SNAPSHOTS

```
Hadoop1 [Running]
g after reserving slots (ms)=0
16/05/09 04:47:16 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
16/05/09 04:47:16 INFO mapred.JobClient:   Map-Reduce Framework
16/05/09 04:47:16 INFO mapred.JobClient:     Map input records=2
16/05/09 04:47:16 INFO mapred.JobClient:     Map output records=8
16/05/09 04:47:16 INFO mapred.JobClient:     Map output bytes=82
16/05/09 04:47:16 INFO mapred.JobClient:     Input split bytes=212
16/05/09 04:47:16 INFO mapred.JobClient:     Combine input records=8
16/05/09 04:47:16 INFO mapred.JobClient:     Combine output records=6
16/05/09 04:47:16 INFO mapred.JobClient:     Reduce input groups=5
16/05/09 04:47:16 INFO mapred.JobClient:     Reduce shuffle bytes=85
16/05/09 04:47:16 INFO mapred.JobClient:     Reduce input records=6
16/05/09 04:47:16 INFO mapred.JobClient:     Reduce output records=5
16/05/09 04:47:16 INFO mapred.JobClient:     Spilled Records=12
16/05/09 04:47:16 INFO mapred.JobClient:     CPU time spent (ms)=1010
16/05/09 04:47:16 INFO mapred.JobClient:     Physical memory (bytes) snapshot=34
7971584
16/05/09 04:47:16 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=116
3071488
16/05/09 04:47:16 INFO mapred.JobClient:     Total committed heap usage (bytes)=
337780736
[training@localhost java-example]$ hadoop fs -ls /user/csds/output
Found 3 items
-rw-r--r--   1 training supergroup          0 2016-05-09 04:47 /user/csds/output
/_SUCCESS
drwxr-xr-x   - training supergroup          0 2016-05-09 04:47 /user/csds/output
/_logs
-rw-r--r--   1 training supergroup         41 2016-05-09 04:47 /user/csds/output
/part-r-00000
[training@localhost java-example]$ hadoop fs -cat /user/csds/output/part-r-00000
Bye     1
Goodbye 1
Hadoop  2
Hello   2
World   2
[training@localhost java-example]$
```

Figure 1- Map Reduce job in Java

```
[training@localhost python-example]$ ls
mapper.py  reducer.py
[training@localhost python-example]$ gedit mapper.py
[training@localhost python-example]$ hadoop fs -rm -r -f  /user/csds/outputpy/
Deleted /user/csds/outputpy
[training@localhost python-example]$ hadoop fs -rm  /user/csds/input/file2~
rm: `/user/csds/input/file2~': No such file or directory
[training@localhost python-example]$ hs mapper.py reducer.py /user/csds/input/* /user/csds/outputpy
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-training/hadoop-unjar7501281590360115193/] [] /tmp/streamjob86604553634332
24330.jar tmpDir=null
16/05/09 06:31:13 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement To
ol for the same.
16/05/09 06:31:13 WARN snappy.LoadSnappy: Snappy native library is available
16/05/09 06:31:13 INFO snappy.LoadSnappy: Snappy native library loaded
16/05/09 06:31:13 INFO mapred.FileInputFormat: Total input paths to process : 2
16/05/09 06:31:13 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/mapred/local]
16/05/09 06:31:13 INFO streaming.StreamJob: Running job: job_201605090422_0009
16/05/09 06:31:13 INFO streaming.StreamJob: To kill this job, run:
16/05/09 06:31:13 INFO streaming.StreamJob: UNDEF/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201605090422_00
09
16/05/09 06:31:13 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201605090422_0009
16/05/09 06:31:14 INFO streaming.StreamJob:  map 0%  reduce 0%
16/05/09 06:31:18 INFO streaming.StreamJob:  map 100%  reduce 0%
16/05/09 06:31:20 INFO streaming.StreamJob:  map 100%  reduce 100%
16/05/09 06:31:21 INFO streaming.StreamJob: Job complete: job_201605090422_0009
16/05/09 06:31:21 INFO streaming.StreamJob: Output: /user/csds/outputpy
[training@localhost python-example]$ hadoop fs -ls  /user/csds/outputpy/
Found 3 items
-rw-r--r--   1 training supergroup          0 2016-05-09 06:31 /user/csds/outputpy/_SUCCESS
drwxr-xr-x   - training supergroup          0 2016-05-09 06:31 /user/csds/outputpy/_logs
-rw-r--r--   1 training supergroup         41 2016-05-09 06:31 /user/csds/outputpy/part-00000
[training@localhost python-example]$ hadoop fs -cat  /user/csds/outputpy/part-00000
Bye     1
Goodbye 1
Hadoop  2
Hello   2
World   2
[training@localhost python-example]$
```

