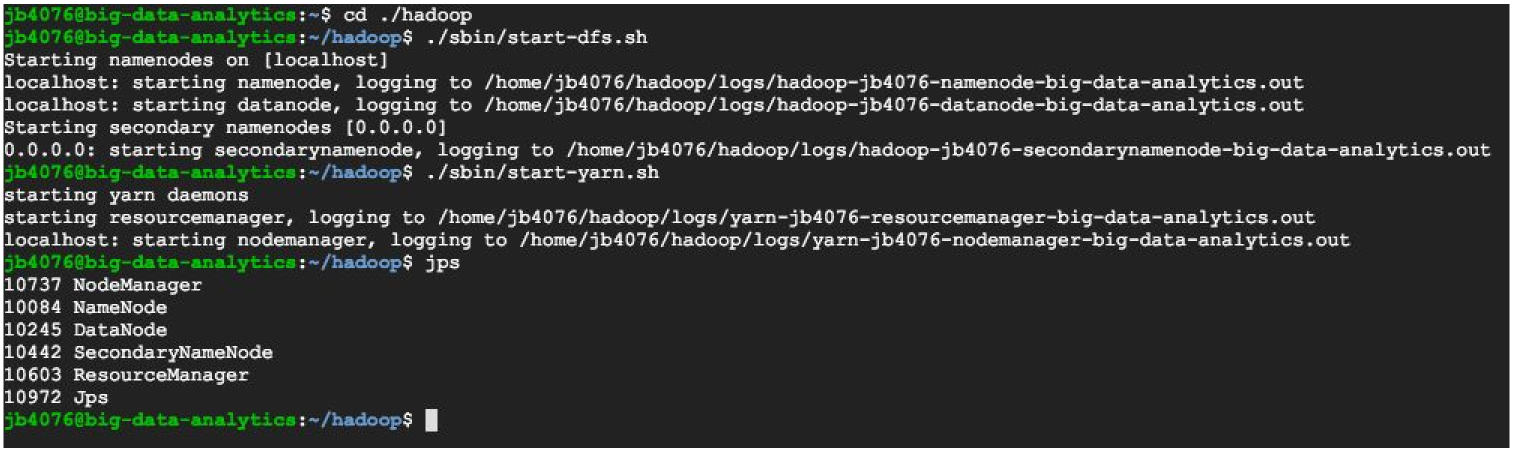
First-try on Hadoop, HBase, Hive, Spark, Jupyter

1. Installation and Setup

1.1 Start hadoop

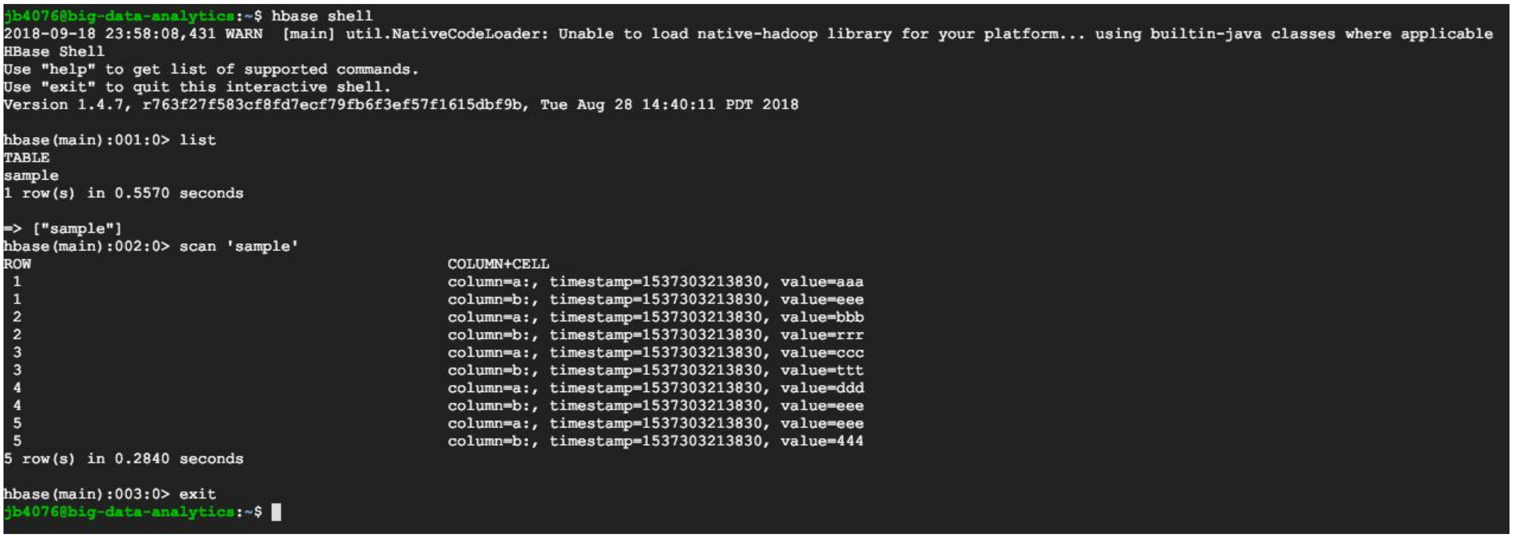


```
jb4076@big-data-analytics:~$ cd ./hadoop
jb4076@big-data-analytics:~/hadoop$ ./sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/jb4076/hadoop/logs/hadoop-jb4076-namenode-big-data-analytics.out
localhost: starting datanode, logging to /home/jb4076/hadoop/logs/hadoop-jb4076-datanode-big-data-analytics.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/jb4076/hadoop/logs/hadoop-jb4076-secondarynamenode-big-data-analytics.out
jb4076@big-data-analytics:~/hadoop$ ./sbin/start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/jb4076/hadoop/logs/yarn-jb4076-resourcemanager-big-data-analytics.out
localhost: starting nodemanager, logging to /home/jb4076/hadoop/logs/yarn-jb4076-nodemanager-big-data-analytics.out
jb4076@big-data-analytics:~/hadoop$ jps
10737 NodeManager
10084 NameNode
10245 DataNode
10442 SecondaryNameNode
10603 ResourceManager
10972 Jps
jb4076@big-data-analytics:~/hadoop$ 
```

script:

```
jb4076@big-data-analytics:~$ cd ./hadoop
jb4076@big-data-analytics:~/hadoop$ ./sbin/stop-dfs.sh
jb4076@big-data-analytics:~/hadoop$ ./sbin/stop-yarn.sh
jb4076@big-data-analytics:~/hadoop$ jps
```

1.2 Start HBase

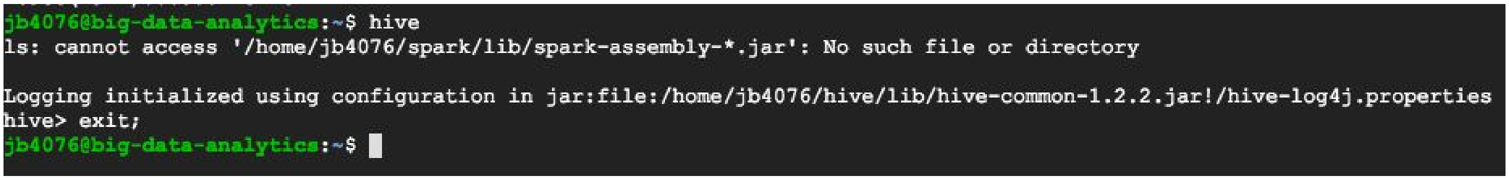


```
jb4076@big-data-analytics:~$ hbase shell
2018-09-18 23:58:08,431 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.7, r763f27f583cf8fd7ecf79fb6f3ef57f1615dbf9b, Tue Aug 28 14:40:11 PDT 2018

hbase(main):001:0> list
TABLE
sample
1 row(s) in 0.5570 seconds

=> ["sample"]
hbase(main):002:0> scan 'sample'
ROW                                             COLUMN+CELL
 1                                              column=a:, timestamp=1537303213830, value=aaa
 1                                              column=b:, timestamp=1537303213830, value=eee
 2                                              column=a:, timestamp=1537303213830, value=bbb
 2                                              column=b:, timestamp=1537303213830, value=rrr
 3                                              column=a:, timestamp=1537303213830, value=ccc
 3                                              column=b:, timestamp=1537303213830, value=ttt
 4                                              column=a:, timestamp=1537303213830, value=ddd
 4                                              column=b:, timestamp=1537303213830, value=eee
 5                                              column=a:, timestamp=1537303213830, value=eee
 5                                              column=b:, timestamp=1537303213830, value=444
5 row(s) in 0.2840 seconds

hbase(main):003:0> exit
jb4076@big-data-analytics:~$ 
```

script:

```
jb4076@big-data-analytics:~$ hbase shell
hbase(main):001:0> list
hbase(main):002:0> scan 'sample'
hbase(main):003:0> exit
```

1.3 Start Hive

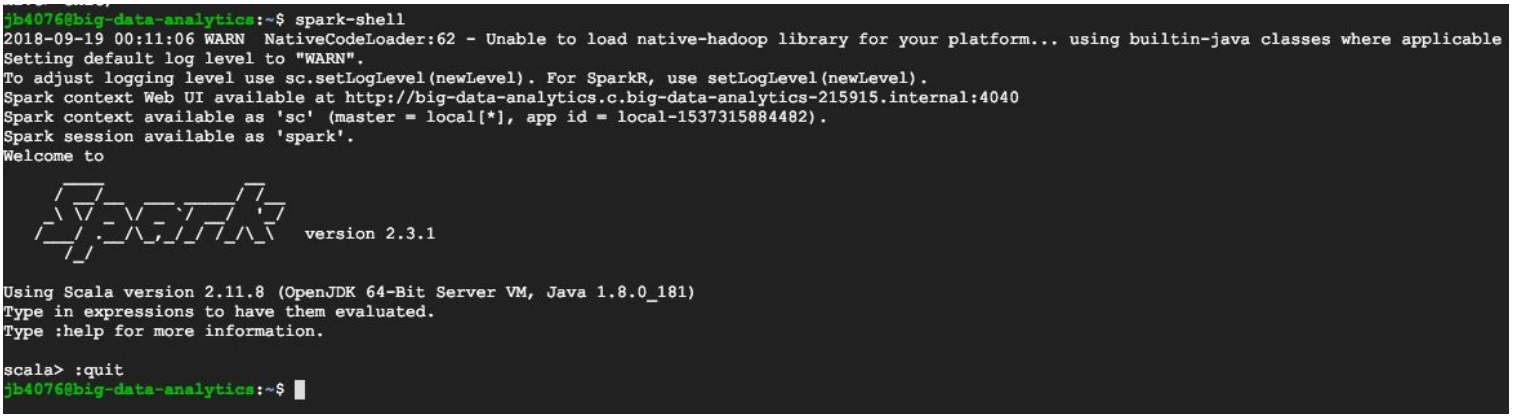


```
jb4076@big-data-analytics:~$ hive
ls: cannot access '/home/jb4076/spark/lib/spark-assembly-*.jar': No such file or directory

Logging initialized using configuration in jar:file:/home/jb4076/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties
hive> exit;
jb4076@big-data-analytics:~$
```

script:

```
jb4076@big-data-analytics:~$ hive
hive> exit;
```

1.4 Start spark

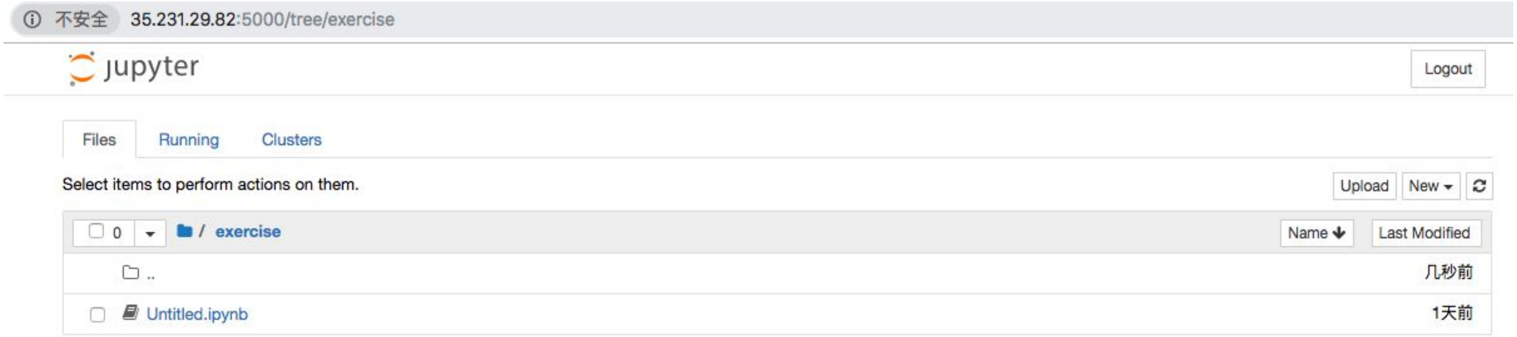


```
jb4076@big-data-analytics:~$ spark-shell
2018-09-19 00:11:06 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://big-data-analytics.c.big-data-analytics-215915.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1537315884482).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.1
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.

scala> :quit
jb4076@big-data-analytics:~$
```

