Databricks Sales Analysis Project

Overview

This project demonstrates a complete Sales Analysis pipeline using Databricks, PySpark, and Delta Lake. It processes sales and product data, applies transformations, and performs analytical aggregations with visualization.

Technologies Used

- Databricks (Community Edition)
- Apache Spark & PySpark
- Delta Lake for optimized storage
- Matplotlib for visualization
- GitHub for version control

Data Workflow

1. **Data Upload**: Products and Sales data (CSV files) are uploaded to DBFS.

2. **Data Processing**:
    - Load CSV data into Spark DataFrames.
    - Perform transformations (Joins, Aggregations, etc.)
    - Handle missing values.

3. **Data Storage**:
    - Convert DataFrames to **Delta Tables** for better performance.

4. **Data Analysis**:
    - Total Sales Per Month
    - Total Sales Per Brand
    - Total Sales Per Product
    - Total Order Value Per Customer

5. **Visualization**:
    - Line plot for **Monthly Sales Trend**.
    - Bar plot for **Sales by Product**.

How to Run This Project

Run in Databricks

1. **Import Sales_Analysis.dbc into Databricks**:
    - Go to Databricks → Workspace.
    - Click on "Import" and select Sales_Analysis.dbc.

2. **Attach to a Cluster** and run the notebook.

3. **Upload the CSV files** to DBFS.

Files in This Repository

FILE NAME	DESCRIPTION
Sales_Analysis.dbc	Databricks Notebook export (can be imported in Databricks)
Sales_Analysis.py	Python version of the notebook (for local execution)
README.md	Documentation
Data/Products.csv	Sample Products dataset
Data/Sales.csv	Sample Sales dataset

Sample Queries and Transformations

- Aggregating **Total Sales by Month**
  ```
  df_monthly_sales = df_final.groupBy("Order Month").agg(sum("Total Sales Amount").alias("Total Sales by Month"))
  ```

- **Joining Sales Data with Product Information**
  ```
  df_final = df_sales.join(df_products.select("ProductID", "Product Name", "Category"), on="ProductID", how="left")
  ```

- **Storing Data in Delta Format**
  ```
  df_final.write.format("delta").mode("overwrite").save("/mnt/delta/final_sales_data")
  ```

Visualization Examples

Total Sales Per Month:

1. plt.figure(figsize=(10,5))
2. plt.plot(pdf["Order Month"], pdf["Total Sales by Month"], marker="o", linestyle="-", color="royalblue")
3. plt.xlabel("Order Month")
4. plt.ylabel("Total Sales")
5. plt.title("Total Sales Per Month")
6. plt.xticks(rotation=45)

plt.show()

Author

- **Manish Dhanabalakrishnan**