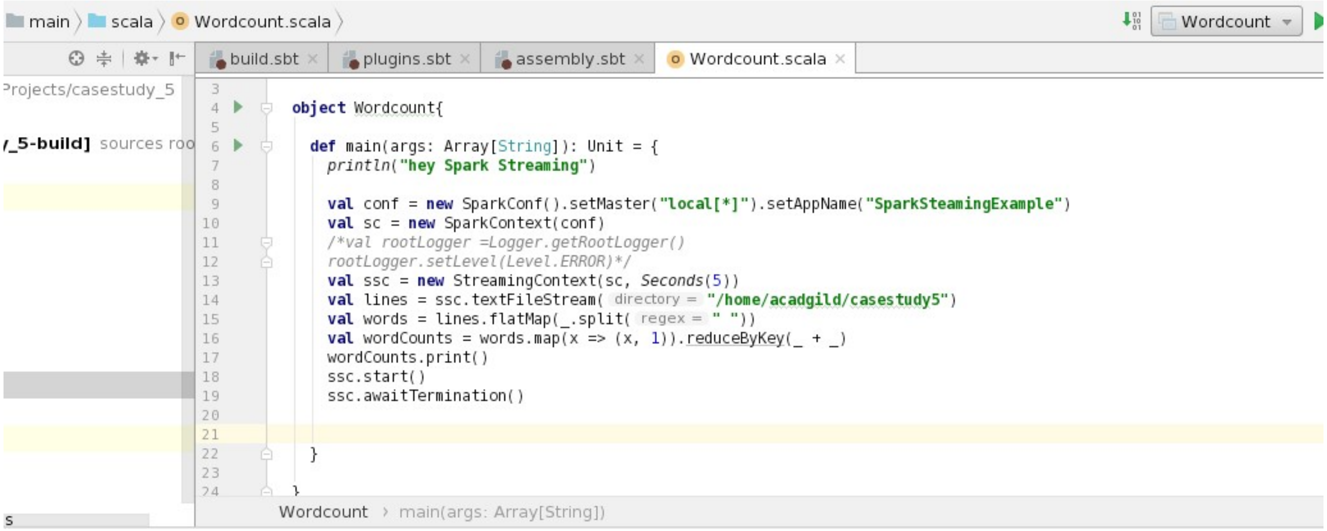
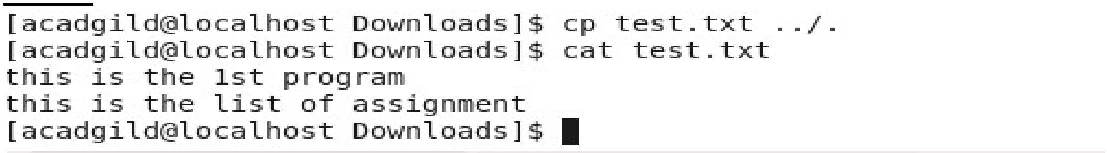
Case study -5

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.

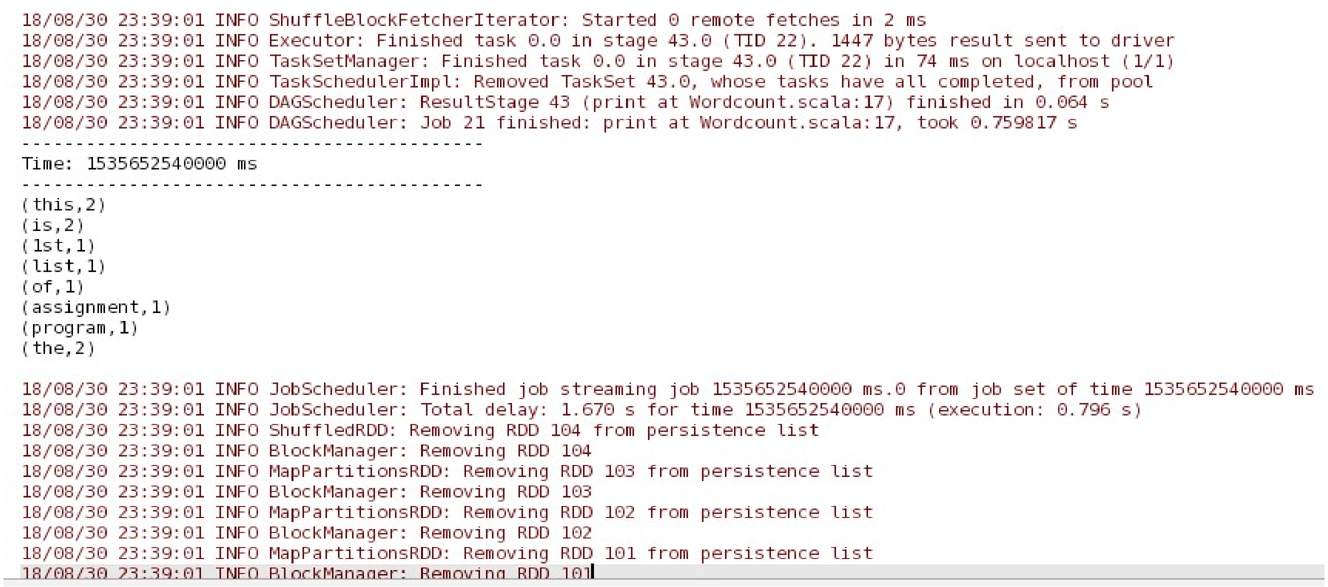


```
object Wordcount{

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

    val conf = new SparkConf().setMaster("local[*]").setAppName("SparkSteamingExample")
    val sc = new SparkContext(conf)
    /*val rootLogger =Logger.getRootLogger()
    rootLogger.setLevel(Level.ERROR)*/
    val ssc = new StreamingContext(sc, Seconds(5))
    val lines = ssc.textFileStream( directory = "/home/acadgild/casestudy5")
    val words = lines.flatMap(_.split( regex = " "))
    val wordCounts = words.map(x => (x, 1)).reduceByKey(_ + _)
    wordCounts.print()
    ssc.start()
    ssc.awaitTermination()


  }

}
```