script:

```
jb4076@big-data-analytics:~$ spark-shell
scala> :quit
```

1.5 Start jupyter notebook



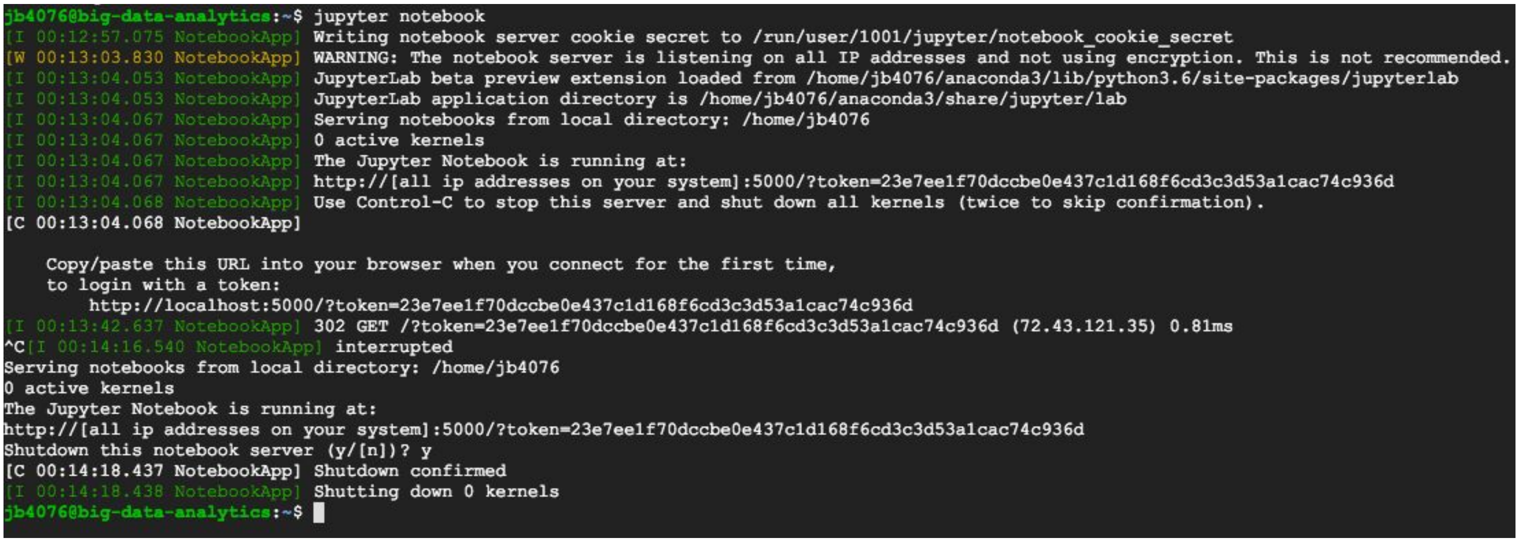
ⓘ 不安全 35.231.29.82:5000/tree/exercise

Select items to perform actions on them.

/ exercise

.. 几秒前

□ Untitled.ipynb 1天前



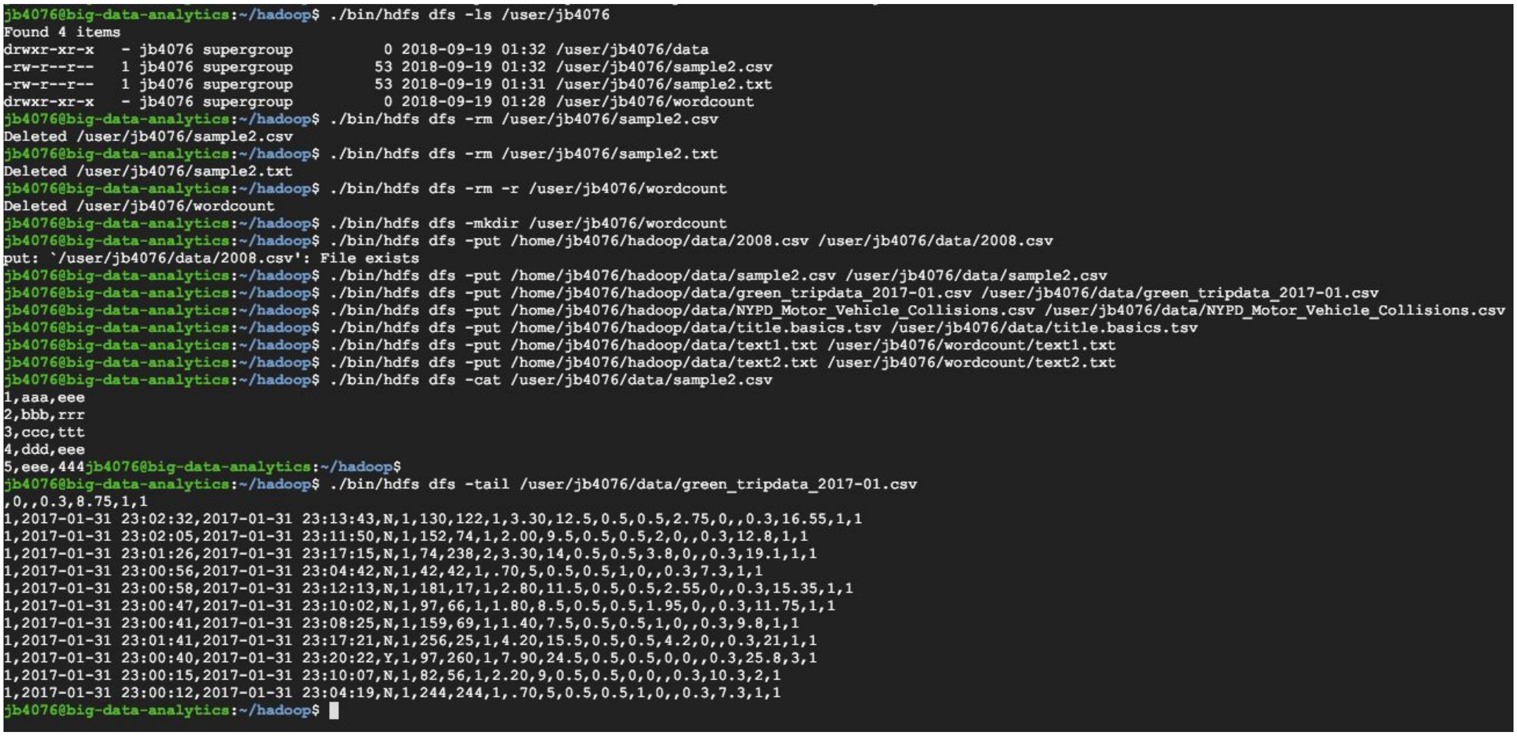
```
jb4076@big-data-analytics:~$ jupyter notebook
[I 00:12:57.075 NotebookApp] Writing notebook server cookie secret to /run/user/1001/jupyter/notebook_cookie_secret
[W 00:13:03.830 NotebookApp] WARNING: The notebook server is listening on all IP addresses and not using encryption. This is not recommended.
[I 00:13:04.053 NotebookApp] JupyterLab beta preview extension loaded from /home/jb4076/anaconda3/lib/python3.6/site-packages/jupyterlab
[I 00:13:04.053 NotebookApp] JupyterLab application directory is /home/jb4076/anaconda3/share/jupyter/lab
[I 00:13:04.067 NotebookApp] Serving notebooks from local directory: /home/jb4076
[I 00:13:04.067 NotebookApp] 0 active kernels
[I 00:13:04.067 NotebookApp] The Jupyter Notebook is running at:
[I 00:13:04.067 NotebookApp] http://[all ip addresses on your system]:5000/?token=23e7ee1f70dccbe0e437c1d168f6cd3c3d53a1cac74c936d
[I 00:13:04.068 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 00:13:04.068 NotebookApp]

    Copy/paste this URL into your browser when you connect for the first time,
    to login with a token:
        http://localhost:5000/?token=23e7ee1f70dccbe0e437c1d168f6cd3c3d53a1cac74c936d
[I 00:13:42.637 NotebookApp] 302 GET /?token=23e7ee1f70dccbe0e437c1d168f6cd3c3d53a1cac74c936d (72.43.121.35) 0.81ms
^C[I 00:14:16.540 NotebookApp] interrupted
Serving notebooks from local directory: /home/jb4076
0 active kernels
The Jupyter Notebook is running at:
http://[all ip addresses on your system]:5000/?token=23e7ee1f70dccbe0e437c1d168f6cd3c3d53a1cac74c936d
Shutdown this notebook server (y/[n])? y
[C 00:14:18.437 NotebookApp] Shutdown confirmed
[I 00:14:18.438 NotebookApp] Shutting down 0 kernels
jb4076@big-data-analytics:~$ 
```

script:

```
jb4076@big-data-analytics:~$ jupyter notebook
```

2. Hadoop

2.1 Demonstrate you can manage your file systems

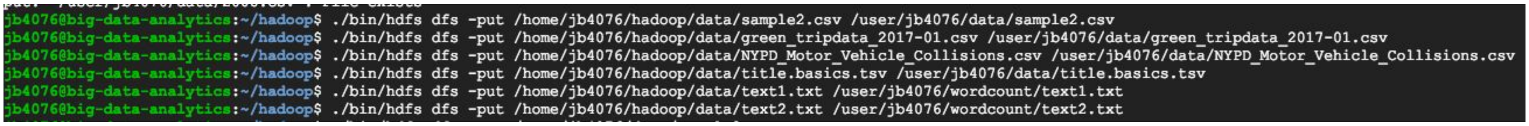


```
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -ls /user/jb4076
Found 4 items
drwxr-xr-x   - jb4076 supergroup          0 2018-09-19 01:32 /user/jb4076/data
-rw-r--r--   1 jb4076 supergroup         53 2018-09-19 01:32 /user/jb4076/sample2.csv
-rw-r--r--   1 jb4076 supergroup         53 2018-09-19 01:31 /user/jb4076/sample2.txt
drwxr-xr-x   - jb4076 supergroup          0 2018-09-19 01:28 /user/jb4076/wordcount
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -rm /user/jb4076/sample2.csv
Deleted /user/jb4076/sample2.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -rm /user/jb4076/sample2.txt
Deleted /user/jb4076/sample2.txt
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -rm -r /user/jb4076/wordcount
Deleted /user/jb4076/wordcount
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -mkdir /user/jb4076/wordcount
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/2008.csv /user/jb4076/data/2008.csv
put: `/user/jb4076/data/2008.csv': File exists
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/sample2.csv /user/jb4076/data/sample2.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/green_tripdata_2017-01.csv /user/jb4076/data/green_tripdata_2017-01.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/NYPD_Motor_Vehicle_Collisions.csv /user/jb4076/data/NYPD_Motor_Vehicle_Collisions.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/title.basics.tsv /user/jb4076/data/title.basics.tsv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/text1.txt /user/jb4076/wordcount/text1.txt
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/text2.txt /user/jb4076/wordcount/text2.txt
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -cat /user/jb4076/data/sample2.csv
1,aaa,eee
2,bbb,rrr
3,ccc,ttt
4,ddd,eee
5,eee,444jb4076@big-data-analytics:~/hadoop$
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -tail /user/jb4076/data/green_tripdata_2017-01.csv
,0,,0.3,8.75,1,1
1,2017-01-31 23:02:32,2017-01-31 23:13:43,N,1,130,122,1,3.30,12.5,0.5,0.5,2.75,0,,0.3,16.55,1,1
1,2017-01-31 23:02:05,2017-01-31 23:11:50,N,1,152,74,1,2.00,9.5,0.5,0.5,2,0,,0.3,12.8,1,1
1,2017-01-31 23:01:26,2017-01-31 23:17:15,N,1,74,238,2,3.30,14,0.5,0.5,3.8,0,,0.3,19.1,1,1
1,2017-01-31 23:00:56,2017-01-31 23:04:42,N,1,42,42,1,.70,5,0.5,0.5,1,0,,0.3,7.3,1,1
1,2017-01-31 23:00:58,2017-01-31 23:12:13,N,1,181,17,1,2.80,11.5,0.5,0.5,2.55,0,,0.3,15.35,1,1
1,2017-01-31 23:00:47,2017-01-31 23:10:02,N,1,97,66,1,1.80,8.5,0.5,0.5,1.95,0,,0.3,11.75,1,1
1,2017-01-31 23:00:41,2017-01-31 23:08:25,N,1,159,69,1,1.40,7.5,0.5,0.5,1,0,,0.3,9.8,1,1
1,2017-01-31 23:01:41,2017-01-31 23:17:21,N,1,256,25,1,4.20,15.5,0.5,0.5,4.2,0,,0.3,21,1,1
1,2017-01-31 23:00:40,2017-01-31 23:20:22,Y,1,97,260,1,7.90,24.5,0.5,0.5,0,0,,0.3,25.8,3,1
1,2017-01-31 23:00:15,2017-01-31 23:10:07,N,1,82,56,1,2.20,9,0.5,0.5,0,0,,0.3,10.3,2,1
1,2017-01-31 23:00:12,2017-01-31 23:04:19,N,1,244,244,1,.70,5,0.5,0.5,1,0,,0.3,7.3,1,1
jb4076@big-data-analytics:~/hadoop$ 
```

script:

```
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -ls /user/jb4076
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -rm /user/jb4076/sample2.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -rm -r /user/jb4076/wordcount
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -mkdir /user/jb4076/wordcount
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put
/home/jb4076/hadoop/data/2008.csv /user/jb4076/data/2008.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -cat
/user/jb4076/data/sample2.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -tail
/user/jb4076/data/green_tripdata_2017-01.csv
```

