Case-Study Day (Spark Streaming)

Objective 1:

1. There are two parts this case study

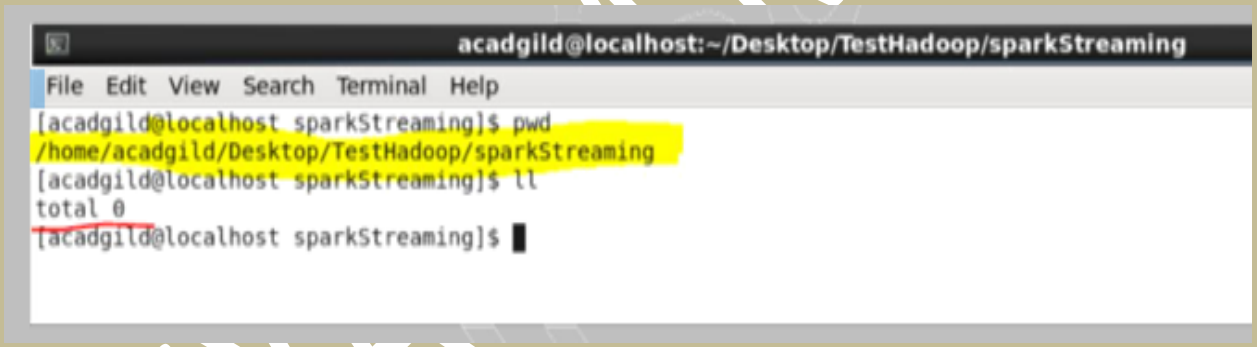
First Part - You need to create a Spark Application which streams data from a file on local directory on your machine and does the word count on the fly. The word count should be done by the spark application in such a way that as soon as you drop the file in your local directory, your spark application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

ANS:-

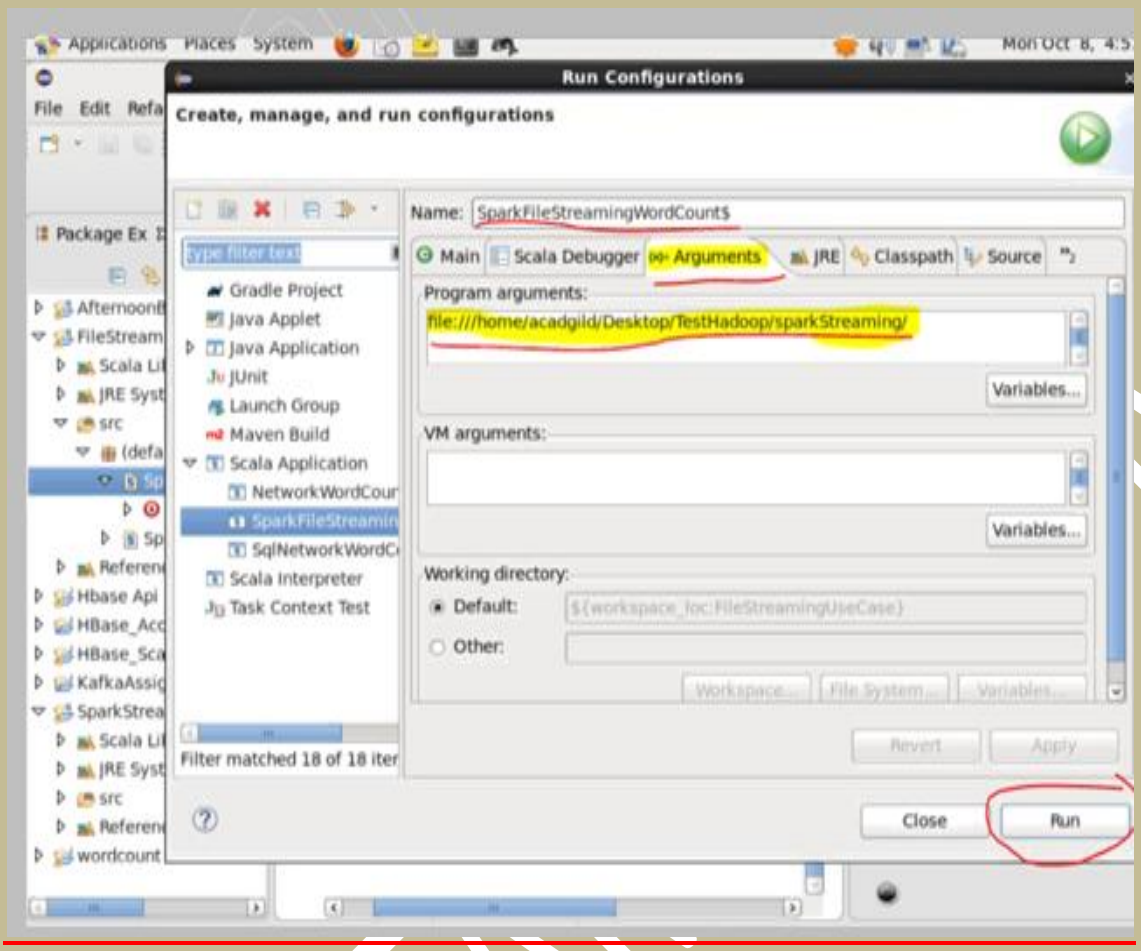
Program files are properly documented for a detailed description of each instruction used within the program.