Result



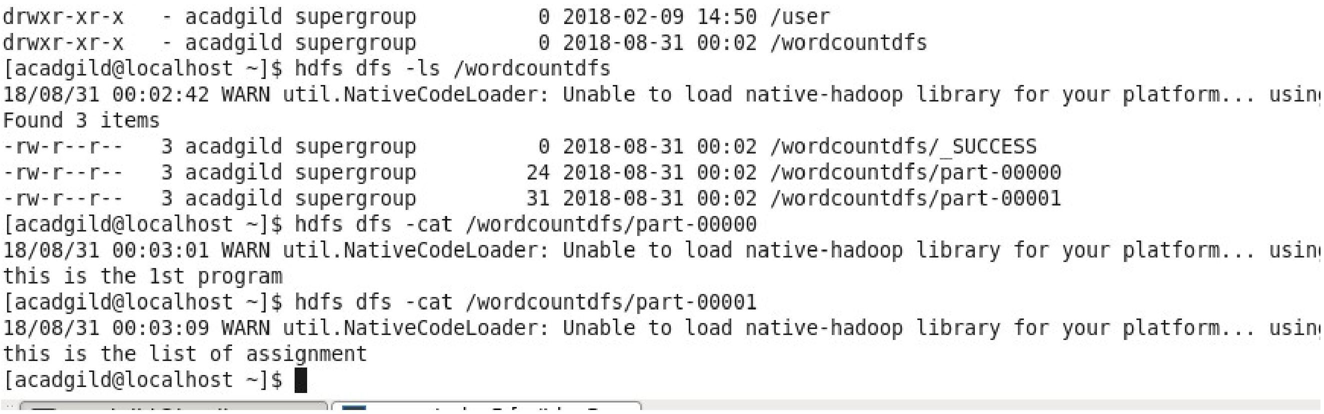
```
[acadgild@localhost Downloads]$ cp test.txt ../.
[acadgild@localhost Downloads]$ cat test.txt
this is the 1st program
this is the list of assignment
[acadgild@localhost Downloads]$
```



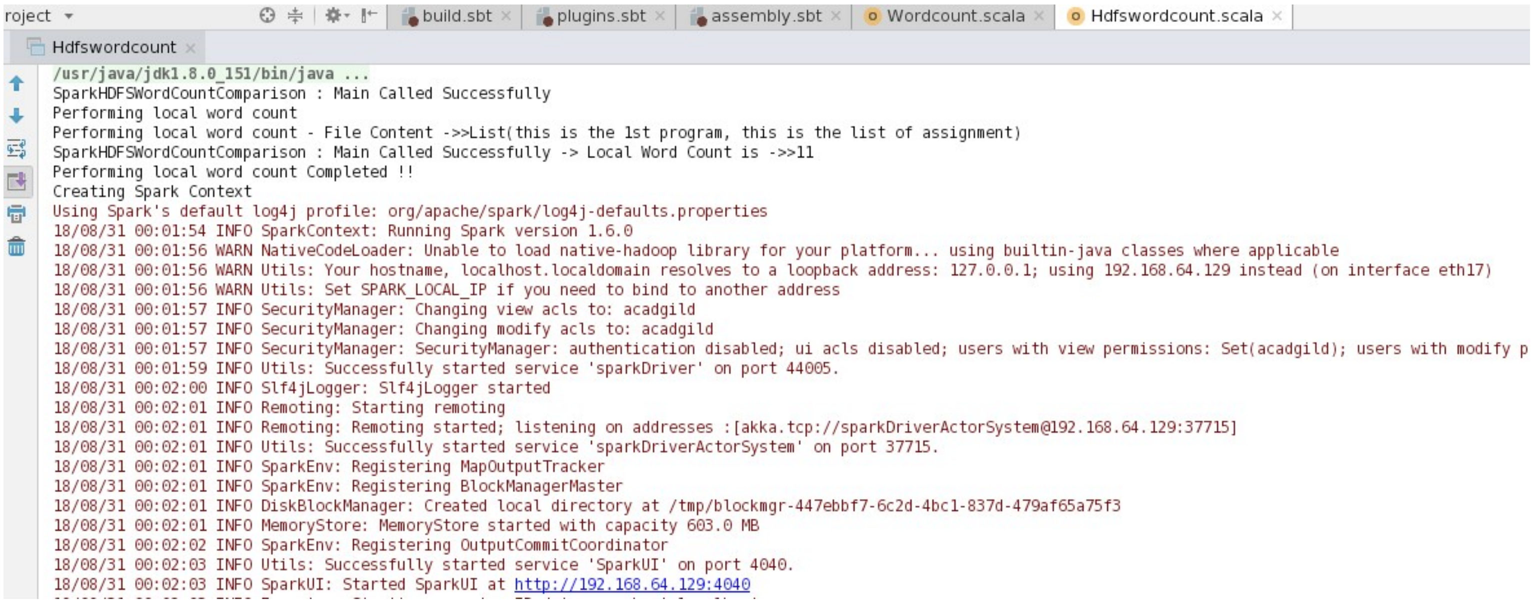
```
18/08/30 23:39:01 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 2 ms
18/08/30 23:39:01 INFO Executor: Finished task 0.0 in stage 43.0 (TID 22). 1447 bytes result sent to driver
18/08/30 23:39:01 INFO TaskSetManager: Finished task 0.0 in stage 43.0 (TID 22) in 74 ms on localhost (1/1)
18/08/30 23:39:01 INFO TaskSchedulerImpl: Removed TaskSet 43.0, whose tasks have all completed, from pool
18/08/30 23:39:01 INFO DAGScheduler: ResultStage 43 (print at Wordcount.scala:17) finished in 0.064 s
18/08/30 23:39:01 INFO DAGScheduler: Job 21 finished: print at Wordcount.scala:17, took 0.759817 s
-------------------------------------------
Time: 1535652540000 ms
-------------------------------------------
(this,2)
(is,2)
(1st,1)
(list,1)
(of,1)
(assignment,1)
(program,1)
(the,2)

18/08/30 23:39:01 INFO JobScheduler: Finished job streaming job 1535652540000 ms.0 from job set of time 1535652540000 ms
18/08/30 23:39:01 INFO JobScheduler: Total delay: 1.670 s for time 1535652540000 ms (execution: 0.796 s)
18/08/30 23:39:01 INFO ShuffledRDD: Removing RDD 104 from persistence list
18/08/30 23:39:01 INFO BlockManager: Removing RDD 104
18/08/30 23:39:01 INFO MapPartitionsRDD: Removing RDD 103 from persistence list
18/08/30 23:39:01 INFO BlockManager: Removing RDD 103
18/08/30 23:39:01 INFO MapPartitionsRDD: Removing RDD 102 from persistence list
18/08/30 23:39:01 INFO BlockManager: Removing RDD 102
18/08/30 23:39:01 INFO MapPartitionsRDD: Removing RDD 101 from persistence list
18/08/30 23:39:01 INFO BlockManager: Removing RDD 101
```

