*Spark*

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

Session 5: Spark Core Part 2

Assignment 2*Spark*

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

*Session 5: Spark Core Part 2*

*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**

https://drive.google.com/open?id=0ByJLBTmJojjzZ3NzM2llZnJ5QVU

The above dataset consists of the details of movies. The dataset description is as follows

Metadata header: id,name,year,rating,duration

**5. Problem Statement**

1. Generate movie name and its duration

movieFile=sc.textFile(“tmp/movie.txt”);

val split = movieFile.filter { x => {if(x.toString().split(",").length >= 6) true else false} }.map(line=>{line.toString().split(",")})

val mov=split.map(x=>(x(1)+" "+x(4)+”,1)).reduceByKey(\_+\_).collect

2. How many distinct movies are there?

Val dis=mov.distinct.collect()foreach(println).

**6. Approximate Time to Complete Task**

30 min