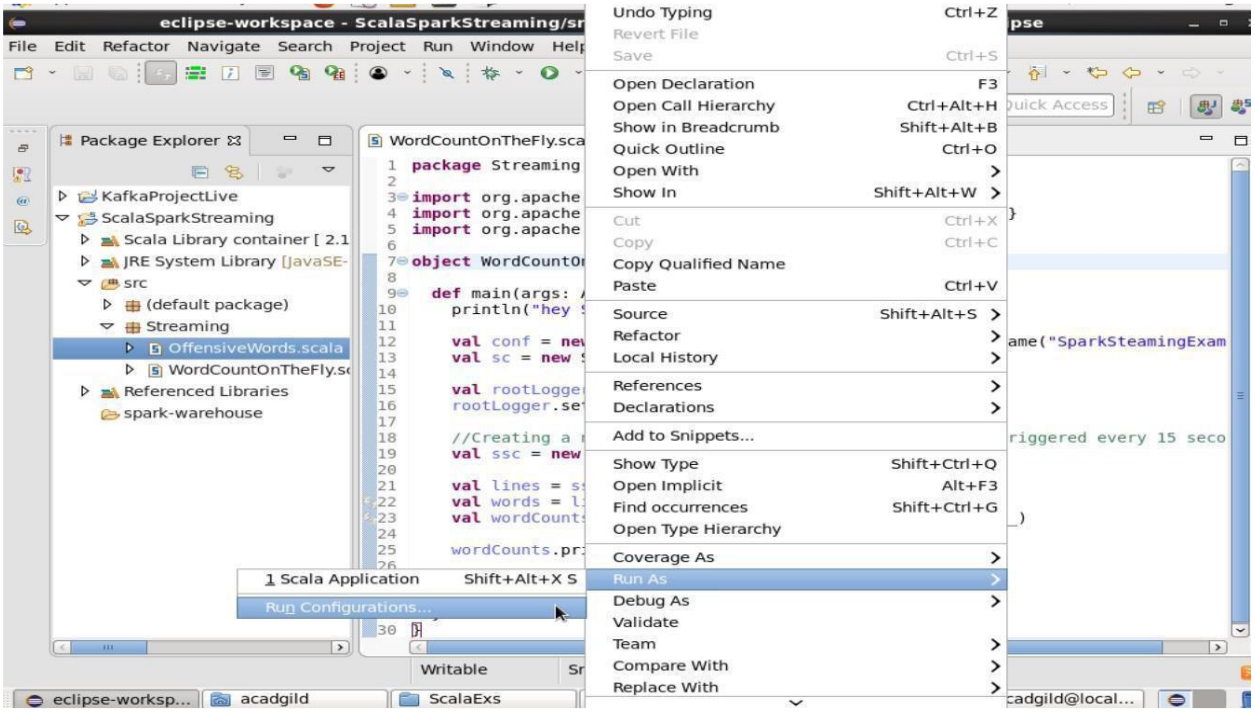
Case Study- IV Spark Streaming

1. There are two parts in the case study

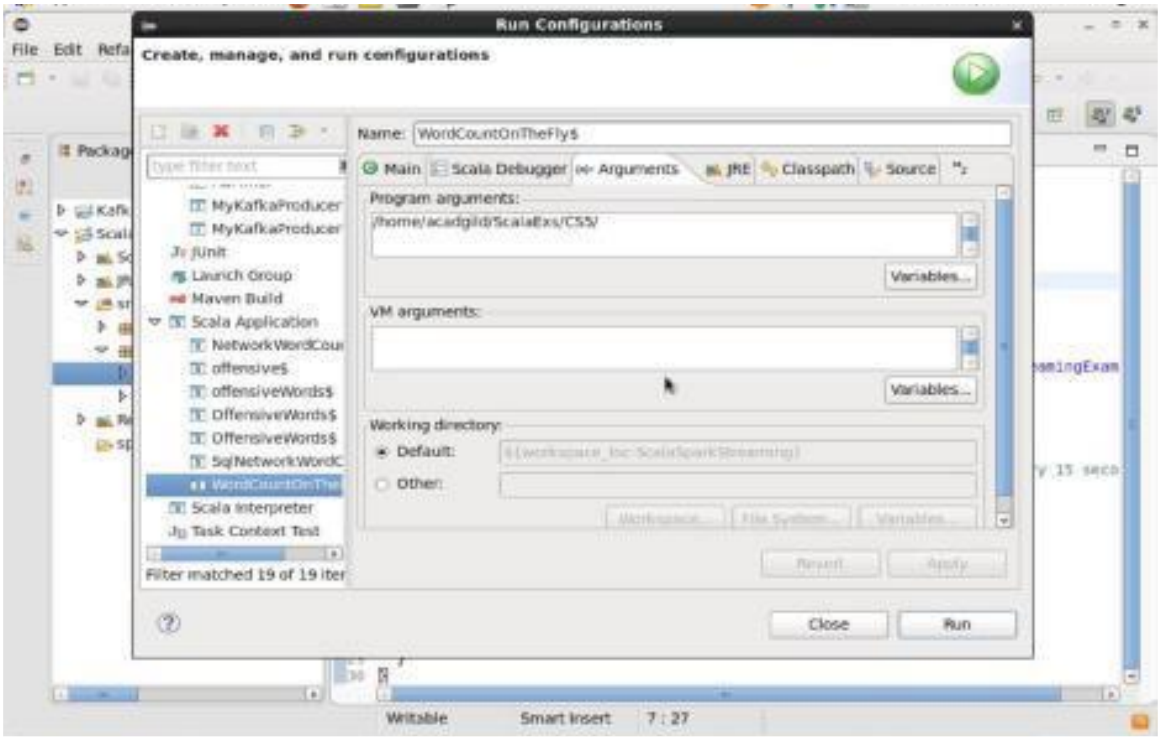
   First Part - You have to create a Spark Application which streams data from a file on local directory on your machine and does the word count on the fly. The word should be done by the spark