EXPERIMENT 10

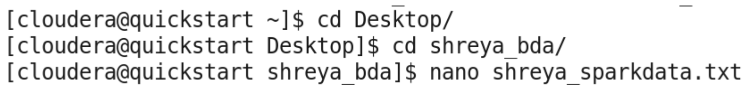
Name	Shreya Shetty
UID	20191410059
Batch	A
Class	TE IT
Subject	BDA

AIM: Implement word count using Apache Spark

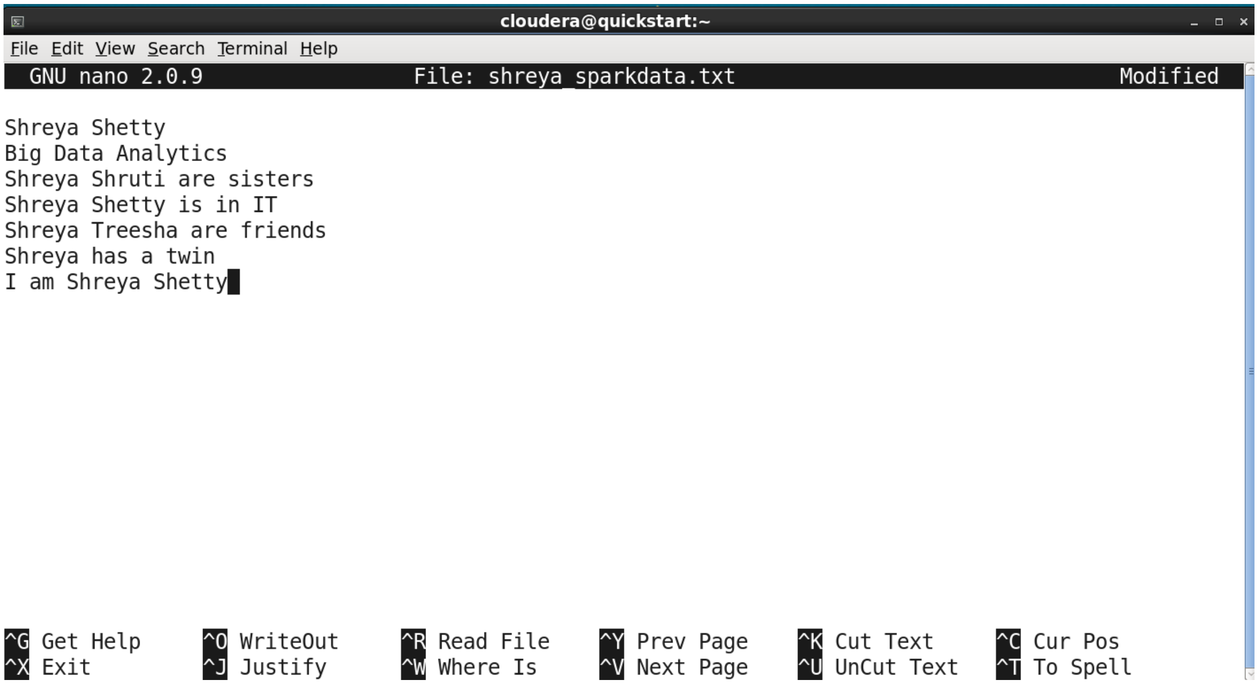
STEPS FOLLOWED:

1. **Create a text file in your local machine and write some text into it**

   $ nano shreya_sparkdata.txt



```
[cloudera@quickstart ~]$ cd Desktop/
[cloudera@quickstart Desktop]$ cd shreya_bda/
[cloudera@quickstart shreya_bda]$ nano shreya_sparkdata.txt
```

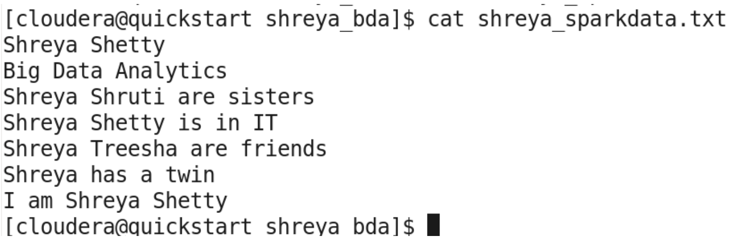


```
  GNU nano 2.0.9              File: shreya_sparkdata.txt                   Modified

Shreya Shetty
Big Data Analytics
Shreya Shruti are sisters
Shreya Shetty is in IT
Shreya Treesha are friends
Shreya has a twin
I am Shreya Shetty
```

2. **Check the text written in the shreya_sparkdata.txt file**

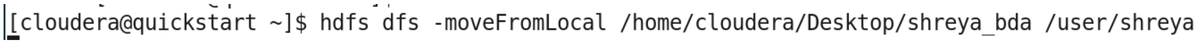
   $ cat shreya_sparkdata.txt



```
[cloudera@quickstart shreya_bda]$ cat shreya_sparkdata.txt
Shreya Shetty
Big Data Analytics
Shreya Shruti are sisters
Shreya Shetty is in IT
Shreya Treesha are friends
Shreya has a twin
I am Shreya Shetty
[cloudera@quickstart shreya_bda]$ █
```

