ORACLE®

ORACLE

ORACLE OPEN WORLD

Programming with Oracle Big Data Connectors

Melli Annamalai
Rob Abbott

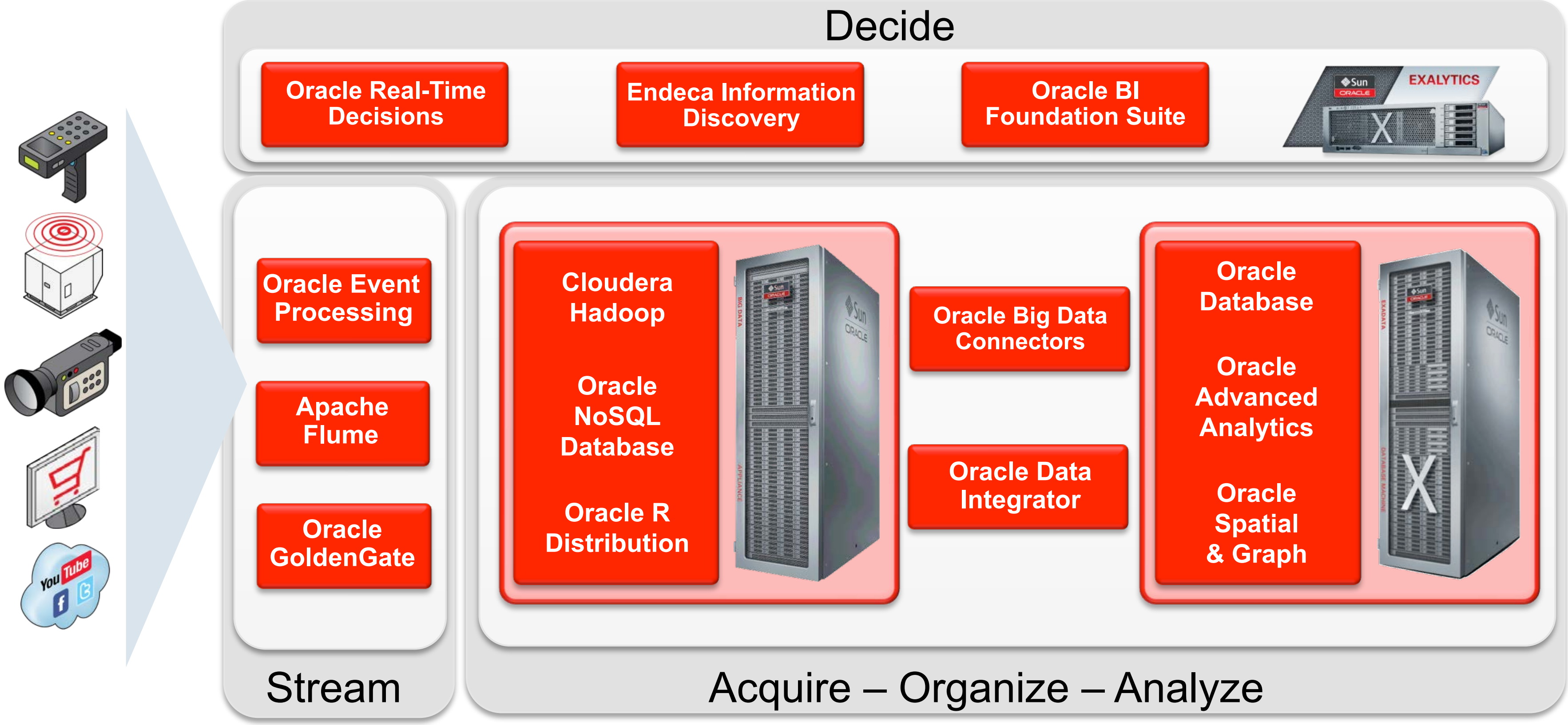
HARDWARE AND SOFTWARE ENGINEERED TO WORK TOGETHER

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

- Oracle Big Data Connectors Overview
- Oracle Loader for Hadoop
- Oracle SQL Connector for HDFS
- Performance Tuning
- Summary

Oracle Big Data Solution



Decide

Oracle Real-Time Decisions

Endeca Information Discovery

Oracle BI Foundation Suite

Oracle Event Processing

Apache Flume

Oracle GoldenGate

Stream

Cloudera Hadoop

Oracle NoSQL Database

Oracle R Distribution

Oracle Big Data Connectors

Oracle Data Integrator

Oracle Database

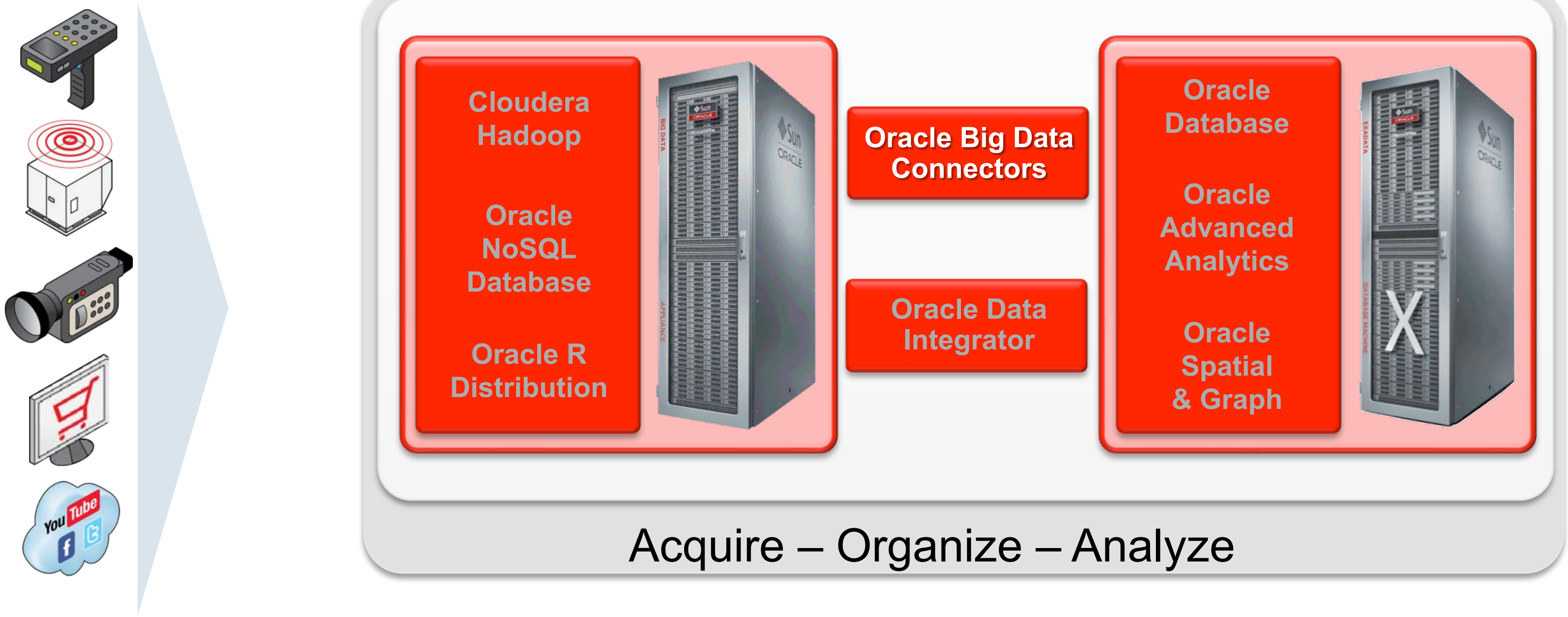
Oracle Advanced Analytics

Oracle Spatial & Graph

Acquire – Organize – Analyze

Oracle Big Data Connectors

Connecting Hadoop to Oracle Database



Cloudera Hadoop

Oracle NoSQL Database

Oracle R Distribution

Oracle Big Data Connectors

Oracle Data Integrator

Oracle Database

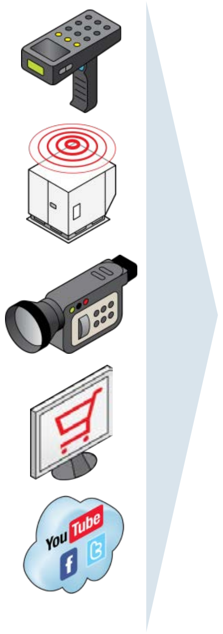
Oracle Advanced Analytics

Oracle Spatial & Graph

Acquire – Organize – Analyze

Oracle Big Data Connectors

Connecting Hadoop to Oracle Database



Hadoop	Database
Batch oriented	Real-Time
Transform input data	Fast access to a specific record
Schema on read	Schema on write
Unstructured data, less useful after relevant data is extracted	High availability, reliability, security
Write once, read many times	Read, write, delete update

Oracle Big Data Connectors

– Organize –

Oracle Big Data Connectors

Licensed Together

- Oracle SQL Connector for HDFS
- Oracle Loader for Hadoop
- Oracle R Connector for Hadoop
- Oracle Data Integrator Application Adapters for Hadoop
- **Announcing at OOW 2013**: Oracle XQuery for Hadoop

