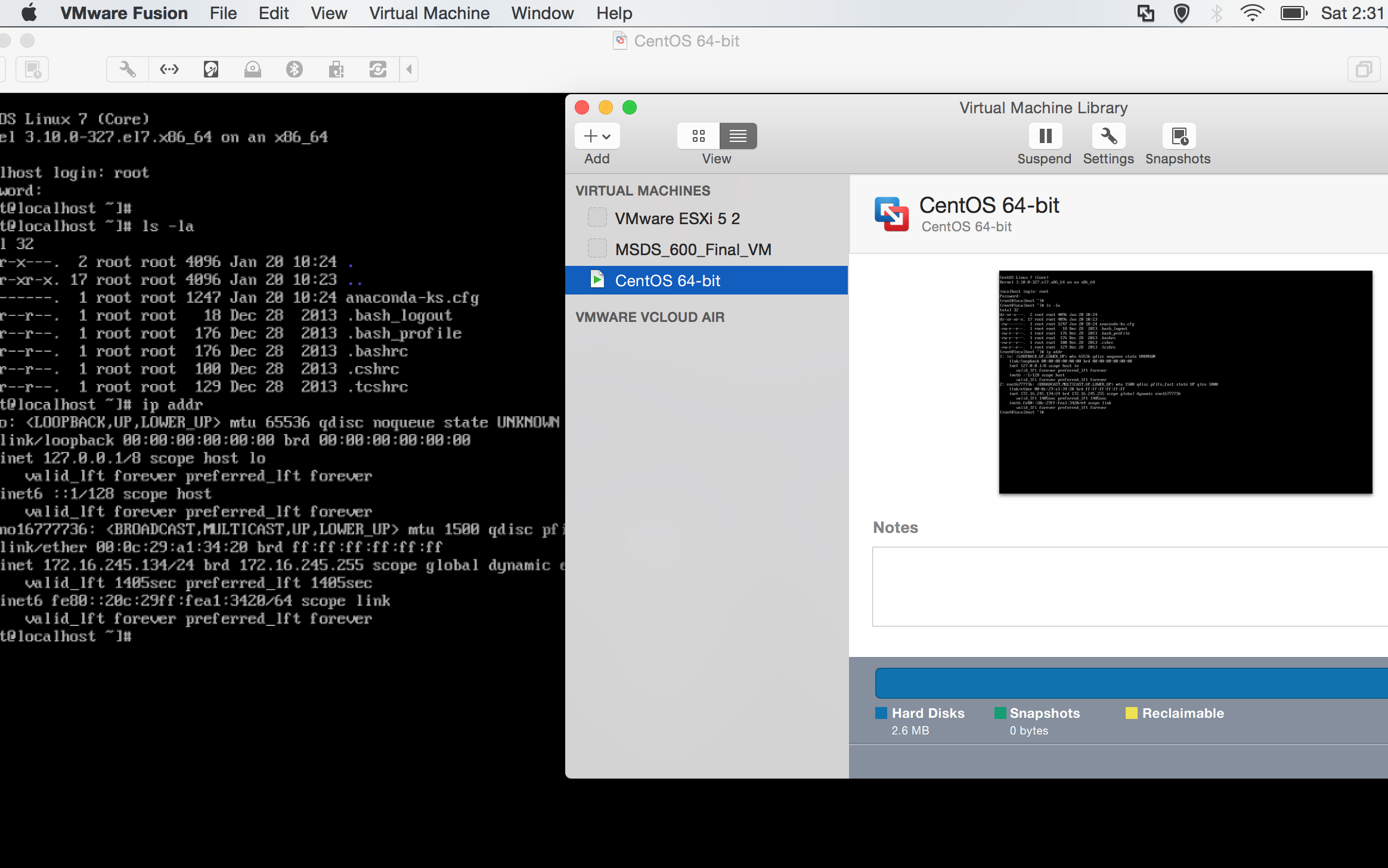
Sean O’Malley

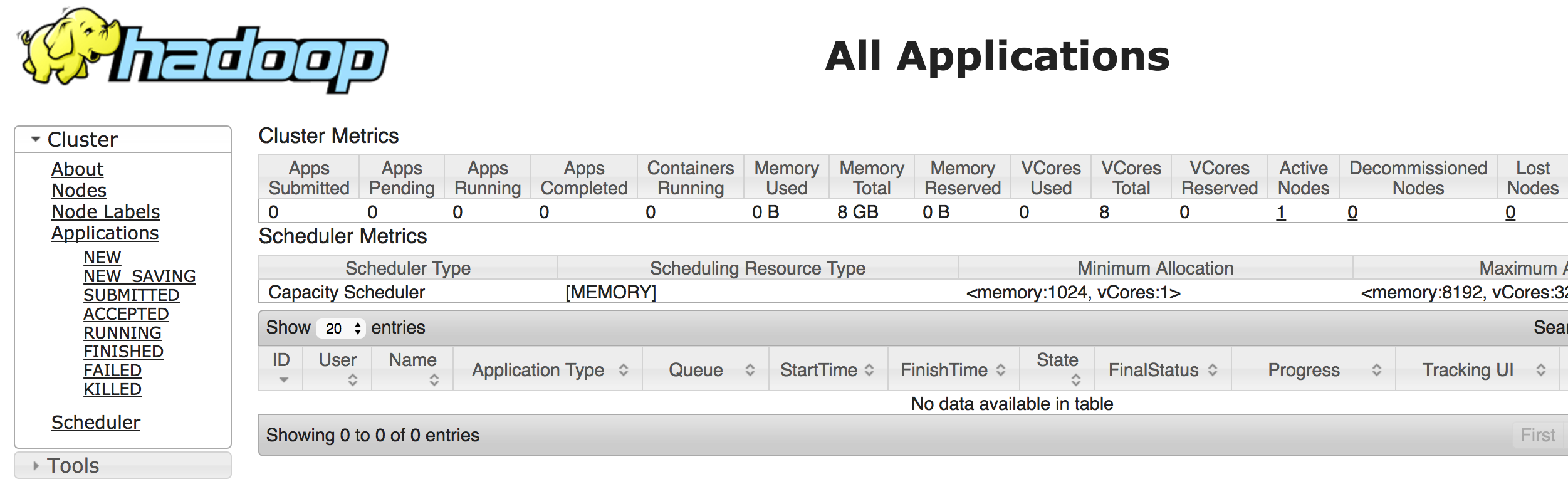