3. **Moving the text file into hdfs in directory “/user/shreya”**

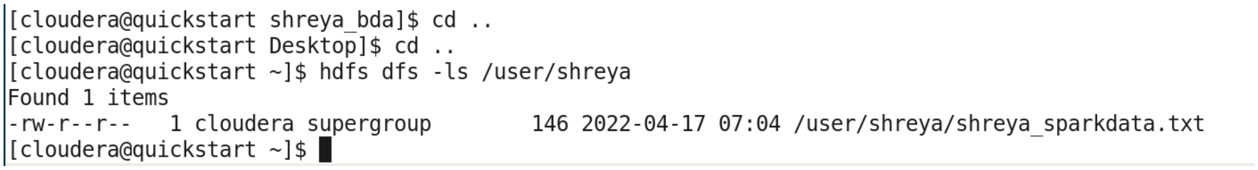
   $ hdfs dfs –moveFromLocal /ome/cloudera/Desktop/shreya_bda /user/shreya



```
[cloudera@quickstart ~]$ hdfs dfs -moveFromLocal /home/cloudera/Desktop/shreya_bda /user/shreya
```

4. **Checking the contents of the directory, to check if the file has been moved to hdfs**

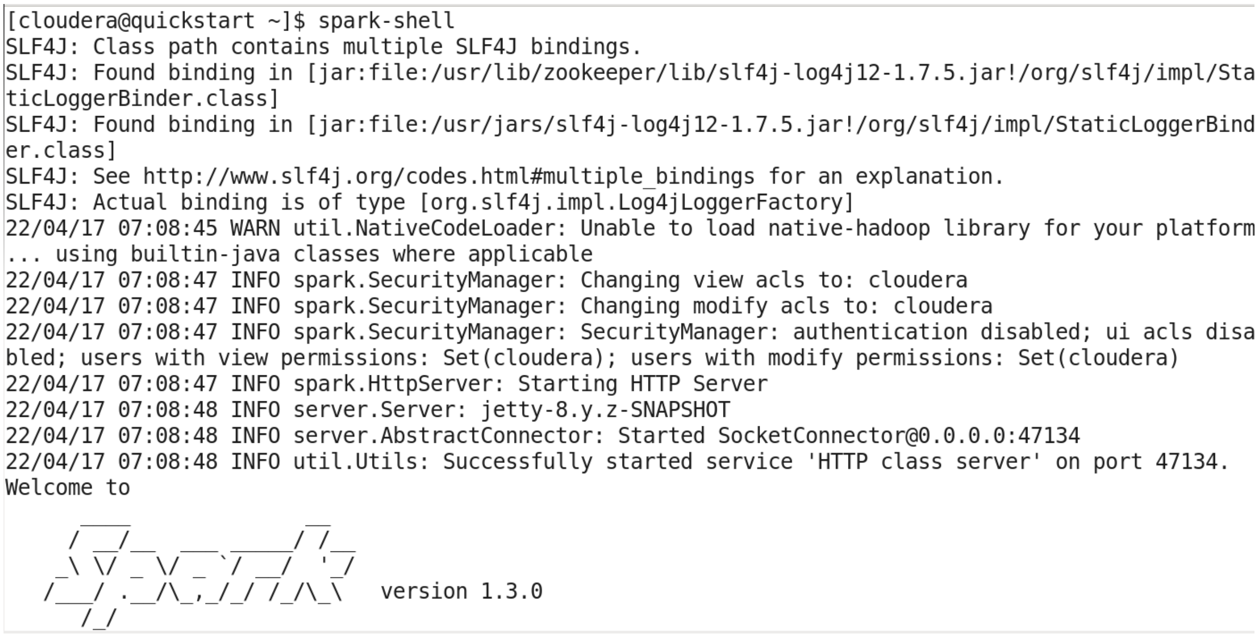
   $ hdfs dfs –ls /user/shreya



```
[cloudera@quickstart shreya_bda]$ cd ..
[cloudera@quickstart Desktop]$ cd ..
[cloudera@quickstart ~]$ hdfs dfs -ls /user/shreya
Found 1 items
-rw-r--r--   1 cloudera supergroup        146 2022-04-17 07:04 /user/shreya/shreya_sparkdata.txt
[cloudera@quickstart ~]$ █
```

5. **Open the spark in Scala mode**

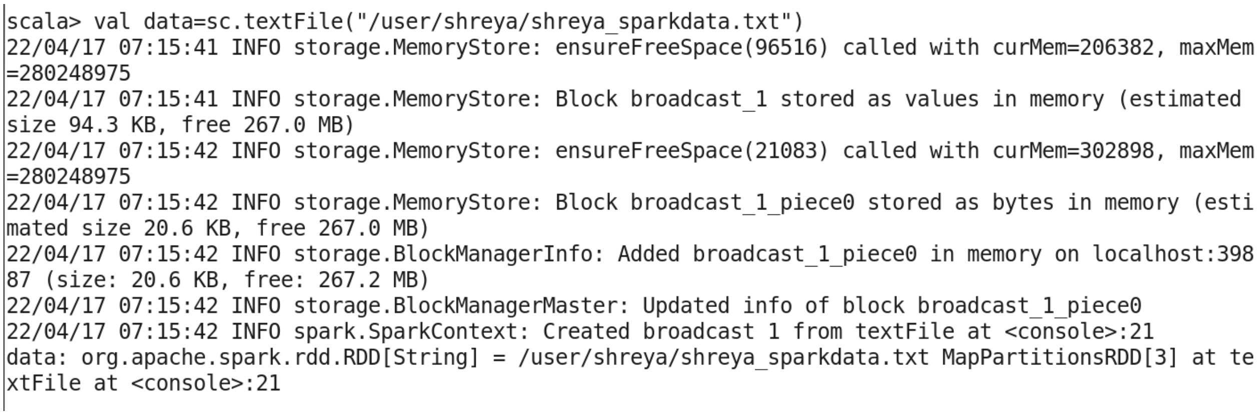
   $ spark-shell



```
[cloudera@quickstart ~]$ spark-shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/jars/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBind
er.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/04/17 07:08:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform
... using builtin-java classes where applicable
22/04/17 07:08:47 INFO spark.SecurityManager: Changing view acls to: cloudera
22/04/17 07:08:47 INFO spark.SecurityManager: Changing modify acls to: cloudera
22/04/17 07:08:47 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disa
bled; users with view permissions: Set(cloudera); users with modify permissions: Set(cloudera)
22/04/17 07:08:47 INFO spark.HttpServer: Starting HTTP Server
22/04/17 07:08:48 INFO server.Server: jetty-8.y.z-SNAPSHOT
22/04/17 07:08:48 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:47134
22/04/17 07:08:48 INFO util.Utils: Successfully started service 'HTTP class server' on port 47134.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/
```

