

FINAL PROJECT

● K E L O M P O K 1 5

December 2023

HELLO EVERYONE!



UNIVERSITAS
MIKROSKIL

MHD. FISABILUDDIN



SEKOLAH TINGGI ILMU
EKONOMI
BINA KARYA

EGIDIA N TAMBUNAN



UNIVERSITAS
NEGERI SURABAYA

RENITA FITRIANI



UNIVERSITAS
AMIKOM
PURWOKERTO

ZUHRIYATUL LUBNA



DISCUSSION

FINAL PROJECT PRESENTATION



01 Hello Friends!

02 Study Case

03 Technical Details

04 Tools & Reasons

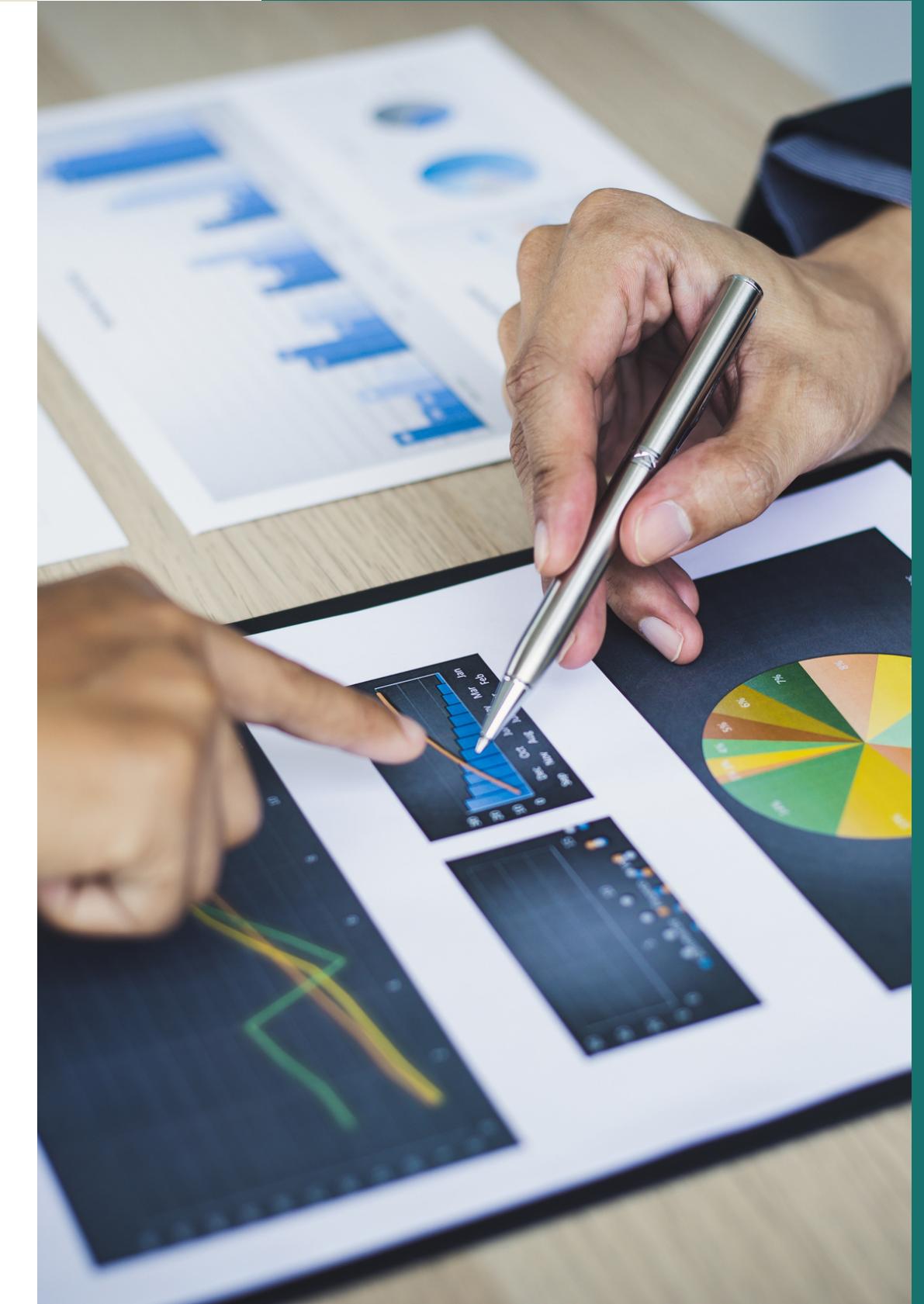
05 Entity Relationship
Diagram (ERD)

