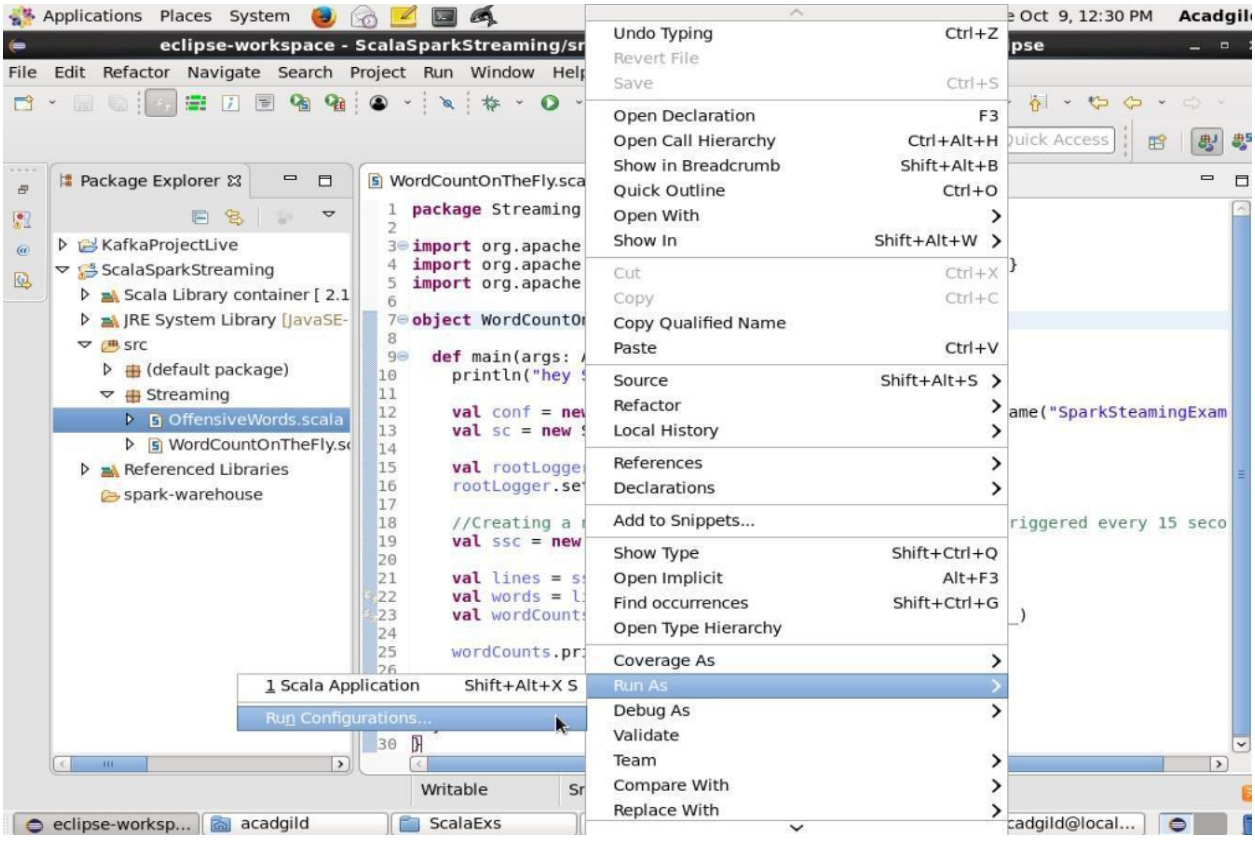
1. There are two parts this case study

- 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.

Output:

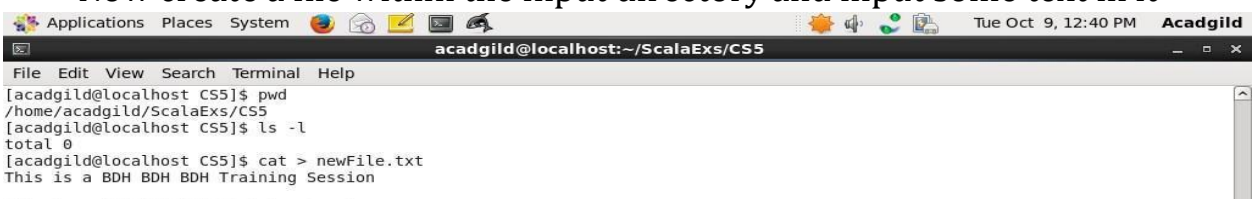
- Go to "Run Configurations" of the program



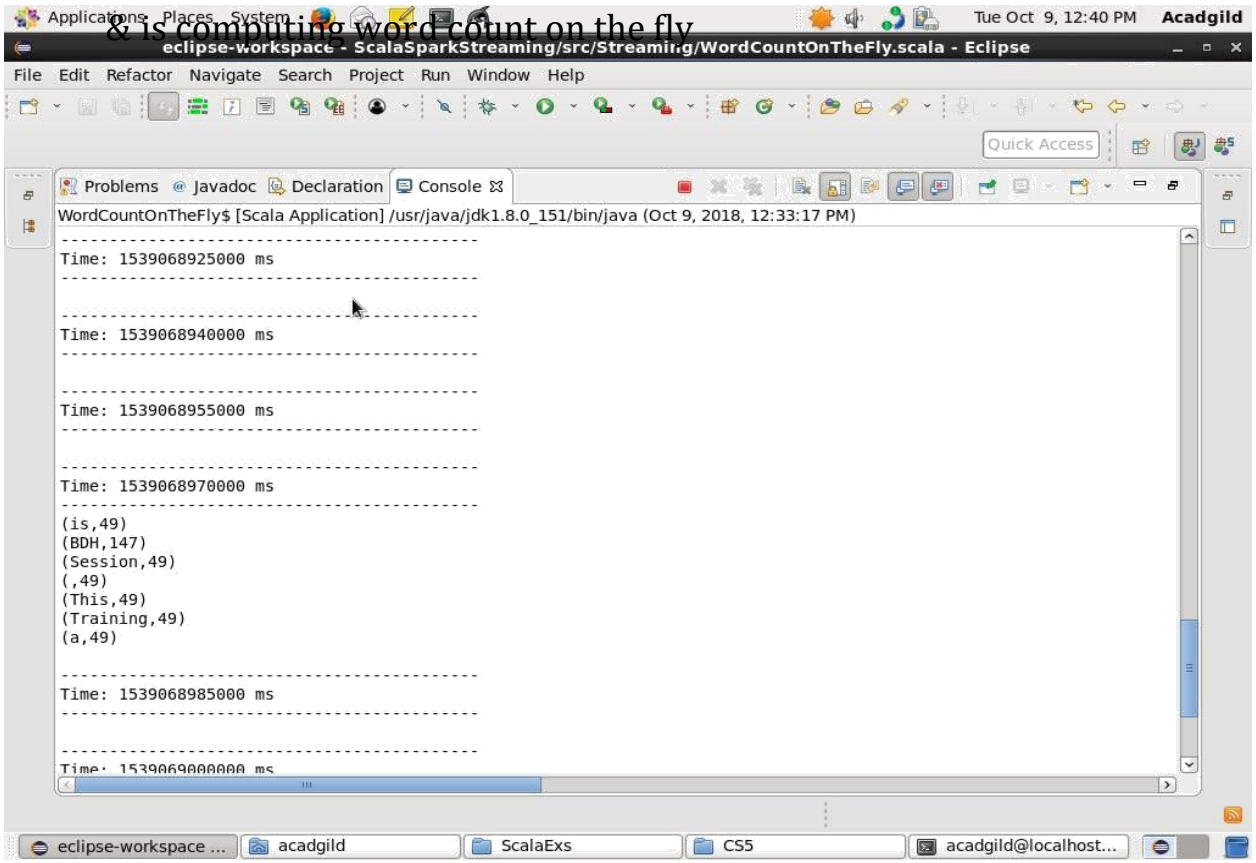
- On the "Arguments" tab Pass the arguments and click on "Run" as shown below

– The application is streaming now

– Now create a file within the input directory and input some text in it



– The contents of newFile.txt are being read by Spark Streaming application & is computing word count on the fly