ScreenShot:

The Directory created for streaming file is:
/home/acadgild/Desktop/TestHadoop/sparkStreaming/

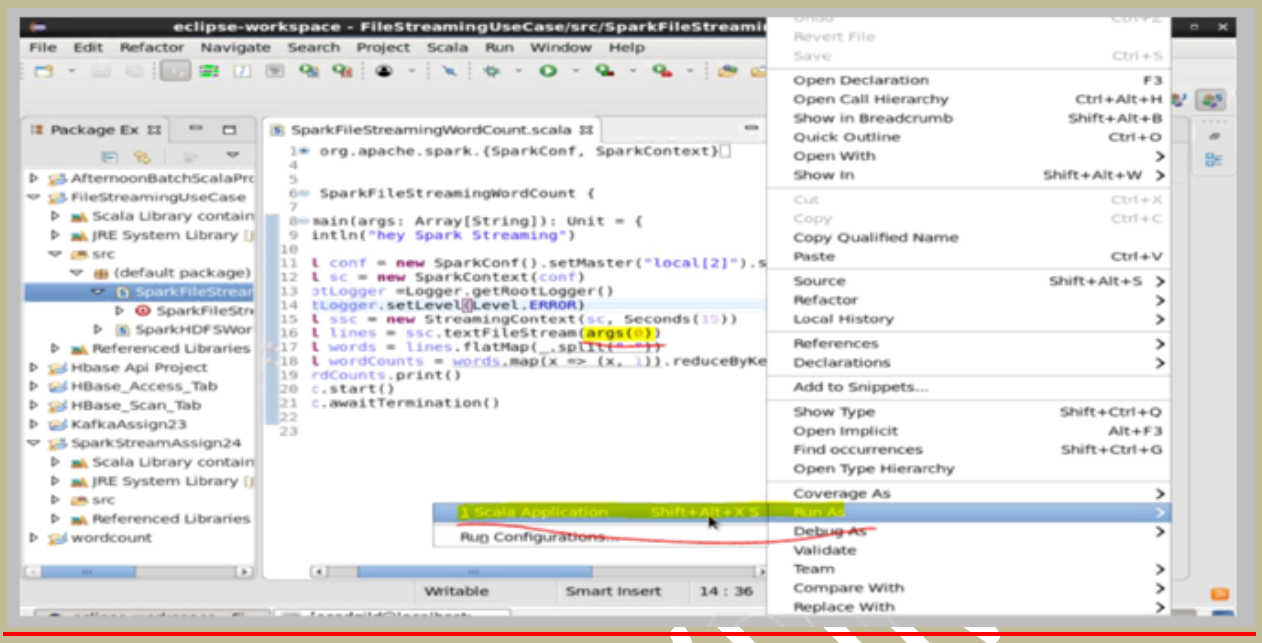


Specifying the above created directory as input to program arguments.

