

## Homework\_1

Name: Rahul Purushottam Gaonkar(rpg283)

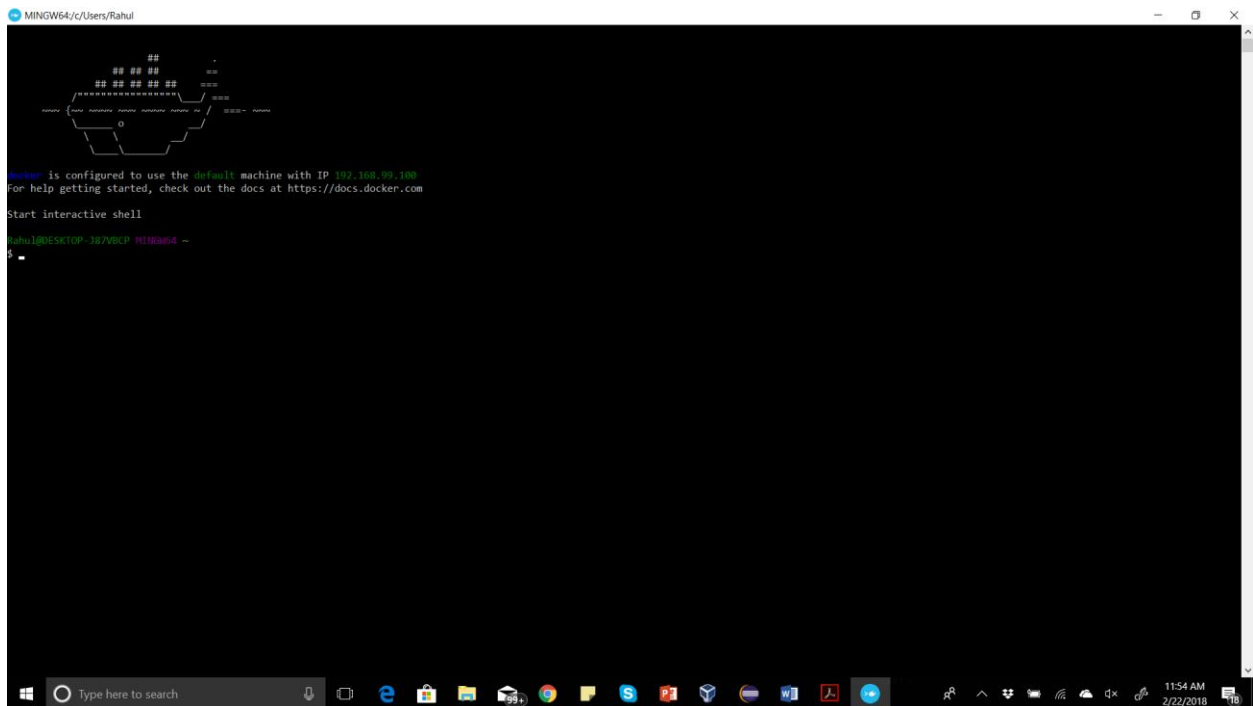


Figure1: Docker Running

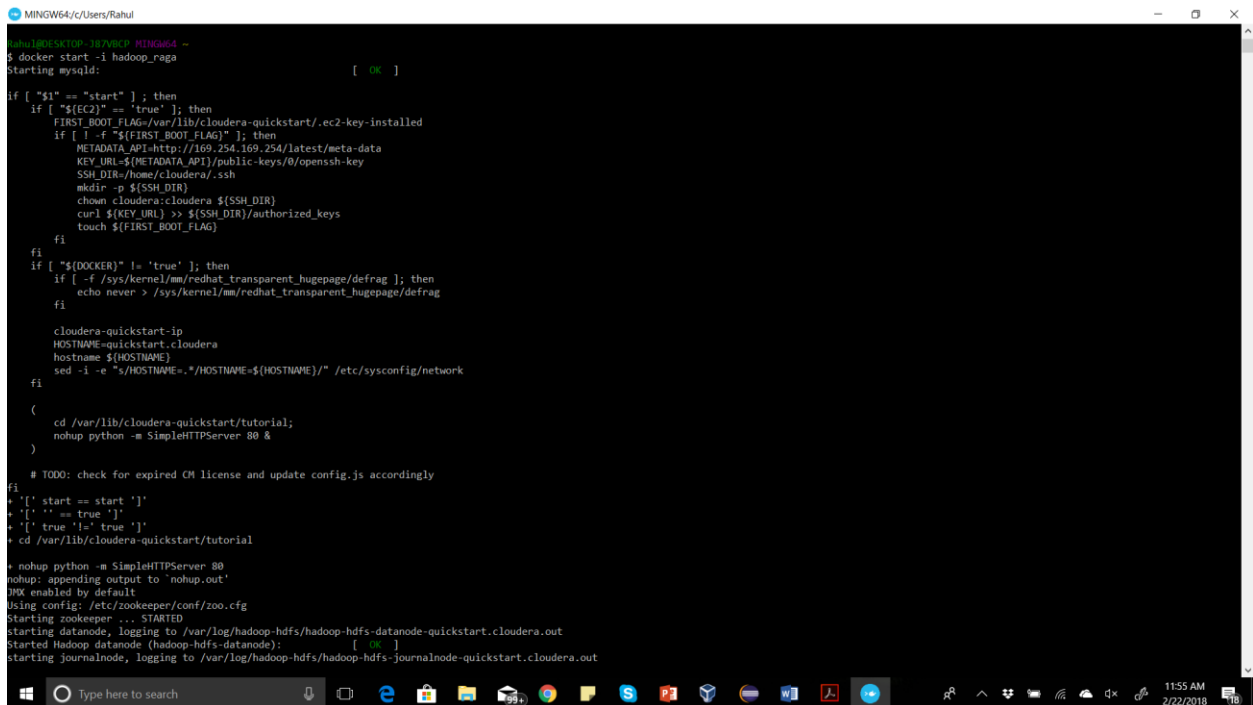


Figure 2: Starting Hadoop Container

```
@quickstart/
setting CATALINA_BASE=/var/lib/oozie/tomcat-deployment
setting OOOIE_HTTPS_PORT=11443
setting OOOIE_HTTPS_KEYSTORE_PASS=password
setting CATALINA_OPTS="$CATALINA_OPTS -Doozie.https.port=${OOOIE_HTTPS_PORT}"
setting CATALINA_OPTS="$CATALINA_OPTS -Doozie.https.keystore.pass=${OOOIE_HTTPS_KEYSTORE_PASS}"
setting OOOIE_CONF_DIR=/etc/oozie/conf
setting OOOIE_LOG=/var/log/oozie
Setting OOOIE_CONFIG_FILE: oooie-site.xml
Using OOOIE_DATA: /var/lib/oozie
Using OOOIE_LOG: /var/log/oozie
Setting OOOIE_LOG4J_FILE: oooie-log4j.properties
Setting OOOIE_LOG4J_RELOAD: 10
Setting OOOIE_HTTP_HOSTNAME: quickstart.cloudera
Setting OOOIE_HTTP_PORT: 11000
Setting OOOIE_ADMIN_PORT: 11001
Using OOOIE_HTTPS_PORT: 11443
Setting OOOIE_BASE_URL: http://quickstart.cloudera:11000/oozie
Using CATALINA_BASE: /var/lib/oozie/tomcat-deployment
Setting OOOIE_HTTPS_KEYSTORE_FILE: /var/lib/oozie/.keystore
Using OOOIE_HTTPS_KEYSTORE_PASS: password
Setting OOOIE_INSTANCE_ID: quickstart.cloudera
Setting CATALINA_OUT: /var/log/oozie/catalina.out
Using CATALINA_PID: /var/run/oozie/oozie.pid
Using CATALINA_OPTS: -Doozie.https.port=11443 -Doozie.https.keystore.pass=password -Xmx1024m -Doozie.https.port=11443 -Doozie.https.keystore.pass=password -Xmx1024m -Dderby.stream.error.file=/var/log/oozie/derby.log
