End To End Data Engineering Project using Healthcare Data

25 October 2024

Auni Khairina Azman

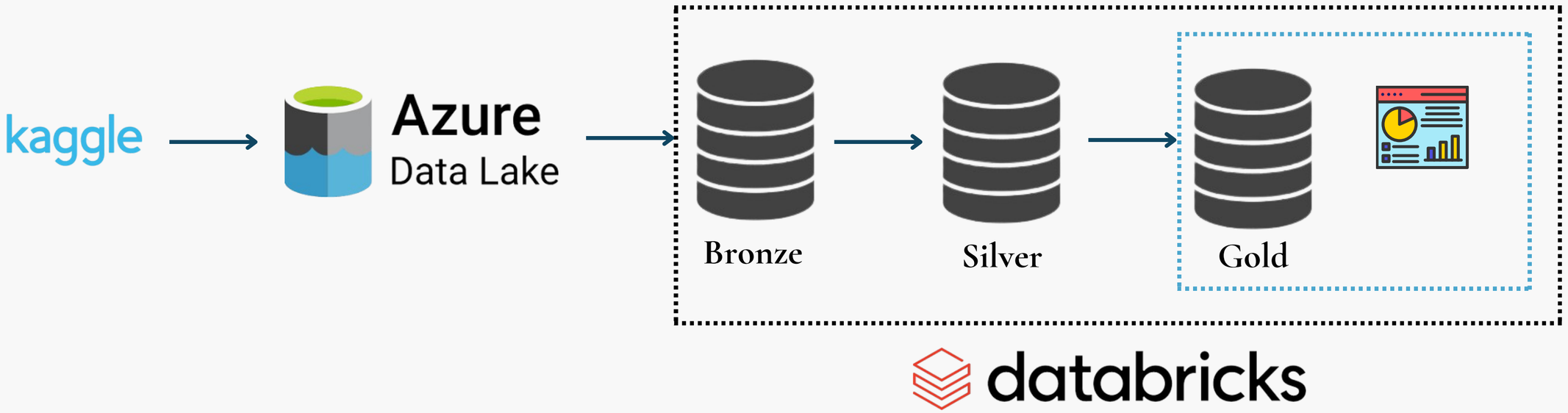
Problem Statement

Assessing Vaccine Coverage Trends

To analyze the trends in vaccination coverage for BCG, Hep B, and DTP among children under 1 year old in Malaysia from 2010 to 2018, determining if there has been a consistent increase, decrease, or fluctuation over the years.

Data Source : Kaggle (https://www.kaggle.com/datasets/thedevastator/who-malaysia-health-indicators)

Architecture Diagram



Detail explanation in each processing notebook

Bronze

- Get adls connection
- Read csv file from adls
- Write data as parquet file in Bronze

