FACULTY OF ENGINEERING

Word Count Application using Spark

Distributed Systems

Presented by

1- **Islam Yousry Abdelwahid 14**
2- **Andrew Adel Sanad 17**
3- **Seif Eldeen Ehab Mostafa 32**

2/4/2022

Table of Contents

1 Problem Definition

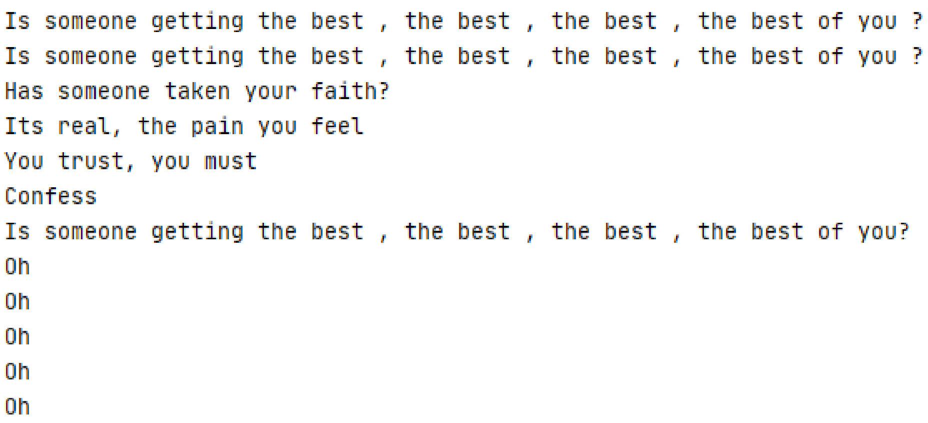
After installing Hadoop in pseudo distributed mode, it is required to install Spark and Scala and create a Java Word Count application that runs as a Spark Job on a created text file in order to find the count of each word found in the document.

2 Algorithms

- Hadoop installation is done on Windows 10 (done in a previous Lab)
- Spark and Scala installation is done on Windows 10
- The WordCount Java application is implemented as described for the Spark Job
- A text file is created to test the WordCount program.
- Output is created to show the results

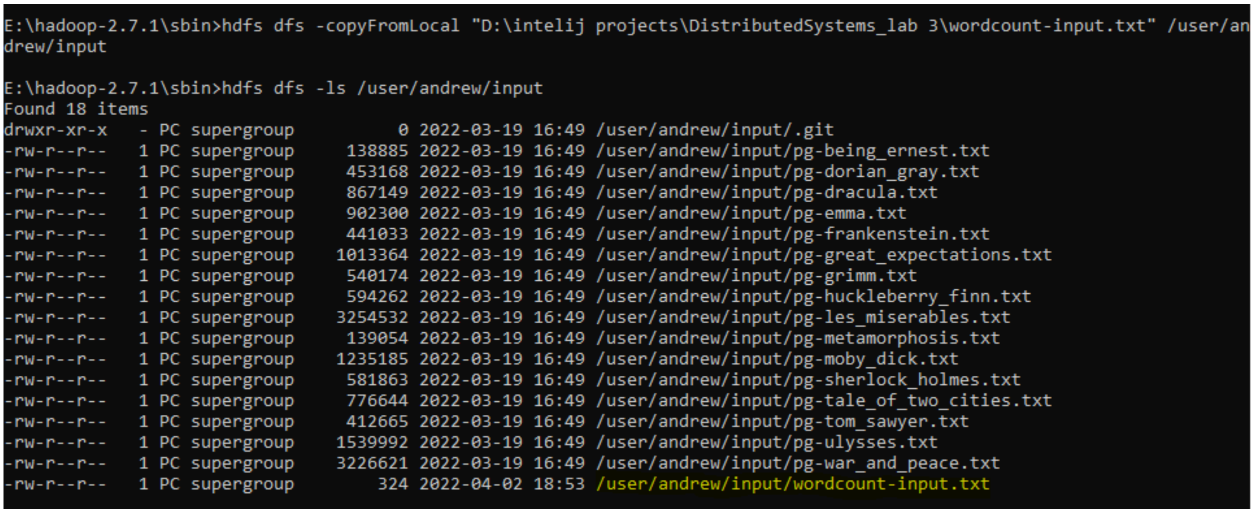
4 Implementation

input file:



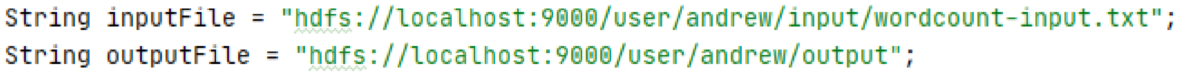
```
Is someone getting the best , the best , the best , the best of you ?
Is someone getting the best , the best , the best , the best of you ?
Has someone taken your faith?
Its real, the pain you feel
You trust, you must
Confess
Is someone getting the best , the best , the best , the best of you?
Oh
Oh
Oh
Oh
Oh
```