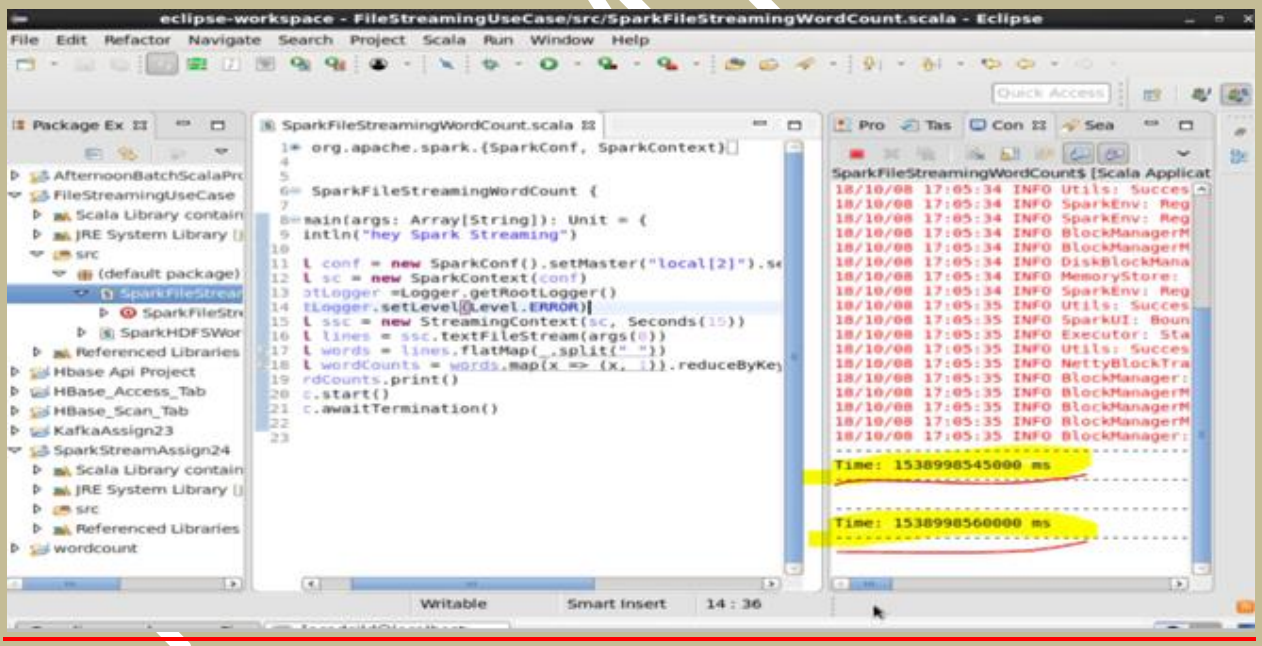


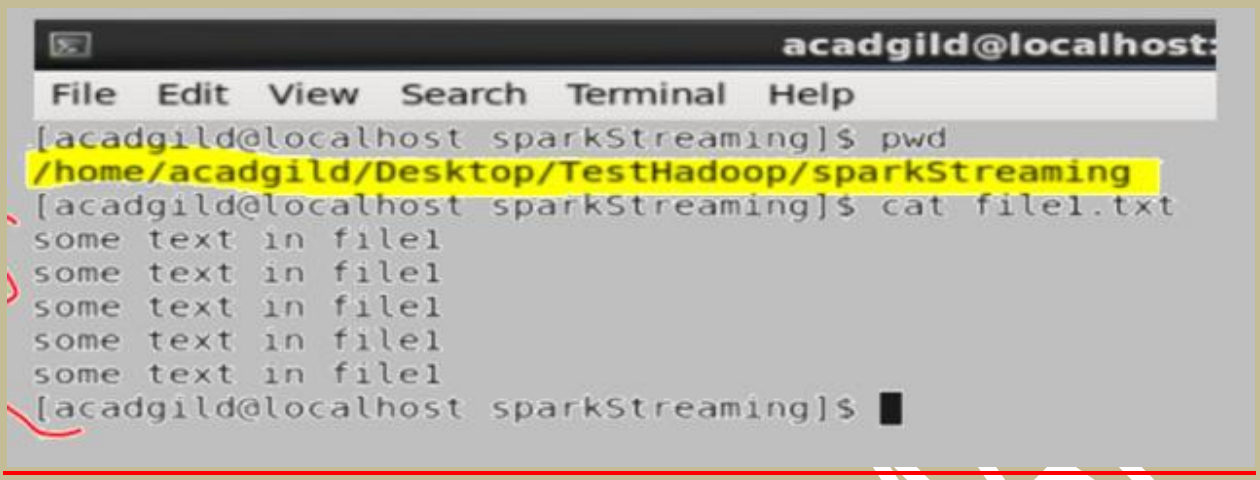
arg(0) specified within the program reads entire files that are getting added