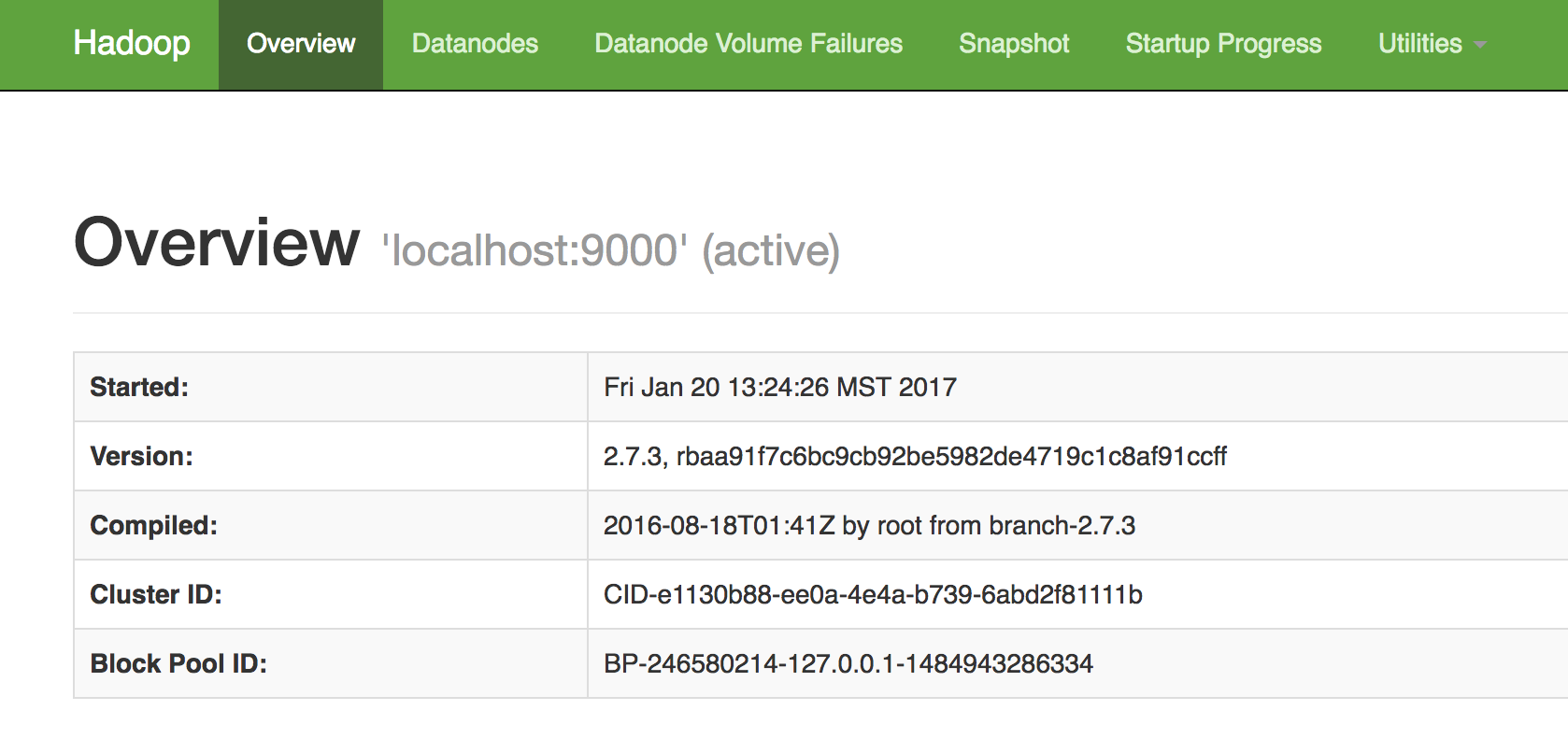
Data Engineering

