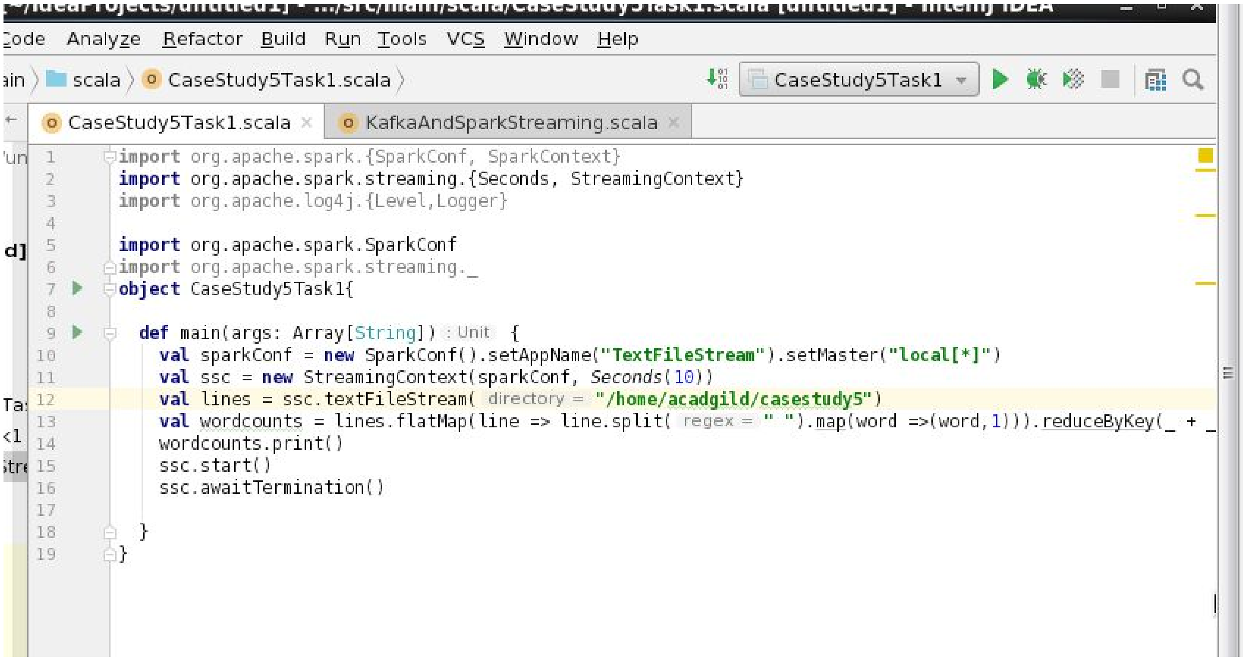
1. There are two parts this case study

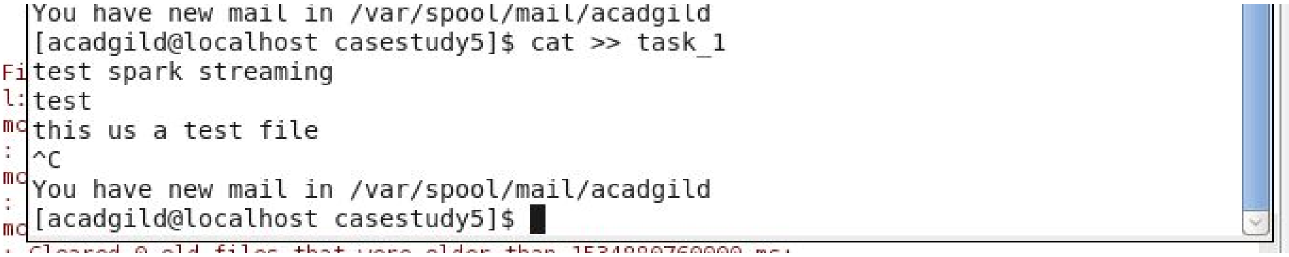
- First Part- You have to create a Spark Application which streams data from a file on local directory on your machine and does the word count on the fly. The word should be done by the spark application in such a way that as soon as you drop the file in your local directory, your spark application should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

Code:


~/IdeaProjects/untitled1] - .../src/main/scala/CaseStudy5Task1.scala [untitled1] - IntelliJ IDEA
Code Analyze Refactor Build Run Tools VCS Window Help
ain scala CaseStudy5Task1.scala
CaseStudy5Task1
CaseStudy5Task1.scala
KafkaAndSparkStreaming.scala

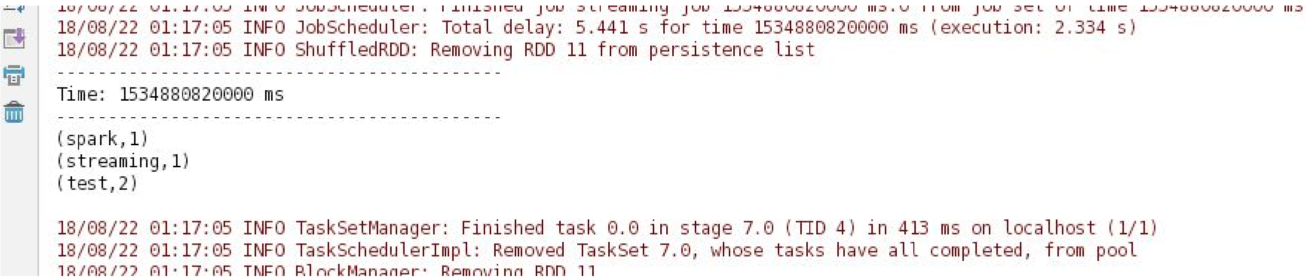
```
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.log4j.{Level,Logger}

import org.apache.spark.SparkConf
import org.apache.spark.streaming._
object CaseStudy5Task1{

  def main(args: Array[String]) : Unit  {
    val sparkConf = new SparkConf().setAppName("TextFileStream").setMaster("local[*]")
    val ssc = new StreamingContext(sparkConf, Seconds(10))
    val lines = ssc.textFileStream( directory = "/home/acadgild/casestudy5")
    val wordcounts = lines.flatMap(line => line.split( regex = " ").map(word =>(word,1))).reduceByKey(_ + _
    wordcounts.print()
    ssc.start()
    ssc.awaitTermination()

  }
}
```

