ASSIGNMENT: CASE STUDY SPARK STREAMING - WORD COUNT

=> There are two parts this case study:-

1) First Part - You have to create a Spark Application which streams data from a file on local directory on your machine and does the word count on the fly. The word should be done by the spark application in such a way that as soon as you drop the file in your local directory, your spark application should immediately do the word count for you.

2) Second Part - In this part, you will have to create a Spark Application which should do the following

A) Pick up a file from the local directory and do the word count
B) Then in the same Spark Application, write the code to put the same file on HDFS.
C) Then in same Spark Application, do the word count of the file copied on HDFS in step 2
D) Lastly, compare the word count of step 1 and 2. Both should match, other throw an error

TASK 1 : LOCAL WORD COUNT

SPARK STREAMING CODE FOR WORD COUNT:

```
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.log4j.{Level,Logger}

object StreamingSparkWordCount {
 def main(args: Array[String]): Unit = {
    println("hey Spark Streaming")

    val conf = new
SparkConf().setMaster("local[2]").setAppName("SparkSteamingExample")
    val sc = new SparkContext(conf)
    val rootLogger = Logger.getRootLogger()
   rootLogger.setLevel(Level.ERROR)
    val ssc = new StreamingContext(sc, Seconds(15))
    val lines = ssc.textFileStream(args(0))
    val words = lines.flatMap(_.split(" "))
    val wordCounts = words.map(x => (x, 1)).reduceByKey(_ +   _).count
    wordCounts.print()
    ssc.start()
    ssc.awaitTermination()
 }
}
```

EXPLANATION: SO FIRST WE CREATE A MAVEN PROJECT AND ADDED ALL THE MAVEN DEPENDENCIES FOR SPARK STREAMING TO WORK. THEN WE IMPORTED THE NECESSARY SPARK AND STREAMING CONTEXTS AND CREATED SPARK SESSION AS "conf" AND SET THE MASTER TO "local[2]". THEN CREATED A SPARK CONTEXT OBJECT.

CODE: val conf = new
SparkConf().setMaster("local[2]").setAppName("SparkSteaming
Example")

val sc = new SparkContext(conf)

NOW WE NEED TO SET ROOT LOGGER AS IN STANDALONE MODE, THE SPARK STREAMING DRIVER IS RUNNING ON THE MACHINE WHERE YOU SUBMIT THE JOB AND EACH SPARK WORKING NODE WILL RUN AN EXECUTOR FOR THE JOB. SO WE NEED TO SETUP "log4j". SO HERE WE ARE SETTING THE LOGGER AND LEVEL.

CODE: val rootLogger = Logger.getRootLogger()
rootLogger.setLevel(Level.ERROR)

THEN WE CREATED A OBJECT FOR STREAMING CONTEXT AS "ssc" WITH WORKING THREAD AND BATCH INTERVAL SET TO 15s.

CODE: val ssc = new StreamingContext(sc, Seconds(15))

NOW THIS STREAMING CONTEXT WILL READ THE INPUT STREAM GIVEN BY THE ARGUMENT WHICH IS THE PATH FOR THE TEXT FILE PRESENT IN THE LOCAL DIRECTORY.

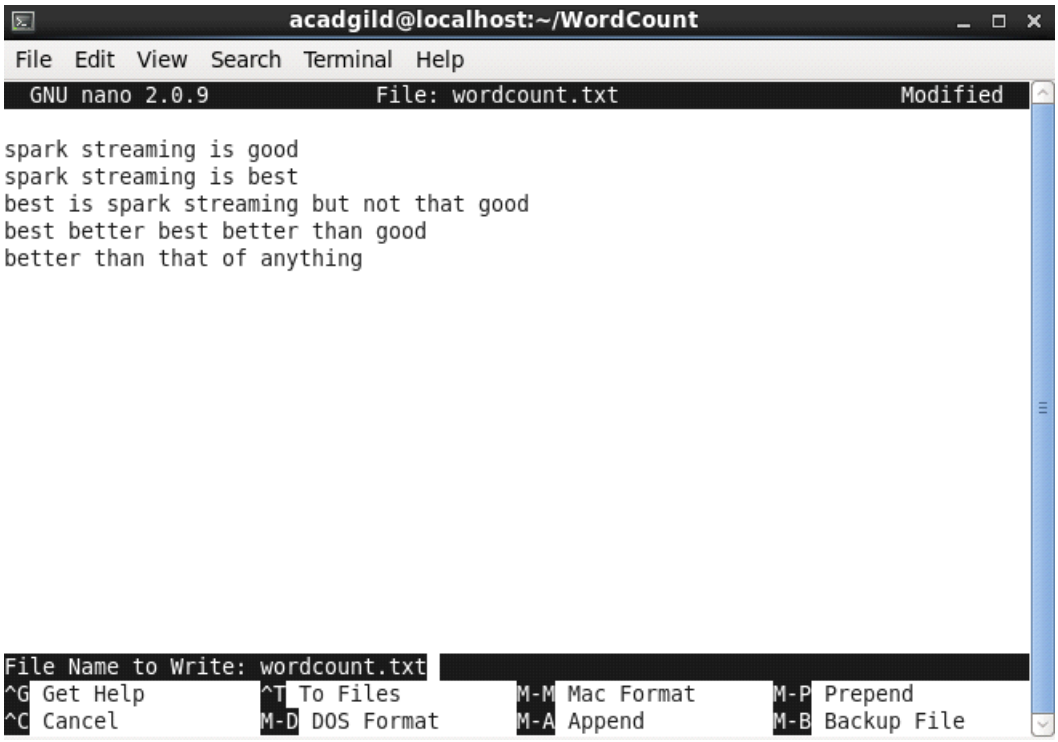
CODE: val lines = ssc.textFileStream(args(0))

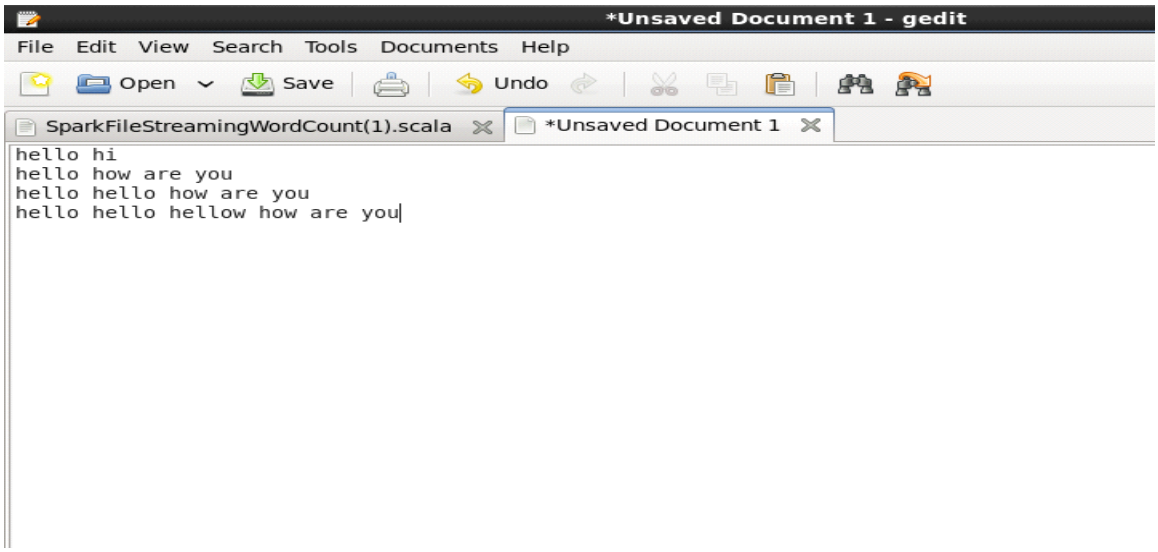
AFTER THE INPUT STREAM IS READ BY THE SPARK STREAMING. WE USE "flatMap" FUNCTION TO SPLIT THE WORDS WITH DELIMTER("") AND THEN WE MAP EACH WORD INTO "words,1" THEN WITH THE HELP OF "reduceByKey" FUNCTION AND USING "count" WE GET THE OUTPUT.

