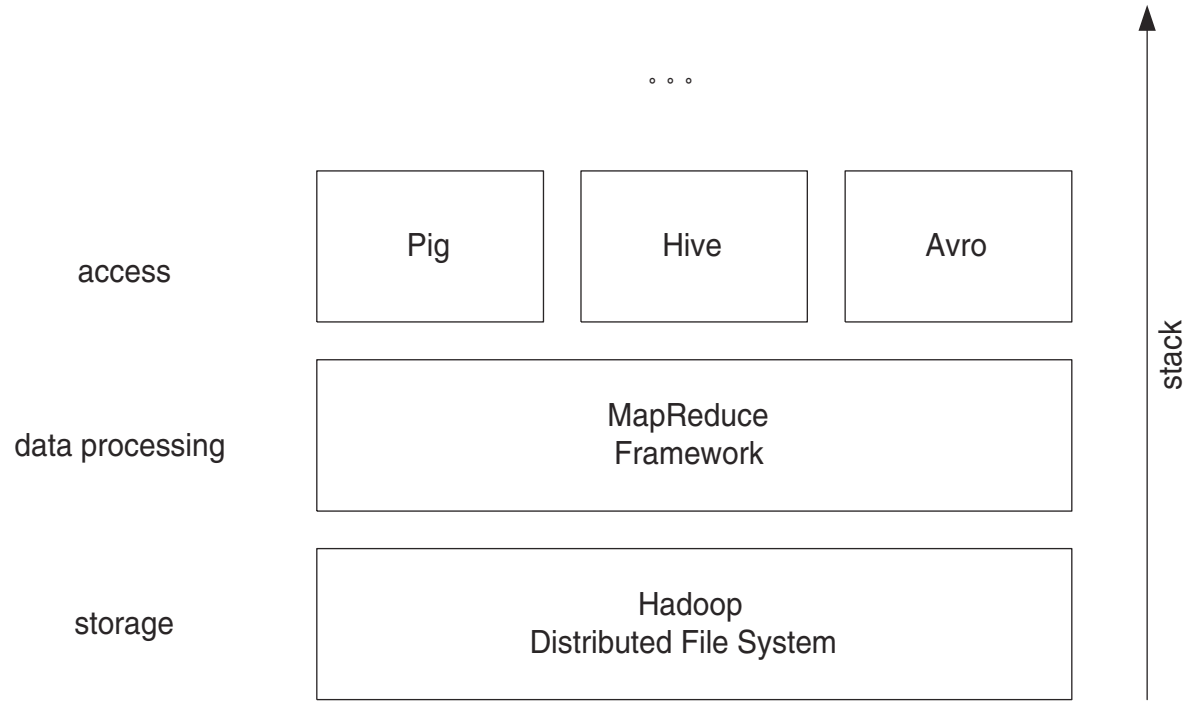


Oracle Big Data Connectors

Objectives :

- **introduce Oracle Big Data Connectors**
- **compare Connectors**
- **[some] user experiences**

Oracle and Cloudera



Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - Hadoop

HDFS [Hadoop Distributed File System] :