application in such a way that as soon as you drop the file in your local directory, your spark application should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

   Output:

   - Go to Run Configurations of the program



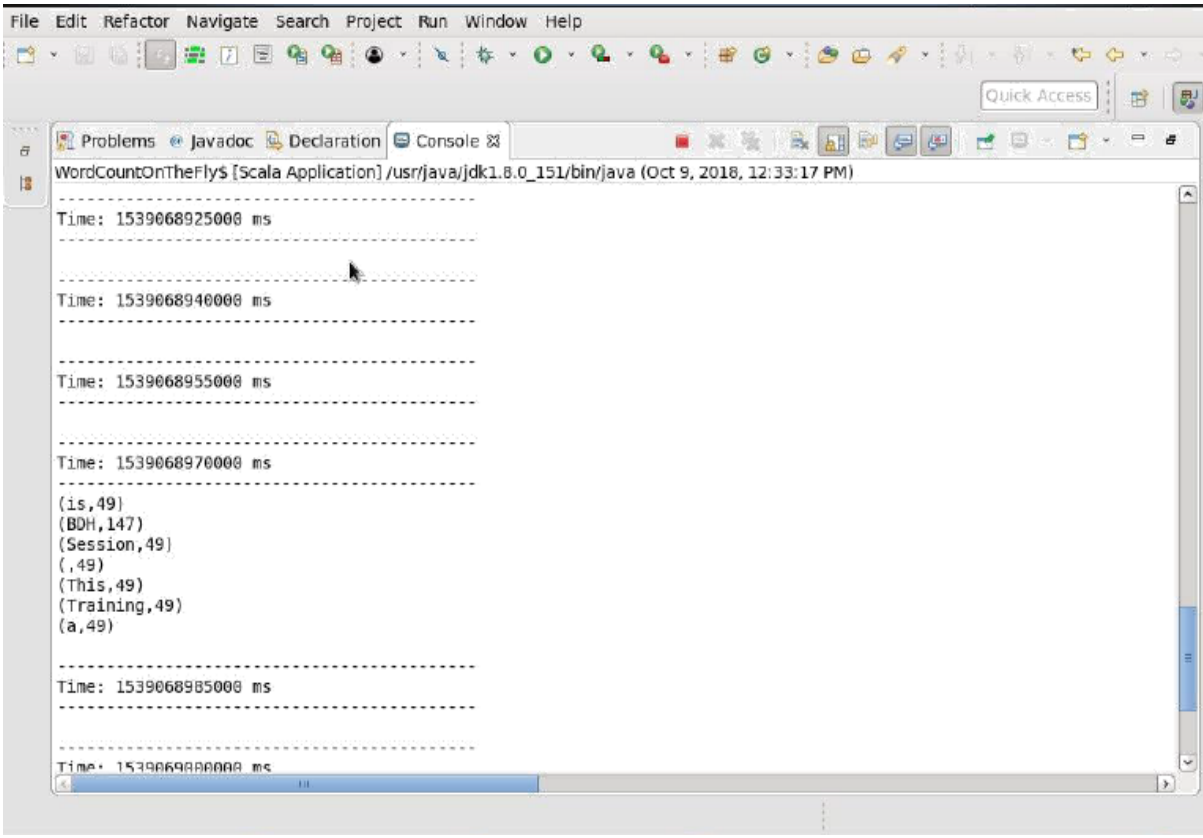
   - On the “Arguments” tab Pass the arguments and click on “Run” as shown below



