Class: Hadoop: Distributed Processing of Big Data
Instructor: Venkat Mavram
Student: Sheetal Gangakhedkar
Date: 03/18/2017
Homework: Assignment 3

Importing Database Tables into HDFS with Sqoop

```
$ unzip Assignment3.zip

$ sqoop import-all-tables -m 1 --connect
jdbc:mysql://quickstart.cloudera:3306/retail_db --username=retail_dba
--password cloudera --compression-codec=snappy --as-avrodatafile
--warehouse-dir=/user/hive/warehouse
log file: sqoop-import-avro-data.log

$ hadoop fs -ls /user/hive/warehouse
Found 6 items
drwxr-xr-x   - cloudera supergroup          0 2017-03-16 20:41
/user/hive/warehouse/categories
drwxr-xr-x   - cloudera supergroup          0 2017-03-16 20:41
/user/hive/warehouse/customers
drwxr-xr-x   - cloudera supergroup          0 2017-03-16 20:42
/user/hive/warehouse/departments
drwxr-xr-x   - cloudera supergroup          0 2017-03-16 20:43
/user/hive/warehouse/order_items
drwxr-xr-x   - cloudera supergroup          0 2017-03-16 20:44
/user/hive/warehouse/orders
drwxr-xr-x   - cloudera supergroup          0 2017-03-16 20:44
/user/hive/warehouse/products

$ hadoop fs -ls /user/hive/warehouse/categories
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2017-03-16 20:41
/user/hive/warehouse/categories/_SUCCESS
-rw-r--r--   1 cloudera supergroup       1378 2017-03-16 20:41
/user/hive/warehouse/categories/part-m-00000.avro

$ ls -l *.avsc
-rwxrwxrwx 1 cloudera cloudera  594 Mar 16 20:39 categories.avsc
-rwxrwxrwx 1 cloudera cloudera 1509 Mar 16 20:39 customers.avsc
-rwxrwxrwx 1 cloudera cloudera  440 Mar 16 20:39 departments.avsc
-rwxrwxrwx 1 cloudera cloudera 1099 Mar 16 20:39 order_items.avsc
-rwxrwxrwx 1 cloudera cloudera  707 Mar 16 20:39 orders.avsc
-rwxrwxrwx 1 cloudera cloudera 1041 Mar 16 20:39 products.avsc
```

```
$ sudo -u hdfs hadoop fs -mkdir /user/examples
$ sudo -u hdfs hadoop fs -chmod +rw /user/examples
$ hadoop fs -copyFromLocal ~/*.avsc /user/examples

$ hive
2017-03-16 20:48:33,751 WARN  [main] mapreduce.TableMapReduceUtil: The
hbase-prefix-tree module jar containing PrefixTreeCodec is not present.
Continuing without it.

Logging initialized using configuration in
file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.

hive> CREATE EXTERNAL TABLE categories ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.avro.AvroSerDe' STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat' LOCATION
'hdfs:///user/hive/warehouse/categories' TBLPROPERTIES
('avro.schema.url'='hdfs:///user/examples/categories.avsc');
OK
Time taken: 15.709 seconds
hive> CREATE EXTERNAL TABLE customers ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.avro.AvroSerDe' STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat' LOCATION
'hdfs:///user/hive/warehouse/customers' TBLPROPERTIES
('avro.schema.url'='hdfs:///user/examples/customers.avsc');
OK
Time taken: 0.451 seconds
hive> CREATE EXTERNAL TABLE departments ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.avro.AvroSerDe' STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat' LOCATION
'hdfs:///user/hive/warehouse/departments' TBLPROPERTIES
('avro.schema.url'='hdfs:///user/examples/departments.avsc');
OK
Time taken: 0.144 seconds
hive> CREATE EXTERNAL TABLE orders ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.avro.AvroSerDe' STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat' LOCATION
'hdfs:///user/hive/warehouse/orders' TBLPROPERTIES
('avro.schema.url'='hdfs:///user/examples/orders.avsc');
OK
Time taken: 0.148 seconds
hive> CREATE EXTERNAL TABLE order_items ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.avro.AvroSerDe' STORED AS INPUTFORMAT
```

```
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat' LOCATION
'hdfs:///user/hive/warehouse/order_items' TBLPROPERTIES
('avro.schema.url'='hdfs:///user/examples/order_items.avsc');
OK
Time taken: 0.278 seconds
hive> CREATE EXTERNAL TABLE products ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.avro.AvroSerDe' STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat' OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat' LOCATION
'hdfs:///user/hive/warehouse/products' TBLPROPERTIES
('avro.schema.url'='hdfs:///user/examples/products.avsc');
OK
Time taken: 0.181 seconds

hive> show tables;
OK
categories
customers
departments
order_items
orders
products
Time taken: 0.04 seconds, Fetched: 6 row(s)
```