Oracle Loader for Hadoop and Oracle Direct Connector for HDFS

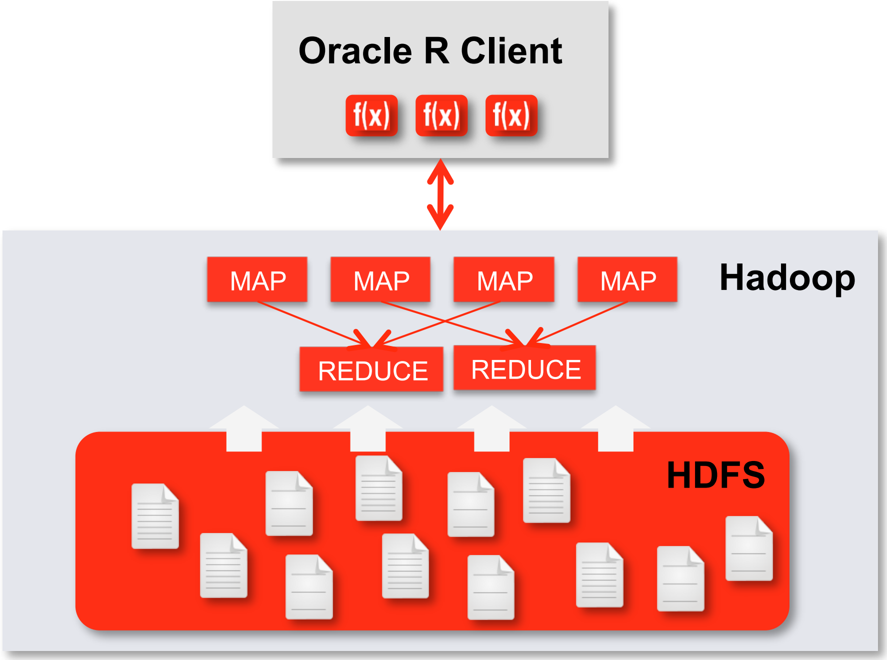
Load speed:
~~12~~ 15 TB/hour

- High speed load from Hadoop to Oracle Database
- Access data on HDFS from Oracle Database
- Aggregate data from both Hadoop and Oracle Database

Oracle R Connector for Hadoop

R Analytics leveraging Hadoop and HDFS

5x faster in BDC 3.0



Linearly Scale a Robust Set of R Algorithms

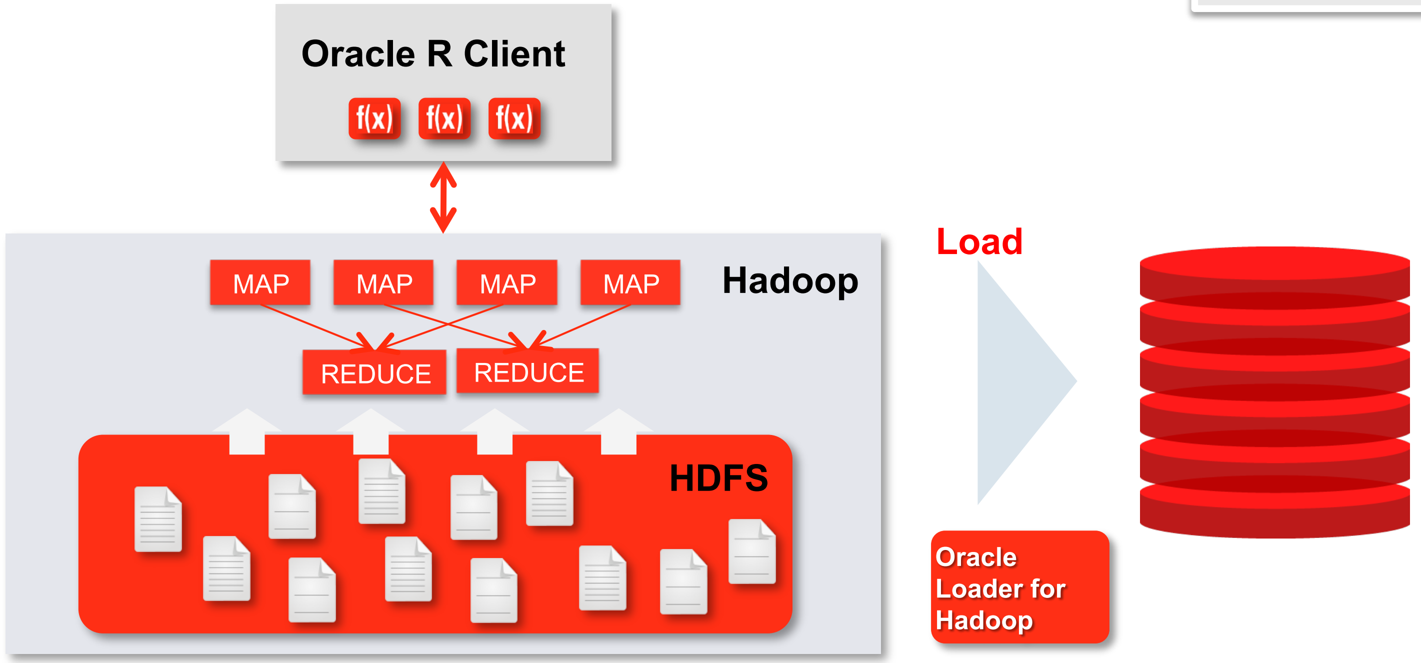
Leverage MapReduce for R Calculations

Compute Intensive Parallelism for Simulations

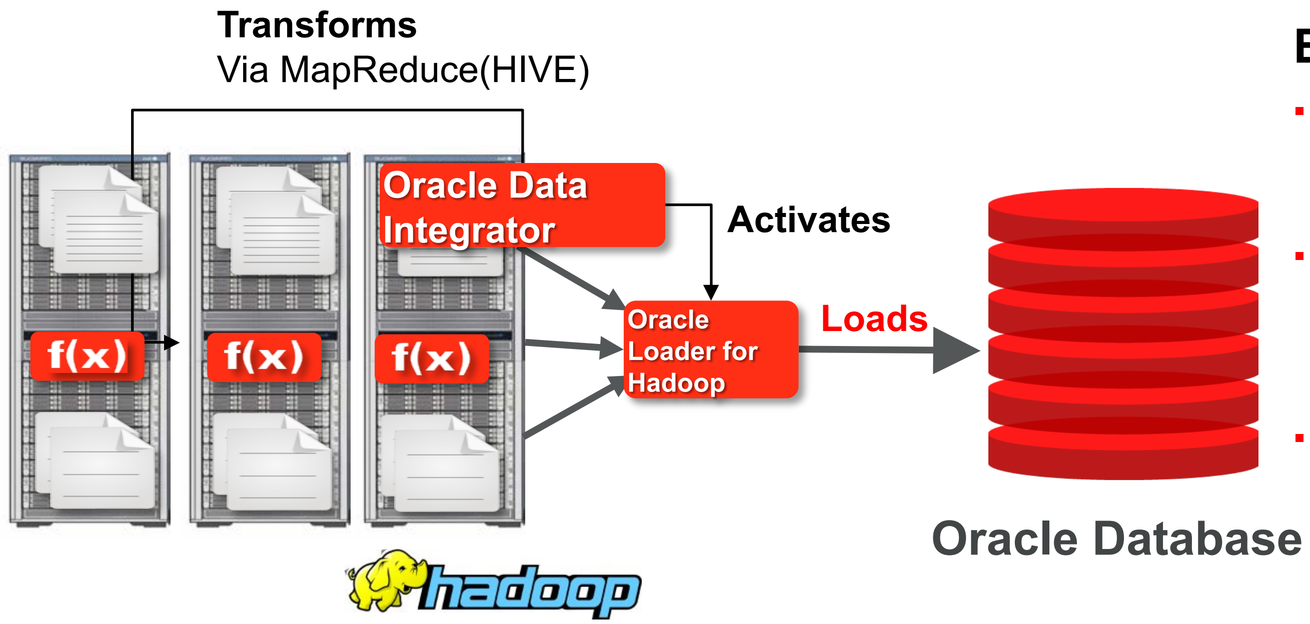
Oracle R Connector for Hadoop

R Analytics leveraging Hadoop and HDFS

5x faster in BDC 3.0



Oracle Data Integrator Application Adapters for Hadoop



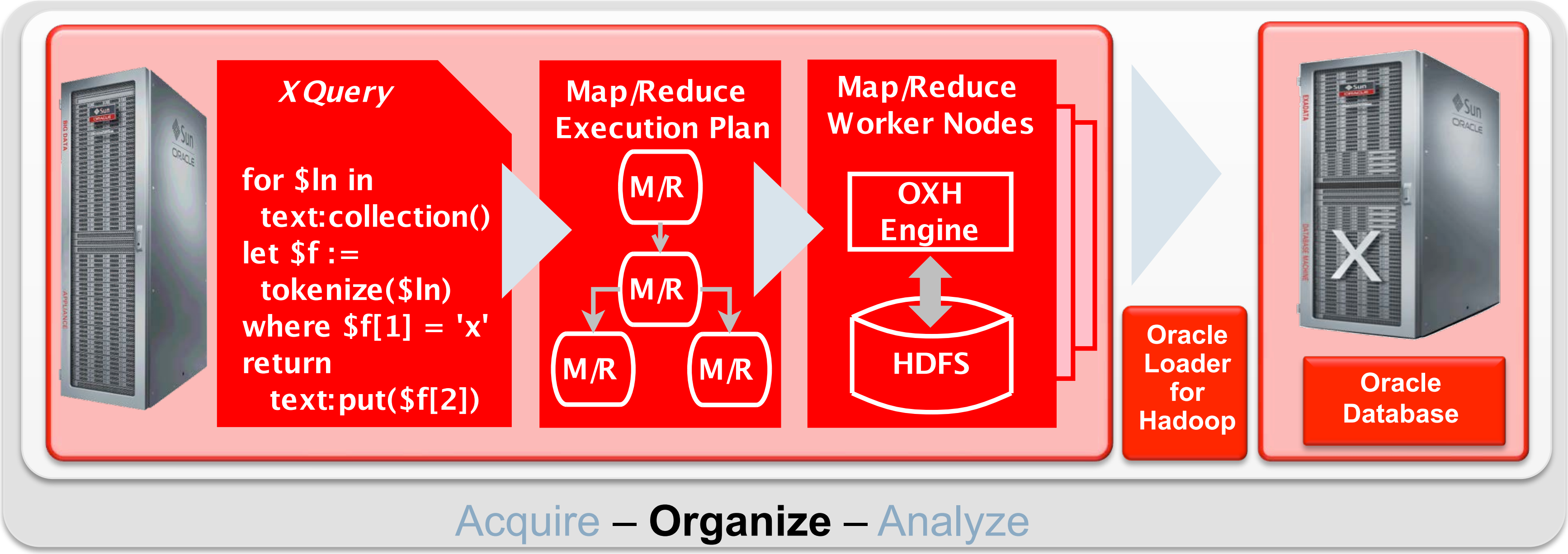
Benefits

- Consistent tooling across BI/DW, SOA, Integration and Big Data
- Reduce complexities of processing Hadoop through graphical tooling
- Improves productivity when processing Big Data (Structured + Unstructured)

