LAB 04: Delta Lake Optimization

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

---

Scenario

> *“The Bronze layer is growing with many small files. Your job is to optimize the Delta tables — compact files, apply Z-ORDER for faster queries, clean up with VACUUM, explore Liquid Clustering, and handle a data skew scenario.”*

Objectives

After completing this lab you will be able to: - Inspect table metrics with `DESCRIBE DETAIL` - Compact small files with `OPTIMIZE` - Apply Z-ORDER for query optimization - Clean obsolete files with `VACUUM` - Create a Liquid Clustered table - Detect and handle data skew with `broadcast()` join

---

Prerequisites

- Cluster running and attached to notebook
- Setup cell creates test tables automatically

Tasks Overview

Open `LAB_04_code.ipynb` and complete the `# TODO` cells.

Task	What to do	Key concept
Task 1	Inspect Table Metrics	`DESCRIBE DETAIL` — check `numFiles` , `sizeInBytes`
Task 2	OPTIMIZE	Compact small files into larger ones
Task 3	ZORDER BY	`OPTIMIZE ... ZORDER BY (customer_id)`
Task 4	VACUUM	`DRY RUN` first, then execute with 0 hours retention
Task 5	Liquid Clustering	`CREATE TABLE ... CLUSTER BY (col)`
Task 6	Detect and Handle Data Skew	Identify skew with GROUP BY, fix with `broadcast()` join

---

Detailed Hints

Task 1: DESCRIBE DETAIL

- Command: `DESCRIBE DETAIL table_name`
- Look at `numFiles` and `sizeInBytes` columns

Task 2: OPTIMIZE

- Command: `OPTIMIZE table_name`
- Compare `numFiles` before and after

Task 3: ZORDER

- Syntax: `OPTIMIZE table_name ZORDER BY (column_name)`

Task 4: VACUUM

- First: `VACUUM table_name RETAIN 0 HOURS DRY RUN` (preview only)
- Then: `VACUUM table_name RETAIN 0 HOURS` (execute)
- Must disable safety check first for 0 hours retention

Task 5: Liquid Clustering

- Syntax: `CREATE TABLE ... CLUSTER BY (col) AS SELECT ...`
- Verify with `DESCRIBE DETAIL` — check `clusteringColumns`

Task 6: Data Skew

- Detect: `SELECT customer_id, COUNT(*) ... GROUP BY ... ORDER BY count DESC`
- Fix: `from pyspark.sql.functions import broadcast` → `df.join(broadcast(small_df), ...)`

Summary

In this lab you: - Inspected table metrics with DESCRIBE DETAIL - Compacted small files with OPTIMIZE - Applied Z-ORDER for query optimization - Cleaned obsolete files with VACUUM - Created a Liquid Clustered table - Detected data skew and resolved it with broadcast join

> **Exam Tip:** Liquid Clustering replaces both partitioning and Z-ORDER. Use `ALTER TABLE ... CLUSTER BY (new_cols)` to change clustering columns without rewriting data. For data skew, `broadcast()` works when one side is small (< 10MB by default). AQE handles skew automatically in most cases.

> **What’s next:** In LAB 05 you will set up Auto Loader for streaming ingestion and explore Change Data Feed (CDF).