Adding to CATALINA_OPTS: -Doozie.home.dir=/usr/lib/oozie -Doozie.config.dir=/etc/oozie/conf -Doozie.log.dir=/var/log/oozie -Doozie.data.dir=/var/lib/oozie -Doozie.instance.id=quickstart.cloudera -Doozie.config.file=oozie-site.xml -Doozie.log4j.file=oozie-log4j.properties -Doozie.log4j.reload=10 -Doozie.http.hostname=quickstart.cloudera -Doozie.admin.port=11001 -Doozie.http.port=11000 -Doozie.https.port=11443 -Doozie.base.url=http://quickstart.cloudera:11000/oozie -Doozie.https.keystore.file=/var/lib/oozie/.keystore -Doozie.https.keystore.pass=password -Djava.library.path=/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/native
Using CATALINA_BASE: /var/lib/oozie/tomcat-deployment
Using CATALINA_HOME: /usr/lib/bigtop-tomcat
Using CATALINA_TMPDIR: /var/lib/oozie
Using JRE_HOME: /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH: /usr/lib/bigtop-tomcat/bin/bootstrap.jar
Using CATALINA_PID: /var/run/oozie/oozie.pid
Starting Solr server daemon: [ OK ]
Using CATALINA_BASE: /var/lib/solr/tomcat-deployment
Using CATALINA_HOME: /usr/lib/solr/.bigtop-tomcat
Using CATALINA_TMPDIR: /var/lib/solr/
Using JRE_HOME: /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH: /usr/lib/solr/.bigtop-tomcat/bin/bootstrap.jar
Using CATALINA_PID: /var/run/solr/solr.pid
Existing PID file found during start.
Removing/clearing stale PID file.
Started Impala Catalog Server (catalogd): [ OK ]
Started Impala Server (impalad): [ OK ]
[root@quickstart /]#
```

Figure 3: Hadoop Container Started

```
@quickstart/
[root@quickstart /]# hadoop fs -ls
[root@quickstart /]# hadoop fs -mkdir rahul
[root@quickstart /]# hadoop fs -ls
Found 1 items
drwxr-xr-x - root supergroup 0 2018-02-22 16:59 rahul
[root@quickstart /]#
```

Figure 4: Created a directory of name Rahul in Hadoop filesystem

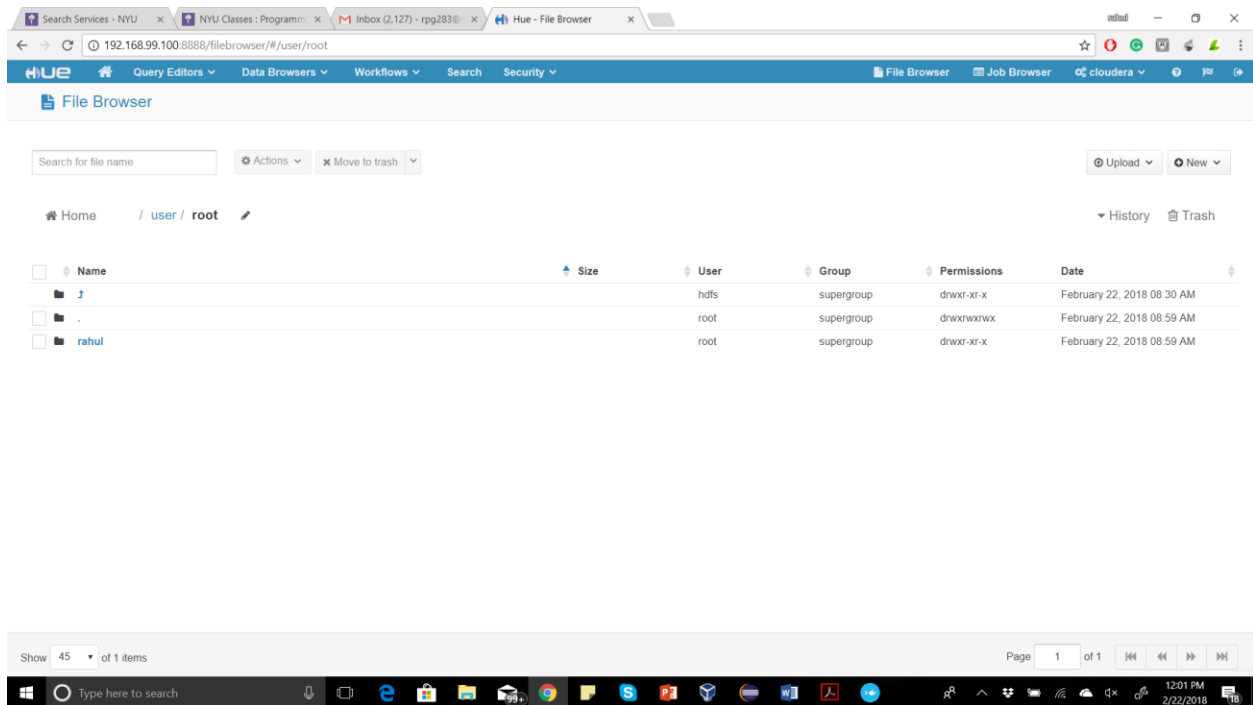


Figure 5: Directory name Rahul reflected in Hue

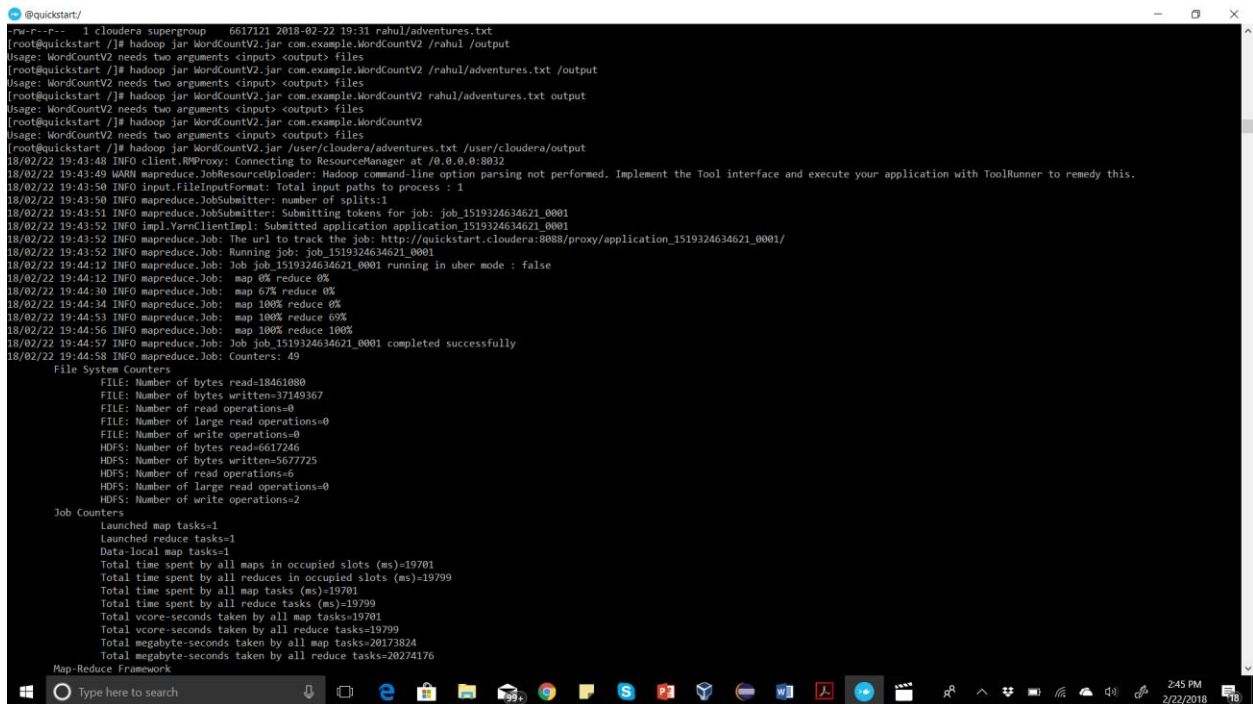


Figure 6: Running the MapReduce program(Job) of WordCountV2 to get the Phrase Count (i.e. Word Pair Count)

