**Session 17: RDD'S IN SPARK**

**Assignment 17.1**

Problem Statement

1. Write a program to read a text file and print the number of rows of data in the document.

2. Write a program to read a text file and print the number of words in the document.

3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

**Sample document:**

This-is-my-first-assignment.It-will-count-the-number-of-lines-in-this-document.The-total-number-of-lines-is-3

Task1 - Write a program to read a text file and print the number of rows of data in the document.

In this task, we are using a text file **“television.txt”** which has **72** rows as shown below,

****

**Spark Operation**

Read the text file,

***scala>val rows= sc.textFile("/home/acadgild/hadoop/television.txt")***

***scala>rows.count()***

***res0: Long = 72***

****

Task2 - Write a program to read a text file and print the number of words in the document.

In this task, we are using a text file “***Spark\_numberofwords.txt”*** which we created and it has number of words as 83, please see below,

***cat Spark\_numberofwords.txt***

***wc -w Spark\_numberofwords.txt***

****

**Spark Operation**

Read the text file,

***scala>val base = sc.textFile("/home/acadgild/hadoop/Spark\_numberofwords.txt")***

***scala>val words = base.flatMap(word=>word.split(" "))***

***scala>words.count()***

***res5: Long = 83***

****

Task3 - We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

The same file “***Spark\_numberofwords.txt”*** has been modified by placing **“-“between** the words, please see below.

***scala>val base1 = sc.textFile("/home/acadgild/hadoop/Spark\_numberofwords.txt")***

***scala>val words = base1.flatMap(word=>word.split("-"))***

***scala>words.count()***

***res12: Long = 83***

****

***words.collect()***

****