Part I: Develop and Run Simple Queries

```
hive> set hive.cli.print.header=true;
```

Question: Which customers did your query identify as the winner of the $5000 prize?

```
hive> select customer_fname, customer_lname from customers where
(customer_fname='Brian' or customer_fname='Bryan') and customer_city='Chicago';
OK
Bryan Smith
Brian Wilson
Time taken: 2.434 seconds, Fetched: 2 row(s)

[cloudera@quickstart ~]$ hive -e 'SELECT product_id, product_price,
product_name
> FROM products ORDER BY product_price LIMIT 10'

2017-03-16 22:45:55,913 WARN  [main] mapreduce.TableMapReduceUtil: The
hbase-prefix-tree module jar containing PrefixTreeCodec is not present.
Continuing without it.

Logging initialized using configuration in
```

```
file:/etc/hive/conf.dist/hive-log4j.properties
Query ID = cloudera_20170316224646_0e9b7485-404a-48a1-9bc0-02514ccab483
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0035, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0035/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0035
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-16 22:46:30,809 Stage-1 map = 0%,  reduce = 0%
2017-03-16 22:46:48,267 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.62
sec
2017-03-16 22:47:03,704 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.32
sec
MapReduce Total cumulative CPU time: 6 seconds 320 msec
Ended Job = job_1487633215935_0035
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.32 sec   HDFS Read: 72605
HDFS Write: 485 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 320 msec
OK
1284    0.0     Nike Men's Hypervenom Phantom Premium FG Socc
517     0.0     Nike Men's Hypervenom Phantom Premium FG Socc
414     0.0     Nike Men's Hypervenom Phantom Premium FG Socc
934     0.0     Callaway X Hot Driver
547     0.0     Nike Men's Hypervenom Phantom Premium FG Socc
388     0.0     Nike Men's Hypervenom Phantom Premium FG Socc
38      0.0     Nike Men's Hypervenom Phantom Premium FG Socc
624     4.99    adidas Batting Helmet Hardware Kit
815     4.99    Zero Friction Practice Golf Balls - 12 Pack
336     5.0     Nike Swoosh Headband - 2"
Time taken: 58.194 seconds, Fetched: 10 row(s)
WARN: The method class
org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
```

Question: Which two product names have a price of zero?

```
[cloudera@quickstart ~]$ hive -e 'SELECT DISTINCT product_name FROM products
WHERE product_price=0.0'
2017-03-16 22:51:51,214 WARN  [main] mapreduce.TableMapReduceUtil: The
hbase-prefix-tree module jar containing PrefixTreeCodec is not present.
Continuing without it.
```

```
Logging initialized using configuration in
file:/etc/hive/conf.dist/hive-log4j.properties
Query ID = cloudera_20170316225252_c8c1c4ea-0272-4749-895e-4b4ba19ab960
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0036, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0036/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0036
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-16 22:52:25,094 Stage-1 map = 0%,  reduce = 0%
2017-03-16 22:52:41,998 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.72
sec
2017-03-16 22:52:57,162 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.46
sec
MapReduce Total cumulative CPU time: 6 seconds 460 msec
Ended Job = job_1487633215935_0036
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.46 sec   HDFS Read: 73579
HDFS Write: 68 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 460 msec
OK
Callaway X Hot Driver
Nike Men's Hypervenom Phantom Premium FG Socc
Time taken: 54.81 seconds, Fetched: 2 row(s)
WARN: The method class
org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
```

Question: How many customers ids are in the customers table?

```
hive> SELECT count(*) customer_id FROM customers;
Query ID = cloudera_20170316225757_6f8d0874-9cf8-4ad8-b579-db092da8d013
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
```

```
Starting Job = job_1487633215935_0039, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0039/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0039
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-16 23:02:49,608 Stage-1 map = 0%,  reduce = 0%
2017-03-16 23:03:05,965 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.93
sec
2017-03-16 23:03:20,868 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.48
sec
MapReduce Total cumulative CPU time: 7 seconds 480 msec
Ended Job = job_1487633215935_0039
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.48 sec   HDFS Read: 491393
HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 480 msec
OK
12435
Time taken: 46.97 seconds, Fetched: 1 row(s)
```