Command Used: `hadoop jar WordCountV2.jar /user/cloudera/adventures.txt /user/cloudera/output`

Placed the input file in user/cloudera and set the output path as /user/cloudera/output.

```
@quickstart/
HDFS: Number of bytes written=5677725
HDFS: Number of read operations=0
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)=19701
  Total time spent by all reduces in occupied slots (ms)=19799
  Total time spent by all map tasks (ms)=19701
  Total time spent by all reduce tasks (ms)=19799
  Total vcore-seconds taken by all map tasks=19701
  Total vcore-seconds taken by all reduce tasks=19799
  Total megabyte-seconds taken by all map tasks=20173824
  Total megabyte-seconds taken by all reduce tasks=20274176
Map-Reduce Framework
  Map input records=128457
  Map output records=1115031
  Map output bytes=16231012
  Map output materialized bytes=18461080
  Input split bytes=125
  Combine input records=0
  Combine output records=0
  Reduce input groups=384423
  Reduce shuffle bytes=18461080
  Reduce input records=1115031
  Reduce output records=384423
  Spilled Records=2230062
  Shuffled Maps=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=483
  CPU time spent (ms)=17060
  Physical memory (bytes) snapshot=400404480
  Virtual memory (bytes) snapshot=261183616
  Total committed heap usage (bytes)=325783552
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=6617121
File Output Format Counters
  Bytes Written=5677725
Job was successful
[root@quickstart ~]#
```

