

## **CS5543 Real-Time Big Data Analytics**

### **LAB ASSIGNMENT #2**

**Akhilesh Gattu**

#### **1) Question**

##### **Spark Programming:**

**Write a spark program with an interesting use case using text data as the input and program should have at least Two Spark Transformations and Two Spark Actions.**

##### **Transformations:**

###### **Map:**

Map function returns a new distributed dataset where each element is given as input through a function.

###### **SortByKey:**

SortByKey function pairs up the key value pairs where in K is ordered. This transformation returns the key value pairs in ascending or descending order as specified.

###### **ReduceByKey**

This transformation returns key value pairs where all the values for each key are reduced. Number of reduce functions could be specified in the augment.

##### **Actions:**

###### **Take(n):**

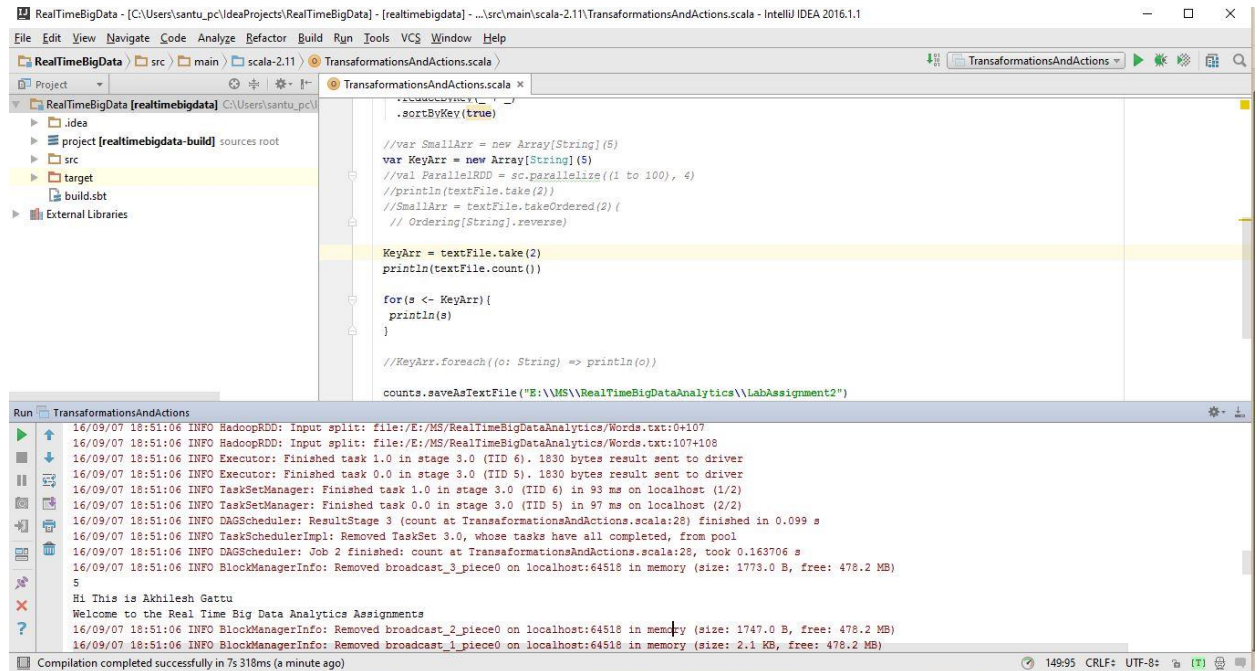
The Take action in spark returns the array with first n elements in the dataset.

###### **Count():**

The count action in spark returns the number of elements in the dataset.

## Screenshots:

The actions used are count and take



The screenshot shows the IntelliJ IDEA IDE with a Scala file named `TransaformationsAndActions.scala` open. The file contains the following code:

```
.sortByKey(true)

//var SmallArr = new Array[String](5)
var KeyArr = new Array[String](5)
//val ParallelRDD = sc.parallelize((1 to 100), 4)
//println(textFile.take(2))
//SmallArr = textFile.takeOrdered(2){
// Ordering[String].reverse)

KeyArr = textFile.take(2)
println(textFile.count())

for(s <- KeyArr){
  println(s)
}

//KeyArr.foreach((o: String) => println(o))

counts.saveAsTextFile("E:\\MS\\RealTimeBigDataAnalytics\\LabAssignment2")
```