Input :



```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost casestudy5]$ cat >> task_1
test spark streaming
test
this us a test file
^C
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost casestudy5]$ 
```

Output

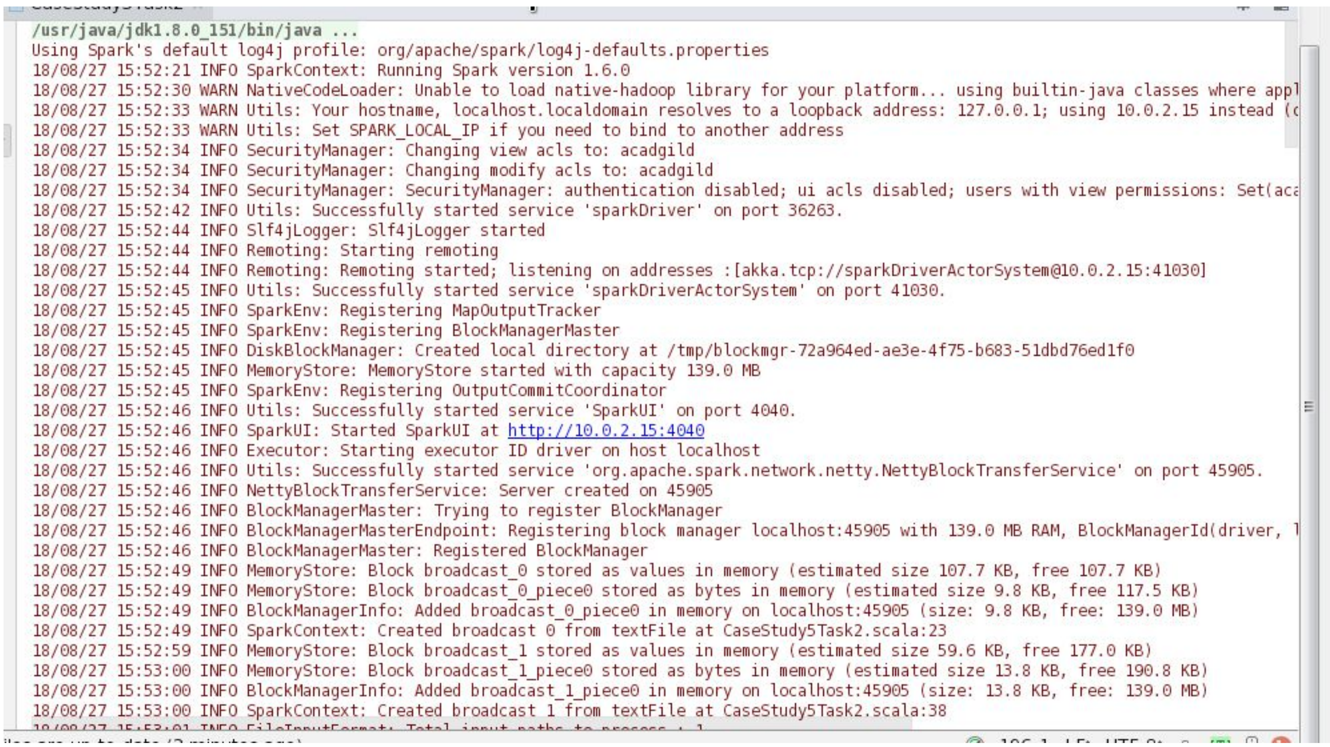


```
18/08/22 01:17:05 INFO JobScheduler: Finished job streaming job 1534880820000 ms.0 from job set of time 1534880820000 ms
18/08/22 01:17:05 INFO JobScheduler: Total delay: 5.441 s for time 1534880820000 ms (execution: 2.334 s)
18/08/22 01:17:05 INFO ShuffledRDD: Removing RDD 11 from persistence list
-------------------------------------------
Time: 1534880820000 ms
-------------------------------------------
(spark,1)
(streaming,1)
(test,2)

18/08/22 01:17:05 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 4) in 413 ms on localhost (1/1)
18/08/22 01:17:05 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
18/08/22 01:17:05 INFO BlockManager: Removing RDD 11
```

