

Produkt-Update für die Oracle-Big-Data-Plattform

Jean-Pierre Dijcks, Oracle Corp.

Rund um das Thema „Big Data“ war es und ist es wohl nie langweilig bei Oracle. Nun gibt es aus dem Development wieder eine Reihe von spannenden neuen Features, Produkten und Services rund um das Big-Data-Management-System und seine Komponenten zu vermelden. Hier ein kurzes Update. Zur DOAG Konferenz im November werden sicherlich noch weitere vorgestellt.

Bezüglich Big Data strebt Oracle eine ganzheitliche Lösung an. So integriert man Hadoop-Technologie, NoSQL- und relationale Datenbank in einem zusammenhängenden System. Zusätzlich kann man diesen Verbund auch als Public Cloud nutzen.

Big-Data-Cloud-Service

Die Big-Data-Cloud steht jetzt offiziell zur Verfügung, Sie ist besonders interessant, weil man jetzt das komplette Big-Data-Management-System sowohl On-Premise als auch in der Cloud nutzen kann – und zwar ohne funktionalen Unterschied. Unternehmen können also das Big-Data-Management-System in der Cloud unter den gleichen technischen Rahmenbedingungen nutzen, als wenn sie die Lösung im eigenen Rechenzentrum stehen hätten (siehe *Abbildung 1*). Darüber hinaus sind weitere Services wie die Big-Data-Discovery-Cloud und der Big-Data-Preparation-Cloud-Service im Angebot.

Big Data Appliance

Die Big Data Appliance wurde bereits zu Beginn des Jahres unter dem Namen „X5-2“ ausgebaut. Die Kapazität der Platten ist auf 8 TB gestiegen. Die Maschine verfügt also jetzt über 1,7 PB Rohdaten-Kapazität. *Abbildung 2* zeigt eine kurze Übersicht über die Erweiterung.

Noch eindrucksvoller als der riesige Speicher-Ausbau ist die CPU-Erweiterung auf die neueste Intel-Xeon-Prozessor-Generation. Heutige Big-Data-Anwendungen

haben häufig sehr große Anforderungen: MapReduce für ETL, SQL für Ad-hoc-Abfragen und Spark für spezielle, hochperformante Rechenoperationen verlangen nach enorm viel CPU-Rechen-Power und das vor allem auf gewaltigen Datenmengen. Die 658-CPU-Kerne der neuen Intel-Prozessoren ermöglichen dies.

Auf der Software-Seite ist es jetzt möglich, von der Big Data Appliance aus auf Daten in der Oracle-Datenbank zuzugreifen („Oracle Table Access for Hadoop“). Damit können beispielsweise Hive-Anwendungen Joins zwischen Daten auf der BDA und der Oracle-Datenbank bilden, ohne dass die Datenbank-Daten zunächst auf die BDA kopiert werden müssen. Das erleichtert viele Abläufe auf der BDA enorm. Zudem gibt es jetzt eine Reihe weiterer Automatismen, mit denen sich Daten sehr leicht von der Datenbank in das Hadoop-System auf der BDA kopieren lassen. Neue Interfaces für den SQL Developer unterstützen diese Vorgänge grafisch.

Big Data SQL

Big Data SQL in der Version 2 liefert mit dem schon von der Exadata her bekannten Storage Index ein wichtigstes neues Feature. Der automatisch durch das System im Hauptspeicher gebildete Index dient primär der I/O-Minimierung. Im Gegensatz zur Exadata wird auf der Big Data Appliance ein Index pro 256 MB Block und auf 32 Spalten gebildet. Er ist für die Oracle-Datenbank und für das HDFS völlig transparent und wird automatisch ungültig, wenn sich an den betroffenen Daten etwas ändert (siehe *Abbildung 3*).

Der Storage Index ist an die jeweiligen Spalten geknüpft, die der Benutzer über eine External Table vom Typ „ORACLE_HIVE“ oder „ORACLE_HDFS“ abfragt. Wird eine Spalte in einer Abfrage angefasst, schaltet sich der Index sofort ein.

Big Data Spatial and Graph

Big Data Spatial und Graph steht seit geraumer Zeit auch für die Hadoop-Umgebung zur Verfügung. Eine weitere neue Kompo-



Abbildung 1: Oracle-Big-Data-Cloud-Service: Private und Public Cloud

nente hier ist „Multimedia Analytics“. Damit lassen sich Filme oder Bildarchive scannen und Suchmuster beziehungsweise bestimmte Bildinhalte erkennen. Gesichtserkennung ist beispielsweise eine Lösung, die out of the box damit machbar ist. Multimedia-Analysen nutzen in optimaler Weise die extreme Rechenleistung eines Hadoop-Clusters aus. Bilder können in nahezu unbegrenzter Menge vorliegen und sie werden in einer akzeptablen Zeit analysiert.

Diejenigen, die bislang noch nicht mit Spatial- oder Graph-Daten in Berührung kamen, sollten unbedingt einen Blick auf diese Technologie werfen. Geografische Informationen kommen bei fast allen Daten vor, die wir heute in einem Hadoop-System speichern. Auch Spatial- und Graph-Daten sind ideale Kandidaten für das Skalierungskonzept von Hadoop. Daten, kombiniert mit diesen Spatial-Informationen, lassen sich auch in großer Menge sehr leicht visualisieren, sind dann leichter zu verstehen und wir erhalten eine Reihe zusätzlicher Analyseverfahren.

