LAB 04: Delta Lake Optimization

Duration: ~30 min
Day: 2
After module: M04: Delta Lake Optimization
Difficulty: Intermediate

Scenario

> *"RetailHub's orders table has grown to millions of rows. Dashboard queries are slowing down. Your task: apply optimization techniques – OPTIMIZE, Z-ORDER, VACUUM, and evaluate Liquid Clustering – to bring query times back to acceptable levels."*

Objectives

After completing this lab you will be able to: - Run `OPTIMIZE` to compact small files - Apply `ZORDER BY` for query-specific optimization - Run `VACUUM` to clean up obsolete files - Use `DESCRIBE DETAIL` to inspect table metrics - Understand Liquid Clustering configuration

---

Part 1: Analyze Current State (~5 min)

Task 1: Check Table Metrics

Use `DESCRIBE DETAIL` on your orders table to check: - Number of files - Total size in bytes - Partitioning columns

<screen = DESCRIBE DETAIL output showing numFiles, sizeInBytes, and partitionColumns for the orders table>

Part 2: OPTIMIZE & ZORDER (~10 min)

Task 2: Run OPTIMIZE

Run `OPTIMIZE` on the orders table. Compare `numFiles` before and after.

> **Exam Tip:** `OPTIMIZE` compacts small files into larger ones (target ~1GB). It does NOT remove obsolete files – that's VACUUM's job.

Task 3: ZORDER BY

Run `OPTIMIZE ... ZORDER BY (customer_id)` for queries that frequently filter by customer_id.

> **Exam Tip:** Z-ORDER co-locates related data in the same files, reducing the amount of data scanned. Best for high-cardinality, frequently filtered columns. Cannot be combined with Liquid Clustering.

Part 3: VACUUM (~10 min)

Task 4: Check Obsolete Files

After OPTIMIZE, old files still exist. Check table history.

Task 5: VACUUM

Run `VACUUM` to remove files older than the retention threshold.

> **Exam Tip:** Default retention is 7 days. Setting `delta.retentionDurationCheck.enabled = false` bypasses the safety check (NOT recommended in production).

<screen = VACUUM command output showing number of files deleted>

Part 4: Liquid Clustering (~5 min)

Task 6: Create a Liquid Clustered Table

Create a NEW table with Liquid Clustering enabled:

```
CREATE TABLE ... CLUSTER BY (column)
```

Compare physical layout with the Z-ORDER table.

> **Exam Tip:** Liquid Clustering replaces partitioning AND Z-ORDER. It's incremental (OPTIMIZE triggers it automatically) and supports column changes via `ALTER TABLE ... CLUSTER BY`.

Summary

In this lab you: - Analyzed table metrics with DESCRIBE DETAIL - Compacted small files with OPTIMIZE - Applied Z-ORDER for query optimization - Cleaned up obsolete files with VACUUM - Created a Liquid Clustered table

> **What's next:** LAB 05 - Set up streaming ingestion with Auto Loader.