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

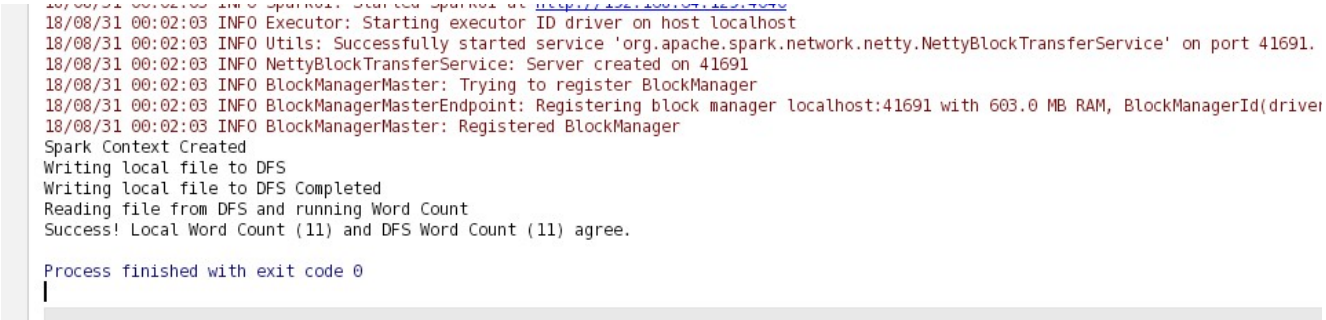
1. Pick up a file from the local directory and do the word count
2. Then in the same Spark Application, write the code to put the same file on HDFS.
3. Then in same Spark Application, do the word count of the file copied on HDFS in step 2
4. Lastly, compare the word count of step 1 and 2. Both should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



```
drwxr-xr-x   - acadgild supergroup          0 2018-02-09 14:50 /user
drwxr-xr-x   - acadgild supergroup          0 2018-08-31 00:02 /wordcountdfs
[acadgild@localhost ~]$ hdfs dfs -ls /wordcountdfs
18/08/31 00:02:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
Found 3 items
-rw-r--r--   3 acadgild supergroup          0 2018-08-31 00:02 /wordcountdfs/_SUCCESS
-rw-r--r--   3 acadgild supergroup         24 2018-08-31 00:02 /wordcountdfs/part-00000
-rw-r--r--   3 acadgild supergroup         31 2018-08-31 00:02 /wordcountdfs/part-00001
[acadgild@localhost ~]$ hdfs dfs -cat /wordcountdfs/part-00000
18/08/31 00:03:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
this is the 1st program
[acadgild@localhost ~]$ hdfs dfs -cat /wordcountdfs/part-00001
18/08/31 00:03:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
this is the list of assignment
[acadgild@localhost ~]$
```



```
/usr/java/jdk1.8.0_151/bin/java ...
SparkHDFSWordCountComparison : Main Called Successfully
Performing local word count
Performing local word count - File Content ->>List(this is the 1st program, this is the list of assignment)
SparkHDFSWordCountComparison : Main Called Successfully -> Local Word Count is ->>11
Performing local word count Completed !!
Creating Spark Context
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
18/08/31 00:01:54 INFO SparkContext: Running Spark version 1.6.0
18/08/31 00:01:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/08/31 00:01:56 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.64.129 instead (on interface eth17)
18/08/31 00:01:56 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/08/31 00:01:57 INFO SecurityManager: Changing view acls to: acadgild
18/08/31 00:01:57 INFO SecurityManager: Changing modify acls to: acadgild
18/08/31 00:01:57 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(acadgild); users with modify p
18/08/31 00:01:59 INFO Utils: Successfully started service 'sparkDriver' on port 44005.
18/08/31 00:02:00 INFO Slf4jLogger: Slf4jLogger started
18/08/31 00:02:01 INFO Remoting: Starting remoting
18/08/31 00:02:01 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@192.168.64.129:37715]
18/08/31 00:02:01 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 37715.
18/08/31 00:02:01 INFO SparkEnv: Registering MapOutputTracker
18/08/31 00:02:01 INFO SparkEnv: Registering BlockManagerMaster
18/08/31 00:02:01 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-447ebbf7-6c2d-4bc1-837d-479af65a75f3
18/08/31 00:02:01 INFO MemoryStore: MemoryStore started with capacity 603.0 MB
18/08/31 00:02:02 INFO SparkEnv: Registering OutputCommitCoordinator
18/08/31 00:02:03 INFO Utils: Successfully started service 'SparkUI' on port 4040.
18/08/31 00:02:03 INFO SparkUI: Started SparkUI at http://192.168.64.129:4040
```



```
18/08/31 00:02:03 INFO SparkUI: Started SparkUI at http://192.168.64.129:4040
18/08/31 00:02:03 INFO Executor: Starting executor ID driver on host localhost
18/08/31 00:02:03 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 41691.
18/08/31 00:02:03 INFO NettyBlockTransferService: Server created on 41691
18/08/31 00:02:03 INFO BlockManagerMaster: Trying to register BlockManager
18/08/31 00:02:03 INFO BlockManagerMasterEndpoint: Registering block manager localhost:41691 with 603.0 MB RAM, BlockManagerId(driver
18/08/31 00:02:03 INFO BlockManagerMaster: Registered BlockManager
Spark Context Created
Writing local file to DFS
Writing local file to DFS Completed
Reading file from DFS and running Word Count
Success! Local Word Count (11) and DFS Word Count (11) agree.

Process finished with exit code 0
```