High Level Design: Synapsis Coal Mining Company Data Warehouse

Context

Objective

Make SSOT (Single Sources of Truth) of data in Synapsis Coal Mining Company.

Design

Overview

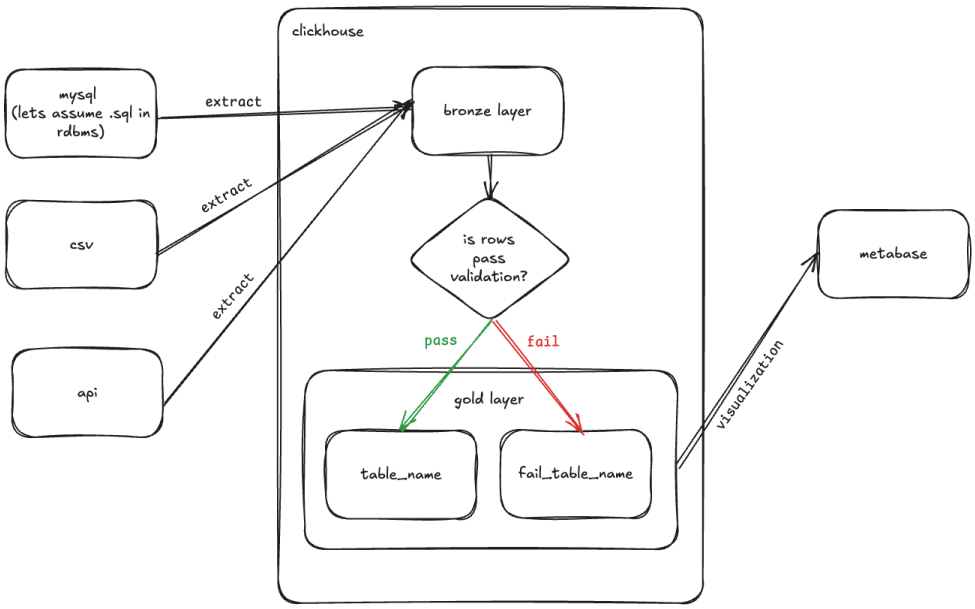

Image 1: the data flow

Data Sources

1. Synapsis Coal Mining Company
   - MySQL database [production_logs, mines]
   - Files (CSV) [equipment_sensors.csv]
   - API [https://api.open-meteo.com/v1/forecast]

Detailed Design

Details

1. Description
   Based on the data flow at *Image 1: the data flow*, data will be extracted from various sources (mysql database, csv file, and api). Implement a medallion architecture and slow changing dimension concept, data from sources will be injected to *layer_bronze* database as is like the sources, no transformation or another process do in this process just add some metadata like when rows inserted and surrogate key for id.
   Then, created and inserted data into a summary table called *summary_coal_mining* that performs transformation and aggregation data to meet the business need. This table will be created in the layer_gold database.
   In *summary_coal_mining* table in *layer_gold* database, also perform a validation process that checks conditions that are explained in the **Data Quality** stage. If rows data don't meet the criterias, it will be deleted from *summary_coal_mining* table and inserted to *fail_summary_coal_mining*.

Components and Tech Stack

1. Data sources
   a. Synapsis Coal Mining
      - MySQL database
        - coal_mining.production_logs
        - coal_mining.mines
      - Files (CSV)
        - equipment_sensors.csv
      - API
        - https://api.open-meteo.com/v1/forecast
2. ETL process
   a. ETL image — Python
      Custom image that performs extraction from data sources that are mentioned in point *1. Data Sources*, perform transformation and aggregation, and perform validation
3. Data visualization
   a. Metabase
      Visualization of data
4. Docker
   Infrastructure deployment for *point* 1, 2, 3

Data Quality

1. total_production_daily > 0.
2. equipment_utilization is between 0 and 100.
3. weather data is not null for production day.