

# Behavioral Analysis of Financial Transaction

Muhammad Khairunnas





# Muhammad Khairunnas

## Education

- UIN SUSKA RIAU (2017–2021)
- dibimbing.id (2024– sekarang)

## Working

- PT. Lintas Negara Express (2023–sekarang)
- CIDSCI (2021–2022)

## Overview Project

### ● Web Scraping Product Furniture

Scraping product using beautifulsoup and request GraphQL

### ● Data modeling Amazon Sale Report

Create dimension table and fact table using DBT from data set amazon sales

### ● ETL Process Using Airflow

Implementation airflow for ETL process with batch processing

### ● Data Analysis Using Airflow and Pyspark

Transformation data and analysis pyspark and airflow as ETL


On the left side of the slide, there are three overlapping geometric shapes: a large black parallelogram at the top, a medium-sized light orange parallelogram in the middle, and a smaller dark orange parallelogram at the bottom. All shapes are tilted to the right.

# Project Background

Proyek ini bertujuan untuk merancang pipeline ETL (Extract, Transform, Load) untuk menganalisis perilaku pelanggan saat menggunakan kartu untuk transaksi. Tujuan proyek ini adalah menarik pelanggan untuk sering melakukan transaksi dengan memberikan diskon atau promo kepada pelanggan yang jarang bertransaksi di merchant yang menjadi mitra layanan kartu.

Proyek ini penting karena:

1. Penyedia layanan kartu dapat meningkatkan jumlah transaksi saat pelanggan berbelanja di merchant mitra.
2. Pelanggan mendapatkan keuntungan berupa promo atau diskon.
3. Merchant memperoleh lebih banyak pelanggan melalui promosi tersebut.

A large black parallelogram is positioned on the left side of the slide. Overlapping its bottom edge are two smaller parallelograms: a light orange one in front and a darker orange one behind it, both tilted at the same angle.

# Problem Statement

Masalah teknis yang akan dihadapi oleh project ini adalah:

1. Bagaimana cara melakukan ekstraksi dari berbagai sumber data kedalam satu tempat dimana datanya sudah siap di analisis
2. Bagaimana cara melakukan transformasi terhadap data yang memiliki jumlah besar dengan cepat
3. Bagaimana cara melakukan data modeling sehingga datanya bisa digunakan sesuai subject yang membutuhkan seperti tim marketing atau tim operasional

Tujuan teknis dari project ini adalah:

1. Memastikan data yang diekstrak menjadi data yang lebih terstruktur dan rapi
2. Memastikan data yang dibutuhkan sesuai dengan permintaan subject
3. Memastikan data tersimpan ke data warehouse dengan cepat sesuai dengan jadwal yang telah ditentukan



Masalah analisis yang akan dihadapi oleh project ini adalah:

1. Merchant di kota apa yang paling banyak dikunjungi setiap gender?
2. berapa rata-rata umur pengguna yang sering melakukan transaksi dengan kartu?
3. Berapa kartu yang belum menggunakan chip?
4. Kapan transaksi banyak terjadi?
5. Brand kartu apa yang masih sedikit penggunaanya?
6. Dimana kota yang paling sedikit penggunaan kartu untuk transaksi berdasarkan brand tertentu?

Tujuan analisis dari project ini adalah:

1. Menentukan kota merchant yang paling banyak dikunjungi setiap gender
2. Menentukan rata-rata umur yang sering melakukan transaksi dengan kartu
3. Menentukan jumlah kartu yang masih belum menggunakan chip
4. Menentukan bulan yang sering terjadi transaksi
5. Menentukan Brand kartu apa yang masih sedikit penggunaanya
6. Menentukan kota yang paling sedikit penggunaan kartu untuk transaksi berdasarkan brand tertentu

Matriks yang akan menjadi acuan dalam pengukuran ini adalah matriks summary. Matriks summary diambil berdasarkan transaksi setiap bulannya. dimana promo akan diberikan kepada merchant, dan client berdasarkan umur, dan jenis kelamin. Agar promosi bisa diberikan di bulan depannya.

A large black parallelogram is positioned on the left side of the slide. Below it, two overlapping parallelograms in shades of orange and yellow are also positioned on the left, creating a layered, geometric effect.

# Data Platform Understanding

### **Pipeline data**

Pipeline merupakan konsep yang menggambarkan alur dari aliran data dari awal hingga akhir. Pipeline yang digunakan adalah batch processing, dengan proses ETL.

