Integrating Big Data for the Enterprise

Melli Annamalai
Product Manager

Rob Abbott
Consulting Engineer

Oracle Big Data Development
October 1, 2014
Safe Harbor Statement

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 remains at the sole discretion of Oracle.
Program Agenda

1 Customer Use Cases

2 Oracle Big Data Connectors
   Load, SQL Query, Analyze

3 Oracle Big Data SQL
   Optimized SQL Access for Engineered Systems

4 Oracle Data Integrator
   Comprehensive Data integration

5 Oracle Golden Gate
   Real-time Replication
Big Data Analytic Services

- Measure effect of new marketing campaign
- **Quick access to weblogs in Hadoop, combine with data in database**

Business Transformation

- Leading Spanish Bank > 13M customers interact via ATMs, web, mobile, branches, numerous acquisitions over the years
- **Collect & unify all relevant information in ‘Data Pool’**

Network Performance

- Clean and process network monitoring data on Hadoop
- Load into database

Usage-based Insurance

- Track driving parameters integrated with location data on Hadoop
- Load into database
Measure effect of marketing campaign
• Customer subscriptions in database
• Online activity in weblogs in Hadoop

Collect and unify in Data ‘Pool’
• > 13M customers interact via ATMs, web, mobile, branches, numerous acquisitions over the years

Usage-based insurance
• Track driving parameters integrated with location data on Hadoop
• Driver profiles, policies in database

Monitor network performance
• Clean and process network monitoring data in Hadoop
• Load into database for further analysis
Driving Business Value from Technology Innovation

Use the Right Tool for the Job and benefit from the Power of “AND”

Hadoop
- Change the Business
  - Disrupt competitors
  - Disintermediate supply chains
  - Leverage new paradigms
  - Exploit new analyses

NoSQL
- Scale the Business
  - Serve data faster
  - Meet mobile challenges
  - Scale-out economically

Relational
- Run the Business
  - Integrate existing systems
  - Support mission-critical tasks
  - Protect existing expenditures
  - Ensure skills relevance
Focus in this Session

**Hadoop**
- Data in files
- Schema on read
- Simple programming model for large scale data processing
- Append only
- Sequential access of blocks

**Relational**
- Data organized for fast query
- Structured schema
- Complex programming models
- Read, write, delete, update
- Access specific record
“The implementation of this Big Data solution will help CaixaBank remain at the forefront of innovation in the financial sector, delivering the best and most competitive services to our customers”

– Juan Maria Nin, Chief Executive Officer, CaixaBank
Integrating Big Data

- Data Preparation
- Data Staging
- Data Access
- Data Load
- Real-time Replication
- Exploratory Analysis
- Data Reservoir
- Deep Analytics

Required features:
- Fewer new interfaces
- Uniform access methods
- Easy to use
- Performance
Oracle Big Data Connectors
Oracle Big Data Connectors

Optimized for Hadoop: Maximise parallelism
Fast performance
Analyze data on Hadoop using familiar client tools
## Certified Hadoop and Database Versions

<table>
<thead>
<tr>
<th>Database versions (on any operating system*)</th>
<th>Hadoop versions</th>
<th>Certified by</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.2.0.5 and greater</td>
<td>Apache Hadoop 2.x</td>
<td>Oracle</td>
</tr>
<tr>
<td>11.2.0.3 and greater</td>
<td>CDH 4.x (Cloudera)</td>
<td>Oracle</td>
</tr>
<tr>
<td>12c</td>
<td>CDH 5.x (Cloudera)</td>
<td>Oracle</td>
</tr>
<tr>
<td></td>
<td>HDP 1.3 (Hortonworks)</td>
<td>Hortonworks</td>
</tr>
<tr>
<td></td>
<td>HDP 2.1 (Hortonworks)</td>
<td>Hortonworks</td>
</tr>
</tbody>
</table>

* Oracle SQL Connector for HDFS requires Hadoop client to be supported on the operating system
Analyze on Hadoop

