

## Assignment 17.1

Created a text file Test.txt under /home/acadgild/kapil

### 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/kapil/Test.txt")
```

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

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

```
baseRDD.count()
```

### 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/kapil/Test.txt")
```

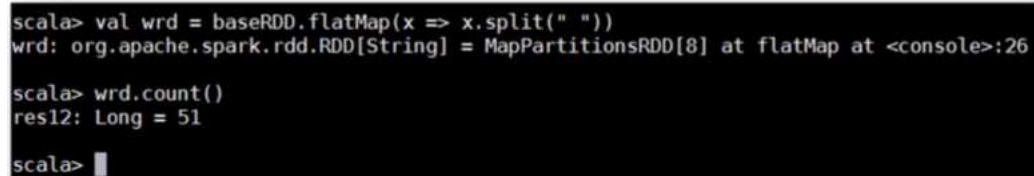
Command used to count the number of words in the text file:

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

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/kapil/Sample.txt")`

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("-"))
```

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

```
countRDD.count()
```

```
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> █
```