

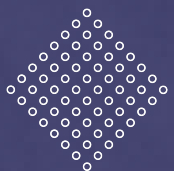
Final Project

Data Engineer

M.Abdurrahman Shidiq

22 March, 2023

Digitalskola, Batch 11





Problem Statement

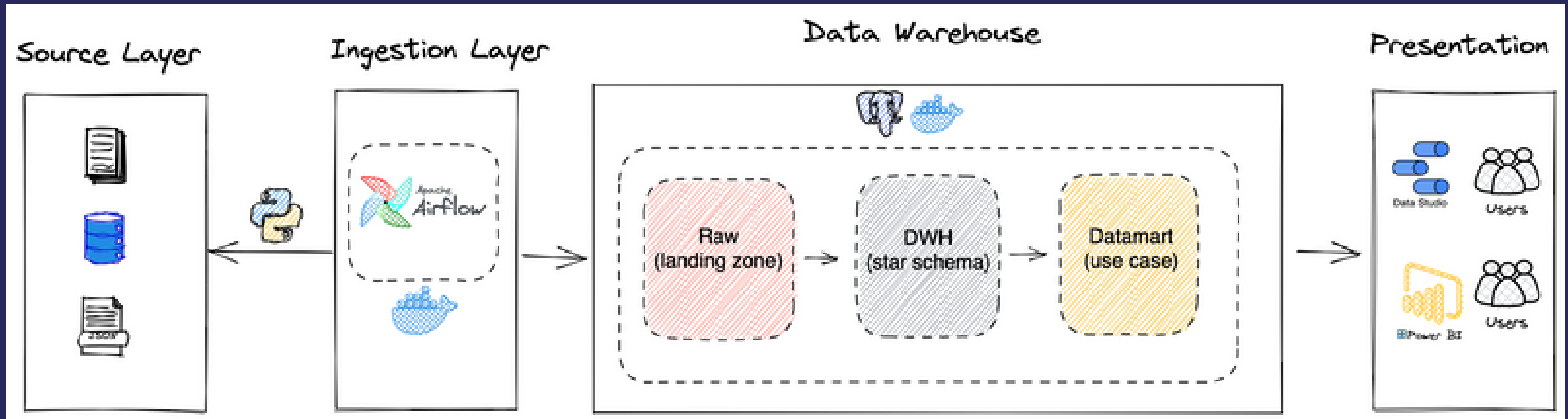
Sebagai Data Engineer, pada project ini kita diminta membangun sebuah data warehouse untuk kebutuhan analytics. Sehingga kebutuhan analisis bisa dilakukan pada OLAP bukan OLTP.

User ingin mendapatkan insight : Bagaimana pengaruh curah hujan & suhu terhadap review / rating sebuah restaurant