06 Technical Improvement
& Next Plan

STUDY CASE

LATAR BELAKANG

Sebagai seorang data engineer baru yang dipekerjakan di sebuah e-commerce di Amerika Serikat dimana dalam perusahaan tersebut menjual produk secara langsung kepada pelanggan dimana produk-produk tersebut berasal dari berbagai pemasok di seluruh dunia. Seorang data engineer tersebut diberikan tantangan untuk membangun infrastruktur data menggunakan data yang sebelumnya telah dibuat sama dengan data yang ada di perusahaan e-commerce.



TECHNICAL DETAILS

PENJELASAN WORKFLOW

1

Pemahaman Data

2

Desain DAGs untuk proses ETL di Airflow

3

Alur kerja DAGs

Pemahaman Data

Mulai dengan memahami isi data yang akan diolah, termasuk formatnya (JSON, CSV, Avro, Parquet, XLS/XLSX) dan sumber daya data yang berbeda.

Desain DAGs untuk proses ETL di Airflow

DAG yang telah kami bangun memiliki peran penting dalam mengelola siklus lengkap ekstraksi, transformasi, dan pemuatan data (ETL/ELT) ke dalam lingkungan PostgreSQL. Setiap fungsi ekstraksi didesain untuk menangani format data yang berbeda, mulai dari file JSON, CSV, Avro, Parquet, hingga XLS/XLSX. Masing-masing fungsi mengelola sumber daya data yang berbeda dan memprosesnya ke dalam skema tabel yang telah ditentukan dalam database PostgreSQL.

Alur kerja DAGs

Alur kerja DAG telah dirancang dengan kejelasan alur: misalnya, data coupons yang pertama diekstrak, diikuti dengan pemrosesan data customer, yang kemudian digunakan dalam proses pembuatan data orders, dan seterusnya. Ketergantungan antara tugas-tugas ini disusun secara hati-hati untuk memastikan bahwa setiap langkah memanfaatkan data yang telah diproses sebelumnya, menjaga integritas dan konsistensi data.

TECHNICAL DETAILS

PENJELASAN WORKFLOW

4 Penjadwalan Harian DAG

5 Fungsi Kunci dalam DAG

Penjadwalan harian DAG

Jadwal harian DAG memastikan bahwa proses ETL/ELT terjadi secara teratur, memperbarui dan memasukkan data terbaru ke dalam sistem untuk memenuhi kebutuhan analisis bisnis yang kontinu. Dengan memanfaatkan berbagai jenis data dari sumber yang berbeda, kemampuan untuk mengubah formatnya ke dalam struktur yang konsisten dalam database, dan menjalankan tugas-tugas secara otomatis, DAG yang dibuat menjadi inti dari keseluruhan proses analisis data yang efisien, memastikan ketersediaan data aktual dan terpercaya bagi pengambilan keputusan bisnis yang tepat waktu.

Fungsi Kunci dalam DAG

1. Beberapa fungsi kunci yang dibuat dalam DAG :

- Fungsi `load_data_to_postgres` mengelola pengiriman data dari Pandas DataFrame ke PostgreSQL. Fungsi ini membangun koneksi menggunakan SQLAlchemy dan memuat atau mengganti data dalam tabel yang sudah ada.
- fungsi `create_table_in_postgres` bertugas membuat tabel baru di PostgreSQL, menggunakan koneksi yang sama dengan fungsi sebelumnya.
- fungsi `extract_and_load_coupons_to_postgres` dan `extract_and_load_customers_to_postgres` dan lainnya.

TOOLS AND REASONS

PEMILIHAN TOOLS DAN ALASANNYA

Apache Airflow untuk Proses ETL

1. Skalabilitas dan Keterjadwalan: Airflow memungkinkan untuk membuat workflows yang kompleks dan bisa dijadwalkan. Ini sangat berguna dalam manajemen ETL yang melibatkan banyak sumber data dan proses yang berbeda.
2. Keterbukaan dan Ekstensibilitas: Airflow merupakan platform open-source yang memungkinkan penggunaannya dalam berbagai macam lingkungan dan integrasi dengan berbagai sumber data dan alat lain.
3. Monitoring dan Manajemen Error: Airflow memiliki fitur monitoring yang baik untuk melacak kinerja workflows dan mendeteksi masalah dalam proses ETL.

