Oracle’s Big Data solutions
Roger Wullschleger
DBTA Workshop on Big Data, Cloud Data Management and NoSQL
10. October 2012, Stade de Suisse, Berne
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.

The development, release, and timing of any features or functionality described for Oracle’s products remain at the sole discretion of Oracle.

Agenda

- Big Data
- Solution Spectrum
- Inside the Big Data Appliance
- Big Data Applications Software
- Big Data Analytics
- Conclusions
Big Data
Why Everyone Should Care

Big Data: Acting on New Data

"I found it"

60%

Potential increase in retailers’ operating margins possible with Big Data

McKinsey Global Institute: Big DataThe next frontier for innovation, competition and productivity (May 2011)
Tapping into Diverse Data Sets

Big Data:
Decisions based on all your data

Information Architectures
Today:
Decisions based on database data

Drive Value from Big Data
Building a Big Data Platform
Big Data: Infrastructure Requirements

- Deep Analytics
- Agile Development
- Massive Scalability
- Real Time Results
- High Throughput
- In-Place Preparation
- All Data Sources/Structures
- Low, predictable Latency
- High Transaction Count
- Flexible Data Structures

Divided Solution Spectrum

Acquire Analyze Organize

Data Variety

Unstructured

DBMS (OLTP)

Transaction (Key-Value) Stores

MapReduce Solutions

Schema-less

Distributed File Systems

DBMS (DW)

Advanced Analytics

Schema

ETL

NoSQL

Flexible Specialized Developer Centric

SQL

Trusted Secure Administered

ORACLE
Hadoop to Oracle – Bridging the Gap

Oracle Integrated Software Solution
Inside the Big Data Appliance
Overview

Oracle Engineered Solutions

Data Variety

Unstructured

Schema

Information Density

Acquire
Organize
Analyze

Big Data Appliance
- Hadoop
- NoSQL Database
- Oracle Loader for Hadoop
- Oracle Data Integrator

Oracle Exadata
- OLTP & DW
- Data Mining & Oracle R

Exalytics
- Speed of Thought
- Analytics
Big Data Appliance

Usage Model

Oracle Big Data Appliance

Oracle Exadata

Oracle Exalytics

Stream | Acquire | Organize | Analyze & Visualize

Why build a Hadoop Appliance?

- Time to Build?
- Required Expertise?
- Cost and Difficulty Maintaining?
### Oracle Big Data Appliance Hardware

- 18 Sun X4270 M2 Servers
  - 48 GB memory per node = 864 GB memory
  - 12 Intel cores per node = 216 cores
  - 24 TB storage per node = 432 TB storage
- 40 Gb p/sec InfiniBand
- 10 Gb p/sec Ethernet

### Big Data Appliance

Cluster of industry standard servers for Hadoop and NoSQL Database
- Focus on **Scalability** and **Availability** at low cost

**InfiniBand Network**
- Redundant 40Gb/s switches
- [IB connectivity to Exadata](https://www.oracle.com)

**10GigE Network**
- 8 10GigE ports
- [Datacenter connectivity](https://www.oracle.com)
Scale Out to Infinity

Scale out by connecting racks to each other using Infiniband
- Expand up to eight racks without additional switches
- Scale beyond eight racks by adding an additional switch

Oracle Big Data Appliance Software

- Oracle Enterprise Linux 5.6
- Oracle Hotspot Java VM
- Cloudera's Distribution including Apache Hadoop
- Cloudera Manager
- Open Source Distribution of R
- Oracle NoSQL Database Community Edition
Why Open-Source Apache Hadoop?

- Fast evolution in critical features
  - Built by the Hadoop experts in the community
  - Practical instead of esoteric
  - Focus on what is needed for large clusters
- Proven at very large scale
  - In production at all the large consumers of Hadoop
  - Extremely stable in those environments
  - Well-understood by practitioners

Software Layout

- Node 1:
  - M: Name Node, Balancer & HBase Master
  - S: HDFS Data Node, NoSQL DB Storage Node
- Node 2:
  - M: Secondary Name Node, Management, Zookeeper, MySQL Slave
  - S: HDFS Data Node, NoSQL DB Storage Node
- Node 3:
  - M: JobTracker, MySQL Master, ODI Agent, Hive Server
  - S: HDFS Data Node, NoSQL DB Storage Node
- Node 4 – 18:
  - S: HDFS Data Nodes, Task Tracker, HBase Region Server, NoSQL DB Storage Nodes
  - Your MapReduce runs here!
Big Data Application Software
Acquire New Information

Key-Value Store Workloads

- Large dynamic schema based data repositories
- Data capture
  - Web applications (click-through capture)
  - Online retail
  - Sensor/statistics/network capture (factory automation for example)
  - Backup services for mobile devices
- Data services
  - Scalable authentication
  - Real-time communication (MMS, SMS, routing)
  - Personalization
  - Social Networks
Oracle NoSQL DB
A distributed, scalable key-value database

• Simple Data Model
  • Key-value pair with major-sub-key paradigm
  • Read/insert/update/delete operations