Question: How many customers ids are in the customers table? (with DISTINCT)

```
hive> SELECT count(DISTINCT customer_id) FROM customers;
Query ID = cloudera_20170316225757_6f8d0874-9cf8-4ad8-b579-db092da8d013
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0040, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0040/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0040
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-16 23:10:39,129 Stage-1 map = 0%,  reduce = 0%
2017-03-16 23:10:57,909 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.93
sec
2017-03-16 23:11:16,094 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU
11.48 sec
MapReduce Total cumulative CPU time: 11 seconds 480 msec
Ended Job = job_1487633215935_0040
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.48 sec   HDFS Read:
491737 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 480 msec
OK
12435
```

```
Time taken: 53.757 seconds, Fetched: 1 row(s)
```

Question: How may households are in the customers table? (Hint, try concatenating the customers' address and zipcode and count the number of distinct households)

```
hive> SELECT COUNT(DISTINCT CONCAT_WS(',', customer_street, customer_zipcode))
FROM customers;
OK
11508
Time taken: 55.19 seconds, Fetched: 1 row(s)
```

Question: How may households are in the customers table? (Hint, try concatenating the customers' address and zipcode and count the number of distinct households) - (with address=street, city, state, zipcode)

```
hive> SELECT COUNT(DISTINCT CONCAT_WS(',', customer_street, customer_city,
customer_state, customer_zipcode)) FROM customers;
Query ID = cloudera_20170316225757_6f8d0874-9cf8-4ad8-b579-db092da8d013
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0044, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0044/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0044
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-16 23:45:20,873 Stage-1 map = 0%,  reduce = 0%
2017-03-16 23:45:39,675 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.09
sec
2017-03-16 23:45:57,844 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU
11.73 sec
MapReduce Total cumulative CPU time: 11 seconds 730 msec
Ended Job = job_1487633215935_0044
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.73 sec   HDFS Read:
493406 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 730 msec
OK
11508
Time taken: 52.566 seconds, Fetched: 1 row(s)
```

Question: **Using customer_id, which state has the most customers?**

```
hive> SELECT customer_state, count(DISTINCT customer_id) as cnt FROM customers
GROUP BY customer_state ORDER BY cnt DESC LIMIT 1;
Query ID = cloudera_20170316235151_9f9caaa1-a636-4622-92f0-70e3e3462d97
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0050, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0050/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0050
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-17 00:02:57,210 Stage-1 map = 0%,  reduce = 0%
2017-03-17 00:03:17,087 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 7.31
sec
2017-03-17 00:03:18,230 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.71
sec
2017-03-17 00:03:37,038 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU
12.35 sec
MapReduce Total cumulative CPU time: 12 seconds 350 msec
Ended Job = job_1487633215935_0050
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0051, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0051/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0051
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-17 00:03:55,994 Stage-2 map = 0%,  reduce = 0%
2017-03-17 00:04:08,735 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.82
sec
2017-03-17 00:04:23,844 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.59
sec
MapReduce Total cumulative CPU time: 4 seconds 590 msec
Ended Job = job_1487633215935_0051
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.35 sec   HDFS Read:
```

```
491582 HDFS Write: 1043 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.59 sec   HDFS Read: 5799
HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 940 msec
OK
PR    4771
Time taken: 103.921 seconds, Fetched: 1 row(s)
```

Question: Which top three product_ids had the most orders? Show your query.

```
hive> SELECT order_item_product_id, SUM(order_item_quantity) as total FROM
order_items GROUP BY order_item_product_id ORDER BY total DESC LIMIT 3;
Query ID = cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0066, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0066/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0066
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-18 09:49:10,527 Stage-1 map = 0%,  reduce = 0%
2017-03-18 09:49:29,361 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.99
sec
2017-03-18 09:49:44,122 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.41
sec
MapReduce Total cumulative CPU time: 8 seconds 410 msec
Ended Job = job_1487633215935_0066
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0067, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0067/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0067
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-18 09:50:00,488 Stage-2 map = 0%,  reduce = 0%
2017-03-18 09:50:12,988 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.73
sec
2017-03-18 09:50:27,988 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.3
```

```
sec
MapReduce Total cumulative CPU time: 4 seconds 300 msec
Ended Job = job_1487633215935_0067
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.41 sec   HDFS Read:
1546176 HDFS Write: 2304 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.3 sec   HDFS Read: 7082
HDFS Write: 31 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 710 msec
OK
365   73698
502   62956
1014  57803
Time taken: 94.133 seconds, Fetched: 3 row(s)
```