2.2 Upload a file to HDFS



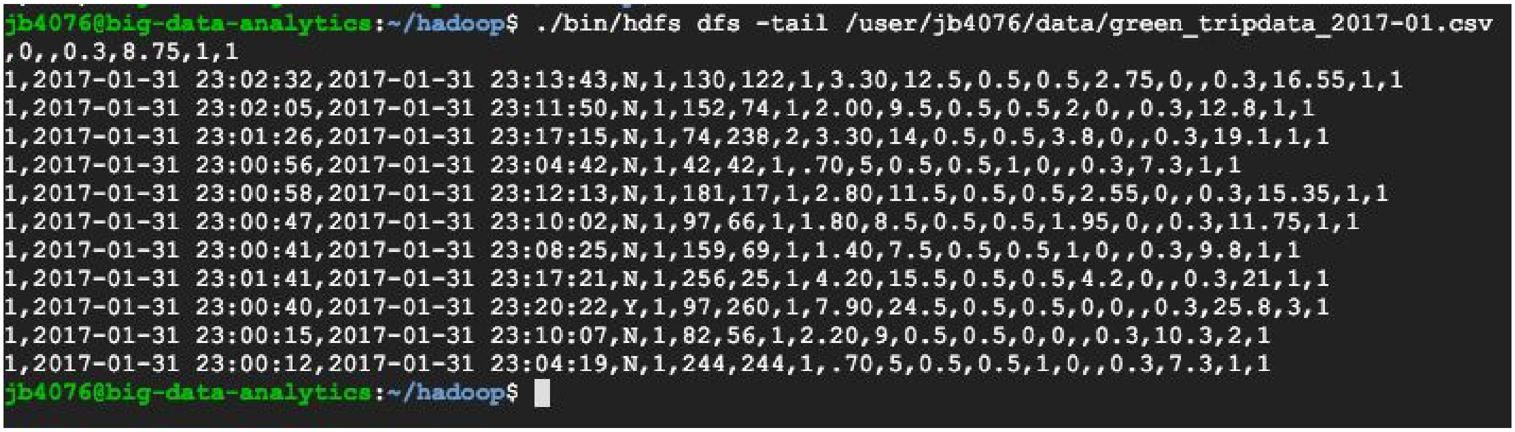
```
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/sample2.csv /user/jb4076/data/sample2.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/green_tripdata_2017-01.csv /user/jb4076/data/green_tripdata_2017-01.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/NYPD_Motor_Vehicle_Collisions.csv /user/jb4076/data/NYPD_Motor_Vehicle_Collisions.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/title.basics.tsv /user/jb4076/data/title.basics.tsv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/text1.txt /user/jb4076/wordcount/text1.txt
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/text2.txt /user/jb4076/wordcount/text2.txt
```

script:

```
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put
/home/jb4076/hadoop/data/sample2.csv /user/jb4076/data/sample2.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put
/home/jb4076/hadoop/data/green_tripdata_2017-01.csv
/user/jb4076/data/green_tripdata_2017-01.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put
/home/jb4076/hadoop/data/NYPD_Motor_Vehicle_Collisions.csv
/user/jb4076/data/NYPD_Motor_Vehicle_Collisions.csv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put
/home/jb4076/hadoop/data/title.basics.tsv /user/jb4076/data/title.basics.tsv
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put
/home/jb4076/hadoop/data/text1.txt /user/jb4076/wordcount/text1.txt
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put
/home/jb4076/hadoop/data/text2.txt /user/jb4076/wordcount/text2.txt
```

2.3 Inspect the last kilobytes of content of the file

Using NYC TLC Trip Data (2017 January Green Taxi) (green_tripdata_2017-01.csv)

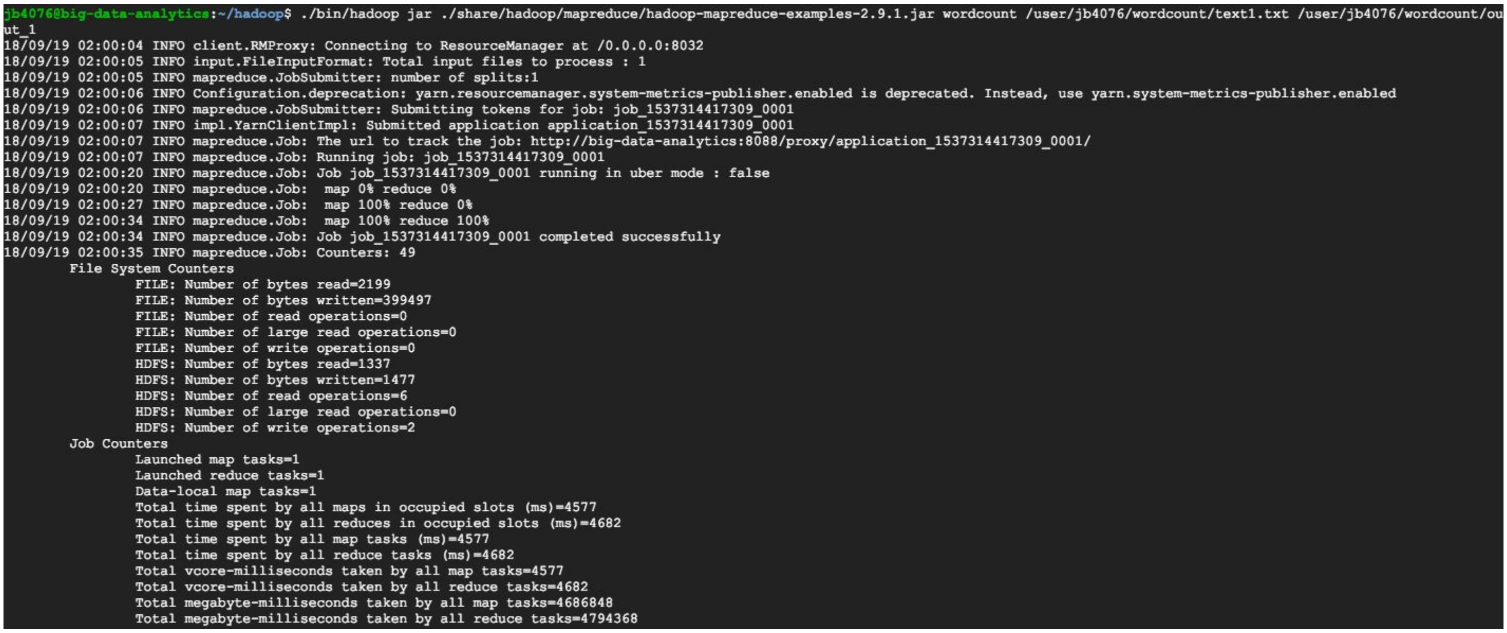


```
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -tail /user/jb4076/data/green_tripdata_2017-01.csv
,0,,0.3,8.75,1,1
1,2017-01-31 23:02:32,2017-01-31 23:13:43,N,1,130,122,1,3.30,12.5,0.5,0.5,2.75,0,,0.3,16.55,1,1
1,2017-01-31 23:02:05,2017-01-31 23:11:50,N,1,152,74,1,2.00,9.5,0.5,0.5,2,0,,0.3,12.8,1,1
1,2017-01-31 23:01:26,2017-01-31 23:17:15,N,1,74,238,2,3.30,14,0.5,0.5,3.8,0,,0.3,19.1,1,1
1,2017-01-31 23:00:56,2017-01-31 23:04:42,N,1,42,42,1,.70,5,0.5,0.5,1,0,,0.3,7.3,1,1
1,2017-01-31 23:00:58,2017-01-31 23:12:13,N,1,181,17,1,2.80,11.5,0.5,0.5,2.55,0,,0.3,15.35,1,1
1,2017-01-31 23:00:47,2017-01-31 23:10:02,N,1,97,66,1,1.80,8.5,0.5,0.5,1.95,0,,0.3,11.75,1,1
1,2017-01-31 23:00:41,2017-01-31 23:08:25,N,1,159,69,1,1.40,7.5,0.5,0.5,1,0,,0.3,9.8,1,1
1,2017-01-31 23:01:41,2017-01-31 23:17:21,N,1,256,25,1,4.20,15.5,0.5,0.5,4.2,0,,0.3,21,1,1
1,2017-01-31 23:00:40,2017-01-31 23:20:22,Y,1,97,260,1,7.90,24.5,0.5,0.5,0,0,,0.3,25.8,3,1
1,2017-01-31 23:00:15,2017-01-31 23:10:07,N,1,82,56,1,2.20,9,0.5,0.5,0,0,,0.3,10.3,2,1
1,2017-01-31 23:00:12,2017-01-31 23:04:19,N,1,244,244,1,.70,5,0.5,0.5,1,0,,0.3,7.3,1,1
jb4076@big-data-analytics:~/hadoop$
```

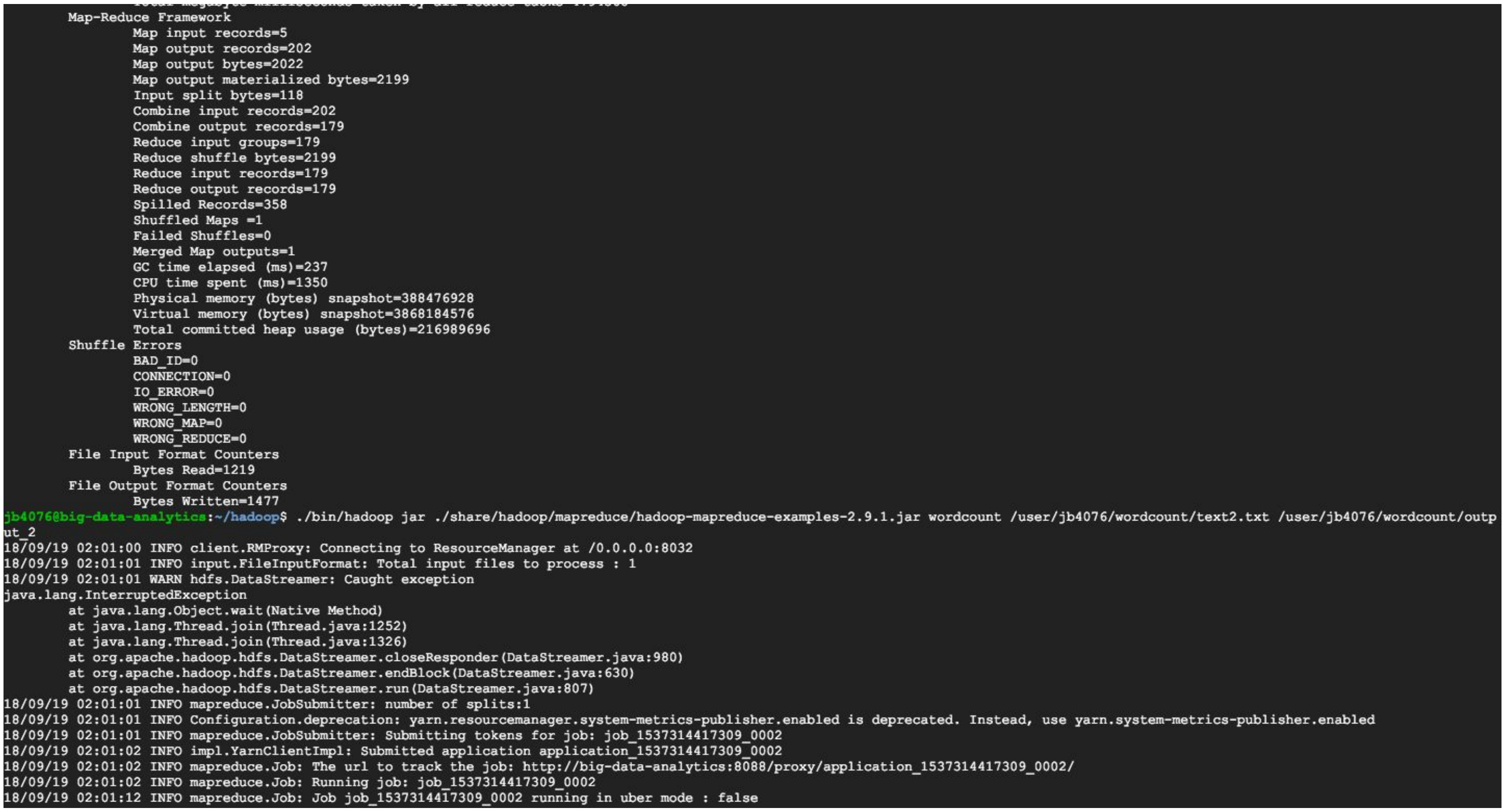
script:

```
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -tail
/user/jb4076/data/green_tripdata_2017-01.csv
```