to the above directory & performs the word count operation on the fly.

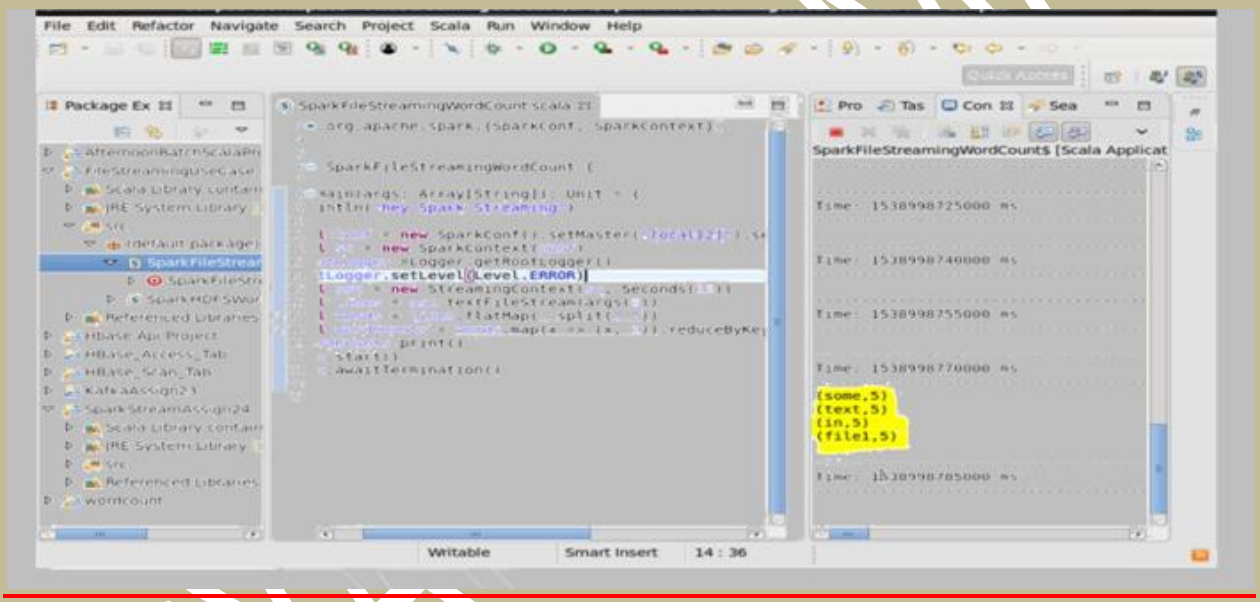
Run the application as Scala Application.