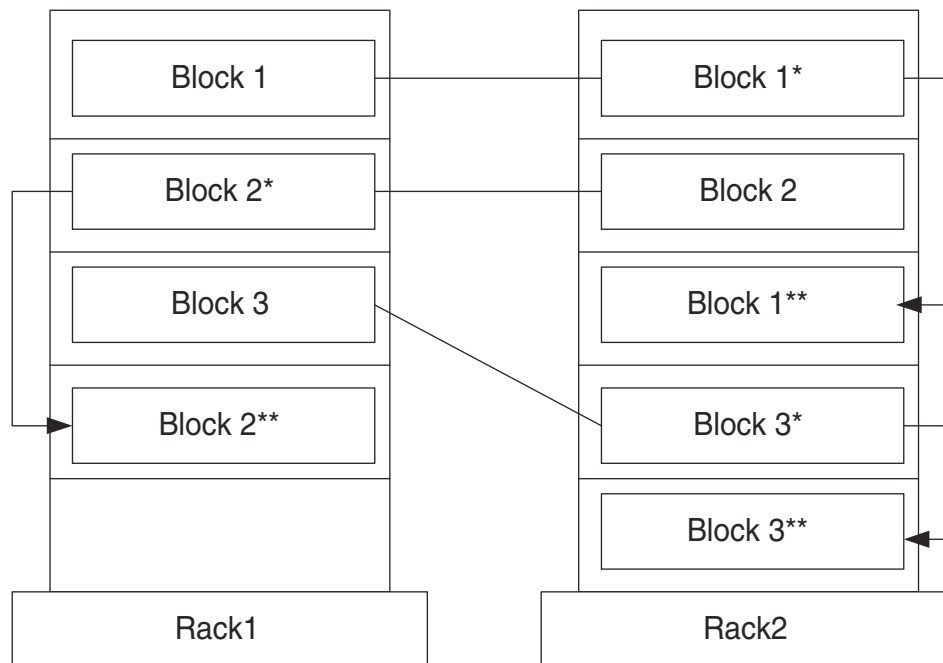
- data storage platform, data stored in clusters on 'inexpensive' hardware (eg. MPP clusters)
- data to be stored in 'known', 'common' files and directories; physically stored in blocks, replicated for fault tolerance [block size, replication factor => configurable]
- main components:
 - master node, or NameNode - one or more [depending on...]
 - workhorses, or DataNode(s)
- command-line interface 'hadoop fs' or 'hadoop dfs' - not POSIX compliant, eg.

```
hadoop fs [-fs <local | file system URI>] [-conf <configuration file>]
  [-D <property=value>] [-ls <path>] [-lsr <path>] [-du <path>]
  [-mv <src> <dst>] [-cp <src> <dst>] [-rm [-skipTrash] <src>]
  [-put <localsrc> ... <dst>] [-copyFromLocal <localsrc> ... <dst>]
  [-moveFromLocal <localsrc> ... <dst>] [-get [-ignoreCrc] [-crc] <src> <localdst>]
  [-copyToLocal [-ignoreCrc] [-crc] <src> <localdst>] [-moveToLocal <src> <localdst>]
  [-mkdir <path>] [-report] [-setrep [-R] [-w] <rep> <path/file>]
  ...
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - Hadoop (II)



Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - MapReduce

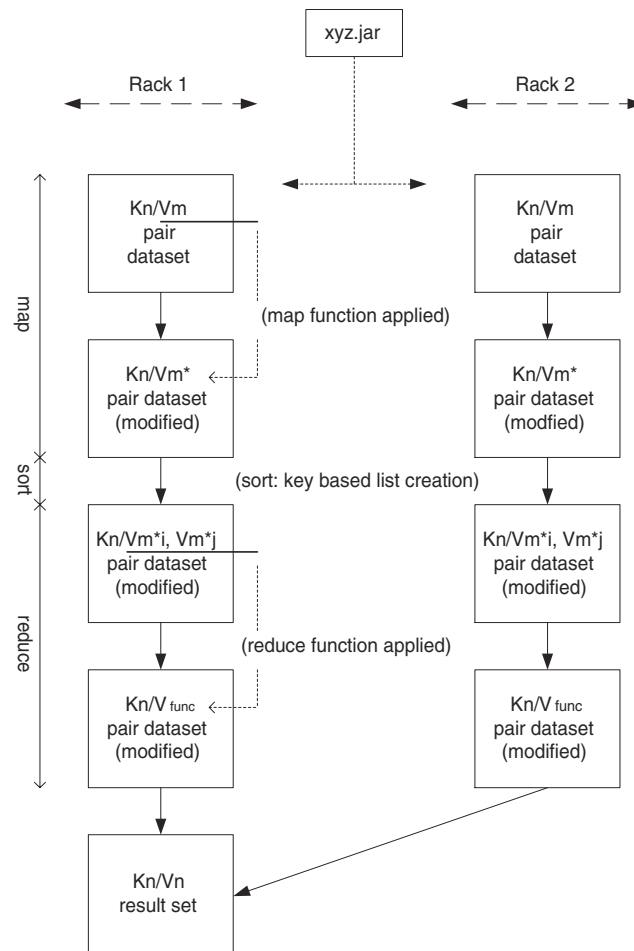
MapReduce :

- processing of datasets across Hadoop nodes/clusters storing key/value pairs
- a compiled .jar file is loaded by Hadoop into HDFS and distributed to the relevant DataNodes, where tasks are executed
- independant *Map* ...
[mapping the input across DataNodes of a cluster]
[data sets are split into smaller data sets; each map tasks *treats* one smaller data set]
[*treats*: applies a mapping function]
- and *Reduce* processes
[result of the Map process reduced to one single, new key/value pair, based on a reduce function]
[reduce tasks distributed across the cluster]
- fault tolerant [in data nodes]

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - MapReduce (II)



Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - Pig

Pig :

- 'high-level' programming language suited for massive parallel computing (based on MapReduce frameworks)
-> focuss on analysis, not on writing MapReduce programs!
- the Pig compiler produces sequences of MapReduce programs - high-level code written in a scripting language called PigLatin
- a PigLatin script:
 - LOADs data to be analysed
 - TRANSFORMs data (FILTER, JOIN, GROUP, ORDER)
 - DUMPs - STOREs the result
- run it from: Java, a script, or the Pig runtime grunt

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - Pig (II) - example

