USN:1RVU23CSE153
Name:Ekshu DP

Ex No: 4 **Date: 11-09-2025**	**Building and Automating Pipeline in Databricks for an E-Commerce Dataset**

Objective:

To design and implement a data pipeline on Databricks using a healthcare dataset, following the Medallion architecture (Bronze → Silver → Gold). The pipeline should ingest raw data, clean and transform it, perform aggregations, and finally automate execution with dashboards and notifications.

Outcomes:

1. A fully functional end-to-end pipeline in Databricks.
2. Clean and aggregated data stored in Bronze, Silver, and Gold tables.
3. Automated dashboards for revenue insights.
4. Scheduled execution with email notifications for monitoring.

Materials:

Software & Platform

- Databricks Community Edition / Enterprise Edition
- Apache Spark (PySpark API)
- Databricks SQL and Notebooks
- Databricks Jobs (for scheduling & automation)

Dataset:-

- `healthcare_orders.csv`

USN:1RVU23CSE153
Name:Ekshu DP

Lab Procedure:

Step 1: Ingestion & Cleaning (Bronze → Silver)

1. Upload the healthcare_orders.csv dataset into Databricks File System (DBFS).
2. Create a Bronze table (healthcare_orders) to store the raw data.
3. Use PySpark to clean the data:
    - Convert order_date to a standard format (yyyy-MM-dd).
    - Compute total_value = quantity × price.
4. Save the cleaned dataset as a Silver table (silver_healthcare_orders).

Step 2: Aggregation & Enrichment (Silver → Gold)

1. Load the Silver table into a DataFrame.
2. Perform aggregations:
    - Calculate total revenue by service category → store as gold_healthcare_service_sales.
    - Calculate daily revenue trends → store as gold_healthcare_daily_sales.
3. Save both results as Gold tables for reporting and analytics.

Step 3: Dashboard & Visualization

1. Use Databricks Notebooks to create interactive dashboards.
2. Add the following visualizations:

    - Bar Chart → Revenue by service category.
    - Line Chart → Daily revenue trends.
    - Filters → City and payment method filters for drill-down analysis.

Step 4: Automation & Scheduling

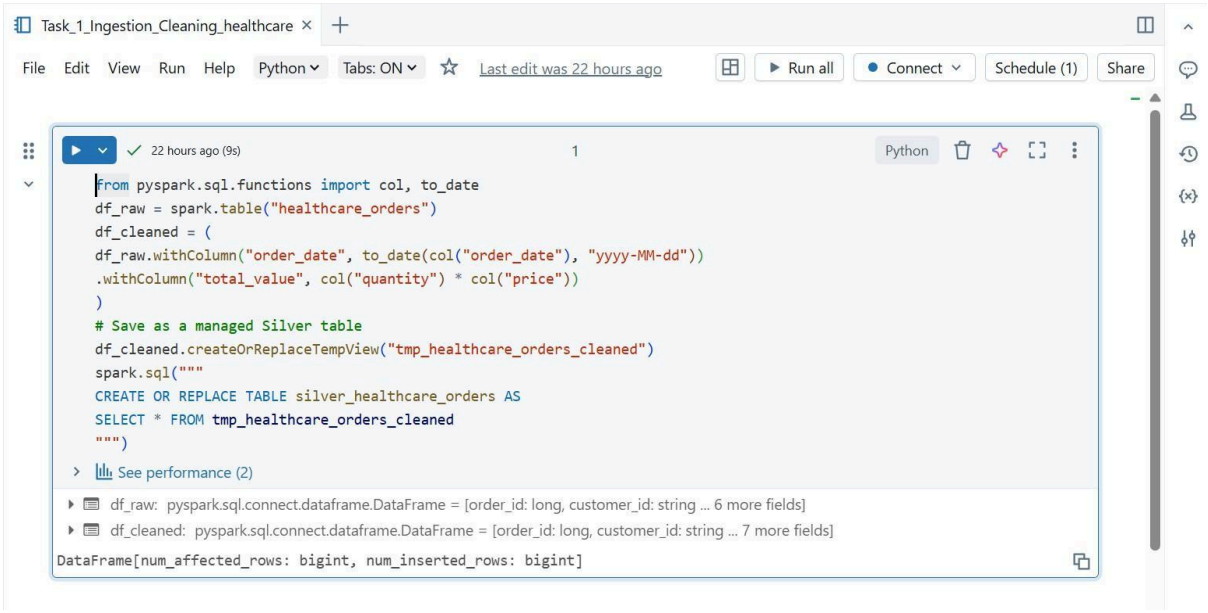
1. Convert the pipeline into a Databricks Job.
2. Configure a schedule (daily/weekly) for automatic pipeline execution.
3. Ensure that Bronze → Silver → Gold tables are refreshed automatically.
4. Automate dashboard updates after each run.