### **Data Resource**

Data yang digunakan adalah data di kaggle sebelumnya dimana skenario yang dibuat disesuaikan dengan platform berikut:

- FastAPI merupakan platform framework REST API yang digunakan untuk membuat API sederhana. Untuk penerapannya diskenario yang dibangun adalah data yang akan ditampilkan di platform FastAPI ini adalah data transactions\_data.csv yang difilter berdasarkan tanggal sehingga hasil kembaliannya struktur JSON.
- PostgreSQL merupakan platform database yang berbasis RDBMS yang mampu, dan handle penyimpanan data dalam jumlah besar. Database ini biasanya digunakan untuk OLTP proses karena kecepatan dalam melakukan manajemen data lebih baik. Pada skenario penerapannya adalah data cards\_data.csv dan users\_data.csv yang dari kaggle akan di ingestion dari PostgreSQL berdasarkan filter tanggal menggunakan SQL

## **Orchestration**

Platform yang digunakan untuk mengatur penjadwalan pada kasus ini adalah airflow. Airflow merupakan salah satu platform penjadwalan yang berfungsi menjalankan berbagai tugas yang saling terhubung dan bergantung dengan menggunakan operator yang menggunakan konsep DAGs yang berarti alur tugas yang saling terhubung dieksekusi secara sekuensial yaitu tugas yang sudah selesai tidak kan dijalankan ulang kembali.

## **Storage**

Platform storage (penyimpanan) yang digunakan ada 2 yaitu:

- minio adalah platform yang dikembangkan oleh Amazon Web Service digunakan untuk menyimpan file atau data dalam berbagai format. Penerapan pada skenario ini adalah minio akan dijadikan tempat penyimpanan file, atau format lain sebagai staging area dari proses ingestion data tujuan dari staging data ini adalah jika ada proses gagal di tengah jalan dan ingin menjalankan ulang data yang bisa langsung diambil adalah dari minio tanpa perlu melakukan ingestion ke API atau ke postgresSQL.
- postgresSQL adalah platform RDBMS yang dirancang untuk handle jumlah data yang besar platform ini akan dijadikan tempat penyimpanan data warehouse

### **Transformation**

Platform transform yang digunakan adalah spark menggunakan library pyspark dari python sebagai jembatan untuk menghubungkan service dari spark. Spark adalah salah satu platform dari apache yang digunakan untuk transformasi data yang terdistribusi memanfaatkan beberapa atau seluruh node pada spark cluster (selain spark manager) untuk pemrosesan data secara paralel yang mempercepat proses transform data.

### **Visualization**

Platform visualization yang digunakan adalah Power BI yang merupakan platform analisis yang tersedia beberapa koneksi dengan beberapa platform data seperti bigQuery, database, csv, dan lain-lain.

A large black parallelogram is positioned on the left side of the slide. Below it, two overlapping parallelograms in shades of orange and yellow are also positioned on the left, creating a layered, geometric effect.

# Data Understanding



Dataset yang digunakan adalah dataset dari kaggle [Financial Transactions Dataset: Analytics](#), data yang diambil sebanyak 3 data masing-masing adalah:

**cards\_data.csv (6146 record)**

merupakan data yang berisi informasi dari kartu yang digunakan untuk transaksi termasuk users/clients yang menggunakan kartu tersebut.

terdiri dari kolom:

- id : type biginteger, sebagai primary key
- client\_id big integer, sebagai foreign key yang terhubung dengan data users\_data.csv
- card\_brand character, berisi brand kartu yang digunakan dengan 3 jenis yaitu mastercard, visa, other
- card\_type character, berisi tipe kartu debit, kredit, dan other
- card\_number integer, berisi nomor kartu
- expires date, berisi tanggal berlakunya kartu yang digunakan
- cvv integer, berisi nomor cvv yang berada di kartu baik debit, maupun kartu kredit
- has\_chip boolean, tanda apakah kartu yang digunakan sudah menggunakan chip atau tidak
- num\_cards\_issued integer, menandakan edisi atau penerbitan kartu yang digunakan
- credit\_limit biginteger, limit yang dimiliki oleh kartu yang digunakan

### **users\_data.csv (2000 record)**

merupakan data yang berisi informasi dari pengguna kartu sebagai pemilik kartu.  
terdiri dari kolom:

- id : type biginteger, sebagai primary key
- current\_age integer, umur dari pengguna kartu
- retirement\_age integer, umur pensiun dari pengguna kartu
- birth\_year integer, tahun lahir dari pengguna
- birth\_month integer, bulan lahir dari pengguna
- gender character, berisi gender dari pengguna kartu laki-laki dan perempuan
- address character, alamat dari pengguna kartu
- latitude float, titik koordinat latitude pengguna kartu
- longitude float, titik koordinat longitude pengguna kartu
- per\_capita\_income integer, pendapatan pengguna per kapita

### **transactions\_data.csv (13.3 juta record)**

merupakan data yang berisi informasi transaksi kartu dengan merchant yang dilakukan oleh pengguna.

