LINKS

Lecture: https://umkc.app.box.com/s/1rrlhhoyqeeslurdpm37yxm8z62y3kg8 Lesson Plan: https://umkc.app.box.com/s/1rrlhhoyqeeslurdpm37yxm8z62y3kg8 Sample Data: https://umkc.app.box.com/s/r4jtmjnoip7g0q8tzyqb2naa78u50t3c

---

TASKS:

1. Counting the frequency of words in the given input with MapReduce algorithm

    a. Create Java WordCount class

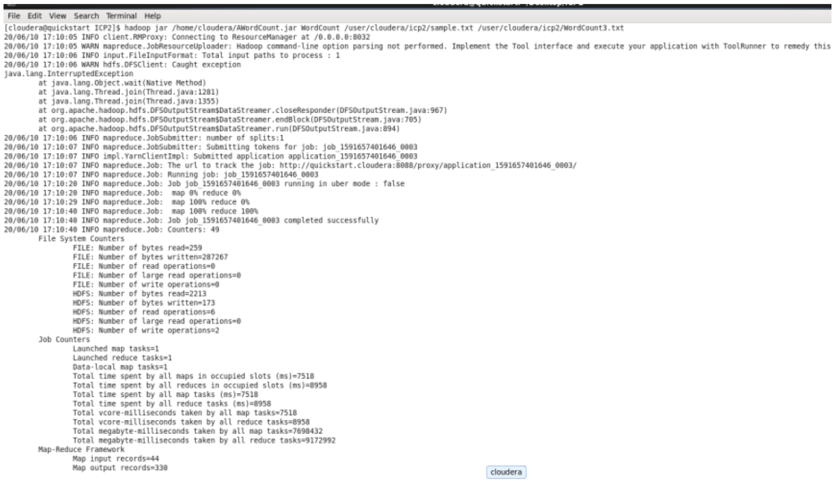
    b. Add external libraries JARS

    c. Export the Jar

    d. Input data file in hdfs

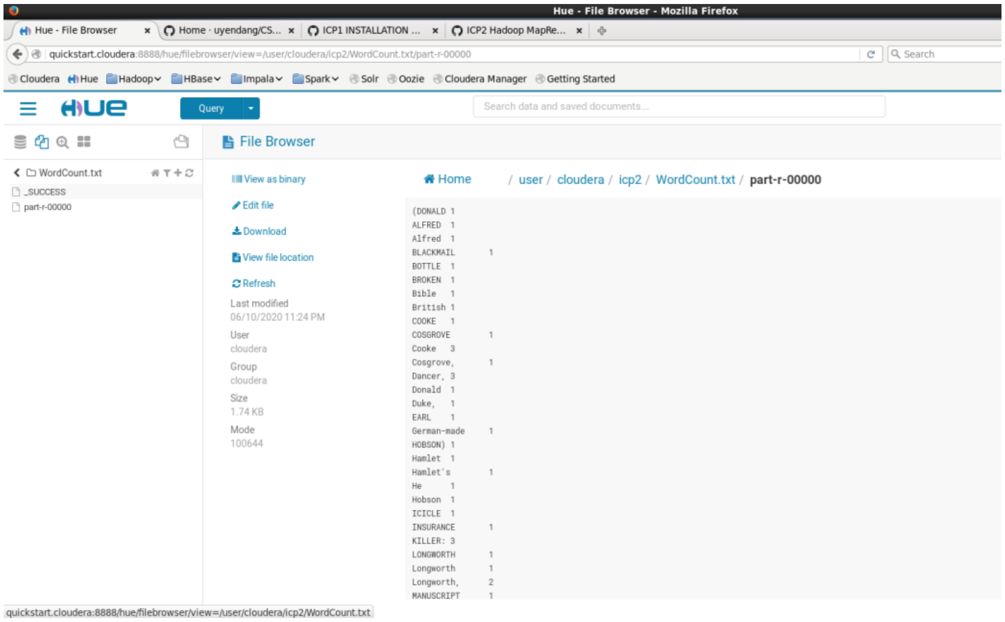
    e. Run MapReduce Job

    hadoop jar /home/cloudera/WordCount.jar WordCount /user/cloudera/icp2/sample.txt /user/cloudera/icp2/WordCount.txt



```
[cloudera@quickstart ICP2]$ hadoop jar /home/cloudera/AWordCount.jar WordCount /user/cloudera/icp2/sample.txt /user/cloudera/icp2/WordCount3.txt
20/06/10 17:10:05 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/06/10 17:10:05 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this
20/06/10 17:10:06 INFO input.FileInputFormat: Total input paths to process : 1
20/06/10 17:10:06 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
20/06/10 17:10:06 INFO mapreduce.JobSubmitter: number of splits:1
20/06/10 17:10:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1591657401646_0003
20/06/10 17:10:07 INFO impl.YarnClientImpl: Submitted application application_1591657401646_0003
20/06/10 17:10:07 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1591657401646_0003/
20/06/10 17:10:07 INFO mapreduce.Job: Running job: job_1591657401646_0003
20/06/10 17:10:20 INFO mapreduce.Job: Job job_1591657401646_0003 running in uber mode : false
20/06/10 17:10:20 INFO mapreduce.Job:  map 0% reduce 0%
20/06/10 17:10:29 INFO mapreduce.Job:  map 100% reduce 0%
20/06/10 17:10:40 INFO mapreduce.Job:  map 100% reduce 100%
20/06/10 17:10:40 INFO mapreduce.Job: Job job_1591657401646_0003 completed successfully
20/06/10 17:10:40 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=259
                FILE: Number of bytes written=287267
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2213
                HDFS: Number of bytes written=173
                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)=7518