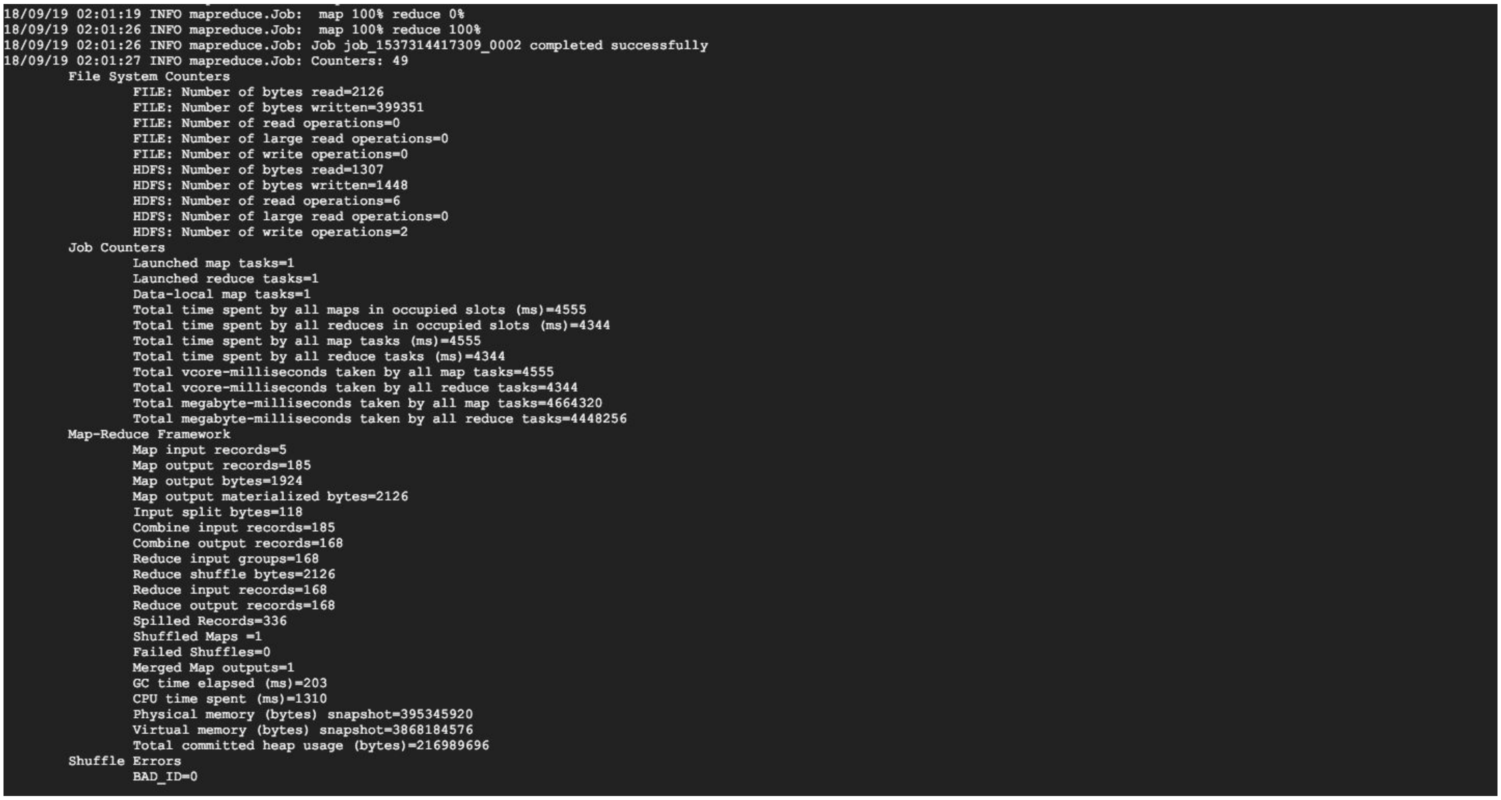
2.4 Run the mapreduce word count example with the provided 2 text files and find the 3 most frequent words



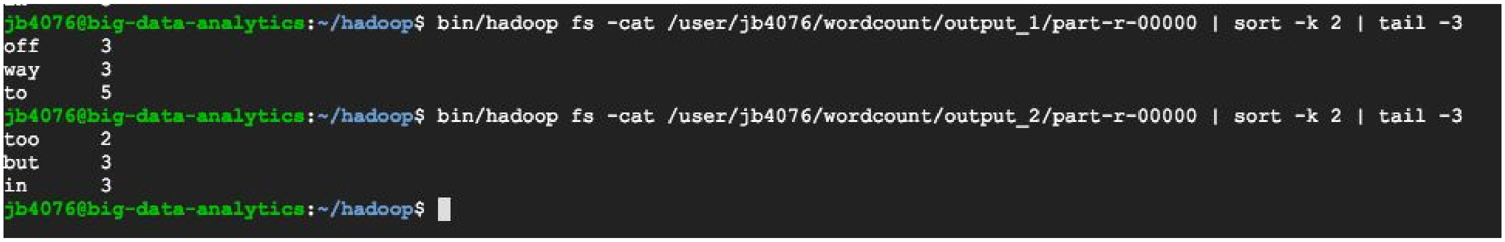
```
jb4076@big-data-analytics:~/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.1.jar wordcount /user/jb4076/wordcount/text1.txt /user/jb4076/wordcount/ou
ut_1
18/09/19 02:00:04 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/09/19 02:00:05 INFO input.FileInputFormat: Total input files to process : 1
18/09/19 02:00:05 INFO mapreduce.JobSubmitter: number of splits:1
18/09/19 02:00:06 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
18/09/19 02:00:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1537314417309_0001
18/09/19 02:00:07 INFO impl.YarnClientImpl: Submitted application application_1537314417309_0001
18/09/19 02:00:07 INFO mapreduce.Job: The url to track the job: http://big-data-analytics:8088/proxy/application_1537314417309_0001/
18/09/19 02:00:07 INFO mapreduce.Job: Running job: job_1537314417309_0001
18/09/19 02:00:20 INFO mapreduce.Job: Job job_1537314417309_0001 running in uber mode : false
18/09/19 02:00:20 INFO mapreduce.Job:  map 0% reduce 0%
18/09/19 02:00:27 INFO mapreduce.Job:  map 100% reduce 0%
18/09/19 02:00:34 INFO mapreduce.Job:  map 100% reduce 100%
18/09/19 02:00:34 INFO mapreduce.Job: Job job_1537314417309_0001 completed successfully
18/09/19 02:00:35 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=2199
                FILE: Number of bytes written=399497
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1337
                HDFS: Number of bytes written=1477
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4577
                Total time spent by all reduces in occupied slots (ms)=4682
                Total time spent by all map tasks (ms)=4577
                Total time spent by all reduce tasks (ms)=4682
                Total vcore-milliseconds taken by all map tasks=4577
                Total vcore-milliseconds taken by all reduce tasks=4682
                Total megabyte-milliseconds taken by all map tasks=4686848
                Total megabyte-milliseconds taken by all reduce tasks=4794368
```



```
        Map-Reduce Framework
                Map input records=5
                Map output records=202
                Map output bytes=2022
                Map output materialized bytes=2199
                Input split bytes=118
                Combine input records=202
                Combine output records=179
                Reduce input groups=179
                Reduce shuffle bytes=2199
                Reduce input records=179
                Reduce output records=179
                Spilled Records=358
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=237
                CPU time spent (ms)=1350
                Physical memory (bytes) snapshot=388476928
                Virtual memory (bytes) snapshot=3868184576
                Total committed heap usage (bytes)=216989696
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1219
        File Output Format Counters
                Bytes Written=1477
jb4076@big-data-analytics:~/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.1.jar wordcount /user/jb4076/wordcount/text2.txt /user/jb4076/wordcount/outp
ut_2
18/09/19 02:01:00 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/09/19 02:01:01 INFO input.FileInputFormat: Total input files to process : 1
18/09/19 02:01:01 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
18/09/19 02:01:01 INFO mapreduce.JobSubmitter: number of splits:1
18/09/19 02:01:01 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
18/09/19 02:01:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1537314417309_0002
18/09/19 02:01:02 INFO impl.YarnClientImpl: Submitted application application_1537314417309_0002
18/09/19 02:01:02 INFO mapreduce.Job: The url to track the job: http://big-data-analytics:8088/proxy/application_1537314417309_0002/
18/09/19 02:01:02 INFO mapreduce.Job: Running job: job_1537314417309_0002
18/09/19 02:01:12 INFO mapreduce.Job: Job job_1537314417309_0002 running in uber mode : false
```



```
18/09/19 02:01:19 INFO mapreduce.Job:  map 100% reduce 0%
18/09/19 02:01:26 INFO mapreduce.Job:  map 100% reduce 100%
18/09/19 02:01:26 INFO mapreduce.Job: Job job_1537314417309_0002 completed successfully
18/09/19 02:01:27 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=2126
                FILE: Number of bytes written=399351
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1307
                HDFS: Number of bytes written=1448
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4555
                Total time spent by all reduces in occupied slots (ms)=4344
                Total time spent by all map tasks (ms)=4555
                Total time spent by all reduce tasks (ms)=4344
                Total vcore-milliseconds taken by all map tasks=4555
                Total vcore-milliseconds taken by all reduce tasks=4344
                Total megabyte-milliseconds taken by all map tasks=4664320
                Total megabyte-milliseconds taken by all reduce tasks=4448256
        Map-Reduce Framework
                Map input records=5
                Map output records=185
                Map output bytes=1924
                Map output materialized bytes=2126
                Input split bytes=118
                Combine input records=185
                Combine output records=168
                Reduce input groups=168
                Reduce shuffle bytes=2126
                Reduce input records=168
                Reduce output records=168
                Spilled Records=336
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=203
                CPU time spent (ms)=1310
                Physical memory (bytes) snapshot=395345920
                Virtual memory (bytes) snapshot=3868184576
                Total committed heap usage (bytes)=216989696
        Shuffle Errors
                BAD_ID=0
```



```
jb4076@big-data-analytics:~/hadoop$ bin/hadoop fs -cat /user/jb4076/wordcount/output_1/part-r-00000 | sort -k 2 | tail -3
off     3
way     3
to      5
jb4076@big-data-analytics:~/hadoop$ bin/hadoop fs -cat /user/jb4076/wordcount/output_2/part-r-00000 | sort -k 2 | tail -3
too     2
but     3
in      3
jb4076@big-data-analytics:~/hadoop$ 
```

script:

```
jb4076@big-data-analytics:~/hadoop$ ./bin/hadoop jar
./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.1.jar wordcount
/user/jb4076/wordcount/text1.txt /user/jb4076/wordcount/output_1

jb4076@big-data-analytics:~/hadoop$ ./bin/hadoop jar
./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.1.jar wordcount
/user/jb4076/wordcount/text2.txt /user/jb4076/wordcount/output_2

jb4076@big-data-analytics:~/hadoop$ bin/hadoop fs -cat
/user/jb4076/wordcount/output_1/part-r-00000 | sort -k 2 | tail -3

jb4076@big-data-analytics:~/hadoop$ bin/hadoop fs -cat
/user/jb4076/wordcount/output_2/part-r-00000 | sort -k 2 | tail -3
```

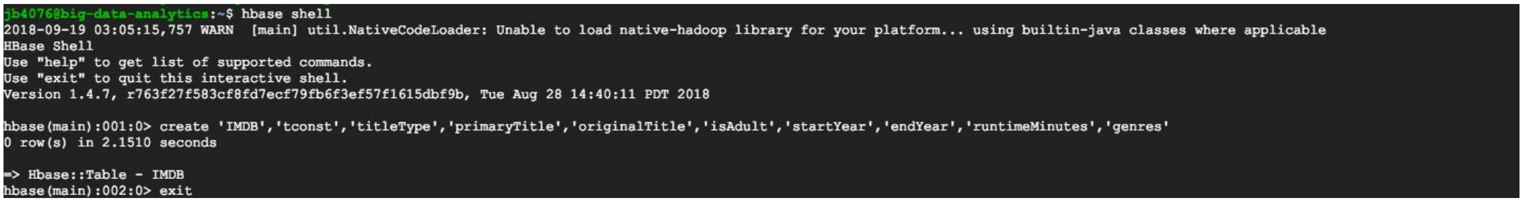
result:

The 3 most frequent words of text1.txt are “to”, “way”, and “off”.
The 3 most frequent words of text2.txt are “in”, “but”, and “to”.