The Run console shows the following output:

```
16/09/07 18:51:06 INFO HadoopRDD: Input split: file:/E:/MS/RealTimeBigDataAnalytics/Words.txt:0+107
16/09/07 18:51:06 INFO HadoopRDD: Input split: file:/E:/MS/RealTimeBigDataAnalytics/Words.txt:107+108
16/09/07 18:51:06 INFO Executor: Finished task 1.0 in stage 3.0 (TID 6). 1830 bytes result sent to driver
16/09/07 18:51:06 INFO Executor: Finished task 0.0 in stage 3.0 (TID 5). 1830 bytes result sent to driver
16/09/07 18:51:06 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 6) in 93 ms on localhost (1/2)
16/09/07 18:51:06 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 5) in 97 ms on localhost (2/2)
16/09/07 18:51:06 INFO DAGScheduler: ResultStage 3 (count at TransaformationsAndActions.scala:28) finished in 0.099 s
16/09/07 18:51:06 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
16/09/07 18:51:06 INFO DAGScheduler: Job 2 finished: count at TransaformationsAndActions.scala:28, took 0.163706 s
16/09/07 18:51:06 INFO BlockManagerInfo: Removed broadcast_3_piece0 on localhost:64518 in memory (size: 1773.0 B, free: 478.2 MB)
5
Hi This is Akhilesh Gattu
Welcome to the Real Time Big Data Analytics Assignments
16/09/07 18:51:06 INFO BlockManagerInfo: Removed broadcast_2_piece0 on localhost:64518 in memory (size: 1747.0 B, free: 478.2 MB)
16/09/07 18:51:06 INFO BlockManagerInfo: Removed broadcast_1_piece0 on localhost:64518 in memory (size: 2.1 KB, free: 478.2 MB)
Compilation completed successfully in 7s 318ms (a minute ago)
```

The input given is

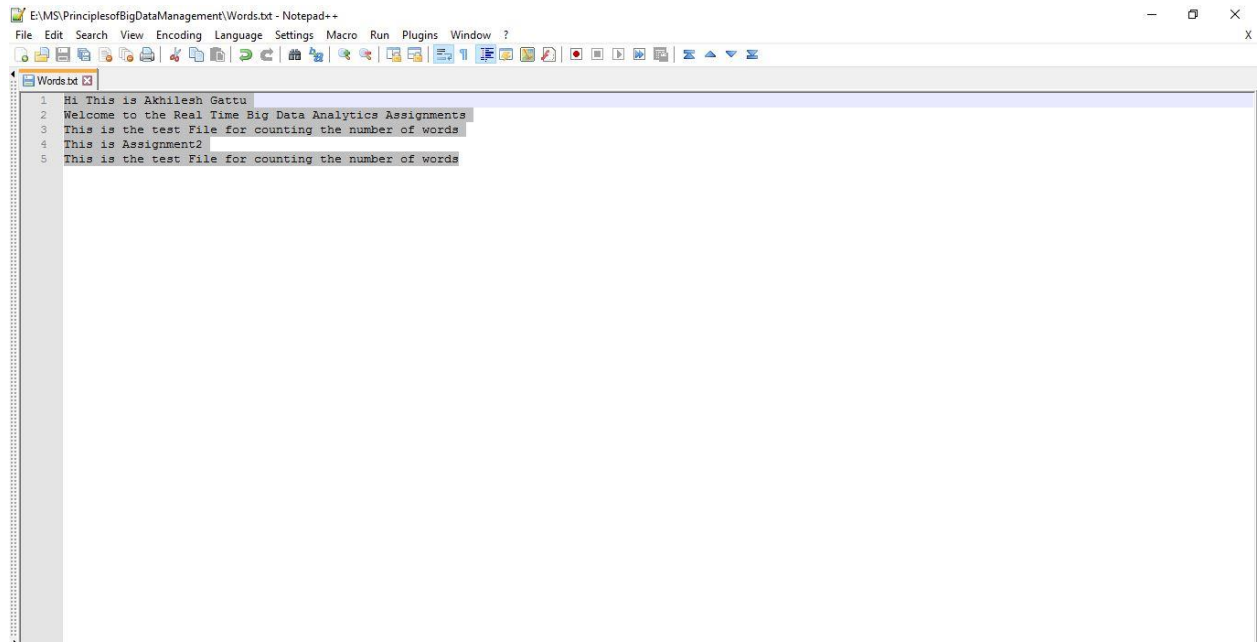
Hi This is Akhilesh Gattu

Welcome to Big Data Analytics Assignments

This is the text File for counting the number of words

This is Assignment2

This is the text File for counting the number of words



The screenshot shows a Notepad++ window with the file `Words.txt` open. The file contains the following text:

```
1 Hi This is Akhilesh Gattu
2 Welcome to the Real Time Big Data Analytics Assignments
3 This is the test File for counting the number of words
4 This is Assignment2
5 This is the test File for counting the number of words
```

### Map Reduce Diagram:

