Lab 03 — Spark Programming

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

In this lab, we practice Spark Programming.

Here also we work in Eclipse, and you create a new Java Maven project in the same way we did in MR Lab. However, you use dependencies as following. Let the project for this lab be named as `Lab03`. Copy and paste following dependencies in your `pom.xml` file.

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

Here you should use JDK 1.8, and later. This you can set in “Java Build Path” under project properties.

Attempt running following Spark programs in Java. Source and data files are given to you. Place all source files in package “`nosql.lab03`”. You shall require changing the package name accordingly. Also, place all data files in “`data`” folder under your project folder.

1. [`WordCount.java`] is given in file WordCount.java. May use given input text file “`article1.txt`” as input.
2. [`SalarySums.java`]. Use “`employee.csv`” file as input.
3. [`SalaryAverages.java`]. Use “`employee.csv`” file as input.
4. [`ImageCounter.java`]. Use same web access log of Lab01/02.

Submission Required:

For each exercise list down (1) RDDs used (2) Transformations and Actions performed on RDDs in their occurring sequence. Put all answers in a single document file and submit as PDF.