Goals

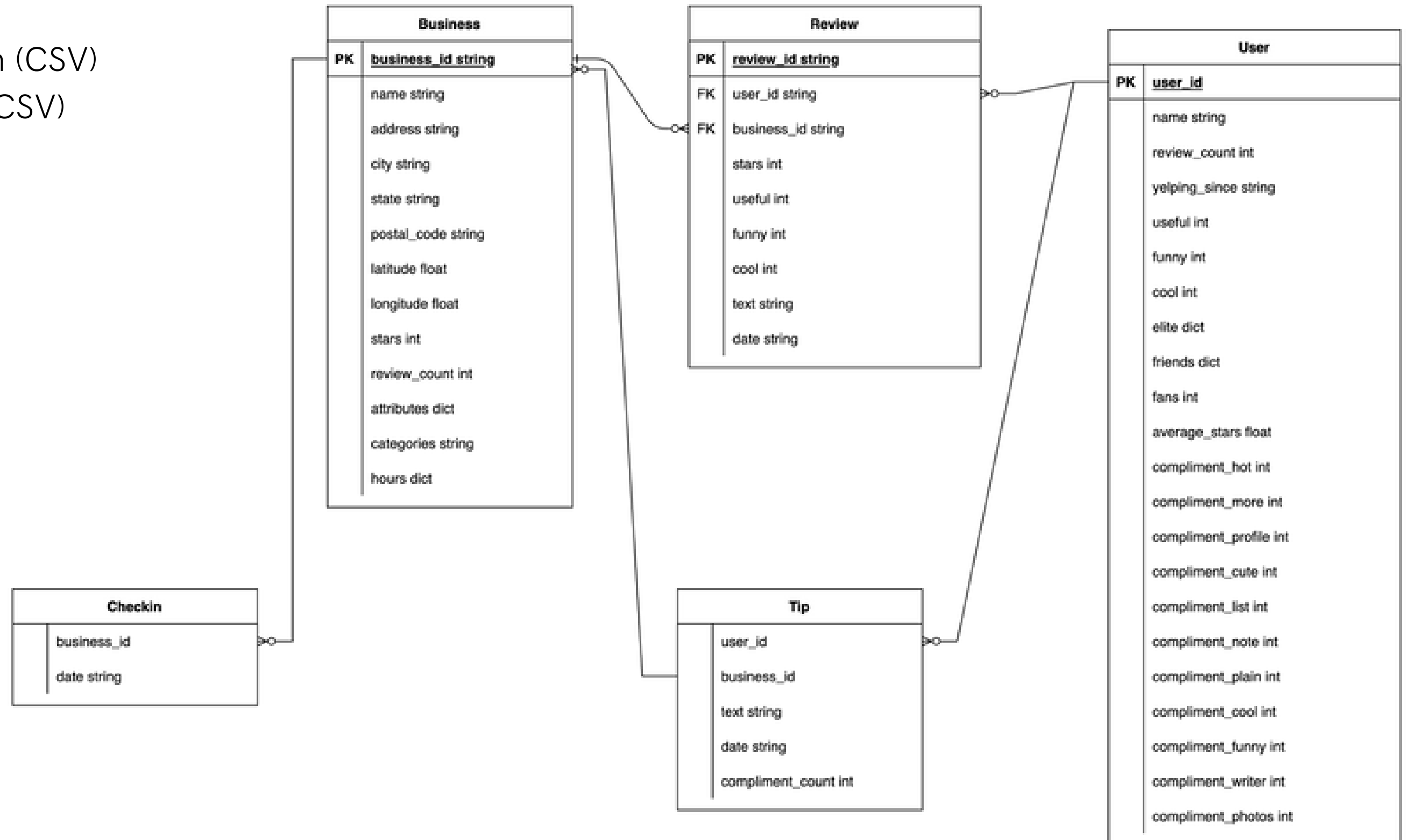
- Data Ingestion
- Data Warehousing (star schema)
- Presentation Layer (Dashboard / insight)

