Functions

This describes how to use the functions to build a few different kinds of pipelines.

General Cases

These are for the silver layer. They can be done with streaming, but it might be easier to use batch. Streaming assumes no duplicates in the incoming stream, whereas batch must deduplicate. Streaming seems easier until there are problems, then batch becomes easier due to fewer variables.

1. **SCD2 model**:
   * Heavier on compute/storage.
   * Maintains full history.
   * Common in regulated industries, finance, HR, audit needs.
   * Idempotent: Yes but fragile (needs to run every day).
2. **Overwrite/snapshot model** (stateless):
   * Silver always reflects the latest state.
   * No historical tracking.
   * Simple, fast, cost-efficient.
   * Cheaper than upsert for small tables.
   * Idempotent: Yes.
3. **Upsert (MERGE) model**:
   * Latest changes merged into silver.
   * No historical tracking.
   * Suitable in CDC pipelines without SCD2 needs.
   * Cheaper than snapshot for big tables with small updates.
   * Idempotent: Yes with proper business key.

SCD2 Streaming

Bronze

    "read\_function": "functions.stream\_read\_cloudfiles",

    "transform\_function": "functions.bronze\_standard\_transform",

    "write\_function": "functions.stream\_write\_table",

    "build\_history": "true",

Silver

    "read\_function": "functions.stream\_read\_table",

    "transform\_function": "functions.silver\_scd2\_transform",

    "write\_function": "functions.stream\_write\_scd2\_table",

    "upsert\_function": "functions.scd2\_upsert",

    "build\_history": "false",

    "ingest\_time\_column": "derived\_ingest\_time",

SCD2 Batch

Bronze

    "read\_function": "functions.stream\_read\_cloudfiles",

    "transform\_function": "functions.bronze\_standard\_transform",

    "write\_function": "functions.stream\_write\_table",

    "build\_history": "true",

Silver

    "read\_function": "functions.read\_table",

    "transform\_function": "functions.silver\_scd2\_transform",

    "write\_function": "functions.write\_scd2\_table",

    "upsert\_function": "functions.scd2\_upsert",

    "build\_history": "false",

    "ingest\_time\_column": "derived\_ingest\_time",

Overwrite/Snapshot Streaming

Bronze

    "read\_function": "functions.stream\_read\_cloudfiles",

    "transform\_function": "functions.bronze\_standard\_transform",

    "write\_function": "functions.stream\_write\_table",

    "build\_history": "true",

Silver

Overwrite/Snapshot Batch

Bronze  
Silver

Upsert (Merge) Streaming

Bronze

Silver

Upsert (Merge) Batch

Bronze

Silver

|  |  |
| --- | --- |
| **Function Name** | **Description** |
| stream\_read\_cloudfiles | Reads streaming data using Auto Loader (cloudFiles). |
| stream\_read\_table | Reads a Delta table as a streaming source. |
| read\_table | Reads a Delta table in batch mode. |
| read\_windowed\_snapshot | Reads latest row per key using window and row\_number. |
| read\_latest\_ingest | Reads only rows with the latest ingest\_time value. |
| bronze\_standard\_transform | Applies common bronze transforms. |
| silver\_standard\_transform | Renames, casts types, and adds hash/file metadata. |
| silver\_scd2\_transform | Applies standard silver transforms and adds SCD2 columns. |
| silver\_scd2\_catchup\_transform | Filters input to only include rows newer than the target table. |
| add\_scd2\_columns | Adds SCD2 columns. |
| add\_source\_metadata | Adds source\_metadata struct column. |
| add\_row\_hash | Adds a SHA-256 hash column if enabled. |
| clean\_column\_names | Normalizes column names. |
| rename\_columns | Renames columns based on a mapping. |
| cast\_data\_types | Casts columns based on a mapping. |
| overwrite\_table | Overwrites destination table in batch mode. |
| stream\_write\_table | Writes DataFrame to table using streaming sink. |
| stream\_upsert\_table | Uses foreachBatch to perform streaming upsert. |
| upsert\_microbatch | Returns a function that performs standard upsert merge logic. |
| stream\_write\_scd2\_table | Streaming SCD2 table write using upsert with deduplication. |
| scd2\_upsert | Upsert function for batch SCD2 tables. |
| stream\_scd2\_catchup\_write | Streaming catch-up SCD2 write, to apply late arriving data. |
| scd2\_catchup\_outer\_upsert | Loops through distinct ingest times, calling scd2\_upsert. |
| scd2\_write | Batch SCD2 write using deduplicated input. |