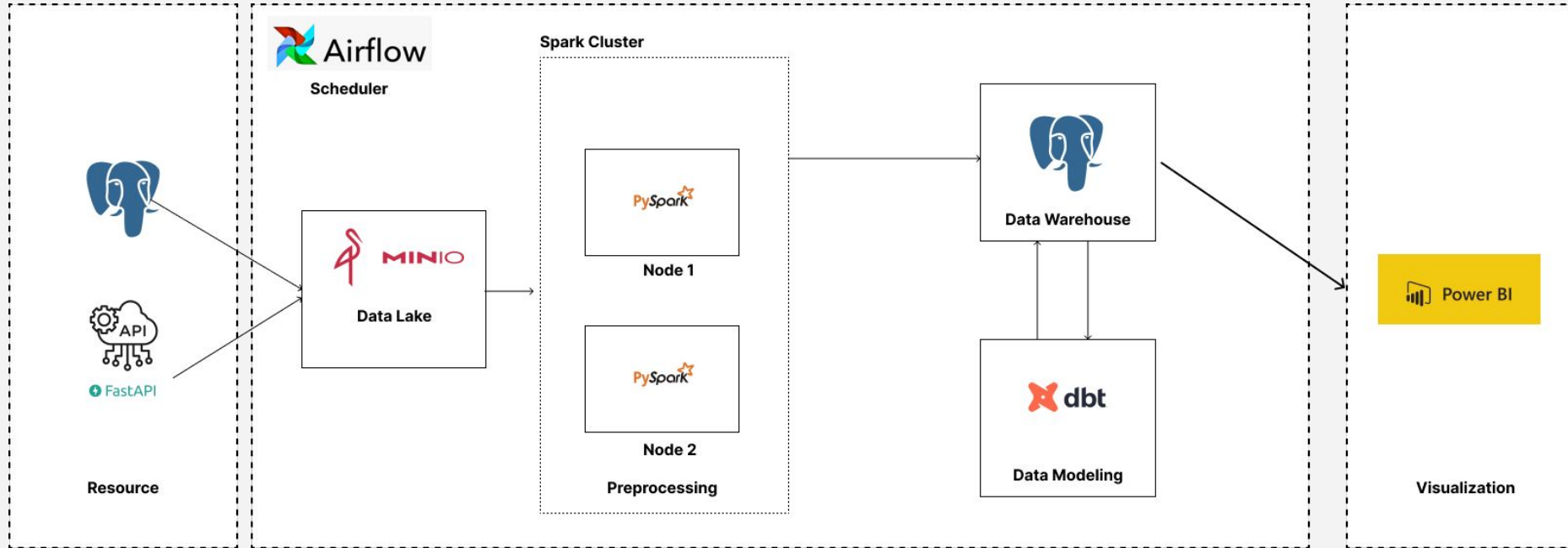
terdiri dari kolom:

- id : typebig integer, sebagai primary key
- date date: waktu pengguna melakukan transaksi
- client\_id biginteger, sebagai foreign key yang terhubung dengan data users\_data.csv
- card\_id biginteger, sebagai foreign key yang terhubung dengan data cards\_data.csv
- amount biginteger, jumlah transaksi yang dilakukan di merchant
- merchant\_id biginteger, merchant tempat terjadinya transaksi
- merchant\_city character, kota merchant yang terjadi transaksi
- merchant\_state character, kode negara bagian yang berada di wilayah Amerika Serikat
- zip float, berisi kode pos dari merchant

Data dikumpulkan dengan menggunakan API dan dari database PostgreSQL. Untuk kualitas data beberapa data di **transactions\_data.csv** ada nilai null, beberapa tipe data dari data yang dikumpulkan tidak sesuai seperti kodepos yang menggunakan tipe float.

A large black parallelogram is positioned on the left side of the slide. Overlapping its bottom edge are two orange parallelograms, one in a lighter shade and one in a darker shade, both pointing towards the right.

