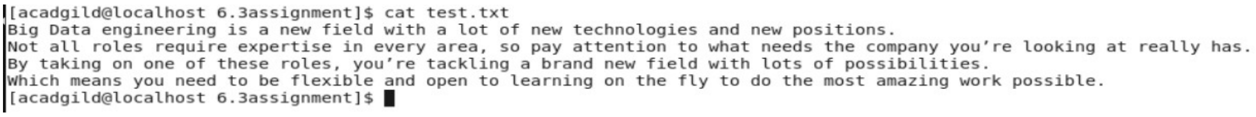
Bigdata Assignment 6.3

Created a text file 'test.txt' . The content was checked with -
cat test.txt

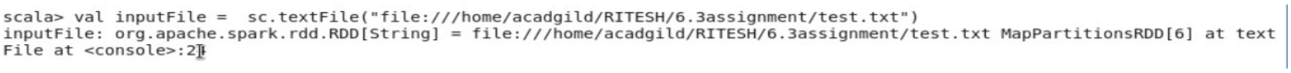


```
[acadgild@localhost 6.3assignment]$ cat test.txt
Big Data engineering is a new field with a lot of new technologies and new positions.
Not all roles require expertise in every area, so pay attention to what needs the company you’re looking at really has.
By taking on one of these roles, you’re tackling a brand new field with lots of possibilities.
Which means you need to be flexible and open to learning on the fly to do the most amazing work possible.
[acadgild@localhost 6.3assignment]$ █
```

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

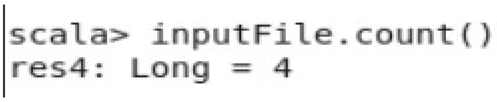
Solution -

- RDD is created from a text file.
  val inputFile = sc.text(“file:///home/acadgild/RITESH/6.3assignment/test.txt”)



```
scala> val inputFile =  sc.textFile("file:///home/acadgild/RITESH/6.3assignment/test.txt")
inputFile: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/RITESH/6.3assignment/test.txt MapPartitionsRDD[6] at text
File at <console>:24
```

- For calculating rows we have to use count function.
  inputFile.count()



```
scala> inputFile.count()
res4: Long = 4
```

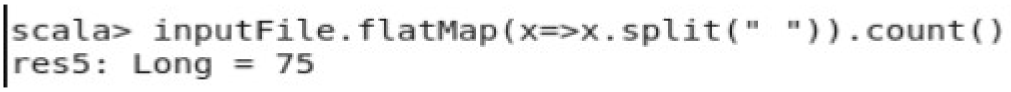
Output - 4
From the 1st scrrenshot it is evident that , there are 4 rows in the text file.

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

Solution -

- Splitted the strings where space is available and used the flatMap function to flattened the list of strings RDD and count function is used to calculate no of strings.

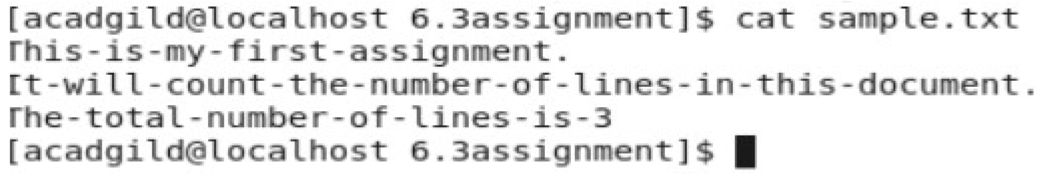
inputFile.flatMap(x=>x.split(“ “)).count()



```
scala> inputFile.flatMap(x=>x.split(" ")).count()
res5: Long = 75
```

Output - 75

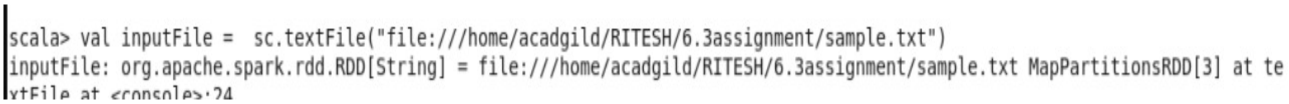
3.



```
[acadgild@localhost 6.3assignment]$ 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 6.3assignment]$ █
```

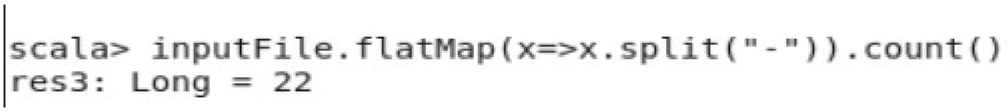
Solution -

- RDD is created from a text file.
  val inputFile = sc.text(“file:///home/acadgild/RITESH/6.3assignment/sample.txt”)



```
scala> val inputFile =  sc.textFile("file:///home/acadgild/RITESH/6.3assignment/sample.txt")
inputFile: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/RITESH/6.3assignment/sample.txt MapPartitionsRDD[3] at te
xtFile at <console>:24
```

- Splitted the strings where '-' is available and used the flatMap function to flattened the list of strings RDD and count function is used to calculate no of strings.



```
scala> inputFile.flatMap(x=>x.split("-")).count()
res3: Long = 22
```

Output - 22