*Spark*

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

Session 7: Using Spark’s API in Scala, Java and Python

Assignment 1*Spark*

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

*Session 7: Spark’s API in Scala, Java and Python*

*Assignment 1*

**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**

APIs in Spark

**4. Associated Data Files**

N/A

**5. Problem Statement**

1. Store data in sequence file format

|  |
| --- |
| val data = sc.parallelize(List(("RAM", 1), ("RAHIM", 2), ("RAJ", 2))) |
|  |

data.saveAsSequenceFile("/tmp/output")

2. Read sequence file data

|  |
| --- |
| import org.apache.hadoop.io.Text |
|  |

|  |
| --- |
| import org.apache.hadoop.io.IntWritable |
|  |

|  |
| --- |
| val outputResult = sc.sequenceFile("/tmp/output/part-00001", classOf[Text], classOf[IntWritable]). map{case (x, y) => (x.toString, y.get())} |
|  |

outputResult.collect

**6. Approximate Time to Complete Task**

30 min