# Transformation & Consideration



Arsitektur ETL yang digunakan adalah batch processing

**Transformation:** spark dengan 2 node di dalam 1 cluster

**Orchestration:** airflow dengan dynamic task dengan multi processing

**Staging Area:** minio dengan menyimpan data berbentuk parquet

**Data Warehouse:** postgresSQL sebagai untuk data yang sudah ditransformasi

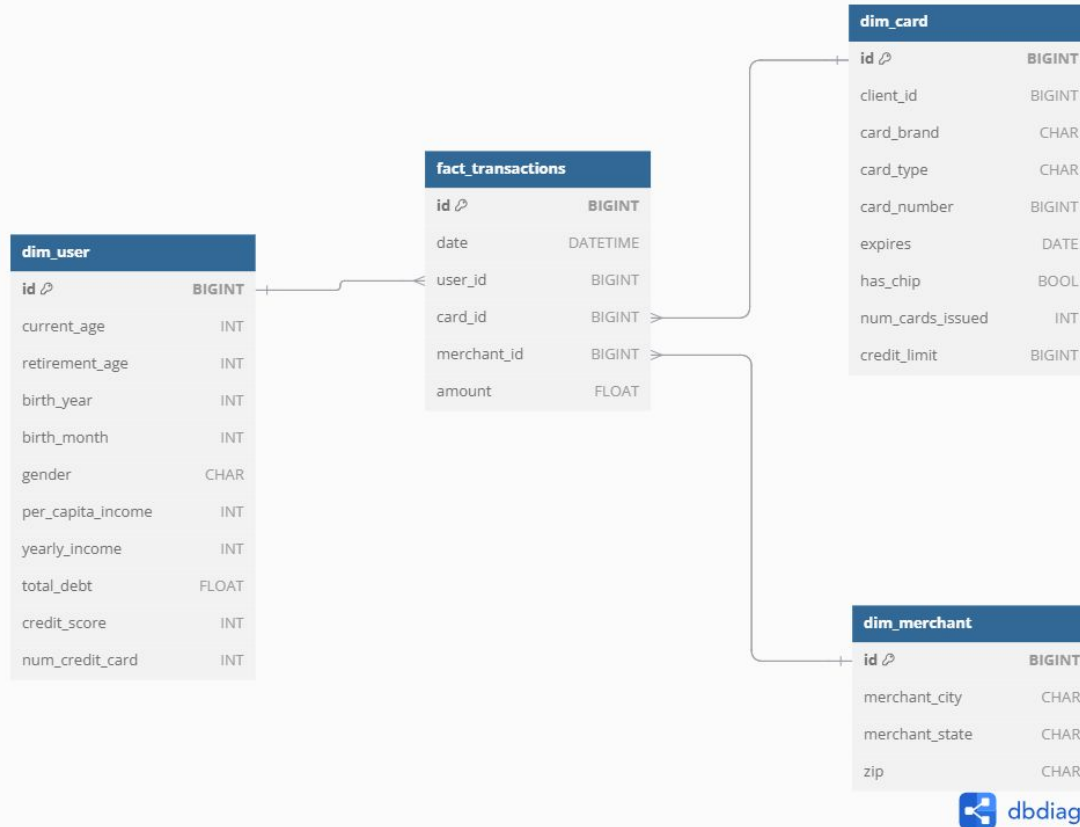
**Data Modeling:** dbt membuat fact table, dan dimensional table

**Data Visualization:** power BI visualisasi dari hasil tabel yang sudah di modeling

Decorative geometric shapes on the left side of the slide: a large black parallelogram, a light orange parallelogram, and a darker orange parallelogram, all slanted to the right.