CODE: val words = lines.flatMap(_.split(" "))
val wordCounts = words.map(x => (x, 1)).reduceByKey(_ +
_).count

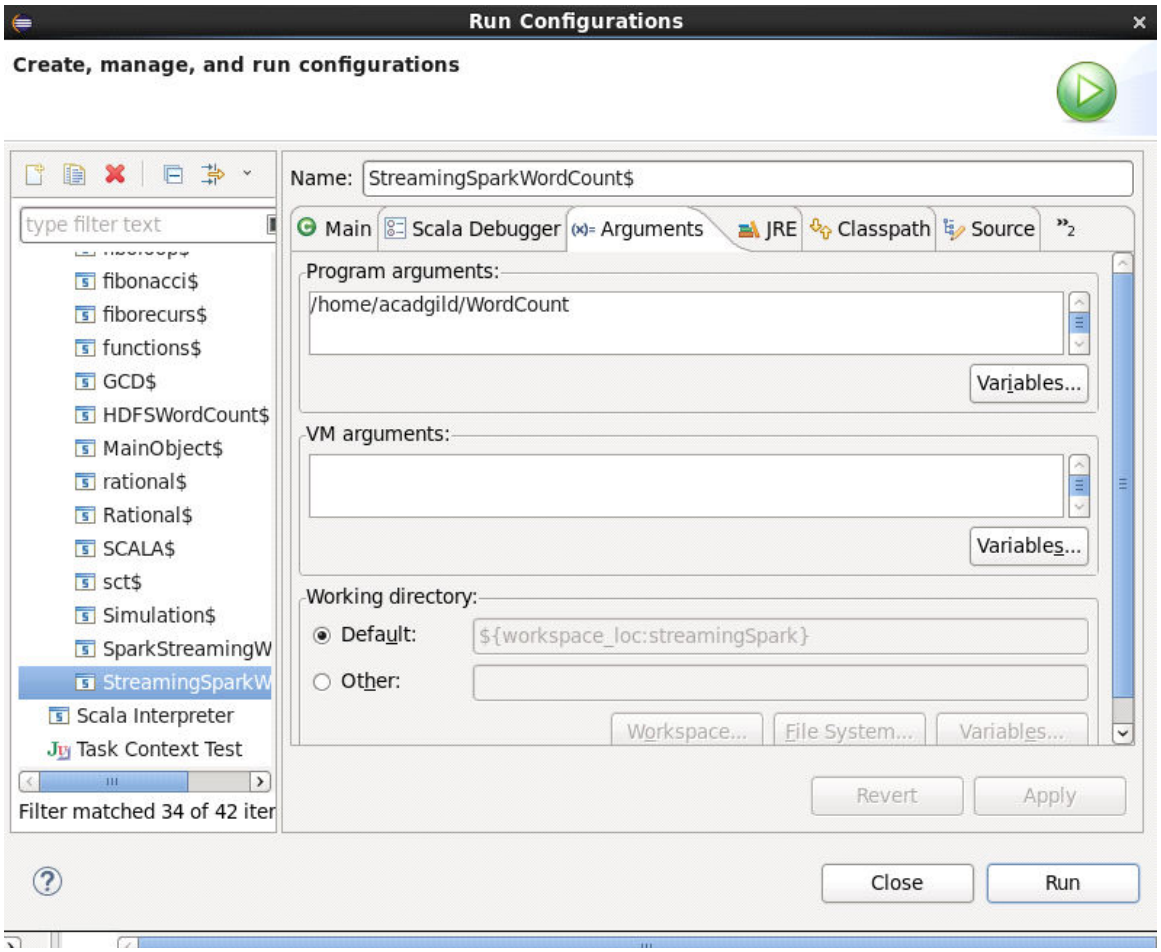
EXPLANATION: NOW WE RUN THE CODE BY RIGHT CLICKING THE STREAMING OBJECT WHICH IS CREATED THEN WE GO TO RUN CONFIGURATION AS SHOWN BELOW. THEN WE GO TO ARGUMENTS AND GIVE THE PATH TO THE STREAM SO THAT SPARK STREAMING CAN START THE WORD COUNT. SO BEFORE THAT WE NEED TO CREATE A DIRECTORY IN LOCAL DIRECTORY AS "WordCount" AS SHOWN IN THE SCREENSHOT. NOW AS SOON AS WE DROP A FILE WITH SOME CONTENTS WRITTEN INTO THIS DIRECTORY THE SPARK STREAMING WILL AUTOMATICALLY START THE WORD COUNT WHICH WILL SPLIT THE WORDS AS "Words,1" THEN COUNT THE OCCURANCES OF EACH WORDS. THIS IS SHOWN BELOW IN THE SOLUTION REPORT. SO BELOW ARE THE FILES WHICH WE CREATED INSIDE THE DIRECTORY "WordCount" IN LOCAL.

PATH : /home/acadgild/WordCount/





RUN CONFIGURATION:



SOLUTION REPORT:

```
hey Spark Streaming
Using Spark's default log4j profile:
org/apache/spark/log4j-defaults.properties
19/03/14 10:53:34 INFO SparkContext: Running Spark version 2.2.1
19/03/14 10:53:37 WARN NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
19/03/14 10:53:38 WARN Utils: Your hostname, localhost.localdomain
resolves to a loopback address: 127.0.0.1; using 192.168.43.62 instead (on
interface eth17)
19/03/14 10:53:38 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to
another address
19/03/14 10:53:38 INFO SparkContext: Submitted application:
SparkSteamingExample
19/03/14 10:53:38 INFO SecurityManager: Changing view acls to: acadgild
19/03/14 10:53:38 INFO SecurityManager: Changing modify acls to: acadgild
19/03/14 10:53:38 INFO SecurityManager: Changing view acls groups to:
19/03/14 10:53:38 INFO SecurityManager: Changing modify acls groups to:
19/03/14 10:53:38 INFO SecurityManager: SecurityManager: authentication
disabled; ui acls disabled; users  with view permissions: Set(acadgild);
groups with view permissions: Set(); users  with modify permissions:
Set(acadgild); groups with modify permissions: Set()
```

```
19/03/14 10:53:39 INFO Utils: Successfully started service 'sparkDriver'
on port 41417.
19/03/14 10:53:40 INFO SparkEnv: Registering MapOutputTracker
19/03/14 10:53:40 INFO SparkEnv: Registering BlockManagerMaster
19/03/14 10:53:40 INFO BlockManagerMasterEndpoint: Using
org.apache.spark.storage.DefaultTopologyMapper for getting topology
information
19/03/14 10:53:40 INFO BlockManagerMasterEndpoint:
BlockManagerMasterEndpoint up
19/03/14 10:53:40 INFO DiskBlockManager: Created local directory at
/tmp/blockmgr-28c3882d-7cec-42b9-802b-ffa65bf8efd6
19/03/14 10:53:40 INFO MemoryStore: MemoryStore started with capacity 111.2
MB
19/03/14 10:53:40 INFO SparkEnv: Registering OutputCommitCoordinator
19/03/14 10:53:41 INFO Utils: Successfully started service 'SparkUI' on
port 4040.
19/03/14 10:53:41 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at
http://192.168.43.62:4040
19/03/14 10:53:41 INFO Executor: Starting executor ID driver on host
localhost
19/03/14 10:53:42 INFO Utils: Successfully started service
'org.apache.spark.network.netty.NettyBlockTransferService' on port
46204.
19/03/14 10:53:42 INFO NettyBlockTransferService: Server created on
192.168.43.62:46204
19/03/14 10:53:42 INFO BlockManager: Using
org.apache.spark.storage.RandomBlockReplicationPolicy for block
replication policy
19/03/14 10:53:42 INFO BlockManagerMaster: Registering BlockManager
BlockManagerId(driver, 192.168.43.62, 46204, None)
19/03/14 10:53:42 INFO BlockManagerMasterEndpoint: Registering block
manager 192.168.43.62:46204 with 111.2 MB RAM, BlockManagerId(driver,
192.168.43.62, 46204, None)
19/03/14 10:53:42 INFO BlockManagerMaster: Registered BlockManager
BlockManagerId(driver, 192.168.43.62, 46204, None)
19/03/14 10:53:42 INFO BlockManager: Initialized BlockManager:
BlockManagerId(driver, 192.168.43.62, 46204, None)
-------------------------------------------
Time: 1552541040000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541055000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541070000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541085000 ms
-------------------------------------------
```

```
-------------------------------------------
Time: 1552541100000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541115000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541130000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541145000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541160000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541175000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541190000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541205000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541220000 ms
-------------------------------------------
(is,3)
(anything,1)
(best,4)
(better,3)
(streaming,3)
(not,1)
(spark,3)
(than,2)
(that,2)
(of,1)
...

-------------------------------------------
Time: 1552541235000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541250000 ms
```

```
-------------------------------------------

-------------------------------------------
Time: 1552541265000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541280000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541295000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541310000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541325000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541340000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541355000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541370000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541385000 ms
-------------------------------------------

-------------------------------------------
Time: 1552541395000 ms
-------------------------------------------
(are,6)
(how,6)
(hello,12)
(you,6)
(hi,2)
(hellow,2)

-------------------------------------------
Time: 1552541410000 ms
-------------------------------------------
```

