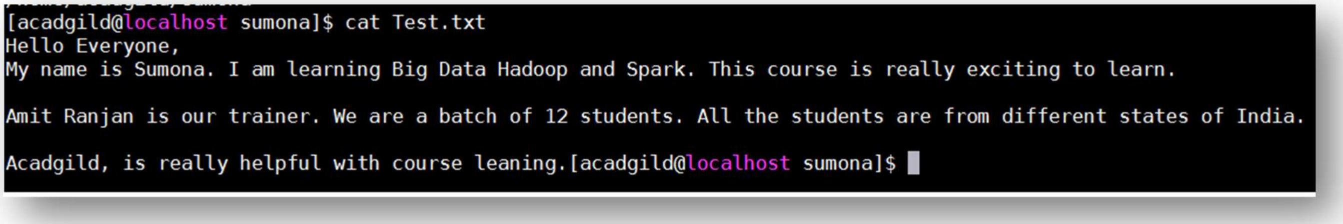
Assignment 17.1

Created a text file **Test.txt** under **/home/acadgild/Sumona**



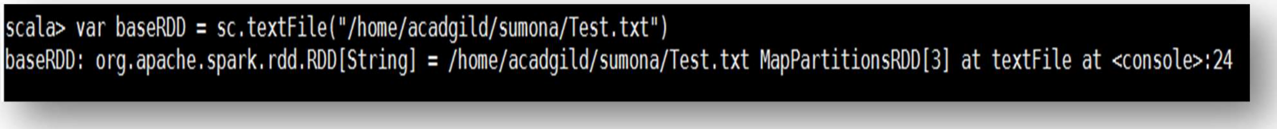
```
[acadgild@localhost sumona]$ cat Test.txt
Hello Everyone,
My name is Sumona. I am learning Big Data Hadoop and Spark. This course is really exciting to learn.

Amit Ranjan is our trainer. We are a batch of 12 students. All the students are from different states of India.

Acadgild, is really helpful with course leaning.[acadgild@localhost sumona]$ 
```

1. Write a program to read a text file and print the number of rows of data in the document.

The command used to read the text file is:

var baseRDD = sc.textFile("/home/acadgild/sumona/Test.txt")

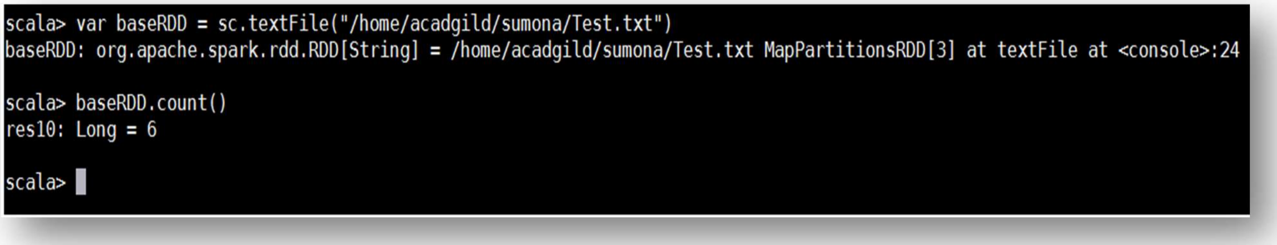
baseRDD is a variable into which we read the text file from the file location.



```
scala> var baseRDD = sc.textFile("/home/acadgild/sumona/Test.txt")
baseRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/sumona/Test.txt MapPartitionsRDD[3] at textFile at <console>:24
```

Command used to count the number of Lines in the text file is

baseRDD.count()



```
scala> var baseRDD = sc.textFile("/home/acadgild/sumona/Test.txt")
baseRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/sumona/Test.txt MapPartitionsRDD[3] at textFile at <console>:24

scala> baseRDD.count()
res10: Long = 6

scala> 
```

2. Write a program to read a text file and print the number of words in the document.

   The command used to read the text file is:

var baseRDD = sc.textFile("/home/acadgild/sumona/Test.txt")