# Data Modeling (Business)





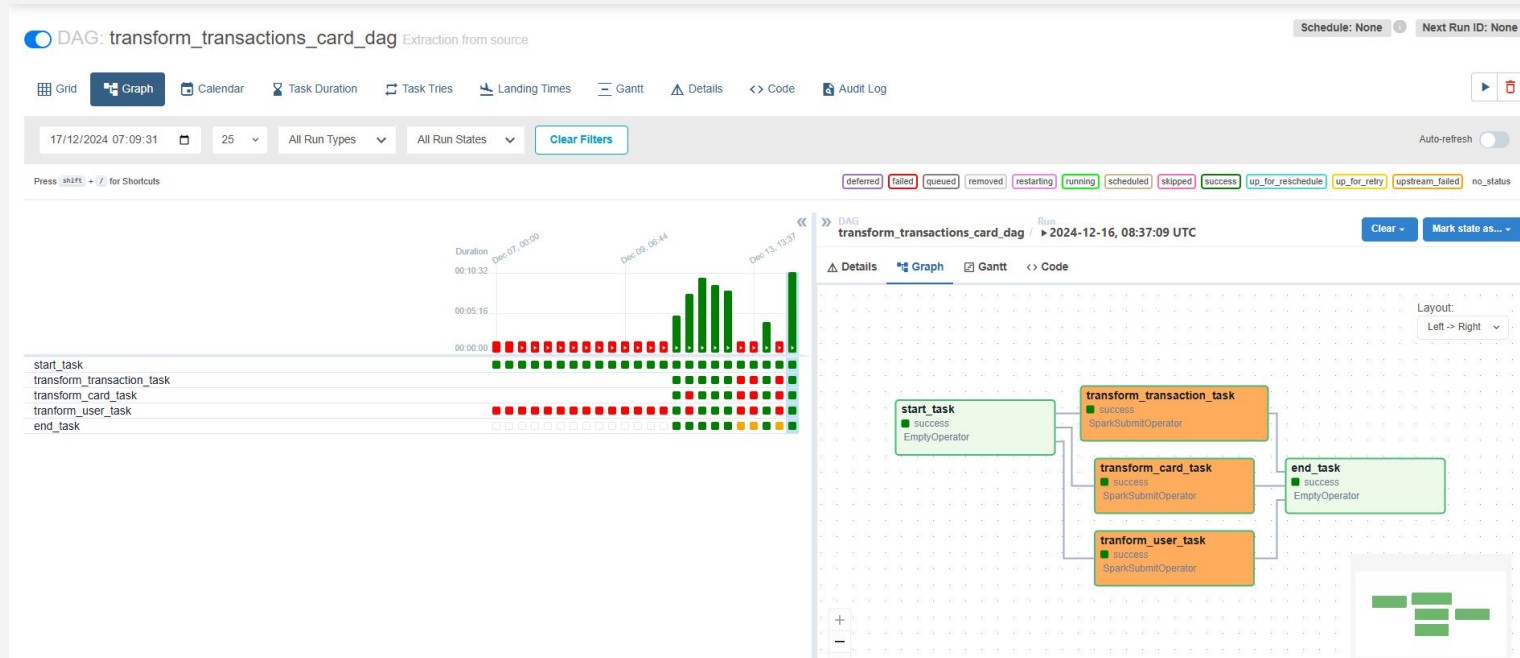
Data modeling yang digunakan adalah versi kimball model dengan star schema

A large black parallelogram on the left side of the slide, with two overlapping orange parallelograms positioned below it, creating a modern, abstract geometric design.