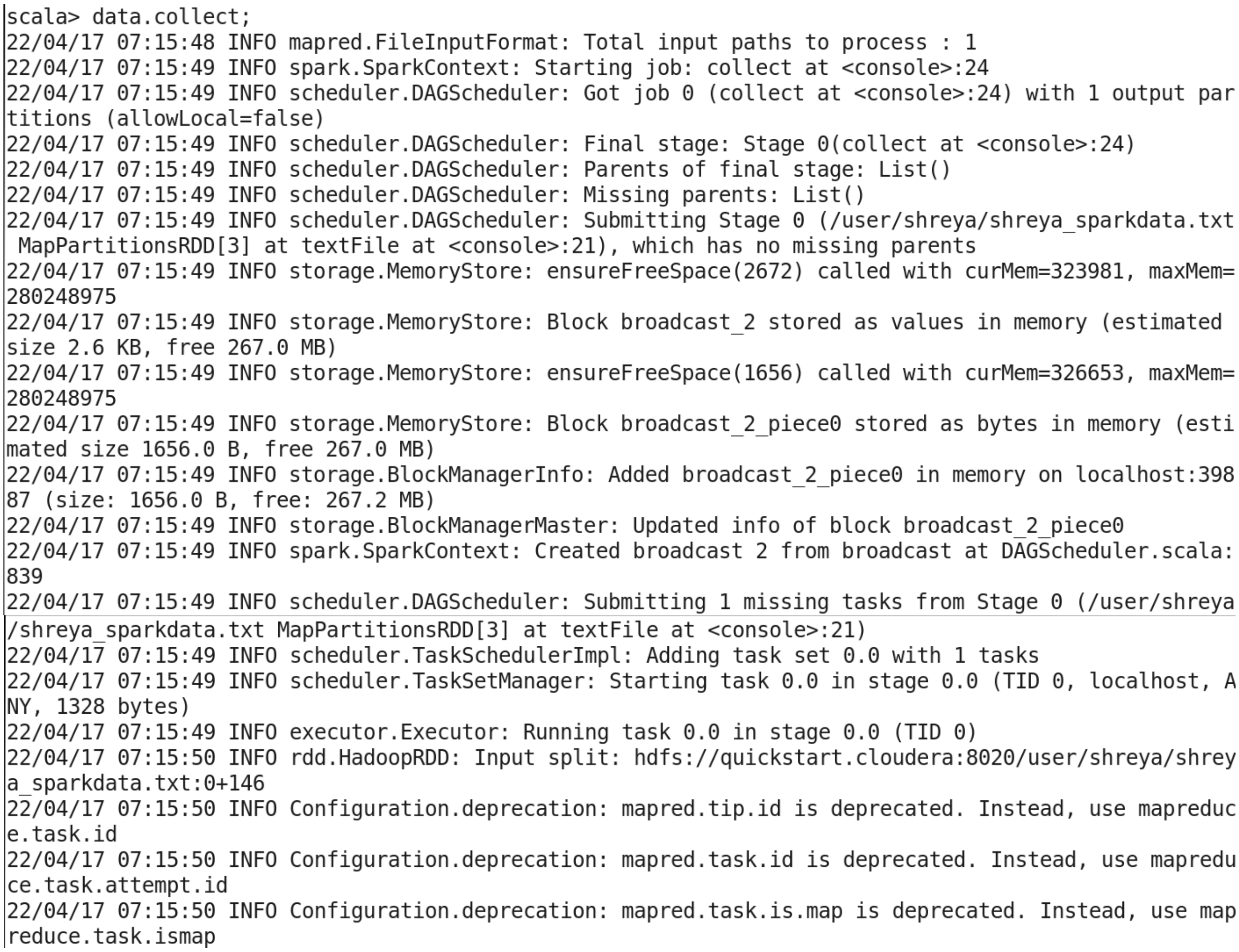
6. **Create an RDD, pass the file that contains the data**

scala> val data=sc.textFile("/user/shreya/shreya_sparkdata.txt")