Question: Which top three product_ids had the most orders? Show your query.
Extra Credit: What were the product names? Show your query.

```
hive> SELECT p.product_id, p.product_name FROM (SELECT order_item_product_id,
SUM(order_item_quantity) as total FROM order_items GROUP BY
order_item_product_id ORDER BY total DESC LIMIT 3) o JOIN products p ON
o.order_item_product_id = p.product_id;
Query ID = cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0063, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0063/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0063
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-18 09:45:52,410 Stage-1 map = 0%,  reduce = 0%
2017-03-18 09:46:09,424 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.32
sec
2017-03-18 09:46:23,202 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.59
sec
MapReduce Total cumulative CPU time: 7 seconds 590 msec
Ended Job = job_1487633215935_0063
Launching Job 2 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
```

```
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0064, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0064/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0064
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-18 09:46:41,153 Stage-2 map = 0%,  reduce = 0%
2017-03-18 09:46:53,742 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.73
sec
2017-03-18 09:47:09,541 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.81
sec
MapReduce Total cumulative CPU time: 4 seconds 810 msec
Ended Job = job_1487633215935_0064
Execution log at:
/tmp/cloudera/cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e.log
2017-03-18 09:47:22     Starting to launch local task to process map join;
maximum memory = 1013645312
2017-03-18 09:47:25     Dump the side-table for tag: 1 with group count: 1345
into file:
file:/tmp/cloudera/2ee2596b-51c7-450a-b064-4d76c88409be/hive_2017-03-18_09-45-3
7_929_7434517338800007868-1/-local-10005/HashTable-Stage-5/MapJoin-mapfile11--.
hashtable
2017-03-18 09:47:26     Uploaded 1 File to:
file:/tmp/cloudera/2ee2596b-51c7-450a-b064-4d76c88409be/hive_2017-03-18_09-45-3
7_929_7434517338800007868-1/-local-10005/HashTable-Stage-5/MapJoin-mapfile11--.
hashtable (81198 bytes)
2017-03-18 09:47:26     End of local task; Time Taken: 4.113 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1487633215935_0065, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0065/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0065
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 0
2017-03-18 09:47:43,202 Stage-5 map = 0%,  reduce = 0%
2017-03-18 09:47:56,989 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 2.47
sec
MapReduce Total cumulative CPU time: 2 seconds 470 msec
Ended Job = job_1487633215935_0065
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.59 sec   HDFS Read:
1546189 HDFS Write: 2304 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.81 sec   HDFS Read: 6799
HDFS Write: 156 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 2.47 sec   HDFS Read: 5582 HDFS Write:
116 SUCCESS
```

```
Total MapReduce CPU Time Spent: 14 seconds 870 msec
OK
365   Perfect Fitness Perfect Rip Deck
502   Nike Men's Dri-FIT Victory Golf Polo
1014  O'Brien Men's Neoprene Life Vest
Time taken: 140.242 seconds, Fetched: 3 row(s)
```

Question: Using the orders_corrected table, count the number of orders (using order_id) that had a status of COMPLETE, on May 17, 2014. Show your query. (Notice, you can specify MONTH, YEAR and DAY as built-in functions to retrieve the month, year or day from a date string).

```
hive> create table orders_corrected as select *,
from_unixtime(cast(substring(order_date,0,10) as INT)) as order_dateStr
from orders;
hive> SELECT COUNT(order_id) FROM orders_corrected WHERE
order_status='COMPLETE' AND YEAR(order_datestr)=2014 AND
MONTH(order_datestr)=05 AND DAY(order_datestr)=17;
Query ID = cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0070, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0070/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0070
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-18 10:35:04,798 Stage-1 map = 0%,  reduce = 0%
2017-03-18 10:35:20,839 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.18
sec
2017-03-18 10:35:36,534 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.92
sec
MapReduce Total cumulative CPU time: 7 seconds 920 msec
Ended Job = job_1487633215935_0070
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.92 sec   HDFS Read:
3835308 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 920 msec
OK
61
Time taken: 48.337 seconds, Fetched: 1 row(s)
```

Question: What was Dualcore's total revenue from completed orders on May 17,

2014? (Hint: use a left semi join). Show your query.

```
hive> SELECT SUM(i.order_item_subtotal) FROM order_items i LEFT SEMI JOIN
(SELECT * FROM orders_corrected WHERE order_status='COMPLETE' AND
YEAR(order_datestr)=2014 AND MONTH(order_datestr)=05 AND DAY(order_datestr)=17)
c ON c.order_id = i.order_item_order_id;
Query ID = cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e
Total jobs = 1
Execution log at:
/tmp/cloudera/cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e.log
2017-03-18 10:55:42     Starting to launch local task to process map join;
maximum memory = 1013645312
2017-03-18 10:55:48     Dump the side-table for tag: 1 with group count: 61
into file:
file:/tmp/cloudera/2ee2596b-51c7-450a-b064-4d76c88409be/hive_2017-03-18_10-55-3
1_725_7082137311885508918-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile21--.
hashtable
2017-03-18 10:55:48     Uploaded 1 File to:
file:/tmp/cloudera/2ee2596b-51c7-450a-b064-4d76c88409be/hive_2017-03-18_10-55-3
1_725_7082137311885508918-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile21--.
hashtable (1502 bytes)
2017-03-18 10:55:48     End of local task; Time Taken: 5.397 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0071, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0071/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0071
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-18 10:56:04,652 Stage-2 map = 0%,  reduce = 0%
2017-03-18 10:56:22,751 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.93
sec
2017-03-18 10:56:37,640 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.67
sec
MapReduce Total cumulative CPU time: 8 seconds 670 msec
Ended Job = job_1487633215935_0071
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 8.67 sec   HDFS Read:
1549928 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 670 msec
OK
29198.830507278442
```

```
Time taken: 68.139 seconds, Fetched: 1 row(s)
```

Question: The result of the above query is in scientific notation. Rewrite the last query to format the value in dollars and cents (e.g., $2000000.00). To do this, format the result using the PRINTF function and the format string "$%.2f". Show your query.

```
hive> SELECT PRINTF("$%.2f", SUM(i.order_item_subtotal)) FROM order_items i
LEFT SEMI JOIN (SELECT * FROM orders_corrected WHERE order_status='COMPLETE'
AND YEAR(order_datestr)=2014 AND MONTH(order_datestr)=05 AND
DAY(order_datestr)=17) c ON c.order_id = i.order_item_order_id;
Query ID = cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e
Total jobs = 1
Execution log at:
/tmp/cloudera/cloudera_20170318082828_ca2ff272-907d-4025-b138-151904dc667e.log
2017-03-18 11:00:59     Starting to launch local task to process map join;
maximum memory = 1013645312
2017-03-18 11:01:04     Dump the side-table for tag: 1 with group count: 61
into file:
file:/tmp/cloudera/2ee2596b-51c7-450a-b064-4d76c88409be/hive_2017-03-18_11-00-4
8_961_4681465441794058518-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile41--.
hashtable
2017-03-18 11:01:04     Uploaded 1 File to:
file:/tmp/cloudera/2ee2596b-51c7-450a-b064-4d76c88409be/hive_2017-03-18_11-00-4
8_961_4681465441794058518-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile41--.
hashtable (1502 bytes)
2017-03-18 11:01:04     End of local task; Time Taken: 5.436 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0073, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0073/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0073
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-18 11:01:21,014 Stage-2 map = 0%,  reduce = 0%
2017-03-18 11:01:39,025 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.87
sec
2017-03-18 11:01:55,137 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 9.37
sec
MapReduce Total cumulative CPU time: 9 seconds 370 msec
Ended Job = job_1487633215935_0073
```

```
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 9.37 sec   HDFS Read:
1550272 HDFS Write: 10 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 370 msec
OK
$29198.83
Time taken: 67.37 seconds, Fetched: 1 row(s)
```

This is the end of Part I:

Develop and Run Simple Queries

Part II: Data Management with Hive

Question: Create a table named ratings for storing tab-delimited records using this structure (posted: TIMESTAMP, cust_id: INT, prod_id: INT, rating: TINYINT, message: STRING)

```
hive> CREATE TABLE IF NOT EXISTS ratings (posted TIMESTAMP, cust_id INT,
prod_id INT, rating TINYINT, message STRING) COMMENT 'Product Ratings Table'
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n' STORED
AS TEXTFILE;
OK
Time taken: 0.545 seconds

hive> DESCRIBE ratings;
OK
posted                  timestamp
cust_id                 int
prod_id                 int
rating                  tinyint
message                 string
Time taken: 0.197 seconds, Fetched: 5 row(s)

[cloudera@quickstart ~]$ hadoop fs -put ~/datasets/ratings_2012.txt
/user/hive/warehouse/ratings

hive> set hive.cli.print.header=true;
hive> SELECT * FROM ratings LIMIT 20;
OK
ratings.posted    ratings.cust_id   ratings.prod_id   ratings.rating
ratings.message
2012-05-21 12:52:48     1043182     1274362     5     This is truly fantastic!
2012-10-14 01:36:07     1242853     1273879     2     The product quality was
OK
```

```
2012-10-14 02:41:50     1047430     1273799     2     Shoddy quality
2012-10-14 10:10:05     1087455     1274476     4     Quality was passable
2012-10-14 10:42:41     1170230     1273964     2     It was OK
2012-10-14 19:12:33     1063130     1274734     4     It was OK
2012-10-14 22:00:56     1031378     1274616     4     Quality was passable
2012-10-15 00:27:47     1203215     1273850     5     Awesome product
2012-10-15 01:14:26     1135616     1274218     4     Value of product was
just alright
2012-10-15 01:18:58     1145446     1274304     3     Average quality
2012-10-15 04:49:00     1211187     1273654     3     It was just alright
2012-10-15 05:01:38     1026707     1273964     2     OK but not great
2012-10-15 05:25:30     1166507     1273732     1     I would never buy this
again
2012-10-15 06:20:16     1228815     1274149     2     Cheap quality
2012-10-15 13:34:01     1229606     1274522     4     Alright but not great
2012-10-15 14:37:04     1182384     1274628     4     Average quality
2012-10-15 17:14:28     1086291     1274157     3     Quality was passable
2012-10-15 17:54:47     1166286     1274151     4     The item was decent
2012-10-15 23:42:48     1025997     1274210     3     Alright but nothing
special
2012-10-16 01:43:55     1057881     1274179     2     Poor quality
Time taken: 2.246 seconds, Fetched: 20 row(s)

[cloudera@quickstart ~]$ hadoop fs -put ~/datasets/ratings_2013.txt
ratings_2013.txt
[cloudera@quickstart ~]$ hadoop fs -ls ratings_2013.txt
-rw-r--r--   1 cloudera cloudera    1240550 2017-03-18 12:57 ratings_2013.txt

hive> LOAD DATA INPATH '/user/cloudera/ratings_2013.txt' INTO TABLE ratings;
Loading data to table default.ratings
chgrp: changing ownership of
'hdfs://quickstart.cloudera:8020/user/hive/warehouse/ratings/ratings_2013.txt':
User does not belong to supergroup
Table default.ratings stats: [numFiles=2, totalSize=1267575]
OK
Time taken: 0.842 seconds

[cloudera@quickstart ~]$ hadoop fs -ls ratings_2013.txt
ls: `ratings_2013.txt': No such file or directory

[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/ratings
Found 2 items
-rw-r--r--   1 cloudera supergroup      27025 2017-03-18 12:51
/user/hive/warehouse/ratings/ratings_2012.txt
-rwxrwxrwx   1 cloudera cloudera      1240550 2017-03-18 12:57
/user/hive/warehouse/ratings/ratings_2013.txt
```