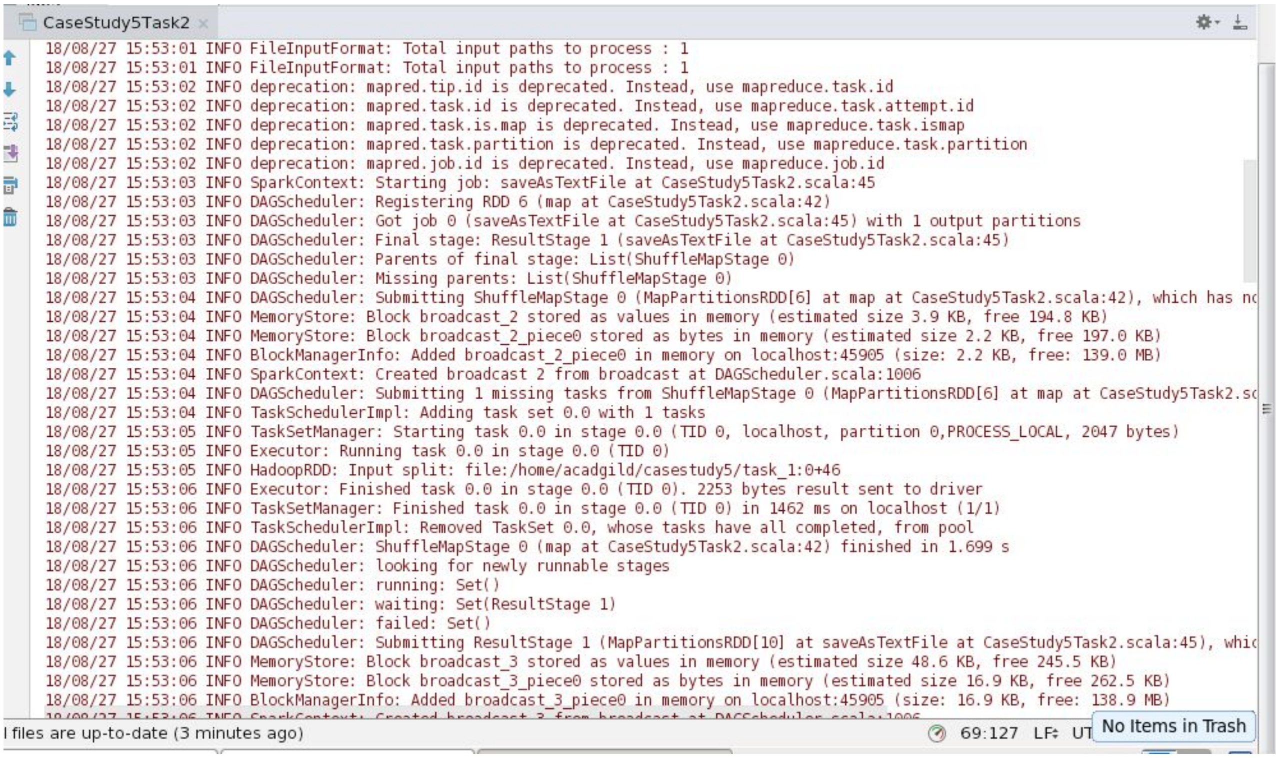
- Second Part - In this part, you will have to create a Spark Application which should do the following

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

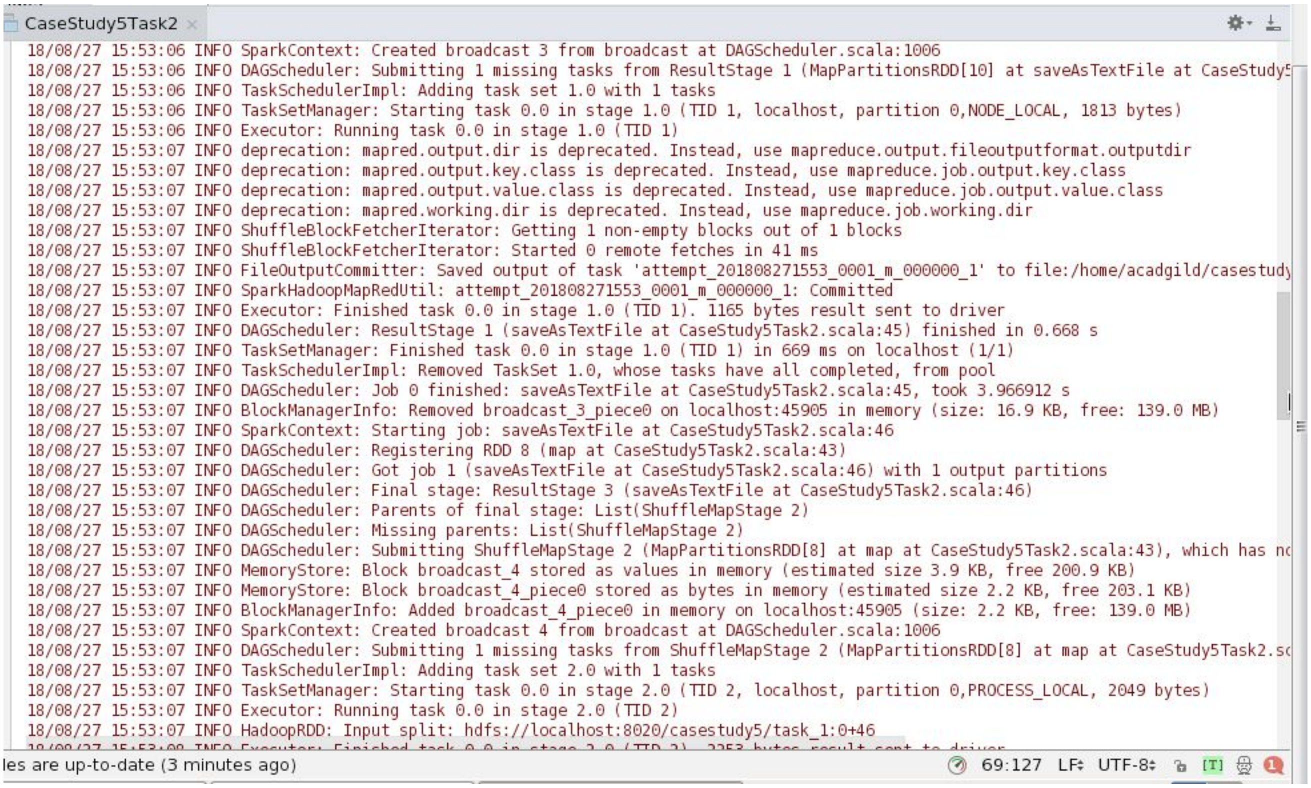
Output:



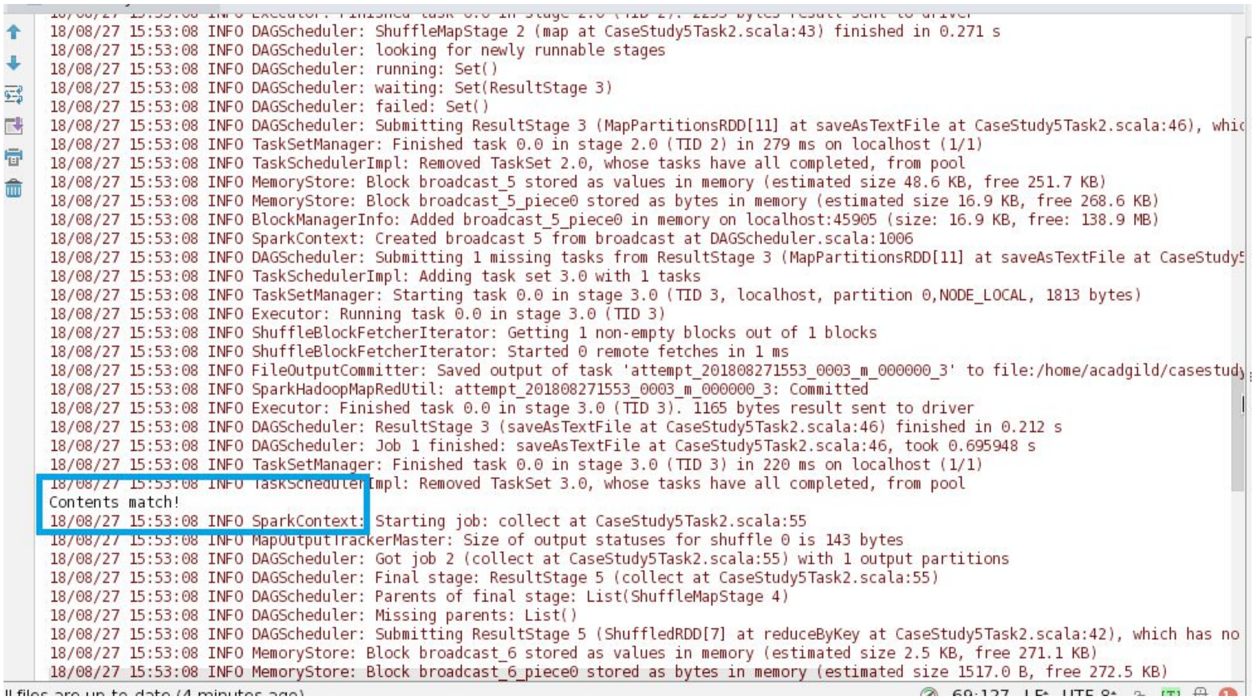
```
/usr/java/jdk1.8.0_151/bin/java ...
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
18/08/27 15:52:21 INFO SparkContext: Running Spark version 1.6.0
18/08/27 15:52:30 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appl
18/08/27 15:52:33 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 10.0.2.15 instead (
18/08/27 15:52:33 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/08/27 15:52:34 INFO SecurityManager: Changing view acls to: acadgild
18/08/27 15:52:34 INFO SecurityManager: Changing modify acls to: acadgild
18/08/27 15:52:34 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(aca
18/08/27 15:52:42 INFO Utils: Successfully started service 'sparkDriver' on port 36263.
18/08/27 15:52:44 INFO Slf4jLogger: Slf4jLogger started
18/08/27 15:52:44 INFO Remoting: Starting remoting
18/08/27 15:52:44 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:41030]
18/08/27 15:52:45 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 41030.
18/08/27 15:52:45 INFO SparkEnv: Registering MapOutputTracker
18/08/27 15:52:45 INFO SparkEnv: Registering BlockManagerMaster
18/08/27 15:52:45 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-72a964ed-ae3e-4f75-b683-51dbd76ed1f0
18/08/27 15:52:45 INFO MemoryStore: MemoryStore started with capacity 139.0 MB
18/08/27 15:52:45 INFO SparkEnv: Registering OutputCommitCoordinator
18/08/27 15:52:46 INFO Utils: Successfully started service 'SparkUI' on port 4040.
18/08/27 15:52:46 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
18/08/27 15:52:46 INFO Executor: Starting executor ID driver on host localhost
18/08/27 15:52:46 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 45905.
18/08/27 15:52:46 INFO NettyBlockTransferService: Server created on 45905
18/08/27 15:52:46 INFO BlockManagerMaster: Trying to register BlockManager
18/08/27 15:52:46 INFO BlockManagerMasterEndpoint: Registering block manager localhost:45905 with 139.0 MB RAM, BlockManagerId(driver, l
18/08/27 15:52:46 INFO BlockManagerMaster: Registered BlockManager
18/08/27 15:52:49 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 107.7 KB, free 107.7 KB)
18/08/27 15:52:49 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 9.8 KB, free 117.5 KB)
18/08/27 15:52:49 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:45905 (size: 9.8 KB, free: 139.0 MB)
18/08/27 15:52:49 INFO SparkContext: Created broadcast 0 from textFile at CaseStudy5Task2.scala:23
18/08/27 15:52:59 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 59.6 KB, free 177.0 KB)
18/08/27 15:53:00 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 13.8 KB, free 190.8 KB)
18/08/27 15:53:00 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:45905 (size: 13.8 KB, free: 139.0 MB)
18/08/27 15:53:00 INFO SparkContext: Created broadcast 1 from textFile at CaseStudy5Task2.scala:38
```



