CaseStudy5:(Spark Streaming)

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.

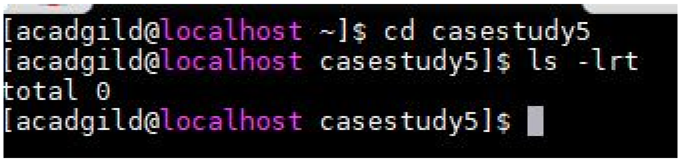
Steps:
----

1. Created a new empty folder with no files in it
2. Run the scala code
3. Create a new file inside casestudy5 while the code is running.
4. Code will count the words and prints the output in console.

Note : Source code is uploaded separately.

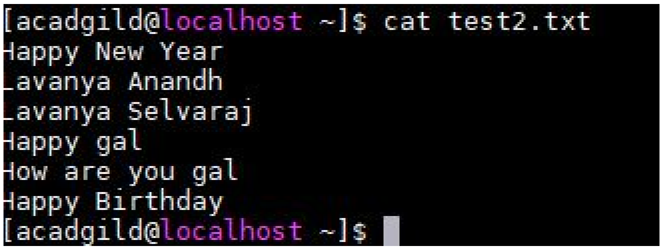
Output screen shots:

New empty folder:



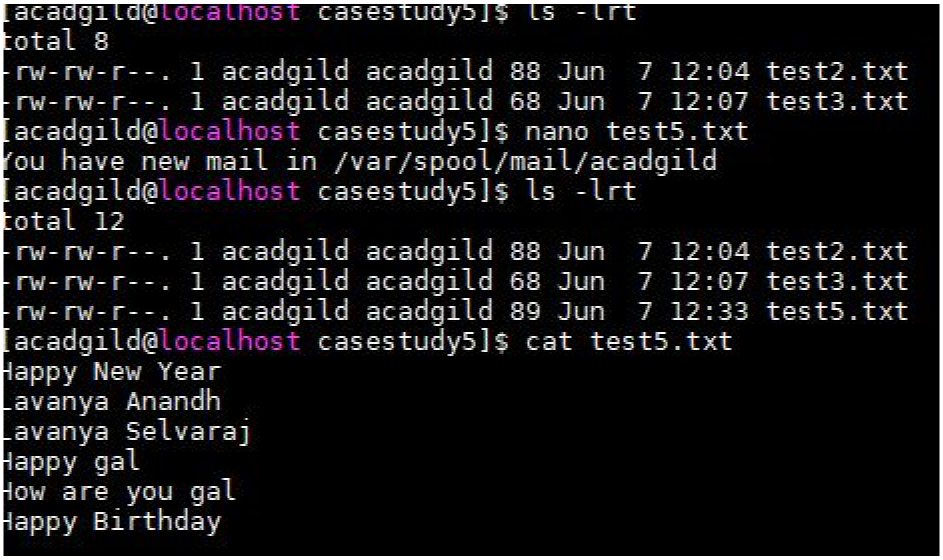
```
[acadgild@localhost ~]$ cd casestudy5
[acadgild@localhost casestudy5]$ ls -lrt
total 0
[acadgild@localhost casestudy5]$
```

File in the local folder:



```
[acadgild@localhost ~]$ cat test2.txt
Happy New Year
Lavanya Anandh
Lavanya Selvaraj
Happy gal
How are you gal
Happy Birthday
[acadgild@localhost ~]$
```