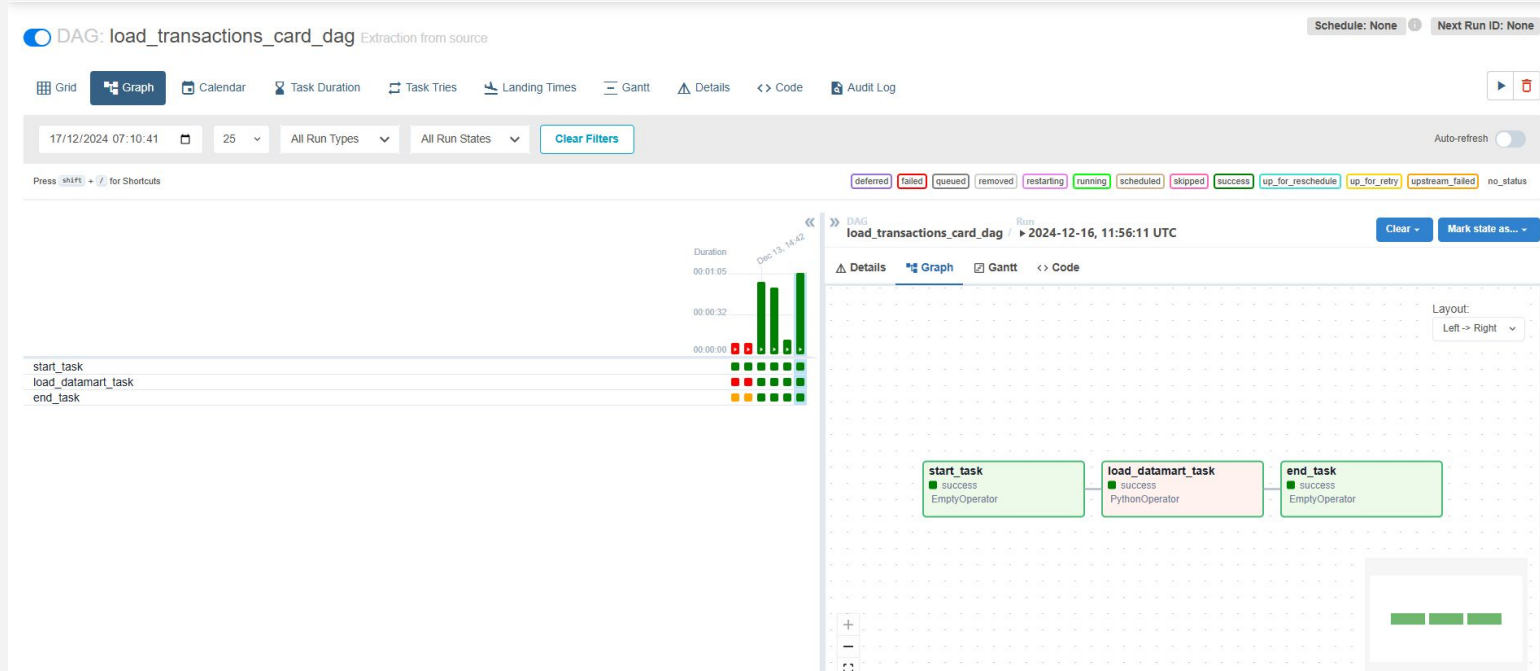
# Conclusion & Recommendation



# Transform Data dengan spark



# Load Data ke postgresQL sebagai data warehouse



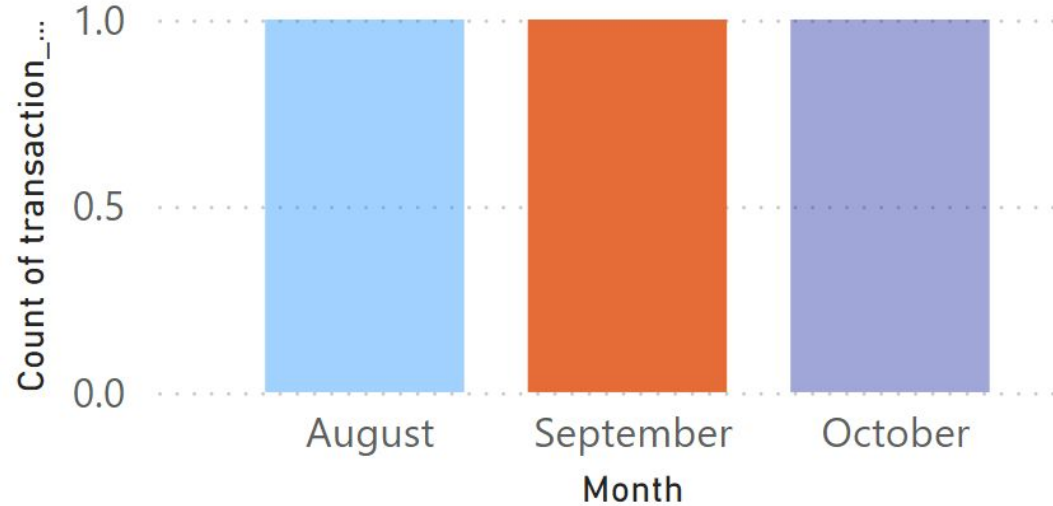
## Conclusion and Recommendation for architecture

- Kesimpulan
  - Sistem mampu mengolah data yang terdiri dari 13++ juta baris data
  - Sistem ETL mampu melakukan proses ingestion data hingga tahap data modeling dengan DBT
  - Ada beberapa kendala terutama penyesuaian versi spark airflow dan spark di cluster
  - Data storage mampu menyimpan data parquet yang sudah dikelompokkan berdasarkan bulan-tahun
- Keterbatasan
  - Beberapa proses dijalankan masih dengan manual terutama ketika membuat visualisasi dengan power BI
  - Proses berjalan lumayan lama karena proses data yang cukup banyak
- Rekomendasi
  - Menyimpan data setelah ditransformasi ke platform data warehouse sebenarnya
  - Menerapkan monitoring untuk sistem yang sedang berjalan
  - Menerapkan metadata, dan data lineage untuk implementasi data governance