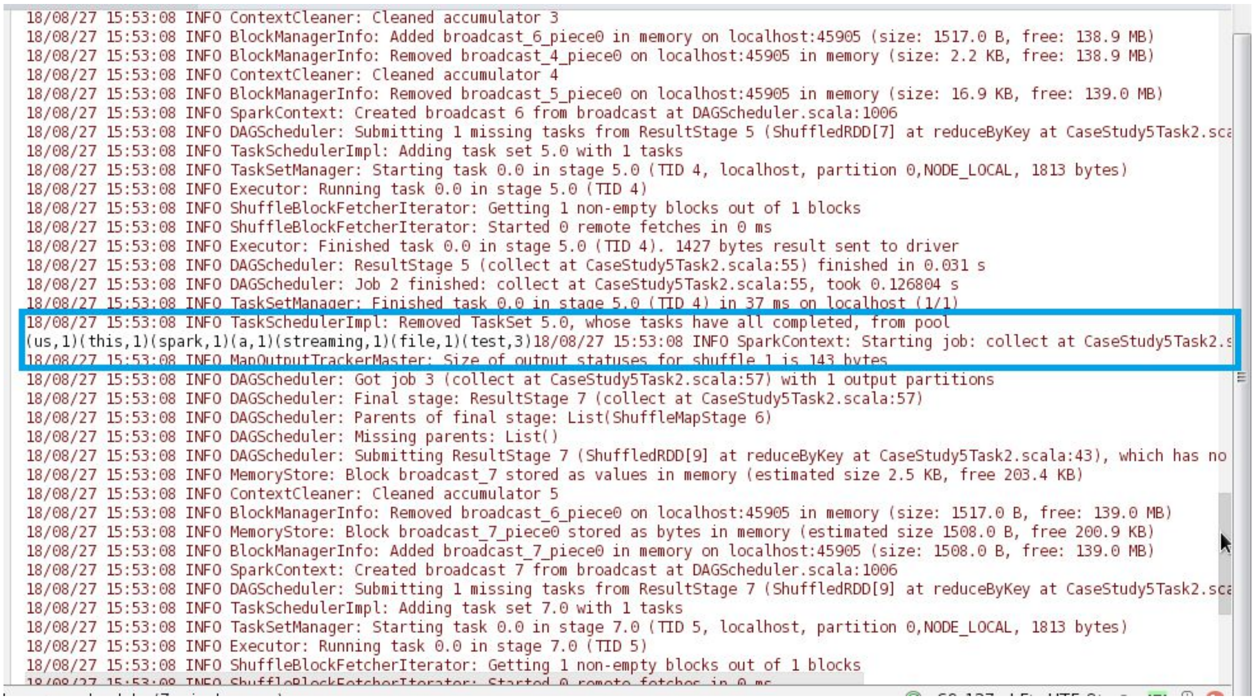
```
18/08/27 15:53:01 INFO FileInputFormat: Total input paths to process : 1
18/08/27 15:53:01 INFO FileInputFormat: Total input paths to process : 1
18/08/27 15:53:02 INFO deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
18/08/27 15:53:02 INFO deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
18/08/27 15:53:02 INFO deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
18/08/27 15:53:02 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
18/08/27 15:53:02 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
18/08/27 15:53:03 INFO SparkContext: Starting job: saveAsTextFile at CaseStudy5Task2.scala:45
18/08/27 15:53:03 INFO DAGScheduler: Registering RDD 6 (map at CaseStudy5Task2.scala:42)
18/08/27 15:53:03 INFO DAGScheduler: Got job 0 (saveAsTextFile at CaseStudy5Task2.scala:45) with 1 output partitions
18/08/27 15:53:03 INFO DAGScheduler: Final stage: ResultStage 1 (saveAsTextFile at CaseStudy5Task2.scala:45)
18/08/27 15:53:03 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
18/08/27 15:53:03 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 0)
18/08/27 15:53:04 INFO DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[6] at map at CaseStudy5Task2.scala:42), which has nc
18/08/27 15:53:04 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 3.9 KB, free 194.8 KB)
18/08/27 15:53:04 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 2.2 KB, free 197.0 KB)
18/08/27 15:53:04 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:45905 (size: 2.2 KB, free: 139.0 MB)
18/08/27 15:53:04 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006
18/08/27 15:53:04 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[6] at map at CaseStudy5Task2.sc
18/08/27 15:53:04 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
18/08/27 15:53:05 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,PROCESS_LOCAL, 2047 bytes)
18/08/27 15:53:05 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
18/08/27 15:53:05 INFO HadoopRDD: Input split: file:/home/acadgild/casestudy5/task_1:0+46
18/08/27 15:53:06 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 2253 bytes result sent to driver
18/08/27 15:53:06 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1462 ms on localhost (1/1)
18/08/27 15:53:06 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/08/27 15:53:06 INFO DAGScheduler: ShuffleMapStage 0 (map at CaseStudy5Task2.scala:42) finished in 1.699 s
18/08/27 15:53:06 INFO DAGScheduler: looking for newly runnable stages
18/08/27 15:53:06 INFO DAGScheduler: running: Set()
18/08/27 15:53:06 INFO DAGScheduler: waiting: Set(ResultStage 1)
18/08/27 15:53:06 INFO DAGScheduler: failed: Set()
18/08/27 15:53:06 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[10] at saveAsTextFile at CaseStudy5Task2.scala:45), whic
18/08/27 15:53:06 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 48.6 KB, free 245.5 KB)
18/08/27 15:53:06 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 16.9 KB, free 262.5 KB)
18/08/27 15:53:06 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:45905 (size: 16.9 KB, free: 138.9 MB)
```