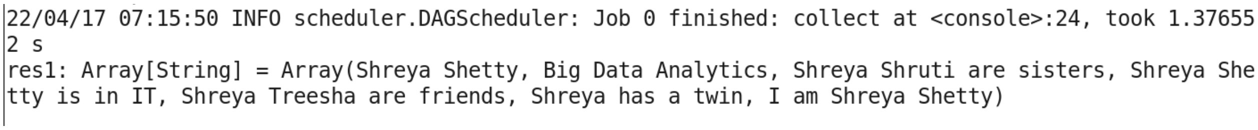


```
scala> val data=sc.textFile("/user/shreya/shreya_sparkdata.txt")
22/04/17 07:15:41 INFO storage.MemoryStore: ensureFreeSpace(96516) called with curMem=206382, maxMem
=280248975
22/04/17 07:15:41 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated
size 94.3 KB, free 267.0 MB)
22/04/17 07:15:42 INFO storage.MemoryStore: ensureFreeSpace(21083) called with curMem=302898, maxMem
=280248975
22/04/17 07:15:42 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (esti
mated size 20.6 KB, free 267.0 MB)
22/04/17 07:15:42 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:398
87 (size: 20.6 KB, free: 267.2 MB)
22/04/17 07:15:42 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
22/04/17 07:15:42 INFO spark.SparkContext: Created broadcast 1 from textFile at <console>:21
data: org.apache.spark.rdd.RDD[String] = /user/shreya/shreya_sparkdata.txt MapPartitionsRDD[3] at te
xtFile at <console>:21
```

7. **Read the generated result by using the following command**
scala> data.collect;



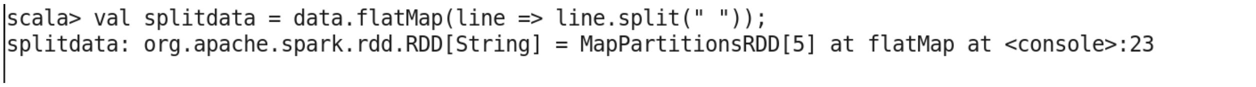
```
scala> data.collect;
22/04/17 07:15:48 INFO mapred.FileInputFormat: Total input paths to process : 1
22/04/17 07:15:49 INFO spark.SparkContext: Starting job: collect at <console>:24
22/04/17 07:15:49 INFO scheduler.DAGScheduler: Got job 0 (collect at <console>:24) with 1 output par
titions (allowLocal=false)
22/04/17 07:15:49 INFO scheduler.DAGScheduler: Final stage: Stage 0(collect at <console>:24)
22/04/17 07:15:49 INFO scheduler.DAGScheduler: Parents of final stage: List()
22/04/17 07:15:49 INFO scheduler.DAGScheduler: Missing parents: List()
22/04/17 07:15:49 INFO scheduler.DAGScheduler: Submitting Stage 0 (/user/shreya/shreya_sparkdata.txt
 MapPartitionsRDD[3] at textFile at <console>:21), which has no missing parents
22/04/17 07:15:49 INFO storage.MemoryStore: ensureFreeSpace(2672) called with curMem=323981, maxMem=
280248975
22/04/17 07:15:49 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated
size 2.6 KB, free 267.0 MB)
22/04/17 07:15:49 INFO storage.MemoryStore: ensureFreeSpace(1656) called with curMem=326653, maxMem=
280248975
22/04/17 07:15:49 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (esti
mated size 1656.0 B, free 267.0 MB)
22/04/17 07:15:49 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:398
87 (size: 1656.0 B, free: 267.2 MB)
22/04/17 07:15:49 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
22/04/17 07:15:49 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:
839
22/04/17 07:15:49 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 0 (/user/shreya
/shreya_sparkdata.txt MapPartitionsRDD[3] at textFile at <console>:21)
22/04/17 07:15:49 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
22/04/17 07:15:49 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, A
NY, 1328 bytes)
22/04/17 07:15:49 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
22/04/17 07:15:50 INFO rdd.HadoopRDD: Input split: hdfs://quickstart.cloudera:8020/user/shreya/shrey
a_sparkdata.txt:0+146
22/04/17 07:15:50 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduc
e.task.id
22/04/17 07:15:50 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapredu
ce.task.attempt.id
22/04/17 07:15:50 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use map
reduce.task.ismap
```



```
22/04/17 07:15:50 INFO scheduler.DAGScheduler: Job 0 finished: collect at <console>:24, took 1.37655
2 s
res1: Array[String] = Array(Shreya Shetty, Big Data Analytics, Shreya Shruti are sisters, Shreya She
tty is in IT, Shreya Treesha are friends, Shreya has a twin, I am Shreya Shetty)
```