- The Application is streaming now
- Now create a file within the input directory and input some text in it.

```
acadgild@localhost:~/ScalaExs/CS5
File  Edit  View  Search  Terminal  Help
[acadgild@localhost CS5]$ pwd
/home/acadgild/ScalaExs/CS5
[acadgild@localhost CS5]$ ls -l
total 0
[acadgild@localhost CS5]$ cat > newFile.txt
This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session

This is a BDH BDH BDH Training Session
```

- The contents of newFile.txt are being read by Spark Streaming application and is word counting on the fly



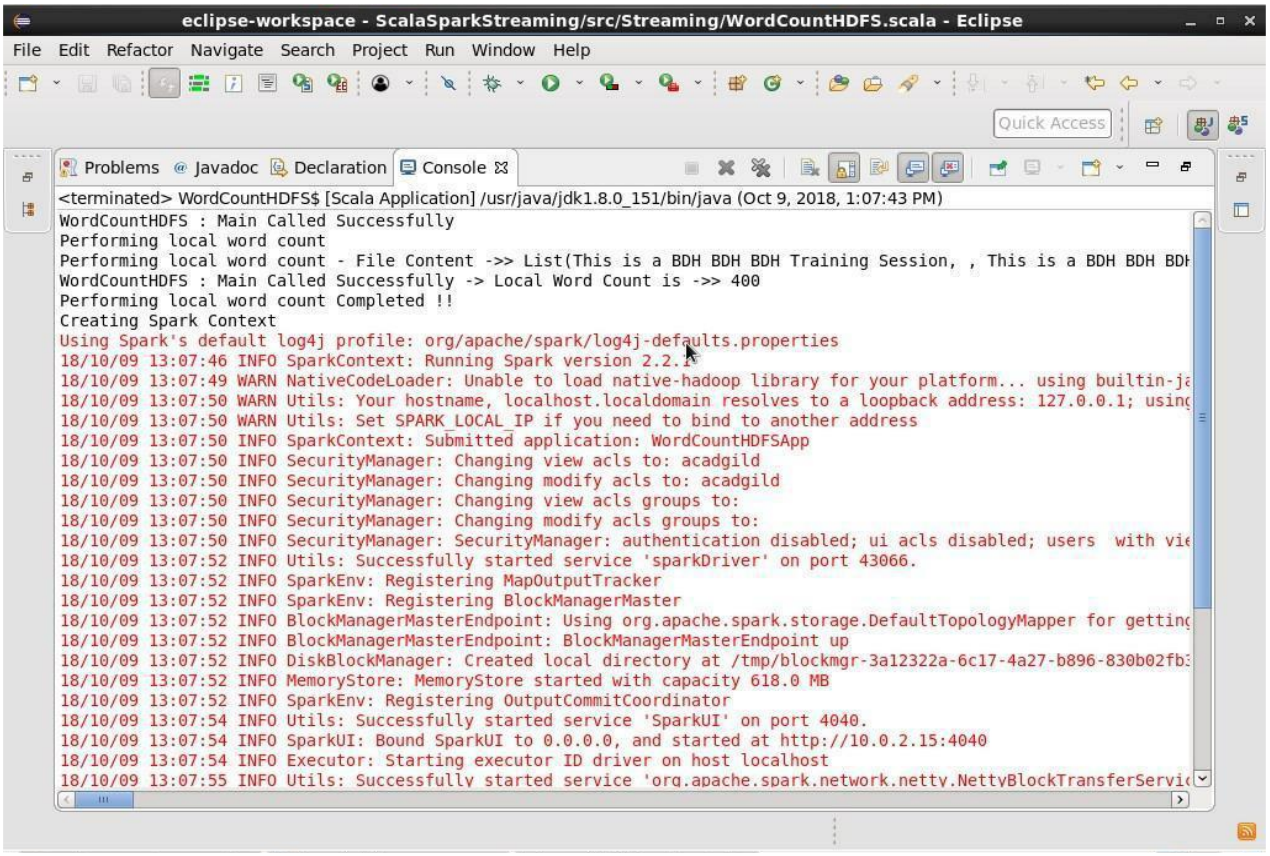
Second Part - In this part, you will have to create a Spark Application which should do the following

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

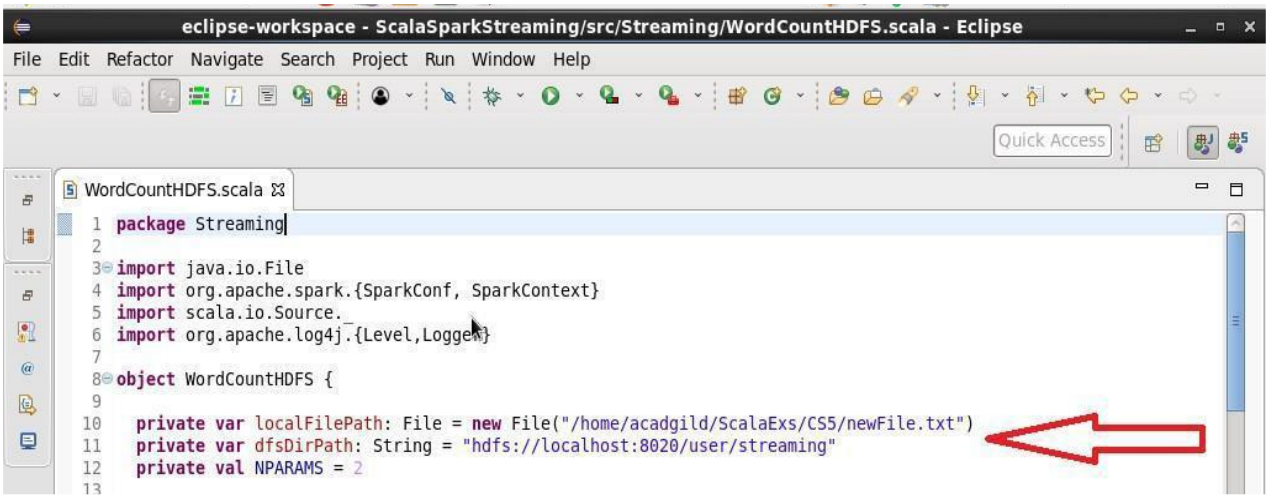
Output:

- 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.