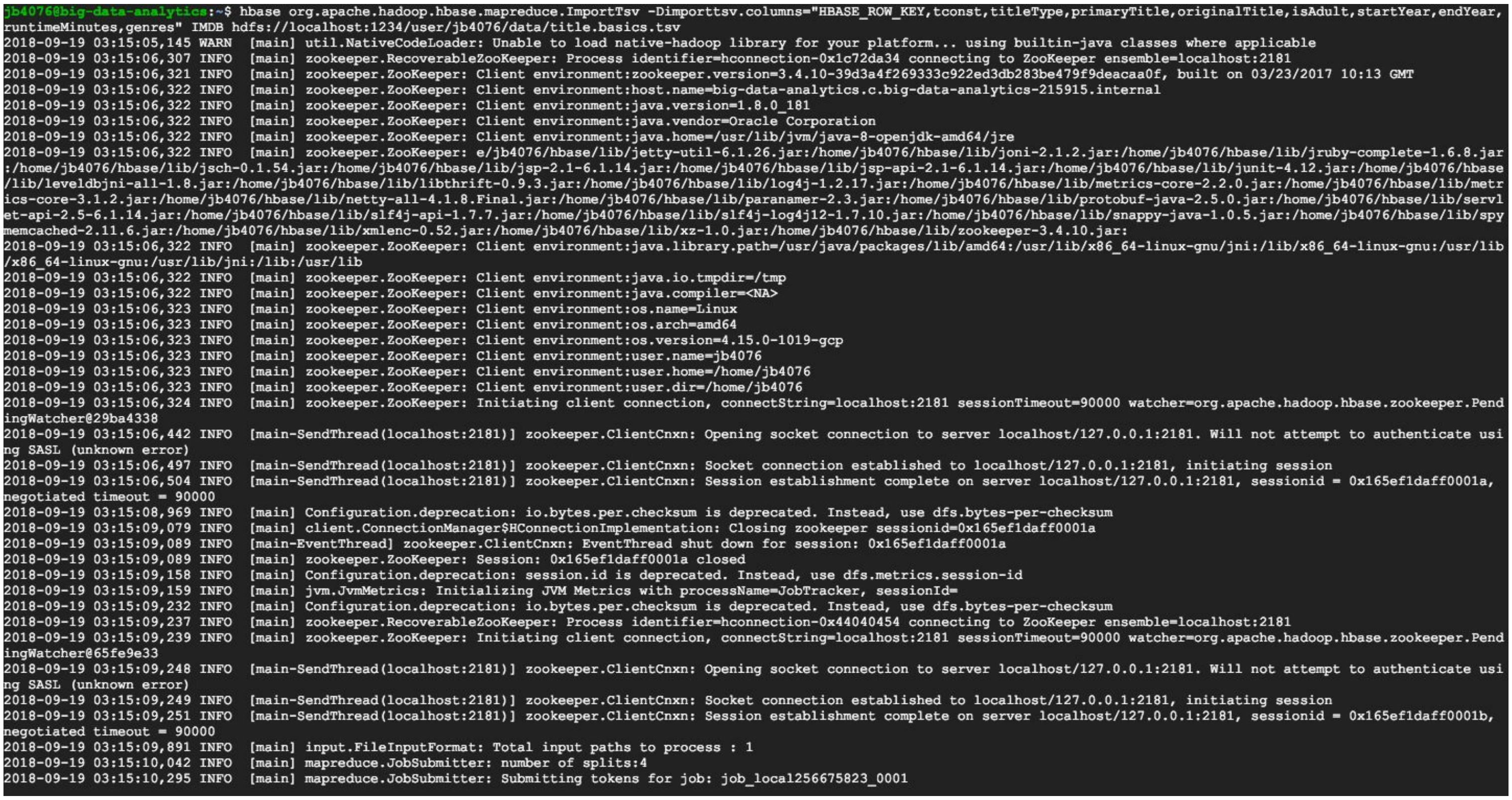
3. HBase

Using IMDB dataset (title.basics.tsv)

3.1 Import a table from an external file in HDFS



```
jb4076@big-data-analytics:~$ hbase shell
2018-09-19 03:05:15,757 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.7, r763f27f583cf8fd7ecf79fb6f3ef57f1615dbf9b, Tue Aug 28 14:40:11 PDT 2018

hbase(main):001:0> create 'IMDB','tconst','titleType','primaryTitle','originalTitle','isAdult','startYear','endYear','runtimeMinutes','genres'
0 row(s) in 2.1510 seconds

=> Hbase::Table - IMDB
hbase(main):002:0> exit
```

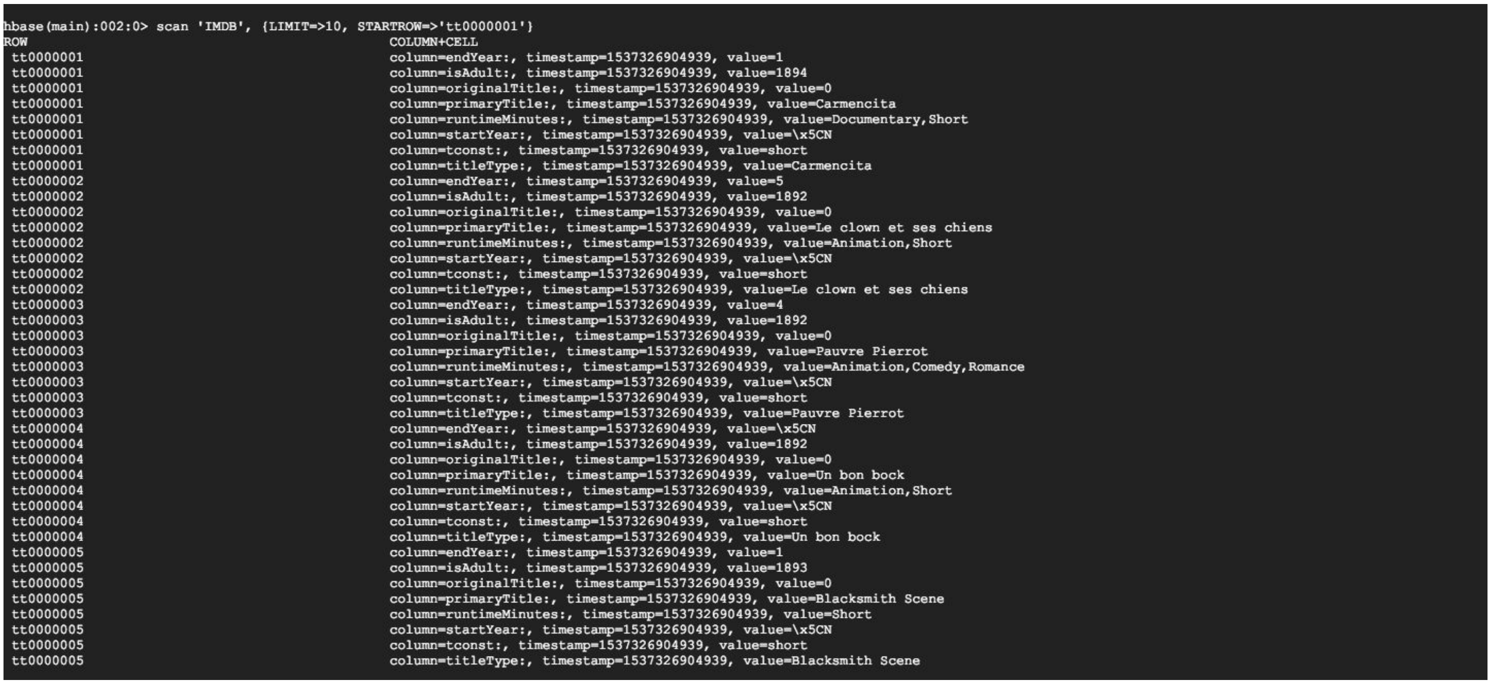


```
jb4076@big-data-analytics:~$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.columns="HBASE_ROW_KEY,tconst,titleType,primaryTitle,originalTitle,isAdult,startYear,endYear,
runtimeMinutes,genres" IMDB hdfs://localhost:1234/user/jb4076/data/title.basics.tsv
2018-09-19 03:15:05,145 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2018-09-19 03:15:06,307 INFO  [main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x1c72da34 connecting to ZooKeeper ensemble=localhost:2181
2018-09-19 03:15:06,321 INFO  [main] zookeeper.ZooKeeper: Client environment:zookeeper.version=3.4.10-39d3a4f269333c922ed3db283be479f9deacaa0f, built on 03/23/2017 10:13 GMT
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: Client environment:host.name=big-data-analytics.c.big-data-analytics-215915.internal
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: Client environment:java.version=1.8.0_181
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: Client environment:java.vendor=Oracle Corporation
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: Client environment:java.home=/usr/lib/jvm/java-8-openjdk-amd64/jre
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: e/jb4076/hbase/lib/jetty-util-6.1.26.jar:/home/jb4076/hbase/lib/joni-2.1.2.jar:/home/jb4076/hbase/lib/jruby-complete-1.6.8.jar
:/home/jb4076/hbase/lib/jsch-0.1.54.jar:/home/jb4076/hbase/lib/jsp-2.1-6.1.14.jar:/home/jb4076/hbase/lib/jsp-api-2.1-6.1.14.jar:/home/jb4076/hbase/lib/junit-4.12.jar:/home/jb4076/hbase
/lib/leveldbjni-all-1.8.jar:/home/jb4076/hbase/lib/libthrift-0.9.3.jar:/home/jb4076/hbase/lib/log4j-1.2.17.jar:/home/jb4076/hbase/lib/metrics-core-2.2.0.jar:/home/jb4076/hbase/lib/metr
ics-core-3.1.2.jar:/home/jb4076/hbase/lib/netty-all-4.1.8.Final.jar:/home/jb4076/hbase/lib/paranamer-2.3.jar:/home/jb4076/hbase/lib/protobuf-java-2.5.0.jar:/home/jb4076/hbase/lib/servl
et-api-2.5-6.1.14.jar:/home/jb4076/hbase/lib/slf4j-api-1.7.7.jar:/home/jb4076/hbase/lib/slf4j-log4j12-1.7.10.jar:/home/jb4076/hbase/lib/snappy-java-1.0.5.jar:/home/jb4076/hbase/lib/spy
memcached-2.11.6.jar:/home/jb4076/hbase/lib/xmlenc-0.52.jar:/home/jb4076/hbase/lib/xz-1.0.jar:/home/jb4076/hbase/lib/zookeeper-3.4.10.jar:
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: Client environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib/x86_64-linux-gnu/jni:/lib/x86_64-linux-gnu:/usr/lib
/x86_64-linux-gnu:/usr/lib/jni:/lib:/usr/lib
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: Client environment:java.io.tmpdir=/tmp
2018-09-19 03:15:06,322 INFO  [main] zookeeper.ZooKeeper: Client environment:java.compiler=<NA>
2018-09-19 03:15:06,323 INFO  [main] zookeeper.ZooKeeper: Client environment:os.name=Linux
2018-09-19 03:15:06,323 INFO  [main] zookeeper.ZooKeeper: Client environment:os.arch=amd64
2018-09-19 03:15:06,323 INFO  [main] zookeeper.ZooKeeper: Client environment:os.version=4.15.0-1019-gcp
2018-09-19 03:15:06,323 INFO  [main] zookeeper.ZooKeeper: Client environment:user.name=jb4076
2018-09-19 03:15:06,323 INFO  [main] zookeeper.ZooKeeper: Client environment:user.home=/home/jb4076
2018-09-19 03:15:06,323 INFO  [main] zookeeper.ZooKeeper: Client environment:user.dir=/home/jb4076
2018-09-19 03:15:06,324 INFO  [main] zookeeper.ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionTimeout=90000 watcher=org.apache.hadoop.hbase.zookeeper.Pend
ingWatcher@29ba4338
2018-09-19 03:15:06,442 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Opening socket connection to server localhost/127.0.0.1:2181. Will not attempt to authenticate usi
ng SASL (unknown error)
2018-09-19 03:15:06,497 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Socket connection established to localhost/127.0.0.1:2181, initiating session
2018-09-19 03:15:06,504 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Session establishment complete on server localhost/127.0.0.1:2181, sessionid = 0x165ef1daff0001a,
negotiated timeout = 90000
2018-09-19 03:15:08,969 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-09-19 03:15:09,079 INFO  [main] client.ConnectionManager$HConnectionImplementation: Closing zookeeper sessionid=0x165ef1daff0001a
2018-09-19 03:15:09,089 INFO  [main-EventThread] zookeeper.ClientCnxn: EventThread shut down for session: 0x165ef1daff0001a
2018-09-19 03:15:09,089 INFO  [main] zookeeper.ZooKeeper: Session: 0x165ef1daff0001a closed
2018-09-19 03:15:09,158 INFO  [main] Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
2018-09-19 03:15:09,159 INFO  [main] jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
2018-09-19 03:15:09,232 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-09-19 03:15:09,237 INFO  [main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x44040454 connecting to ZooKeeper ensemble=localhost:2181
2018-09-19 03:15:09,239 INFO  [main] zookeeper.ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionTimeout=90000 watcher=org.apache.hadoop.hbase.zookeeper.Pend
ingWatcher@65fe9e33
2018-09-19 03:15:09,248 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Opening socket connection to server localhost/127.0.0.1:2181. Will not attempt to authenticate usi
ng SASL (unknown error)
2018-09-19 03:15:09,249 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Socket connection established to localhost/127.0.0.1:2181, initiating session
2018-09-19 03:15:09,251 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Session establishment complete on server localhost/127.0.0.1:2181, sessionid = 0x165ef1daff0001b,
negotiated timeout = 90000
2018-09-19 03:15:09,891 INFO  [main] input.FileInputFormat: Total input paths to process : 1
2018-09-19 03:15:10,042 INFO  [main] mapreduce.JobSubmitter: number of splits:4
2018-09-19 03:15:10,295 INFO  [main] mapreduce.JobSubmitter: Submitting tokens for job: job_local256675823_0001
```