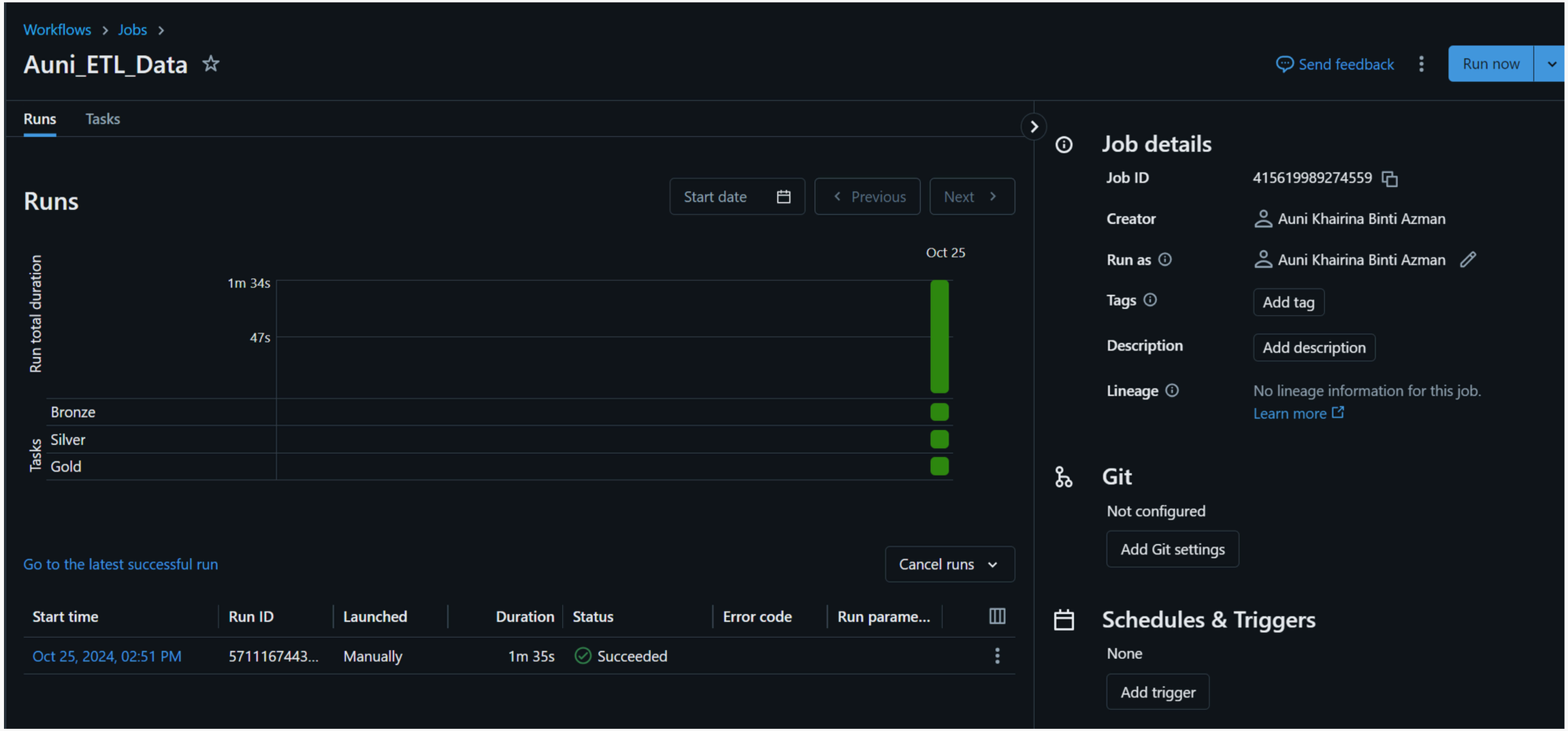
Silver

- Read parquet file from Bronze
- Remove anomaly from dataset (invalid data)
- Remove special characters & white space from variable column
- Convert value from String into Integer
- Select few columns to make analysis
- Write data as parquet file in Silver

