Streaming SCD2 Rescue Procedure

This will show you how you can use the derived ingest time column to rebuild or recreate or rescue a silver streaming SCD2 table, and subsequently turn streaming back on so that the next job run works correctly.

Overview of Procedure

1. Update JSON settings: ingest time.
2. Rescue the table.
3. Update JSON settings: starting version.
4. Run job to re-establish streaming.
5. Update JSON settings: remove starting version.

Procedure

Update JSON Settings for Ingest Time

The rescue script works by looping through an ingest time to figure out SCD2 fields at each iteration*.* Typically I will use a column called derived ingest time which is found by looking at something in the path that gives away when the file was created. Look in transforms.py to see how this works. This is one way to do it, and there is also the real ingest time, which you should also have since bronze is assumed to be intact.

Add this (or something that will work) to your settings:

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

Rescue the Table

Run the rescue\_silver\_table script found in utilities. It does drop the table and removes the checkpoint folder; then it loops through each ingest time it finds and does normal table upserts.

Update JSON Settings for Starting Version

This will tell your next job to stream starting at this version. Do not use the versionAsOf option since that will get you a static dataframe and cause problems with your writestream function.

First, find the latest version:

describe history edsm.bronze.systemsPopulated

Now, add latest version from bronze to your settings. Instead of 5 put your next-to-latest version from describe history. The reason is you want to establish the stream that went from that version to the latest.

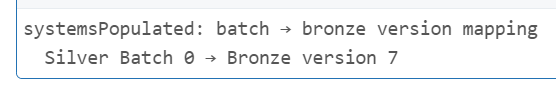
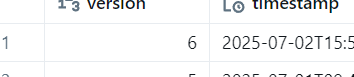
"startingVersion": "5"

Run the Job

Find your job in Jobs & Pipelines and run it.

Inspect the Tables and Offsets

Find a notebook in utilities called inspect\_checkpoint\_folder. Run it against your silver table and it will print output that shows which silver batch load came from which bronze version. Having done that, also run describe history on the bronze table. Compare these and verify. *Note: I have not figured out why the* ***reservoirVersion*** *from the offset is always one higher than the bronze table version. “Some new ideas are here needed.”*

 **→** 

Remove Starting Version

Go back to the settings and remove this line (or whatever you set it to).

"startingVersion": "5"

The End