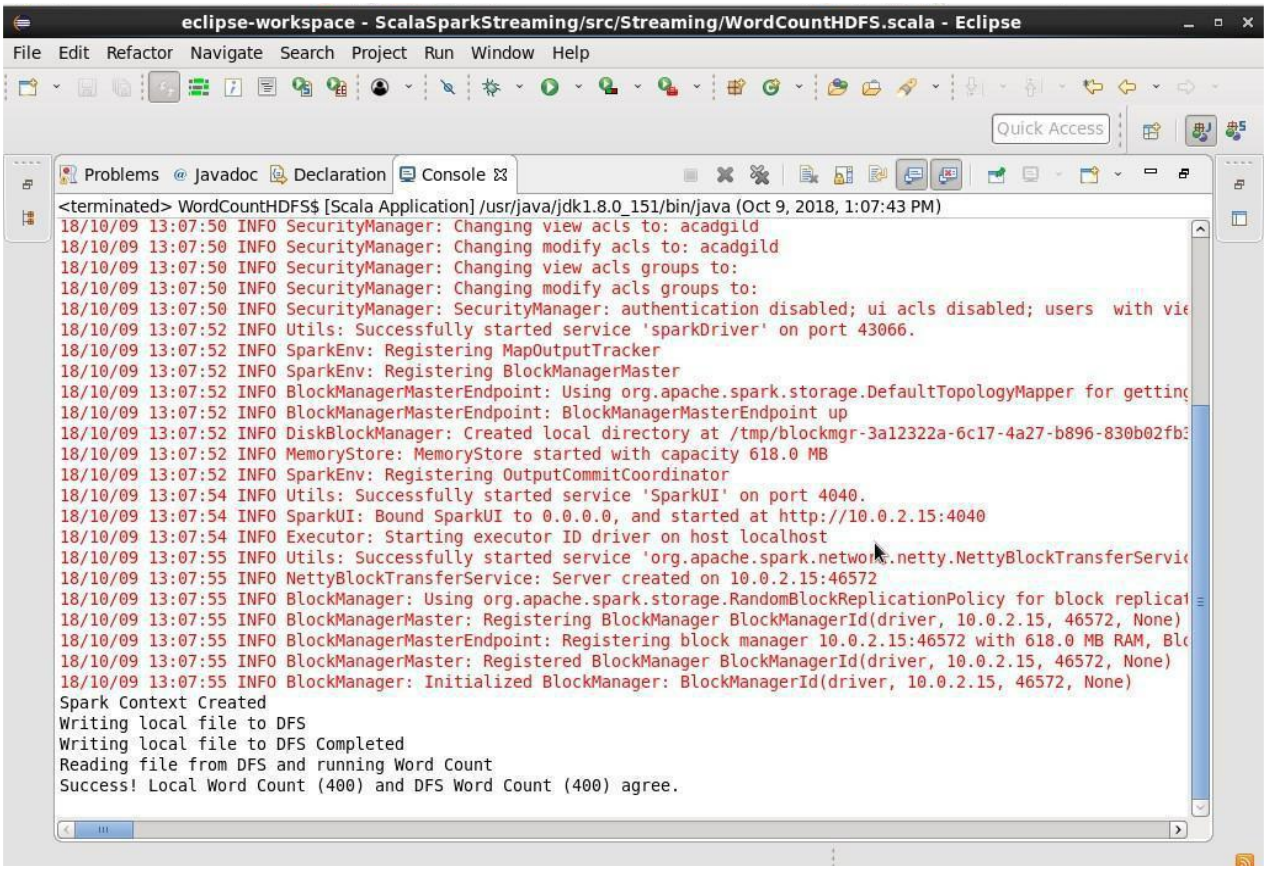


Step 2: Then in the same spark Application, write the code to put the same file on HDFS