                Total time spent by all reduces in occupied slots (ms)=8958
                Total time spent by all map tasks (ms)=7518
                Total time spent by all reduce tasks (ms)=8958
                Total vcore-milliseconds taken by all map tasks=7518
                Total vcore-milliseconds taken by all reduce tasks=8958
                Total megabyte-milliseconds taken by all map tasks=7698432
                Total megabyte-milliseconds taken by all reduce tasks=9172992
        Map-Reduce Framework
                Map input records=44
                Map output records=330
```

Map-Reduce Job



quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera/icp2/WordCount.txt/part-r-00000

WordCount.txt
- _SUCCESS
- part-r-00000

Home / user / cloudera / icp2 / WordCount.txt / part-r-00000

Last modified
06/10/2020 11:24 PM

User
cloudera

Group
cloudera

Size
1.74 KB

Mode
100644

```
(DONALD 1
ALFRED  1
Alfred  1
BLACKMAIL       1
BOTTLE  1
BROKEN  1
Bible   1
British 1
COOKE   1
COSGROVE        1
Cooke   3
Cosgrove,       1
Dancer, 3
Donald  1
Duke,   1
EARL    1
German-made     1
HOBSON) 1
Hamlet  1
Hamlet's        1
He      1
Hobson  1
ICICLE  1
INSURANCE       1
KILLER: 3
LONGWORTH       1
Longworth       1
Longworth,      2
MANUSCRIPT      1
```

quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera/icp2/WordCount.txt

Result

2. Counting the frequency of words in given text file that starts with letter ‘a’

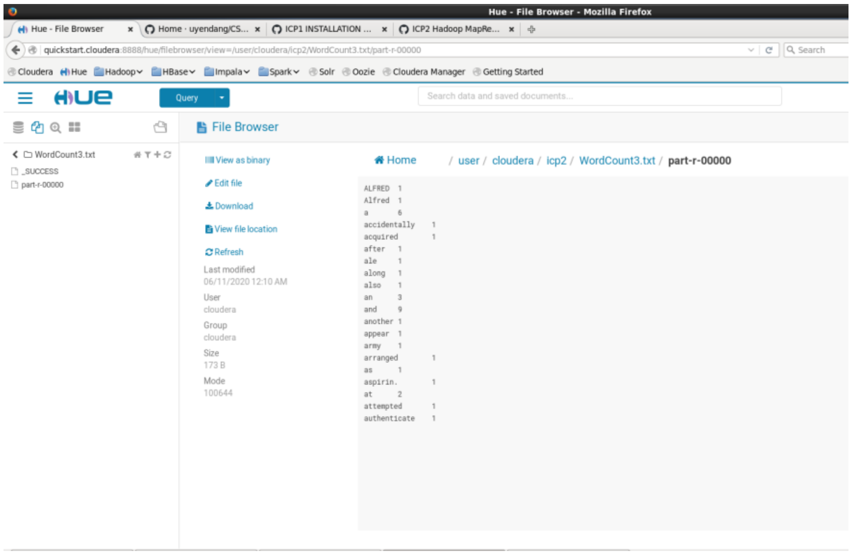
a. Add AWordCount function

b. Export the Jar

c. Input data file in hdfs

d. Run MapReduce Job

```
hadoop jar /home/cloudera/AWordCount.jar WordCount
/user/cloudera/icp2/sample.txt /user/cloudera/icp2/WordCount3.txt
```



Last modified
06/11/2020 12:10 AM
User
cloudera
Group
cloudera
Size
173 B
Mode
100644

```
ALFRED  1
Alfred  1
a       6
accidentally    1
acquired        1
after   1
ale     1
along   1
also    1
an      3
and     9
another 1
appear  1
army    1
arranged        1
as      1
aspirin.        1
at      2
attempted       1
authenticate    1
```

Result

REFERENCES: https://umkc.app.box.com/s/xk4jj0do3p7fa3swx6utcuazx3wny8j8