Improving Productivity and Efficiency for Big Data

Announcing: Oracle XQuery for Hadoop (OXH)

- OXH is a transformation engine for Big Data
- XQuery language executed on the Map/Reduce framework



XQuery

```
for $ln in
  text:collection()
let $f :=
  tokenize($ln)
where $f[1] = 'x'
return
   text:put($f[2])
```

Oracle Loader for Hadoop Oracle SQL Connector for HDFS

High speed load from Hadoop to Oracle Database

Load Data into the Database

Two Options

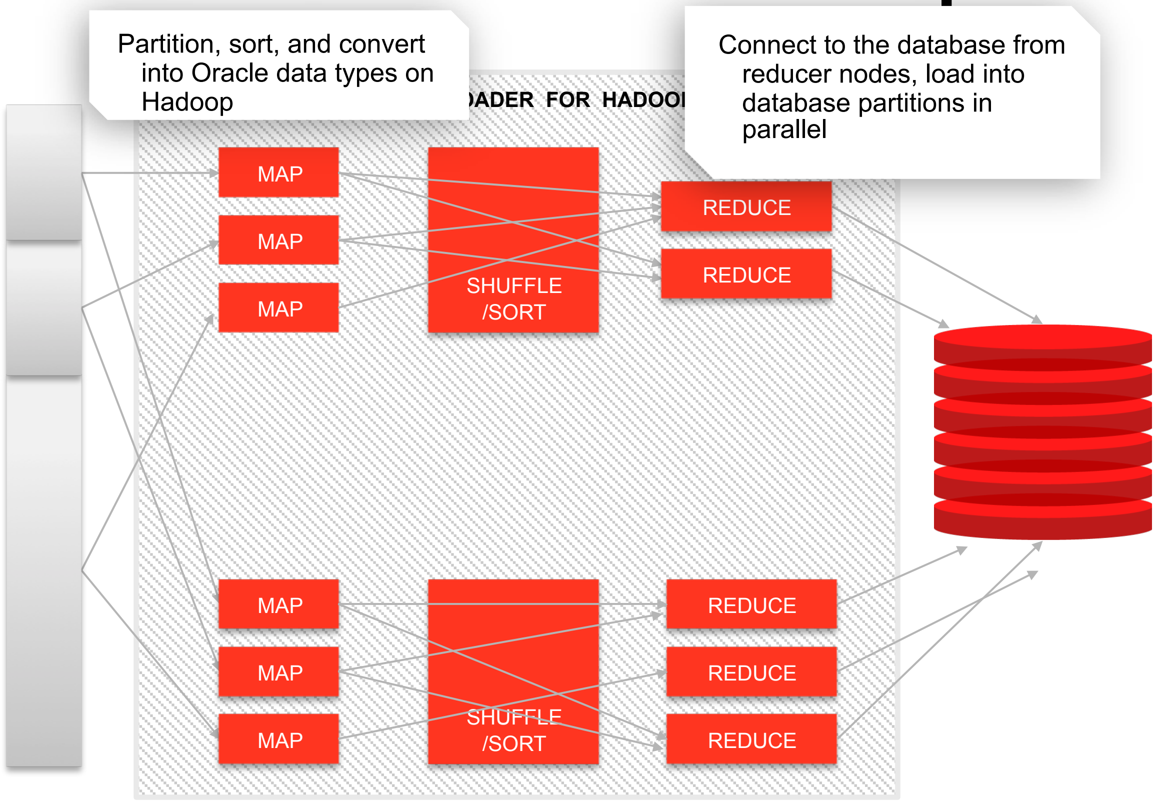
- Oracle Loader for Hadoop
  - Map Reduce job transforms data on Hadoop into Oracle-ready data types
  - Use more Hadoop compute resources
- Oracle SQL Connector for HDFS
  - Oracle SQL access to data on Hadoop via external tables
  - Use more database compute resources
  - Includes option to query in-place

Performance

- 15 TB / HOUR
- *25* TIMES FASTER THAN THIRD PARTY PRODUCTS
- REDUCED DATABASE CPU USAGE IN COMPARISON

	Oracle Loader for Hadoop	Oracle SQL Connector for HDFS
Use Case	Continuous or frequent load into production database, requiring reduced use of database CPU resources	Bulk load of large volumes of data Uses more database CPU resources
Input Data Formats	Load various types of input data: HBase, JSON files, Weblogs, sequence files, custom formats, etc.	Load text (HDFS files, and Hive table files) Load Oracle Data Pump files: *Generated by Oracle Loader for Hadoop from HBase, JSON files, Weblogs, sequence files, custom formats, etc.*
Functionality	Load	Load and also query in place (Note: Query requires full table scans since data files are external to the database)
Performance	Uses more time on Hadoop for pre-processing data.	End-to-end time is faster because no time is spent processing on Hadoop. Trade-off is more database CPU resources are used.
Usability	Likely to be preferred by Hadoop developers	Likely to be preferred by Oracle developers

Oracle Loader for Hadoop



Partition, sort, and convert into Oracle data types on Hadoop

Connect to the database from reducer nodes, load into database partitions in parallel

Features

Offloads data pre-processing from the database server to Hadoop

Works with a range of input data formats

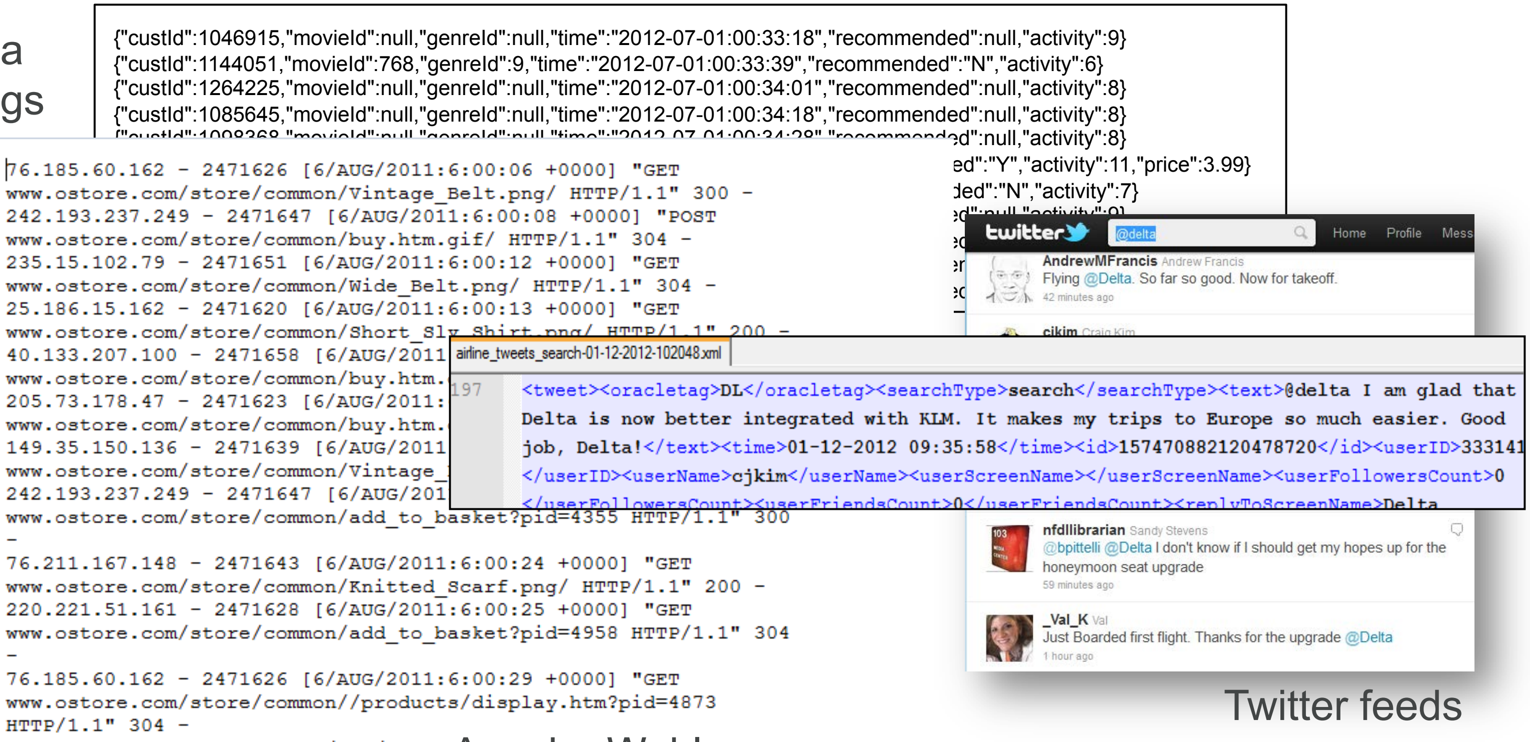
Automatic balancing in case of skew in input data