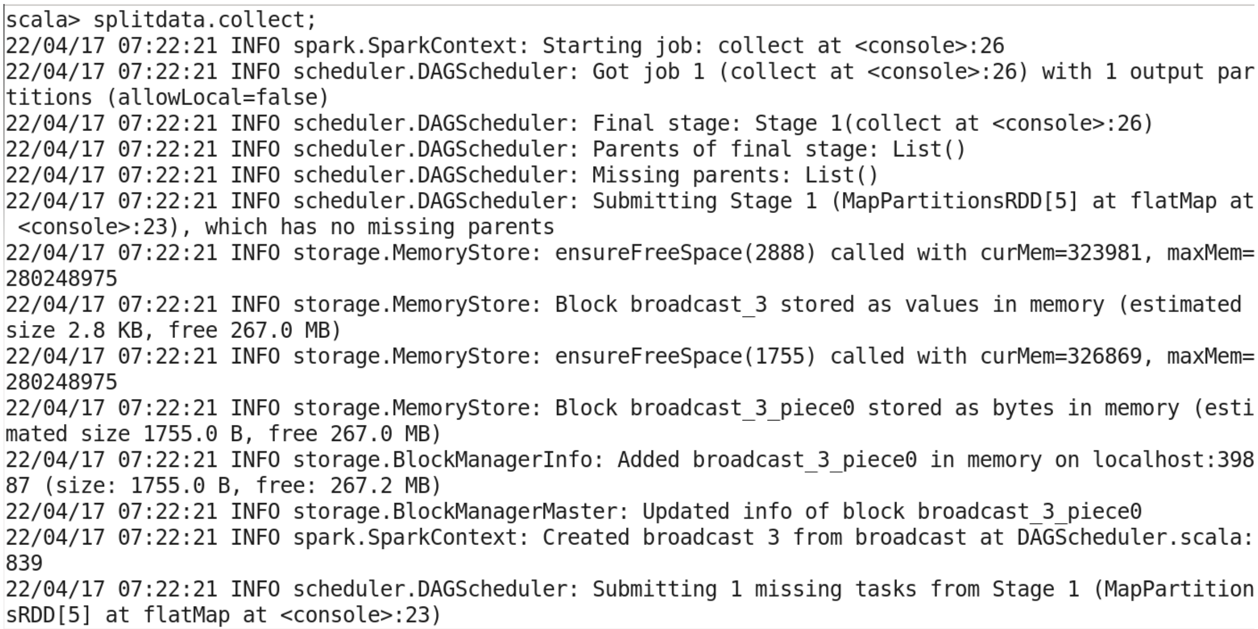
8. **Split the existing data in the form of individual words using the following command**

   scala> val splitdata = data.flatMap(line => line.split(" "));

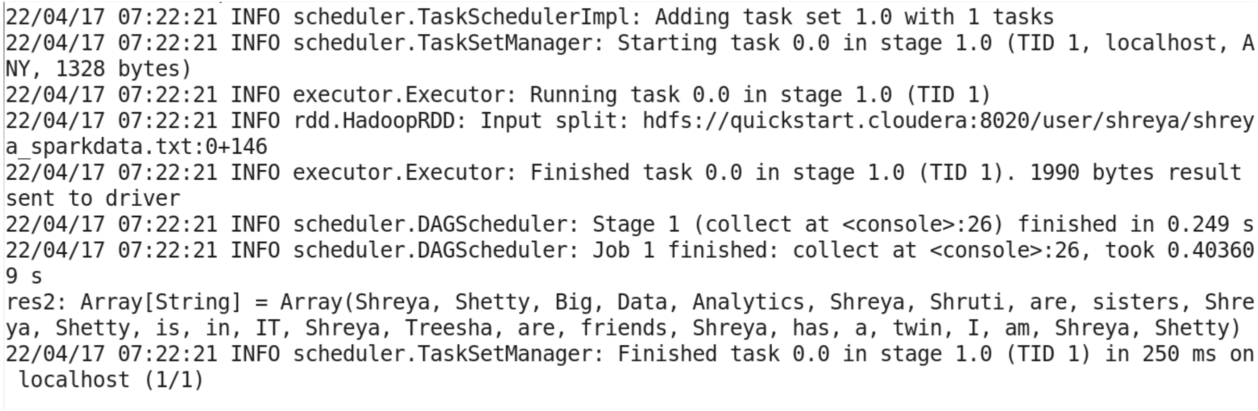


```
scala> val splitdata = data.flatMap(line => line.split(" "));
splitdata: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[5] at flatMap at <console>:23
```

9. **Read the generated result by using the following command**
   scala> splitdata.collect;



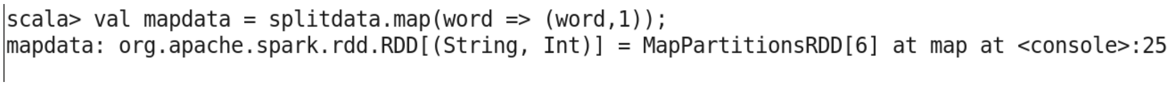
```
scala> splitdata.collect;
22/04/17 07:22:21 INFO spark.SparkContext: Starting job: collect at <console>:26
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Got job 1 (collect at <console>:26) with 1 output par
titions (allowLocal=false)
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Final stage: Stage 1(collect at <console>:26)
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Parents of final stage: List()
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Missing parents: List()
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Submitting Stage 1 (MapPartitionsRDD[5] at flatMap at
 <console>:23), which has no missing parents
22/04/17 07:22:21 INFO storage.MemoryStore: ensureFreeSpace(2888) called with curMem=323981, maxMem=
280248975
22/04/17 07:22:21 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated
size 2.8 KB, free 267.0 MB)
22/04/17 07:22:21 INFO storage.MemoryStore: ensureFreeSpace(1755) called with curMem=326869, maxMem=
280248975
22/04/17 07:22:21 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (esti
mated size 1755.0 B, free 267.0 MB)
22/04/17 07:22:21 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:398
87 (size: 1755.0 B, free: 267.2 MB)
22/04/17 07:22:21 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
22/04/17 07:22:21 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:
839
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 1 (MapPartition
sRDD[5] at flatMap at <console>:23)
```



```
22/04/17 07:22:21 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
22/04/17 07:22:21 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, A
NY, 1328 bytes)
22/04/17 07:22:21 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
22/04/17 07:22:21 INFO rdd.HadoopRDD: Input split: hdfs://quickstart.cloudera:8020/user/shreya/shrey
a_sparkdata.txt:0+146
22/04/17 07:22:21 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 1990 bytes result
sent to driver
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Stage 1 (collect at <console>:26) finished in 0.249 s
22/04/17 07:22:21 INFO scheduler.DAGScheduler: Job 1 finished: collect at <console>:26, took 0.40360
9 s
res2: Array[String] = Array(Shreya, Shetty, Big, Data, Analytics, Shreya, Shruti, are, sisters, Shre
ya, Shetty, is, in, IT, Shreya, Treesha, are, friends, Shreya, has, a, twin, I, am, Shreya, Shetty)
22/04/17 07:22:21 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 250 ms on
 localhost (1/1)
```

