Big Data Hadoop and Spark Developer

Retail Business Analytics

Powered by simplilearn

UCI

Objectives

- The objective is to analyze the “**retail_db”** dataset, provide reports on the total completed orders, and perform customer and product analytics.

Prerequisites

- Python
- PySpark
- Hadoop
- Spark

Industry Relevance

- **Python:** It is used for web development, data science and data analysis, machine learning, startups, and the finance industry.
- **PySpark**: It writes Spark apps in Python.
- **Hadoop**: It is used in public sectors, such as intelligence, science, cyber security, and defense.
- **Spark:** It is used for machine learning and streaming data.

Problem Statement

Customers can purchase products or services from Amazon for consumption and usage. Amazon usually sells products and services in-store. However, some may be sold online or over the phone and shipped to the customer. Clothing, medicine, supermarkets, and convenience stores are examples of their retail operations.

Tasks to Perform

- Perform the following tasks on the dataset provided using PySpark:
- 1. Explore the customer records saved in the "customers-tab-delimited“ directory on HDFS
  - Show the client information for those who live in California
  - Save the results in the result/scenario1/solution folder
  - Include the customer's entire name in the output
- 2. Explore the order records saved in the “orders parquet” directory on HDFS
  - Show all orders with the order status value "COMPLETE“
  - Save the data in the "result/scenario2/solution" directory on HDFS
  - Include order number, order date, and current situation in the output

Tasks to Perform

- Perform the following tasks on the dataset provided using PySpark:
- 3. Explore the customer records saved in the "customers-tab-delimited“ directory on HDFS
  - Produce a list of all consumers who live in the city of "Caguas"
  - Save the results in the result/scenario3/solution folder
  - The result should only contain records with the value "Caguas" for the customer city
- 4. Explore the order records saved in the “categories” directory on HDFS
  - Save the result files in CSV format
  - Save the data in the result/scenario4/solution directory on HDFS
  - Use lz4 compression to compress the output

Tasks to Perform

- Perform the following tasks on the dataset provided using PySpark:
- 5. Explore the customer records saved in the “products_avro" directory on HDFS
  - Include the products with a price of more than 1000.0 in the output
  - Remove data from the table if the product price is greater than 1000.0
  - Save the results in the result/scenario5/solution folder
- 6. Explore the order records saved in the “products_avro” directory on HDFS
  - Only products with a price of more than 1000.0 should be in the output
  - The pattern "Treadmill" appears in the product name
  - Save the data in the result/scenario6/solution directory on HDFS

Tasks to Perform

- Perform the following tasks on the dataset provided using PySpark:
- 7. Explore the customer records saved in the “orders parquet" directory on HDFS
  - Output all PENDING orders in July 2013
  - Only entries with the order status value of "PENDING" should be included in the result
  - Order date should be in the YYY-MM-DD format
  - Save the results in the result/scenario7/solution folder

Project Outcome

- This project is designed to help understand the retail database and generate reports on the completed orders.

- You should be able to analyze the dataset for this project to create a report. You will be able to use PySpark, do analyses, and obtain the desired results.

Submission Process

1. Complete the project in the Simplilearn lab
2. Complete each task listed in the problem statement
3. Take screenshots of the results for each question and the corresponding code
4. It should be saved as a document and submitted using the assessment tab.
5. Tap the "Submit" button (this will present you with three choices)
6. Attach three files and then click "Submit“

- **Note:** Be sure to include screenshots of the output

Thank You