Question: Finally, count the records in the ratings table to ensure that all

21,997 are available. How many ratings are there?

```
hive> SELECT COUNT(*) FROM ratings;
Query ID = cloudera_20170318125656_d13c904c-285b-4992-b1b4-2dda6ff9844a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1487633215935_0074, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1487633215935_0074/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1487633215935_0074
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-18 13:01:24,066 Stage-1 map = 0%,  reduce = 0%
2017-03-18 13:01:38,547 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.35
sec
2017-03-18 13:01:52,570 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.82
sec
MapReduce Total cumulative CPU time: 4 seconds 820 msec
Ended Job = job_1487633215935_0074
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.82 sec   HDFS Read:
1274422 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 820 msec
OK
_c0
21997
Time taken: 47.896 seconds, Fetched: 1 row(s)

hive> CREATE TABLE loyalty_program (cust_id INT, fname STRING, lname STRING,
email STRING, level STRING, phone MAP<STRING, STRING>, order_ids ARRAY<INT>,
order_value STRUCT<min: INT, max: INT, avg: INT, total: INT>) ROW FORMAT
DELIMITED FIELDS TERMINATED BY '|' COLLECTION ITEMS TERMINATED BY ',' MAP KEYS
TERMINATED BY ':';
OK
Time taken: 0.476 seconds

hive> DESCRIBE loyalty_program;
OK
col_name    data_type   comment
cust_id                 int
fname                   string
lname                   string
email                   string
level                   string
```

```
phone                   map<string,string>
order_ids               array<int>
order_value             struct<min:int,max:int,avg:int,total:int>
Time taken: 0.199 seconds, Fetched: 8 row(s)

hive> LOAD DATA LOCAL INPATH 'loyalty_data.txt' INTO TABLE
loyalty_program;
hive> SELECT * FROM loyalty_program LIMIT 5;
OK
loyalty_program.cust_id loyalty_program.fname   loyalty_program.lname
loyalty_program.email   loyalty_program.level   loyalty_program.phone
loyalty_program.order_ids     loyalty_program.order_value
1000238     Christy     Herrinchristy.herrin@example.com    SILVER
{"MOBILE":"918-555-1162"}
[5179798,5346469,5517663,5783754,5811408,5828487,5838891,5854423,5864072,590449
6,5927566,5930762,5933463,5939057,5989823,6086708,6122093,6136912,6196295,62236
70,6258692,6327973,6373285,6384529,6481652,6484329,6485890,6531025,6612924]
{"min":409,"max":83726,"avg":15579,"total":451794}
1000279     Casey Francofranco81@example.com    SILVER{"WORK":"916-555-2791"}
[5262426,5307405,5477142,5507578,5609963,5640325,5714567,5809293,5854218,586447
8,6052001,6111750,6116797,6128791,6134800,6136527,6153194,6161430,6287762,63972
09,6474968,6544432]     {"min":529,"max":74626,"avg":19800,"total":435615}
1000810     Adam  Montoya     amontoya@example.com    SILVER
{"WORK":"415-555-4950"}
[5006384,5057993,5220993,5401325,5633591,5650325,5836641,5874250,5903066,598785
4,6196986,6254265]{"min":409,"max":418177,"avg":46873,"total":562486}
1001219     Ervin Groff eg1981@example.comSILVER{"WORK":"915-555-7945"}
[5104660,5287498,5679336,5790583,5791777,5795316,5844532,5855379,5901252,593823
9,5951563,5952299,5955489,5988409,5994480,6003787,6118472,6151719,6163195,62133
75,6220044,6231296,6326088,6360650,6391878]
{"min":808,"max":62818,"avg":18312,"total":457810}
1001661     Jody  Culverjody.culver@example.com SILVER
{"MOBILE":"515-555-8686","HOME":"515-555-2233"}
[5002071,5132660,5459665,5478462,5546890,5588290,5782846,5854706,6050353,607076
9,6097523,6168986,6171973,6208380,6238245,6372287,6383168,6485238,6503277,65856
55,6608572] {"min":119,"max":71857,"avg":26562,"total":557817}
Time taken: 0.107 seconds, Fetched: 5 row(s)
```

Show the 3 queries that you ran.

Question: 1. Select the HOME phone number (Hint: Map keys are case-sensitive) for customer ID 1200866. You should see 408-555-4914 as the result.

```
hive> SELECT phone["HOME"] FROM loyalty_program WHERE cust_id=1200866;
OK
_c0
408-555-4914
Time taken: 0.242 seconds, Fetched: 1 row(s)
```

Question: 2. Select the third element from the order_ids array for customer ID 1200866 (Hint: Elements are indexed from zero). The query should return 5278505.

```
hive> SELECT order_ids[2] FROM loyalty_program WHERE cust_id=1200866;
OK
_c0
5278505
Time taken: 0.139 seconds, Fetched: 1 row(s)
```

Quesstion: 3. Select the total attribute from the order_value struct for customer ID 1200866. The query should return 401874.

```
hive> SELECT order_value.total FROM loyalty_program WHERE cust_id=1200866;
OK
total
401874
Time taken: 0.101 seconds, Fetched: 1 row(s)
```

Alter and Drop a Table

Show the queries that you ran for steps 1 - 5.

