DIBIMBING

Designing a Scalable ETL Data Pipeline for Airbnb Amsterdam Insights

Mochamad Reza Rahadi

DIBIMBING

Introduction

Mochamad Reza Rahadi

Education

- Universitas Indonesia

Working Experience

- Data and AI Optimization - Eklipse
- Cyber Security Analyst - Mata Elang Team

Project Overview!

1. Real-time Game Voucher Aggregation with Spark and Kafka
2. Kafka rolling aggregation with ksqlDB for Financial Transaction
3. Using PySpark for Customer Flight Activity Analysis
4. Spark Transform & Analysis E-commerce with Airflow
5. Utilizing Dynamic Airflow to ETL E-commerce data
6. Simulated Streaming Stocks Data to S3 with Kafka and Crawled by AWS Glue
7. Building Star Scheme with DBT on Big query for Insight Analysis E-commerce Data
8. Scraping Finvizz to Gain Stocks data with BeautifulSoup

Project Background

AIRBNB IS ONE OF THE WORLD'S LARGEST ONLINE PLATFORMS FOR BOOKING ACCOMMODATIONS, CONNECTING PROPERTY OWNERS WITH GUESTS FROM ALL OVER THE GLOBE.

AMSTERDAM, AS ONE OF THE MOST POPULAR TOURIST DESTINATIONS IN EUROPE, HAS THOUSANDS OF AIRBNB LISTINGS OFFERING A WIDE RANGE OF ACCOMMODATIONS, FROM APARTMENTS TO LUXURY HOMES.

Problem Statement

AS THE AMOUNT OF DATA CONTINUES TO GROW, COMPANIES LIKE AIRBNB NEED TO BUILD EFFECTIVE SYSTEMS TO MANAGE, ANALYZE, AND DERIVE INSIGHTS FROM THIS LARGE DATASET. UNSTRUCTURED DATA, OFTEN SPREAD ACROSS DIFFERENT SOURCES, CAN BE CHALLENGING TO PROCESS AND ANALYZE EFFICIENTLY.

DATA SOURCES & INGESTION

- AIR BNB DIRECT CSV INGESTION WITH KAFKA (SIMULATED STREAMING)

DATA STORAGE LAYER

- GOOGLE CLOUD STORAGE FOR DATA LAKE
- GOOGLE BIG QUERY FOR DATA WAREHOUSE

DATA TRANSFORMATION LAYER

- DATAPROC WITH APACHE SPARK FOR PROCESSING DATA

DATA ORCHESTRATION LAYER

- APACHE AIRFLOW

DATA VISUALIZATION LAYER

- POWER BI

DATA MONITORING

- PROMETHEUS + GRAFANA

Data Platform Understanding

Data Understanding

calendars

listing_id	integer
date	date
available	boolean
price	float

listings_details

id	integer
listing_url	varchar
scrape_id	integer
last_scraped	date
host_name	varchar
name	varchar
summary	varchar
space	varchar
description	varchar
experiences_offered	varchar
neighbourhood_overview	varchar

listings

id	integer
name	varchar
host_id	integer
host_name	varchar
neighbourhood_group	integer
neighbourhood	varchar
latitude	float
longitude	float
room_type	varchar
price	float

reviews_details

listing_id	integer
id	integer
date	date
reviewer_id	integer
reviewer_name	varchar
comments	varchar

reviews

listing_id	integer
date	date

neighbourhoods

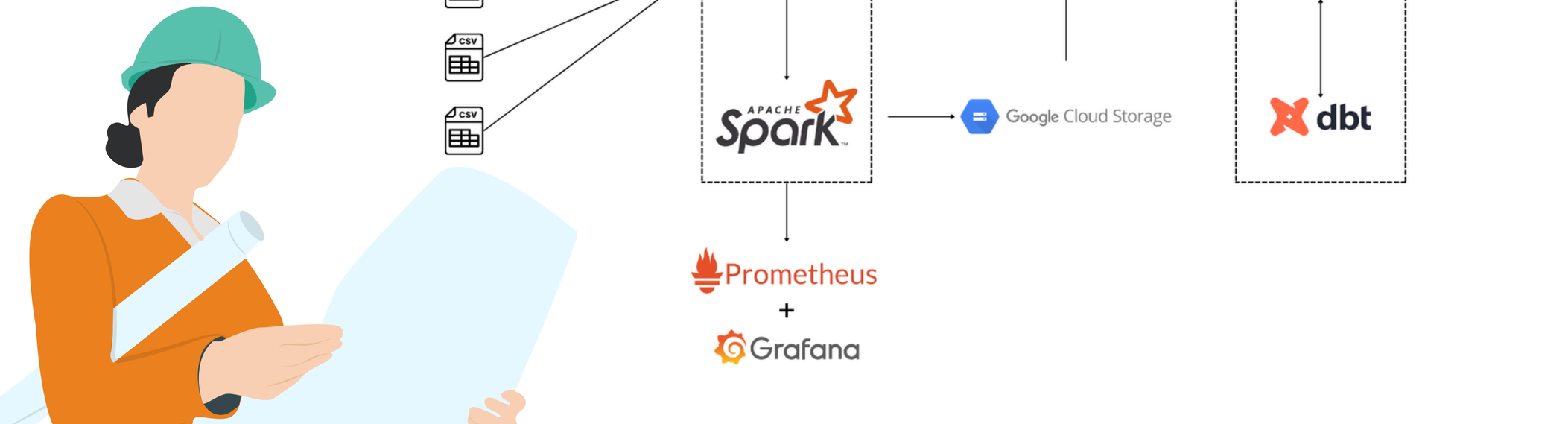
neighbourhood_group	integer
neighbourhood	varchar

Data Understanding

THE 'LISTINGS' FILE CONTAINS ALL THE ADVERTISEMENTS IN AMSTERDAM ON DECEMBER 6TH, 2018 UNTIL 2022 (20K). THE LISTINGS_DETAILS FILE CONTAINS ADDITIONAL VARIABLES. THE CALENDAR HAS 365 RECORDS FOR EACH LISTING. IT SPECIFIES THE WHETHER THE LISTING IS AVAILABLE ON A PARTICULAR DAY (365 DAYS AHEAD), AND THE PRICE ON THAT DAY.

THERE IS ONE DATA THAT HAS GEOJSON TYPE WHICH CONTAINS GEOSPATIAL DATA RELATED TO THE NEIGHBORHOODS IN AMSTERDAM. IT IS TYPICALLY USED TO REPRESENT GEOGRAPHICAL BOUNDARIES, WHICH CAN BE EXTREMELY USEFUL IN VARIOUS TYPES OF ANALYSES AND VISUALIZATIONS.

Transformation and Consideration



WIP AND THANK YOU!

Contact
Mochamad Reza Rahadi
morezarahadi@gmail.com

ON GOING