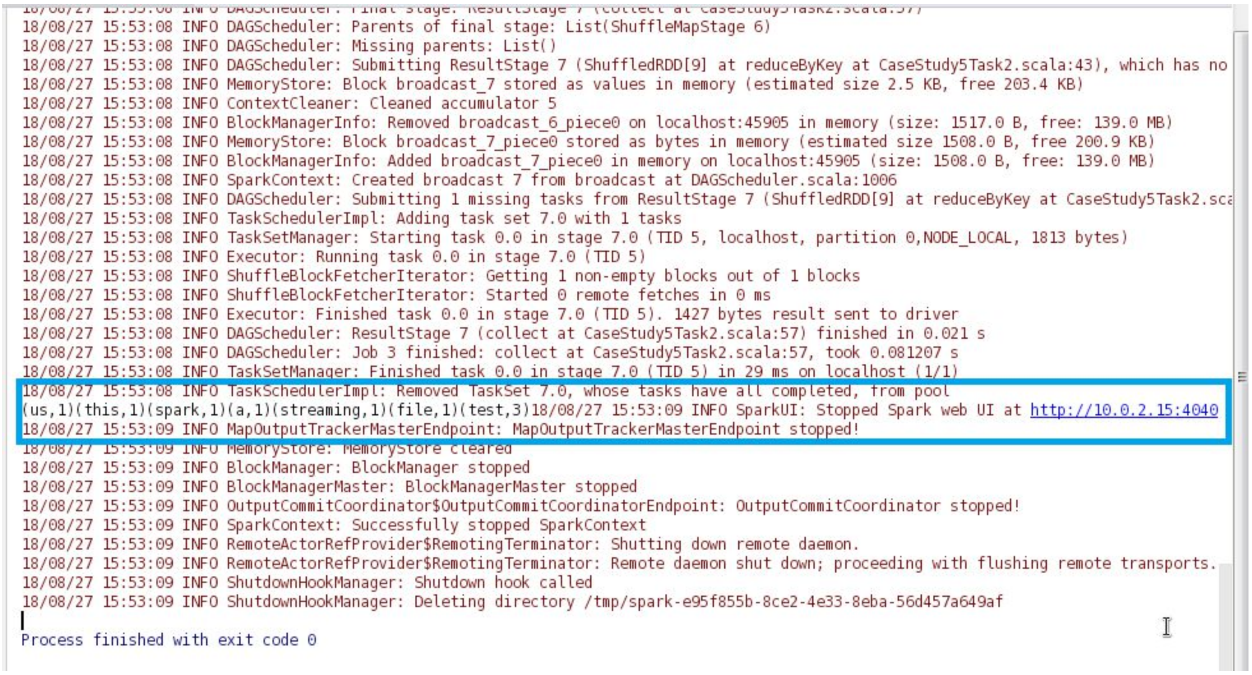
```
18/08/27 15:53:06 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1006
18/08/27 15:53:06 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[10] at saveAsTextFile at CaseStudy5
18/08/27 15:53:06 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
18/08/27 15:53:06 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, partition 0,NODE_LOCAL, 1813 bytes)
18/08/27 15:53:06 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
18/08/27 15:53:07 INFO deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
18/08/27 15:53:07 INFO deprecation: mapred.output.key.class is deprecated. Instead, use mapreduce.job.output.key.class
18/08/27 15:53:07 INFO deprecation: mapred.output.value.class is deprecated. Instead, use mapreduce.job.output.value.class
18/08/27 15:53:07 INFO deprecation: mapred.working.dir is deprecated. Instead, use mapreduce.job.working.dir
18/08/27 15:53:07 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/27 15:53:07 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 41 ms
18/08/27 15:53:07 INFO FileOutputCommitter: Saved output of task 'attempt_201808271553_0001_m_000000_1' to file:/home/acadgild/casestudy
18/08/27 15:53:07 INFO SparkHadoopMapRedUtil: attempt_201808271553_0001_m_000000_1: Committed
18/08/27 15:53:07 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1165 bytes result sent to driver
18/08/27 15:53:07 INFO DAGScheduler: ResultStage 1 (saveAsTextFile at CaseStudy5Task2.scala:45) finished in 0.668 s
18/08/27 15:53:07 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 669 ms on localhost (1/1)
18/08/27 15:53:07 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/08/27 15:53:07 INFO DAGScheduler: Job 0 finished: saveAsTextFile at CaseStudy5Task2.scala:45, took 3.966912 s
18/08/27 15:53:07 INFO BlockManagerInfo: Removed broadcast_3_piece0 on localhost:45905 in memory (size: 16.9 KB, free: 139.0 MB)
18/08/27 15:53:07 INFO SparkContext: Starting job: saveAsTextFile at CaseStudy5Task2.scala:46
18/08/27 15:53:07 INFO DAGScheduler: Registering RDD 8 (map at CaseStudy5Task2.scala:43)
18/08/27 15:53:07 INFO DAGScheduler: Got job 1 (saveAsTextFile at CaseStudy5Task2.scala:46) with 1 output partitions
18/08/27 15:53:07 INFO DAGScheduler: Final stage: ResultStage 3 (saveAsTextFile at CaseStudy5Task2.scala:46)
18/08/27 15:53:07 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 2)
18/08/27 15:53:07 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 2)
18/08/27 15:53:07 INFO DAGScheduler: Submitting ShuffleMapStage 2 (MapPartitionsRDD[8] at map at CaseStudy5Task2.scala:43), which has nc
18/08/27 15:53:07 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 3.9 KB, free 200.9 KB)
18/08/27 15:53:07 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 2.2 KB, free 203.1 KB)
18/08/27 15:53:07 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:45905 (size: 2.2 KB, free: 139.0 MB)
18/08/27 15:53:07 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1006
18/08/27 15:53:07 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 2 (MapPartitionsRDD[8] at map at CaseStudy5Task2.sc
18/08/27 15:53:07 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
18/08/27 15:53:07 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, partition 0,PROCESS_LOCAL, 2049 bytes)
18/08/27 15:53:07 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
18/08/27 15:53:07 INFO HadoopRDD: Input split: hdfs://localhost:8020/casestudy5/task_1:0+46
```



```
18/08/27 15:53:08 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 2235 bytes result sent to driver
18/08/27 15:53:08 INFO DAGScheduler: ShuffleMapStage 2 (map at CaseStudy5Task2.scala:43) finished in 0.271 s
18/08/27 15:53:08 INFO DAGScheduler: looking for newly runnable stages
18/08/27 15:53:08 INFO DAGScheduler: running: Set()
18/08/27 15:53:08 INFO DAGScheduler: waiting: Set(ResultStage 3)
18/08/27 15:53:08 INFO DAGScheduler: failed: Set()
18/08/27 15:53:08 INFO DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[11] at saveAsTextFile at CaseStudy5Task2.scala:46), whic
18/08/27 15:53:08 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 279 ms on localhost (1/1)
18/08/27 15:53:08 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 48.6 KB, free 251.7 KB)
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 16.9 KB, free 268.6 KB)
18/08/27 15:53:08 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on localhost:45905 (size: 16.9 KB, free: 138.9 MB)
18/08/27 15:53:08 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1006
18/08/27 15:53:08 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[11] at saveAsTextFile at CaseStudy5
18/08/27 15:53:08 INFO TaskSchedulerImpl: Adding task set 3.0 with 1 tasks
18/08/27 15:53:08 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, localhost, partition 0,NODE_LOCAL, 1813 bytes)
18/08/27 15:53:08 INFO Executor: Running task 0.0 in stage 3.0 (TID 3)
