Session 18:

INTRODUCTION TO SPARK

Assignment 1

PROBLEM STATEMENT –

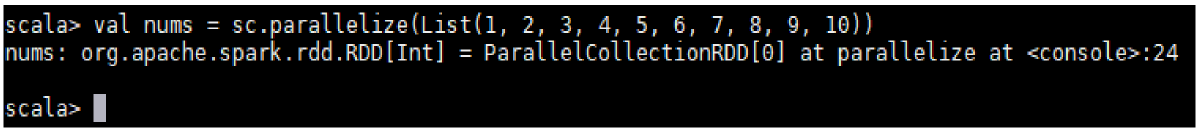
TASK 1 –

Given a list of numbers - List[Int] (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

1. find the sum of all numbers
2. find the total elements in the list
3. calculate the average of the numbers in the list
4. find the sum of all the even numbers in the list
5. find the total number of elements in the list divisible by both 5 and 3

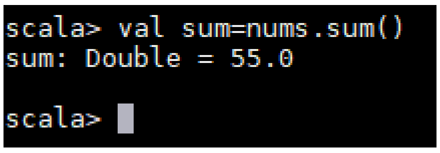
SOLUTION –

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



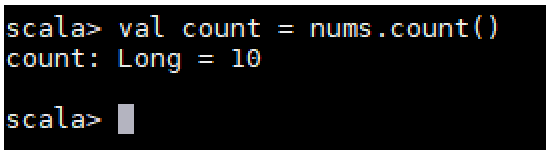
```
scala> val nums = sc.parallelize(List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10))
nums: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:24

scala>
```

1. val sum=nums.sum()



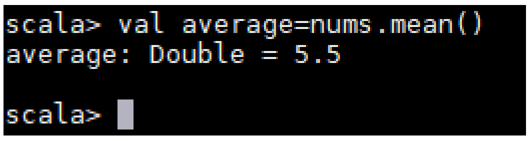
```
scala> val sum=nums.sum()
sum: Double = 55.0

scala>
```

2. val count = nums.count()



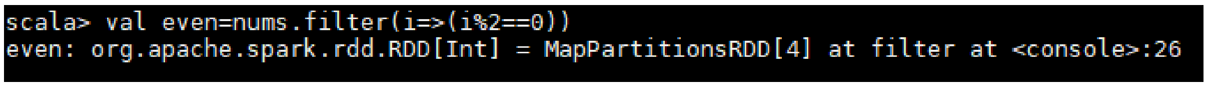
```
scala> val count = nums.count()
count: Long = 10

scala>
```

3. val average=nums.mean()



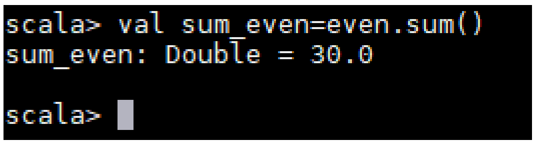
```
scala> val average=nums.mean()
average: Double = 5.5

scala>
```

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



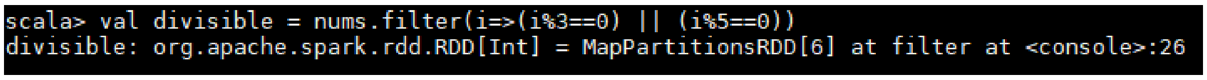
```
scala> val even=nums.filter(i=>(i%2==0))
even: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[4] at filter at <console>:26
```

val sum_even=even.sum()



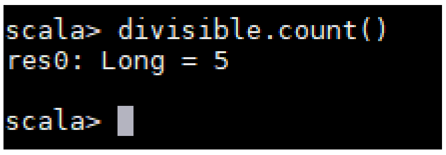
```
scala> val sum_even=even.sum()
sum_even: Double = 30.0

scala>
```