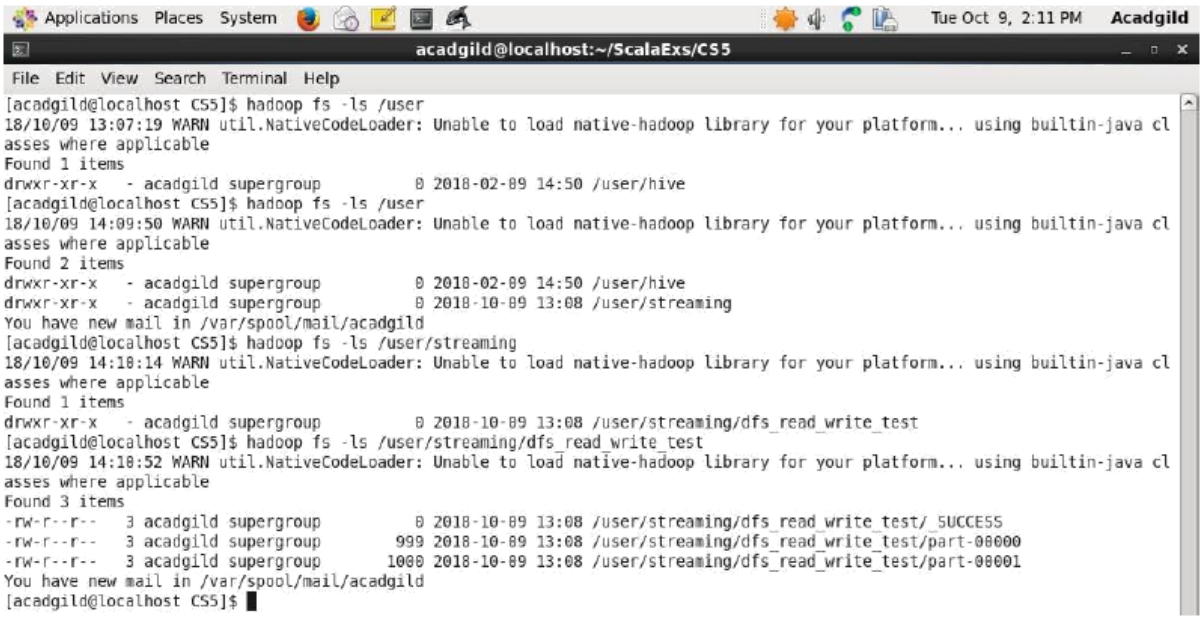


Then in same Spark Application, do the word count of the file copied on to HDFS in step 2

Lastly, compare the word count of step 1 and step 2. Both should match, otherwise 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


Applications Places System
Tue Oct 9, 2:11 PM Acadgild
acadgild@localhost:~/ScalaExs/CS5
File Edit View Search Terminal Help
[acadgild@localhost CS5]$ hadoop fs -ls /user
18/10/09 13:07:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x - acadgild supergroup 0 2018-02-09 14:50 /user/hive
[acadgild@localhost CS5]$ hadoop fs -ls /user
18/10/09 14:09:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x - acadgild supergroup 0 2018-02-09 14:50 /user/hive
drwxr-xr-x - acadgild supergroup 0 2018-10-09 13:08 /user/streaming
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost CS5]$ hadoop fs -ls /user/streaming
18/10/09 14:10:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x - acadgild supergroup 0 2018-10-09 13:08 /user/streaming/dfs_read_write_test
[acadgild@localhost CS5]$ hadoop fs -ls /user/streaming/dfs_read_write_test
18/10/09 14:10:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r-- 3 acadgild supergroup 0 2018-10-09 13:08 /user/streaming/dfs_read_write_test/_SUCCESS
-rw-r--r-- 3 acadgild supergroup 999 2018-10-09 13:08 /user/streaming/dfs_read_write_test/part-00000
-rw-r--r-- 3 acadgild supergroup 1000 2018-10-09 13:08 /user/streaming/dfs_read_write_test/part-00001
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost CS5]$