Figure 7: Job was Successful

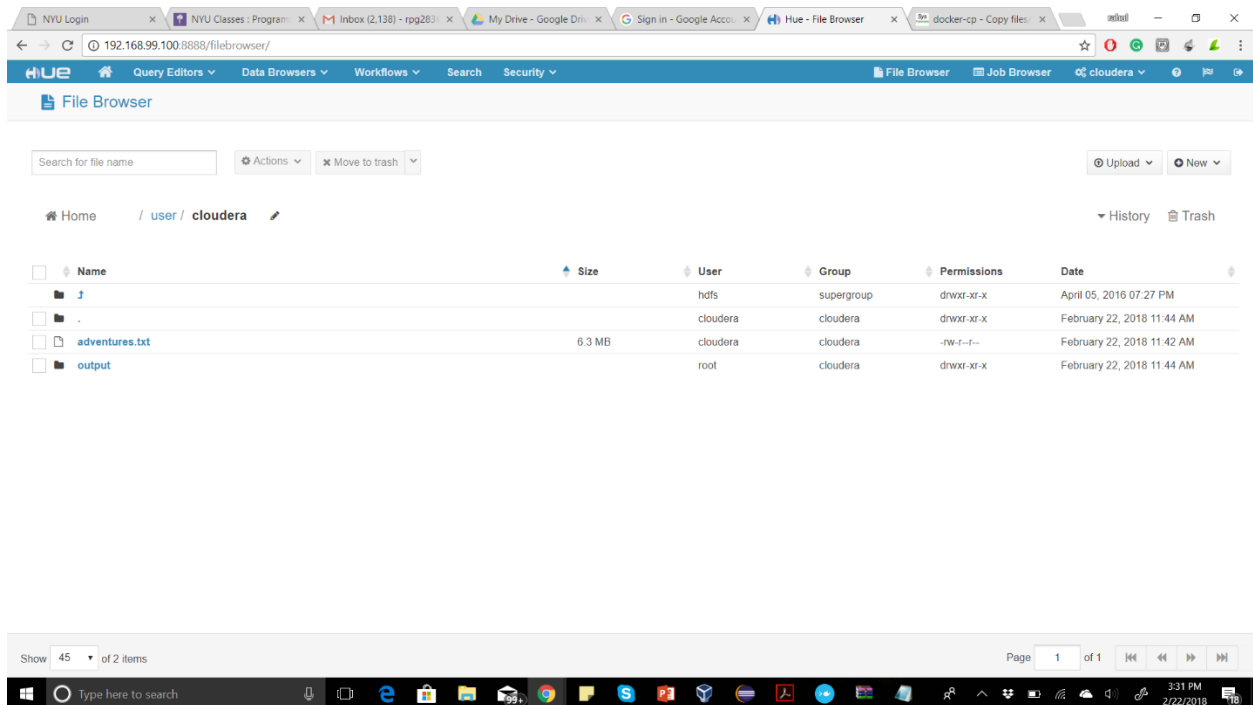


Figure 8: Output folder generated in file HDFS (shown in hue)

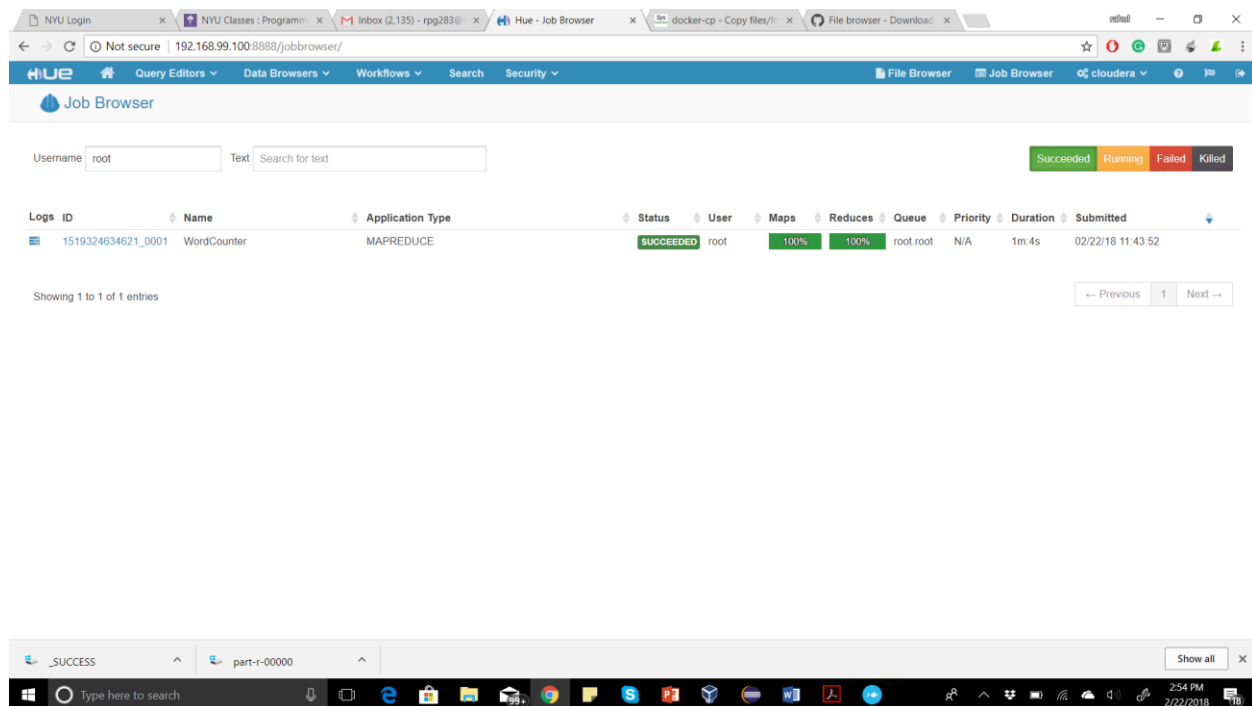


Figure 9: Job Successful Status in hue

Steps to run the Job:

1. Place the WordCountV2.jar file in the hadoop container.
2. Place the adventures.txt file in the HDFS. I had placed it in user/cloudera/adventures.txt
3. Then run the job by executing the command **hadoop jar WordCountV2.jar /user/cloudera/adventures.txt /user/cloudera/output**
4. The output path argument should be a unique path, or it will throw an exception. I have mentioned **/user/cloudera/output** where I will get the output file generated.