```
grunt> dividends = load 'MySecondDir/NYSE_dividends' as (exchange, symbol, date, dividend);
grunt> grouped = group dividends by symbol;
grunt> avg = foreach grouped generate group, AVG(dividends.dividend);
grunt> sorted = order avg by $1 DESC;
grunt> store sorted into 'average_dividend';
INFO [Thrd-22] org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process: 1
INFO [Thrd-31] org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process: 1
INFO [Thrd-42] org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process: 1
```

(1) NYSE_dividends is a Hadoop stored file

(2) computations are performed using standard libraries - custom libraries can be loaded (REGISTERed) if desired - jar file based

(3) output is stored in a Hadoop stored file, average_dividend

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - Hive

Hive :

- provides ad hoc query and data aggregation for Hadoop based data
- HiveQL: SQL-like language, broken down by the Hive service into MapReduce jobs executed across the Hadoop cluster
- can be run from:
 - Hive Shell, JDBC, ODBC, Hive Trift Clients ['remote' clients]
- long sequential scans, 'never' used for write operations

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - Hive (II) - Example

```
hive> create table dividends (exchange string, symbol string, year string, dividend string) row format delimited fields terminated by '\t';
```

```
OK
```

```
Time taken: 4.489 seconds
```

```
hive> show tables;
```

```
OK
```

```
dividends
```

```
Time taken: 0.169 seconds
```

```
hive> load data inpath '/user/biadmin/MySecondDir/NYSE_dividends' into table dividends;
```

```
Loading data to table default.dividends
```

```
OK
```

```
Time taken: 0.427 seconds
```

```
hive> select * from dividends;
```

```
OK
```

```
nyse  cpo    2009-12-30    0.14
nyse  cpo    2009-09-28    0.15
nyse  ccs     2009-07-04    0.413
nyse  ccs     2009-07-05    0.42
nyse  cif     2009-12-09    0.44
```

```
Time taken: 0.202 seconds
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Introduction - Avro

- **Avro :**

data serialisation

converts data into a binary format + schema, stored in a file

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

An efficient loader for fast data movement from a Hadoop cluster into an Oracle table.

- a MapReduce application
- the last stage in a MapReduce workflow, invoked through a command-line utility
- prepartitioning step - partitioned and non-partitioned tables are supported
- transformation step - prepares 'database-ready format'
- sorts records (primary key, user-specified columns)
- online and offline loads

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - Steps/Example

Oracle Loader steps:

1. reads/verifies input configuration
2. retrieves/verifies table and column metadata for the target table, used during the actual load process
3. internal configuration data and dependent Java libraries are stored in a distributed cache, available to map and reduce tasks throughout the cluster
4. submits MapReduce job(s)
5. create an integrated log file for the job after the map and reduce tasks are complete - oraloader-report.txt.

