Assignment 18.1

Associated Data Files

Link: https://drive.google.com/file/d/0Bxr27gVaXO5sU3g5ZUFhVDYxczQ/view?usp=sharing

Problem Statement

1 Create a HBase table 'customer' with one column family 'details'.

2 Write a shells script that loads the content of the file customers.dat in the HBase table.

3 The structure of file looks like:

id, name, location, age

Step 1: Create a HBase table 'customers' with column_family 'customers_data' from HBase shell.

create 'customers', 'details'

describe 'customer'

list

```
hbase(main):006:0> create 'customer','details'
0 row(s) in 0.3960 seconds

=> Hbase::Table - customer
hbase(main):007:0> describe 'customer'
Table customer is ENABLED
customer
COLUMN FAMILIES DESCRIPTION
{NAME => 'details', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL =>
'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.0320 seconds

hbase(main):008:0> list
TABLE
clicks
customer
2 row(s) in 0.0070 seconds

=> ["clicks", "customer"]
hbase(main):009:0>
```

Step 2: Write the following PIG script to load data into the 'customers' table in HBase

--create pig script Load_HBase_Customers.pig

--Load dataset 'customers' from local location

```
raw_data = LOAD '/home/acadgild/dataset/customers.dat' USING PigStorage(',') AS (
                id:chararray,
                name:chararray,
                location:chararray,
                age:int,
);
```

-- Use HBase storage handler to map data from PIG to HBase

--NOTE: In this case, id (first unique column) will be considered as row key.

STORE raw_data INTO 'hbase://customers' USING org.apache.pig.backend.hadoop.hbase.HBaseStorage(

'details:name

details:location

details:age'

);

```
[acadgild@localhost dataset]$ cat Load_HBase_Customers.pig
raw_data = LOAD '/home/acadgild/dataset/customers.dat' USING PigStorage(',') AS (
           id:chararray,
           name:chararray,
           location:chararray,
           age:int
);

STORE raw_data INTO 'hbase://customer' USING org.apache.pig.backend.hadoop.hbase.HBaseStorage(
'details:name
 details:location
 details:age'
);

[acadgild@localhost dataset]$
```

Step 3: set PIG_CLASSPATH with the jars required to load files in the hbase.

PIG_CLASSPATH=/usr/lib/hbase/hbase.jar:/usr/lib/zookeeper/zookeeper-3.4.5-cdh4.4.0.jar

```
[acadgild@localhost dataset]$ PIG_CLASSPATH=/usr/lib/hbase/hbase.jar:/usr/lib/zookeeper/zookeeper-3.4.5-cdh4.4.0.jar
[acadgild@localhost dataset]$ echo $PIG_CLASSPATH
/usr/lib/hbase/hbase.jar:/usr/lib/zookeeper/zookeeper-3.4.5-cdh4.4.0.jar
```

Step 4: scan 'customer' table in hbase shell. It shows zero rows present in the table.

scan 'customer'

```
hbase(main):012:0> scan 'customer'
ROW                                           COLUMN+CELL
0 row(s) in 0.3240 seconds

hbase(main):013:0>
```

Step 5: Run PIG script

pig -x local Load_HBase_Customers.pig

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.2.0   0.14.0  acadgild        2017-08-17 22:34:16     2017-08-17 22:34:18     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local259188163_0001 1       0       n/a     n/a     n/a     n/a     0       0       0       0       raw_data        MAP_ONLY        hbase://customer,

Input(s):
Successfully read 4 records from: "/home/acadgild/dataset/customers.dat"

Output(s):
Successfully stored 4 records in: "hbase://customer"

Counters:
Total records written : 4
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local259188163_0001


2017-08-17 22:34:18,795 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-08-17 22:34:18,795 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-08-17 22:34:18,801 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-08-17 22:34:18,815 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-08-17 22:34:18,845 [main] INFO  org.apache.pig.Main - Pig script completed in 4 seconds and 542 milliseconds (4542 ms)
```

Step 6: Run below command in hbase shell.

scan 'customer'

```
hbase(main):013:0> scan 'customer'
ROW                                             COLUMN+CELL
 1                                              column=details:age, timestamp=1502989458189, value=18
 1                                              column=details:location, timestamp=1502989458189, value=IND
 1                                              column=details:name, timestamp=1502989458189, value=Amit
 2                                              column=details:age, timestamp=1502989458200, value=20
 2                                              column=details:location, timestamp=1502989458200, value=PAK
 2                                              column=details:name, timestamp=1502989458200, value=Sumit
 3                                              column=details:age, timestamp=1502989458201, value=26
 3                                              column=details:location, timestamp=1502989458201, value=AUS
 3                                              column=details:name, timestamp=1502989458201, value=Rohit
 4                                              column=details:age, timestamp=1502989458201, value=24
 4                                              column=details:location, timestamp=1502989458201, value=UK
 4                                              column=details:name, timestamp=1502989458201, value=Namit
4 row(s) in 0.0510 seconds

hbase(main):014:0>
```