TASK 2- HDFS WORDCOUNT COMPARISON

HDFS WORD COUNT CODE:

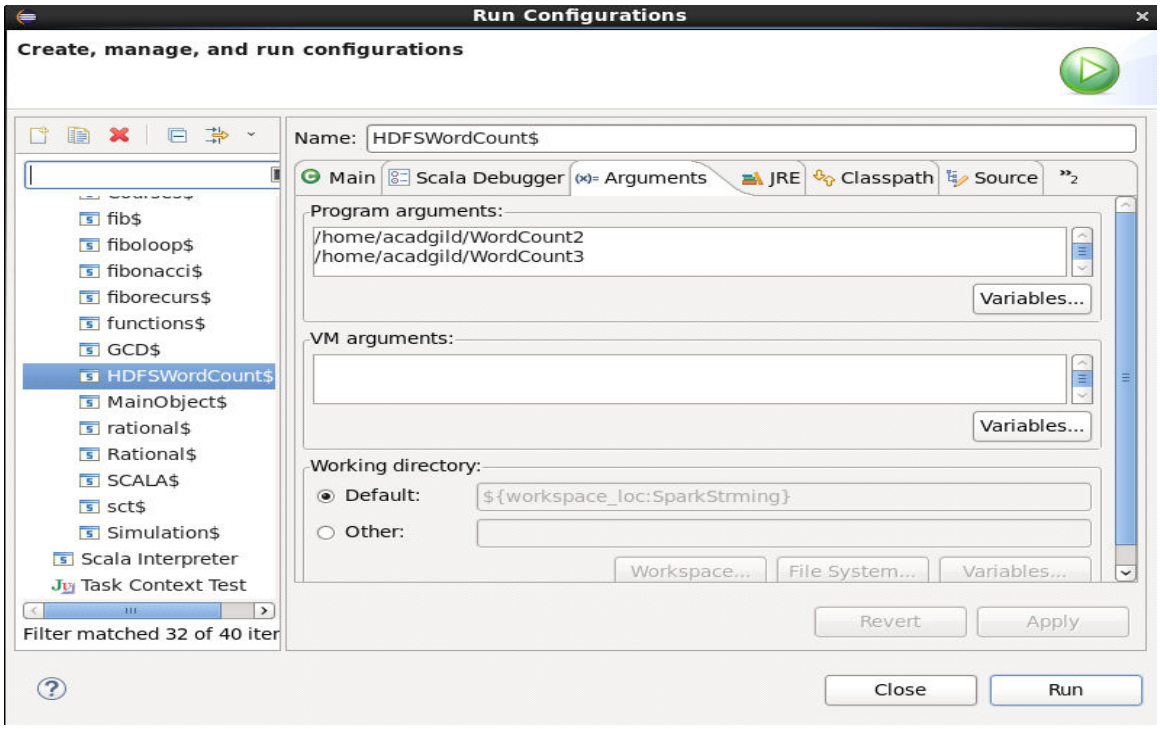
```
package NewSparkStreaming

import org.apache.spark
import org.apache.spark.SparkContext
import org.apache.spark.SparkConf
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.fs.{FileAlreadyExistsException, FileSystem,
FileUtil, Path}
import scala.io.Source

object HDFSWordCount {

   def main(args:Array[String]) {
  val conf = new SparkConf().setMaster("local[*]").setAppName("HDFS WORD
COUNT COMPARISON")
  val sc = new SparkContext(conf)
  val hadoopConf = new Configuration()
   if(args.length < 2) {
   println("Usage:ScalaWordCount<output1><output2>")
  System.exit(1)
  }

//PICKING UP THE FILE FROM LOCAL DIRECTORY

  val rawData = sc.textFile("/home/acadgild/WordCount/")

//THEN WE ADD THE core-site.xml AND hdfs-site.xml FOR COPYING DATA INTO HDFS
FROM LOCAL FILE SYSTEM
//CODE TO PUT THE SAME FILE INTO HDFS

  hadoopConf.addResource(new
Path("/home/acadgild/install/hadoop/hadoop-2.6.5/etc/hadoop/core-site.x
ml"))

  hadoopConf.addResource(new
Path("/home/acadgild/install/hadoop/hadoop-2.6.5/etc/hadoop/hdfs-site.x
ml"))

//ADDING HADOOP CONFIGURATION TO FILE SYSTEM FOR COPYING DATA FROM LOCAL
TO HDFS

  val fs = FileSystem.get(hadoopConf);
  val sourcePath = new Path("/home/acadgild/WordCount")
  val destPath = new Path("hdfs://localhost:8020/")

  if((fs.exists(destPath)))
  {
```

```
      System.out.println("Such destination exists"+destPath);
  }

//COPYING FILE FROM SOURCE TO DESTINATION PATH

  fs.copyFromLocalFile(sourcePath, destPath)

//WORD COUNT ON THE SAME FILE COPIED IN HDFS

  val words = rawData.flatMap(line => line.split(" "))
  val hdfsfile = sc.textFile("hdfs://localhost:8020/WordCount/test")
  val hdfswords = hdfsfile.flatMap(line => line.split(" "))

  val wordCount = words.map(x => (x,1)).reduceByKey(_+_)
  val hdfsWC = hdfswords.map(x => (x,1))reduceByKey(_+_)

//SAVING RESULT IN THE PATH MENTIONED BY THE ARGUMENTS

  wordCount.saveAsTextFile(args(0))
  hdfsWC.saveAsTextFile(args(1))

//COMPARING HDFS FILE WORD COUNT WITH LOCAL FILE WORD COUNT AND PRINTING
THE RESULT

  val LFSWCfile =
Source.fromFile("/home/acadgild/WordCount2/part-00000").getLines().toAr
ray
  val hdfsWCfile =
Source.fromFile("/home/acadgild/WordCount3/part-00000").getLines().toAr
ray

  val elem = LFSWCfile.sameElements(hdfsWCfile)
  if(elem == false) {
  println("Error!:Contents mismatch")
  }
  else     println("Contents match!")

  wordCount.collect().foreach(print)
  hdfsWC.collect().foreach(print)

//STOPPING SPARK CONTEXT

  sc.stop

 }
}
```

EXPLANATION: SO HERE WE DO THE SAME BY CREATING THE MAVEN PROJECT AND ADDING THE DEPENDENCIES. THEN WE IMPORT THE NECESSARY PACKAGES AS SHOWN IN THE CODE. THEN WE CREATE A SPARK CONF OBJECT THEN A SPARK CONTEXT OBJECT. THEN WE CREATE A CONFIGURATION OBJECT FOR HADOOP. AFTER CREATING ALL THE CONTEXTS AND CONFIGURATION OBJECTS. NEXT WE PICK THE FILE FROM THE LOCAL DIRECTORY WHICH IS THE PATH "WordCount" AND THE FILE IS "test". THEN WE PUT THE SAME FILE

IN THE HDFS USING "core-site.xml" AND "hdfs-site.xml". AFTER COPYING THE SAME FILE INTO HDFS WE THEN DO THE WORD COUNT FOR THE FILE IN HDFS. THIS RESULT IS SAVED AND MENTIONED IN BY THE ARGUMENTS. LASTLY WE COMPARE BOTH FILES IN DIFFERENT FILE SYSTEM AND PRINT THE RESULT. BELOW IS THE CODE FOR THE SAME.