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - Installation, Setup

- **standard installation**

standard unzip, no specific installation routines required

- **once installed...**

- define a variable `$OLH_HOME` and set it to the installation directory
- add `$OLH_HOME/jlib/*` to the `HADOOP_CLASSPATH` variable
- [optional] add `$KVHOME/lib/kvstore.jar` to the `HADOOP_CLASSPATH` variable, if accessing Oracle NoSQL database

....

```
export OSCH_HOME=/home/biadmin/orahdfs-2.0.1
export OLH_HOME=/home/biadmin/oraloader-2.0.1-1
export KVHOME=/home/biadmin/kv-2.0.26
export JAVA_HOME=$BIGINSIGHTS_HOME/jdk
export HADOOP_HOME=$BIGINSIGHTS_HOME/IHC
export HADOOP_CONF_DIR=$BIGINSIGHTS_HOME/hadoop-conf
export HADOOP_CLASSPATH=$BIGINSIGHTS_HOME/IHC/hadoop-core-1.0.0.jar:$OSCH_HOME/jlib/*:$OLH_HOME/jlib/*:$KVHOME/lib/kvstore.jar
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - example (I)

```
hadoop jar $OLH_HOME/jlib/oraloader.jar oracle.hadoop.loader.OraLoader \  
-libjars $OLH_HOME/jlib/avro-1.6.3.jar,$OLH_HOME/jlib/oraloader-examples.jar \  
-D mapreduce.outputformat.class=oracle.hadoop.loader.lib.output.DelimitedTextOutputFormat \  
-D mapreduce.inputformat.class=oracle.hadoop.loader.examples.CSVInputFormat \  
-D mapred.output.dir=/user/biadmin/ldr \  
-D oracle.hadoop.loader.loaderMapFile=file:/home/biadmin/ldrMap.xml \  
-D oracle.hadoop.loader.connection.url=jdbc:oracle:thin:scott/scott@//imtebi:1521/xe \  
-D mapred.input.dir=/user/biadmin/NYSE_dividends \  
-D oracle.hadoop.loader.sampler.enableSampling=true \  
-D oracle.hadoop.loader.enableSorting=true \  
-D mapred.job.name=OraLoader \  
-fs hdfs://imtebi.imte.com:9000
```

- a hadoop job executed using 'standard' hadoop tools
- the loadermap document, ldrMap.xml
- libraries describing the Oracle loader input formats to be used, additional jars can/need to be specified

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle loader for Hadoop - example (II)

One needs to specify:

- the Hadoop environment available
[-fs ...]
- possibly, the class generating the hadoop 'output'
- input and output directories
input typically a file on the local filesystem
output a HDFS file
- a loader mapping file
- the 'nature' or 'type' of output to be generated by the 'loader'

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - loadermap document

[optional - depending on ...]

XML - document indicating:

- schema/table to be loaded
- columns to be loaded - mapping fields <> columns
- date value specifics

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - supported import formats

| Supported format mapreduce.outputformat.class= ... | Description |
|--|---|
| Hive table sources ...lib.input.HiveToAvroInputFormat | reads data from Hive table Hive table's column names |
| Delimited text files ...lib.output.DelimitedTextOutputFormat | reads delimited text files, comma-separated values files or tab-separated values files |
| Text files [regex interpretation] ..lib.input.RegexInputFormat | Comma separated list from the property oracle.hadoop.loader.input.fieldNames (or fld0, fld11, ... if property not defined) |
| Binary format Avro record files ...lib.input.AvroInputFormat | Field names from input files' Avro schemas |
| Simple, delimited text files ...lib.output.DelimitedTextOutputFormat | fld1, fld2, ... mapreduce.inputformat.class= (sample) |
| Oracle NoSQL Database ...kv.hadoop.KVAvroInputFormat | Field names from value portion of key value pairs for Oracle NoSQL database Avro records [available when Oracle NoSQL database is available] |

Depending on the format, additional properties can be specified, for example:

oracle.hadoop.loader.input.fieldTerminator

oracle.hadoop.loader.input.initialFieldEncloser

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - Meta data

Metadata required for loading the target table is required.

- **retrieved from the database whenever a JDBC connection can be made to that database**
- **the Oracle loader looks for metadata stored in an XML file (location oracle.hadoop.loader.tableMetadataFile property)**