Big Data Graph bringt eine neue Graphen-Datenbank in das Eco-System. In-Memory lassen sich Graphen-Analysen auf großen Datenmengen im HDFS durchführen. Hier ist der augenfälligste Anwendungsfall sicherlich die Analyse von Social-Media-Kommunikation und -Beziehungen. Man speichert die Kommunikations- beziehungsweise Freundschaftsbeziehungen in der Datenbank und kann wie in einem Netzwerk von Person zu Person Verbindungen erkennen. Anwen-

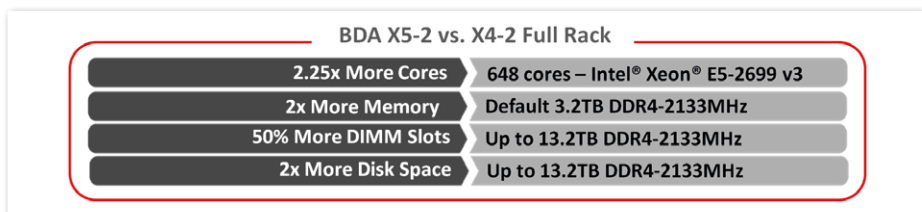


Abbildung 2: Vergleich der Big-Data-Appliance-Komponenten X4-2 mit dem jüngsten Release X5-2

dungsfälle sind aber auch Internet-of-Things-Lösungen mit ihren Interaktionen zwischen den Dingen oder „Cyber Security“ und Verbrechenbekämpfungs-Anwendungen.

Mit den Hadoop-Daten arbeiten

Es gibt noch eine Reihe weiterer wichtiger Neuerungen. Einige davon sind:

- Für Oracle R Advanced Analytics for Hadoop wurden bereits zwei Algorithmen (GLM, Neural Network) auf die Spark-Technologie umgestellt. Das Ergebnis sind um die Faktoren 100 bis 200 verkürzte Laufzeiten. Diese Umstellung mit den Performance-Resultaten zeigen, wo die Reise in der Hadoop-Entwicklung hingeht. Weitere Data-Mining- und Analyse-Algorithmen werden folgen.
- Big Data Discovery steht jetzt in der Version 1.1 zur Verfügung. In dem Werkzeug sind jetzt noch mehr Datenarten (Hadoop und RDBMS) abfragbar.
- Für Streaming Data gibt es ein neues grafisches Interface, den Oracle Stream

Explorer. Hier haben Analysten ein intuitives Tool für Streaming-Analysen.

Die genannten neuen Lösungen machen es immer leichter, auch „Big Daten“ zu analysieren. Solche Tools fehlten ja in den ersten Jahren der Hadoop-Entwicklung.

Von zu Hause aus testen

All diese Features kann man selbst ohne große Hürden ausprobieren. Oracle bietet hierfür die Oracle Big Data Lite Virtual Machine an. Diese wird für jede aktuelle Version der Big Data Appliance angepasst und erweitert. Sie ist perfekt geeignet, um die unterschiedlichen Aspekte von Big Data ohne besonderen Aufwand zu testen und ohne selbst einen eigenen Big-Data-Cluster aufbauen zu müssen (siehe „<http://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html>“).

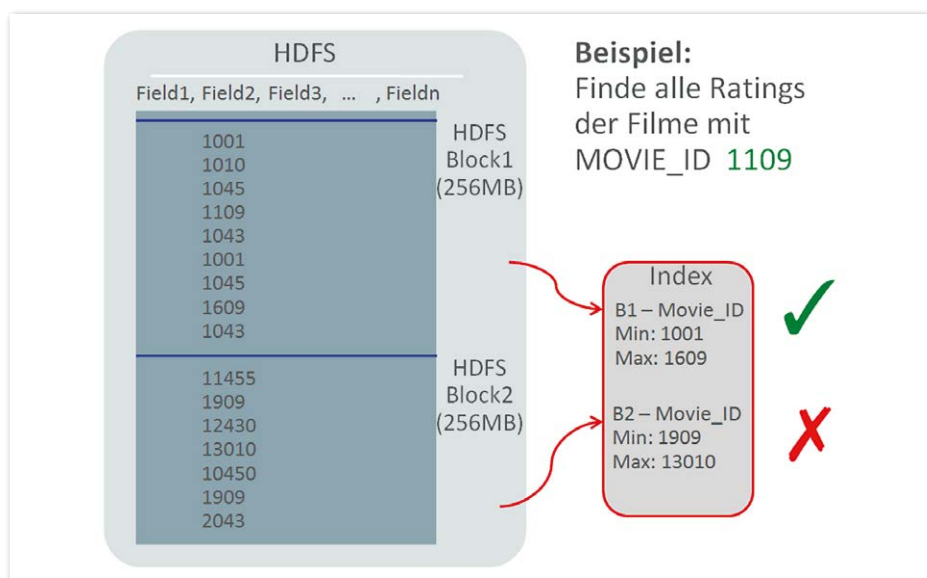


Abbildung 3: Der neue Storage Index arbeitet auf HDFS-Block-Ebene und erzielt eine verbesserte I/O-Performance durch das Überspringen nicht benötigter Blöcke



Jean-Pierre Dijcks



Ins Deutsche übersetzt von Alfred Schlaucher
alfred.schlaucher@oracle.com