Step 5: Email Notifications & Alerts

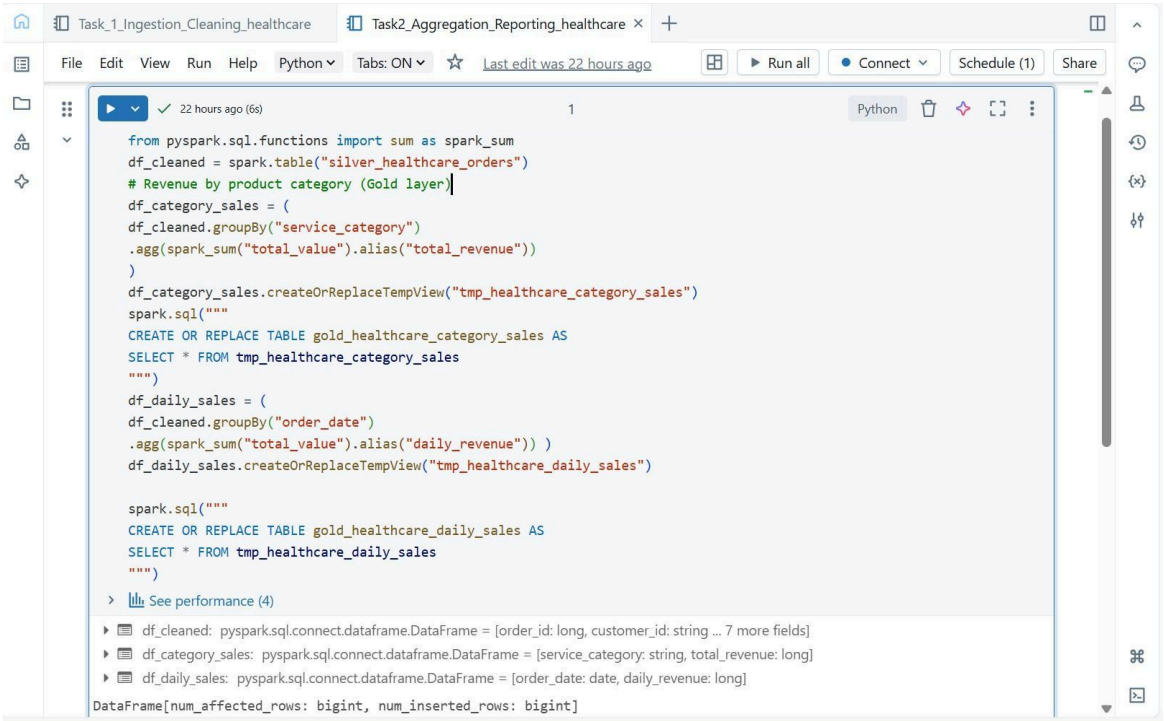
1. Configure email alerts in the Databricks Job settings.
2. Enable:
    - Success notifications → sent to administrators.
    - Failure alerts → sent to data engineers for debugging.

