*Spark*

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

Session 5: Spark Core Part 2

Assignment 1*Spark*

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

*Session 5: Spark Core Part 2*

*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 give brief answers to 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. What is RDD Lineage?

**RDD Lineage** (aka **RDD** operator graph or **RDD** dependency graph) is a graph of all the parent **RDDs** of a **RDD**. It is built as a result of applying transformations to the **RDD** and creates a logical execution plan



2. What is Spark behavior in following Scenario?

Consider a cluster of 100 nodes and your application has 10 transformations. One of the node fails or disconnects from the cluster during the execution.

Ans :Parallelism Control the spark behavior in this senario by partition where the data stored(As RDD).

3. How do you control parallelism in applications?

Ans : by Repartitions and coalesce actually change the number of partitions where the data is stored as RDD.

4. What is the difference between map and mapPartitions?

The method map converts each element of the source RDD into a single element of the result RDD by applying a function. mapPartition converts each partition of the source RDD into multiple elements of the result (possibly none).

5. What are the benefits of Spark Architecture?

Ans :1)**Isolation :**

**-> Application are isolated**

**-> Task scheduling per application**

**2)Low-Overhead**

**->10-100 times faster**

**->Small task mitigate effects of data skew.**

**3)Sharing data**

**->Application cannot share data in memory process.**

**->Use an External storage service**

**4) Resource allocation**

**->Static process provisioning for executor,Even Without active Task.**

**6. Approximate Time to Complete Task**

30 min