MSDS

I installed CentOS on VMWare and completed the necessary configurations for the VM, JS and Hadoop in my Mac’s terminal.



I then progressed to run into a couple issues with the CentOS firewall, but was eventually able to establish a connection between my yarn / Hadoop and my browser.



From there I established a new Hadoop cluster, but eventually ran into the issue of having my datanode disappear. With help from the group I removed the datanode and namenode filepath’s and then reconfigured the format of the hdfs. This fixed my problem and I then ran my analysis of the Shakespeare text file from start to finish, getting a full list of ngram wordcounts from every Shakespeare work. Example below:

[hadoop@localhost ~]$ start-dfs.sh

Starting namenodes on [localhost]

localhost: starting namenode, logging to /home/hadoop/hadoop/logs/hadoop-hadoop-namenode-localhost.localdomain.out

localhost: starting datanode, logging to /home/hadoop/hadoop/logs/hadoop-hadoop-datanode-localhost.localdomain.out

Starting secondary namenodes [0.0.0.0]

0.0.0.0: starting secondarynamenode, logging to /home/hadoop/hadoop/logs/hadoop-hadoop-secondarynamenode-localhost.localdomain.out

[hadoop@localhost ~]$ start-yarn.sh

starting yarn daemons

starting resourcemanager, logging to /home/hadoop/hadoop/logs/yarn-hadoop-resourcemanager-localhost.localdomain.out

localhost: starting nodemanager, logging to /home/hadoop/hadoop/logs/yarn-hadoop-nodemanager-localhost.localdomain.out

[hadoop@localhost ~]$ mr-jobhistory-daemon.sh start historyserver

starting historyserver, logging to /home/hadoop/hadoop/logs/mapred-hadoop-historyserver-localhost.localdomain.out

[hadoop@localhost ~]$ jps

62819 SecondaryNameNode

62979 ResourceManager

62516 NameNode

63429 JobHistoryServer

62650 DataNode

63499 Jps

63086 NodeManager

[hadoop@localhost ~]$ hadoop fs -mkdir -p /user/hadoop

[hadoop@localhost ~]$ hadoop fs -ls

[hadoop@localhost ~]$ wget http://norvig.com/ngrams/shakespeare.txt

--2017-01-20 15:47:31-- http://norvig.com/ngrams/shakespeare.txt

Resolving norvig.com (norvig.com)... 66.96.146.129, 66.96.146.129

Connecting to norvig.com (norvig.com)|66.96.146.129|:80... connected.