NOW FOR RUNNING THE APPLICATION WE FIRST GO INTO RUN CONFIGURATION AND ENTER THE PATH INSIDE THE ARGUMENT "args(0)" AND "args(1)".



SOLUTION REPORT:

```
Using Spark's default log4j profile:
org/apache/spark/log4j-defaults.properties
19/03/12 21:23:18 INFO SparkContext: Running Spark version 2.2.1
19/03/12 21:23:20 WARN NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
19/03/12 21:23:22 WARN Utils: Your hostname, localhost.localdomain
resolves to a loopback address: 127.0.0.1; using 192.168.0.107 instead (on
interface eth17)
19/03/12 21:23:22 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to
another address
19/03/12 21:23:22 INFO SparkContext: Submitted application: HDFS WORD COUNT
COMPARISON
19/03/12 21:23:22 INFO SecurityManager: Changing view acls to: acadgild
19/03/12 21:23:22 INFO SecurityManager: Changing modify acls to: acadgild
19/03/12 21:23:22 INFO SecurityManager: Changing view acls groups to:
19/03/12 21:23:22 INFO SecurityManager: Changing modify acls groups to:
19/03/12 21:23:22 INFO SecurityManager: SecurityManager: authentication
disabled; ui acls disabled; users  with view permissions: Set(acadgild);
```

```
groups with view permissions: Set(); users  with modify permissions:
Set(acadgild); groups with modify permissions: Set()
19/03/12 21:23:23 INFO Utils: Successfully started service 'sparkDriver'
on port 36542.
19/03/12 21:23:24 INFO SparkEnv: Registering MapOutputTracker
19/03/12 21:23:24 INFO SparkEnv: Registering BlockManagerMaster
19/03/12 21:23:24 INFO BlockManagerMasterEndpoint: Using
org.apache.spark.storage.DefaultTopologyMapper for getting topology
information
19/03/12 21:23:24 INFO BlockManagerMasterEndpoint:
BlockManagerMasterEndpoint up
19/03/12 21:23:24 INFO DiskBlockManager: Created local directory at
/tmp/blockmgr-d84bcc15-e99d-42cc-b1c7-e66dc1d500ca
19/03/12 21:23:24 INFO MemoryStore: MemoryStore started with capacity 111.2
MB
19/03/12 21:23:24 INFO SparkEnv: Registering OutputCommitCoordinator
19/03/12 21:23:25 INFO Utils: Successfully started service 'SparkUI' on
port 4040.
19/03/12 21:23:26 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at
http://192.168.0.107:4040
19/03/12 21:23:26 INFO Executor: Starting executor ID driver on host
localhost
19/03/12 21:23:26 INFO Utils: Successfully started service
'org.apache.spark.network.netty.NettyBlockTransferService' on port
38220.
19/03/12 21:23:26 INFO NettyBlockTransferService: Server created on
192.168.0.107:38220
19/03/12 21:23:26 INFO BlockManager: Using
org.apache.spark.storage.RandomBlockReplicationPolicy for block
replication policy
19/03/12 21:23:26 INFO BlockManagerMaster: Registering BlockManager
BlockManagerId(driver, 192.168.0.107, 38220, None)
19/03/12 21:23:26 INFO BlockManagerMasterEndpoint: Registering block
manager 192.168.0.107:38220 with 111.2 MB RAM, BlockManagerId(driver,
192.168.0.107, 38220, None)
19/03/12 21:23:26 INFO BlockManagerMaster: Registered BlockManager
BlockManagerId(driver, 192.168.0.107, 38220, None)
19/03/12 21:23:26 INFO BlockManager: Initialized BlockManager:
BlockManagerId(driver, 192.168.0.107, 38220, None)
19/03/12 21:23:30 INFO MemoryStore: Block broadcast_0 stored as values in
memory (estimated size 214.5 KB, free 111.0 MB)
19/03/12 21:23:31 INFO MemoryStore: Block broadcast_0_piece0 stored as
bytes in memory (estimated size 20.4 KB, free 111.0 MB)
19/03/12 21:23:31 INFO BlockManagerInfo: Added broadcast_0_piece0 in
memory on 192.168.0.107:38220 (size: 20.4 KB, free: 111.2 MB)
19/03/12 21:23:31 INFO SparkContext: Created broadcast 0 from textFile at
HDFSWordCount.scala:22
Such destination existshdfs://localhost:8020/
19/03/12 21:23:36 INFO MemoryStore: Block broadcast_1 stored as values in
memory (estimated size 214.5 KB, free 110.7 MB)
19/03/12 21:23:36 INFO MemoryStore: Block broadcast_1_piece0 stored as
bytes in memory (estimated size 20.4 KB, free 110.7 MB)
19/03/12 21:23:36 INFO BlockManagerInfo: Added broadcast_1_piece0 in
```

```
memory on 192.168.0.107:38220 (size: 20.4 KB, free: 111.1 MB)
19/03/12 21:23:36 INFO SparkContext: Created broadcast 1 from textFile at
HDFSWordCount.scala:39
19/03/12 21:23:36 INFO FileInputFormat: Total input paths to process : 1
19/03/12 21:23:37 INFO FileInputFormat: Total input paths to process : 1
19/03/12 21:23:37 INFO SparkContext: Starting job: saveAsTextFile at
HDFSWordCount.scala:45
19/03/12 21:23:38 INFO DAGScheduler: Registering RDD 6 (map at
HDFSWordCount.scala:42)
19/03/12 21:23:38 INFO DAGScheduler: Got job 0 (saveAsTextFile at
HDFSWordCount.scala:45) with 1 output partitions
19/03/12 21:23:38 INFO DAGScheduler: Final stage: ResultStage 1
(saveAsTextFile at HDFSWordCount.scala:45)
19/03/12 21:23:38 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 0)
19/03/12 21:23:38 INFO DAGScheduler: Missing parents: List(ShuffleMapStage
0)
19/03/12 21:23:38 INFO DAGScheduler: Submitting ShuffleMapStage 0
(MapPartitionsRDD[6] at map at HDFSWordCount.scala:42), which has no
missing parents
19/03/12 21:23:39 INFO MemoryStore: Block broadcast_2 stored as values in
memory (estimated size 4.8 KB, free 110.7 MB)
19/03/12 21:23:39 INFO MemoryStore: Block broadcast_2_piece0 stored as
bytes in memory (estimated size 2.8 KB, free 110.7 MB)
19/03/12 21:23:39 INFO BlockManagerInfo: Added broadcast_2_piece0 in
memory on 192.168.0.107:38220 (size: 2.8 KB, free: 111.1 MB)
19/03/12 21:23:39 INFO SparkContext: Created broadcast 2 from broadcast at
DAGScheduler.scala:1006
19/03/12 21:23:39 INFO DAGScheduler: Submitting 1 missing tasks from
ShuffleMapStage 0 (MapPartitionsRDD[6] at map at HDFSWordCount.scala:42)
(first 15 tasks are for partitions Vector(0))
19/03/12 21:23:39 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
19/03/12 21:23:39 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID
0, localhost, executor driver, partition 0, PROCESS_LOCAL, 4842 bytes)
19/03/12 21:23:39 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
19/03/12 21:23:40 INFO HadoopRDD: Input split:
file:/home/acadgild/WordCount/test:0+86
19/03/12 21:23:40 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0).
1197 bytes result sent to driver
19/03/12 21:23:40 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID
0) in 1049 ms on localhost (executor driver) (1/1)
19/03/12 21:23:40 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks
have all completed, from pool
19/03/12 21:23:40 INFO DAGScheduler: ShuffleMapStage 0 (map at
HDFSWordCount.scala:42) finished in 1.263 s
19/03/12 21:23:40 INFO DAGScheduler: looking for newly runnable stages
19/03/12 21:23:40 INFO DAGScheduler: running: Set()
19/03/12 21:23:40 INFO DAGScheduler: waiting: Set(ResultStage 1)
19/03/12 21:23:40 INFO DAGScheduler: failed: Set()
19/03/12 21:23:40 INFO DAGScheduler: Submitting ResultStage 1
(MapPartitionsRDD[10] at saveAsTextFile at HDFSWordCount.scala:45), which
has no missing parents
19/03/12 21:23:41 INFO MemoryStore: Block broadcast_3 stored as values in
```

```
memory (estimated size 65.3 KB, free 110.7 MB)
19/03/12 21:23:41 INFO MemoryStore: Block broadcast_3_piece0 stored as
bytes in memory (estimated size 23.3 KB, free 110.6 MB)
19/03/12 21:23:41 INFO BlockManagerInfo: Added broadcast_3_piece0 in
memory on 192.168.0.107:38220 (size: 23.3 KB, free: 111.1 MB)
19/03/12 21:23:41 INFO SparkContext: Created broadcast 3 from broadcast at
DAGScheduler.scala:1006
19/03/12 21:23:41 INFO DAGScheduler: Submitting 1 missing tasks from
ResultStage 1 (MapPartitionsRDD[10] at saveAsTextFile at
HDFSWordCount.scala:45) (first 15 tasks are for partitions Vector(0))
19/03/12 21:23:41 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
19/03/12 21:23:41 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID
1, localhost, executor driver, partition 0, ANY, 4621 bytes)
19/03/12 21:23:41 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
19/03/12 21:23:41 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty
blocks out of 1 blocks
19/03/12 21:23:41 INFO ShuffleBlockFetcherIterator: Started 0 remote
fetches in 31 ms
19/03/12 21:23:42 INFO FileOutputCommitter: Saved output of task
'attempt_20190312212337_0001_m_000000_1' to
file:/home/acadgild/WordCount2/_temporary/0/task_20190312212337_0001_m_
000000
19/03/12 21:23:42 INFO SparkHadoopMapRedUtil:
attempt_20190312212337_0001_m_000000_1: Committed
19/03/12 21:23:42 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1).
1267 bytes result sent to driver
19/03/12 21:23:42 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID
1) in 1389 ms on localhost (executor driver) (1/1)
19/03/12 21:23:42 INFO DAGScheduler: ResultStage 1 (saveAsTextFile at
HDFSWordCount.scala:45) finished in 1.391 s
19/03/12 21:23:42 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks
have all completed, from pool
19/03/12 21:23:42 INFO DAGScheduler: Job 0 finished: saveAsTextFile at
HDFSWordCount.scala:45, took 5.105623 s
19/03/12 21:23:42 INFO SparkContext: Starting job: saveAsTextFile at
HDFSWordCount.scala:46
19/03/12 21:23:42 INFO DAGScheduler: Registering RDD 8 (map at
HDFSWordCount.scala:43)
19/03/12 21:23:42 INFO DAGScheduler: Got job 1 (saveAsTextFile at
HDFSWordCount.scala:46) with 1 output partitions
19/03/12 21:23:42 INFO DAGScheduler: Final stage: ResultStage 3
(saveAsTextFile at HDFSWordCount.scala:46)
19/03/12 21:23:42 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 2)
19/03/12 21:23:42 INFO DAGScheduler: Missing parents: List(ShuffleMapStage
2)
19/03/12 21:23:42 INFO DAGScheduler: Submitting ShuffleMapStage 2
(MapPartitionsRDD[8] at map at HDFSWordCount.scala:43), which has no
missing parents
19/03/12 21:23:42 INFO MemoryStore: Block broadcast_4 stored as values in
memory (estimated size 4.8 KB, free 110.6 MB)
19/03/12 21:23:42 INFO MemoryStore: Block broadcast_4_piece0 stored as
bytes in memory (estimated size 2.7 KB, free 110.6 MB)
```