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/08/27 15:53:08 INFO FileOutputCommitter: Saved output of task 'attempt_201808271553_0003_m_000000_3' to file:/home/acadgild/casestudy
18/08/27 15:53:08 INFO SparkHadoopMapRedUtil: attempt_201808271553_0003_m_000000_3: Committed
18/08/27 15:53:08 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 1165 bytes result sent to driver
18/08/27 15:53:08 INFO DAGScheduler: ResultStage 3 (saveAsTextFile at CaseStudy5Task2.scala:46) finished in 0.212 s
18/08/27 15:53:08 INFO DAGScheduler: Job 1 finished: saveAsTextFile at CaseStudy5Task2.scala:46, took 0.695948 s
18/08/27 15:53:08 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 220 ms on localhost (1/1)
18/08/27 15:53:08 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
Contents match!
18/08/27 15:53:08 INFO SparkContext: Starting job: collect at CaseStudy5Task2.scala:55
18/08/27 15:53:08 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 143 bytes
18/08/27 15:53:08 INFO DAGScheduler: Got job 2 (collect at CaseStudy5Task2.scala:55) with 1 output partitions
18/08/27 15:53:08 INFO DAGScheduler: Final stage: ResultStage 5 (collect at CaseStudy5Task2.scala:55)
18/08/27 15:53:08 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 4)
18/08/27 15:53:08 INFO DAGScheduler: Missing parents: List()
18/08/27 15:53:08 INFO DAGScheduler: Submitting ResultStage 5 (ShuffledRDD[7] at reduceByKey at CaseStudy5Task2.scala:42), which has no
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 2.5 KB, free 271.1 KB)
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 1517.0 B, free 272.5 KB)
```



```
18/08/27 15:53:08 INFO ContextCleaner: Cleaned accumulator 3
18/08/27 15:53:08 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on localhost:45905 (size: 1517.0 B, free: 138.9 MB)
18/08/27 15:53:08 INFO BlockManagerInfo: Removed broadcast_4_piece0 on localhost:45905 in memory (size: 2.2 KB, free: 138.9 MB)
18/08/27 15:53:08 INFO ContextCleaner: Cleaned accumulator 4
18/08/27 15:53:08 INFO BlockManagerInfo: Removed broadcast_5_piece0 on localhost:45905 in memory (size: 16.9 KB, free: 139.0 MB)
18/08/27 15:53:08 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1006
18/08/27 15:53:08 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (ShuffledRDD[7] at reduceByKey at CaseStudy5Task2.sca
18/08/27 15:53:08 INFO TaskSchedulerImpl: Adding task set 5.0 with 1 tasks
18/08/27 15:53:08 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 4, localhost, partition 0,NODE_LOCAL, 1813 bytes)
18/08/27 15:53:08 INFO Executor: Running task 0.0 in stage 5.0 (TID 4)
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/08/27 15:53:08 INFO Executor: Finished task 0.0 in stage 5.0 (TID 4). 1427 bytes result sent to driver
18/08/27 15:53:08 INFO DAGScheduler: ResultStage 5 (collect at CaseStudy5Task2.scala:55) finished in 0.031 s
18/08/27 15:53:08 INFO DAGScheduler: Job 2 finished: collect at CaseStudy5Task2.scala:55, took 0.126804 s
18/08/27 15:53:08 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 4) in 37 ms on localhost (1/1)
18/08/27 15:53:08 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
(us,1)(this,1)(spark,1)(a,1)(streaming,1)(file,1)(test,3)18/08/27 15:53:08 INFO SparkContext: Starting job: collect at CaseStudy5Task2.s
18/08/27 15:53:08 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 1 is 143 bytes
18/08/27 15:53:08 INFO DAGScheduler: Got job 3 (collect at CaseStudy5Task2.scala:57) with 1 output partitions
18/08/27 15:53:08 INFO DAGScheduler: Final stage: ResultStage 7 (collect at CaseStudy5Task2.scala:57)
18/08/27 15:53:08 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 6)
18/08/27 15:53:08 INFO DAGScheduler: Missing parents: List()
18/08/27 15:53:08 INFO DAGScheduler: Submitting ResultStage 7 (ShuffledRDD[9] at reduceByKey at CaseStudy5Task2.scala:43), which has no
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 2.5 KB, free 203.4 KB)
18/08/27 15:53:08 INFO ContextCleaner: Cleaned accumulator 5
18/08/27 15:53:08 INFO BlockManagerInfo: Removed broadcast_6_piece0 on localhost:45905 in memory (size: 1517.0 B, free: 139.0 MB)
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 1508.0 B, free 200.9 KB)
18/08/27 15:53:08 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on localhost:45905 (size: 1508.0 B, free: 139.0 MB)
18/08/27 15:53:08 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1006
18/08/27 15:53:08 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 7 (ShuffledRDD[9] at reduceByKey at CaseStudy5Task2.sca
18/08/27 15:53:08 INFO TaskSchedulerImpl: Adding task set 7.0 with 1 tasks
18/08/27 15:53:08 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 5, localhost, partition 0,NODE_LOCAL, 1813 bytes)
18/08/27 15:53:08 INFO Executor: Running task 0.0 in stage 7.0 (TID 5)
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```