Moving it to hdfs:

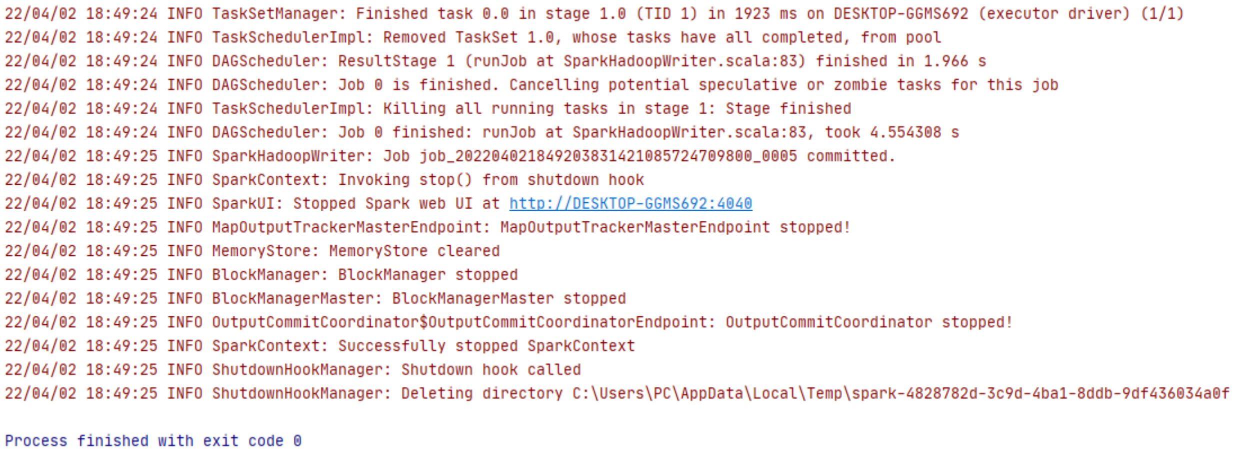


```
E:\hadoop-2.7.1\sbin>hdfs dfs -copyFromLocal "D:\intelij projects\DistributedSystems_lab 3\wordcount-input.txt" /user/an
drew/input

E:\hadoop-2.7.1\sbin>hdfs dfs -ls /user/andrew/input
Found 18 items
drwxr-xr-x   - PC supergroup          0 2022-03-19 16:49 /user/andrew/input/.git
-rw-r--r--   1 PC supergroup     138885 2022-03-19 16:49 /user/andrew/input/pg-being_ernest.txt
-rw-r--r--   1 PC supergroup     453168 2022-03-19 16:49 /user/andrew/input/pg-dorian_gray.txt
-rw-r--r--   1 PC supergroup     867149 2022-03-19 16:49 /user/andrew/input/pg-dracula.txt
-rw-r--r--   1 PC supergroup     902300 2022-03-19 16:49 /user/andrew/input/pg-emma.txt
-rw-r--r--   1 PC supergroup     441033 2022-03-19 16:49 /user/andrew/input/pg-frankenstein.txt
-rw-r--r--   1 PC supergroup    1013364 2022-03-19 16:49 /user/andrew/input/pg-great_expectations.txt
-rw-r--r--   1 PC supergroup     540174 2022-03-19 16:49 /user/andrew/input/pg-grimm.txt
-rw-r--r--   1 PC supergroup     594262 2022-03-19 16:49 /user/andrew/input/pg-huckleberry_finn.txt
-rw-r--r--   1 PC supergroup    3254532 2022-03-19 16:49 /user/andrew/input/pg-les_miserables.txt
-rw-r--r--   1 PC supergroup     139054 2022-03-19 16:49 /user/andrew/input/pg-metamorphosis.txt
-rw-r--r--   1 PC supergroup    1235185 2022-03-19 16:49 /user/andrew/input/pg-moby_dick.txt
-rw-r--r--   1 PC supergroup     581863 2022-03-19 16:49 /user/andrew/input/pg-sherlock_holmes.txt
-rw-r--r--   1 PC supergroup     776644 2022-03-19 16:49 /user/andrew/input/pg-tale_of_two_cities.txt
-rw-r--r--   1 PC supergroup     412665 2022-03-19 16:49 /user/andrew/input/pg-tom_sawyer.txt
-rw-r--r--   1 PC supergroup    1539992 2022-03-19 16:49 /user/andrew/input/pg-ulysses.txt
-rw-r--r--   1 PC supergroup    3226621 2022-03-19 16:49 /user/andrew/input/pg-war_and_peace.txt
-rw-r--r--   1 PC supergroup        324 2022-04-02 18:53 /user/andrew/input/wordcount-input.txt
```

