DAY 5 – Delta Lake Advanced

1️⃣ Time Travel (Version History)

What it is:

Delta Lake keeps every change made to a table. You can go back in time and query old versions of data.

Why it matters:

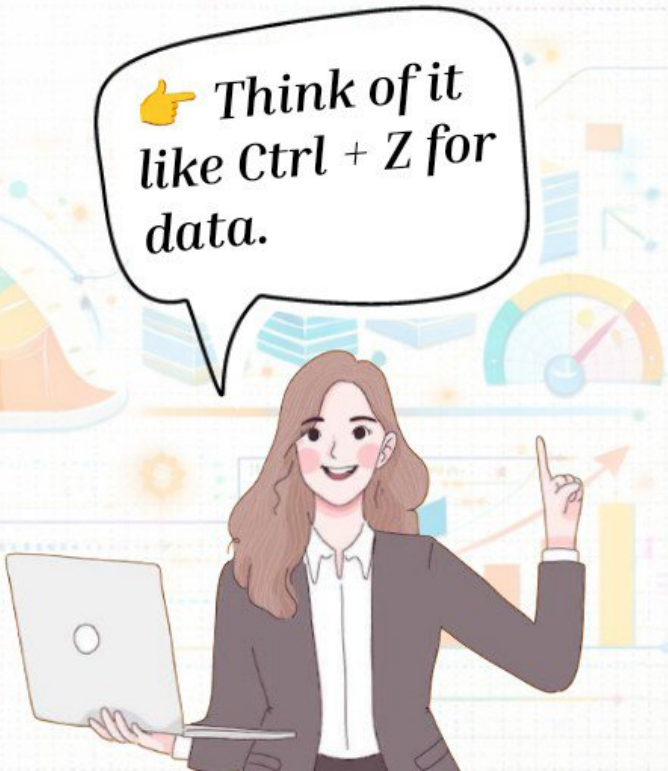
- *Recover deleted/updated data*
- *Debug data issues*
- *Compare past vs current data*

Example (SQL):

```sql
-- View table history
DESCRIBE HISTORY sales_delta;

-- Query a specific version
SELECT * FROM sales_delta VERSION AS
OF 2;

-- Query by timestamp
SELECT * FROM sales_delta TIMESTAMP AS
OF '2024-01-10';
```



2️⃣ MERGE Operations (Upserts)

What it is:

MERGE allows you to INSERT, UPDATE, or DELETE data in one command.

Why it matters:

- *Used for incremental loads*
- *Avoids full table overwrite*
- *Very common in real-world pipelines*

Example:

```
MERGE INTO target_table t
USING source_table s
ON t.id = s.id
WHEN MATCHED THEN
  UPDATE SET *
WHEN NOT MATCHED THEN
  INSERT *
```

3️⃣ OPTIMIZE & ZORDER

🔹 *OPTIMIZE*

What it does:

·Combines many small files into bigger files

·Improves query performance

```
OPTIMIZE sales_delta;
```

🔹 *ZORDER*

What it does:

·Physically organizes data based on frequently filtered columns

```
OPTIMIZE sales_delta
ZORDER BY (customer_id);
```

👉 *Faster queries when filtering on customer_id.*

4️⃣ VACUUM (Cleanup Old Files)

What it is:

Deletes old, unused data files that Delta no longer needs.

Why it matters:

- *Saves storage*
- *Keeps tables clean*

```
VACUUM sales_delta RETAIN 168 HOURS;
```

⚠️ *Default retention = 7 days (168 hours)*
⚠️ *After VACUUM → old versions cannot be recovered*

Day 5 Key Takeaway

💡 *Delta Lake = Reliable Data Engineering*

✔ *Version control for data*
✔ *Safe upserts*
✔ *High-performance queries*
✔ *Storage optimization*

👉 *On to Day 6* 🚀