
Capstone Project: Retail Data Analytics Platform using Azure Databricks

Objective

Design and build an **end-to-end data engineering pipeline** for a fictional retail company — **Retail360** — using **Azure Databricks**. The goal is to simulate real-world scenarios: ingest raw data, clean and transform it, store it in Delta Lake, perform aggregations, enable incremental loads, and build a Gold-layer analytical view.

Project Overview

Layer	Description	What Participants Will Learn
Raw	Ingest CSV, JSON, and streaming-like data	Data ingestion, schema inference
Bronze	Store raw data “as-is” into Delta tables	Basic Delta Lake
Silver	Clean, transform, and enrich data	Transformations, Joins, Aggregations
Gold	Business analytics (e.g., revenue dashboard)	Aggregations, window functions
DLT	Automate Bronze → Silver → Gold	Delta Live Tables
Advanced	Time Travel, Merge, Incremental loads	Real-world operations

---

1. Dataset Creation

Simulate three raw data sources (you can generate inline or upload to `/FileStore/tables/`):

a) `customers.csv`

```
customer_id,name,region,email
1,Arjun Rao,North,arjun@example.com
2,Sneha Patel,South,sneha@example.com
3,Rahul Sharma,East,rahul@example.com
4,Neha Iyer,West,neha@example.com
```

b) `orders_day1.csv`

```
order_id,customer_id,product,quantity,price,status,order_date
1001,1,Laptop,2,55000,Completed,2024-01-15
1002,2,Mobile,3,25000,Completed,2024-01-16
1003,3,Book,10,700,Pending,2024-01-16
1004,1,Headphones,5,3000,Completed,2024-01-17
```

c) `products.json`

```
[
  {"product_id": "P001", "product_name": "Laptop", "category": "Electronics"},
  {"product_id": "P002", "product_name": "Mobile", "category": "Electronics"},
  {"product_id": "P003", "product_name": "Book", "category": "Stationery"},
  {"product_id": "P004", "product_name": "Headphones", "category": "Accessories"}
]
```

---

2. Bronze Layer – Data Ingestion

Tasks:

- Read CSV, JSON, and other data into DataFrames.
- Perform schema inference.
- Write raw data into **Delta tables** (`bronze_customers` , `bronze_orders` , `bronze_products`).

---

3. Silver Layer – Data Cleansing & Transformation

Tasks:

- Remove invalid records (e.g., null emails or Pending orders).
- Add a new column `total_amount = quantity * price` .
- Join orders with customer and product info.
- Store results as `silver_orders` .

---

4. Gold Layer – Business Aggregations

Tasks:

- Calculate total revenue by region.
- Find top-selling products.
- Use **window functions** to rank products by sales.
- Store final results as `gold_sales_summary` .

---

5. Incremental Load Simulation

- Create a new file `orders_day2.csv` with new orders.
- Use **MERGE** or **Upsert** to update the `silver_orders` table with new data.
- Demonstrate how Delta Lake handles late-arriving data.

---

6. Time Travel & Vacuum

- Query a historical version of the `gold_sales_summary` table.
- Use `VACUUM` to clean up old versions.
- Demonstrate rollback with `VERSION AS OF` .

---

7. Delta Live Tables (Optional but Recommended)

- Automate the Bronze → Silver → Gold transformations using **DLT**.
- Show lineage and schema evolution automatically.

□ 8. Advanced Features (Optional for Pro Learners)

- **Z-Ordering** for query performance.
- **OPTIMIZE** commands.
- **Incremental Load Pattern** with `cloud_files()` ingestion.

□ Expected Outcomes

By the end, learners should be able to:

□ Ingest and process raw data into Delta tables □ Perform transformations and joins with PySpark □ Build multi-layer architecture (Bronze → Silver → Gold) □ Implement MERGE for incremental loads □ Explore Delta Lake Time Travel and Vacuum □ Automate pipelines using Delta Live Tables

---