Final Project Bootcamp
Data Engineering

Geospatial Analysis NYC Yellow Taxi Trip

Introduction

- Self-overview
  - Passionate and experienced backend developer with a robust background in Golang, Node.js, TS, and PHP. Interested in data science especially data engineering.
- Education
  - Undergraduate, Information System Brawijaya University
- Working
  - GovTech Edu, Belajar.id and ANBK

Overview Project

1. PySpark
    a. Using pyspark to cleaning, transform, analyze, and viz data.
2. Batch Processing
    a. Using pyspark for batch processing. ETL with pyspark, load it to Postgres, and using airflow for as orchestration tools.
3. Kafka
    a. Create random data with faker for kafka producer, then consume it with kafka consumer.
4. Streaming
    a. Create random data with faker for kafka producer, then consume and ETL data using apache spark streaming.

Main Project

Project Background

This project analyzes the NYC Yellow Cab trip data, a publicly available dataset provided by the New York City Taxi and Limousine Commission (TLC). This dataset contains detailed information about millions of individual taxi trips within New York City, including:

Trip timestamps: Pickup and dropoff dates and times. Location data: Pickup and dropoff locations, often specified by taxi zones. Trip distance: The distance covered during the trip. Fare information: Fare amount, payment type, and any additional charges. Passenger count: The number of passengers on the trip.

This rich dataset offers valuable insights into urban mobility patterns, traffic dynamics, and passenger behavior within a major metropolitan area. By applying data analysis and visualization techniques, we can extract meaningful information from this raw data to address specific questions and challenges faced by taxi services and city planners.

Problem Statement

Taxi services need to optimize their operations and improve customer satisfaction by understanding passenger demand and accessibility patterns. This requires the ability to:

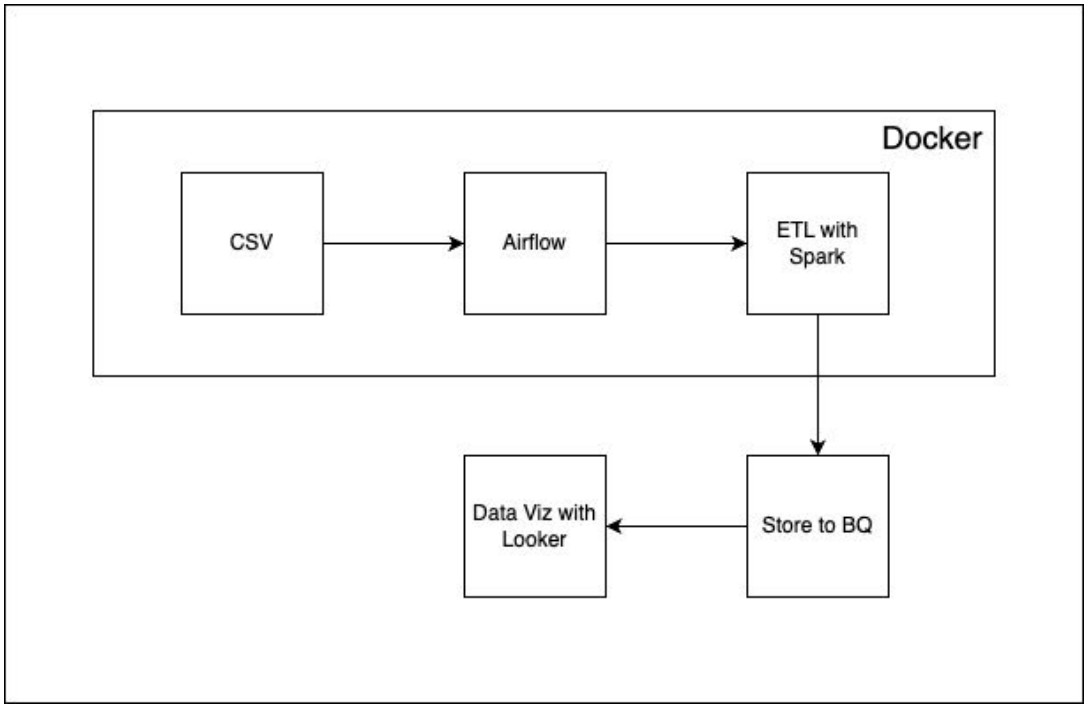
- Identify high-demand areas: Where are the most frequent pickup and dropoff locations? Are there any spatial patterns or hotspots that emerge?
- Assess location accessibility: How easy is it to reach various locations by taxi, considering travel time and distance?