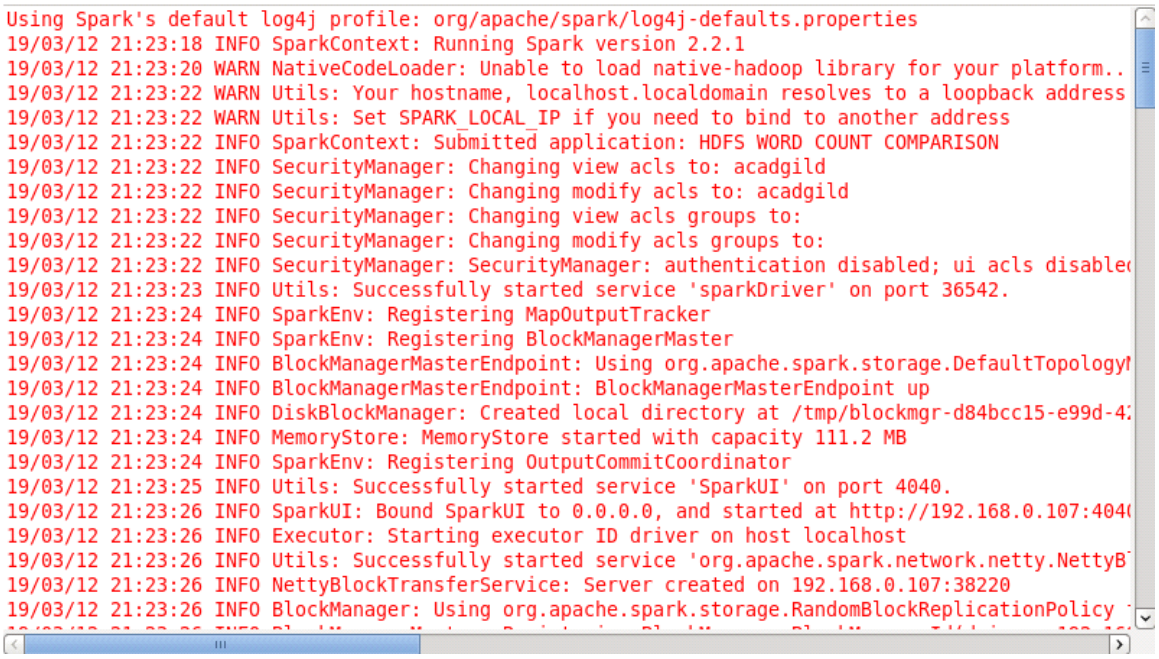
```
19/03/12 21:23:42 INFO BlockManagerInfo: Added broadcast_4_piece0 in
memory on 192.168.0.107:38220 (size: 2.7 KB, free: 111.1 MB)
19/03/12 21:23:42 INFO SparkContext: Created broadcast 4 from broadcast at
DAGScheduler.scala:1006
19/03/12 21:23:42 INFO DAGScheduler: Submitting 1 missing tasks from
ShuffleMapStage 2 (MapPartitionsRDD[8] at map at HDFSWordCount.scala:43)
(first 15 tasks are for partitions Vector(0))
19/03/12 21:23:42 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
19/03/12 21:23:42 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID
2, localhost, executor driver, partition 0, PROCESS_LOCAL, 4844 bytes)
19/03/12 21:23:42 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
19/03/12 21:23:42 INFO HadoopRDD: Input split:
hdfs://localhost:8020/WordCount/test:0+86
19/03/12 21:23:43 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2).
1111 bytes result sent to driver
19/03/12 21:23:43 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID
2) in 404 ms on localhost (executor driver) (1/1)
19/03/12 21:23:43 INFO DAGScheduler: ShuffleMapStage 2 (map at
HDFSWordCount.scala:43) finished in 0.398 s
19/03/12 21:23:43 INFO DAGScheduler: looking for newly runnable stages
19/03/12 21:23:43 INFO DAGScheduler: running: Set()
19/03/12 21:23:43 INFO DAGScheduler: waiting: Set(ResultStage 3)
19/03/12 21:23:43 INFO DAGScheduler: failed: Set()
19/03/12 21:23:43 INFO DAGScheduler: Submitting ResultStage 3
(MapPartitionsRDD[11] at saveAsTextFile at HDFSWordCount.scala:46), which
has no missing parents
19/03/12 21:23:43 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks
have all completed, from pool
19/03/12 21:23:43 INFO MemoryStore: Block broadcast_5 stored as values in
memory (estimated size 65.3 KB, free 110.6 MB)
19/03/12 21:23:43 INFO MemoryStore: Block broadcast_5_piece0 stored as
bytes in memory (estimated size 23.3 KB, free 110.5 MB)
19/03/12 21:23:43 INFO BlockManagerInfo: Added broadcast_5_piece0 in
memory on 192.168.0.107:38220 (size: 23.3 KB, free: 111.1 MB)
19/03/12 21:23:43 INFO SparkContext: Created broadcast 5 from broadcast at
DAGScheduler.scala:1006
19/03/12 21:23:43 INFO DAGScheduler: Submitting 1 missing tasks from
ResultStage 3 (MapPartitionsRDD[11] at saveAsTextFile at
HDFSWordCount.scala:46) (first 15 tasks are for partitions Vector(0))
19/03/12 21:23:43 INFO TaskSchedulerImpl: Adding task set 3.0 with 1 tasks
19/03/12 21:23:43 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID
3, localhost, executor driver, partition 0, ANY, 4621 bytes)
19/03/12 21:23:43 INFO Executor: Running task 0.0 in stage 3.0 (TID 3)
19/03/12 21:23:43 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty
blocks out of 1 blocks
19/03/12 21:23:43 INFO ShuffleBlockFetcherIterator: Started 0 remote
fetches in 0 ms
19/03/12 21:23:43 INFO FileOutputCommitter: Saved output of task
'attempt_20190312212342_0003_m_000000_3' to
file:/home/acadgild/WordCount3/_temporary/0/task_20190312212342_0003_m_
000000
19/03/12 21:23:43 INFO SparkHadoopMapRedUtil:
attempt_20190312212342_0003_m_000000_3: Committed
```