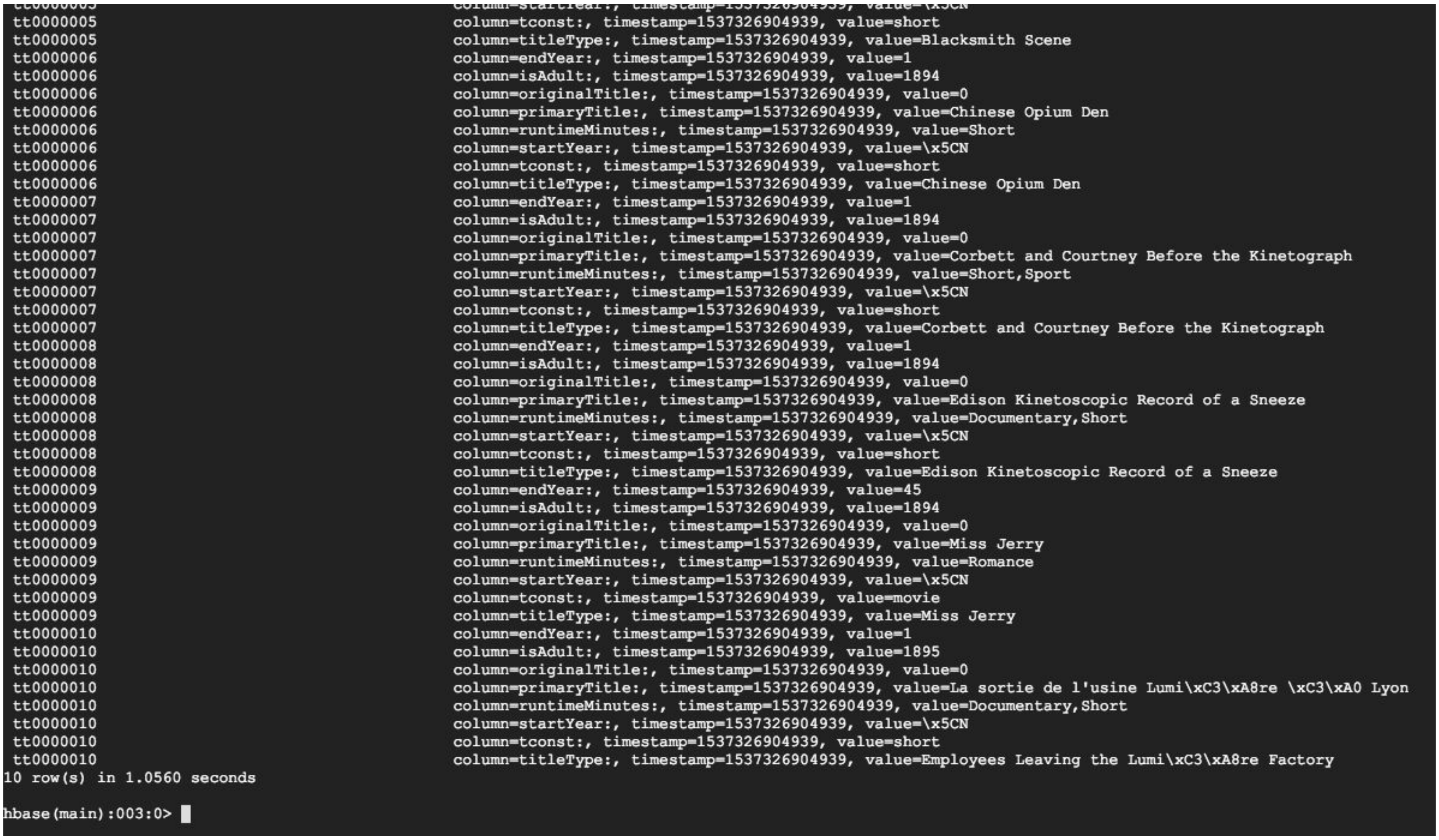
script:

```
jb4076@big-data-analytics:~$ hbase shell
hbase(main):001:0> list
hbase(main):002:0> create
'IMDB','tconst','titleType','primaryTitle','originalTitle','isAdult','startYear','endYear','
runtimeMinutes','genres'
hbase(main):003:0> exit

hbase org.apache.hadoop.hbase.mapreduce.ImportTsv
-Dimporttsv.columns="HBASE_ROW_KEY,tconst,titleType,primaryTitle,originalTitle,isAdult,start
Year,endYear,
runtimeMinutes,genres" IMDB hdfs://localhost:1234/user/jb4076/data/title.basics.tsv
```

3.2 Display the top 10 rows of content of the table



```
hbase(main):002:0> scan 'IMDB', {LIMIT=>10, STARTROW=>'tt0000001'}
ROW                                          COLUMN+CELL
 tt0000001                                   column=endYear:, timestamp=1537326904939, value=1
 tt0000001                                   column=isAdult:, timestamp=1537326904939, value=1894
 tt0000001                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000001                                   column=primaryTitle:, timestamp=1537326904939, value=Carmencita
 tt0000001                                   column=runtimeMinutes:, timestamp=1537326904939, value=Documentary,Short
 tt0000001                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000001                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000001                                   column=titleType:, timestamp=1537326904939, value=Carmencita
 tt0000002                                   column=endYear:, timestamp=1537326904939, value=5
 tt0000002                                   column=isAdult:, timestamp=1537326904939, value=1892
 tt0000002                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000002                                   column=primaryTitle:, timestamp=1537326904939, value=Le clown et ses chiens
 tt0000002                                   column=runtimeMinutes:, timestamp=1537326904939, value=Animation,Short
 tt0000002                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000002                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000002                                   column=titleType:, timestamp=1537326904939, value=Le clown et ses chiens
 tt0000003                                   column=endYear:, timestamp=1537326904939, value=4
 tt0000003                                   column=isAdult:, timestamp=1537326904939, value=1892
 tt0000003                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000003                                   column=primaryTitle:, timestamp=1537326904939, value=Pauvre Pierrot
 tt0000003                                   column=runtimeMinutes:, timestamp=1537326904939, value=Animation,Comedy,Romance
 tt0000003                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000003                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000003                                   column=titleType:, timestamp=1537326904939, value=Pauvre Pierrot
 tt0000004                                   column=endYear:, timestamp=1537326904939, value=\x5CN
 tt0000004                                   column=isAdult:, timestamp=1537326904939, value=1892
 tt0000004                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000004                                   column=primaryTitle:, timestamp=1537326904939, value=Un bon bock
 tt0000004                                   column=runtimeMinutes:, timestamp=1537326904939, value=Animation,Short
 tt0000004                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000004                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000004                                   column=titleType:, timestamp=1537326904939, value=Un bon bock
 tt0000005                                   column=endYear:, timestamp=1537326904939, value=1
 tt0000005                                   column=isAdult:, timestamp=1537326904939, value=1893
 tt0000005                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000005                                   column=primaryTitle:, timestamp=1537326904939, value=Blacksmith Scene
 tt0000005                                   column=runtimeMinutes:, timestamp=1537326904939, value=Short
 tt0000005                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000005                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000005                                   column=titleType:, timestamp=1537326904939, value=Blacksmith Scene
```

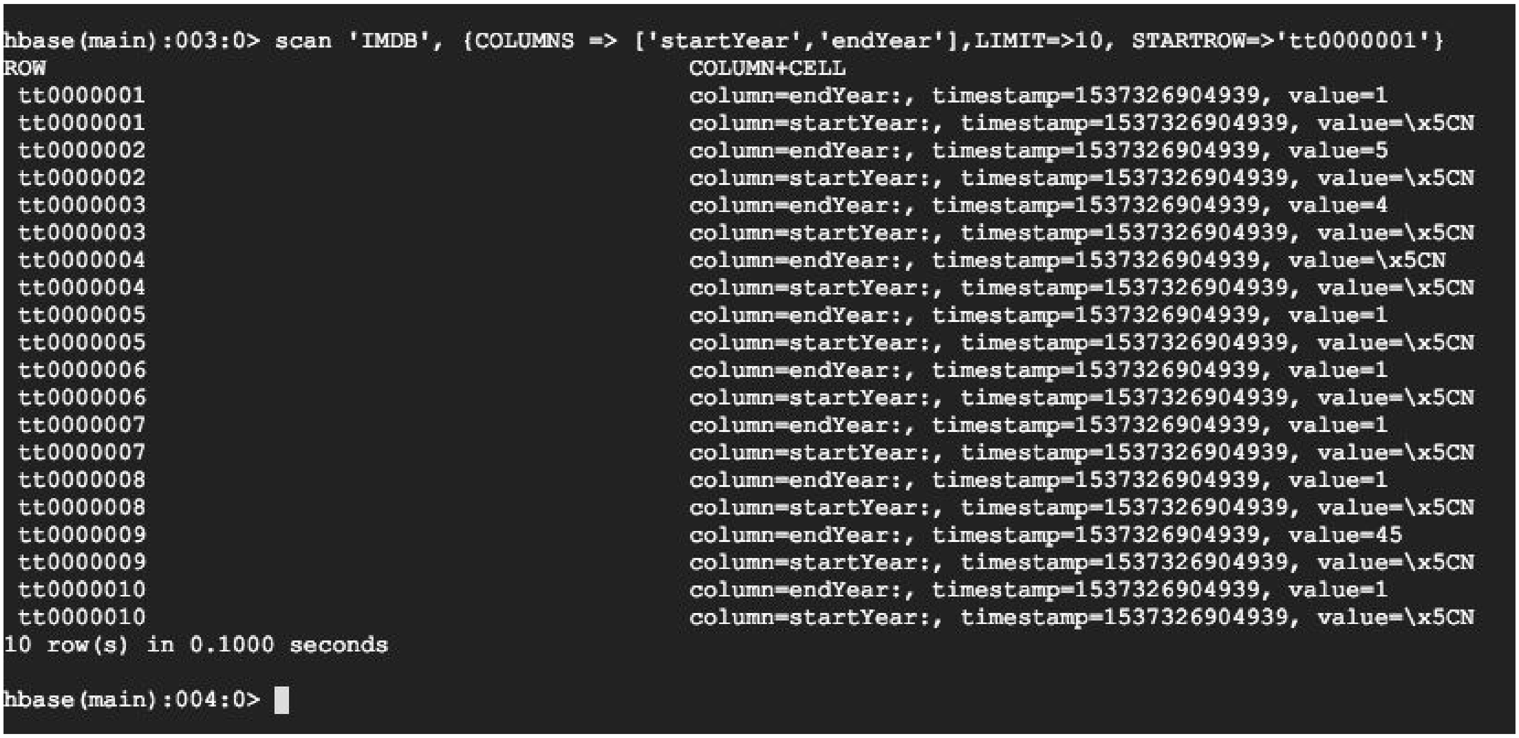


```
 tt0000005                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000005                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000005                                   column=titleType:, timestamp=1537326904939, value=Blacksmith Scene
 tt0000006                                   column=endYear:, timestamp=1537326904939, value=1
 tt0000006                                   column=isAdult:, timestamp=1537326904939, value=1894
 tt0000006                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000006                                   column=primaryTitle:, timestamp=1537326904939, value=Chinese Opium Den
 tt0000006                                   column=runtimeMinutes:, timestamp=1537326904939, value=Short
 tt0000006                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000006                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000006                                   column=titleType:, timestamp=1537326904939, value=Chinese Opium Den
 tt0000007                                   column=endYear:, timestamp=1537326904939, value=1
 tt0000007                                   column=isAdult:, timestamp=1537326904939, value=1894
 tt0000007                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000007                                   column=primaryTitle:, timestamp=1537326904939, value=Corbett and Courtney Before the Kinetograph
 tt0000007                                   column=runtimeMinutes:, timestamp=1537326904939, value=Short,Sport
 tt0000007                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000007                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000007                                   column=titleType:, timestamp=1537326904939, value=Corbett and Courtney Before the Kinetograph
 tt0000008                                   column=endYear:, timestamp=1537326904939, value=1
 tt0000008                                   column=isAdult:, timestamp=1537326904939, value=1894
 tt0000008                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000008                                   column=primaryTitle:, timestamp=1537326904939, value=Edison Kinetoscopic Record of a Sneeze
 tt0000008                                   column=runtimeMinutes:, timestamp=1537326904939, value=Documentary,Short
 tt0000008                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000008                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000008                                   column=titleType:, timestamp=1537326904939, value=Edison Kinetoscopic Record of a Sneeze
 tt0000009                                   column=endYear:, timestamp=1537326904939, value=45
 tt0000009                                   column=isAdult:, timestamp=1537326904939, value=1894
 tt0000009                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000009                                   column=primaryTitle:, timestamp=1537326904939, value=Miss Jerry
 tt0000009                                   column=runtimeMinutes:, timestamp=1537326904939, value=Romance
 tt0000009                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000009                                   column=tconst:, timestamp=1537326904939, value=movie
 tt0000009                                   column=titleType:, timestamp=1537326904939, value=Miss Jerry
 tt0000010                                   column=endYear:, timestamp=1537326904939, value=1
 tt0000010                                   column=isAdult:, timestamp=1537326904939, value=1895
 tt0000010                                   column=originalTitle:, timestamp=1537326904939, value=0
 tt0000010                                   column=primaryTitle:, timestamp=1537326904939, value=La sortie de l'usine Lumi\xC3\xA8re \xC3\xA0 Lyon
 tt0000010                                   column=runtimeMinutes:, timestamp=1537326904939, value=Documentary,Short
 tt0000010                                   column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000010                                   column=tconst:, timestamp=1537326904939, value=short
 tt0000010                                   column=titleType:, timestamp=1537326904939, value=Employees Leaving the Lumi\xC3\xA8re Factory
10 row(s) in 1.0560 seconds

hbase(main):003:0> █
```