Online and offline modes

Data Samples

Sensor data
Machine logs

JSON files



```
{"custId":1046915,"movieId":null,"genreId":null,"time":"2012-07-01:00:33:18","recommended":null,"activity":9}
{"custId":1144051,"movieId":768,"genreId":9,"time":"2012-07-01:00:33:39","recommended":"N","activity":6}
{"custId":1264225,"movieId":null,"genreId":null,"time":"2012-07-01:00:34:01","recommended":null,"activity":8}
{"custId":1085645,"movieId":null,"genreId":null,"time":"2012-07-01:00:34:18","recommended":null,"activity":8}
{"custId":1098368,"movieId":null,"genreId":null,"time":"2012-07-01:00:34:28","recommended":null,"activity":8}
ed":"Y","activity":11,"price":3.99}
ded":"N","activity":7}
ed":null,"activity":9}
```

```
76.185.60.162 - 2471626 [6/AUG/2011:6:00:06 +0000] "GET
www.ostore.com/store/common/Vintage_Belt.png/ HTTP/1.1" 300 -
242.193.237.249 - 2471647 [6/AUG/2011:6:00:08 +0000] "POST
www.ostore.com/store/common/buy.htm.gif/ HTTP/1.1" 304 -
235.15.102.79 - 2471651 [6/AUG/2011:6:00:12 +0000] "GET
www.ostore.com/store/common/Wide_Belt.png/ HTTP/1.1" 304 -
25.186.15.162 - 2471620 [6/AUG/2011:6:00:13 +0000] "GET
www.ostore.com/store/common/Short_Sly_Shirt.png/ HTTP/1.1" 200 -
40.133.207.100 - 2471658 [6/AUG/2011
www.ostore.com/store/common/buy.htm.
205.73.178.47 - 2471623 [6/AUG/2011:
www.ostore.com/store/common/buy.htm.
149.35.150.136 - 2471639 [6/AUG/2011
www.ostore.com/store/common/Vintage_
242.193.237.249 - 2471647 [6/AUG/201
www.ostore.com/store/common/add_to_basket?pid=4355 HTTP/1.1" 300
-
76.211.167.148 - 2471643 [6/AUG/2011:6:00:24 +0000] "GET
www.ostore.com/store/common/Knitted_Scarf.png/ HTTP/1.1" 200 -
220.221.51.161 - 2471628 [6/AUG/2011:6:00:25 +0000] "GET
www.ostore.com/store/common/add_to_basket?pid=4958 HTTP/1.1" 304
-
76.185.60.162 - 2471626 [6/AUG/2011:6:00:29 +0000] "GET
www.ostore.com/store/common//products/display.htm?pid=4873
HTTP/1.1" 304 -
```

airline_tweets_search-01-12-2012-102048.xml

```
<tweet><oracletag>DL</oracletag><searchType>search</searchType><text>@delta I am glad that
Delta is now better integrated with KLM. It makes my trips to Europe so much easier. Good
job, Delta!</text><time>01-12-2012 09:35:58</time><id>157470882120478720</id><userID>333141
</userID><userName>cjkim</userName><userScreenName></userScreenName><userFollowersCount>0
</userFollowersCount><userFriendsCount>0</userFriendsCount><replyToScreenName>Delta
```

AndrewMFrancis Andrew Francis
Flying @Delta. So far so good. Now for takeoff.
42 minutes ago

cjkim Craig Kim

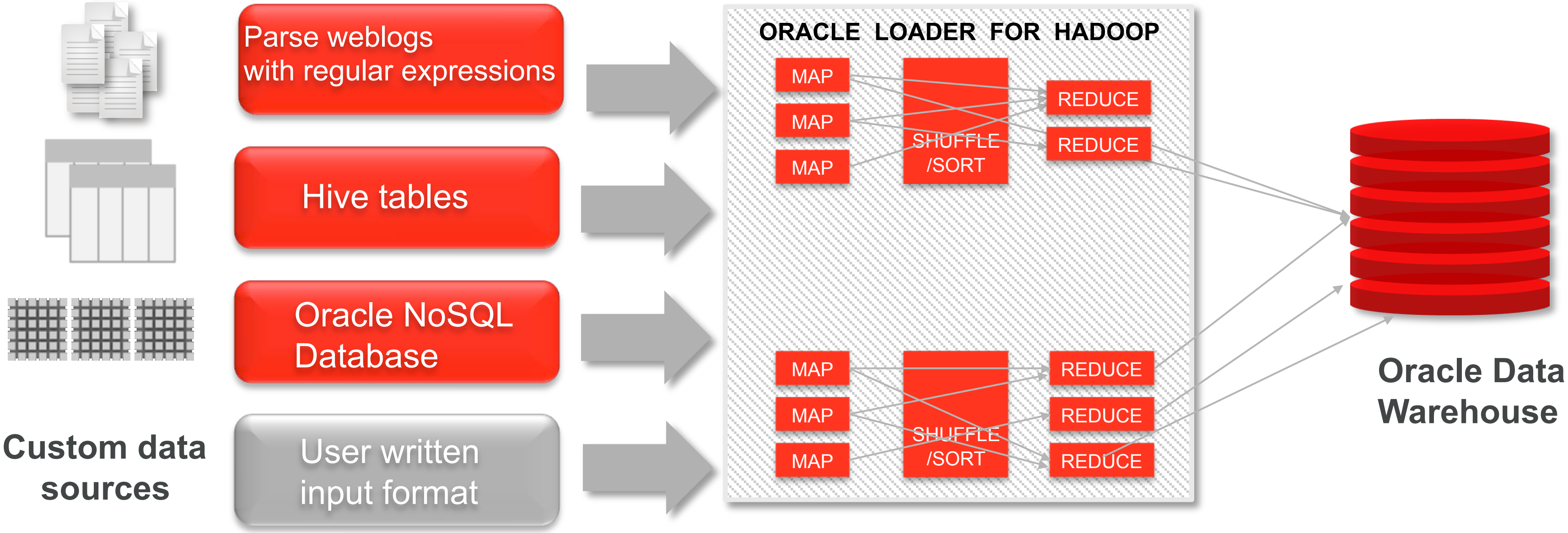
nfdllibrarian Sandy Stevens
@bpittelli @Delta I don't know if I should get my hopes up for the honeymoon seat upgrade
59 minutes ago

_Val_K Val
Just Boarded first flight. Thanks for the upgrade @Delta
1 hour ago

Twitter feeds

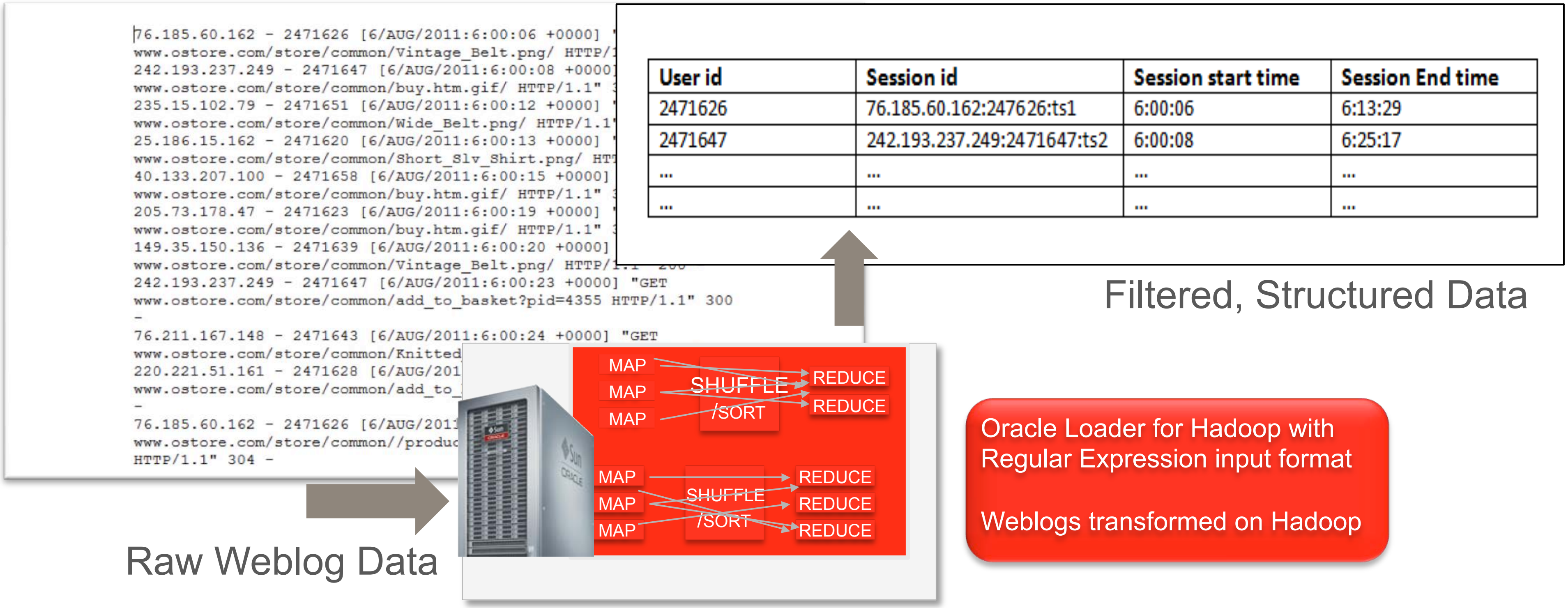
Apache Weblogs

1. Load by Reading Data through Input Format Interface



Load Weblog Data

Using Regular Expression Input Format



```
76.185.60.162 - 2471626 [6/AUG/2011:6:00:06 +0000] "
www.ostore.com/store/common/Vintage_Belt.png/ HTTP/1
242.193.237.249 - 2471647 [6/AUG/2011:6:00:08 +0000]
www.ostore.com/store/common/buy.htm.gif/ HTTP/1.1" 3
235.15.102.79 - 2471651 [6/AUG/2011:6:00:12 +0000] "
www.ostore.com/store/common/Wide_Belt.png/ HTTP/1.1"
25.186.15.162 - 2471620 [6/AUG/2011:6:00:13 +0000] "
www.ostore.com/store/common/Short_Slv_Shirt.png/ HTT
40.133.207.100 - 2471658 [6/AUG/2011:6:00:15 +0000]
www.ostore.com/store/common/buy.htm.gif/ HTTP/1.1" 3
205.73.178.47 - 2471623 [6/AUG/2011:6:00:19 +0000] "
www.ostore.com/store/common/buy.htm.gif/ HTTP/1.1" 3
149.35.150.136 - 2471639 [6/AUG/2011:6:00:20 +0000]
www.ostore.com/store/common/Vintage_Belt.png/ HTTP/1.1" 200
242.193.237.249 - 2471647 [6/AUG/2011:6:00:23 +0000] "GET
www.ostore.com/store/common/add_to_basket?pid=4355 HTTP/1.1" 300
-
76.211.167.148 - 2471643 [6/AUG/2011:6:00:24 +0000] "GET
www.ostore.com/store/common/Knitted
220.221.51.161 - 2471628 [6/AUG/201
www.ostore.com/store/common/add_to_
-
76.185.60.162 - 2471626 [6/AUG/2011
www.ostore.com/store/common//produc
HTTP/1.1" 304 -
```

User id	Session id	Session start time	Session End time
2471626	76.185.60.162:247626:ts1	6:00:06	6:13:29
2471647	242.193.237.249:2471647:ts2	6:00:08	6:25:17
...	...	...	...
...	...	...	...

Filtered, Structured Data

Raw Weblog Data

Oracle Loader for Hadoop with Regular Expression input format

Weblogs transformed on Hadoop

Submitting an Oracle Loader for Hadoop Job

MyConf.xml

InputFormat:

```
 <property>
mapreduce.inputformat.class
</property>
<value>RegExInputFormat</value>
```

Database connection information

Target table name/schema

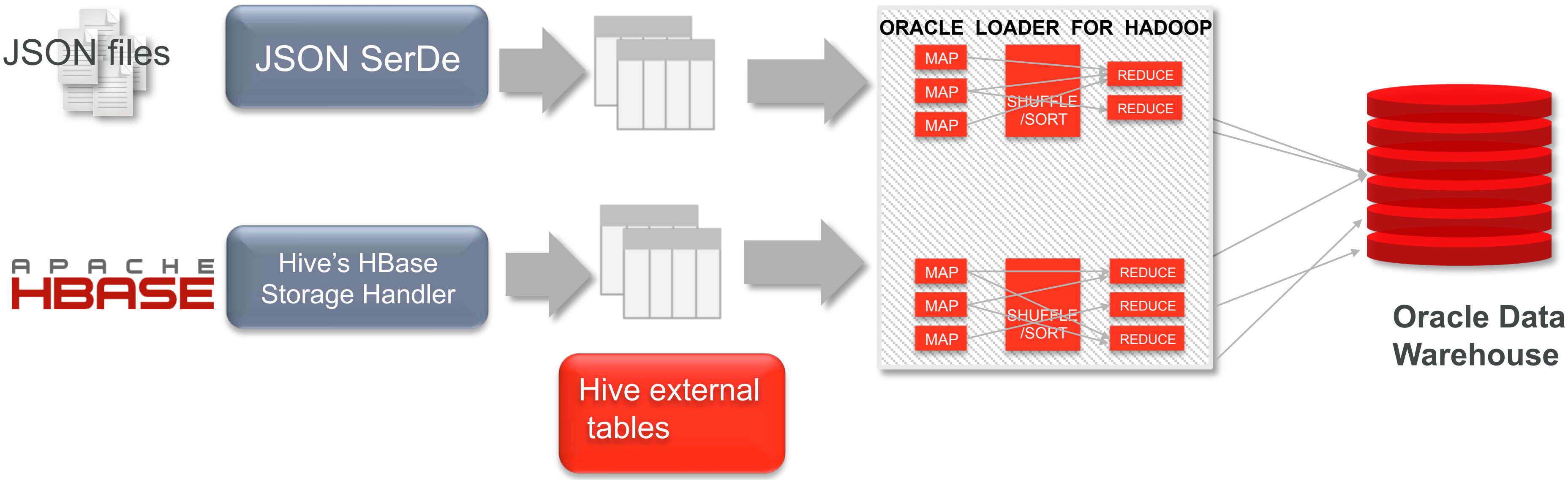
…

```
>hadoop jar                        \
$OLH_HOME/jlib/oraloader.jar       \
oracle.hadoop.loader.OraLoader     \
-conf MyConf.xml
```

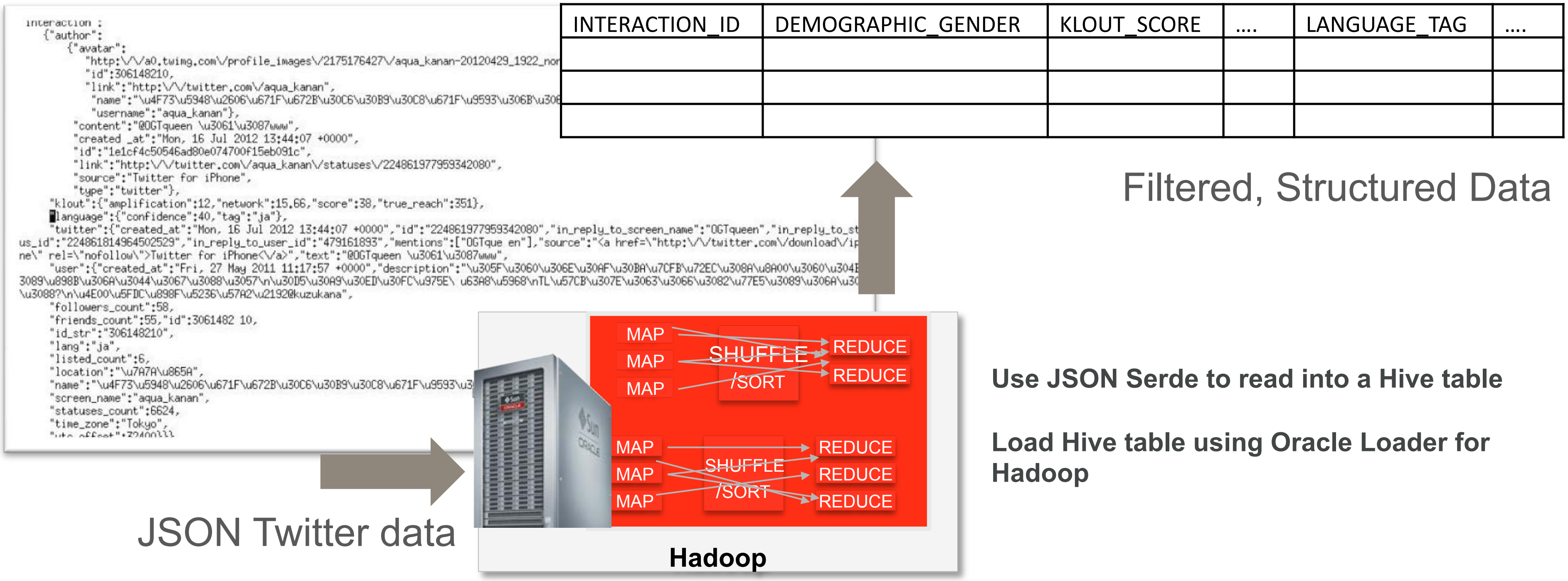
User-written Input Format Implementation

- Oracle Loader for Hadoop reads Avro IndexedRecords
- Input format implementation should read input data records and put into an Avro IndexedRecord
- Sample input format implementation shipped with Oracle Loader for Hadoop kit
- Users can implement input formats for HBase, nosql data stores, custom formats, etc.

2. Use OLH's Connectivity to Hadoop Technologies



Process JSON Files and Load into Database



```
interaction":
   {"author":
       {"avatar":
          "http:\/\/a0.twimg.com\/profile_images\/2175176427\/aqua_kanan-20120429_1922_nor
          "id":306148210,
          "link":"http:\/\/twitter.com\/aqua_kanan",
           "name":"\u4F73\u5948\u2606\u671F\u672B\u30C6\u30B9\u30C8\u671F\u9593\u306B\u306
           "username":"aqua_kanan"},
        "content":"@OGTqueen \u3061\u3087www",
        "created _at":"Mon, 16 Jul 2012 13:44:07 +0000",
        "id":"1e1cf4c50546ad80e074700f15eb091c",
        "link":"http:\/\/twitter.com\/aqua_kanan\/statuses\/224861977959342080",
        "source":"Twitter for iPhone",
        "type":"twitter"},
    "klout":{"amplification":12,"network":15.66,"score":38,"true_reach":351},
    "language":{"confidence":40,"tag":"ja"},
    "twitter":{"created_at":"Mon, 16 Jul 2012 13:44:07 +0000","id":"224861977959342080","in_reply_to_screen_name":"OGTqueen","in_reply_to_st
us_id":"224861814964502529","in_reply_to_user_id":"479161893","mentions":["OGTque en"],"source":"<a href=\"http:\/\/twitter.com\/download\/ip
ne\" rel=\"nofollow\">Twitter for iPhone<\/a>","text":"@OGTqueen \u3061\u3087www",
    "user":{"created_at":"Fri, 27 May 2011 11:17:57 +0000","description":"\u305F\u3060\u306E\u30AF\u30BA\u7CFB\u72EC\u308A\u8A00\u3060\u304E
3089\u898B\u306A\u3044\u3067\u3088\u3057\n\u30D5\u30A9\u30ED\u30FC\u975E\ u63A8\u5968\nTL\u57CB\u307E\u3063\u3066\u3082\u77E5\u3089\u306A\u30
\u3088?\n\u4E00\u5FDC\u898F\u5236\u57A2\u2192@kuzukana",
    "followers_count":58,
    "friends_count":55,"id":3061482 10,
    "id_str":"306148210",
    "lang":"ja",
    "listed_count":6,
    "location":"\u7A7A\u865A",
    "name":"\u4F73\u5948\u2606\u671F\u672B\u30C6\u30B9\u30C8\u671F\u9593\u3
    "screen_name":"aqua_kanan",
    "statuses_count":6624,
    "time_zone":"Tokyo",
    "utc_offset":32400}}}
```

INTERACTION_ID	DEMOGRAPHIC_GENDER	KLOUT_SCORE	….	LANGUAGE_TAG	….