# Result Analysis

Average of transaction count by current age each 3 month ago

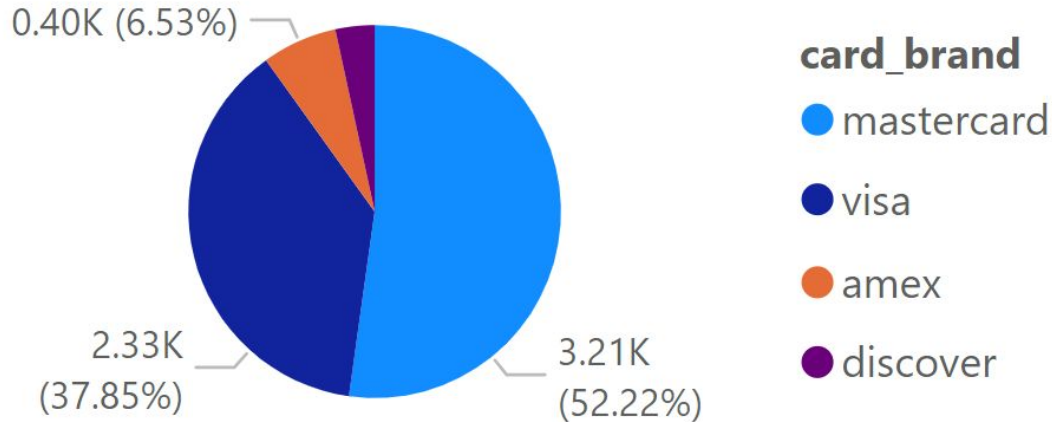
avg\_age ● 53.59 ● 53.66 ● 53.71



Rata-rata umur yang sering melakukan transaksi 3 bulan terakhir

untuk pengguna rata-rata sama

## Total card brands with the least number of users

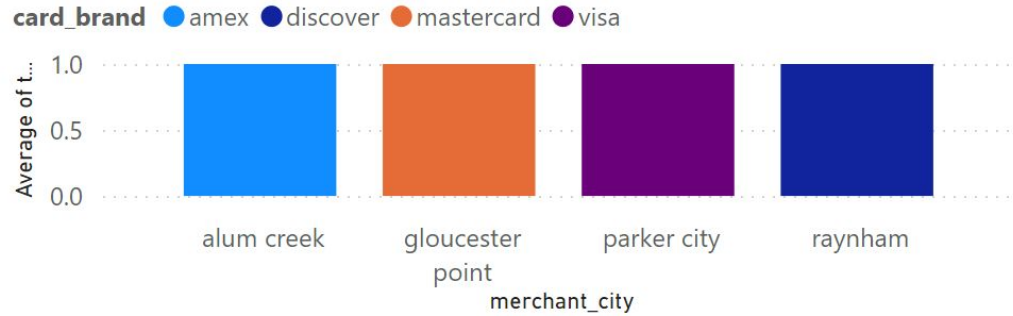


Total pengguna kartu sesuai dengan brandnya

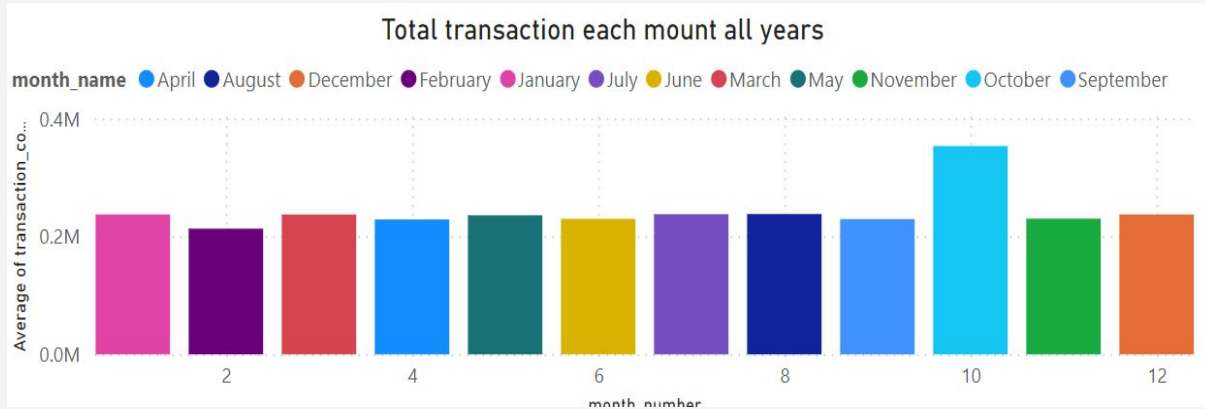
brand discover dan brand amex memiliki pengguna paling sedikit



Cities with the fewest number of card users for each card brand



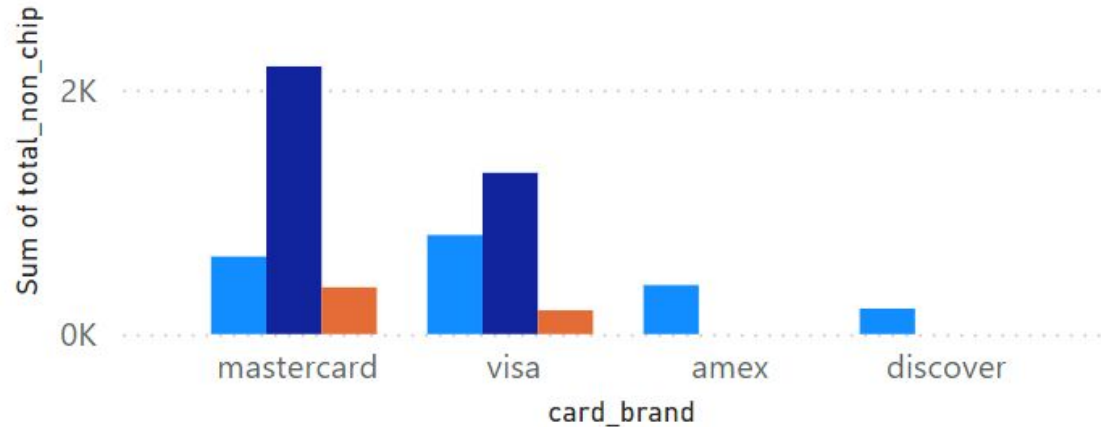
kota dengan jumlah transaksi pengguna paling sedikit untuk tiap brand kartu