Architecture High Level Diagram

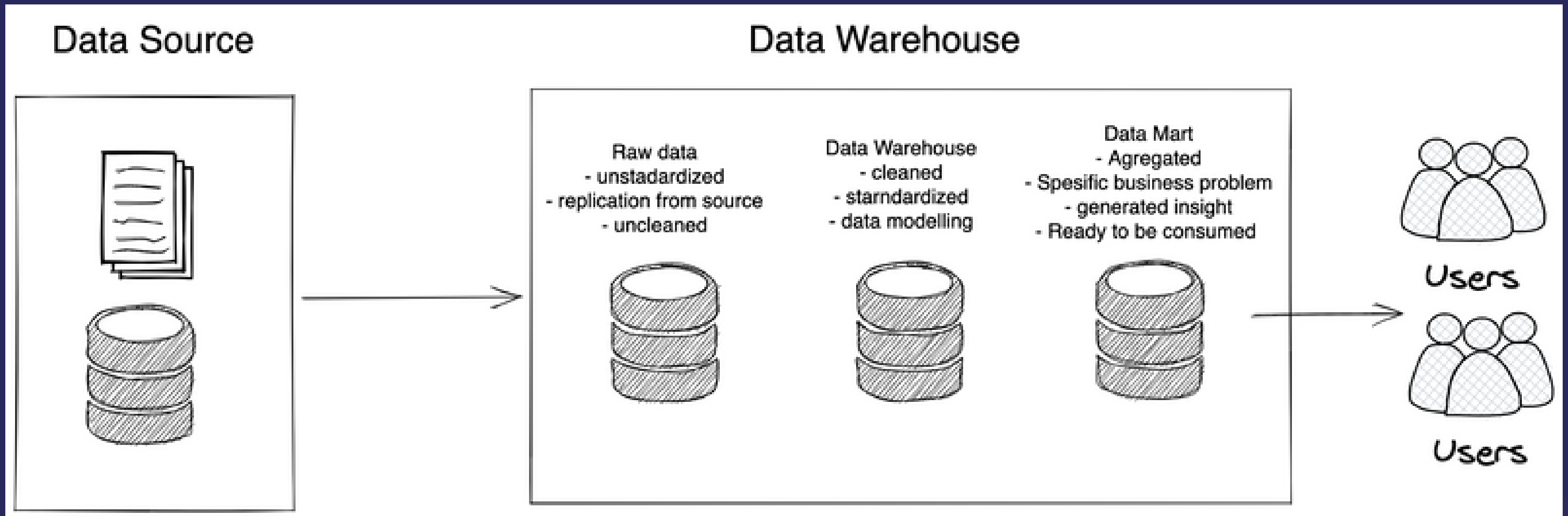


Data Source

- Yelp Kaggle (JSON)
- Las Vegas Percipitation (CSV)
- Temperature Degree (CSV)

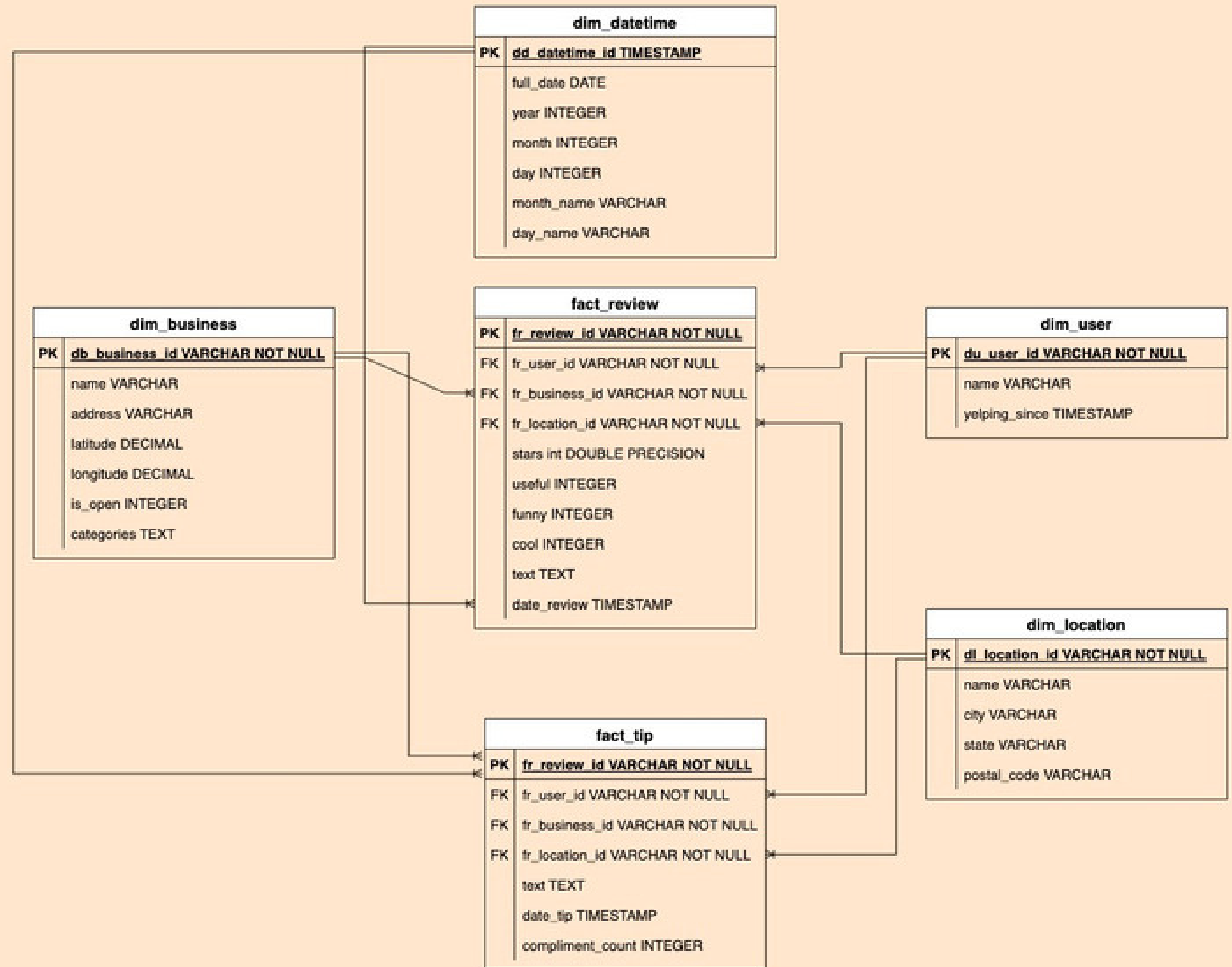


Architecture High Level Diagram



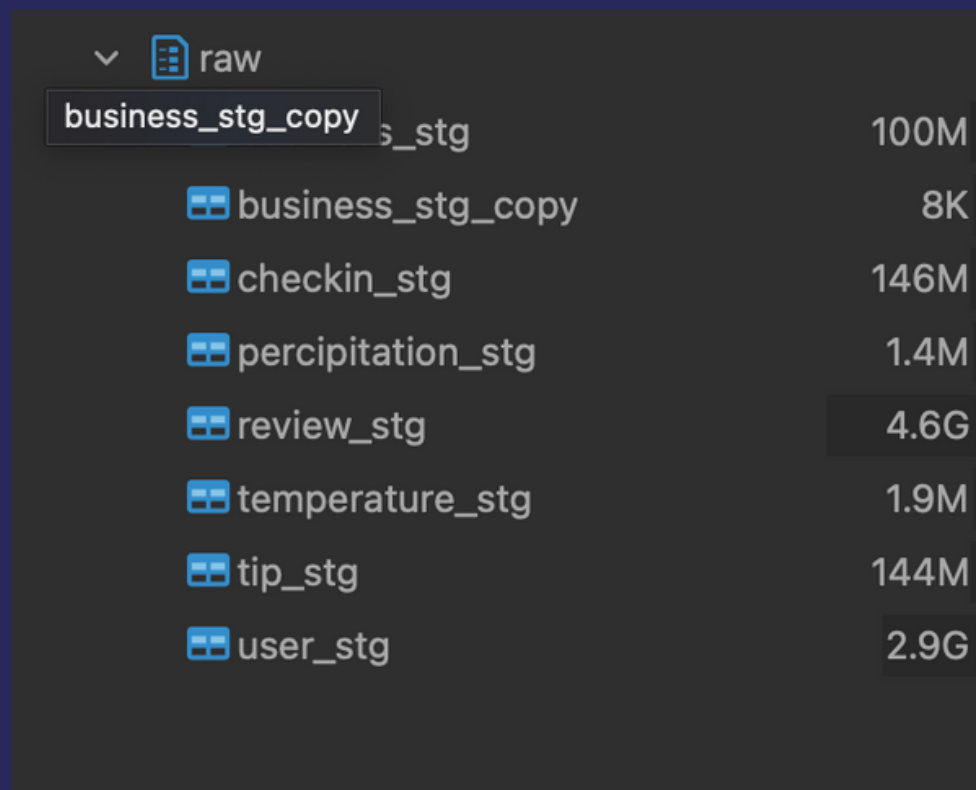
Data Modelling

Star Schema



Data Warehouse Layer on Postgres

RAW Layer

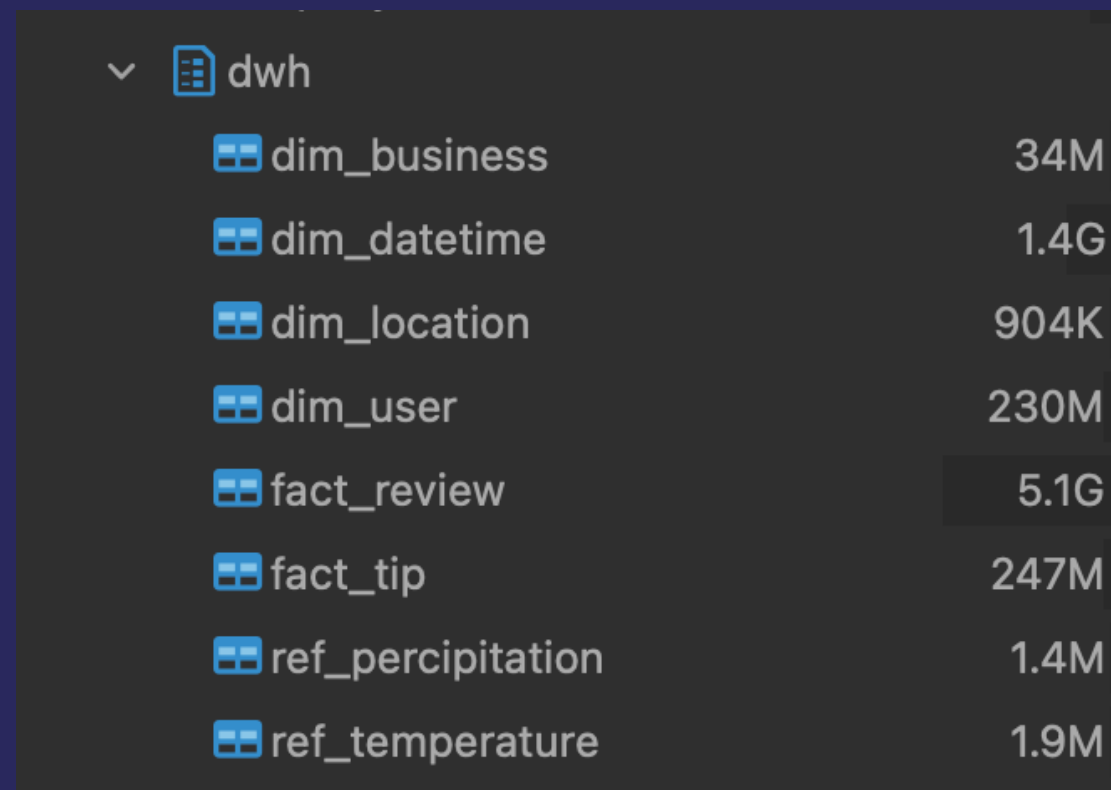


A screenshot of a database interface showing the 'raw' schema. It lists several tables with their sizes: business_stg_copy (100M), business_stg_copy (8K), checkin_stg (146M), percipitation_stg (1.4M), review_stg (4.6G), temperature_stg (1.9M), tip_stg (144M), and user_stg (2.9G).

raw	
business_stg_copy	100M
business_stg_copy	8K
checkin_stg	146M
percipitation_stg	1.4M
review_stg	4.6G
temperature_stg	1.9M
tip_stg	144M
user_stg	2.9G

Data yang masuk apa adanya seperti sourcenya, dan seluruh tipe datanya dibuat `STRING`

DWH Layer

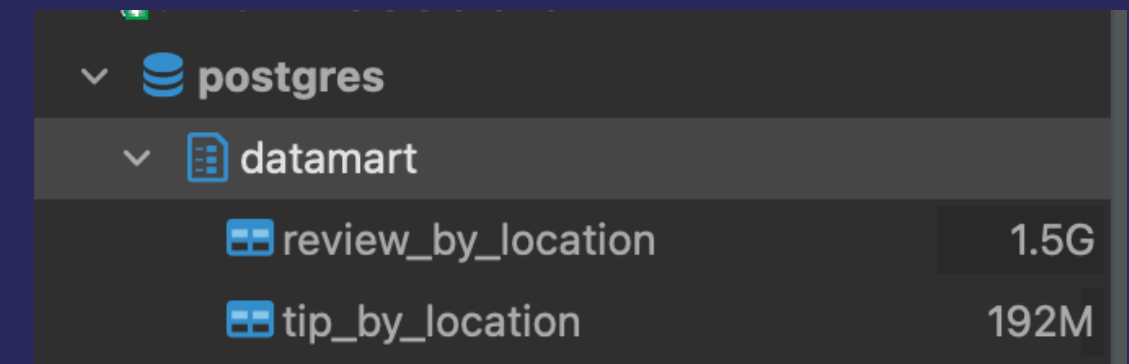


A screenshot of a database interface showing the 'dwh' schema. It lists several tables with their sizes: dim_business (34M), dim_datetime (1.4G), dim_location (904K), dim_user (230M), fact_review (5.1G), fact_tip (247M), ref_percipitation (1.4M), and ref_temperature (1.9M).

dwh	
dim_business	34M
dim_datetime	1.4G
dim_location	904K
dim_user	230M
fact_review	5.1G
fact_tip	247M
ref_percipitation	1.4M
ref_temperature	1.9M

Dari Raw layer, data akan distandarkan tipe datanya, dilakukan modelling & cleansing

Datamart Layer

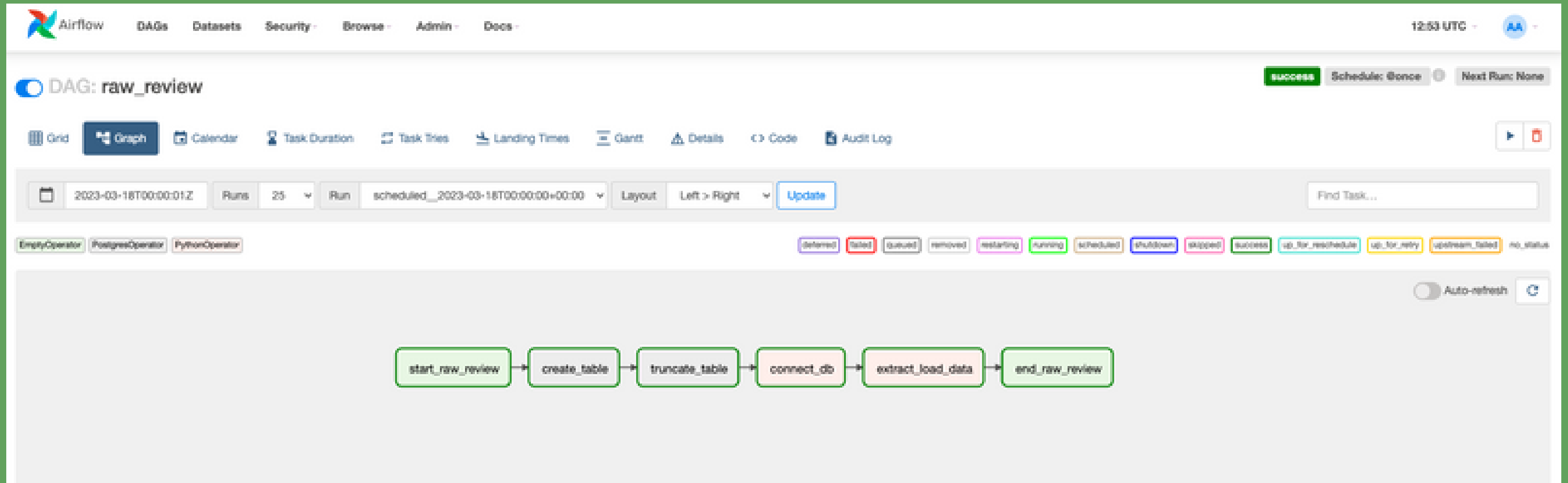


A screenshot of a database interface showing the 'postgres' schema. It lists two tables in the 'datamart' schema: review_by_location (1.5G) and tip_by_location (192M).

