Overview

This project aims to streamline table construction in Databricks by providing ready-made components and example notebooks. The idea was to build a lot of pipelines quickly.

The functions folder implements the pipeline pieces. For example, functions.utility defines the job-type presets used by many settings files. The utilities folder has scripts that do useful things like rescue a table or tell you what version of bronze the silver checkpoint is pointing to. There is a sanity checker and a bad records checker that both run inside the job.

Settings files under layer\_01\_bronze and layer\_02\_silver define how each table is built. If a gold folder and settings are given they will be created. The pipeline merges individual table settings with sensible defaults for each job type, including streaming checkpoint locations and schema locations. Read the Simple Settings Reference for more about simple settings and job types.

.  
├─ layer\_01\_bronze/ # raw ingestion configs  
├─ layer\_02\_silver/ # cleaned & deduplicated configs  
├─ layer\_03\_gold/ # optional gold layer  
├─ functions/ # PySpark helper modules  
│ ├─ read.py, transform.py, write.py, history.py  
├─ utilities/ # notebooks and shell scripts  
├─ dashboards/ # data exploration notebooks  
├─ sandbox/ # ad‑hoc notebooks  
└─ \*.ipynb # workflow notebooks (00\_job\_settings …)

In the databricks-utilities repo there is a little utility called get-schema.py which determines some information by inspecting a datafile. This can be copy/pasted directly into a table’s job settings.

In a Documentation repo, there are procedures for doing the most common tasks like building a table pipeline, resetting the checkpoint folder, or rescuing a silver table. There are also reference documents.