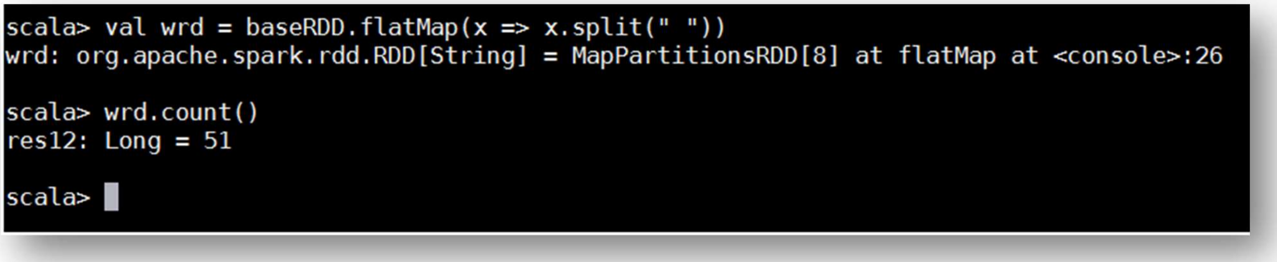
Command used to count the number of words in the text file:

- First, since the words in the text file are separated with a space(“ “), hence we will split the text file using the **flatMap** split command

val wrd = baseRDD.flatMap(x => x.split(" "))

- And then now, we shall get the count of words in the text file using the following command:

wrd.count()



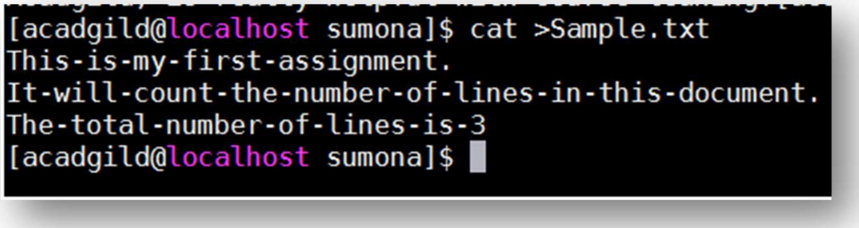
```
scala> val wrd = baseRDD.flatMap(x => x.split(" "))
wrd: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[8] at flatMap at <console>:26

scala> wrd.count()
res12: Long = 51

scala> 
```

3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

Sample document :

This-is-my-first-assignment.
It-will-count-the-number-of-lines-in-this-document.
The-total-number-of-lines-is-3

Created the Sample.txt file



```
[acadgild@localhost sumona]$ cat >Sample.txt
This-is-my-first-assignment.
It-will-count-the-number-of-lines-in-this-document.
The-total-number-of-lines-is-3
[acadgild@localhost sumona]$ 
```

a. First we will read the Sample.txt file from the local file system to an RDD using the command

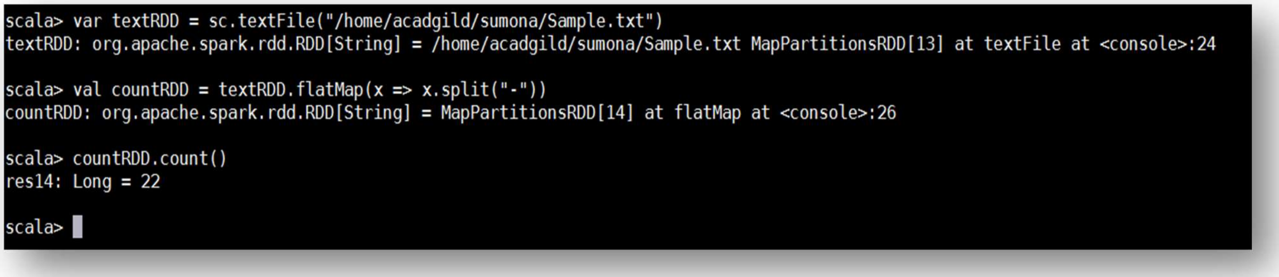
var textRDD = sc.textFile("/home/acadgild/sumona/Sample.txt")

b. Now, since the words in the Sample.txt file are separated by a dash(-), we will first split the text file using the command:

val countRDD = textRDD.flatMap(x => x.split("-"))

c. Now we shall count the number of words using the command;

countRDD.count()



```
scala> var textRDD = sc.textFile("/home/acadgild/sumona/Sample.txt")
textRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/sumona/Sample.txt MapPartitionsRDD[13] at textFile at <console>:24

scala> val countRDD = textRDD.flatMap(x => x.split("-"))
countRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[14] at flatMap at <console>:26

scala> countRDD.count()
res14: Long = 22

scala> 
```