Figure 2- Map reduce output in Python using Hadoop Streaming

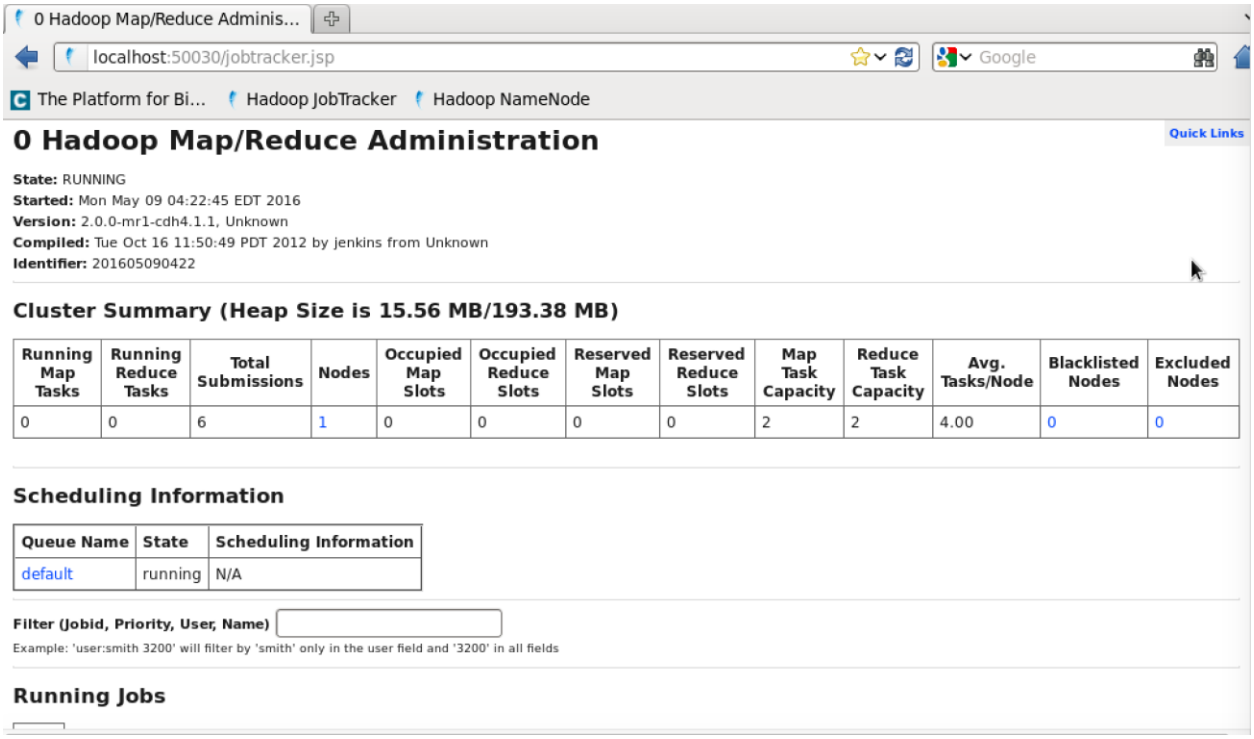
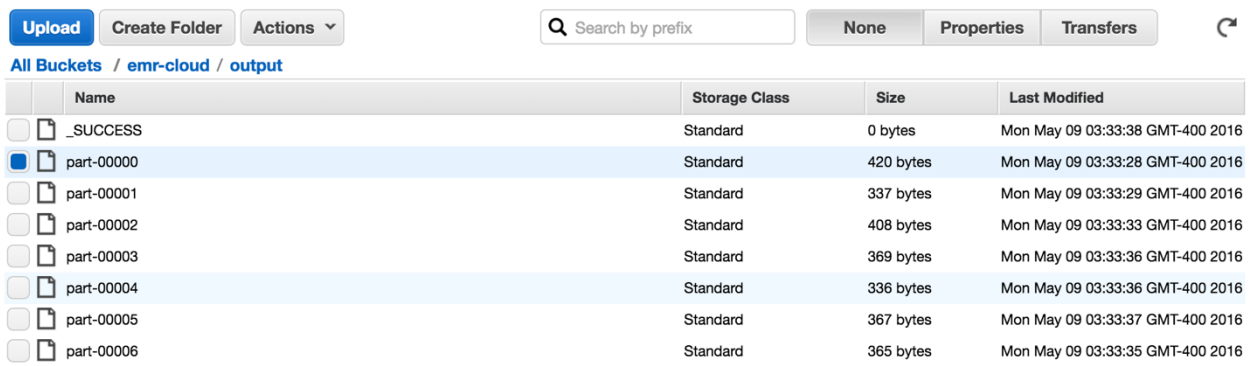