Data Understanding

Data source and structure can be accessed here : Trip Data NYC Yellow Cab

Data cleaning included in ETL processes, cleaning data using dropna for column "tpep_pickup_datetime", "tpep_dropoff_datetime", "pickup_longitude", "pickup_latitude".

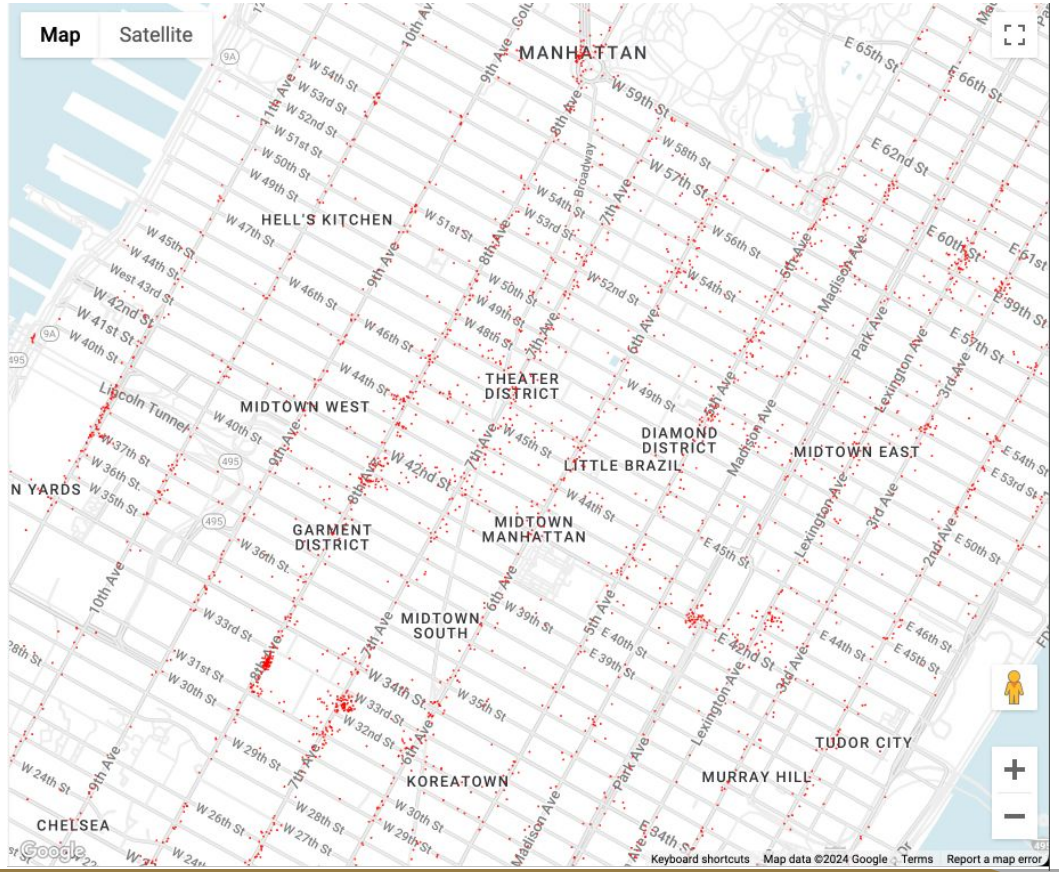
Transformation and Consideration



Conclusion

Platform can do ETL batch processing with the architecture given before, it can process 3 million rows data.It has flaws, like it cannot process data more than 4 million rows, due to insufficient memories.

Pickup heatmap at midtown after office hour on weekend



Pickup heatmap at midtown after office hour on weekday