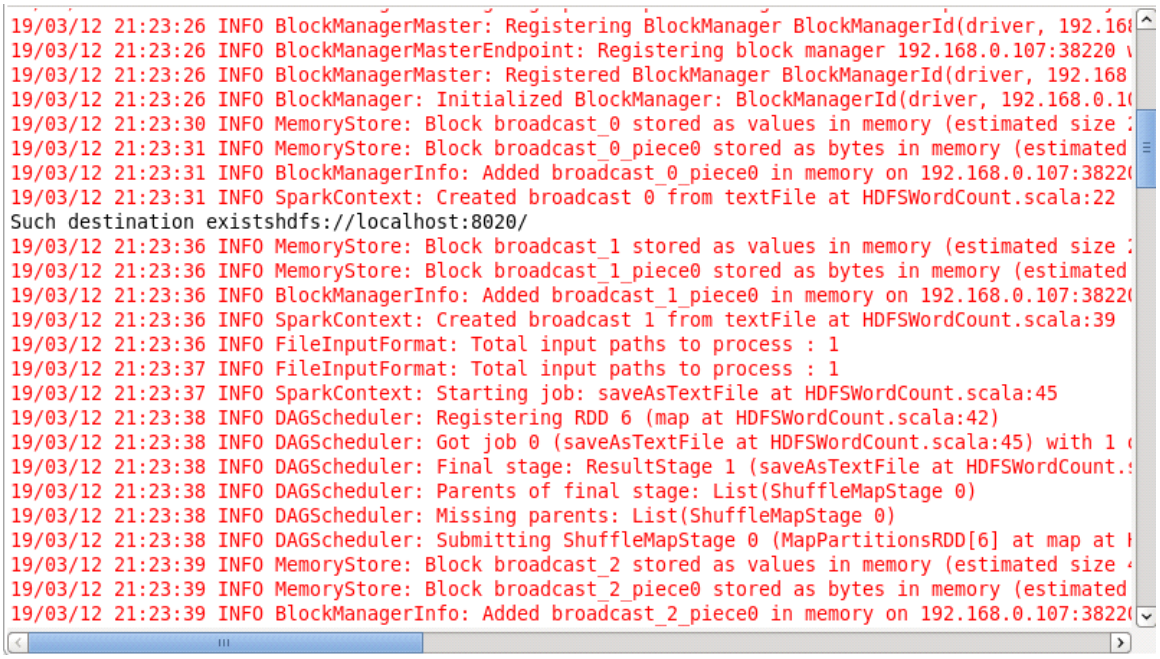
```
19/03/12 21:23:43 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 1181 bytes result sent to driver
19/03/12 21:23:43 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 205 ms on localhost (executor driver) (1/1)
19/03/12 21:23:43 INFO DAGScheduler: ResultStage 3 (saveAsTextFile at HDFSWordCount.scala:46) finished in 0.209 s
19/03/12 21:23:43 INFO DAGScheduler: Job 1 finished: saveAsTextFile at HDFSWordCount.scala:46, took 0.933627 s
19/03/12 21:23:43 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
Contents match!
19/03/12 21:23:43 INFO SparkContext: Starting job: collect at HDFSWordCount.scala:57
19/03/12 21:23:43 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 149 bytes
19/03/12 21:23:44 INFO DAGScheduler: Got job 2 (collect at HDFSWordCount.scala:57) with 1 output partitions
19/03/12 21:23:44 INFO DAGScheduler: Final stage: ResultStage 5 (collect at HDFSWordCount.scala:57)
19/03/12 21:23:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 4)
19/03/12 21:23:44 INFO DAGScheduler: Missing parents: List()
19/03/12 21:23:44 INFO DAGScheduler: Submitting ResultStage 5 (ShuffledRDD[7] at reduceByKey at HDFSWordCount.scala:42), which has no missing parents
19/03/12 21:23:44 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 3.2 KB, free 110.5 MB)
19/03/12 21:23:44 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 1977.0 B, free 110.5 MB)
19/03/12 21:23:44 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on 192.168.0.107:38220 (size: 1977.0 B, free: 111.1 MB)
19/03/12 21:23:44 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1006
19/03/12 21:23:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (ShuffledRDD[7] at reduceByKey at HDFSWordCount.scala:42) (first 15 tasks are for partitions Vector(0))
19/03/12 21:23:44 INFO TaskSchedulerImpl: Adding task set 5.0 with 1 tasks
19/03/12 21:23:44 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 4, localhost, executor driver, partition 0, ANY, 4621 bytes)
19/03/12 21:23:44 INFO Executor: Running task 0.0 in stage 5.0 (TID 4)
19/03/12 21:23:44 INFO BlockManagerInfo: Removed broadcast_5_piece0 on 192.168.0.107:38220 in memory (size: 23.3 KB, free: 111.1 MB)
19/03/12 21:23:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
19/03/12 21:23:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
19/03/12 21:23:44 INFO Executor: Finished task 0.0 in stage 5.0 (TID 4). 1393 bytes result sent to driver
19/03/12 21:23:44 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 4) in 120 ms on localhost (executor driver) (1/1)
19/03/12 21:23:44 INFO DAGScheduler: ResultStage 5 (collect at HDFSWordCount.scala:57) finished in 0.108 s
19/03/12 21:23:44 INFO DAGScheduler: Job 2 finished: collect at
```

