DSCI5350 Spark Activity 1-Databrciks Satyasriharsha Buddha

1)Click New cluster under Common Tasks

A screenshot of a computer

Description automatically generated with medium confidence

2)Enter the name of your cluster, select DataBricks run time version and click create cluster

A screenshot of a computer

Description automatically generated with medium confidence

3)Click Event log and verify the cluster status

A screenshot of a computer

Description automatically generated with medium confidence

4)In your workspace, create new folder named RDD

A screenshot of a computer

Description automatically generated with medium confidence

5) Create scala or python notebook in RDD folder.

A screenshot of a computer

Description automatically generated with medium confidence

6) Verify app Name, Spark session, Spark context, SQL context

A screenshot of a computer

Description automatically generated with medium confidence

7) Create an RDD using parallelize method available in spark context

Creating RDD of three elements from a Scala sequence with 2 partitions by using parallelize method

A screenshot of a computer

Description automatically generated with low confidence

8) Perform collect action in RDD and finding the number of partitions using getNumPartitions action  
Collect () -It returns all items in the RDD to the driver in a single list.  
Glom ()-Return an RDD created by coalescing all elements within each partition into a list

A screenshot of a computer

Description automatically generated with medium confidence

9) Perform “take” action in RDD

take(n)-It returns an array with first n elements of RDD

A screenshot of a computer

Description automatically generated with low confidence

10)Transform RDD by map to make another RDD

Map transformation returns a new RDD that is formed by passing each element of the source RDD through a function

A screenshot of a computer

Description automatically generated with medium confidence

11)Transform RDD by filter to make another RDD

Filter transformation returns a new RDD that is formed by selecting elements for which the function returns true.

A screenshot of a computer

Description automatically generated with medium confidence

12) Perform reduce action in RDD  
Reduce aggregates all elements of RDD by applying user function pairwise to elements and partial results and returns result to driver

A screenshot of a computer

Description automatically generated with low confidence

13) Transform RDD by flat map to make another RDD

Return a new RDD by first applying function to all elements of this RDD and then flattening the result.

A close-up of a computer screen

Description automatically generated with low confidence

14) Create a pair RDD

A screenshot of a computer

Description automatically generated with low confidence

15)Perform transformations in pair RDD

a. Reduce by key

It takes an RDD and returns new RDD of key value pairs. The values of each key are aggregated using the reduced function

A screenshot of a computer program

Description automatically generated with low confidence

b. Sort by key

This returns a new RDD of key-value pairs that’s sorted by keys in ascending order.

A screenshot of a computer

Description automatically generated with medium confidence

c. Group by key

This returns a new RDD consisting of key and iterable-valued pairs.

A screenshot of a computer

Description automatically generated with medium confidence

16) Simple computation

A screenshot of a chat

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence