
Assignment: Online Retail Data Analysis Using PySpark

Datasets

1. **Customers.csv** – `customer_id, name, country, age, gender`
2. **Orders.csv** – `order_id, customer_id, product, category, quantity, price, order_date`

Objective: Perform ETL, analytics, and reporting using PySpark and PySpark SQL.

Exercise Breakdown

1. Setup & Data Ingestion

- Initialize SparkSession in Jupyter Notebook.
- Load the two CSVs into DataFrames.
- Print schemas and first 10 records.
- Persist them as temporary views for SQL queries.

2. Data Cleaning

- Remove duplicates and null values.
- Cast `age` to integer, `price` to double, and `order_date` to `date`.
- Ensure proper column trimming (remove leading/trailing spaces).

3. Exploratory Data Analysis (EDA)

- **Find number of customers by country.**
- **Find age distribution of customers.**
- **Top 5 countries by total customers.**

- **Top 5 categories by order count.**

4. Business Insights Using DataFrame API

- **Total revenue per category.**
- **Average order value per country.**
- **Top 10 customers by revenue.**
- **Number of distinct products sold.**
- **Most frequently purchased product.**

5. Business Insights Using PySpark SQL

- Register both DataFrames as temp views.
- Write SQL queries to:
    - Find total quantity and revenue by category.
    - Find the highest-spending customer.
    - Find monthly revenue trends.
    - Find customers who purchased more than 5 distinct products.

6. Joins & Advanced Operations

- Perform an **inner join** between Customers and Orders to get enriched order data.
- Perform **left join** to see customers with zero orders.
- Group by `country` and find:
    - Total revenue
    - Average revenue per customer
    - Highest revenue customer in each country

7. Window Functions

- Use ranking to find **top 3 customers by revenue in each country.**
- Use lead/lag to compare each month's revenue with the previous month.

8. Writing Data

- Save the cleaned Orders DataFrame as Parquet.
- Save the country-wise revenue summary as CSV.

9. Bonus Tasks (Optional)

- Use **aggregate functions** to find median order value.
- Use **broadcast join** for faster country-wise analysis if the Customers table is small.
- Use **accumulators** to count total records processed.

Deliverables for Participants

1. **Final saved files** in Parquet/CSV format.
