*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

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| val movieFile = sc.textFile("hdfs://manish.com:8020/data/movies.txt") |
|  |

|  |
| --- |
| val header = movieFile.first() |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| val withoutheader= movieFile.filter(row => row != header) |
|  |

|  |
| --- |
| val nameFile = withoutheader.filter { x => {if(x.toString().split(",").length >= 5) true else false} }.map(line=>{line.toString().split(",")}) // splits input line on comma |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| val file = nameFile.map(x=> { (x(1),x(4))}) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| 2>How many distinct movies are there? |
|  |
|  |

|  |
| --- |
| val unique = file.map(i=>i.\_1).distinct |
|  |

|  |
| --- |
| unique.count() |
|  |

|  |
| --- |
| 10 |
|  |

Array(Night Tide, Orphans of the Storm, Nick of Time, The Nightmare Before Christmas, Mother's Boys, The Object of Beauty, Muriel's Wedding, One Magic Christmas, Nosferatu: Original Version)

**6. Approximate Time to Complete Task**

30 min