TOOLS AND REASONS

PEMILIHAN TOOLS DAN ALASANNYA

PostgreSQL untuk Modeling Data

1. Reliability dan Performa: Sebagai sistem manajemen basis data (DBMS) yang handal, PostgreSQL menawarkan performa tinggi dan stabilitas yang baik dalam menyimpan dan mengelola data.
2. Fleksibilitas: Dengan dukungan untuk berbagai jenis data dan dukungan untuk operasi kompleks, PostgreSQL cocok digunakan dalam proyek yang membutuhkan model data yang kompleks.
3. Skalabilitas: Dalam konteks analisis data yang melibatkan banyak entitas, PostgreSQL dapat dikembangkan dengan baik untuk menangani perubahan dan pertumbuhan yang berkelanjutan.

TOOLS AND REASONS

PEMILIHAN TOOLS DAN ALASANNYA

Metabase di Docker untuk Visualisasi

- 1.Kemudahan Penggunaan: Metabase adalah alat visualisasi data yang sederhana namun powerful. Menjalankannya di Docker membuat penggunaan dan pengelolaannya lebih mudah.
- 2.Interaktif dan Customizable: Metabase menawarkan visualisasi data yang interaktif dan bisa disesuaikan. Ini memungkinkan pengguna untuk mengeksplorasi data dengan lebih baik dan membuat laporan yang sesuai dengan kebutuhan.

TOOLS AND REASONS

PEMILIHAN LIBRARY DAN ALASANNYA



Pandas (v2.1.0)

Manipulasi data luas dari format CSV, JSON, dan Excel (XLS/XLSX).

Fastavro (v1.4.0)

mendukung ekstraksi data dari file Avro



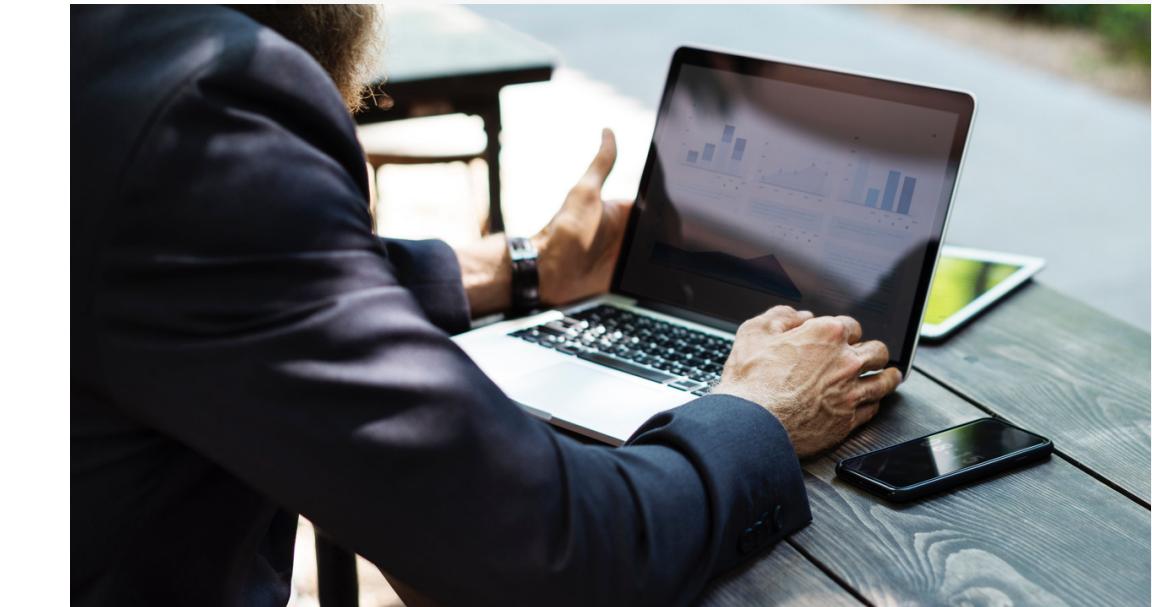
SQLAlchemy (v1.4.36)

berfungsi sebagai ORM (Object-Relational Mapping) yang kuat untuk berinteraksi dengan database PostgreSQL, memfasilitasi pembuatan tabel dan eksekusi query