USN:1RVU23CSE153
Name:Ekshu DP

Task_1_Ingestion_Cleaning_healthcare:-

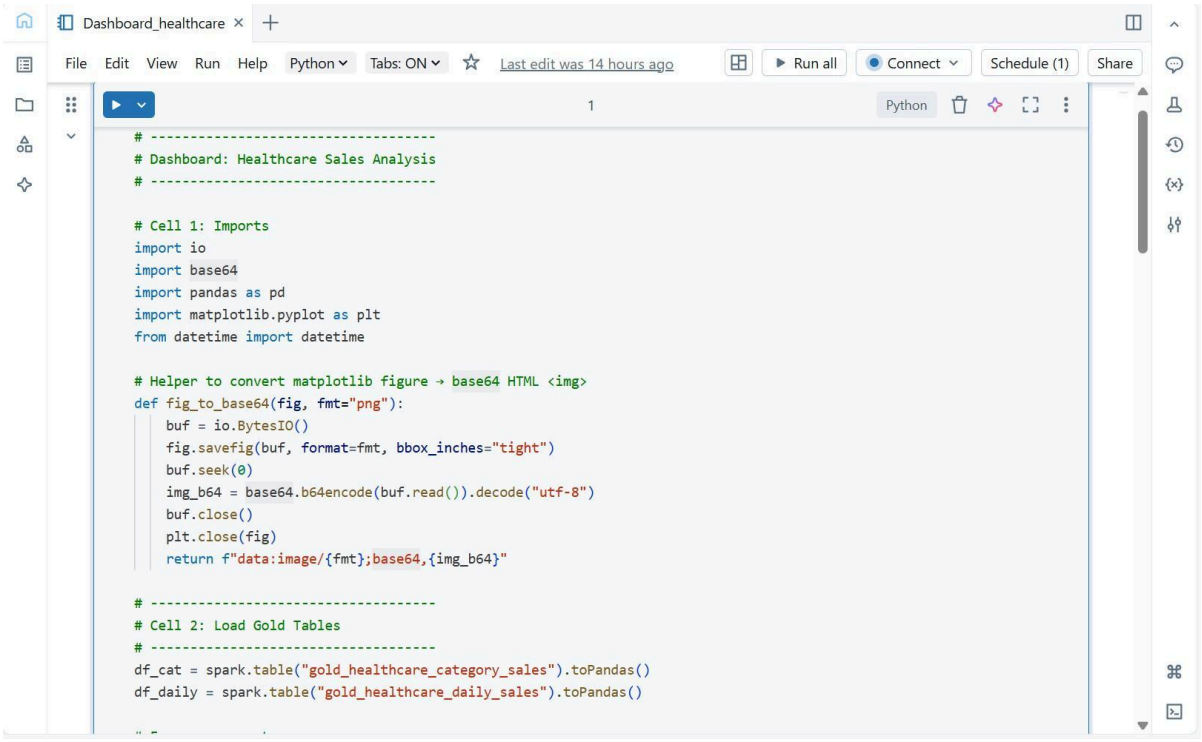


Task2_Aggregation_Reporting_healthcare:-