10. Perform the map operation. Here, we are assigning a value 1 to each word.

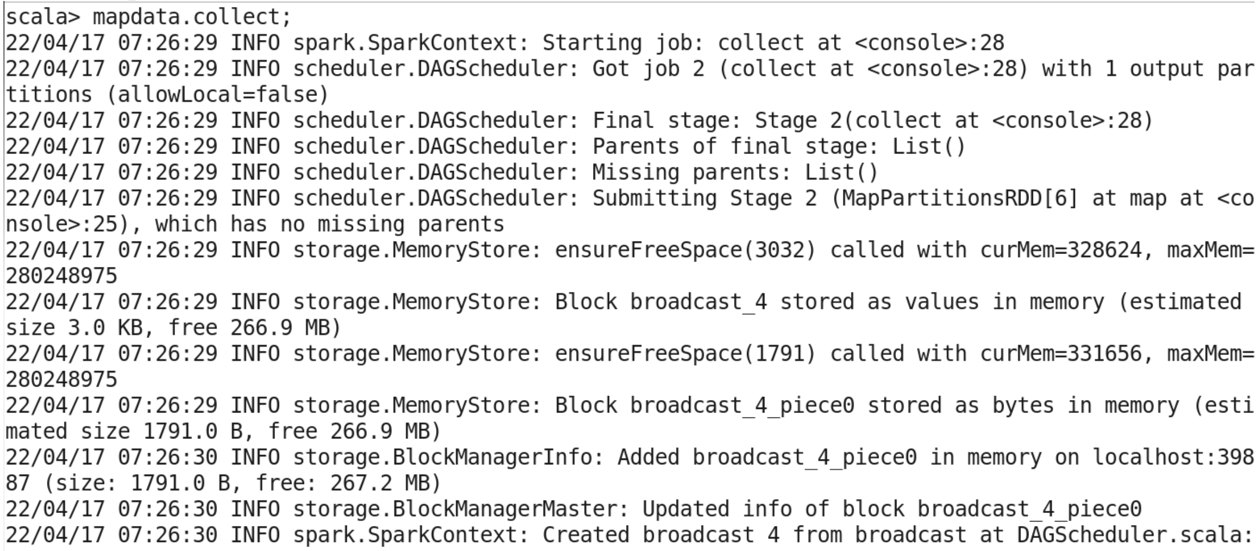
scala> val mapdata = splitdata.map(word => (word,1));



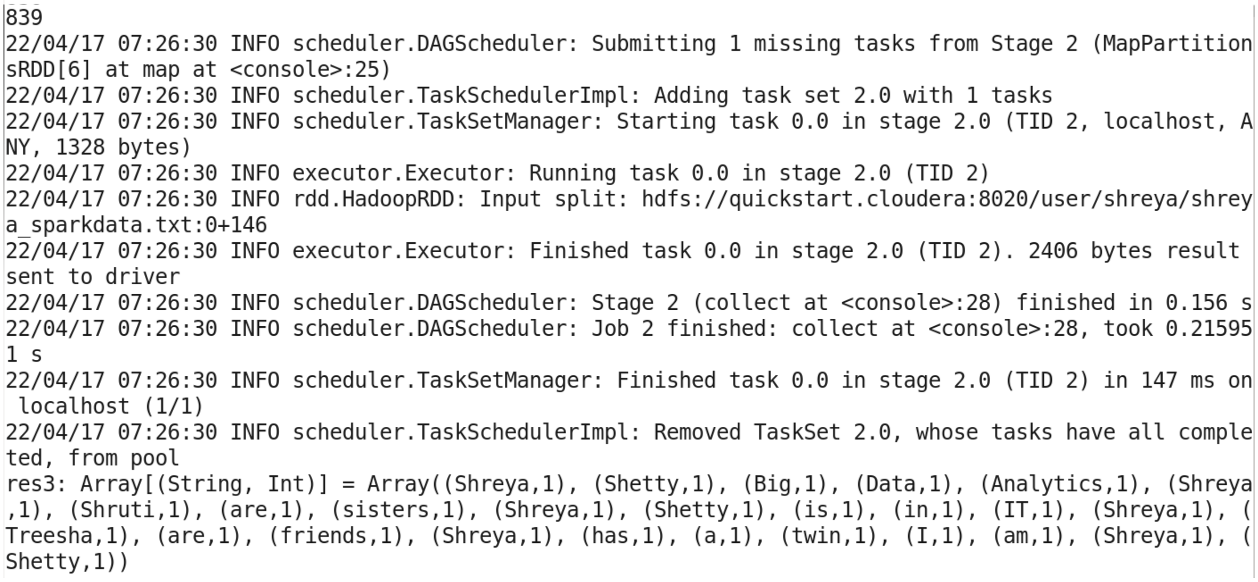
```
scala> val mapdata = splitdata.map(word => (word,1));
mapdata: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[6] at map at <console>:25
```

11. Read the generated result by using the following command

scala> mapdata.collect;