- Second Part - In this part, you will have to create a Spark Application

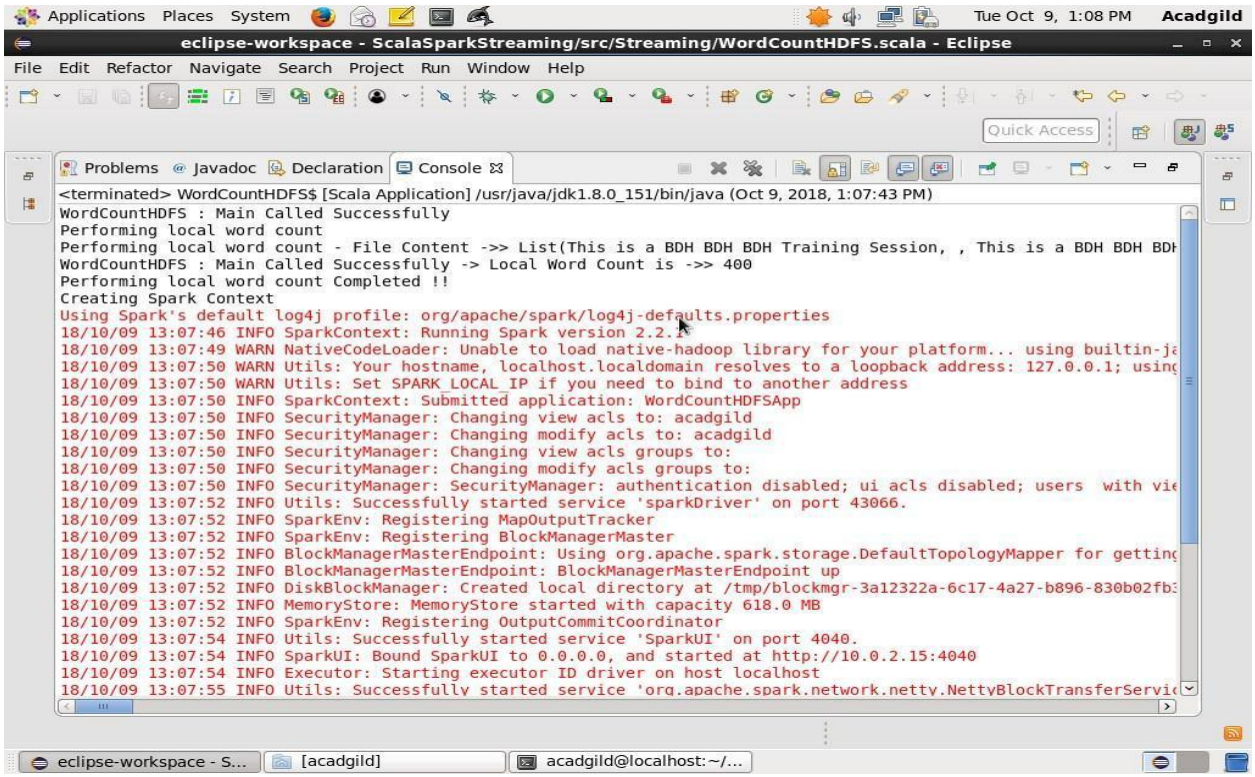
which should do the following

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

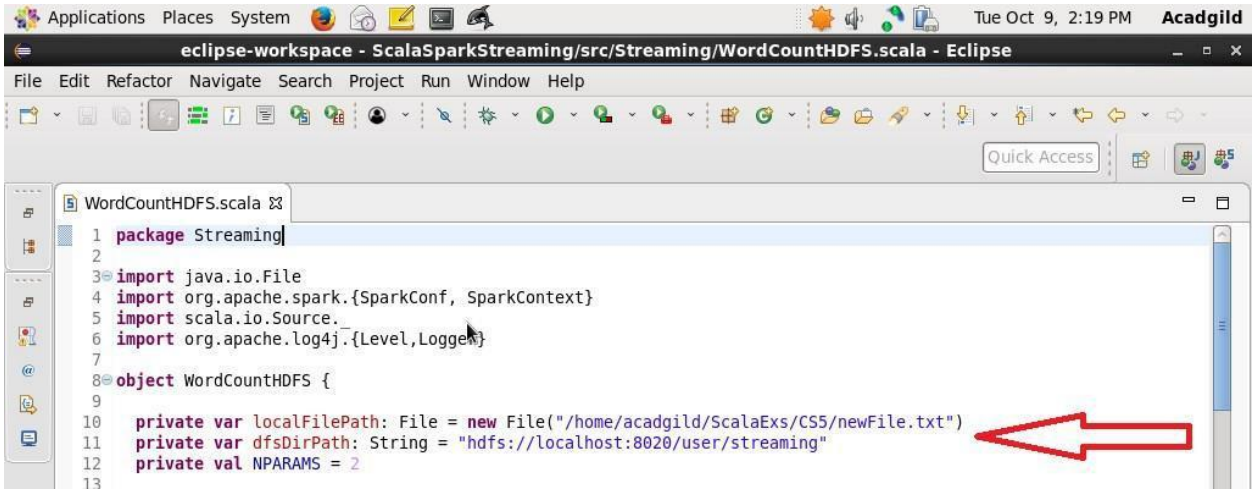
Output:

- HDFS does not contain and streaming directory before the application is run

- **Step 1**: Use newFile.txt from the local directory and do the word count

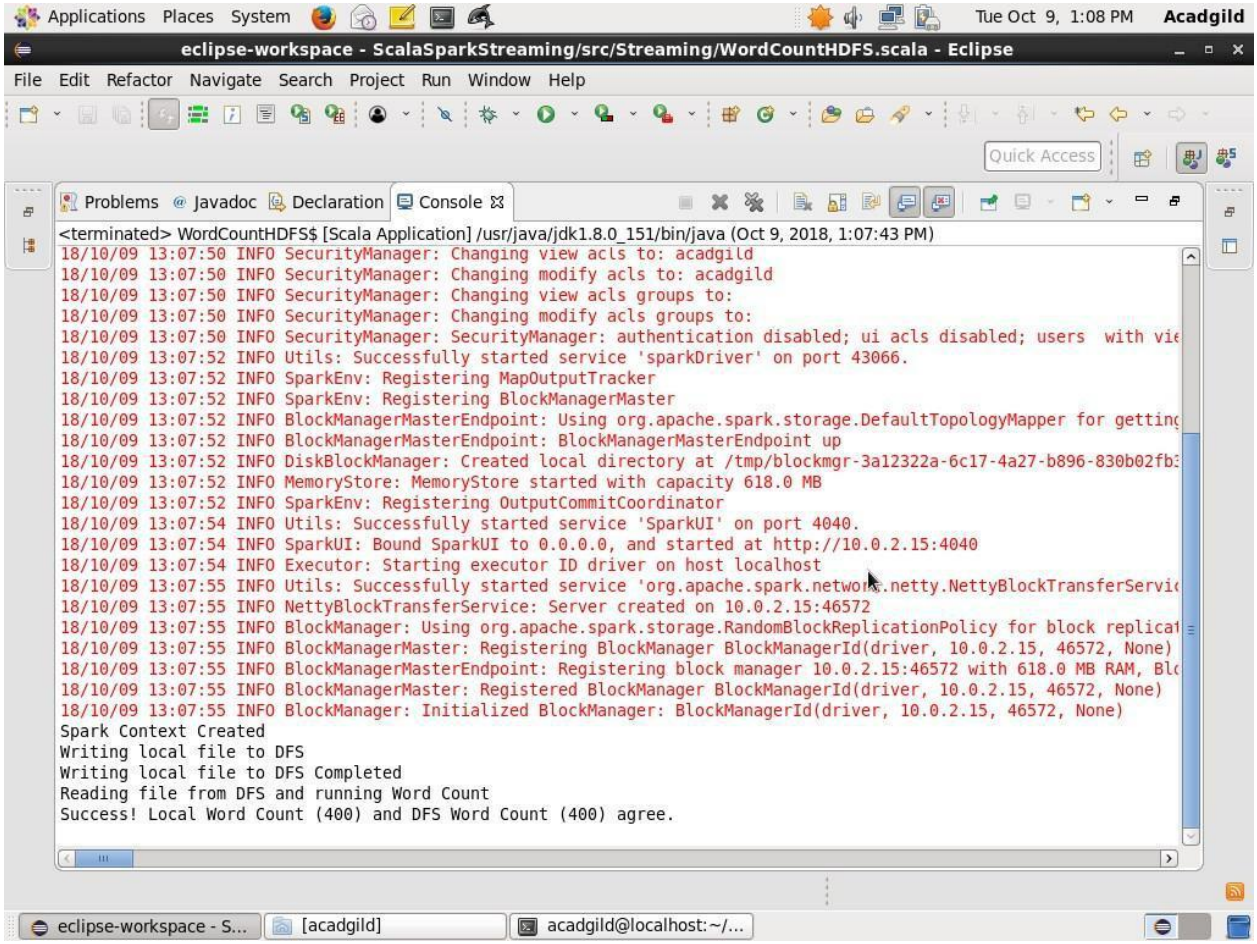


- **Step 2**: Then in the same Spark Application, write the code to put the same file on HDFS



- Then in same Spark Application, do the word count of the file copied on HDFS in step 2

- Lastly, compare the word count of step 1 and 2. Both should match, other throw an error



- Here we see that a directory **“Streaming”** was created in HDFS, which contains another directory **“dsf_read_write_test”** which contains 2 files as a result of the job performed by Spark Streaming program