```
java oracle.hadoop.loader.metadata.OraLoaderMetadata
-user <username>
-connection_url <connection URL>
[-schema <schemaName>] \
-table <tableName>
-output <output filename>
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - ONLINE

Data load format/approach:

- **JDBC based conventional load**

[output is directly loaded into target table; no additional processing required; jdbc connection; batching]

```
-D oracle.hadoop.loader.connection.url=jdbc:oracle:thin:scott/scott@//imtebi:1521/tst \  
-D oracle.hadoop.loader.sampler.enableSampling=true \  
-D oracle.hadoop.loader.enableSorting=true \  

```

- **OCI Direct path load**

[standard direct path technology; no additional processing; requires table to be partitioned; only on linux x64; datatype restrictions apply]

Steps performed

1. read target table metadata (from DB)
2. partitioning, sorting, and data conversion
3. connect to the DB from reducer nodes, load into partitions (JDBC or OCI)

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Loader for Hadoop - OFFLINE

Data load/format approach:

- **delimited text files**

- **generated by reduce jobs**

- **loaded by SQL Loader**

generates: a datafile, loader cntl file, a load script

- **using external table technology**

generates binary format to be loaded using external table technology (ORACLE_DATAPUMP) - creates external table script [data files NOT on a HDFS filesystem]

Steps performed

1. read target table metadata (from DB)

2. partitioning, sorting, and data conversion

3. write from reducer nodes to Oracle Data Pump files

4. copy files from HDFS to a location where database can access them

5. import into the database in parallel using external table mechanism

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

[aka. Oracle direct connector for HDFS]

- **Directly access data files on HDFS using Oracle external table technologies!**
 - Create an external table pointing to file location on HDFS
 - Query data from database using SQL
 - Load data into database when required
- **Fast data movement: parallel, optimized, automatic load balancing**
- **Traditional external table limitations apply:** read-only, no indexes, always scanned
- **Data files can be:**
 - Raw data in delimited text file format stored in HDFS
 - Data Pump files created by Oracle Loader for Hadoop in HDFS
 - Hive tables

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - Installation, Setup (I)

- **standard installation**

standard unzip, no specific installation routines required

- **once installed...**

- **find *hdfs_stream*** - the pre-processor, invoked by the external loader utility

```
#!/bin/bash
```

```
# hdfs_stream
```

```
# Copyright (c) 2011, 2012, Oracle and/or its affiliates. All rights reserved.
```

```
#
```

```
# set OSCH_HOME to the absolute path of the Oracle SQL Connector installation directory
```

```
export OSCH_HOME=/home/biadmin/orahdfs-2.0.1
```

```
#
```

```
# Add the hadoop bin directory to the PATH variable in this script
```

```
export PATH=/opt/ibm/biginsights/IHC/bin:/usr/bin:/bin:$PATH
```

```
# Set the OSCH_LOG_DIR to the absolute path of a log directory for error logging of the pre-processor.
```

```
OSCH_LOG_DIR=${OSCH_HOME}/log
```

```
...
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - Installation, Setup (II)

- **create an Oracle directory object, mapping to the pre-processor file:**

```
CREATE OR REPLACE DIRECTORY osch_bin_path AS '/home/biadmin/orahdfs-2.0.1'
```

- **any Oracle user using OSCH will require at least the following privs:**

- CREATE SESSION
- CREATE TABLE
- EXECUTE on UTL_FILE PL/SQL
[file generation]
- READ and EXECUTE on the OSCH_BIN_PATH
[use of pre-processor]

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - ExternalTable tool (I)

Purpose - facilitate creation/management of the external table environment accessing Hadoop resources!