postgres	
datamart	
review_by_location	1.5G
tip_by_location	192M

Model Datamart yang digunakan adalah One Big Table (OBT). Table-table di Data mart dibuat spesifik menurut use casenya

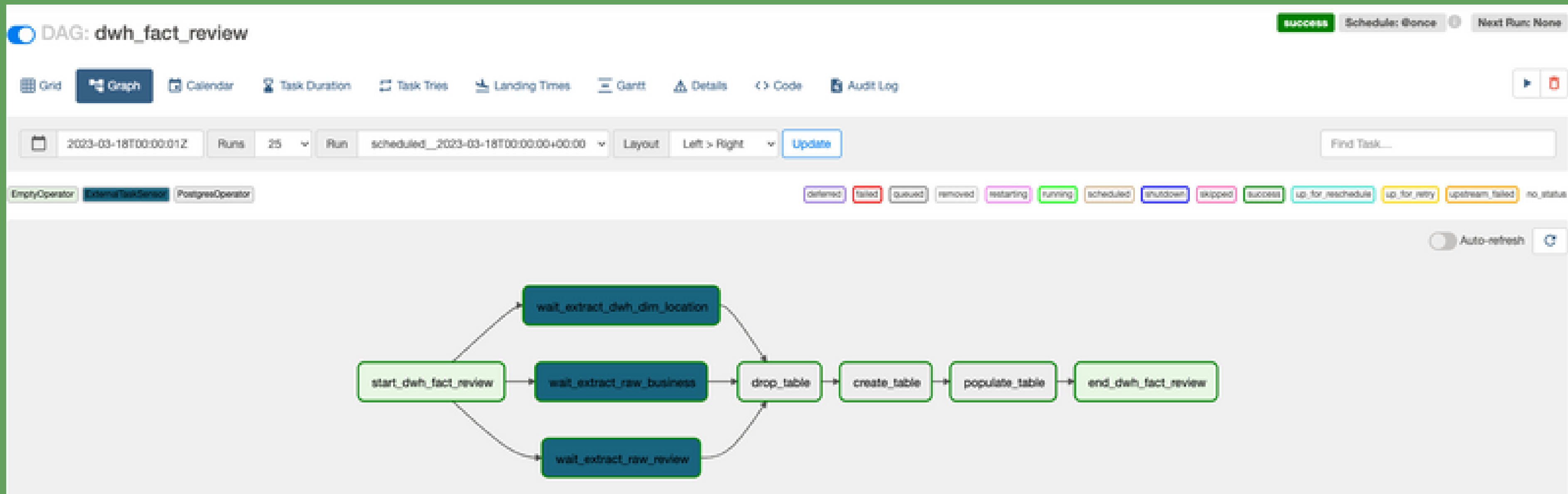
Data Ingestion DAG (RAW Layer)



Airflow Operator

- EmptyOperator / DummyOperator
- PythonOperator
- PostgresOperator

Data Model DAG (Data Warehouse Layer)