TOOLS AND REASONS

PEMILIHAN LIBRARY DAN ALASANNYA



Psycopg2-binary (v2.9.9)

Sebagai driver database

Library xlrd (v2.0.1):

Membantu dalam membaca
file Excel (XLS/XLSX)

Python-snappy (v0.6.1)

Memberikan kemampuan
kompresi dan dekompresi data
Snappy yang mempercepat proses
manipulasi data dalam skala besar

EXECUTE DAGS IN AIRFLOW

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 03:05 UTC Log In

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

11/25/2023 03:03:41 AM 25 All Run Types All Run States Clear Filters Auto-refresh

Press `shift + /` for Shortcuts

deferred failed queued removed restarting running scheduled shutdown skipped success up_for_reschedule up_for_retry upstream_failed no_status

« » DAG extract_and_load_data_to_postgres

Duration: 00:01:37

Details Graph Gantt Code

DAG Runs Summary

Total Runs Displayed	2
Total success	1
Total running	1

First Run Start: 2023-11-25, 03:03:40 UTC

Last Run Start: 2023-11-25, 03:03:40 UTC

Max Run Duration: 00:01:37

extract_and_load_coupons_to_postgres

extract_and_load_customers_to_postgres

extract_and_load_login_attempts_to_postgres

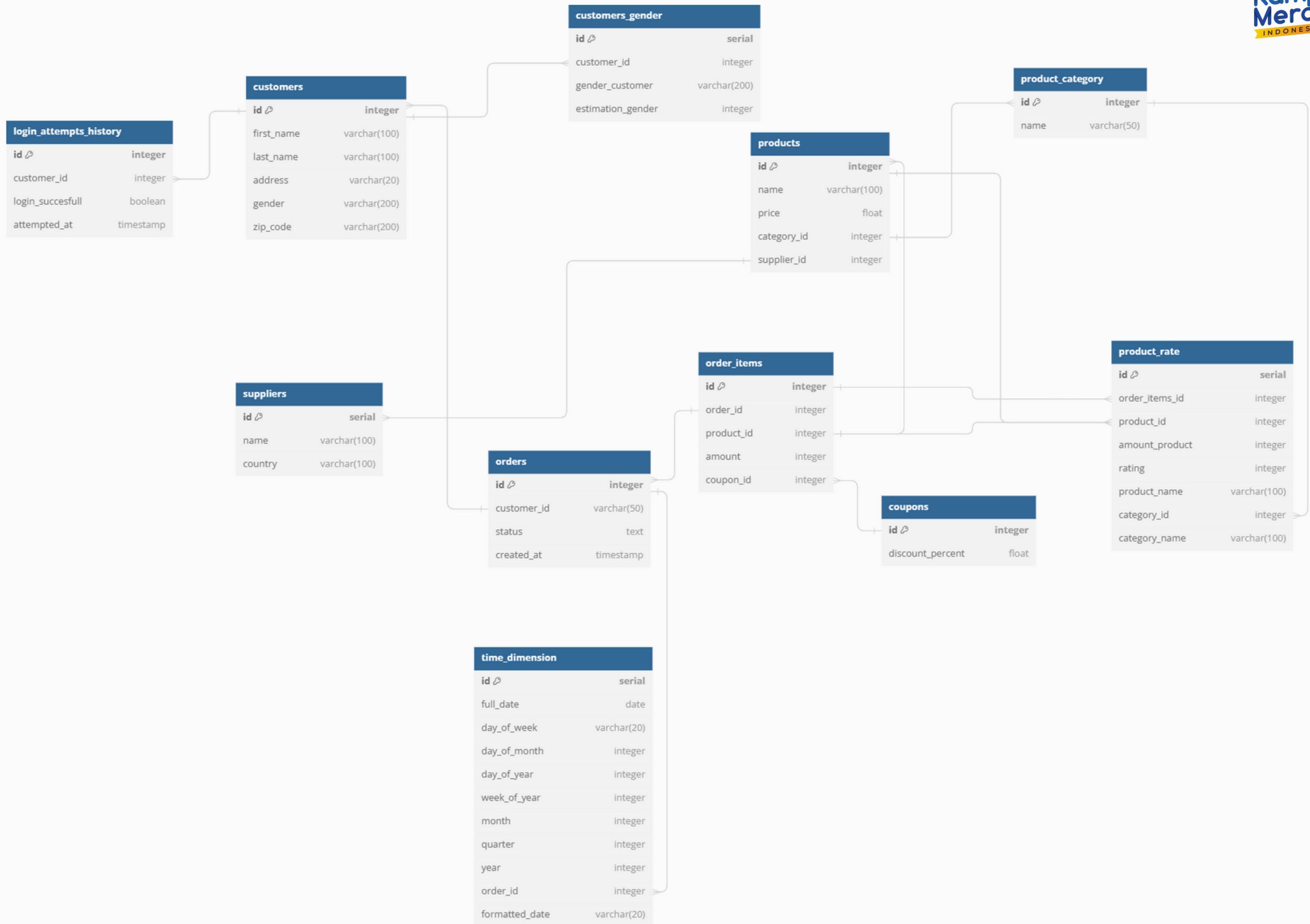
extract_and_load_orders_to_postgres

extract_and_load_product_category_to_postgres

extract_and_load_products_to_postgres

extract_and_load_suppliers_to_postgres

extract_and_load_order_items_to_postgres



ENTITY RELATIONSHIP DIAGRAM

time_dimension	
id ↗	serial
full_date	date
day_of_week	varchar(20)
day_of_month	integer
day_of_year	integer
week_of_year	integer
month	integer
quarter	integer
year	integer
order_id	integer
formatted_date	varchar(20)

product_rate	
id ↗	serial
order_items_id	integer
product_id	integer
amount_product	integer
rating	integer
product_name	varchar(100)
category_id	integer
category_name	varchar(100)

customers_gender	
id ↗	serial
customer_id	integer
gender_customer	varchar(200)
estimation_gender	integer

01

time_dimension

Entitas tambahan time_dimension berfungsi sebagai dimensi waktu untuk analisis mendalam

02

product_rate

Entitas tambahan product_rate berfungsi sebagai entitas yang memuat informasi rating produk untuk analisis kinerja produk

03

customers_gender

Entitas tambahan customers_gender berfokus pada data jenis kelamin yang berasal dari table customers dengan kemungkinan inklusi hasil prediksi atau estimasi

SQL :

https://drive.google.com/file/d/1DdQH1Shi8dkPPtwB_LPbVGP2BLy2S4Er/view?usp=drive_link

