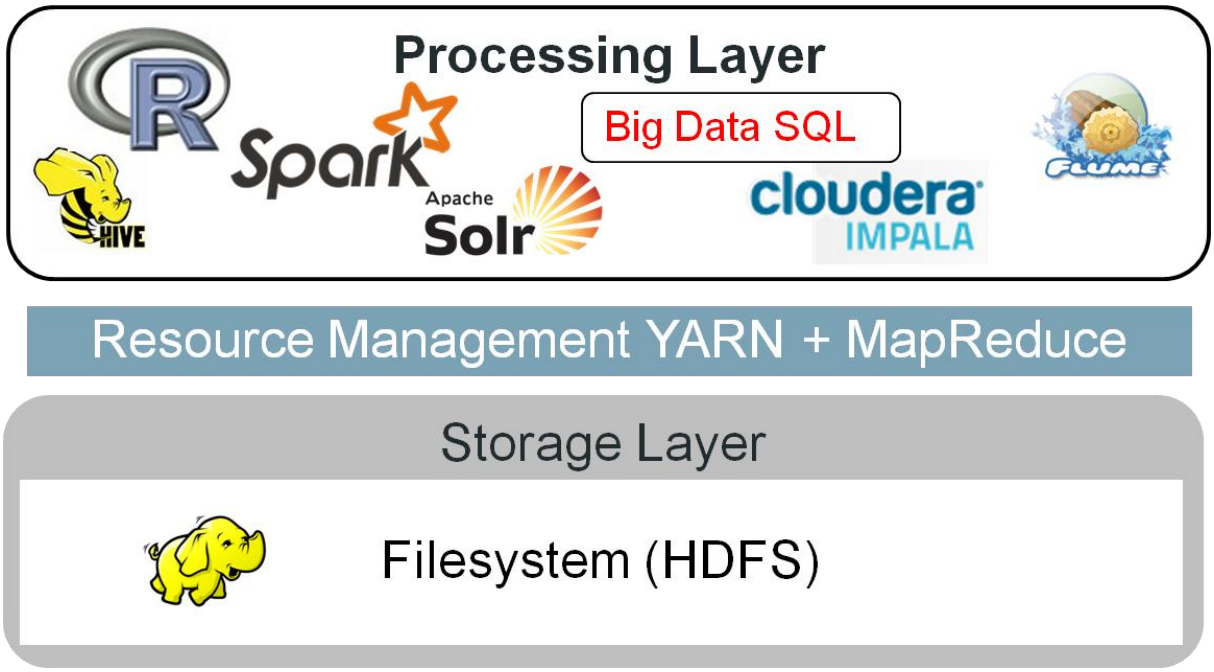

Abb. 1: Hadoop Komponenten

Hadoop und DB im Zusammenspiel

Das Zusammenspiel zwischen einem Hadoop Cluster und einer Oracle Datenbank wird sich im Rahmen von Decision-support Systems (DSS) abspielen, sprich in einem Data Warehouse. Im Rahmen eines Online transaction processing (OLTP) wird die Integration zwischen Datenbank und Hadoop nur eine geringe Rolle einnehmen. Mögliche Einsatzszenarien sind (Auswahl):

- Benutzung als Filesystem
- Zugriff aus der der DB in einen Hadoop Cluster
- Auslagern von Prozessen in ein Hadoop Cluster
- Zugriff auf unstrukturierte Daten
- Und …

Benutzung als Filesystem - HDFS

Das sogenannte HDFS (Hadoop Filesystem) basiert auf dem Apache Hadoop framework. Es ist ein geclustertes Dateiystem, was über fast beliebig viele Knoten betrieben werden kann. Es entstand in Anlehnung an das Google Dateisystem. Das Filesystem ist daraus ausgelegt Daten verteilt abzulegen mit meistens einer 3 fachen Spiegelung. Die Anzahl der Kopien ist konfigurierbar. Ebenso ist eine Standortspiegelung möglich. Die Blockgröße, in den die Dateien zerlegt werden, ist meist 128 MB. Das Filesystem ist auf eine hohe Schreib und Leseperformance optimiert, wobei Änderungen von Daten/Blöcken, sprich das update in einer DB, in der API nicht vorgesehen ist.
In einem Data Warehouse kann somit HDFS als Staging Ebene für eingehende Datenlieferungen dienen. Vorteil ist, die Redundante Speicherung und die Vorhaltung von langen Historien der Lieferungen. Der Zugriff aus der Datenbank auf diese Daten kann nun auf mehreren Wegen Erfolgen.

Zugiff aus der DB in einen Hadoop Cluster

Sind Daten in einem Hadoop Cluster vorhanden können diese auf vielfältige Weise aus der Datenbank zugegriffen werden. Oracle und die Standard open Source Lösungen bieten, je nach eingesetztem System, eine Fülle von Möglichkeiten die zwei Systeme zu verbinden.

Oracle stellt mit den Connectoren Lösungen bereit, die auf Apache Hadoop 2.2.0 und Cloudera CDH 4/5 Systemen in Kombination mit einer 12c, 11g oder 10g Datenbank verwendet werden kann. Eine Lizensierung ist notwendig. Mit den Connectoren ist ein Zugriff auf Daten im HDFS möglich. Das Laden von Daten zwischen den System, auch in Kombination mit dem Oracle Data Integrator ist ebenso implementiert. Mit X-Query Transformationen in Hadoop in Richtung ORacle DB sind auch erweiterte Möglichkeiten unter anderem zu Oracle NoSQL möglich. Angepasste Oracle R Pakete für die Verwendung in Hadoop sind ebenso vorhanden.
Neben den Connectoren gibt es eine weitere Möglichkeit die beiden Systeme zu verbindenb, Big Data SQL. Diese ist momentan nur zwischen Exadata und Big Data Appliance möglich. Zusätzlich zu den Connectoren ist beim Zugriffen aus der DB in die Big Data Appliance ein Offloading, wie bei Exadata Zellen, in die Big Data Appliance möglich.
Neben diesen zusätzlichen Optionen ist auch ein Zugriff auf die Daten in einen Datenbank auf Basis einer ODBC Connection möglich.
Im Vortrag werden Beispiele mit BigDataSQL präsentiert.

