*Spark*

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

Session 3: Spark Core

Assignment 2*Spark*

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

*Session 3: Spark Core*

*Assignment 2*

**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 questions.

**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. Create a simple pairRDD of (1, 2), (3, 4), (3, 6).

scala> val input=sc.parallelize(List((1,2),(3,4),(3,6)))

input: org.apache.spark.rdd.RDD[(Int, Int)] = ParallelCollectionRDD[0] at parallelize at <console>:27

2. Transform an RDD of ("a","b","c","d","e") to PairRDD (a,0), (b,1), (c,2), (d,3), (e,4)

scala> val first=sc.parallelize(List("a","b","c","d","e"),3)

first: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[21] at parallelize at <console>:27

scala>val second=first.zipWithIndex

second: org.apache.spark.rdd.RDD[(String, Long)] = ZippedWithIndexRDD[1] at zipWithIndex at <console>:29

scala>second.foreach(println)

[Stage 1:> (0 + 0) / 3](a,0)

(b,1)

(c,2)

(d,3)

(e,4)

**6. Approximate Time to Complete Task**

30 min