script:

```
hbase(main):002:0> scan 'IMDB', {LIMIT=>10, STARTROW=>'tt0000001'}
```

3.3 Display the top 10 rows of content with some specific values



```
hbase(main):003:0> scan 'IMDB', {COLUMNS => ['startYear','endYear'],LIMIT=>10, STARTROW=>'tt0000001'}
ROW                                               COLUMN+CELL
 tt0000001                                        column=endYear:, timestamp=1537326904939, value=1
 tt0000001                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000002                                        column=endYear:, timestamp=1537326904939, value=5
 tt0000002                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000003                                        column=endYear:, timestamp=1537326904939, value=4
 tt0000003                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000004                                        column=endYear:, timestamp=1537326904939, value=\x5CN
 tt0000004                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000005                                        column=endYear:, timestamp=1537326904939, value=1
 tt0000005                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000006                                        column=endYear:, timestamp=1537326904939, value=1
 tt0000006                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000007                                        column=endYear:, timestamp=1537326904939, value=1
 tt0000007                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000008                                        column=endYear:, timestamp=1537326904939, value=1
 tt0000008                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000009                                        column=endYear:, timestamp=1537326904939, value=45
 tt0000009                                        column=startYear:, timestamp=1537326904939, value=\x5CN
 tt0000010                                        column=endYear:, timestamp=1537326904939, value=1
 tt0000010                                        column=startYear:, timestamp=1537326904939, value=\x5CN
10 row(s) in 0.1000 seconds

hbase(main):004:0>
```

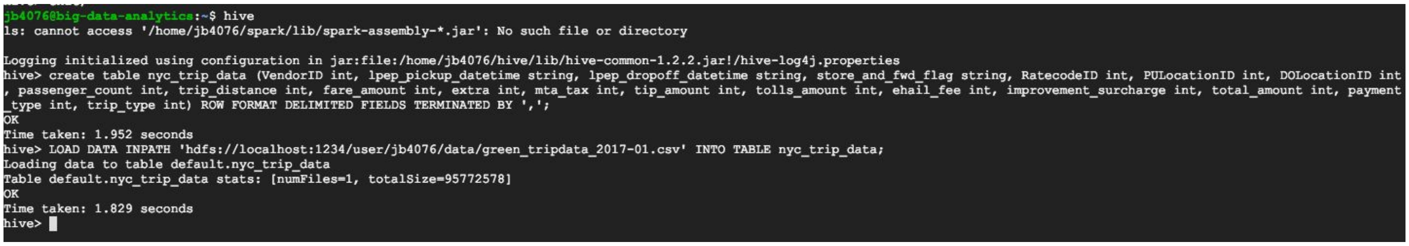
script:

```
hbase(main):003:0> scan 'IMDB', {COLUMNS => ['startYear','endYear'], LIMIT=>10,
STARTROW=>'tt0000001'}
```

4. Hive

Using NYC TLC Trip Data (2017 January Green Taxi) (green_tripdata_2017-01.csv)

4.1 Import a table from an external file in HDFS

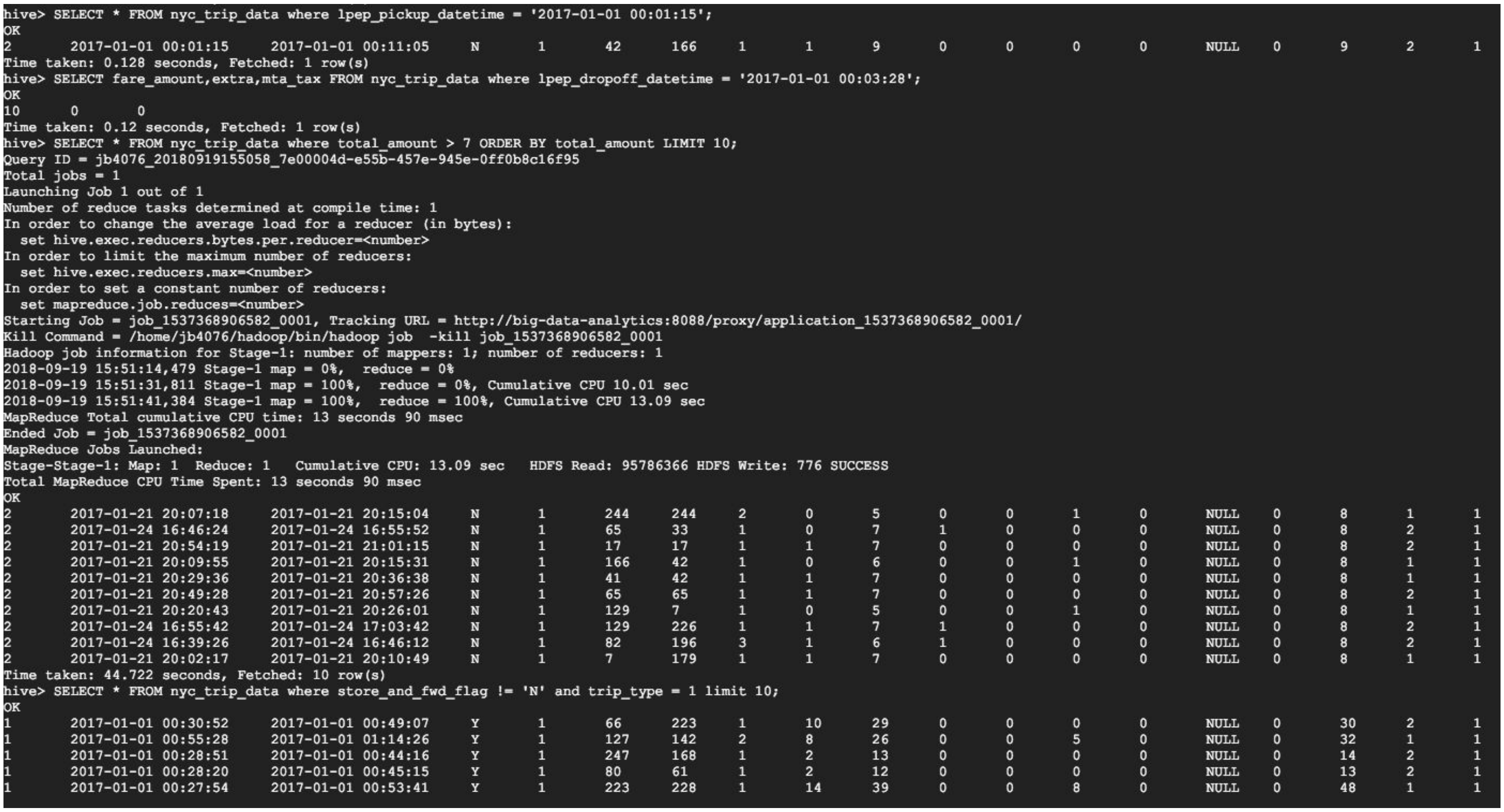


```
jb4076@big-data-analytics:~$ hive
ls: cannot access '/home/jb4076/spark/lib/spark-assembly-*.jar': No such file or directory

Logging initialized using configuration in jar:file:/home/jb4076/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties
hive> create table nyc_trip_data (VendorID int, lpep_pickup_datetime string, lpep_dropoff_datetime string, store_and_fwd_flag string, RatecodeID int, PULocationID int, DOLocationID int
, passenger_count int, trip_distance int, fare_amount int, extra int, mta_tax int, tip_amount int, tolls_amount int, ehail_fee int, improvement_surcharge int, total_amount int, payment
_type int, trip_type int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 1.952 seconds
hive> LOAD DATA INPATH 'hdfs://localhost:1234/user/jb4076/data/green_tripdata_2017-01.csv' INTO TABLE nyc_trip_data;
Loading data to table default.nyc_trip_data
Table default.nyc_trip_data stats: [numFiles=1, totalSize=95772578]
OK
Time taken: 1.829 seconds
hive>
```

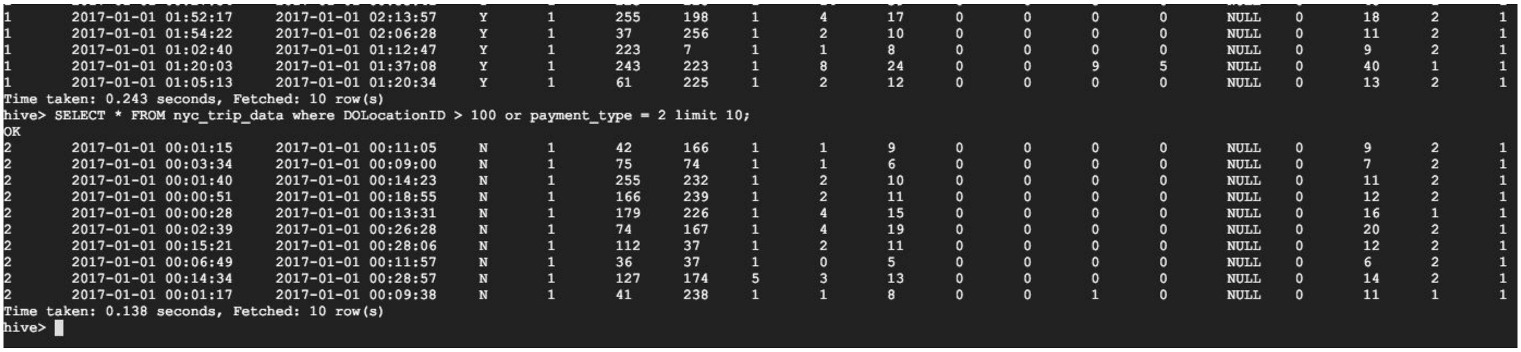
script:

```
jb4076@big-data-analytics:~$ hive
hive> create table nyc_trip_data (VendorID int, lpep_pickup_datetime string,
lpep_dropoff_datetime string, store_and_fwd_flag string, RatecodeID int,
PULocationID int, DOLocationID int, passenger_count int, trip_distance int,
fare_amount int, extra int, mta_tax int, tip_amount int, tolls_amount int, ehail_fee
int, improvement_surcharge int, total_amount int, payment_type int, trip_type int)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
hive> LOAD DATA INPATH
'hdfs://localhost:1234/user/jb4076/data/green_tripdata_2017-01.csv' INTO TABLE
nyc_trip_data;
```