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



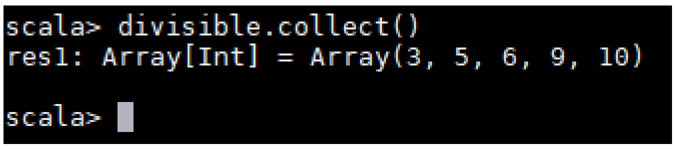
```
scala> val divisible = nums.filter(i=>(i%3==0) || (i%5==0))
divisible: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[6] at filter at <console>:26
```

divisible.count()



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

scala>
```

divisible.collect()



```
scala> divisible.collect()
res1: Array[Int] = Array(3, 5, 6, 9, 10)

scala>
```

TASK 2 –

1. Pen down the limitations of MapReduce.
2. What is RDD? Explain few features of RDD?
3. List down few Spark RDD operations and explain each of them.

SOLUTION –

1. MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.
   • It's based on disk computing
   • Suitable for single pass computations - non iterative computations.
   • Needs a sequence of MR jobs to run iterative tasks
   • Hadoop Map Reduce supports batch processing only, it does not process streamed data, and hence overall performance is slower. MapReduce framework of Hadoop does not leverage the memory of the Hadoop cluster to the maximum.
   • Slow Processing Speed.
   • No Real-time Data Processing.
   • Lengthy Line of Code.

2. RDD stands for Resilient Distributed Datasets are Apache Spark's data abstraction, RDD is a logical reference of a dataset which is partitioned across many server machines in the cluster. RDDs are Immutable and are self-recovered in case of failure. Dataset could be the data loaded externally by the user. RDDs can only be created by reading data from a stable storage such as HDFS or by transformations on existing RDDs.

   Features of RDD,

   In-memory computation
   The data inside RDD are stored in memory for as long as you want to store. Keeping the data in-memory improves the performance by an order of magnitudes.
   Lazy Evaluation
   The data inside RDDs are not evaluated on the go. The changes or the computation is performed only after an action is triggered. Thus, it limits how much work it has to do.
   Fault Tolerance
   Upon the failure of worker node, using lineage of operations we can re-compute the lost partition of RDD from the original one. Thus, we can easily recover the lost data.
   Partitioning
   RDD partition the records logically and distributes the data across various nodes in the cluster. The logical divisions are only for processing and internally it has no division. Thus, it provides parallelism.

3. Apache Spark RDD supports two types of operations -
   - Transformation
   - Action

   RDD Transformation -

   Spark Transformation is a function that produces new RDD from the existing RDDs. It takes RDD as input and produces one or more RDD as output. Each time it creates new RDD when we apply any transformation. Thus, the so input RDDs, cannot be changed since RDD are immutable in nature.

   Applying transformation built an RDD lineage, with the entire parent RDDs of the final RDD(s). RDD lineage, also known as RDD operator graph or RDD dependency graph. It is a logical execution plan i.e., it is **Directed Acyclic Graph (DAG)** of the entire parent RDDs of RDD.

   Transformations are lazy in nature i.e., they get execute when we call an action. They are not executed immediately. Two most basic type of transformations is a **map(), filter()**.

   After the transformation, the resultant RDD is always different from its parent RDD. It can be smaller (e.g. filter, count, distinct, sample), bigger (e.g. **flatMap(), union(), Cartesian()**) or the same size (e.g. map).

   RDD Action –

   Transformations create RDDs from each other, but when we want to work with the actual dataset, at that point action is performed. When the action is triggered after the result, new RDD is not formed like transformation. Thus, Actions are Spark RDD operations that give non-RDD values. The values of action are stored to drivers or to the external storage system. It brings laziness of RDD into motion.

   An action is one of the ways of sending data from Executer to the driver. Executors are agents that are responsible for executing a task. While the driver is a JVM process that coordinates workers and execution of the task.

   Some of the actions of Spark are:

   count(), collect(), take(n), top(),countByValue(),reduce(),fold(),aggregate() and foreach()