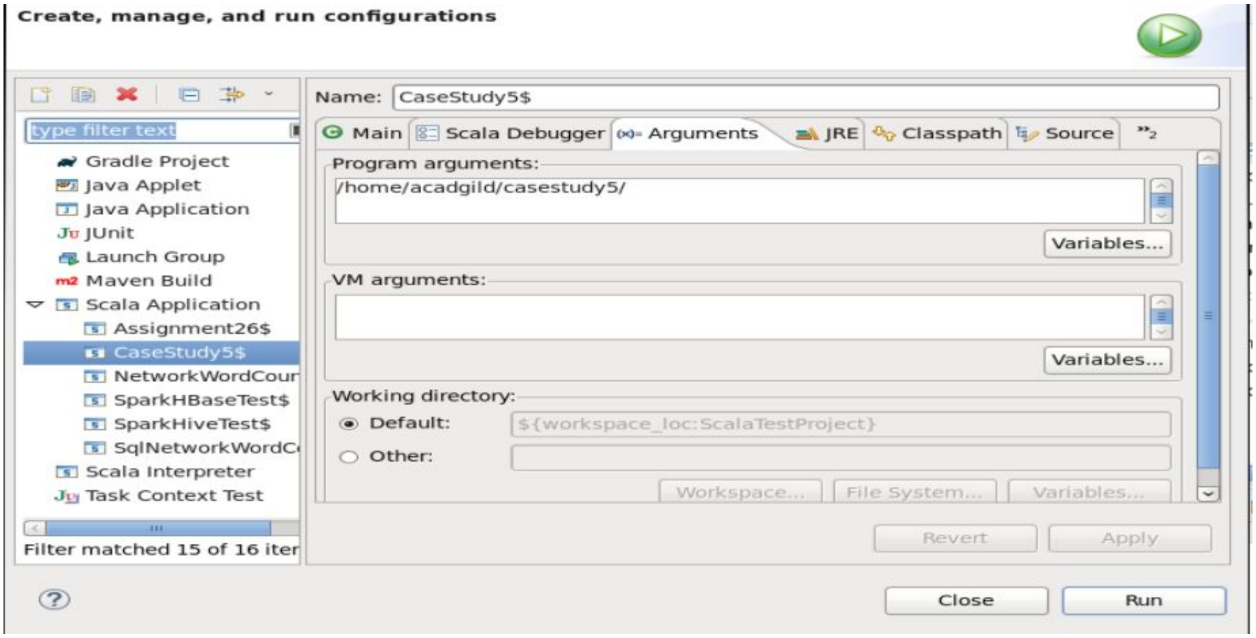
Created new file in the casestudy5 folder while the code is running:



```
[acadgild@localhost casestudy5]$ ls -lrt
total 8
-rw-rw-r--. 1 acadgild acadgild 88 Jun  7 12:04 test2.txt
-rw-rw-r--. 1 acadgild acadgild 68 Jun  7 12:07 test3.txt
[acadgild@localhost casestudy5]$ nano test5.txt
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost casestudy5]$ ls -lrt
total 12
-rw-rw-r--. 1 acadgild acadgild 88 Jun  7 12:04 test2.txt
-rw-rw-r--. 1 acadgild acadgild 68 Jun  7 12:07 test3.txt
-rw-rw-r--. 1 acadgild acadgild 89 Jun  7 12:33 test5.txt
[acadgild@localhost casestudy5]$ cat test5.txt
Happy New Year
Lavanya Anandh
Lavanya Selvaraj
Happy gal
How are you gal
Happy Birthday
```

Output after running the scala code:

Adding the input parameter of newly created folder path:



Create, manage, and run configurations

Name: CaseStudy5$

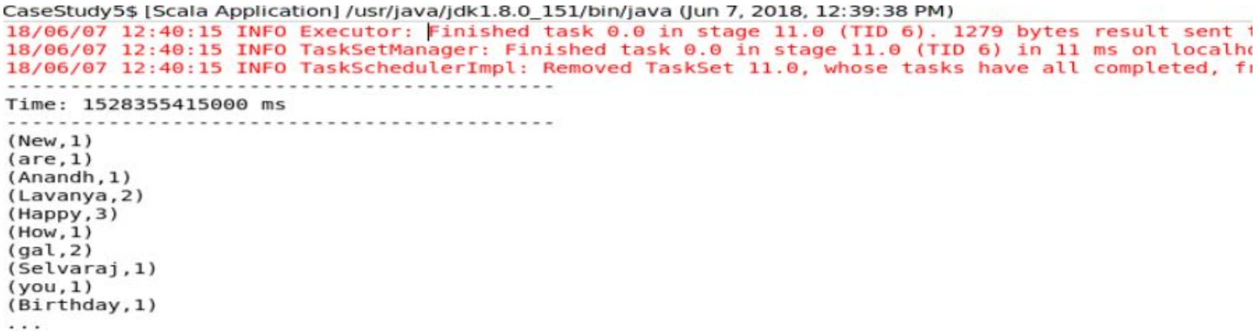
Program arguments: /home/acadgild/casestudy5/

