End to End Lakehouse Pipeline

Input file folders

Folder Name	File Name	Purpose
cust_inbound	customers_20260120_01.csv customers_20260121_01.csv customers_20260122_late.csv customers_20260123_01.csv	These files have customer data (name, email, address). This data act as primary record for customers.
cust_comm_inbound	customers_comm_pref_20260123_01.csv	This file is for the referential integrity validation.

Attributes

File Name	Attribute list
customers_*.csv	cust_id, event_type, event_ts, first_name, last_name, email, phone, addr_city, addr_state, addr_postal
customers_comm_pref_*.csv	cust_id, event_type, event_ts, reach_via

Delta Tables

- bronze_cust
- bronze_comm
- silver_cust
- silver_comm
- quarantine_cust
- quarantine_comm
- dim_customer_scd2

Data Quality Checks

Required fields (i.e., cannot be NULL)	cust_id, event_ts
Length check (value to be 2)	addr_state
Referential Integrity	**cust_id** from **customers_*.csv** are valid entries. These are considered as 'primary/parent' record, over, record in customers_comm_pref_*.csv

Code files

Bronze.py	Two additional attributes are added: ingest_ts, source_file
Silver.py	cust_id is cast to integer event_ts is cast to timestamp rest of fields are cleaned by trim () and lower ()
SCD2.py	eff_end is NULL for current record Table is written with mode("overwrite") Logic is written using temporary views

Note: Notebooks are also provided.