This project is to provide Mid level data engineers a hands on experience to learn basic data engineering principles.

Context

Imagine, you are working for a Health system, where they want to process data from a source relational Database like Oracle Database. However, you are not permitted to access the Oracle tables directly. You are expected to ingest the data from an output of a real time Change Data Capture Tool like Qlik Replicate. This Qlik Replicate (CDC tool) reads the logs of the Oracle tables and writes the data in the landing zone of your data lake. The CDC tool can write the full dataset of the table and incremental records (captures deleted records too). The business request here it to create a pipeline using databricks and ADF that will process the data in the landing zone of the data lake and write it to a delta table target in a refined zone in the data lake. Your pipeline should run every 1 hr and be able to process any new files it sees in the landing zone. After processing a file in the landing zone, it should move it to the archive folder. Your pipeline and databricks notebooks should be parameterized and driven by a config file stored in a location(governance zone) in your data lake. This config file will have the table names and the primary key columns of that table that needs to be processed from the landing zone.

Prerequiste tools needed

1. Azure Data lake Gen 2
2. Data factory
3. Databricks

The goal is to build an ETL pipeline that can process both full and incremental load from the Data lake folder source table and load to the target delta table in data lake

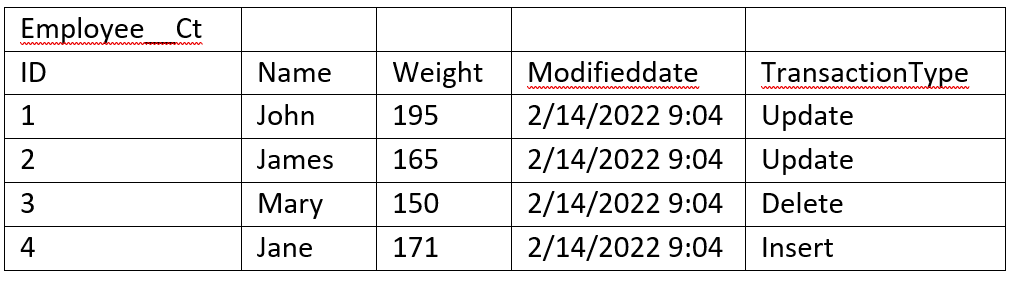
Prerequisite steps

1. Practice your azure skills but first provisioning a Azure data lake gen 2, data factory and databricks
2. Make sure you set up the write permissions and network settings so data factory can connect to the data lake and databricks.
3. In the Azure data lake gen 2 storage account create 3 containers – landing, refined and governance
4. In the landing container, create two directories – input and archive.
5. Upload the files attached to the input directory – it will represent 3 tables one is the Employee table, another is for the EmployeeAddress, then last is the Address table. They are in json format
6. Inside the input directory, we will have sub directory for the 3 tables. This is where the CDC tool will write the data for the respective tables both full and incremental load files
7. The config file with the tables name and primary key columns will be stored in the governance zone container
8. After processing write the data of each table into the refined zone container as a delta file format ( use databricks)
9. The entire pipeline needs to be one meta data driven pipeline, with one parameterized databricks notebook to process all tables…driven by the config file
10. Note that the incremental files can have duplicate records, so you need to ensure you only process the last transaction based on the primary key
11. The three tables you will have are
    1. Customer – Pk = CustomerID
    2. CustomerAddress--- PK –CustomerID,AddressID
    3. Address – Pk – AddressID

PK means Primary key

1. Create a configuration file with the 3 tables and primary keys column names so it can be used to control the pipeline

Example of the final table output

[](https://user-images.githubusercontent.com/20451211/175101501-12abfbd2-34f1-4872-b5b1-be5e5418a2c4.png)