Programming for Big Data - Spark Lab 2

In this lab you will use Spark's functionality to programme RDDs.

1. Introduction

You can use your data bricks cluster or a locally installed SPARK version.

Cluster:
https://community.cloud.databricks.com/

Local:
Open the command line and navigate to your Spark installation. Then type ./bin/spark-shell.

Using the notes from today's class please complete the following tasks. Please copy the line of code you use to complete the task coupled with the result.

Task 1: Create RDD with the following elements:

2,4,6,8,10,12,14,15,13,11,9,7,5,3,1,22

Task 1A: Add 1 to each element:

Code:

Result:

Task 1B: Subtract 2 from each element

Code:

Result:

Task 1C: Sum all the elements

Code:

Result:

Task 1D: Remove all elements less than 10 and sum the result

Code:

Result:

Task 2: Create the following RDDs:

RDD1 - Spark, Apache, Stream, Batch, Programming, Spark, Apache
RDD2 - Bread, Batch, Sodastream, Bread, Bread, Butter

Task 2A: What are the distinct values in each RDD?

Code:

Result:

Task 2B: What are the common values in each RDD?

Code:

Result:

Task 2C: What values are in the first RDD that are absent from the second RDD?

Code:

Result:

Task 3: Create the following RDD:

RDD - [john, 20], [joe, 22], [tom, 77], [bob, 90], [bill, 99], [Ador, 98], [john, 22], [joe, 1], [tom,2]

Task 3A: Add 10 to each value in the data and print the new values

Code:

Result:

Task 3B: Print all the keys

Code:

Result:

Task 3C: Print all the values

Code:

Result:

Task 3D: Add all the values together and print the result

Code:

Result:

Task 3E: Sort the data by key and then print

Code:

Result:

Task 4: Create the following RDDs:

RDD1 - [joe, 22], [tom, 77], [bob, 910], [bill, 99], [Ador, 981], [john, 22], [joe, 1], [tom,2]
RDD2 - [john, 20], [joe, 122], [bob, 77], [bob, 920], [bill, 99], [john, 212], [joe, 1], [Ador,1]

Task 4A: Remove duplicate keys and print

Code:

Result:

Task 4B: Join both rdds and add values for all keys that are in both

Code:

Result:

4. Summary

In this lab we have looked at array and sequence operations for programming spark RDDs.