Filtered, Structured Data

JSON Twitter data

Use JSON Serde to read into a Hive table

Load Hive table using Oracle Loader for Hadoop

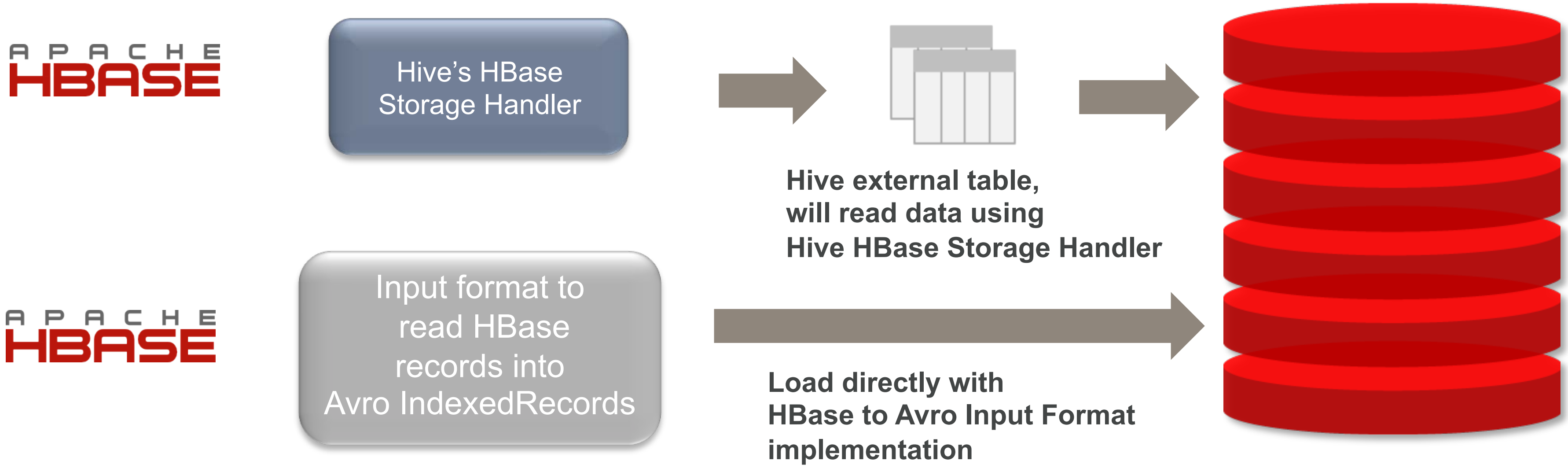
Use JSON SerDe to Access Through Hive

```
CREATE EXTERNAL TABLE tweets
  (……
  )
ROW FORMAT SERDE ‘com.cloudera.serde.JSONSerDe’
STORED AS TEXTFILE
LOCATION ‘/user/oracle/tweets’
```

- Load data from Hive tables using Oracle Loader for Hadoop

Load Data from HBase

Use Hive HBase Storage Handler or Input Format Implementation



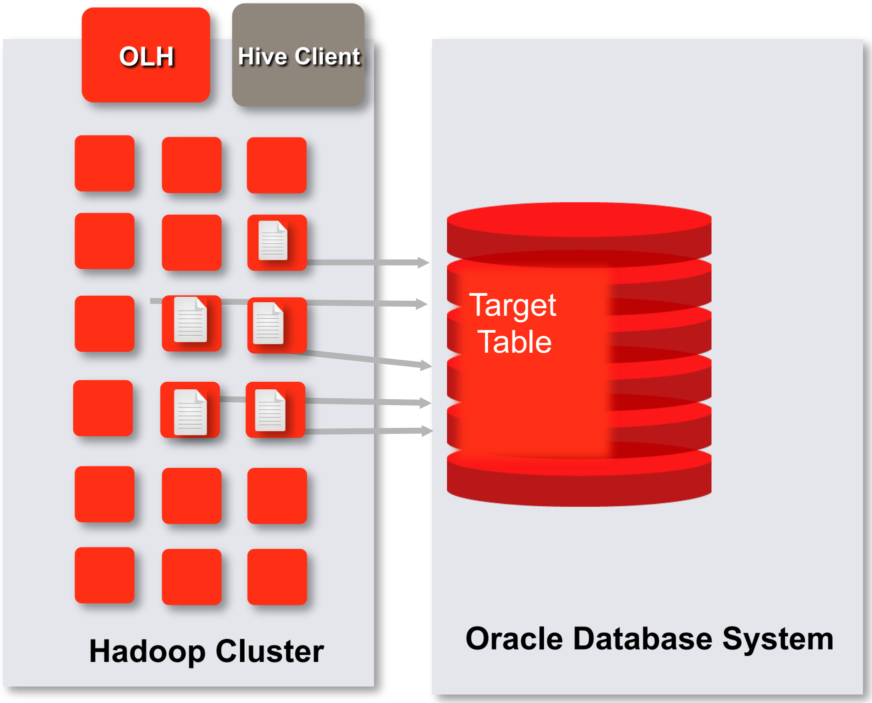
Hive's HBase Storage Handler

Hive external table, will read data using Hive HBase Storage Handler

Input format to read HBase records into Avro IndexedRecords

Load directly with HBase to Avro Input Format implementation

Install of Oracle Loader for Hadoop



Key Benefits

- Load from a wide range of input sources
- Performance
  - 10x faster than comparable third party products
- Offload database server processing on to Hadoop
  - Reduced impact on database during load
- Easy to use
- Developed and supported by Oracle

Oracle SQL Connector for HDFS

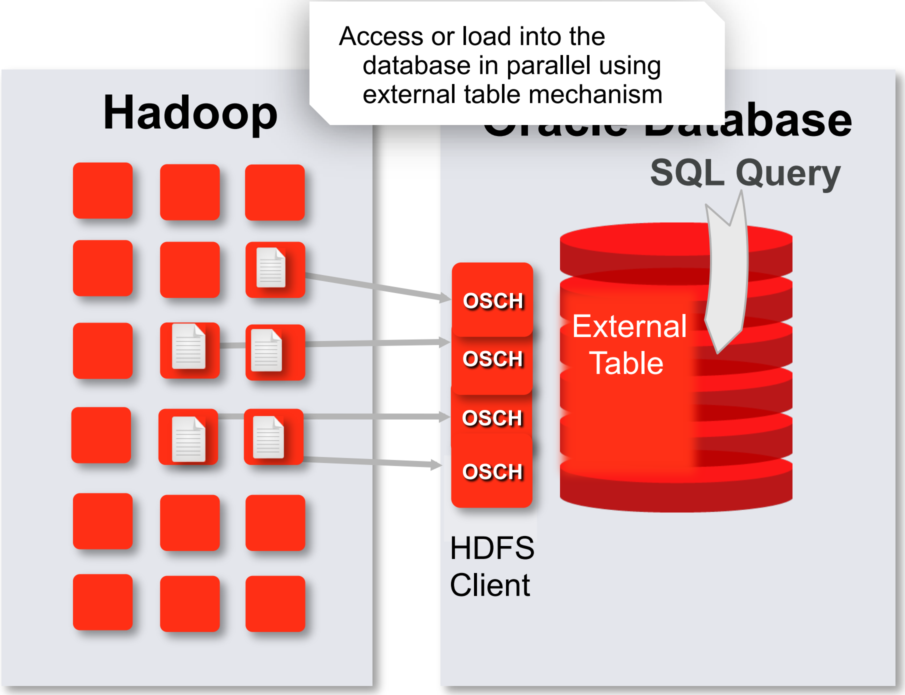
Use Oracle SQL to Load or Access Data on HDFS

Load into the database using SQL

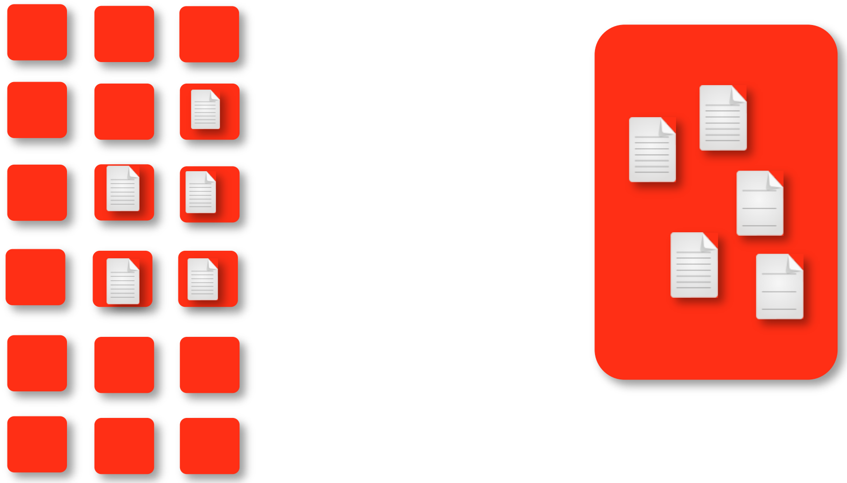
Option to access and analyze data in place on HDFS