With Oracle Big Data Connectors
Oracle XQuery for Hadoop

- Massively scalable XQuery processing in Hadoop
- XQueries processed in parallel with MapReduce
- Query XML with Hive with XML extensions
- Oozie integration

```xml
for $ln in text:collection()
let $f := tokenize($ln,"","")
where $f[1] = 'x'
return text:put($f[2]))
```
Oracle XQuery for Hadoop

Query Output Options

for $ln in text:collection()
let $f := tokenize($ln, ",")
where $f[1] = 'x'
return text:put($f[2])
Oracle R Advanced Analytics for Hadoop

- Pre-packaged predictive analytics algorithms
- Familiar interface R (to Data Scientists)
- Customer: Credit behavior evaluation
  - Enabled faster analytics, simpler solution, and better behavior model
Oracle Loader for Hadoop
High Speed Load from Hadoop to Oracle Database
Oracle Loader for Hadoop

- Parallel load, optimized for Hadoop
- Automatic load balancing
- Convert to Oracle format on Hadoop
  - Save database CPU
- Load specific Hive partitions
- Kerberos authentication
- Load directly into In-Memory table

- Text
- Avro
- Parquet
- Sequence files
- JSON
- Compressed files
- Hive Log files
- And more …
Oracle Loader for Hadoop

**Performance**

- Extremely fast performance
- Sample numbers (on Oracle Engineered Systems)
  - 4.4 TB/hour end-to-end (load + Hadoop process)
  - 12+ TB/hour load time
- Much higher than typical customer requirements
- Optimized for Oracle Big Data Appliance and Oracle Exadata: **InfiniBand Connectivity**
Oracle Loader for Hadoop

Concurrency

- Uses very few database CPU cycles
- Maximizes concurrency on database
- Enables large and continuous loads concurrently with applications

External table load of Oracle Loader for Hadoop generated Data Pump files
Oracle Loader for Hadoop

**Automatic Load Balancing**

Real data is skewed
When one task loads more rows than others

Time = X

Time = 2...10 X or more
Oracle Loader for Hadoop

**Automatic Load Balancing**

- Intelligent sampling to distribute load evenly across load processes
- Fine tune load properties for data distribution in current job
- Maintain repeatable load performance

*Load time: >10x faster*
Customer Use Case

• > 5000 Hive partitions
• 1 TB of data
• High data skew
• Load into multiple target tables

• Achievable speed: 20 min, well exceeded their target
• Performance improvement with load balancing: 2-3x
Oracle Loader for Hadoop

**Benefits**

- Fast, parallel load of a variety of data formats
- Minimize impact on database during load
- Automatic load balancing
- Works with Kerberos
Oracle SQL Connector for HDFS

Oracle SQL access on Commodity Hardware
Oracle SQL Connector for HDFS

- Parallel query and load
- Load into database or query in place
- Access text or Hive over text
- Access compressed data
- Access specific Hive partitions
- Kerberos authentication

```sql
create table customer_address
    ( ca_customer_id     number(10,0)
    , ca_street_number  char(10)
    , ca_state          char(2)
    , ca_zip            char(10)

organization external (
  TYPE ORACLE_LOADER
  DEFAULT DIRECTORY DEFAULT_DIR
  ACCESS PARAMETERS
  ( ...
  PREPROCESSOR "HDFS_BIN_PATH:hdfs_stream" )
```
Oracle SQL Connector for HDFS

• Includes tool to generate external table

• Performance on Engineered Systems
  – 15 TB/hour load time

• Query and load Oracle Data Pump files
  – Binary file in Oracle format
  – Uses less database CPU cycles during query/load
Oracle SQL Connector for HDFS

Hive Partitioned Tables

- Tool generates external table and view for each partition
- Create a UNION ALL view on all views
- Query
  - Individual view
  - UNION ALL view with Hive partition column WHERE clause to access only relevant views

QUERY ONLY SPECIFIED PARTITIONS:

```sql
T_DATE = TO_DATE('2013-10-01', 'YYYY-MM-DD')
AND
T_DATE = TO_DATE('2013-09-30', 'YYYY-MM-DD')
```
Oracle SQL Connector for HDFS

Benefits

• Fast access
  – Parallel access to data in Hadoop

• Query in-place from database

• Easy to use for Oracle developers

• Works with Kerberos
Oracle Big Data SQL
Optimized for Oracle Engineered Systems