```
$HADOOP_HOME/bin/hadoop jar $OSCH_HOME/jlib/orahdfs.jar \  
    oracle.hadoop.exttab.ExternalTable \  
    [-conf config_file] | [-D property=value]... \  
    -createTable | -publish | -listlocations | -getDDL
```

| option | description |
|-----------------------|---|
| -createTable | creates an external table definition and publishes the data URIs to the location files of the external table |
| -publish | publishes the data URIs to the location files of an existing external table (<i>alter table</i>) |
| -listlocations | debugging/tracing purposes |
| -getDDL | debugging/tracing purposes |

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - ExternalTable tool (II)

The content of the config file/properties set depend on the 'structure' of the Hadoop data source to be accessed ...

| | |
|--------------------------|---|
| Common properties | oracle.hadoop.exctab.tableName oracle.hadoop.exctab.sourceType=datapump text hive oracle.hadoop.exctab.dataPaths (not for Hive) oracle.hadoop.exctab.defaultDirectory oracle.hadoop.connection.url oracle.hadoop.connection.user oracle.hadoop.exctab.logDirectory (*) |
| Datapump | -- |
| External files | oracle.hadoop.exctab.columnCount or oracle.hadoop.exctab.columnNames oracle.hadoop.exctab.recordDelimiter (*) oracle.hadoop.exctab.fieldTerminator (*) oracle.hadoop.exctab.initialFieldEncloser (*) oracle.hadoop.exctab.trailingFieldEncloser (*) oracle.hadoop.exctab.locationFileCount (*) |
| Hive | oracle.hadoop.exctab.hive.tableName oracle.hadoop.exctab.hive.databaseName oracle.hadoop.locationFileCount (*) |
| (*) optional | |

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - ExternalTable tool (III) - Example

```
$HADOOP_HOME/bin/hadoop jar $OSCH_HOME/jlib/orahdfs.jar \  
oracle.hadoop.exctab.ExternalTable \  
-D oracle.hadoop.exctab.tableName=SALES_DT_XTAB \  
-D oracle.hadoop.exctab.locationFileCount=2 \  
-D oracle.hadoop.exctab.dataPaths="hdfs:/user/biadmin/" \  
-D oracle.hadoop.exctab.columnNames=A1,A2,A3,A4 \  
-D oracle.hadoop.exctab.defaultDirectory=SALES_DT_DIR \  
-D oracle.hadoop.connection.url=jdbc:oracle:thin:@//imtebi:1521/xe \  
-D oracle.hadoop.connection.user=SCOTT \  
-D oracle.hadoop.exctab.logDirectory=SALES_DT_DIR \  
-createTable
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - ExternalTable tool (III) - Example (cont.)

```
CREATE TABLE "SCOTT"."SALES_DT_XTAB"  
("A1" VARCHAR2(4000), "A2" VARCHAR2(4000), "A3" VARCHAR2(4000), "A4" VARCHAR2(4000))  
ORGANIZATION EXTERNAL  
(  
  TYPE ORACLE_LOADER DEFAULT DIRECTORY "SALES_DT_DIR"  
  ACCESS PARAMETERS  
  ( RECORDS DELIMITED BY 0X'0A'  
    CHARACTERSET AL32UTF8  
    STRING SIZES ARE IN CHARACTERS  
    PREPROCESSOR "OSCH_BIN_PATH":'hdfs_stream'  
    FIELDS TERMINATED BY 0X'2C'  
    MISSING FIELD VALUES ARE NULL  
    ("A1" CHAR, "A2" CHAR, "A3" CHAR, "A4" CHAR) )  
  LOCATION  
  ( 'osch-20130325104814-1029-1')  
) PARALLEL REJECT LIMIT UNLIMITED;
```

The following location files were created.

osch-20130325104814-1029-1 contains 1 URI, 126 bytes

126 hdfs://imtebi.imte.com:9000/user/biadmin/NYSE_dividends

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - Location files

A location file is:

- **a file specified in** the location clause of the external table
- created/managed/published **by Oracle SQL Connector for HDFS**
- **containing the URIs of the HDFS data files to be accessed**

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
```

```
<locationFile>
```

```
  <header>
```

```
    <version>1.0</version>
```

```
    <fileName>osch-20130325060613-5448-1</fileName>
```

```
    <createDate>2013-03-25T06:06:13</createDate>
```

```
    <publishDate>2013-03-25T06:06:13</publishDate>
```

```
    <productName>Oracle SQL Connector for HDFS Release 2.0.1 - Production</productName>
```

```
    <productVersion>2.0.1</productVersion>
```

```
  </header>
```

```
  <uri_list>
```

```
    <uri_list_item compressionCodec="" size="126">
```

```
      hdfs://imtebi.imte.com:9000/user/biadmin/divid</uri_list_item>
```

```
    </uri_list>
```

```
</locationFile>
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - do not forget...