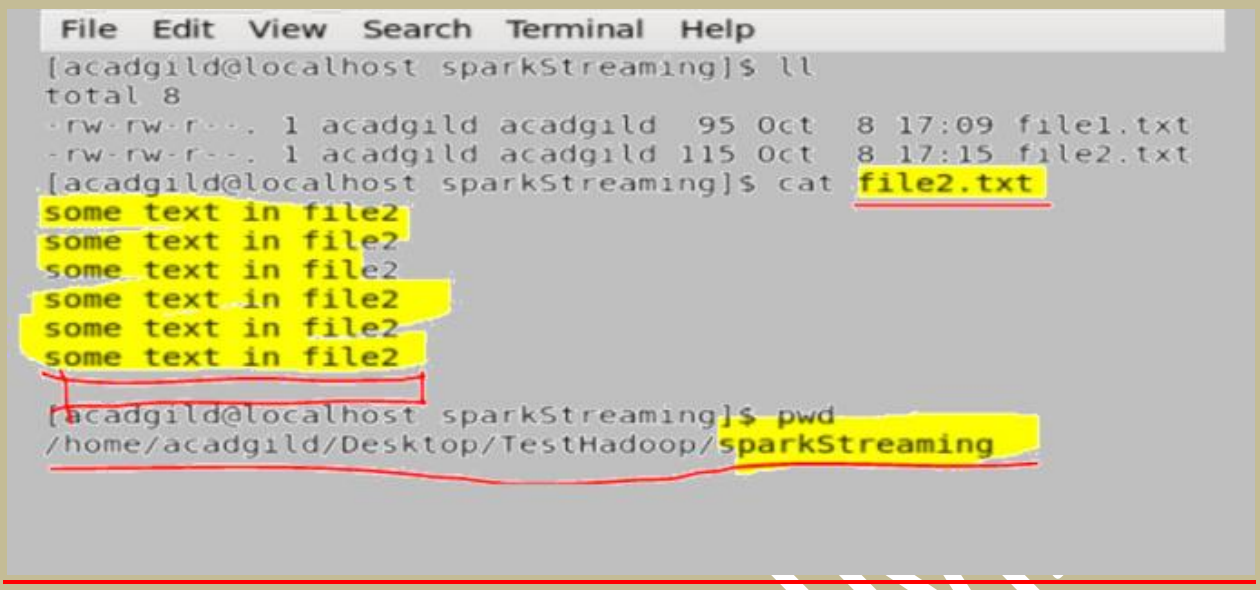
The application is streaming now. Let's create some files within input directory.