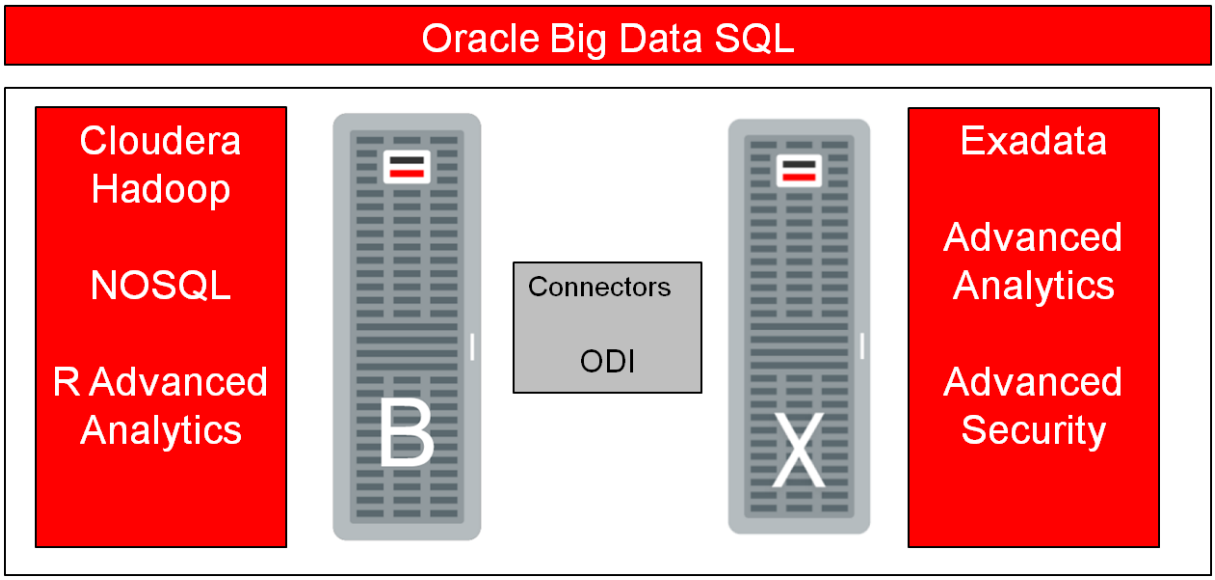

Abb. 2: Connectoren Übersicht

Auslagerung von Prozessen in ein Hadoop Cluster
Neben dem Auslagern von Daten auf ein Hadoop Cluster, ist es auch möglich ganze Prozesse extern in Hadoop laufen zu lassen. Prozesse könnten z.B. Ladeprozesse im Rahmen eines Data Warehouse sein. Dabei werden Berechnung durchgeführt oder Daten verglichen. Im Normalfall sind in einem Hadoop Cluster deutlich mehr Prozessoren und Platten vorhanden, als in einem Datenbank System. Da ebenfalls eine lange Historie im Hadoop Cluster vorhanden ist, können Datenprozesse auf weit zurückliegenden Daten ausgeführt werden. Diese, dann längeren Berechnungsläufe können ohne Einfluß auf das eigentlich Data Warehouse ausgeführt werden. Auch können Teile eines ETL Prozesses, z.B. im Rahmen eine Oracle Data Integrator Ladejobs, einfach auf dem Hadoop Cluster ausgeführt werden. Ziel ist es einerseits die CPU Last in der Oracle DB zu reduzieren um eine schnellere Abfragezeiten dort zu erreichen und andererseits Größere Berechnung ohne Einfluss auf die Data Warehouse Datenbank zu ermöglichen. Im Vortrag wird ein Beispiel mit der Generierung von Hashes präsentiert.

Zugriff auf unstrukturierte Daten
Neben der Ablage von relationalen Daten, wie es in einer RDBMS Datenbank Standard ist, können auch andere Arten von Daten, wie Bilder oder jede Art von Filedaten abgelegt werden. Dadurch, dass viele Tools zur Datenverarbeitung entstanden sind können die Daten analysiert werden und ggf. in eine Oracle DB zurückgeschrieben werden.

Und..
Ein weiteres Problem in der Arbeit mit Hadoop stellt die Security und das Auditing dar. Sicherheitsoptionen, die im Rahmen der Datenbank verwendet werden können auch mit Hadoop verwendet werden. Dazu dient die Datenbank als Zugriffslayer, der die Anfragen an den Cluster weiterleitet. Ein Aufwendiges User und Rollenmangement im Hadoopcluster ist nicht notwendig.

Zusammenfassung

Oracle bietet mit einer Reihe von Integrationsprodukten zwischen Datenbank und Hadoop. Es können Daten und Prozesse ausgelagert werden. Durch die vielen weiteren Komponenten im Hadoop Ökosystem (Zoo) können vielfältige neue Daten und Verarbeitungsschritte erreicht werden.

Kontaktadresse:
Matthias Fuchs
Capgemini Deutschland
Bahnhofstr. 11c
90402 Nürnberg

Telefon:	+49 911 30096 176
Fax:	+49 151 4025 1964
E-Mail	matthias.fuchs@capgemini.com
Internet:	www.de.capgemini.com