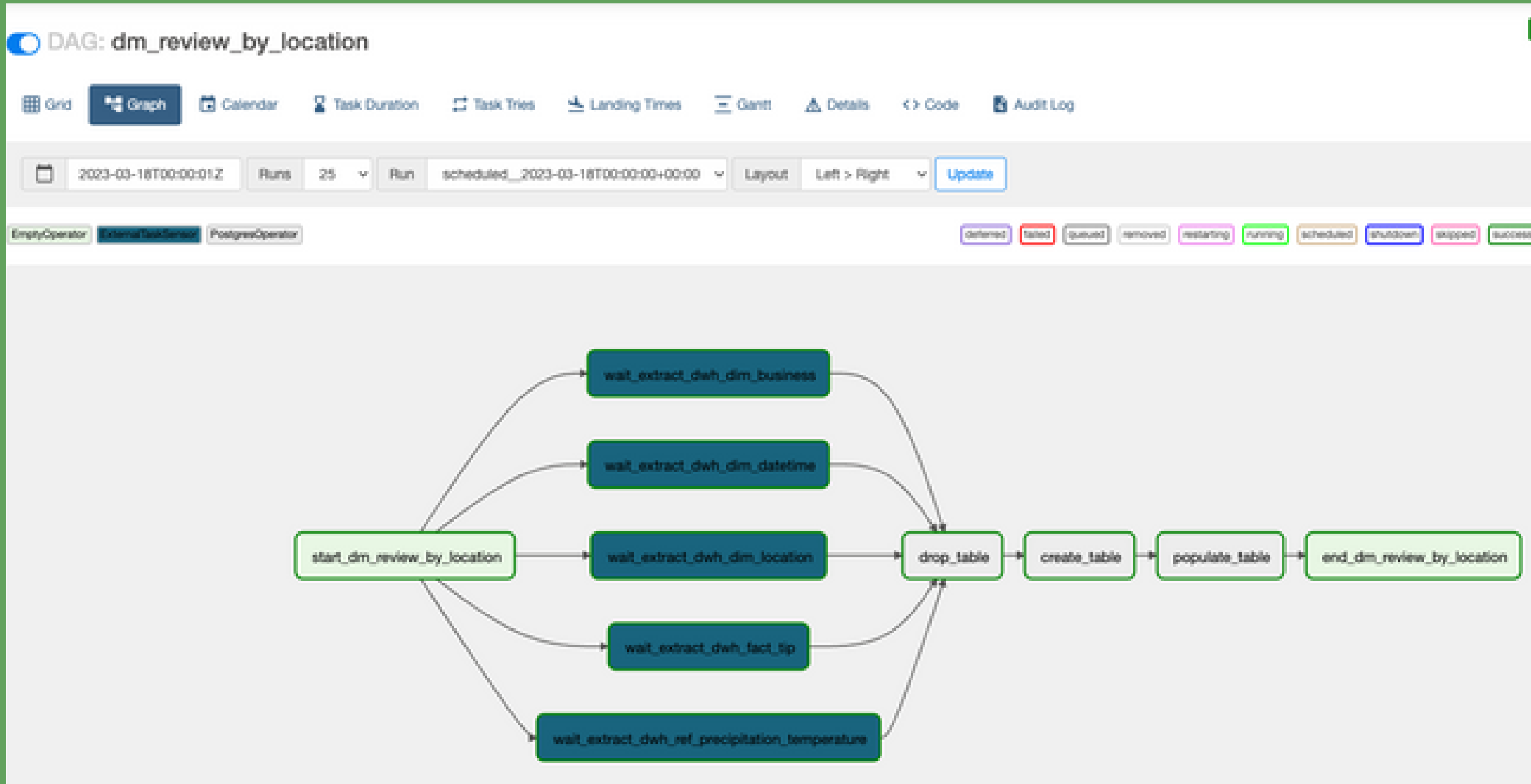


Airflow Operator

- ExternalTaskSensor
- EmptyOperator / DummyOperator
- PostgresOperator

Dengan menggunakan task sensing, maka DAG diatas akan berjalan setelah DAG raw_business, raw_review (ingestion) & dim_location telah selesai. sehingga tidak ada task yang **OVERLAP** satu sama lain

Data Model DAG (Data Mart Layer)



Airflow Operator

- ExternalTaskSensor
- EmptyOperator / DummyOperator
- PostgresOperator

Github Repository

<https://github.com/abdurrahmanshidiq/etl-dwh-project>





Thank you

Terima Kasih