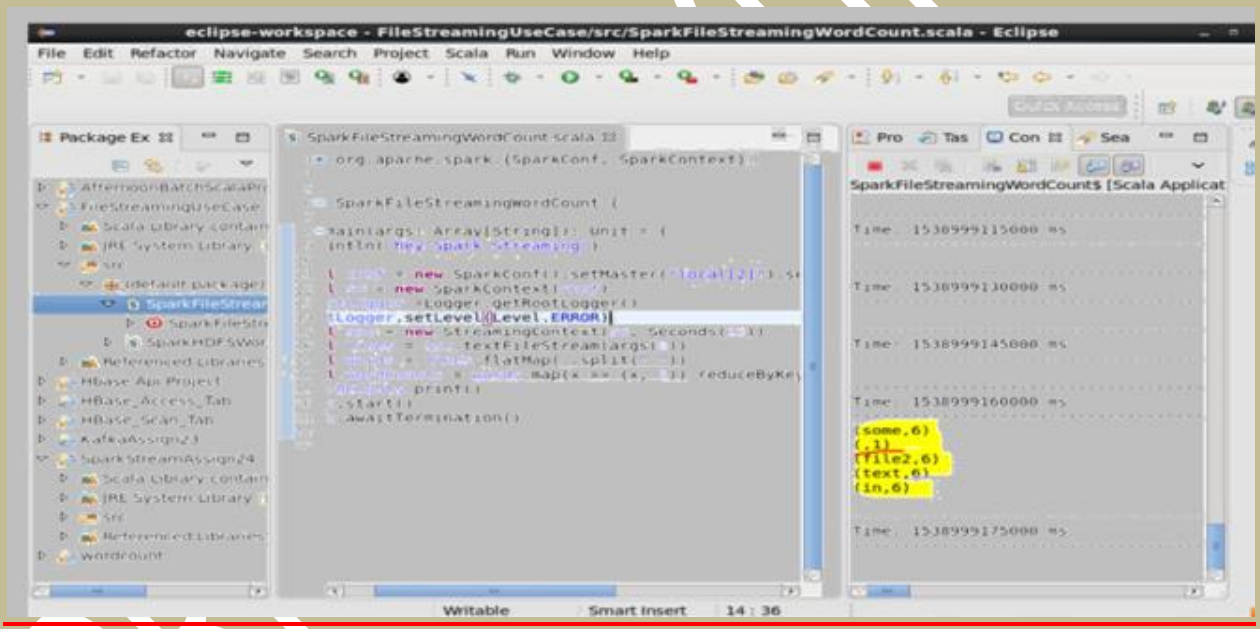




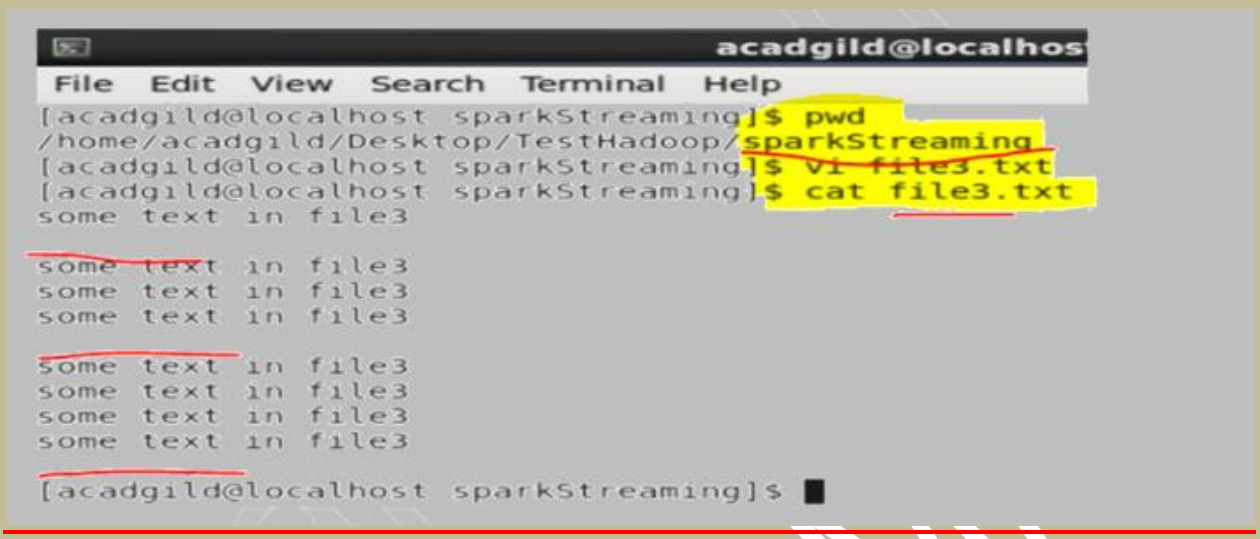


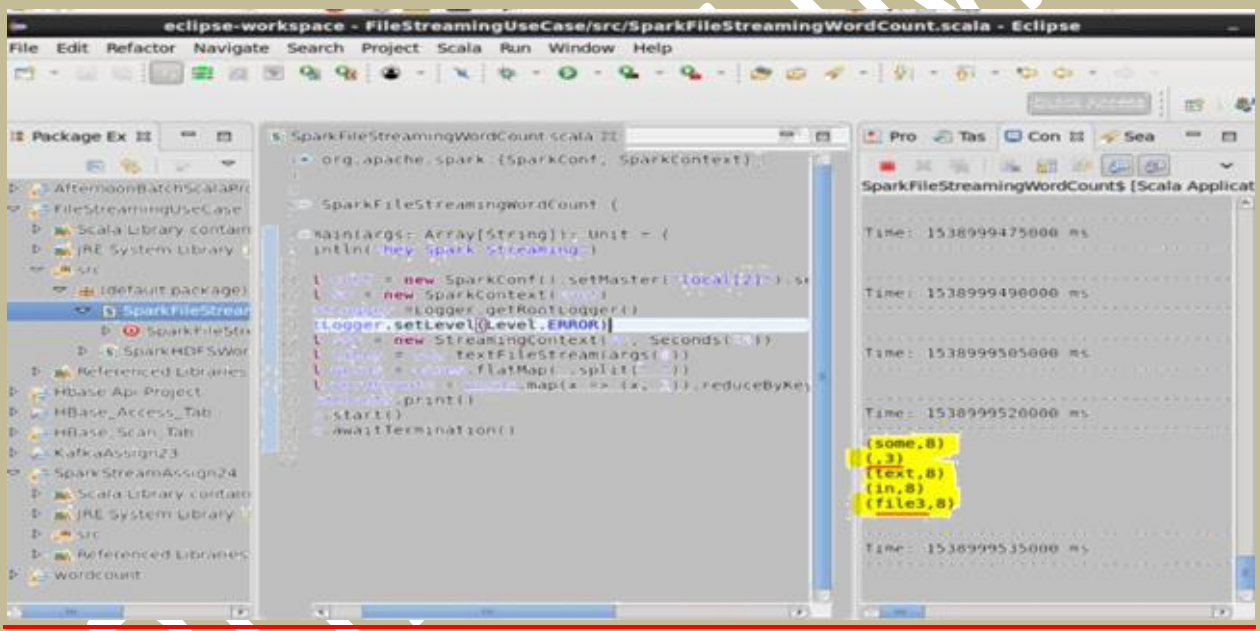
Similarly, creating two other files whose words would be counted by spark streaming application on the fly. (**Counts blank line too**)





Dropping one third file with multiple blank lines.





Objective 2:

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

Ans:-

Note: Program files are properly documented for a detailed description of each instruction used within the program.

