CS 455 – Spring 2017

Word Count Example

Before starting, make sure that you have HDFS and Yarn running, using sbin/start-dfs.sh and sbin/start-yarn.sh

- Download text copies of at least 3 books from Project Gutenberg: (http://www.gutenberg.org/)

```
st-vrain> la
total 1284
-rw------- 1 class  84358 Mar 26 23:47 Indiscreet_Lettert.txt
-rw------- 1 class 792920 Mar 26 23:48 Tale_Of_Two_Cities.txt
-rw------- 1 class 421884 Mar 26 23:49 Tom_Sawyer.txt
```

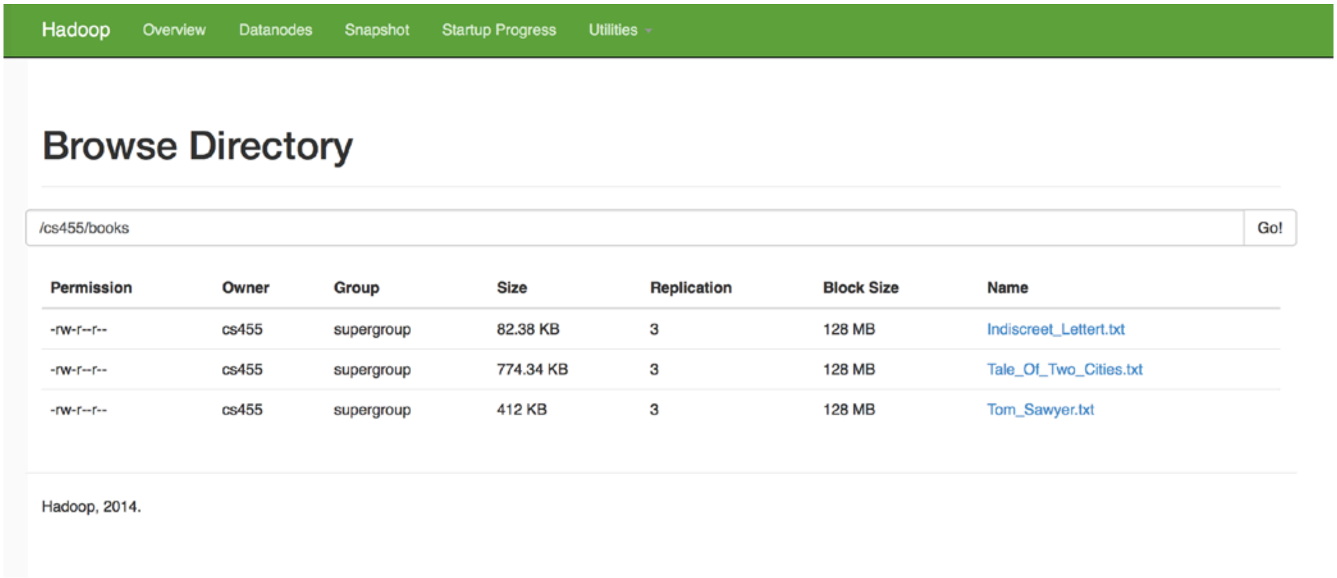
- Create a directory in your local space on HDFS to store these books:

```
st-vrain> $HADOOP_HOME/bin/hdfs dfs -mkdir /cs455
st-vrain> $HADOOP_HOME/bin/hdfs dfs -mkdir /cs455/books
st-vrain> $HADOOP_HOME/bin/hdfs dfs -ls /cs455/
Found 1 items
drwxr-xr-x   - cs455 supergroup          0 2017-03-08 23:51
/cs455/books
```

- Move the books from NFS into HDFS:

```
st-vrain> $HADOOP_HOME/bin/hdfs dfs -put *.txt /cs455/books
st-vrain> $HADOOP_HOME/bin/hdfs dfs -ls /cs455/books
Found 3 items
-rw-r--r--   3 cs455 supergroup      84358 2017-03-08 23:55
/cs455/books/Indiscreet_Lettert.txt
-rw-r--r--   3 cs455 supergroup     792920 2017-03-08 23:55
/cs455/books/Tale_Of_Two_Cities.txt
-rw-r--r--   3 cs455 supergroup     421884 2017-03-08 23:55
/cs455/books/Tom_Sawyer.txt
```

- You can also check that the books are there via the HDFS web portal:



Browse Directory

/cs455/books

Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	cs455	supergroup	82.38 KB	3	128 MB	Indiscreet_Lettert.txt
-rw-r--r--	cs455	supergroup	774.34 KB	3	128 MB	Tale_Of_Two_Cities.txt
-rw-r--r--	cs455	supergroup	412 KB	3	128 MB	Tom_Sawyer.txt

Hadoop, 2014.

