Assessment: Transforming the Breast Cancer Dataset in Databricks

Duration:

4 Hours

Objective:

Assess the understanding and practical implementation of Delta Lake, Dynamic Views, Aggregations, Medallion Architecture, and Data Quality Checks in Databricks to derive Silver and Gold tables from the Bronze layer.

Instructions:

- Use Delta Lake to store transformed data at different layers (Bronze, Silver, Gold).

- Implement Dynamic Views for data governance and row-level security.

- Apply aggregations and transformations to generate insights.

- Ensure data quality checks at each transformation stage.

- Delta Live Tables (DLT) is optional—you may use alternative methods for transformation.

- Submit the Databricks notebook with markdown explanations and code.

Task 1: Understanding the Bronze Layer (30 Marks)

1. Load the breast cancer dataset from data.csv into Databricks and store it in a Delta Lake Bronze Table (bronze.breast_cancer).

2. Inspect the schema, data types, and missing values.

3. Apply basic data profiling to identify inconsistencies (e.g., incorrect data types, missing values, duplicate records).

Deliverable:

- A Delta Table (bronze.breast_cancer) storing raw dataset.

- Markdown cell summarizing schema, missing data, and initial observations.

Task 2: Transforming to Silver Layer (60 Marks)

1. Clean and standardize data (handle missing values, correct data types, remove duplicates).

2. Create a processing job to transform Bronze data into a Silver Table (silver.breast_cancer) with:

  - Categorical variable conversion (e.g., "M" → "Malignant", "B" → "Benign").

- Normalization of numerical fields.

- Removal of redundant columns.

3. Implement schema evolution and enforce data quality constraints.

4. Create Dynamic Views on the Silver Table to restrict access to sensitive data.

5. Track data changes using Delta Lake's versioning (DESCRIBE HISTORY).

Deliverable:

- A Silver Table (silver.breast_cancer) with structured, clean data.

- A processing pipeline for transformation.

- A Dynamic View ensuring restricted access.

Task 3: Creating the Gold Layer with Aggregations & Insights (70 Marks)

1. Create a Gold Table (gold.breast_cancer_insights) with:

- Distribution of tumor classifications.

- Average tumor size per classification.

- Mean, median, and standard deviation for numerical features.

2. Implement row-level security using Dynamic Views.

3. Optimize Gold Table using OPTIMIZE and ZORDER.

4. Use Delta Lake's MERGE operation for incremental updates.

Deliverable:

- A Gold Table (gold.breast_cancer_insights) with aggregated insights.

- Queries for aggregations and performance optimizations.

Task 4: Implementing Data Quality Checks & Monitoring (40 Marks)

1. Define data quality expectations and constraints.

2. Implement error handling for ingestion failures.

3. Generate audit logs for data validation.

4. Create a Databricks SQL Dashboard for monitoring.

Deliverable:

- A data quality framework with validation reports.

- A Databricks SQL Dashboard displaying data quality metrics.

Grading Criteria:

- Bronze Layer Implementation (30 Marks): Correct ingestion, schema analysis.

- Silver Layer Transformation (60 Marks): Data cleaning, Dynamic Views, versioning.

- Gold Layer Aggregations (70 Marks): Insights, security, performance optimization.

- Data Quality Checks & Monitoring (40 Marks): Constraints, error handling, dashboard.