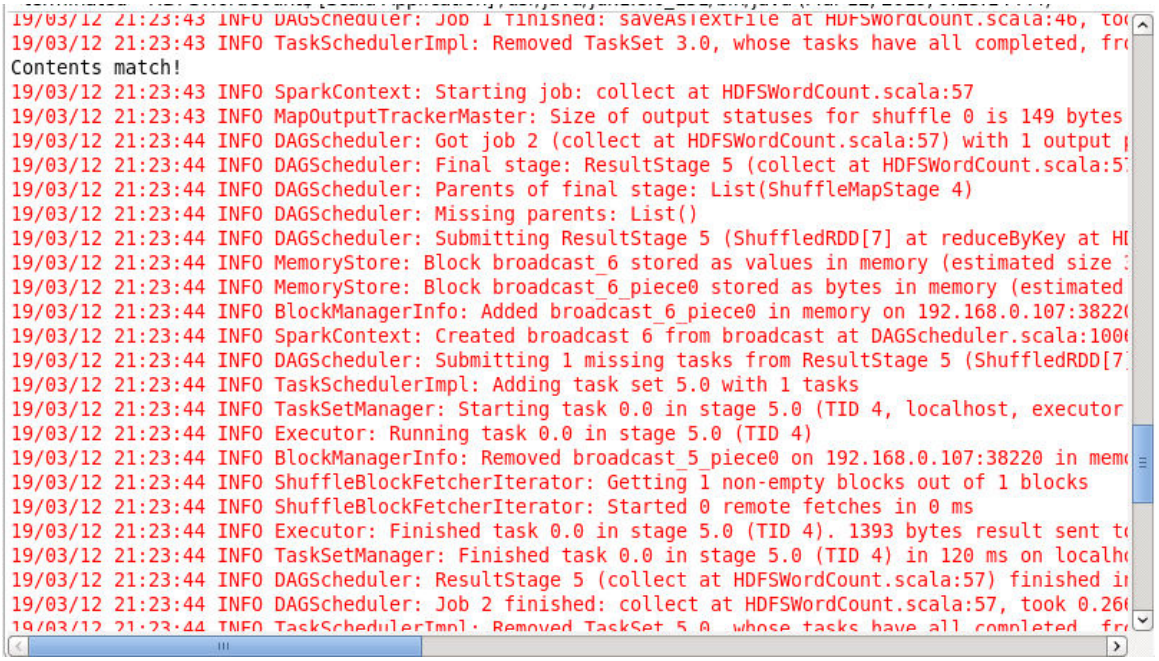
```
HDFSWordCount.scala:57, took 0.266926 s
19/03/12 21:23:44 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks
have all completed, from pool
(are,3)(is,2)(how,5)(u,3)(amit,2)(,1)(and,1)(shravan,2)(yogi,2)19/03/12
21:23:44 INFO SparkContext: Starting job: collect at
HDFSWordCount.scala:58
19/03/12 21:23:44 INFO MapOutputTrackerMaster: Size of output statuses for
shuffle 1 is 149 bytes
19/03/12 21:23:44 INFO DAGScheduler: Got job 3 (collect at
HDFSWordCount.scala:58) with 1 output partitions
19/03/12 21:23:44 INFO DAGScheduler: Final stage: ResultStage 7 (collect
at HDFSWordCount.scala:58)
19/03/12 21:23:44 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 6)
19/03/12 21:23:44 INFO DAGScheduler: Missing parents: List()
19/03/12 21:23:44 INFO DAGScheduler: Submitting ResultStage 7
(ShuffledRDD[9] at reduceByKey at HDFSWordCount.scala:43), which has no
missing parents
19/03/12 21:23:44 INFO MemoryStore: Block broadcast_7 stored as values in
memory (estimated size 3.2 KB, free 110.6 MB)
19/03/12 21:23:44 INFO BlockManagerInfo: Removed broadcast_6_piece0 on
192.168.0.107:38220 in memory (size: 1977.0 B, free: 111.1 MB)
19/03/12 21:23:44 INFO MemoryStore: Block broadcast_7_piece0 stored as
bytes in memory (estimated size 1972.0 B, free 110.6 MB)
19/03/12 21:23:44 INFO BlockManagerInfo: Added broadcast_7_piece0 in
memory on 192.168.0.107:38220 (size: 1972.0 B, free: 111.1 MB)
19/03/12 21:23:44 INFO SparkContext: Created broadcast 7 from broadcast at
DAGScheduler.scala:1006
19/03/12 21:23:44 INFO DAGScheduler: Submitting 1 missing tasks from
ResultStage 7 (ShuffledRDD[9] at reduceByKey at HDFSWordCount.scala:43)
(first 15 tasks are for partitions Vector(0))
19/03/12 21:23:44 INFO TaskSchedulerImpl: Adding task set 7.0 with 1 tasks
19/03/12 21:23:44 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID
5, localhost, executor driver, partition 0, ANY, 4621 bytes)
19/03/12 21:23:44 INFO Executor: Running task 0.0 in stage 7.0 (TID 5)
19/03/12 21:23:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty
blocks out of 1 blocks
19/03/12 21:23:44 INFO ShuffleBlockFetcherIterator: Started 0 remote
fetches in 0 ms
19/03/12 21:23:44 INFO Executor: Finished task 0.0 in stage 7.0 (TID 5).
1393 bytes result sent to driver
19/03/12 21:23:44 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID
5) in 68 ms on localhost (executor driver) (1/1)
19/03/12 21:23:44 INFO DAGScheduler: ResultStage 7 (collect at
HDFSWordCount.scala:58) finished in 0.065 s
19/03/12 21:23:44 INFO DAGScheduler: Job 3 finished: collect at
HDFSWordCount.scala:58, took 0.122533 s
(are,3)(is,2)(how,5)(u,3)(amit,2)(,1)(and,1)(shravan,2)(yogi,2)19/03/12
21:23:44 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all
completed, from pool
19/03/12 21:23:44 INFO SparkUI: Stopped Spark web UI at
http://192.168.0.107:4040
19/03/12 21:23:44 INFO MapOutputTrackerMasterEndpoint:
```

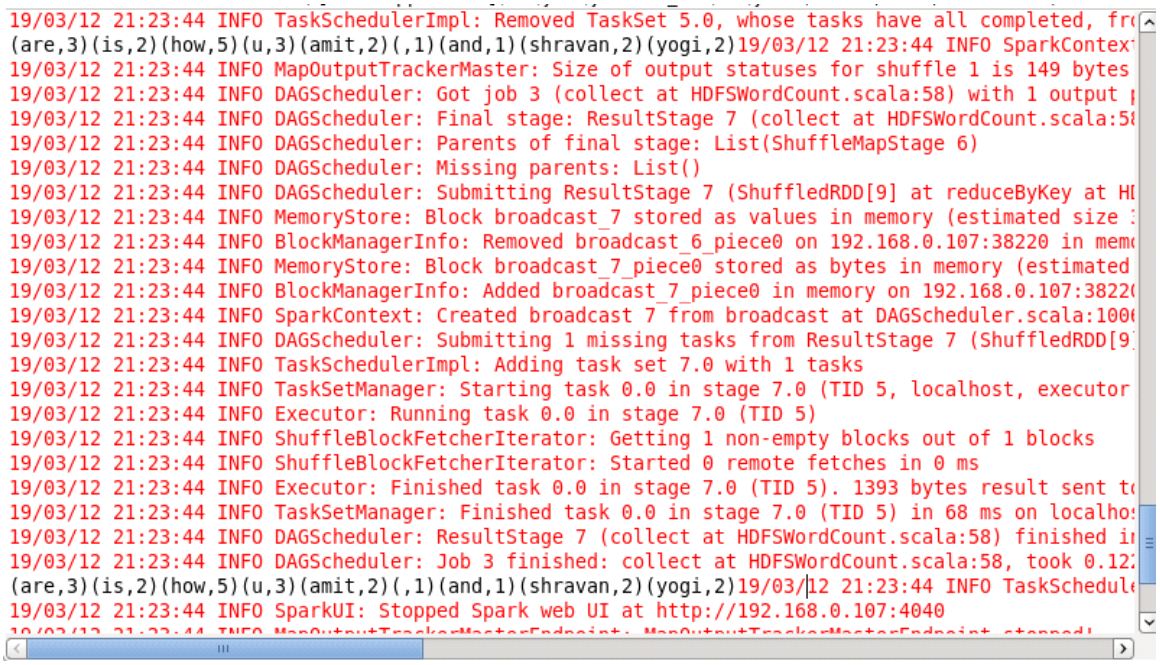
```
MapOutputTrackerMasterEndpoint stopped!
19/03/12 21:23:44 INFO MemoryStore: MemoryStore cleared
19/03/12 21:23:44 INFO BlockManager: BlockManager stopped
19/03/12 21:23:44 INFO BlockManagerMaster: BlockManagerMaster stopped
19/03/12 21:23:44 INFO
OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
OutputCommitCoordinator stopped!
19/03/12 21:23:44 INFO SparkContext: Successfully stopped SparkContext
19/03/12 21:23:44 INFO ShutdownHookManager: Shutdown hook called
19/03/12 21:23:44 INFO ShutdownHookManager: Deleting directory
/tmp/spark-1f1e3497-a474-4fa5-9a5b-6c6f28ddd9b6
```