• Scalability
  • Dynamic data partitioning and distribution
  • Optimized data access via intelligent driver

• High availability
  • One or more replicas
  • Disaster recovery through location of replicas
  • Resilient to partition master failures
  • No single point of failure

• Transparent load balancing
  • Reads from master or replicas
  • Driver is network topology & latency aware

• Elastic (Planned for Release 2)
  • Online addition/removal of Storage Nodes
  • Automatic data redistribution

Oracle NoSQL DB Differentiation

• Commercial Grade Software and Support
  • General-purpose
  • Reliable – Based on proven Berkeley DB JE HA
  • Easy to install and configure

• Scalable throughput, bounded latency

• Simple Programming and Operational Model
  • Simple Major + Sub key and Value data structure
  • ACID transactions
  • Configurable consistency & durability

• Easy Management
  • Web-based console, API accessible
  • Manages and Monitors: Topology; Load; Performance; Events; Alerts

• Completes Oracle large scale data storage offerings
Big Data Application Software
Organizing Data for Analysis

Oracle Loader for Hadoop Features

- Load data into a partitioned or non-partitioned table
  - Single level, composite or interval partitioned table
  - Support for scalar datatypes of Oracle Database
  - Load into Oracle Database 11g Release 2

- Runs as a Hadoop job and supports standard options

- Pre-partitions and sorts data on Hadoop

- Online and offline load modes
Oracle Loader for Hadoop

Read target table metadata from the database
Perform partitioning, sorting, and data conversion
Connect to the database from reducer nodes, load into database partitions in parallel

Oracle Loader for Hadoop: Online Option

Read target table metadata from the database
Perform partitioning, sorting, and data conversion
Connect to the database from reducer nodes, load into database partitions in parallel
Oracle Loader for Hadoop: Offline Option

Oracle Loader for Hadoop Advantages

- Offload database server processing to Hadoop:
  - Convert input data to final database format
  - Compute table partition for row
  - Sort rows by primary key within a table partition
- Generate binary datapump files
- Balance partition groups across reducers
Selection Output Option for Use Case

<table>
<thead>
<tr>
<th>Oracle Loader for Hadoop Output Option</th>
<th>Use Case Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online load with JDBC</td>
<td>The simplest use case for non partitioned tables</td>
</tr>
<tr>
<td>Online load with Direct Path</td>
<td>Fast online load for partitioned tables</td>
</tr>
<tr>
<td>Offline load with datapump files</td>
<td>Fastest load method for external tables</td>
</tr>
<tr>
<td><strong>On Oracle Big Data Appliance</strong></td>
<td><strong>Leave data on HDFS</strong></td>
</tr>
<tr>
<td><strong>Direct HDFS</strong></td>
<td><strong>Parallel access from database</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Import into database when needed</strong></td>
</tr>
</tbody>
</table>

Streaming Access to HDFS

```
HDFS
Datafile_part_1
HDFS
Datafile_part_2
HDFS
Datafile_part_m
HDFS
Datafile_part_n
HDFS
Datafile_part_x

Oracle Database
External Table
View Or Table Function

Map
Reduce
```

Query
Automate Usage of Oracle Loader for Hadoop

Oracle Data Integrator (ODI)

• ODI has knowledge modules to
  – Generate data transformation code to run on Hive/Hadoop
  – Invoke Oracle Loader for Hadoop

• Use the drag-and-drop interface in ODI to
  – Include invocation of Oracle Loader for Hadoop in any ODI packaged flow
**Big Data Analytics**

Real Time Analytics Platform

**R Statistical Programming Language**

- Open source language and environment
- Used for statistical computing and graphics
- Strength in easily producing publication-quality plots
- Highly extensible with open source community R packages
Oracle R Enterprise

- Data and statistical analysis are stored and run in-database
- Same R user experience & same R clients
- Embed in operational systems
- Complements Oracle Data Mining

Drive Value from Big Data

Conclusions
Big Data Appliance
Big Data for the Enterprise

• Optimized and Complete
  • Everything you need to store and integrate your lower information density data
• Integrated with Oracle Exadata
  • Analyze all your data
• Easy to Deploy
  • Risk Free, Quick Installation and Setup
• Single Vendor Support
  • Full Oracle support for the entire system and software set

Oracle Integrated Solution Stack for Big Data

ACQUIRE
HDFS
Oracle NoSQL Database
Enterprise Applications

ORGANIZE
Hadoop (MapReduce)
Oracle Loader for Hadoop

ANALYZE
Data Warehouse
Oracle Data Integrator
H Database Analytics

DECIDE
Oracle Analytic Applications
Oracle: Big Data for the Enterprise

- The most comprehensive solution
  - Includes everything needed to acquire, organize and analyze all your data
- Optimized for Extreme Analytics
  - Deepest analytics portfolio with access to all data
- Engineered to Work Together
  - Eliminate deployment risk and support risk
- Enterprise Ready
  - Deliver extreme performance and scalability

Questions