VM arguments:

Working directory:
● Default: ${workspace_loc:ScalaTestProject}
○ Other:

Final Output showing wordcount:

--



```
CaseStudy5$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (Jun 7, 2018, 12:39:38 PM)
18/06/07 12:40:15 INFO Executor: Finished task 0.0 in stage 11.0 (TID 6). 1279 bytes result sent t
18/06/07 12:40:15 INFO TaskSetManager: Finished task 0.0 in stage 11.0 (TID 6) in 11 ms on localho
18/06/07 12:40:15 INFO TaskSchedulerImpl: Removed TaskSet 11.0, whose tasks have all completed, fr
-------------------------------------------
Time: 1528355415000 ms
-------------------------------------------
(New,1)
(are,1)
(Anandh,1)
(Lavanya,2)
(Happy,3)
(How,1)
(gal,2)
(Selvaraj,1)
(you,1)
(Birthday,1)
...
```

Second Part:

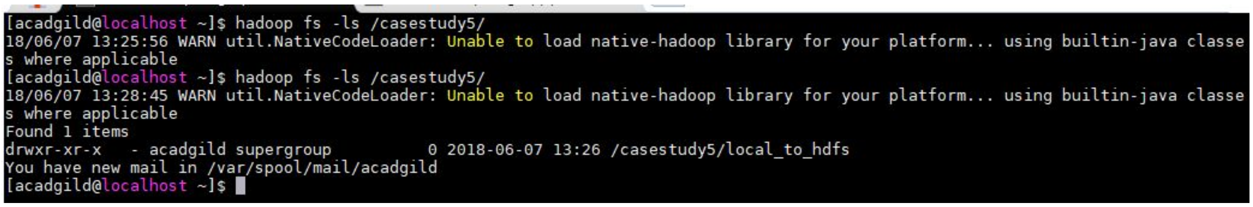
In this part, you will have to create a Spark Application which should do the following
1. Pick up a file from the local directory and do the word count
2. Then in the same Spark Application, write the code to put the same file on HDFS.
3. Then in same Spark Application, do the word count of the file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

Steps:

----

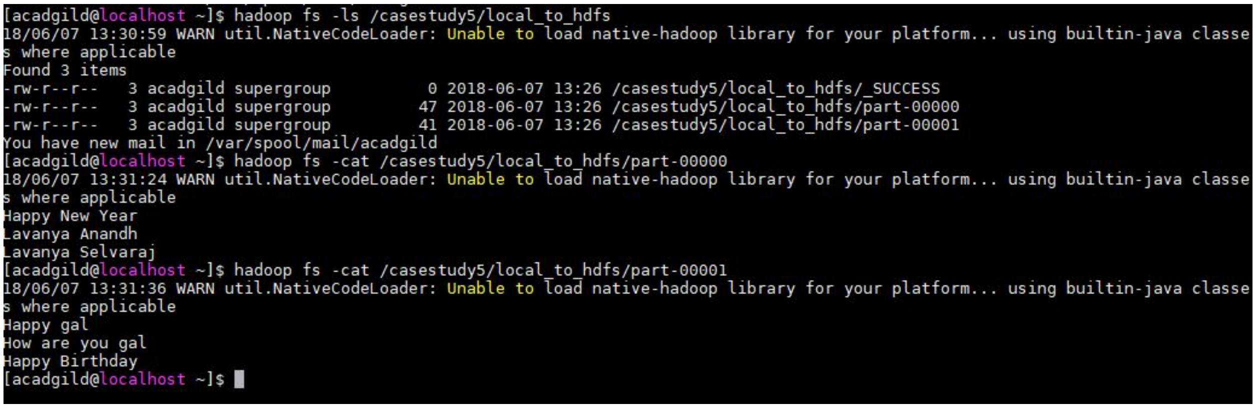
1. **Created a new folder as casestudy5 in hadoop filesystem.**
2. **Had a file in local folder so that it can be moved to hdfs on the fly and used for wordcount.**