Figure 3- Map Reduce Administration Page



Name	Storage Class	Size	Last Modified
_SUCCESS	Standard	0 bytes	Mon May 09 03:33:38 GMT-400 2016
part-00000	Standard	420 bytes	Mon May 09 03:33:28 GMT-400 2016
part-00001	Standard	337 bytes	Mon May 09 03:33:29 GMT-400 2016
part-00002	Standard	408 bytes	Mon May 09 03:33:33 GMT-400 2016
part-00003	Standard	369 bytes	Mon May 09 03:33:36 GMT-400 2016
part-00004	Standard	336 bytes	Mon May 09 03:33:36 GMT-400 2016
part-00005	Standard	367 bytes	Mon May 09 03:33:37 GMT-400 2016
part-00006	Standard	365 bytes	Mon May 09 03:33:35 GMT-400 2016

Figure 4- Output of Map reduce job on Amazon EMR

```
Applications  Places  System                                   Wed May 11, 2:18 AM  trainin
training@localhost:~/udacity_training/data
File  Edit  View  Search  Terminal  Help
[training@localhost data]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
Hive history file=/tmp/training/hive_job_log_training_201605110216_1316099354.txt
hive> show tables;
OK
Time taken: 1.863 seconds
hive> CREATE TABLE test_tables;
FAILED: SemanticException [Error 10043]: Either list of columns or a custom serializer should be specified
hive> CREATE TABLE test_tables (some_text String);
OK
Time taken: 0.462 seconds
hive> show tables;
OK
test_tables
Time taken: 0.078 seconds
hive> select * from test_tables
    > ;
OK
Time taken: 0.15 seconds
hive> [training@localhost data]$ head purchases.txt
2012-01-01      09:00   San Jose        Men's Clothing  214.05  Amex
2012-01-01      09:00   Fort Worth      Women's Clothing        153.57  Visa
2012-01-01      09:00   San Diego       Music   66.08   Cash
2012-01-01      09:00   Pittsburgh      Pet Supplies    493.51  Discover
2012-01-01      09:00   Omaha   Children's Clothing     235.63  MasterCard
2012-01-01      09:00   Stockton        Men's Clothing  247.18  MasterCard
2012-01-01      09:00   Austin  Cameras 379.6   Visa
2012-01-01      09:00   New York        Consumer Electronics    296.8   Cash
2012-01-01      09:00   Corpus Christi  Toys    25.38   Discover
2012-01-01      09:00   Fort Worth      Toys    213.88  Visa
[training@localhost data]$
```

Figure 5- Starting up hive on Cloudera vm

```
File  Edit  View  Search  Terminal  Help
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 3.69 sec   HDFS Read: 0 HDFS
Total MapReduce CPU Time Spent: 3 seconds 690 msec
OK
NULL
Time taken: 9.464 seconds
hive> select * from purchases limit 10;
OK
2012-01-01      09:00   San Jose        Men's Clothing  214.05  Amex
2012-01-01      09:00   Fort Worth      Women's Clothing        153.57
2012-01-01      09:00   San Diego       Music   66.08   Cash    NULL
2012-01-01      09:00   Pittsburgh      Pet Supplies    493.51  Discove
2012-01-01      09:00   Omaha   Children's Clothing     235.63  MasterC
2012-01-01      09:00   Stockton        Men's Clothing  247.18  MasterC
2012-01-01      09:00   Austin  Cameras 379.6   Visa    NULL    NULL
2012-01-01      09:00   New York        Consumer Electronics    296.8
2012-01-01      09:00   Corpus Christi  Toys    25.38   Discover
2012-01-01      09:00   Fort Worth      Toys    213.88  Visa    NULL
Time taken: 0.103 seconds
hive> select price from purchases limit 10;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201605110214_0006, Tracking URL = http://0.0.0.0:500
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0
Hadoop job information for Stage-1: number of mappers: 1; number of red
2016-05-11 03:01:13,864 Stage-1 map = 0%,  reduce = 0%
2016-05-11 03:01:14,869 Stage-1 map = 100%,  reduce = 0%, Cumulative CP
2016-05-11 03:01:15,875 Stage-1 map = 100%,  reduce = 0%, Cumulative CP
2016-05-11 03:01:16,882 Stage-1 map = 100%,  reduce = 100%, Cumulative
MapReduce Total cumulative CPU time: 350 msec
Ended Job = job_201605110214_0006
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 0.35 sec   HDFS Read: 0 HDFS Write: 0 S
Total MapReduce CPU Time Spent: 350 msec
```

Left ⌘

Figure 5- Hive jobs running in Cloudera VM and purchases table filled