```
18/08/27 15:53:08 INFO DAGScheduler: Final stage: ResultStage 7 (collect at CaseStudy5Task2.scala:57)
18/08/27 15:53:08 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 6)
18/08/27 15:53:08 INFO DAGScheduler: Missing parents: List()
18/08/27 15:53:08 INFO DAGScheduler: Submitting ResultStage 7 (ShuffledRDD[9] at reduceByKey at CaseStudy5Task2.scala:43), which has no
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 2.5 KB, free 203.4 KB)
18/08/27 15:53:08 INFO ContextCleaner: Cleaned accumulator 5
18/08/27 15:53:08 INFO BlockManagerInfo: Removed broadcast_6_piece0 on localhost:45905 in memory (size: 1517.0 B, free: 139.0 MB)
18/08/27 15:53:08 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 1508.0 B, free 200.9 KB)
18/08/27 15:53:08 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on localhost:45905 (size: 1508.0 B, free: 139.0 MB)
18/08/27 15:53:08 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1006
18/08/27 15:53:08 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 7 (ShuffledRDD[9] at reduceByKey at CaseStudy5Task2.sca
18/08/27 15:53:08 INFO TaskSchedulerImpl: Adding task set 7.0 with 1 tasks
18/08/27 15:53:08 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 5, localhost, partition 0,NODE_LOCAL, 1813 bytes)
18/08/27 15:53:08 INFO Executor: Running task 0.0 in stage 7.0 (TID 5)
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/27 15:53:08 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/08/27 15:53:08 INFO Executor: Finished task 0.0 in stage 7.0 (TID 5). 1427 bytes result sent to driver
18/08/27 15:53:08 INFO DAGScheduler: ResultStage 7 (collect at CaseStudy5Task2.scala:57) finished in 0.021 s
18/08/27 15:53:08 INFO DAGScheduler: Job 3 finished: collect at CaseStudy5Task2.scala:57, took 0.081207 s
18/08/27 15:53:08 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 5) in 29 ms on localhost (1/1)
18/08/27 15:53:08 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool 
(us,1)(this,1)(spark,1)(a,1)(streaming,1)(file,1)(test,3)18/08/27 15:53:09 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/08/27 15:53:09 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/08/27 15:53:09 INFO MemoryStore: MemoryStore cleared
18/08/27 15:53:09 INFO BlockManager: BlockManager stopped
18/08/27 15:53:09 INFO BlockManagerMaster: BlockManagerMaster stopped
18/08/27 15:53:09 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/08/27 15:53:09 INFO SparkContext: Successfully stopped SparkContext
18/08/27 15:53:09 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
18/08/27 15:53:09 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
18/08/27 15:53:09 INFO ShutdownHookManager: Shutdown hook called
18/08/27 15:53:09 INFO ShutdownHookManager: Deleting directory /tmp/spark-e95f855b-8ce2-4e33-8eba-56d457a649af

Process finished with exit code 0
```