- Download the source code of the word count example from CS 455 course web site. (link: http://www.cs.colostate.edu/~cs455/cs455-wordcount-sp17.tar.gz)

```
wget http://www.cs.colostate.edu/~cs455/cs455-wordcount-sp17.tar.gz
```

- Extract the tarball.

```
tar –xvf cs455-wordcount-sp17.tar.gz
```

- This includes an Ant build file called build.xml. This is used to compile source and package it into a jar. After compiling, it will create the jar file inside the `./dist` directory. You can use this build.xml file as it is for HW3-PC. Type ‘ant’ to compile the source and create the jar file.

```
st-vrain> ant
Buildfile: /s/bach/a/class/cs455/sp15-hadoop/word-count/build.xml
init:
compile:
dist:
BUILD SUCCESSFUL
Total time: 0 seconds
st-vrain> ls ./dist/
wordcount.jar
st-vrain>
```

- Run the jar in yarn:

```
st-vrain> $HADOOP_HOME/bin/hadoop jar dist/wordcount.jar
cs455.hadoop.wordcount.WordCountJob /cs455/books /cs455/wordcount-out
2017-03-08 00:36:53,833 INFO  [main] client.RMProxy
(RMProxy.java:createRMProxy(98)) - Connecting to ResourceManager at
st-vrain/129.82.47.128:46783
2017-03-08 00:36:54,325 WARN  [main] mapreduce.JobSubmitter
(JobSubmitter.java:copyAndConfigureFiles(153)) - Hadoop command-line
option parsing not performed. Implement the Tool interface and
execute your application with ToolRunner to remedy this.
2017-03-08 00:36:54,606 INFO  [main] input.FileInputFormat
(FileInputFormat.java:listStatus(281)) - Total input paths to
process : 3
2017-03-08 00:36:54,696 INFO  [main] mapreduce.JobSubmitter
(JobSubmitter.java:submitJobInternal(494)) - number of splits:3
2017-03-08 00:36:54,909 INFO  [main] mapreduce.JobSubmitter
(JobSubmitter.java:printTokens(583)) - Submitting tokens for job:
job_1427438142863_0001
2017-03-08 00:36:55,231 INFO  [main] impl.YarnClientImpl
(YarnClientImpl.java:submitApplication(251)) - Submitted application
application_1427438142863_0001
2017-03-08 00:36:55,270 INFO  [main] mapreduce.Job
(Job.java:submit(1300)) - The url to track the job: http://st-
vrain:8088/proxy/application_1427438142863_0001/
2017-03-08 00:36:55,271 INFO  [main] mapreduce.Job
(Job.java:monitorAndPrintJob(1345)) - Running job:
job_1427438142863_0001
2017-03-08 00:37:01,455 INFO  [main] mapreduce.Job
(Job.java:monitorAndPrintJob(1366)) - Job job_1427438142863_0001
running in uber mode : false
2017-03-08 00:37:01,456 INFO  [main] mapreduce.Job
(Job.java:monitorAndPrintJob(1373)) -  map 0% reduce 0%
2017-03-08 00:37:07,530 INFO  [main] mapreduce.Job
(Job.java:monitorAndPrintJob(1373)) -  map 100% reduce 0%
2017-03-08 00:37:16,599 INFO  [main] mapreduce.Job
(Job.java:monitorAndPrintJob(1373)) -  map 100% reduce 100%
2017-03-08 00:37:17,631 INFO  [main] mapreduce.Job
(Job.java:monitorAndPrintJob(1384)) - Job job_1427438142863_0001
completed successfully
2017-03-08 00:37:17,773 INFO  [main] mapreduce.Job
(Job.java:monitorAndPrintJob(1391)) - Counters: 49
        File System Counters
                FILE: Number of bytes read=549269
                FILE: Number of bytes written=1522267
                FILE: Number of read operations=0
```