Big Data Appliance
+ Cloudera Hadoop

Exadata
+ Oracle Database
Oracle Big Data SQL

Query All Data without Application Change or Data Conversion

create table customer_address
    ( ca_customer_id     number(10,0)
    , ca_street_number  char(10)
    , ca_state          char(2)
    , ca_zip            char(10))
organization external ( 
    TYPE ORACLE_HIVE
    DEFAULT DIRECTORY DEFAULT_DIR
    ACCESS PARAMETERS
        (com.oracle.bigdata.cluster hadoop_cl_1)
    LOCATION ('hive://customer_address')

Big Data SQL
Query all data with Oracle SQL
Smart scan in Hadoop to optimize data requests

Big Data Appliance + Cloudera Hadoop

Exadata + Oracle Database
Oracle Big Data SQL

Copy to BDA

create table customer_address
( ca_customer_id     number(10,0)
, ca_street_number  char(10)
, ca_state          char(2)
, ca_zip            char(10))
organization external (TYPE ORACLE_DATAPUMP)
DEFAULT DIRECTORY DEFAULT_DIR
LOCATION ('customer_address.dmp')
) AS
SELECT ... FROM ... (can be any Oracle SQL query)
Oracle Big Data SQL

**Copy to BDA**

- Business critical data on Exadata
- Copy older data to BDA
  - Integrate with batch analysis in Hadoop
  - Infrequent query of archive data
- Query data in BDA or database with no application change
Oracle Big Data SQL

Copy to BDA Example Use Case

- Most current data on Exadata
- Older online data in BDA
- Query all online data with no application change

Steps
- Copy older partitions to BDA
- Create views on Exadata + BDA partitions
- Drop older Exadata partitions
Data Integration Platform

Oracle Data Integrator
Oracle Data Integration for Big Data and Hadoop

Comprehensive data integration platform designed to work with all data

Data Replication
  Continuous data staging into Hadoop

Data Transformation
  – Pushdown processing in Hadoop

Data Federation
  – Query Hadoop SQL via JDBC

Data Quality
  – Fix quality at the source or invoke Machine Learning in Hadoop

Metadata Management
  – Lineage and Impact Analysis w/Hadoop

Oracle GoldenGate
  (Data Replication)

Oracle Data Integrator
  (Data Transformation)

Data Service Integrator
  (Data Federation)

Enterprise Data Quality
  (Profile, Cleanse, Match and De-duplicate)

Enterprise Metadata Management
  (Lineage, Impact Analysis and Data Provenance)
Oracle Data Integration Can Help Right Now

#1 – Tools not Spaghetti
- “ETL 101” → avoid complex, costly custom coding

#2 – Non-invasive Capture and Staging
- Move data without inefficient batch extracts

#3 – Processing is Taken to the Data
- No separate ETL engine needed
- Eliminate unnecessary data movement
- Reclaim latency and time from network overhead

#4 – Native Hadoop Execution
- Choose the right Hadoop language for your use case
  - HiveQL, Pig, Spark, Storm, Java/MR2, etc.
- Template driven code gen keeps pace with change on Hadoop platform

#5 – Native SQL Pushdown
- Optimize some join types within the Data Warehouse

#6 – Oracle Optimized
- OGG and ODI certified to run on the Oracle Appliances
Real-time Replication to Hadoop

Oracle Golden Gate
GoldenGate and Streaming Data

Non-invasive Capture and Staging
- Move data without batch extracts

Open OGG APIs for capture of non-DBMS events

Stage DB records for subsequent processing

Leverage DB transactions w/in realtime analytic streams

Apps

Sensors
Summary

• Fast, easy, integration of all data in your Big Data solution

• Oracle Big Data Connectors
• Oracle Big Data SQL (on Oracle Engineered Systems)
• Oracle Data Integrator
• Oracle Golden Gate

Intricate elephant sculptures throughout the base of the Chennakesava temple in Belur, India, symbolizing strength. The temple was built in 1117 CE.
Questions?
Hardware and Software
Engineered to Work Together