Session 16: INTRODUCTION TO APACHE SPARK

Assignment 16.1

Student Name:	Abarajithan SA
Course:	Big Data Hadoop & Spark Training
Start Date:	2017-09-09
End Date:	2017-11-26

Assignment 16.1– Perform Apache Spark operations.

Contents

Introduction

In this assignment, we are going to perform some Spark RDD operation with the given problem statement.

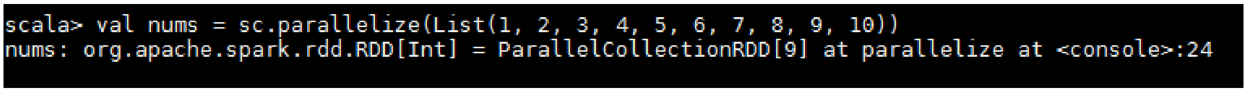
Problem Statement

1. Given a list of numbers - List[Int] (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
   - find the sum of all numbers
   - find the total elements in the list
   - calculate the average of the numbers in the list
   - find the sum of all the even numbers in the list
   - find the total number of elements in the list divisible by both 5 and 3.

Task1 - Find the sum of all numbers

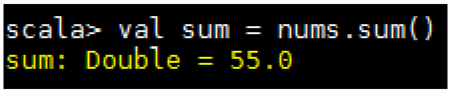
RDD,

val nums = sc.parallelize(List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10))



```
scala> val nums = sc.parallelize(List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10))
nums: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[9] at parallelize at <console>:24
```

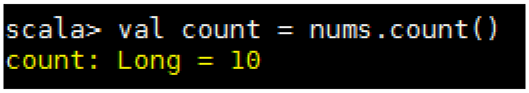
val sum=nums.sum()



```
scala> val sum = nums.sum()
sum: Double = 55.0
```

Task2 - find the total elements in the list

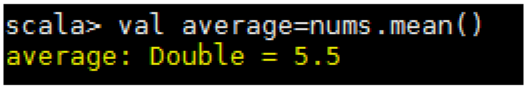
val count = nums.count()



```
scala> val count = nums.count()
count: Long = 10
```

Task3 - calculate the average of the numbers in the list

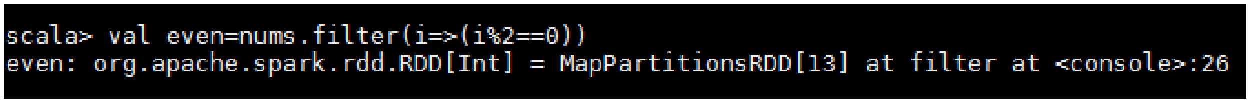
val average=nums.mean()



```
scala> val average=nums.mean()
average: Double = 5.5
```

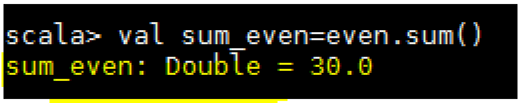
Task4 - find the sum of all the even numbers in the list

val even=nums.filter(i=>(i%2==0))



```
scala> val even=nums.filter(i=>(i%2==0))
even: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[13] at filter at <console>:26
```

val sum_even=even.sum()

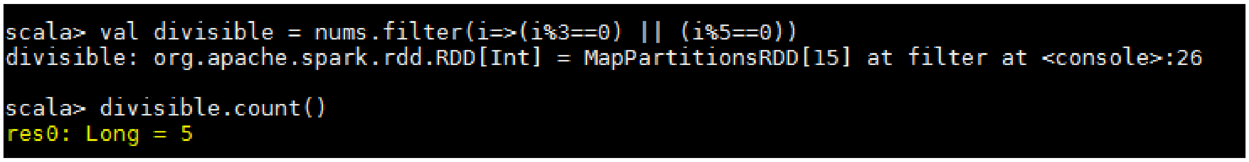


```
scala> val sum_even=even.sum()
sum_even: Double = 30.0
```

Task5 - find the total number of elements in the list divisible by both 5 and 3.

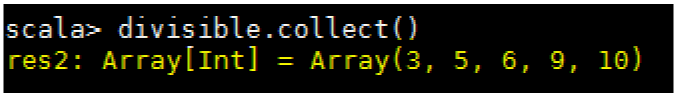
val divisible = nums.filter(i=>(i%3==0) || (i%5==0))

divisible.count()



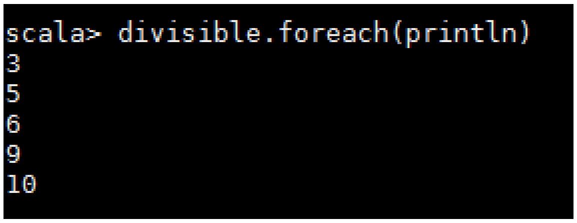
```
scala> val divisible = nums.filter(i=>(i%3==0) || (i%5==0))
divisible: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[15] at filter at <console>:26

scala> divisible.count()
res0: Long = 5
```

divisible.collect()



```
scala> divisible.collect()
res2: Array[Int] = Array(3, 5, 6, 9, 10)
```

divisible.foreach(println)



```
scala> divisible.foreach(println)
3
5
6
9
10
```