Quesstion: 1. Use ALTER TABLE to rename the level column to status.

```
hive> ALTER TABLE loyalty_program CHANGE level status STRING;
OK
Time taken: 0.489 seconds
```

Quesstion: 2. Use the DESCRIBE command on the loyalty_program table to verify the change.

```
hive> DESCRIBE loyalty_program;
OK
col_name    data_type    comment
cust_id                  int
fname                    string
lname                    string
email                    string
status                   string
phone                    map<string,string>
order_ids                array<int>
order_value              struct<min:int,max:int,avg:int,total:int>
Time taken: 0.182 seconds, Fetched: 8 row(s)
```

Quesstion: 3. Use ALTER TABLE to rename the entire table to reward_program.

```
hive> ALTER TABLE loyalty_program RENAME TO reward_program;
OK
Time taken: 0.386 seconds

hive> DESCRIBE reward_program;
OK
```

```
col_name    data_type    comment
cust_id                  int
fname                    string
lname                    string
email                    string
status                   string
phone                    map<string,string>
order_ids                array<int>
order_value              struct<min:int,max:int,avg:int,total:int>
Time taken: 0.198 seconds, Fetched: 8 row(s)
```

Question: 4. Although the ALTER TABLE command often requires that we make a corresponding change to the data in HDFS, renaming a table or column does not. You can verify this by running a query on the table using the new names (the result should be "SILVER").

```
hive> SELECT status FROM reward_program WHERE cust_id=1200866;
OK
status
SILVER
Time taken: 0.255 seconds, Fetched: 1 row(s)
```

Question: 5. As sometimes happens in the corporate world, priorities have shifted and the program is now canceled. Drop the reward_program table.

```
hive> DROP TABLE IF EXISTS reward_program;
OK
Time taken: 1.396 seconds
```

This is the end of Part II

Data Management with Hive

Question: Create a table, using RegexSerde, to load access_log.gz into Hive. Your new table should contain entries for IP address, date_and_time, request, response and bytes_read.

Bonus question: Show the query you ran to create the table.

```
hive> DROP TABLE IF EXISTS access_log;
OK
Time taken: 0.254 seconds

hive> CREATE TABLE access_log (ip_address STRING, date_and_time STRING, request
STRING, status_code STRING, content_length STRING) ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.RegexSerDe' WITH SERDEPROPERTIES ("input.regex"
= "([^ ]*) [^ ]* [^ ]* (-|\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|[0-9]*)
(-|[0-9]*)", "output.format.string" = "%1$s %2$s %3$s %4$s %5$s") STORED AS
TEXTFILE;
OK
Time taken: 0.104 seconds

hive> DESCRIBE access_log;
OK
ip_address              string
date_and_time           string
request                 string
status_code             string
content_length          string
Time taken: 0.121 seconds, Fetched: 5 row(s)

hive> LOAD DATA LOCAL INPATH "file:///home/cloudera/datasets/access_log" INTO
TABLE access_log;
Loading data to table default.access_log
Table default.access_log stats: [numFiles=1, totalSize=56062392]
OK
Time taken: 0.947 seconds

hive> SELECT * FROM access_log LIMIT 10;
OK
10.223.157.186    [15/Jul/2009:14:58:59 -0700]  "GET /favicon.ico HTTP/1.1"
404   209
10.223.157.186    [15/Jul/2009:15:50:35 -0700]  "GET / HTTP/1.1"  200   9157
10.223.157.186    [15/Jul/2009:15:50:35 -0700]  "GET /assets/js/lowpro.js
HTTP/1.1"   200   10469
10.223.157.186    [15/Jul/2009:15:50:35 -0700]  "GET /assets/css/reset.css
```

```
HTTP/1.1"   200   1014
10.223.157.186    [15/Jul/2009:15:50:35 -0700]  "GET /assets/css/960.css
HTTP/1.1"   200   6206
10.223.157.186    [15/Jul/2009:15:50:35 -0700]  "GET
/assets/css/the-associates.css HTTP/1.1"  200   15779
10.223.157.186    [15/Jul/2009:15:50:35 -0700]  "GET
/assets/js/the-associates.js HTTP/1.1"    200   4492
10.223.157.186    [15/Jul/2009:15:50:35 -0700]  "GET /assets/js/lightbox.js
HTTP/1.1"   200   25960
10.223.157.186    [15/Jul/2009:15:50:36 -0700]  "GET
/assets/img/search-button.gif HTTP/1.1"   200   168
10.223.157.186    [15/Jul/2009:15:50:36 -0700]  "GET
/assets/img/dummy/secondary-news-3.jpg HTTP/1.1"200   5604
Time taken: 0.12 seconds, Fetched: 10 row(s)
```