OUR DATA

OUR DATA > POSTGRES DW

 Learn about our data

 Coupons

 Customers

 Customers Gender

 Login Attempts History

 Order Items

 Orders

 Product Category

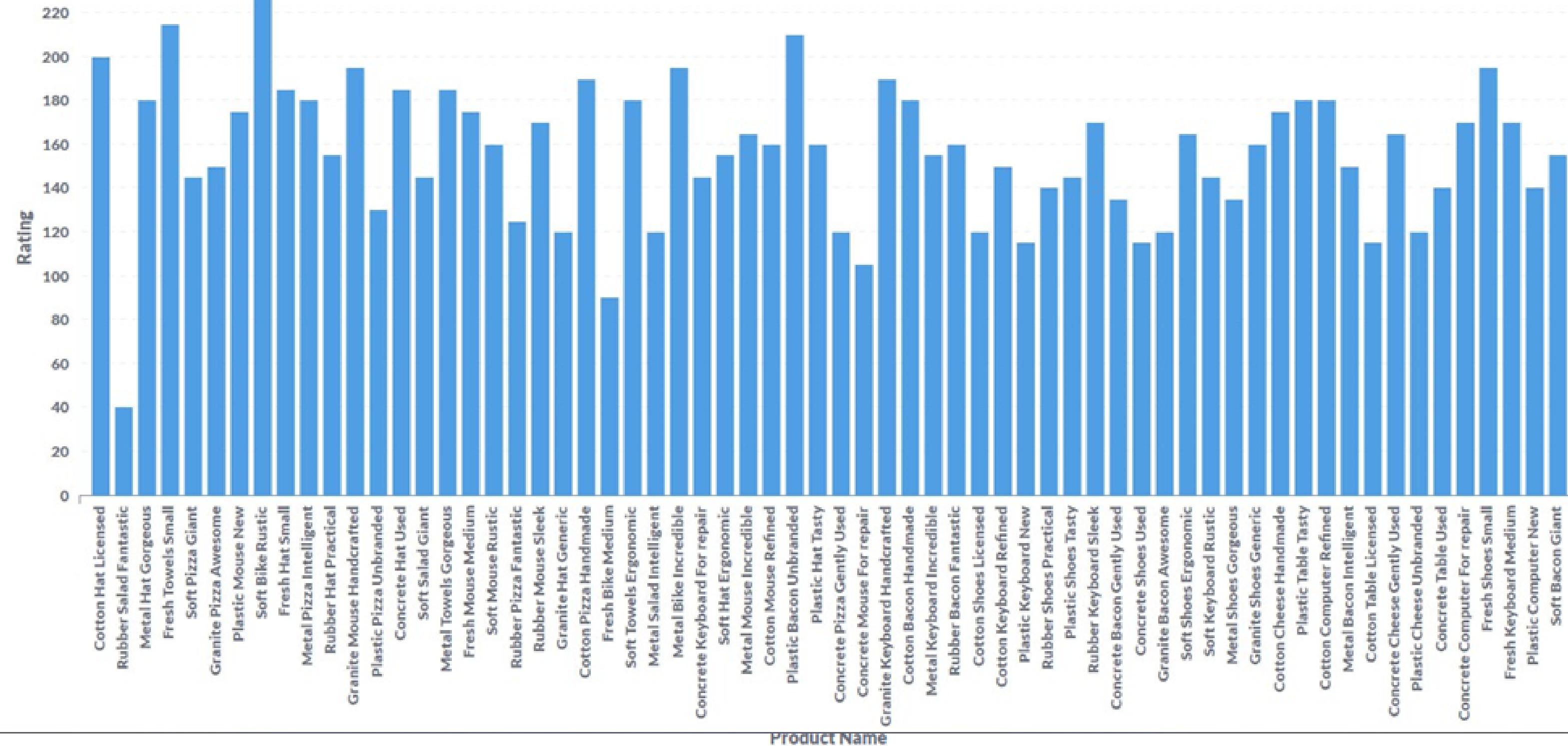
 Product Rate

 Products

 Suppliers

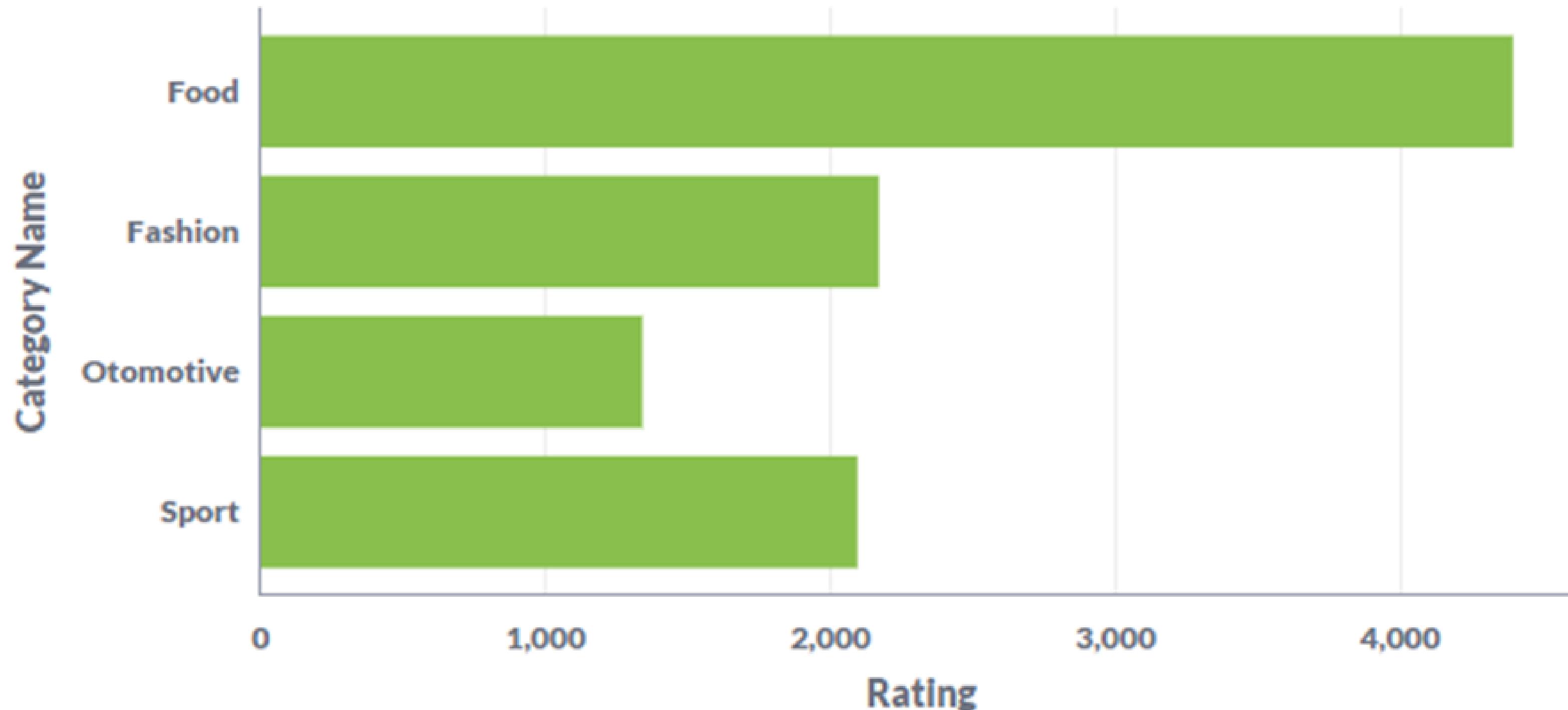
 Time Dimension

VISUALISASI IN METABASE



Visualisasi bar chart untuk mengetahui nama product dengan banyak pembelian dalam sekali pesan (rate)

Category product dengan jumlah rate pembelian terbanyak



Visualisasi row chart untuk mengetahui category dari product yang memiliki jumlah terbanyak pembelian juga jumlah terbanyak dalam sekali pembelian (rate)



Gender pelanggan



Visualisasi bar chart untuk mengetahui jumlah
majoritas gender dari pelanggan

TECHNICAL IMPROVEMENT & NEXT PLAN

- menambahkan library yang mendukung pembacaan file yang akan di extraksi.
- Membuat DAGs untuk proses ETL dalam 1 DAGs dengan penulisan code secara efisien dengan tujuan agar kinerja ETL dapat dimaksimalkan.