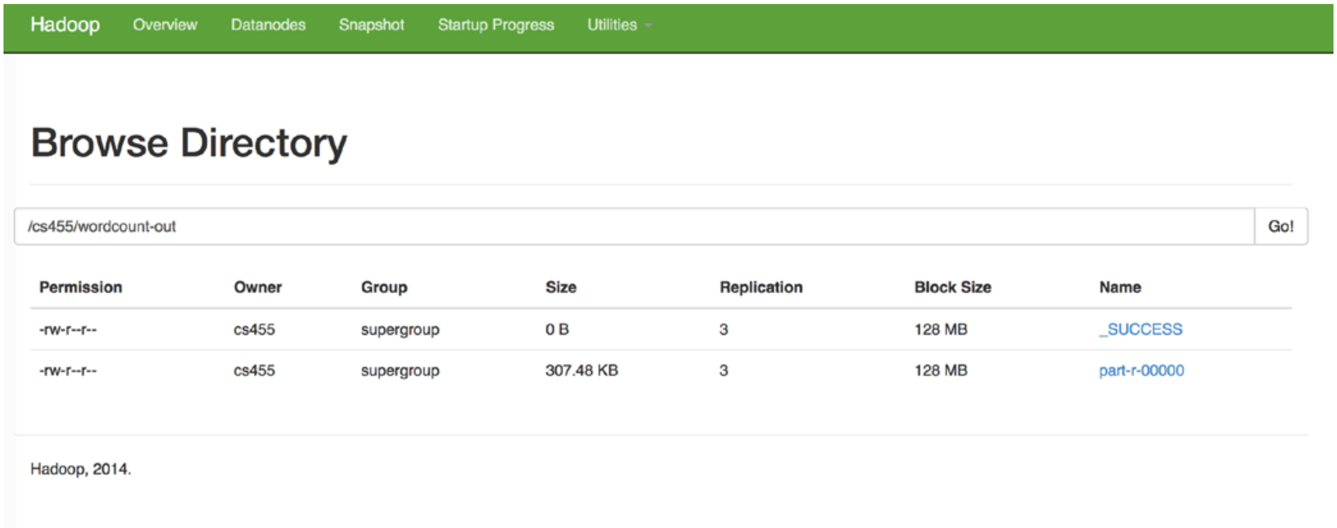
```
           FILE: Number of large read operations=0
           FILE: Number of write operations=0
           HDFS: Number of bytes read=1299517
           HDFS: Number of bytes written=314863
           HDFS: Number of read operations=12
           HDFS: Number of large read operations=0
           HDFS: Number of write operations=2
      Job Counters
           Launched map tasks=3
           Launched reduce tasks=1
           Data-local map tasks=3
           Total time spent by all maps in occupied slots (ms)=10788
           Total time spent by all reduces in occupied slots (ms)=6841
           Total time spent by all map tasks (ms)=10788
           Total time spent by all reduce tasks (ms)=6841
           Total vcore-seconds taken by all map tasks=10788
           Total vcore-seconds taken by all reduce tasks=6841
           Total megabyte-seconds taken by all map tasks=11046912
           Total megabyte-seconds taken by all reduce tasks=7005184
      Map-Reduce Framework


           Map input records=27088
           Map output records=226606
           Map output bytes=2171352
           Map output materialized bytes=549281
           Input split bytes=355
           Combine input records=226606
           Combine output records=38119
           Reduce input groups=29082
           Reduce shuffle bytes=549281
           Reduce input records=38119
           Reduce output records=29082
           Spilled Records=76238
           Shuffled Maps =3
           Failed Shuffles=0
           Merged Map outputs=3
           GC time elapsed (ms)=128
           CPU time spent (ms)=8450
           Physical memory (bytes) snapshot=956612608
           Virtual memory (bytes) snapshot=3741761536
           Total committed heap usage (bytes)=805306368
      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=1299162
      File Output Format Counters
Bytes Written=314863
```

- Check output in HDFS:

```
st-vrain> $HADOOP_HOME/bin/hdfs dfs -ls /cs455/wordcount-out
Found 2 items
-rw-r--r--   3 cs455 supergroup          0 2017-03-08 00:37
/cs455/wordcount-out/_SUCCESS
-rw-r--r--   3 cs455 supergroup     314863 2017-03-08 00:37
/cs455/wordcount-out/part-r-00000
```

- Check the output in the web portal. Click on part-r-00000 file and it will be downloaded.



Browse Directory

Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	cs455	supergroup	0 B	3	128 MB	_SUCCESS
-rw-r--r--	cs455	supergroup	307.48 KB	3	128 MB	part-r-00000

Hadoop, 2014.

- **NOTE:** if you run this repeatedly, you will need to either modify your output folder name, or delete it between runs
- Check the following link for the complete set of HDFS commands. [http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html]