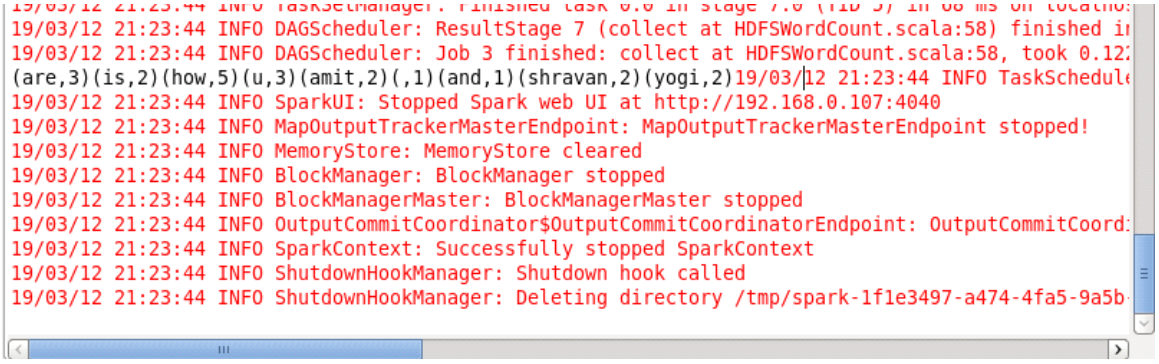
OUTPUT:




19/03/12 21:23:26 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.16
19/03/12 21:23:26 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.0.107:38220
19/03/12 21:23:26 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168
19/03/12 21:23:26 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.0.1
19/03/12 21:23:30 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size
19/03/12 21:23:31 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated
19/03/12 21:23:31 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 192.168.0.107:3822
19/03/12 21:23:31 INFO SparkContext: Created broadcast 0 from textFile at HDFSWordCount.scala:22
Such destination existshdfs://localhost:8020/
19/03/12 21:23:36 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size
19/03/12 21:23:36 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated
19/03/12 21:23:36 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 192.168.0.107:3822
19/03/12 21:23:36 INFO SparkContext: Created broadcast 1 from textFile at HDFSWordCount.scala:39
19/03/12 21:23:36 INFO FileInputFormat: Total input paths to process : 1
19/03/12 21:23:37 INFO FileInputFormat: Total input paths to process : 1
19/03/12 21:23:37 INFO SparkContext: Starting job: saveAsTextFile at HDFSWordCount.scala:45
19/03/12 21:23:38 INFO DAGScheduler: Registering RDD 6 (map at HDFSWordCount.scala:42)
19/03/12 21:23:38 INFO DAGScheduler: Got job 0 (saveAsTextFile at HDFSWordCount.scala:45) with 1
19/03/12 21:23:38 INFO DAGScheduler: Final stage: ResultStage 1 (saveAsTextFile at HDFSWordCount.
19/03/12 21:23:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
19/03/12 21:23:38 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 0)
19/03/12 21:23:38 INFO DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[6] at map at
19/03/12 21:23:39 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size
19/03/12 21:23:39 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated
19/03/12 21:23:39 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 192.168.0.107:3822


19/03/12 21:23:43 INFO DAGScheduler: Job 1 finished: saveAsTextFile at HDFSWordCount.scala:46, to
19/03/12 21:23:43 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, fr
Contents match!
19/03/12 21:23:43 INFO SparkContext: Starting job: collect at HDFSWordCount.scala:57
19/03/12 21:23:43 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 149 bytes
19/03/12 21:23:44 INFO DAGScheduler: Got job 2 (collect at HDFSWordCount.scala:57) with 1 output
19/03/12 21:23:44 INFO DAGScheduler: Final stage: ResultStage 5 (collect at HDFSWordCount.scala:5
19/03/12 21:23:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 4)
19/03/12 21:23:44 INFO DAGScheduler: Missing parents: List()
19/03/12 21:23:44 INFO DAGScheduler: Submitting ResultStage 5 (ShuffledRDD[7] at reduceByKey at HD
19/03/12 21:23:44 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size
19/03/12 21:23:44 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated
19/03/12 21:23:44 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on 192.168.0.107:3822
19/03/12 21:23:44 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:100
19/03/12 21:23:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (ShuffledRDD[7
19/03/12 21:23:44 INFO TaskSchedulerImpl: Adding task set 5.0 with 1 tasks
19/03/12 21:23:44 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 4, localhost, executor
19/03/12 21:23:44 INFO Executor: Running task 0.0 in stage 5.0 (TID 4)
19/03/12 21:23:44 INFO BlockManagerInfo: Removed broadcast_5_piece0 on 192.168.0.107:38220 in mem
19/03/12 21:23:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
19/03/12 21:23:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
19/03/12 21:23:44 INFO Executor: Finished task 0.0 in stage 5.0 (TID 4). 1393 bytes result sent t
19/03/12 21:23:44 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 4) in 120 ms on localh
19/03/12 21:23:44 INFO DAGScheduler: ResultStage 5 (collect at HDFSWordCount.scala:57) finished i
19/03/12 21:23:44 INFO DAGScheduler: Job 2 finished: collect at HDFSWordCount.scala:57, took 0.26
19/03/12 21:23:44 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, fr





EXPLANATION: NOW WE CAN CHECK WHETHER THE WORD COUNT DONE IN BOTH LOCAL AND HDFS ARE SAME OR NOT. SO FOR THAT WE CAN DO AS BELOWS. THE FILE "WordCount2" REPRESENTS THE WORD COUNT DONE IN FILE LOCAL DIRECTORY AND THE FILE "WordCount3" SHOWS THE OUTPUT OF THE WORD COUNT DONE IN HDFS AND ITS OUTPUT IS SAVED IN THE SAME FILE PRESENT IN LOCAL DIRECTORY. SO FROM THE SCREENSHOT WE CAN BOTH THE FILES WORD COUNT MATCHES.

```
[acadgild@localhost ~]$ cd WordCount2
[acadgild@localhost WordCount2]$ ls
part-00000  _SUCCESS
[acadgild@localhost WordCount2]$ cat part-00000
(are,3)
(is,2)
(how,5)
(u,3)
(amit,2)
(,1)
(and,1)
(shravan,2)
(yogi,2)
```

```
[acadgild@localhost ~]$ cd WordCount3
[acadgild@localhost WordCount3]$ cat part-00000
(are,3)
(is,2)
(how,5)
(u,3)
(amit,2)
(,1)
(and,1)
(shravan,2)
(yogi,2)
```