Hadoop empty folder and after moving the file into hadoop folder:



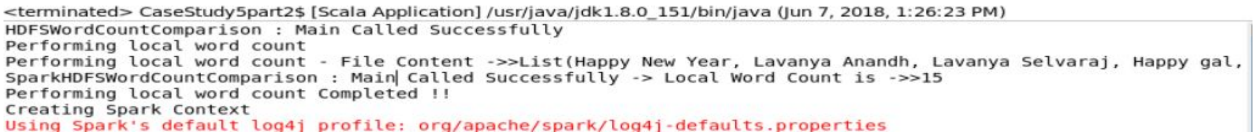
```
[acadgild@localhost ~]$ hadoop fs -ls /casestudy5/
18/06/07 13:25:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
[acadgild@localhost ~]$ hadoop fs -ls /casestudy5/
18/06/07 13:28:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Found 1 items
drwxr-xr-x   - acadgild supergroup          0 2018-06-07 13:26 /casestudy5/local_to_hdfs
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```

Files are written in hdfs:



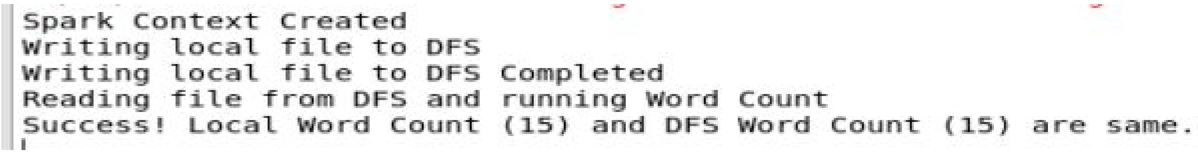
```
[acadgild@localhost ~]$ hadoop fs -ls /casestudy5/local_to_hdfs
18/06/07 13:30:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Found 3 items
-rw-r--r--   3 acadgild supergroup          0 2018-06-07 13:26 /casestudy5/local_to_hdfs/_SUCCESS
-rw-r--r--   3 acadgild supergroup         47 2018-06-07 13:26 /casestudy5/local_to_hdfs/part-00000
-rw-r--r--   3 acadgild supergroup         41 2018-06-07 13:26 /casestudy5/local_to_hdfs/part-00001
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /casestudy5/local_to_hdfs/part-00000
18/06/07 13:31:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Happy New Year
Lavanya Anandh
Lavanya Selvaraj
[acadgild@localhost ~]$ hadoop fs -cat /casestudy5/local_to_hdfs/part-00001
18/06/07 13:31:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Happy gal
How are you gal
Happy Birthday
[acadgild@localhost ~]$
```

Console output:



```
<terminated> CaseStudy5part2$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (Jun 7, 2018, 1:26:23 PM)
HDFSWordCountComparison : Main Called Successfully
Performing local word count
Performing local word count - File Content ->>List(Happy New Year, Lavanya Anandh, Lavanya Selvaraj, Happy gal,
SparkHDFSWordCountComparison : Main Called Successfully -> Local Word Count is ->>15
Performing local word count Completed !!
Creating Spark Context
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
```

File comparison after moving the file contents from local:



```
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 (15) and DFS Word Count (15) are same.
```