4.2 Do five queries and show the results



```
hive> SELECT * FROM nyc_trip_data where lpep_pickup_datetime = '2017-01-01 00:01:15';
OK
2       2017-01-01 00:01:15     2017-01-01 00:11:05     N       1       42      166     1       1       9       0       0       0       0       NULL    0       9       2       1
Time taken: 0.128 seconds, Fetched: 1 row(s)
hive> SELECT fare_amount,extra,mta_tax FROM nyc_trip_data where lpep_dropoff_datetime = '2017-01-01 00:03:28';
OK
10      0       0
Time taken: 0.12 seconds, Fetched: 1 row(s)
hive> SELECT * FROM nyc_trip_data where total_amount > 7 ORDER BY total_amount LIMIT 10;
Query ID = jb4076_20180919155058_7e00004d-e55b-457e-945e-0ff0b8c16f95
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_1537368906582_0001, Tracking URL = http://big-data-analytics:8088/proxy/application_1537368906582_0001/
Kill Command = /home/jb4076/hadoop/bin/hadoop job  -kill job_1537368906582_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-19 15:51:14,479 Stage-1 map = 0%,  reduce = 0%
2018-09-19 15:51:31,811 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.01 sec
2018-09-19 15:51:41,384 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.09 sec
MapReduce Total cumulative CPU time: 13 seconds 90 msec
Ended Job = job_1537368906582_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 13.09 sec   HDFS Read: 95786366 HDFS Write: 776 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 90 msec
OK
2       2017-01-21 20:07:18     2017-01-21 20:15:04     N       1       244     244     2       0       5       0       0       1       0       NULL    0       8       1       1
2       2017-01-24 16:46:24     2017-01-24 16:55:52     N       1       65      33      1       0       7       1       0       0       0       NULL    0       8       2       1
2       2017-01-21 20:54:19     2017-01-21 21:01:15     N       1       17      17      1       1       7       0       0       0       0       NULL    0       8       2       1
2       2017-01-21 20:09:55     2017-01-21 20:15:31     N       1       166     42      1       0       6       0       0       1       0       NULL    0       8       1       1
2       2017-01-21 20:29:36     2017-01-21 20:36:38     N       1       41      42      1       1       7       0       0       0       0       NULL    0       8       1       1
2       2017-01-21 20:49:28     2017-01-21 20:57:26     N       1       65      65      1       1       7       0       0       0       0       NULL    0       8       2       1
2       2017-01-21 20:20:43     2017-01-21 20:26:01     N       1       129     7       1       0       5       0       0       1       0       NULL    0       8       1       1
2       2017-01-24 16:55:42     2017-01-24 17:03:42     N       1       129     226     1       1       7       1       0       0       0       NULL    0       8       2       1
2       2017-01-24 16:39:26     2017-01-24 16:46:12     N       1       82      196     3       1       6       1       0       0       0       NULL    0       8       2       1
2       2017-01-21 20:02:17     2017-01-21 20:10:49     N       1       7       179     1       1       7       0       0       0       0       NULL    0       8       1       1
Time taken: 44.722 seconds, Fetched: 10 row(s)
hive> SELECT * FROM nyc_trip_data where store_and_fwd_flag != 'N' and trip_type = 1 limit 10;
OK
1       2017-01-01 00:30:52     2017-01-01 00:49:07     Y       1       66      223     1       10      29      0       0       0       0       NULL    0       30      2       1
1       2017-01-01 00:55:28     2017-01-01 01:14:26     Y       1       127     142     2       8       26      0       0       5       0       NULL    0       32      1       1
1       2017-01-01 00:28:51     2017-01-01 00:44:16     Y       1       247     168     1       2       13      0       0       0       0       NULL    0       14      2       1
1       2017-01-01 00:28:20     2017-01-01 00:45:15     Y       1       80      61      1       2       12      0       0       0       0       NULL    0       13      2       1
1       2017-01-01 00:27:54     2017-01-01 00:53:41     Y       1       223     228     1       14      39      0       0       8       0       NULL    0       48      1       1
```



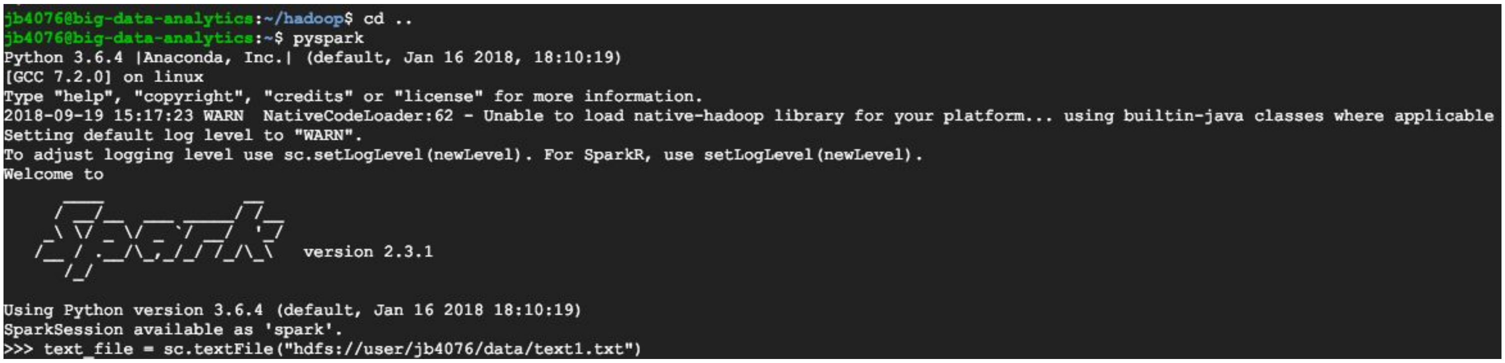
```
1       2017-01-01 01:52:17     2017-01-01 02:13:57     Y       1       255     198     1       4       17      0       0       0       0       NULL    0       18      2       1
1       2017-01-01 01:54:22     2017-01-01 02:06:28     Y       1       37      256     1       2       10      0       0       0       0       NULL    0       11      2       1
1       2017-01-01 01:02:40     2017-01-01 01:12:47     Y       1       223     7       1       1       8       0       0       0       0       NULL    0       9       2       1
1       2017-01-01 01:20:03     2017-01-01 01:37:08     Y       1       243     223     1       8       24      0       0       9       5       NULL    0       40      1       1
1       2017-01-01 01:05:13     2017-01-01 01:20:34     Y       1       61      225     1       2       12      0       0       0       0       NULL    0       13      2       1
Time taken: 0.243 seconds, Fetched: 10 row(s)
hive> SELECT * FROM nyc_trip_data where DOLocationID > 100 or payment_type = 2 limit 10;
OK
2       2017-01-01 00:01:15     2017-01-01 00:11:05     N       1       42      166     1       1       9       0       0       0       0       NULL    0       9       2       1
2       2017-01-01 00:03:34     2017-01-01 00:09:00     N       1       75      74      1       1       6       0       0       0       0       NULL    0       7       2       1
2       2017-01-01 00:01:40     2017-01-01 00:14:23     N       1       255     232     1       2       10      0       0       0       0       NULL    0       11      2       1
2       2017-01-01 00:00:51     2017-01-01 00:18:55     N       1       166     239     1       2       11      0       0       0       0       NULL    0       12      2       1
2       2017-01-01 00:00:28     2017-01-01 00:13:31     N       1       179     226     1       4       15      0       0       0       0       NULL    0       16      1       1
2       2017-01-01 00:02:39     2017-01-01 00:26:28     N       1       74      167     1       4       19      0       0       0       0       NULL    0       20      2       1
2       2017-01-01 00:15:21     2017-01-01 00:28:06     N       1       112     37      1       2       11      0       0       0       0       NULL    0       12      2       1
2       2017-01-01 00:06:49     2017-01-01 00:11:57     N       1       36      37      1       0       5       0       0       0       0       NULL    0       6       2       1
2       2017-01-01 00:14:34     2017-01-01 00:28:57     N       1       127     174     5       3       13      0       0       0       0       NULL    0       14      2       1
2       2017-01-01 00:01:17     2017-01-01 00:09:38     N       1       41      238     1       1       8       0       0       1       0       NULL    0       11      1       1
Time taken: 0.138 seconds, Fetched: 10 row(s)
hive> █
```

script:

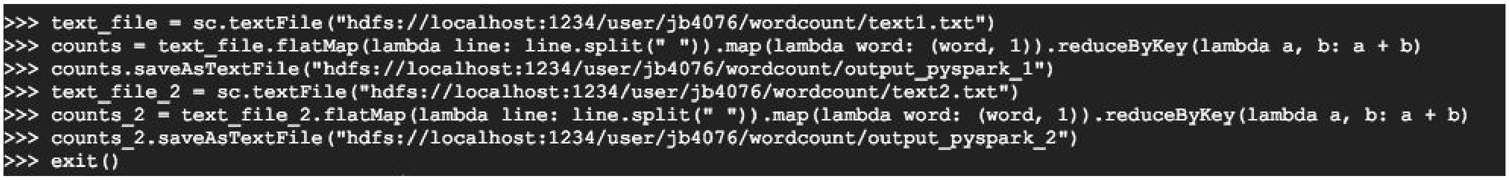
```
hive> SELECT * FROM nyc_trip_data where lpep_pickup_datetime = '2017-01-01 00:01:15';
hive> SELECT fare_amount,extra,mta_tax FROM nyc_trip_data where lpep_dropoff_datetime = '2017-01-01 00:03:28';
hive> SELECT * FROM nyc_trip_data where total_amount > 7 ORDER BY total_amount LIMIT 10;
hive> SELECT * FROM nyc_trip_data where store_and_fwd_flag != 'N' and trip_type = 1 limit 10;
hive> SELECT * FROM nyc_trip_data where DOLocationID > 100 or payment_type = 2 limit 10;
```

5. Spark

5.1 Run the Word Count program with your chosen programming language



```
jb4076@big-data-analytics:~/hadoop$ cd ..
jb4076@big-data-analytics:~$ pyspark
Python 3.6.4 |Anaconda, Inc.| (default, Jan 16 2018, 18:10:19)
[GCC 7.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
2018-09-19 15:17:23 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.3.1
      /_/

Using Python version 3.6.4 (default, Jan 16 2018 18:10:19)
SparkSession available as 'spark'.
>>> text_file = sc.textFile("hdfs://user/jb4076/data/text1.txt")
```



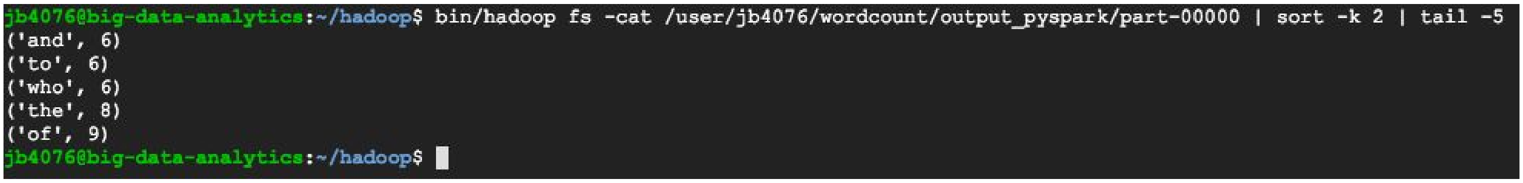
```
>>> text_file = sc.textFile("hdfs://localhost:1234/user/jb4076/wordcount/text1.txt")
>>> counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> counts.saveAsTextFile("hdfs://localhost:1234/user/jb4076/wordcount/output_pyspark_1")
>>> text_file_2 = sc.textFile("hdfs://localhost:1234/user/jb4076/wordcount/text2.txt")
>>> counts_2 = text_file_2.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> counts_2.saveAsTextFile("hdfs://localhost:1234/user/jb4076/wordcount/output_pyspark_2")
>>> exit()
```

script:

```
jb4076@big-data-analytics:~$ gsutil cp gs://6689bigdata/text.txt ./hadoop/data/
jb4076@big-data-analytics:~/hadoop$ ./bin/hdfs dfs -put /home/jb4076/hadoop/data/text.txt /user/jb4076/wordcount/text.txt
jb4076@big-data-analytics:~$ pyspark
```

```
>>> text_file = sc.textFile("hdfs://localhost:1234/user/jb4076/wordcount/text.txt")
>>> counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word,
1)).reduceByKey(lambda a, b: a + b)
>>>
counts.saveAsTextFile("hdfs://localhost:1234/user/jb4076/wordcount/output_pyspark")
>>> exit()
```

5.2 On the provided text, list the top 3 most frequent words



```
jb4076@big-data-analytics:~/hadoop$ bin/hadoop fs -cat /user/jb4076/wordcount/output_pyspark/part-00000 | sort -k 2 | tail -5
('and', 6)
('to', 6)
('who', 6)
('the', 8)
('of', 9)
jb4076@big-data-analytics:~/hadoop$ 
```

script:

```
jb4076@big-data-analytics:~/hadoop$ bin/hadoop fs -cat
/user/jb4076/wordcount/output_pyspark/part-00000 | sort -k 2 | tail -5
```

result:

The 5 most frequent words of text.txt are “of”, “the”, “who”, “to” and “and”.