HTTP request sent, awaiting response... 200 OK

Length: 4538523 (4.3M) [text/plain]

Saving to: ‘shakespeare.txt.3’

100%[========================================>] 4,538,523 1.02MB/s in 4.4s

2017-01-20 15:47:36 (997 KB/s) - ‘shakespeare.txt.3’ saved [4538523/4538523]

[hadoop@localhost ~]$ hadoop fs -mkdir shakespeare

[hadoop@localhost ~]$ hadoop fs -mkdir shakespeare/input

[hadoop@localhost ~]$ hadoop fs -copyFromLocal ~/shakespeare.txt shakespeare/input

[hadoop@localhost ~]$ hadoop fs -ls shakespeare/input

Found 1 items

-rw-r--r-- 1 hadoop supergroup 4538523 2017-01-20 15:48 shakespeare/input/shakespeare.txt

[hadoop@localhost ~]$ hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.3.jar wordcount shakespeare/input shakespeare/output

17/01/20 15:54:26 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

17/01/20 15:54:28 INFO input.FileInputFormat: Total input paths to process : 1

17/01/20 15:54:28 INFO mapreduce.JobSubmitter: number of splits:1

17/01/20 15:54:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1484952245719\_0001

17/01/20 15:54:29 INFO impl.YarnClientImpl: Submitted application application\_1484952245719\_0001

17/01/20 15:54:29 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application\_1484952245719\_0001/

17/01/20 15:54:29 INFO mapreduce.Job: Running job: job\_1484952245719\_0001

17/01/20 15:54:39 INFO mapreduce.Job: Job job\_1484952245719\_0001 running in uber mode : false

17/01/20 15:54:39 INFO mapreduce.Job: map 0% reduce 0%

17/01/20 15:54:49 INFO mapreduce.Job: map 100% reduce 0%

17/01/20 15:54:57 INFO mapreduce.Job: map 100% reduce 100%

17/01/20 15:54:58 INFO mapreduce.Job: Job job\_1484952245719\_0001 completed successfully

17/01/20 15:54:58 INFO mapreduce.Job: Counters: 49

File System Counters

FILE: Number of bytes read=483860

FILE: Number of bytes written=1205087

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=4538655

HDFS: Number of bytes written=356409

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)=7449

Total time spent by all reduces in occupied slots (ms)=4519

Total time spent by all map tasks (ms)=7449

Total time spent by all reduce tasks (ms)=4519

Total vcore-milliseconds taken by all map tasks=7449

Total vcore-milliseconds taken by all reduce tasks=4519

Total megabyte-milliseconds taken by all map tasks=7627776

Total megabyte-milliseconds taken by all reduce tasks=4627456

Map-Reduce Framework

Map input records=129107

Map output records=980637

Map output bytes=8406347

Map output materialized bytes=483860

Input split bytes=132

Combine input records=980637

Combine output records=33505

Reduce input groups=33505

Reduce shuffle bytes=483860

Reduce input records=33505

Reduce output records=33505

Spilled Records=67010

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

GC time elapsed (ms)=290

CPU time spent (ms)=5050

Physical memory (bytes) snapshot=416661504

Virtual memory (bytes) snapshot=4207071232

Total committed heap usage (bytes)=289406976

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=4538523

File Output Format Counters

Bytes Written=356409

[hadoop@localhost ~]$ hadoop fs -ls shakespeare/output

Found 2 items

-rw-r--r-- 1 hadoop supergroup 0 2017-01-20 15:54 shakespeare/output/\_SUCCESS

-rw-r--r-- 1 hadoop supergroup 356409 2017-01-20 15:54 shakespeare/output/part-r-00000

[hadoop@localhost ~]$ hadoop fs -cat shakespeare/output/part-r-00000

! 10526

!' 183

!'By 1

!'t 1

!'twas 1

!, 1

!About 1

!All 1

!As 1

!Ay 1

!Burn 1

!Come 1

!Cuckold 1

!Follow 1

!For 1

!Give't 1

!Handkerchief 1

!Hear 1

!Help 1

!Hoo 1

!How 1

!I 4

!If 1

!Jack 1

!Kill 1

!M 1

!Nay 1

!O 2

!Out 1

!Please 1

!Remain 1

!Seek 1

!Sir 1

!Slay 1

!Speak 1

!Speak-a 1

!Stay 1

!Their 1

!This 1

!Vouchsafe 1

!Wast 1

!We 1

!Well 2

!When 1

!Who's 1

!Why 2

!a 4

!alive 1

!and 1

!any 1

!are 1

!did 1

!do 1

!donc 1

!down 2

!even 2

!for 3

!give 1

!have 2

!he 2

!his 1

!if 1

!it 1

!let 1

!matter 1

!my 1

!nay 1

!on 1

!poor 1

!reveal'd 1

!sixpence 1

!taught 1

!that 1

!the 2

!they 1

!to 3

!videlicet 1

!was 1

!weapons 2

!what 1

!when 1

!who 1

!whose 1

!why 1

!won 1

!would 1

!you 3

"Doubt 1

"To 1

"Upon 1

&c 10

''Tis 5