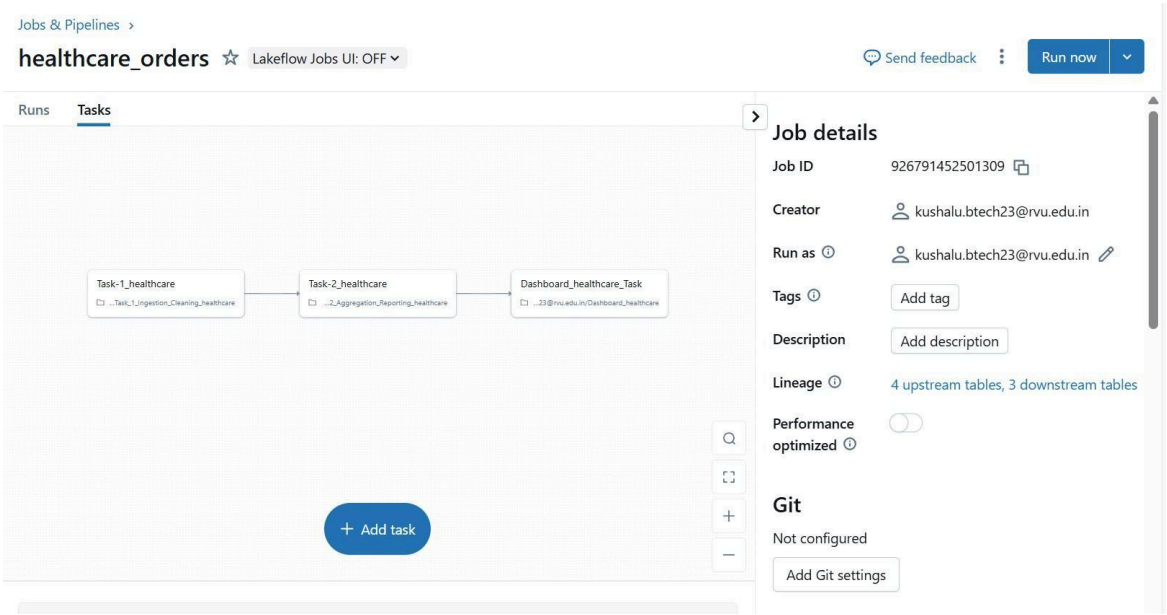


USN:1RVU23CSE153
Name:Ekshu DP

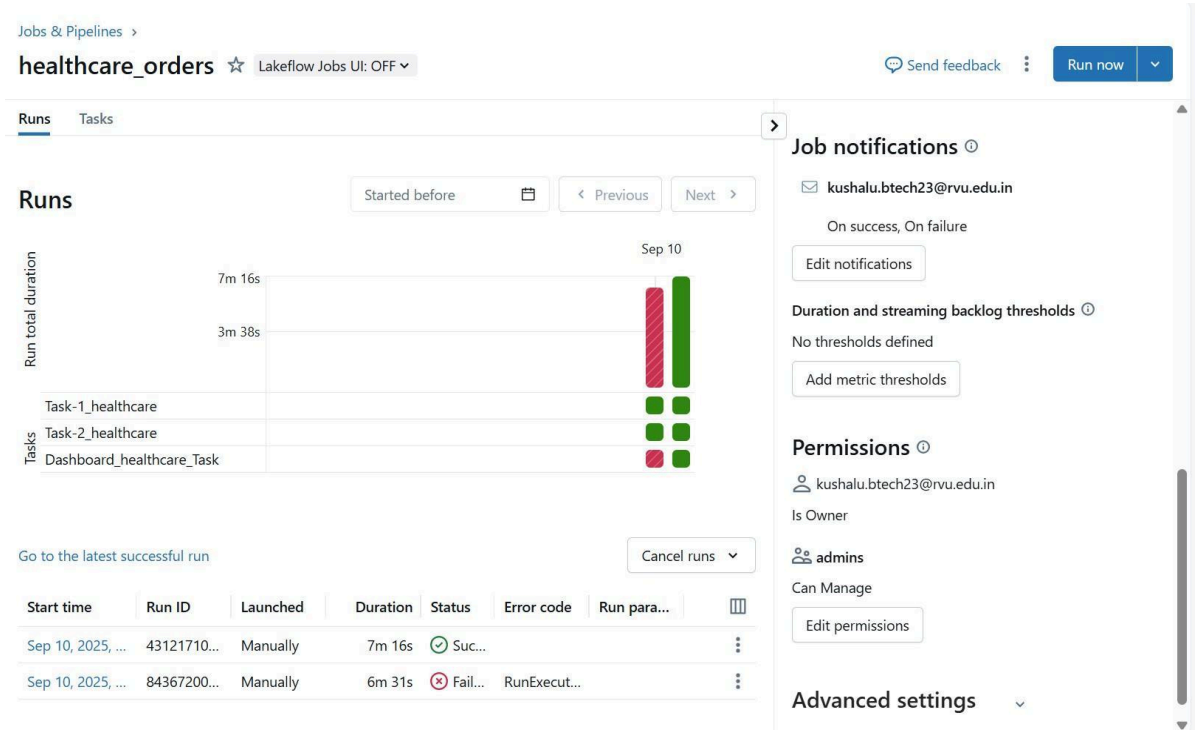
Dashboard_healthcare:-