**ALTER SESSION ENABLE PARALLEL QUERY;
ALTER SESSION ENABLE PARALLEL DDL;
ALTER SESSION ENABLE PARALLEL DML;**

Hints available when using the external table data: APPEND, PQ_DISTRIBUTE

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

OSCH - a scenario

Step 1 - create the data directories the user will access, and grant privs on it

```
mkdir /home/biadmin/scratch
```

```
$ sqlplus / as sysdba
```

```
SQL> CREATE OR REPLACE DIRECTORY sales_dp_dir AS '/home/biadmin/scratch'
```

```
SQL> GRANT READ, WRITE ON DIRECTORY sales_dp_dir TO scott;
```

Step 2 - run the ExternalTable tool, with the correct parameters

Step 3 - access the external table

```
SQL > SELECT * FROM ....
```

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

- **Oracle Loader for Hadoop**
 - **for simple non-partitioned tables, use online load with JDBC**
 - **for partitioned tables, use online direct path load**
 - **offline load (datapump): lesser load on external tables; using datapump technology shifts Oracle workload => Hadoop**
- **OSCH**
 - **possibility to leave data on HDFS**
 - **parallel access**
 - **combine with Oracle Loader for Hadoop to access generated files directly (datapump files)**

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Provides native Hadoop integration within ODI, by means of specific ODI Knowledge Modules included within ODI Application Adapter for Hadoop.

- build Hadoop metadata within ODI
- load data into Hadoop
- transform data within Hadoop
- load data into Oracle Database
[Oracle Loader for Hadoop // Oracle SQL Connector for HDFS]

... while using a standard, uniform ODI programming interface/GUI

.... while using standard ODI features

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Data Integrator Application Adapter for Hadoop - KMs

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

| KM | Description |
|--------------------------------|---|
| IKM File to load | load filesystem data (incl. HDFS) into Hive <i>from: FS</i> <i>to: Hive</i> |
| IKM Hive control append | appends (truncate/insert) data into Hive table (validates) <i>from: Hive</i> <i>to: Hive</i> |
| IKM Hive transform | adds data into a Hive table after (perl/Python) transformation processing <i>from: Hive</i> <i>to: Hive</i> |
| IKM File-Hive to Oracle | HDFS file or Hive source load into Oracle DB using Oracle loader for Hadoop <i>from: FS or Hive</i> <i>to: Oracle DB</i> |
| CKM Hive | Validates data against constraints <i>target: Hive</i> |
| RKM Hive | Reverse engineers Hive tables |

Oracle Data Integrator Application Adapter for Hadoop - properties

- **improve data integration performance**
 - most tasks are processed within Hadoop - use Hadoop cluster resources
 - hive knowledge module available
 - OLH knowledge module available
- **Improve efficiency of development and data integration**
 - uniform ODI programming interface
 - HiveQL language which is similar with SQL - No need to write Map/Reduce programs
 - scheduling Hadoop jobs within ODI

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

- **simple data model**
 - key/value pairs
 - CRUD
- **scalability**
 - dynamic data partitioning and distribution
- **high-availability**
 - through ao. replicas
- **transparent load balancing**
 - reads from master or replicas
- **integrates with Hadoop - use as a source and/or a target**

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Thank you!



ABIS Training & Consulting
Kris Van Thillo
kvanthillo@abis.be

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database

Oracle Big Data Connectors

1. Introduction - setting the scene ...
2. Oracle Loader for Hadoop
3. Oracle SQL Connector for HDFS - OSCH
4. Oracle Loader vs OSCH
5. Oracle Data Integrator Application Adapter for Hadoop
6. Oracle NoSQL Database