Data Warehouse and VACUUM

Q1: What is a Data Warehouse in Databricks?

A data warehouse in Databricks is a central repository that stores structured data from different sources. It supports business intelligence (BI) activities, especially analytics.

Key features of a data warehouse in Databricks:

- Built on a Lakehouse architecture combining data lake and warehouse capabilities.
- Utilizes Delta Lake for ACID transactions, schema enforcement, and time travel.
- Optimized for SQL-based workloads, supporting large-scale analytics.
- Integrated with tools like Databricks SQL and Power BI for interactive reporting.
- Supports ETL pipelines, data modeling, and governance.

Databricks' Lakehouse architecture simplifies traditional data warehousing by using a unified platform for both data engineering and analytics.

Q2: Optimizing Delta Tables, VACUUM, and Data Warehousing in Databricks

Delta Lake optimization techniques improve query performance and storage efficiency.

1. OPTIMIZE Command:
   - Rewrites data files to reduce file count and improve read performance.
   - Syntax: OPTIMIZE table_name [WHERE condition].

2. ZORDER:
   - Sorts data within files by specified columns to speed up selective queries.
   - Syntax: OPTIMIZE table_name ZORDER BY (column1, column2).

3. VACUUM:
   - Cleans up obsolete files no longer in the Delta table version history.
   - Helps free up storage and avoid clutter.
   - Syntax: VACUUM table_name [RETAIN num HOURS].

Data Warehouse and VACUUM

4. Best Practices:

- Run OPTIMIZE on large or frequently queried tables.
- Use ZORDER for high-cardinality filters (e.g., customer_id).
- Schedule VACUUM to maintain storage hygiene without impacting versioning.

5. Data Warehousing with Delta:

- Delta tables support star and snowflake schemas for analytics.
- Support for SQL joins, aggregations, and window functions.
- Integration with BI tools for dashboarding and reports.