- Menambahkan parameter di DAGs tersebut dengan “catchup=false”. dalam kode DAGs Airflow dibuat khusus hanya menjalankan tugas yang terjadwal pada saat itu saja, tanpa menjalankan tugas yang seharusnya sudah berjalan sebelumnya tanpa terlewati.



- Mengubah aturan spinup pada file “makefile” menjadi “@docker-compose -f .docker/docker-compose.yml –env-file .env up -d” Opsi -d pada perintah docker-compose up membuat kontainer Docker berjalan di latar belakang/tidak melihatkan log pada terminal agar dapat digunakan untuk perintah lainnya tanpa terikat ke kontainer yang sedang berjalan

Dalam usaha untuk meningkatkan cara kerja dalam menganalisis data, ada beberapa hal penting yang perlu diperhatikan :

1. Perlu fokus pada pengelolaan alur kerja ETL menggunakan Airflow. Ini dapat membantu melacak kinerja dan menangani masalah dengan cepat. Evaluasi ulang langkah-langkah dalam alur kerja DAG dapat membantu mempercepat proses ETL secara keseluruhan.
2. Pada tahap pemodelan data menggunakan PostgreSQL, harus dipastikan sudah menerapkan indeks dengan benar pada kolom-kolom yang sering diakses. Hal ini akan meningkatkan kinerja query. Selain itu, perlu evaluasi struktur tabel untuk memastikan normalisasi atau denormalisasi yang sesuai, yang dapat meningkatkan efisiensi query dan penggunaan ruang penyimpanan data.

Dari pengerjaan yang sudah kita lakukan, kita sudah memiliki dasar yang kuat untuk analisis data yang baik. Namun, untuk kedepannya ada dua hal yang bisa ditingkatkan, yaitu :

1. Kita bisa memperluas integrasi dengan sumber data baru atau menggali lebih dalam ke dalam sumber data yang sudah ada. Ini akan memberikan kita informasi yang lebih kaya untuk analisis.
2. Meningkatkan cara dalam memvisualisasikan data dengan opsi interaktif yang lebih banyak. Hal ini akan membantu pemahaman dan pengambilan keputusan yang lebih baik. Dengan fokus pada pengembangan integrasi data yang lebih luas dan peningkatan visualisasi yang lebih kuat, kita dapat memberikan wawasan yang lebih baik bagi pengambilan keputusan bisnis di masa depan.



THANK YOU

● FOR YOUR NICE ATTENTION

December 2023