Final Project: Data Analysis using Spark

Estimated time needed: 15 minutes

Introduction

In the final project, you will perform the mentioned tasks on your own. The tasks to be performed are similar to what you did in the practice lab, however, you will not be provided with step-by-step instructions.

This project focuses on mastering Spark SQL, a powerful component of Apache Spark that allows you to work with structured data using SQL-like queries. You will create a DataFrame from a CSV file, define a schema for the data, and leverage Spark SQL to perform transformations and actions on the data.

Scenario

You have been tasked by the HR department of a company to create a data pipeline that can take in employee data in a CSV format. Your responsibilities include analyzing the data, applying any required transformations, and facilitating the extraction of valuable insights from the processed data.

Given your role as a data engineer, you've been requested to leverage Apache Spark components to accomplish the tasks.

Project Overview

Create a DataFrame by loading data from a CSV file and apply transformations and actions using Spark SQL. This needs to be achieved by performing the following tasks:

- Task 1: Generate DataFrame from CSV data.
- Task 2: Define a schema for the data.
- Task 3: Display schema of DataFrame.
- Task 4: Create a temporary view.
- Task 5: Execute an SQL query.
- Task 6: Calculate Average Salary by Department.
- Task 7: Filter and Display **IT** Department Employees.
- Task 8: Add 10% Bonus to Salaries.
- Task 9: Find Maximum Salary by Age.
- Task 10: Self-Join on Employee Data.
- Task 11: Calculate Average Employee Age.
- Task 12: Calculate Total Salary by Department.
- Task 13: Sort Data by Age and Salary.
- Task 14: Count Employees in Each Department.
- Task 15: Filter Employees with the letter **o** in the Name.

You will be provided with a Jupyter notebook environment to complete this project. Please ensure that you follow the provided instructions and use the Python and Spark (PySpark) libraries in your lab environment. The tasks in this project will enable you to effectively work with DataFrames, define schemas, use Spark SQL for querying, and perform data transformations and actions. Upon completion, you will have a solid understanding of Spark SQL and its application in real-world scenarios.

Author(s)

Ragul Ramesh

Other Contributor(s)

Lavanya T S