ScreenShot:

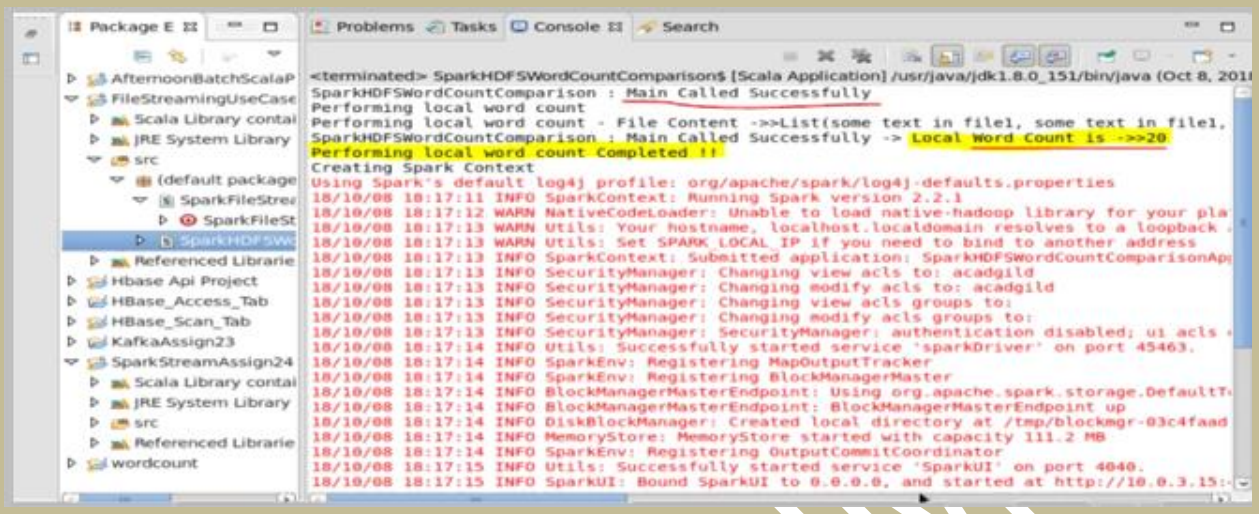
HDFS does not contain streaming directory before the application is run

```
File  Edit  View  Search  Terminal  Help
[acadgild@localhost sparkStreaming]$ hadoop fs -ls /user
18/10/08 17:42:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
drwxr-xr-x   - acadgild supergroup          0 2018-08-31 03:53 /user/acadgild
drwxr-xr-x   - acadgild supergroup          0 2018-02-09 14:50 /user/hive
[acadgild@localhost sparkStreaming]$
```

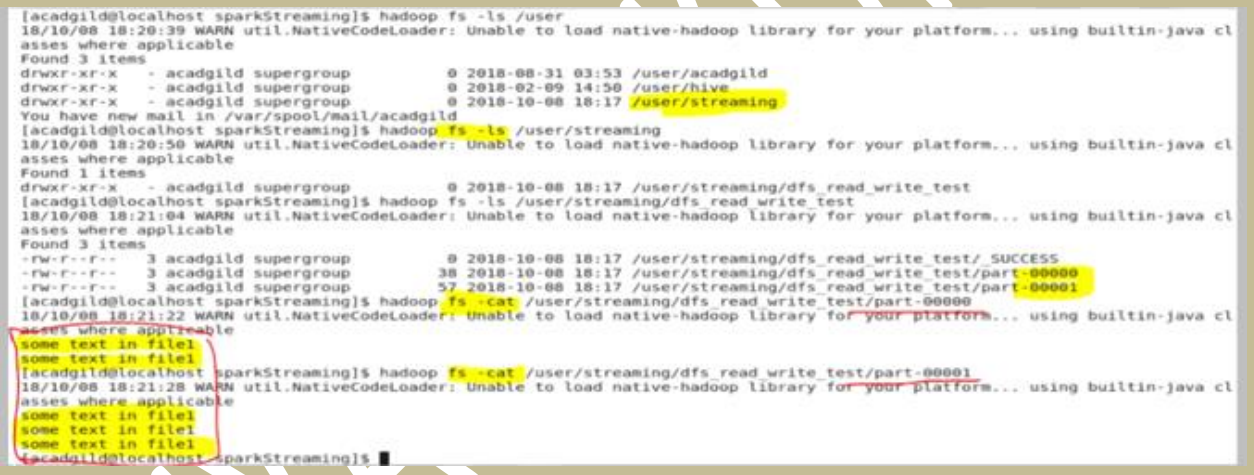
SparkHDFSWordCountComparison.scala

```
1
2
3  ark.{SparkConf, SparkContext}
4
5  ce._
6  g4j.{Level,Logger}
7
8
9
10 lCountComparison {
11
12 ilePath: File = new File("/home/acadgild/Desktop/TestHadoop/sparkStreaming/file1.txt")
13 Path: String = "hdfs://localhost:8020/user/streaming"
14 lS = 2
15
16
17 ay[String]): Unit = {
18 )
19 FSWordCountComparison : Main Called Successfully")
20
21 ing local word count")
22  = readFile(localFilePath.toString())
23
24 ing local word count - File Content ->>"+fileContents)
25 nt = runLocalWordCount(fileContents)
26
27 FSWordCountComparison : Main Called Successfully -> Local Word Count is ->>"+localWord
28
29 ing local word count Completed !!")
30
```

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.