*Spark*

**A C A D G I L D Page 1**

Session 4: Spark Shell

Assignment 3*Spark*

**A C A D G I L D Page 2**

*Session 4: Spark Shell*

*Assignment 3*

**Table of Contents**

1. Introduction .......................................................................................................................................... 3

2. Objective ............................................................................................................................................... 3

3. Prerequisites ......................................................................................................................................... 3

4. Associated Data Files ............................................................................................................................ 3

5. Problem Statement ............................................................................................................................... 3

6. Approximate Time to Complete Task ................................................................................................... 3 *Spark*

**A C A D G I L D Page 3**

**1. Introduction**

In this assignment you need to write the code for the given question.

**2. Objective**

This assignment will help you to understand Big Data basics.

**3. Prerequisites**

None

**4. Associated Data Files**

N/A

**5. Problem Statement**

1. Execute wordcount program in Spark Scala shell.

scala> val input=sc.textFile("/myfolder/DataSet/test.txt")

input: org.apache.spark.rdd.RDD[String] = /myfolder/DataSet/test.txt MapPartitionsRDD[3] at textFile at <console>:24

scala> val words = input.flatMap(line => line.split(" "))

words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[4] at flatMap at <console>:26

scala> val counts = words.map(word => (word, 1)).reduceByKey{case (x, y) => x + y}

counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[6] at reduceByKey at <console>:28

scala> counts.foreach(println)

((CentOS,1)

(vendor.,1)

(derived,1)

(is,2)

(rebuilt,,1)

(Enterprise-class,1)

(small,1)

(American,1)

(contactable,1)

(extensive,1)

(policy,1)

(mirror,1)

**6. Approximate Time to Complete Task**

30 min