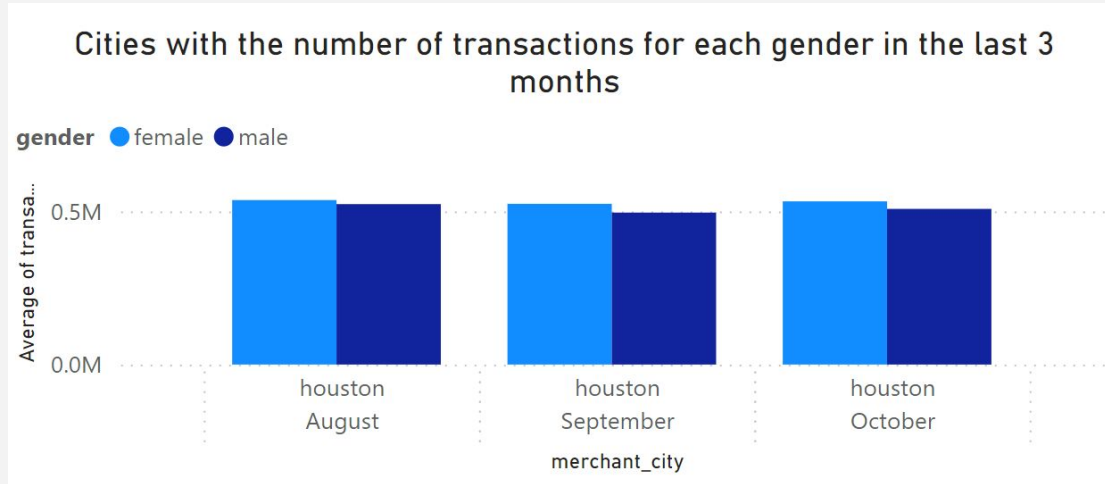
Total transaksi  
pengguna kartu setiap  
bulannya

## Total cards that do not use chips

card\_type ● credit ● debit ● debit (prepaid)



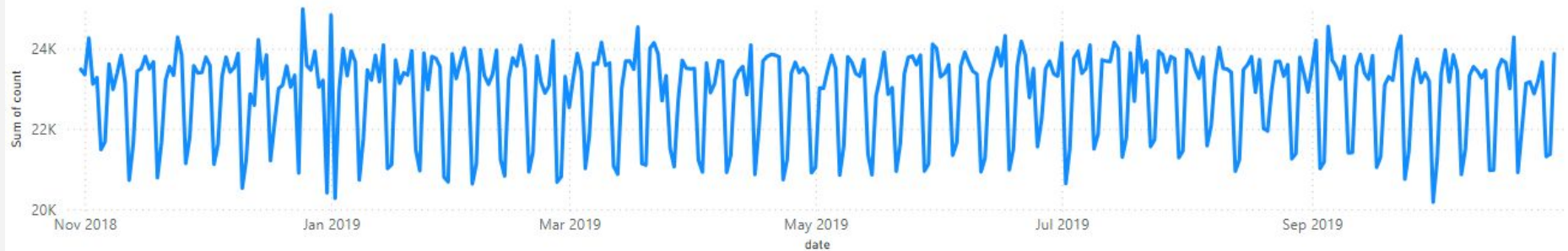
Total kartu yang masih belum menggunakan chip



Kota dengan jumlah transaksi terbanyak untuk 3 bulan terakhir berdasarkan jenis kelamin

Transaksi tertinggi terjadi di akhir tahun  
dari Desember 2018 – Januari 2019

Count transaction from 1 year ago by date



A large, stylized graphic on the left side of the slide. It consists of a blue outline of a person's head and shoulders. Inside the head is a series of concentric circles: a small light orange circle, a medium orange circle, and a large orange circle. The body is a large blue circle, and inside it is a large orange circle, which in turn contains a smaller orange circle.

**Terima  
Kasih.**