*Spark*

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

Session 6: In-Memory Management

Assignment 2*Spark*

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

*Session 6: In-Memory Management*

*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 answer the question given.

**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 the difference between cache and persist operation?

Cache operation is when transformations which lead to intermediary result are stored in-memory so that other operation (transformation and actions) that depend on it can have easy access to it rather than always reading from the file whenever an actions is performed. Persist operation is when transformation which lead to intermediary results can either be stored in-memory or on a disk or even both. Sometimes, our intermediary result might be too big for just the ram so the persist operation is enhanced with storing on also disk also. The beautiful thing about the persist operation is that it has the spill option also. Meaning, when you save in-memory, what was stored has the option to spill over to the disk if in-memory reaches its threshold.

1. The equivalent of accumulators in map reduce is known as counters
2. The equivalent of broadcast variables in map reduce is known as distributed cache
3. How spark allocates the memory during processing? Note, processing takes place on the workers in the executors. When processing is taking place, spark allocates a storage memory that stores Apache Spark cached data and for temporary space serialized data and also all the broadcast variables are stored there as cached blocks. After this pictorially, we have the execution memory which stores the objects required during the execution of Spark tasks. After we have the user memory which is the memory pool that remains after the allocation of Spark Memory, and it is completely up to you to use it in a way you like. You can store your own data structures there that would be used in RDD transformations. Finally we have the reserved memory which is the memory reserved by the system, and its size is hardcoded. As of Spark 1.6.0, its value is 300MB, which means that this 300MB of RAM does not participate in Spark memory region size calculations, and its size cannot be changed in any way without Spark recompilation.
4. You recover data from cached RDD’s by undoing it (rdd.uncache())

**6. Approximate Time to Complete Task**

30 min