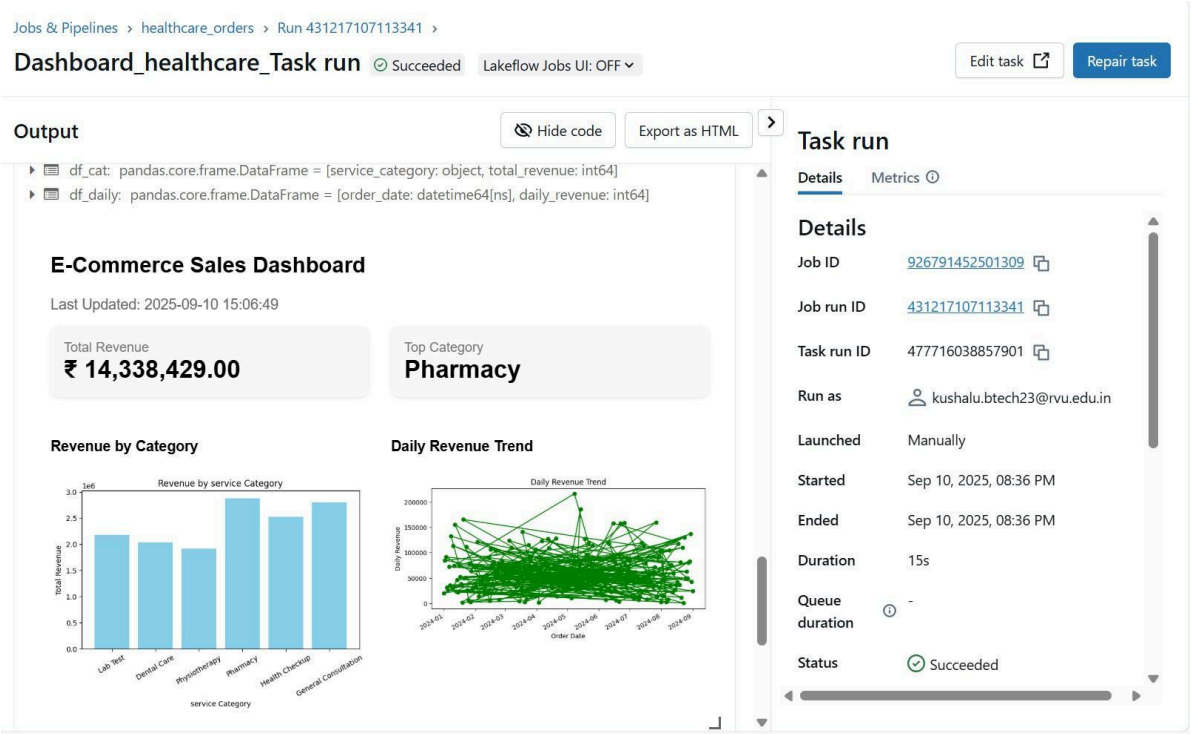


Job Pipeline:-



USN:1RVU23CSE153
Name:Ekshu DP





USN:1RVU23CSE153
Name:Ekshu DP

Email Notification:-