Retail Sales Spark Project - Explanation Report

1. Introduction

The Retail Sales Spark Project is designed to analyze and gain insights from large-scale retail sales data using Apache Spark. It focuses on understanding customer purchasing behavior, sales trends, and product performance. Apache Spark’s distributed processing capability allows handling millions of records efficiently.

2. Objectives

- To clean and preprocess retail sales data using PySpark.
- To perform exploratory data analysis (EDA) for understanding trends.
- To identify top-selling products and revenue patterns.
- To visualize the results for business decision-making.
- To explore predictive modeling for forecasting sales using Spark MLlib.

3. Dataset Description

The dataset includes retail transactional records with attributes such as:

- Invoice Number
- Product ID
- Product Description
- Quantity Sold
- Price
- Transaction Date
- Customer ID
- Country

4. Data Processing with PySpark

The dataset was ingested into a Spark DataFrame. Data preprocessing included handling null values, removing duplicates, and ensuring proper formatting. PySpark functions were used to apply transformations like filtering, grouping, and aggregation to prepare the dataset for analysis.

5. Data Analysis & Insights

Key findings from the analysis include:

- Identification of the most popular products based on sales quantity.
- Revenue trends across different time periods and regions.
- Customer segmentation based on purchase behavior.
- Seasonal sales variations.

6. Visualization

Visualizations were generated using Spark with Python libraries such as Matplotlib and Seaborn. These included bar charts, line graphs, and heatmaps to illustrate revenue trends, sales peaks, and demand distribution.

7. Conclusion & Future Work

The Retail Sales Spark Project highlights the effectiveness of Apache Spark in analyzing large-scale retail datasets. It provides meaningful insights for businesses to make informed decisions. Future improvements may include integrating Spark MLlib for predictive analytics, enabling more accurate sales forecasting and product recommendations.