Lab 04 Programming with Spark RDD

IT413 No SQL Databases, Winter’2021, DAIICT, Gandhinagar; pm_jat

In this lab, we continue on Spark programming. You can work either in Java, Python, or Javascript.

For Java we continue using Eclipse and Maven. Following are dependencies that you would be requiring for Spark.

```
    <dependencies>
         <dependency>
              <groupId>org.apache.spark</groupId>
              <artifactId>spark-core_2.12</artifactId>
              <version>2.4.7</version>
         </dependency>
    </dependencies>
```

Submission Required: Source Code of solutions of all exercises in a single compressed file.

Exercises:

General instruction: Have separate source file for each question. Given question number as source file name. Keep complete solution of one question in single source file.

Perform computation of Lab02 (Question 3, and Question 4 on Web Access Log File) using spark RDD API (Do not use dataframe API)

1. Monthly Summary:

   (a) Total number of requests.

   (b) Total downloaded size

   It should output: <Year-Month, Number of Requests, Download Size> for every months like Dec-2016, Jan-2017, and so!

2. Generate Report that lists Timestamp, URL for which http response status has been 404.

You are given two data file as following

Order (OrderNo, CustomerID, OrderAmount)
Customer(CustomerID, Country, State)

3. Produce a Summary Report that list State, Total Revenue generated in that state. Perform following computation using RDD API (Do not use dataframe API)