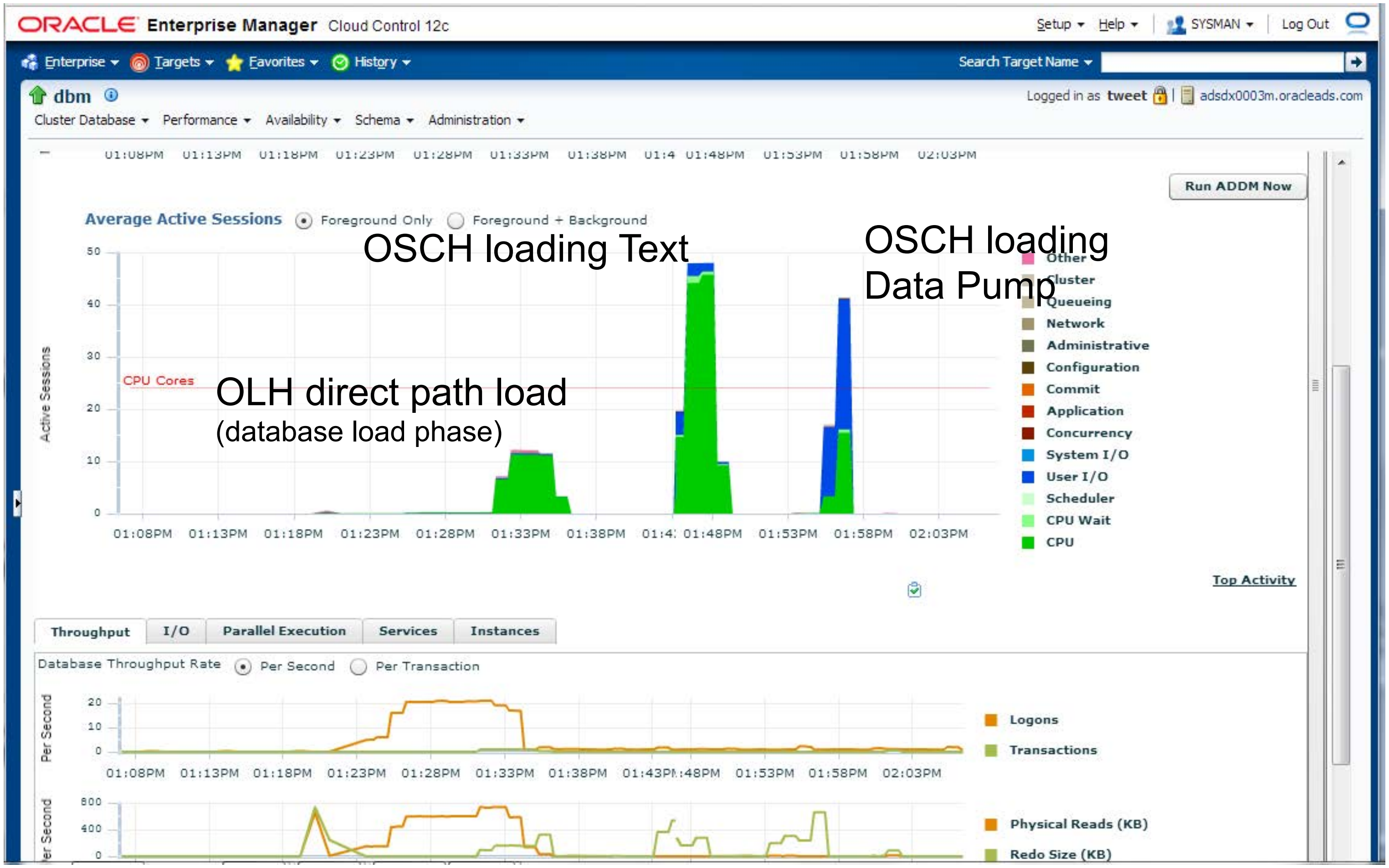
Access Hive (internal and external) tables and HDFS files

Automatic load balancing to maximize performance



Access or load into the database in parallel using external table mechanism


Hadoop Cluster

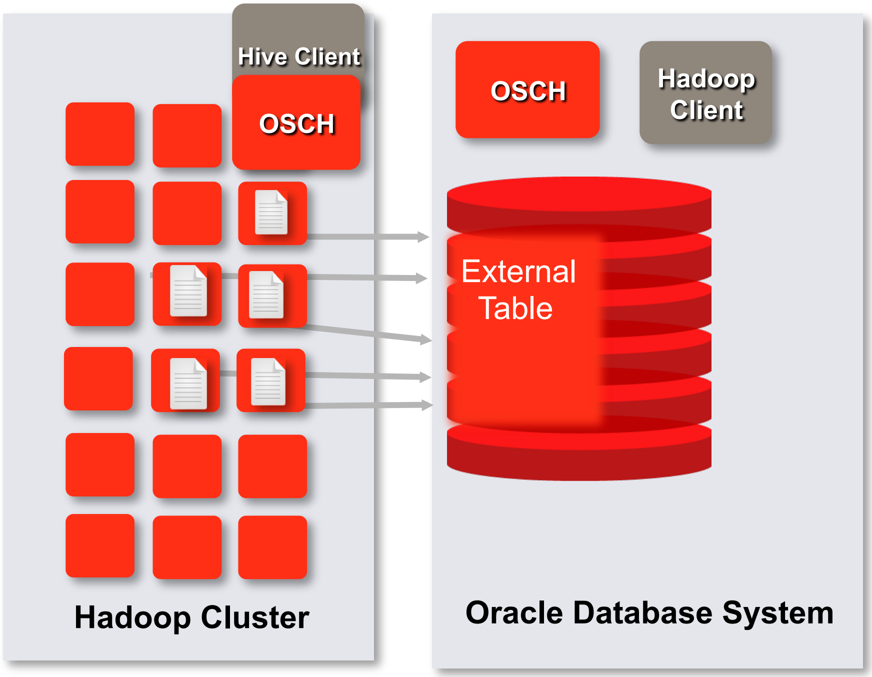

ORACLE Enterprise Manager Cloud Control 12c
Setup ▾ Help ▾ SYSMAN ▾ Log Out
Enterprise ▾ Targets ▾ Favorites ▾ History ▾
Search Target Name ▾
dbm
Logged in as tweet | adsdx0003m.oracleads.com
Cluster Database ▾ Performance ▾ Availability ▾ Schema ▾ Administration ▾
01:08PM 01:13PM 01:18PM 01:23PM 01:28PM 01:33PM 01:38PM 01:4 01:48PM 01:53PM 01:58PM 02:03PM
Run ADDM Now
Average Active Sessions
● Foreground Only ○ Foreground + Background
Active Sessions
50
40
30
20
10
0
CPU Cores
01:08PM 01:13PM 01:18PM 01:23PM 01:28PM 01:33PM 01:38PM 01:4: 01:48PM 01:53PM 01:58PM 02:03PM
Other
Cluster
Queueing
Network
Administrative
Configuration
Commit
Application
Concurrency
System I/O
User I/O
Scheduler
CPU Wait
CPU
Top Activity
Throughput
I/O
Parallel Execution
Services
Instances
Database Throughput Rate
● Per Second ○ Per Transaction
Per Second
20
10
0
01:08PM 01:13PM 01:18PM 01:23PM 01:28PM 01:33PM 01:38PM 01:43PM:48PM 01:53PM 01:58PM 02:03PM
Logons
Transactions
Per Second
800
400
0
Physical Reads (KB)
Redo Size (KB)

OSCH loading Text

OSCH loading Data Pump

OLH direct path load
(database load phase)

Install of Oracle SQL Connector for HDFS



Oracle SQL Connector for HDFS

- Load data from external table with Oracle SQL
  - `INSERT INTO <tablename> AS SELECT * FROM <external tablename>`

- Access data in-place on HDFS with Oracle SQL
  - Note: No indexes, no partitioning, so queries are a full table scan

- Data files are read in parallel
  - Ex: If there are 96 data files and the database can support 96 PQ slaves, all 96 files can be read in parallel
  - OSCH automatically balances the load across the PQ slaves

Oracle SQL Connector for HDFS

- Generates definition and creates external table pointing to data files on HDFS
- When external table is accessed with SQL, data is streamed from HDFS

```
CREATE TABLE "TWEET".”HIVE_ORA_EXT_TAB“
(
 "INTERACTION_ID"                  VARCHAR2(4000),
 "DEMOGRAPHIC_GENDER"              VARCHAR2(4000),
 "KLOUT_SCORE"                     INTEGER,
 "KLOUT_AMPLIFICATION"             INTEGER,
 "KLOUT_NETWORK"                   VARCHAR2(4000),
 "KLOUT_TRUE_REACH"                VARCHAR2(4000),
 "LANGUAGE_TAG"                    VARCHAR2(4000),
 "LANGUAGE_CONFIDENCE"             INTEGER,
 "SALIENCE_CONTENT_SENTIMENT"      INTEGER,
 "DT"                              VARCHAR2(4000)
)
```

```
ORGANIZATION EXTERNAL
(
   TYPE ORACLE_LOADER
   DEFAULT DIRECTORY "ORACLETEST_DIR"
   ACCESS PARAMETERS
   ( …
   )
     …
    LOCATION
   (
     'osch-20130920093955-1225-1',
     'osch-20130920093955-1225-2',
     'osch-20130920093955-1225-3',
       …
    )
```

```
  Location Files contain URIs:
hdfs://…/user/hive/warehouse/dw_augmentation/000000_0
… … …
```

Oracle SQL Connector for HDFS

Input Data Formats

- Text files
- Hive tables
  - Internal and external tables
  - Text data
  - Oracle external table data types map to Hive table data types
- Oracle Data Pump files generated by Oracle Loader for Hadoop

Oracle SQL Connector for HDFS

Data Pump Files

- Oracle Data Pump: Binary format data file
- Oracle Loader for Hadoop generates Oracle Data Pump files for use by Oracle SQL Connector for HDFS
- Load of Oracle Data Pump files is more efficient – uses about 50% less database CPU
  - Hadoop does more of the work, transforming text data into binary data optimized for Oracle

Key Benefits

- Extremely fast load performance
  - 15 TB/hour from Oracle Big Data Appliance to Oracle Exadata
- Load data pump files for reduced database CPU usage
- Unique option to query HDFS data in-place
- Easy to use for Oracle DBAs and Hadoop developers
- Developed and supported by Oracle

Certification with Hadoop Versions

- Certified by Oracle
  - CDH 4.3, CDH 3
  - Apache Hadoop 1.0, 1.1
- Intel announces certification of their distribution at OOW
- Process for third party vendors to certify their distributions

Media Company

Real-time Analytics with 360 Customer View

Objectives

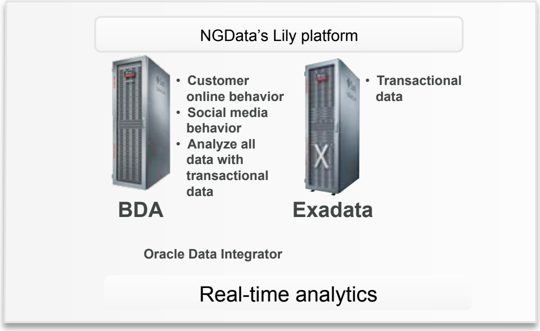
- Find the hidden value in large volumes of online and social media behavior, merged with data in transactional systems

Benefits

- Analysis in real-time instead with a two week lag
- Lower TCO and fast time to value
- BDA , connectors, database: integrated single system for all data for a simplified IT environment

Solution

- Starter rack BDA with connectors for integration of all data for full customer view
- Partner NGData's Lily platform
- Cost-effective storage on the BDA
- Real-time analytics of all data



NGData's Lily platform

- Customer online behavior
- Social media behavior
- Analyze all data with transactional data

- Transactional data

BDA

Exadata

Oracle Data Integrator

Real-time analytics

Financial Services

Risk and Finance

Objectives

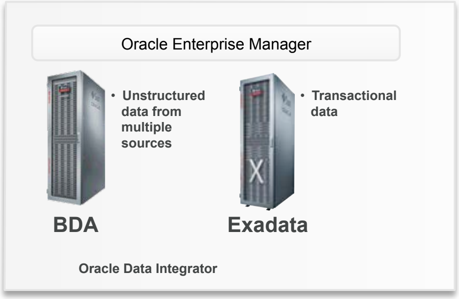
- Aggregate structured and unstructured data from multiple sources
- Scale to increasing volumes of data
- Consolidate existing silos of information for unified access to enterprise data

Solution

- Oracle Big Data Appliance, Oracle Exadata, Oracle Big Data Connectors, Oracle Enterprise Management for comprehensive technology solution
- Oracle Data Integrator for end-to-end data lineage
- Supports data and analytics for risk and finance

Benefits

- Fast and nimble way to get new data into ODS
- Deliver better SLAs to users
- Simplified architecture
- Additional storage and compute resources available for new development projects


Oracle Enterprise Manager
• Unstructured data from multiple sources
• Transactional data
BDA
Exadata
Oracle Data Integrator

Aggregate all data

Some New Features in BDC 2.3

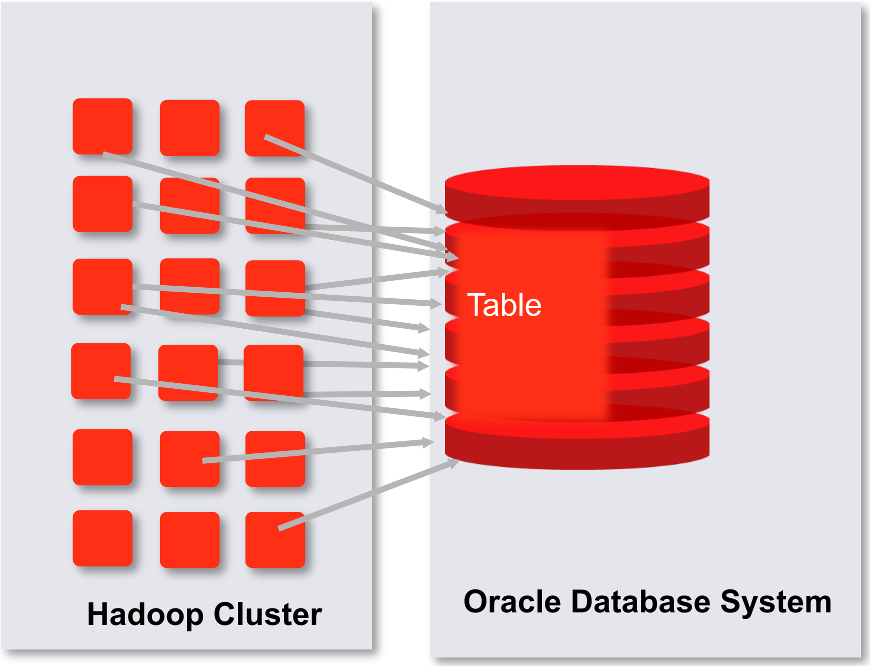
- Performance moves from 12 TB to 15 TB
- Ease of use and flexibility
  - Easier way to map columns to be loaded to target table columns
  - Per column override while mapping Hive column types to external table column types
- Works out-of-the-box with Kerberos authentication protocol

Performance Tuning

Performance Tuning

- Parallelism
- Network bandwidth
- Hadoop property values
- Database target table definition, tablespace parameters, session settings
- Using the sampler (for Oracle Loader for Hadoop)

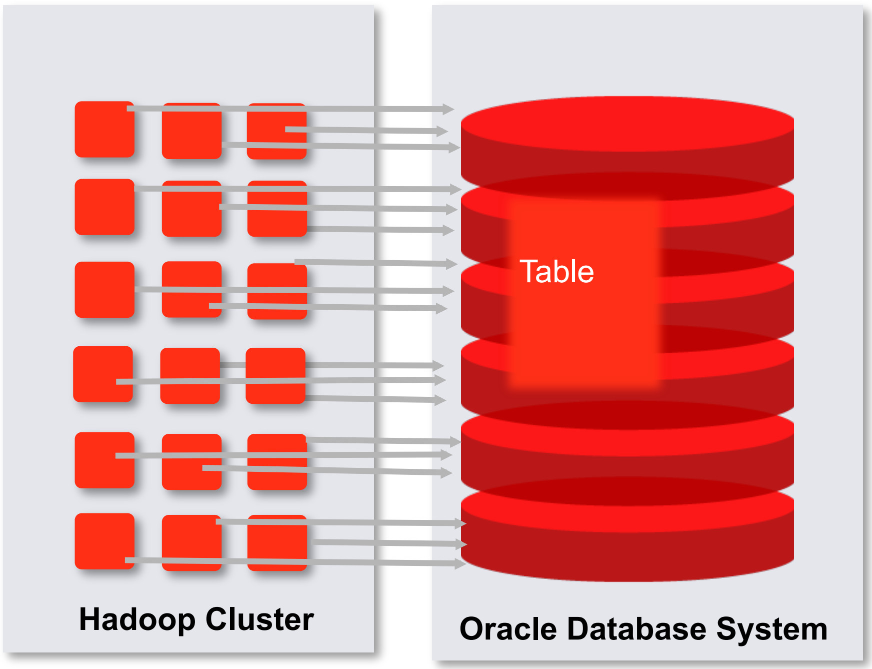
Key: Degree of Parallelism



Key: Degree of Parallelism

Number of reducer slots

OSCH: Number of location files



Number of CPU cores

OLH: Number of partitions in target table

Parallelism

- Oracle Loader for Hadoop
  - Reduce tasks load data in parallel to the database
  - Goal: Number of partitions in the database should be a multiple of number of reduce tasks

- Oracle SQL Connector for HDFS
  - `SQL> alter session enable parallel query` (or `DML`);
  - Number of location files in external table should be a multiple of DOP (which is determined by number of database cores)

Network Bandwidth

- Configure InfiniBand
  - Read and follow Oracle BDA Documentation

- Multi-homing for Hadoop
  - Hadoop needed to support multiple network interfaces to maximize use of InfiniBand bandwidth
  - Enabled by collaboration between Oracle and Cloudera, committed to Apache Hadoop by Cloudera

- For Oracle Loader for Hadoop, configure SDP

Hadoop Property Values

- Batch size: Number of records to be inserted in batch into the database
- Buffer size for direct path load
- Specifying when reduce tasks begin
- Reusing JVM
- …

Database Parameters

- Session parameters
  - Enable parallel query and DML
  - `SQL> alter session enable parallel query` (or `DML`);
- Table definition
  - For maximum throughput: NOLOGGING, PARALLEL, NOCOMPRESS
- Tablespace
  - Use ASM

Sampler for Oracle Loader for Hadoop

- Distributes load evenly among reducer tasks
  - Reduces slow down due to data skew
- Enable Sampling (by config parameter) for this automatic load balancing

Summary

- Oracle Loader for Hadoop and Oracle SQL Connector for HDFS are products for high speed loading from Hadoop to Oracle Database
  - Cover a range of use cases
  - Several input sources
  - Flexible, easy-to-use, developed and supported by Oracle

- The fastest load option loads at 15 TB/hour

Hardware and Software

ORACLE®

Engineered to Work Together

ORACLE®