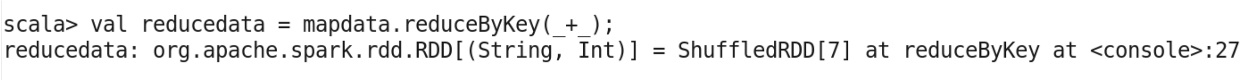
```
scala> mapdata.collect;
22/04/17 07:26:29 INFO spark.SparkContext: Starting job: collect at <console>:28
22/04/17 07:26:29 INFO scheduler.DAGScheduler: Got job 2 (collect at <console>:28) with 1 output par
titions (allowLocal=false)
22/04/17 07:26:29 INFO scheduler.DAGScheduler: Final stage: Stage 2(collect at <console>:28)
22/04/17 07:26:29 INFO scheduler.DAGScheduler: Parents of final stage: List()
22/04/17 07:26:29 INFO scheduler.DAGScheduler: Missing parents: List()
22/04/17 07:26:29 INFO scheduler.DAGScheduler: Submitting Stage 2 (MapPartitionsRDD[6] at map at <co
nsole>:25), which has no missing parents
22/04/17 07:26:29 INFO storage.MemoryStore: ensureFreeSpace(3032) called with curMem=328624, maxMem=
280248975
22/04/17 07:26:29 INFO storage.MemoryStore: Block broadcast_4 stored as values in memory (estimated
size 3.0 KB, free 266.9 MB)
22/04/17 07:26:29 INFO storage.MemoryStore: ensureFreeSpace(1791) called with curMem=331656, maxMem=
280248975
22/04/17 07:26:29 INFO storage.MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (esti
mated size 1791.0 B, free 266.9 MB)
22/04/17 07:26:30 INFO storage.BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:398
87 (size: 1791.0 B, free: 267.2 MB)
22/04/17 07:26:30 INFO storage.BlockManagerMaster: Updated info of block broadcast_4_piece0
22/04/17 07:26:30 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:
```



```
839
22/04/17 07:26:30 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 2 (MapPartition
sRDD[6] at map at <console>:25)
22/04/17 07:26:30 INFO scheduler.TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
22/04/17 07:26:30 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, A
NY, 1328 bytes)
22/04/17 07:26:30 INFO executor.Executor: Running task 0.0 in stage 2.0 (TID 2)
22/04/17 07:26:30 INFO rdd.HadoopRDD: Input split: hdfs://quickstart.cloudera:8020/user/shreya/shrey
a_sparkdata.txt:0+146
22/04/17 07:26:30 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 2). 2406 bytes result
sent to driver
22/04/17 07:26:30 INFO scheduler.DAGScheduler: Stage 2 (collect at <console>:28) finished in 0.156 s
22/04/17 07:26:30 INFO scheduler.DAGScheduler: Job 2 finished: collect at <console>:28, took 0.21595
1 s
22/04/17 07:26:30 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 147 ms on
 localhost (1/1)
22/04/17 07:26:30 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all comple
ted, from pool
res3: Array[(String, Int)] = Array((Shreya,1), (Shetty,1), (Big,1), (Data,1), (Analytics,1), (Shreya
,1), (Shruti,1), (are,1), (sisters,1), (Shreya,1), (Shetty,1), (is,1), (in,1), (IT,1), (Shreya,1), (
Treesha,1), (are,1), (friends,1), (Shreya,1), (has,1), (a,1), (twin,1), (I,1), (am,1), (Shreya,1), (
Shetty,1))
```

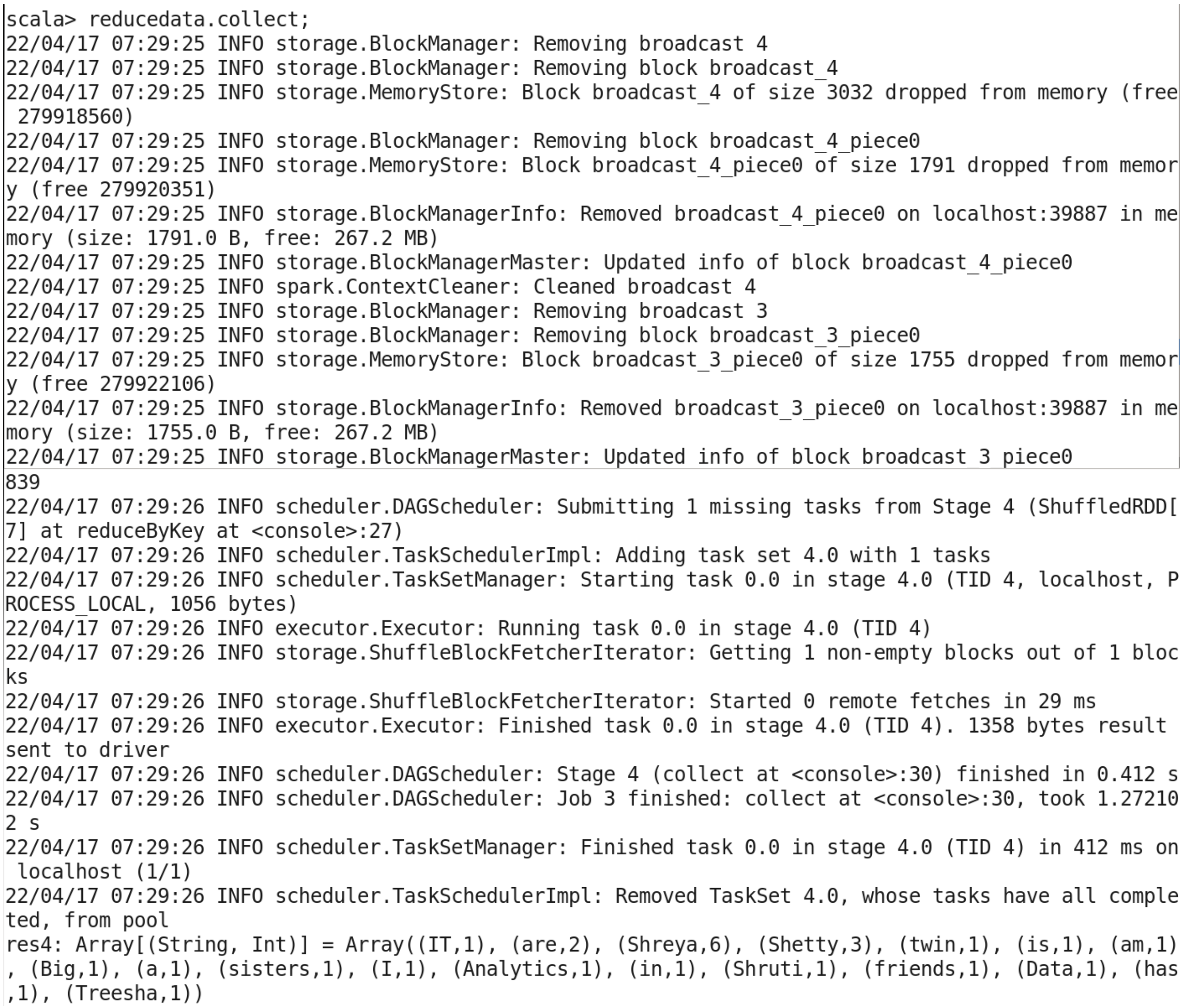
12. Perform the reduce operation. Here, we are summarizing the generated data.