Running spark from intellij with the paths pointing to hdfs:

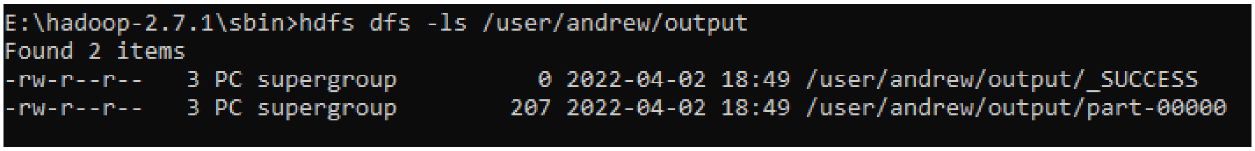


```
String inputFile = "hdfs://localhost:9000/user/andrew/input/wordcount-input.txt";
String outputFile = "hdfs://localhost:9000/user/andrew/output";
```



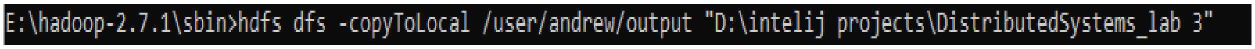
```
22/04/02 18:49:24 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 1923 ms on DESKTOP-GGMS692 (executor driver) (1/1)
22/04/02 18:49:24 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
22/04/02 18:49:24 INFO DAGScheduler: ResultStage 1 (runJob at SparkHadoopWriter.scala:83) finished in 1.966 s
22/04/02 18:49:24 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
22/04/02 18:49:24 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
22/04/02 18:49:24 INFO DAGScheduler: Job 0 finished: runJob at SparkHadoopWriter.scala:83, took 4.554308 s
22/04/02 18:49:25 INFO SparkHadoopWriter: Job job_202204021849203831421085724709800_0005 committed.
22/04/02 18:49:25 INFO SparkContext: Invoking stop() from shutdown hook
22/04/02 18:49:25 INFO SparkUI: Stopped Spark web UI at http://DESKTOP-GGMS692:4040
22/04/02 18:49:25 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/04/02 18:49:25 INFO MemoryStore: MemoryStore cleared
22/04/02 18:49:25 INFO BlockManager: BlockManager stopped
22/04/02 18:49:25 INFO BlockManagerMaster: BlockManagerMaster stopped
22/04/02 18:49:25 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/04/02 18:49:25 INFO SparkContext: Successfully stopped SparkContext
22/04/02 18:49:25 INFO ShutdownHookManager: Shutdown hook called
22/04/02 18:49:25 INFO ShutdownHookManager: Deleting directory C:\Users\PC\AppData\Local\Temp\spark-4828782d-3c9d-4ba1-8ddb-9df436034a0f

Process finished with exit code 0
```

Output directory is created and have the results:



```
E:\hadoop-2.7.1\sbin>hdfs dfs -ls /user/andrew/output
Found 2 items
-rw-r--r--   3 PC supergroup          0 2022-04-02 18:49 /user/andrew/output/_SUCCESS
-rw-r--r--   3 PC supergroup        207 2022-04-02 18:49 /user/andrew/output/part-00000
```

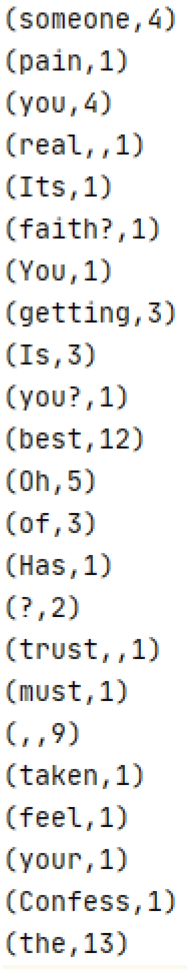
Output is copied to local system:



```
E:\hadoop-2.7.1\sbin>hdfs dfs -copyToLocal /user/andrew/output "D:\intelij projects\DistributedSystems_lab 3"
```

5 Results

The output matches the correct count of each word:



```
(someone,4)
(pain,1)
(you,4)
(real,,1)
(Its,1)
(faith?,1)
(You,1)
(getting,3)
(Is,3)
(you?,1)
(best,12)
(Oh,5)
(of,3)
(Has,1)
(?,2)
(trust,,1)
(must,1)
(,,9)
(taken,1)
(feel,1)
(your,1)
(Confess,1)
(the,13)
```

6 Conclusion

- We can now use a Spark programming framework on top of hadoop in pseudo distributed mode in creating many applications using Java programming language like the WordCount application and test it on as much data as we want.
- We would have changed the application to ignore some unimportant words and symbols and ignore words that count less than a certain threshold.