Session 17 : RDD’s in Spark

Assignment 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

**Solution:-**

Starting Spark shell

Command - spark-shell



Created **file1.txt** (space separated words), **file2.txt**(hyphen separated words) and placed them at **/home/acadgild/assignments/Assignment17\_1** in VM

Files- 



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

Loading the text file to RDD –

**scala>** val textData = sc.textFile(“file1.txt”);

Finding count of data rows –

**scala>** textData.count();

Screenshot:-



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

Loading the text file to RDD –

**scala>** val textData = sc.textFile(“file1.txt”);

Since the RDD has already been created. Re-using the same for solution to this problem.

Splitting a line by space and counting words per line

**Scala>** val wordCount = textData.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(\_ + \_);

**scala>** wordCount.foreach(println);

**scala>** wordCount.map(\_.\_2).sum();

Screenshot:-







The command returns that there are **22** **words** in the file1.txt.

1. **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.**

Loading the text file to RDD –

**scala>** val textData2 = sc.textFile(“file2.txt”);

Splitting a line by hyphen and counting words per line.

**Scala>** val wordCount2 = textData2.flatMap(line => line.split("-")).map(word => (word, 1)).reduceByKey(\_ + \_);

**scala>** wordCount2.foreach(println);

**scala>** wordCount2.map(\_.\_2).sum();

Screenshot:-





The command returns that there are **22 words** in the file2.txt.