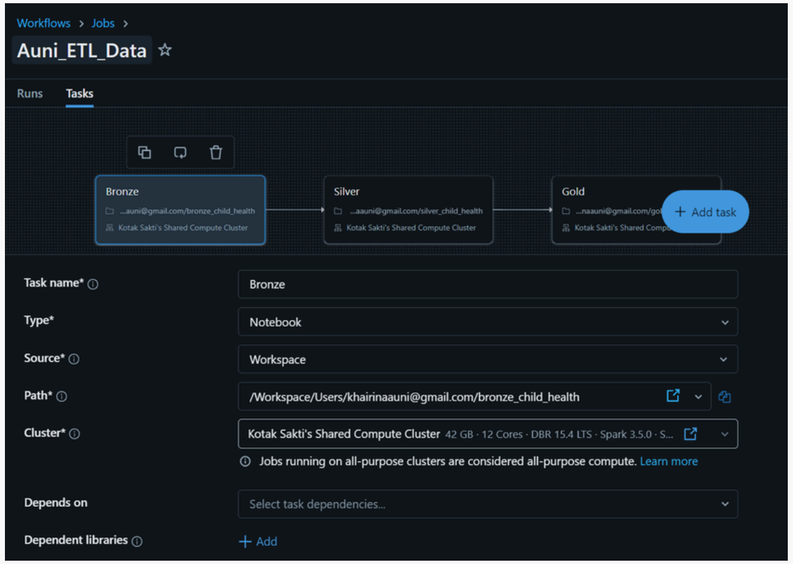
Gold

- Read cleaned BCG, DTaP, Hep B parquet files from Silver
- Combine those 3 files into a dataframe
- Add one more column “Immunization_Type”
- Filter “year” since 2010 until 2018 only
- Write data as parquet file in Gold
- Create line graph to visualize the data

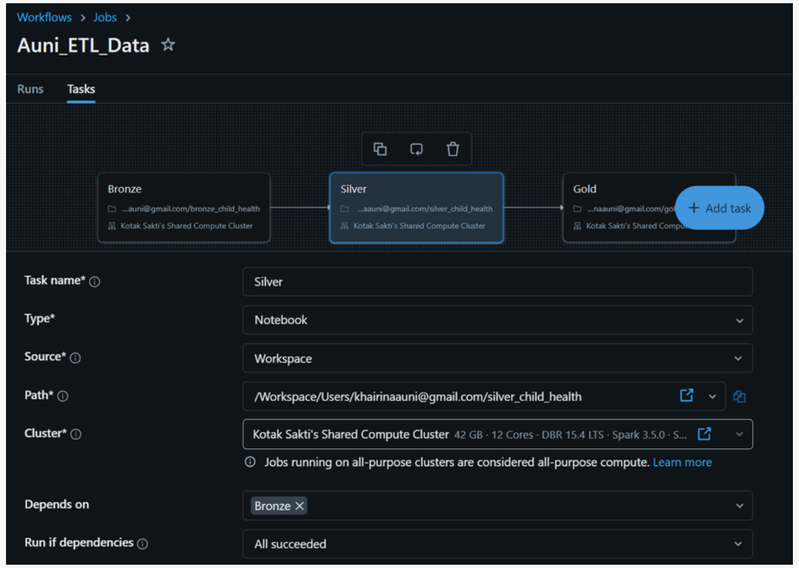
ETL Data Pipeline



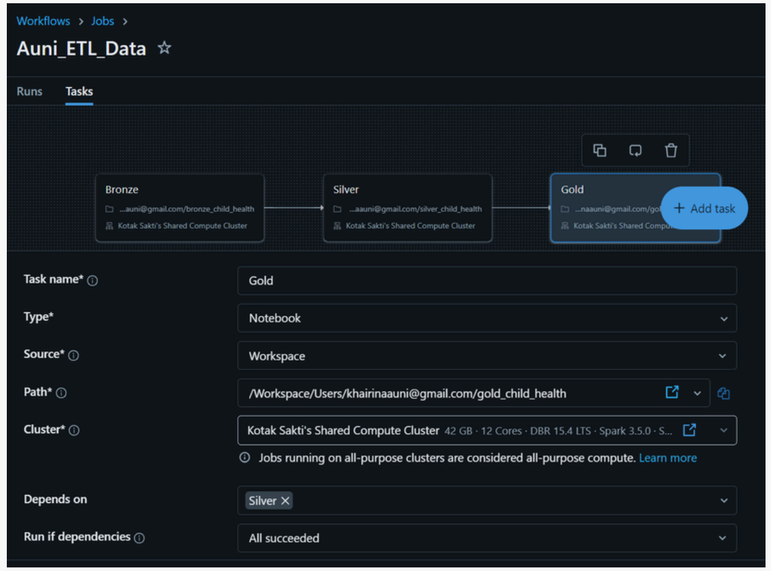
Detail explanation in each task in ETL



Bronze



Silver



Gold

Thank you