scala> val reducedata = mapdata.reduceByKey(_+_);



```
scala> val reducedata = mapdata.reduceByKey(_+_);
reducedata: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[7] at reduceByKey at <console>:27
```

13. **Read the generated result by using the following command.**
scala> reducedata.collect;



```
scala> reducedata.collect;
22/04/17 07:29:25 INFO storage.BlockManager: Removing broadcast 4
22/04/17 07:29:25 INFO storage.BlockManager: Removing block broadcast_4
22/04/17 07:29:25 INFO storage.MemoryStore: Block broadcast_4 of size 3032 dropped from memory (free
 279918560)
22/04/17 07:29:25 INFO storage.BlockManager: Removing block broadcast_4_piece0
22/04/17 07:29:25 INFO storage.MemoryStore: Block broadcast_4_piece0 of size 1791 dropped from memor
y (free 279920351)
22/04/17 07:29:25 INFO storage.BlockManagerInfo: Removed broadcast_4_piece0 on localhost:39887 in me
mory (size: 1791.0 B, free: 267.2 MB)
22/04/17 07:29:25 INFO storage.BlockManagerMaster: Updated info of block broadcast_4_piece0
22/04/17 07:29:25 INFO spark.ContextCleaner: Cleaned broadcast 4
22/04/17 07:29:25 INFO storage.BlockManager: Removing broadcast 3
22/04/17 07:29:25 INFO storage.BlockManager: Removing block broadcast_3_piece0
22/04/17 07:29:25 INFO storage.MemoryStore: Block broadcast_3_piece0 of size 1755 dropped from memor
y (free 279922106)
22/04/17 07:29:25 INFO storage.BlockManagerInfo: Removed broadcast_3_piece0 on localhost:39887 in me
mory (size: 1755.0 B, free: 267.2 MB)
22/04/17 07:29:25 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
839
22/04/17 07:29:26 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 4 (ShuffledRDD[
7] at reduceByKey at <console>:27)
22/04/17 07:29:26 INFO scheduler.TaskSchedulerImpl: Adding task set 4.0 with 1 tasks
22/04/17 07:29:26 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 4.0 (TID 4, localhost, P
ROCESS_LOCAL, 1056 bytes)
22/04/17 07:29:26 INFO executor.Executor: Running task 0.0 in stage 4.0 (TID 4)
22/04/17 07:29:26 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 bloc
ks
22/04/17 07:29:26 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 29 ms
22/04/17 07:29:26 INFO executor.Executor: Finished task 0.0 in stage 4.0 (TID 4). 1358 bytes result
sent to driver
22/04/17 07:29:26 INFO scheduler.DAGScheduler: Stage 4 (collect at <console>:30) finished in 0.412 s
22/04/17 07:29:26 INFO scheduler.DAGScheduler: Job 3 finished: collect at <console>:30, took 1.27210
2 s
22/04/17 07:29:26 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 4.0 (TID 4) in 412 ms on
 localhost (1/1)
22/04/17 07:29:26 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all comple
ted, from pool
res4: Array[(String, Int)] = Array((IT,1), (are,2), (Shreya,6), (Shetty,3), (twin,1), (is,1), (am,1)
, (Big,1), (a,1), (sisters,1), (I,1), (Analytics,1), (in,1), (Shruti,1), (friends,1), (Data,1), (has
,1), (Treesha,1))
```

Hence, we got the desired output i.e. count of each word present in the given text file using Apache Spark.

CONCLUSION:

In this experiment, I implemented word count i.e. count of all the words from given text, using Apache Spark for the given input. I found the frequency of each word that exists in a particular file and used Scala language to perform Spark operations.