# BUILDING DATA PLATFORM FOR ONLINE RETAIL COMPANY

## FINAL PROJECT

# TEAM 6

Elsa Rahma Hidayani

Gabriel Khoirul Huga

Yara Rivayanti

Gadis Syahrani Elhakim

Indah Monica Dewi

# Latar Belakang

Topik ini melibatkan pengelolaan data secara efisien dalam suatu organisasi. Data pipeline memainkan peran kunci dalam mentransfer, mengolah, dan menyimpan data dari berbagai sumber. Entity-Relationship Diagrams (ERD) membantu memodelkan hubungan antara entitas dalam basis data.

Tools ETL (Extract, Transform, Load) digunakan untuk mengambil, mengubah, dan memuat data ke dalam sistem. Ini penting untuk memastikan konsistensi dan integritas data. Data modeling melibatkan perancangan struktur data untuk memahami dan mengorganisir informasi. Sementara itu, data visualization memanfaatkan grafik dan visualisasi untuk menyajikan data dengan cara yang mudah dimengerti.

Pengintegrasian dan penerapan efektif dari elemen-elemen ini mendukung pengambilan keputusan yang lebih baik dan membantu organisasi mengoptimalkan penggunaan data mereka.

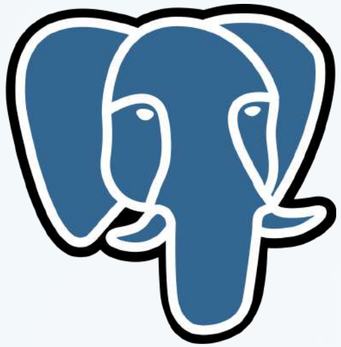
# Tujuan

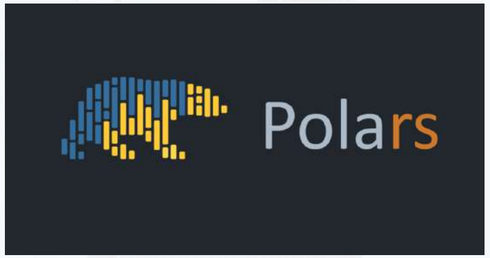
Proyek ini dibentuk untuk merangkum, menyimpan, dan menganalisis data dari berbagai sumber guna memberikan wawasan mendalam tentang operasi bisnis. Dalam konteks ini, kami akan:

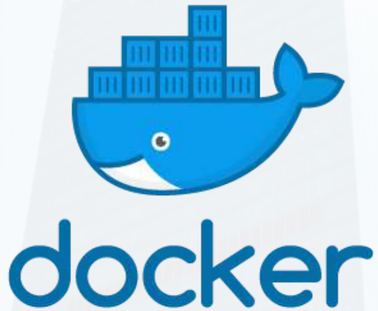
- Mengidentifikasi beragam segmen pelanggan berdasarkan perilaku pembelian mereka.
- Mengevaluasi kinerja produk dengan memperhatikan aspek penjualan dan popularitasnya.
- Memantau serta menganalisis aktivitas login pelanggan untuk pemahaman yang lebih baik.
- Memantau proses pemenuhan pesanan dan melacak status pesanan dengan cermat.
- Menilai efektivitas kampanye kupon untuk meningkatkan strategi pemasaran.
- Mengevaluasi kinerja pemasok dalam hal kualitas produk dan ketepatan pengiriman.

Dengan pendekatan ini, diharapkan proyek ini dapat dilaksanakan secara efektif dan efisien, memberikan kontribusi maksimal terhadap pemahaman mendalam tentang berbagai aspek operasional bisnis yang sedang berlangsung.

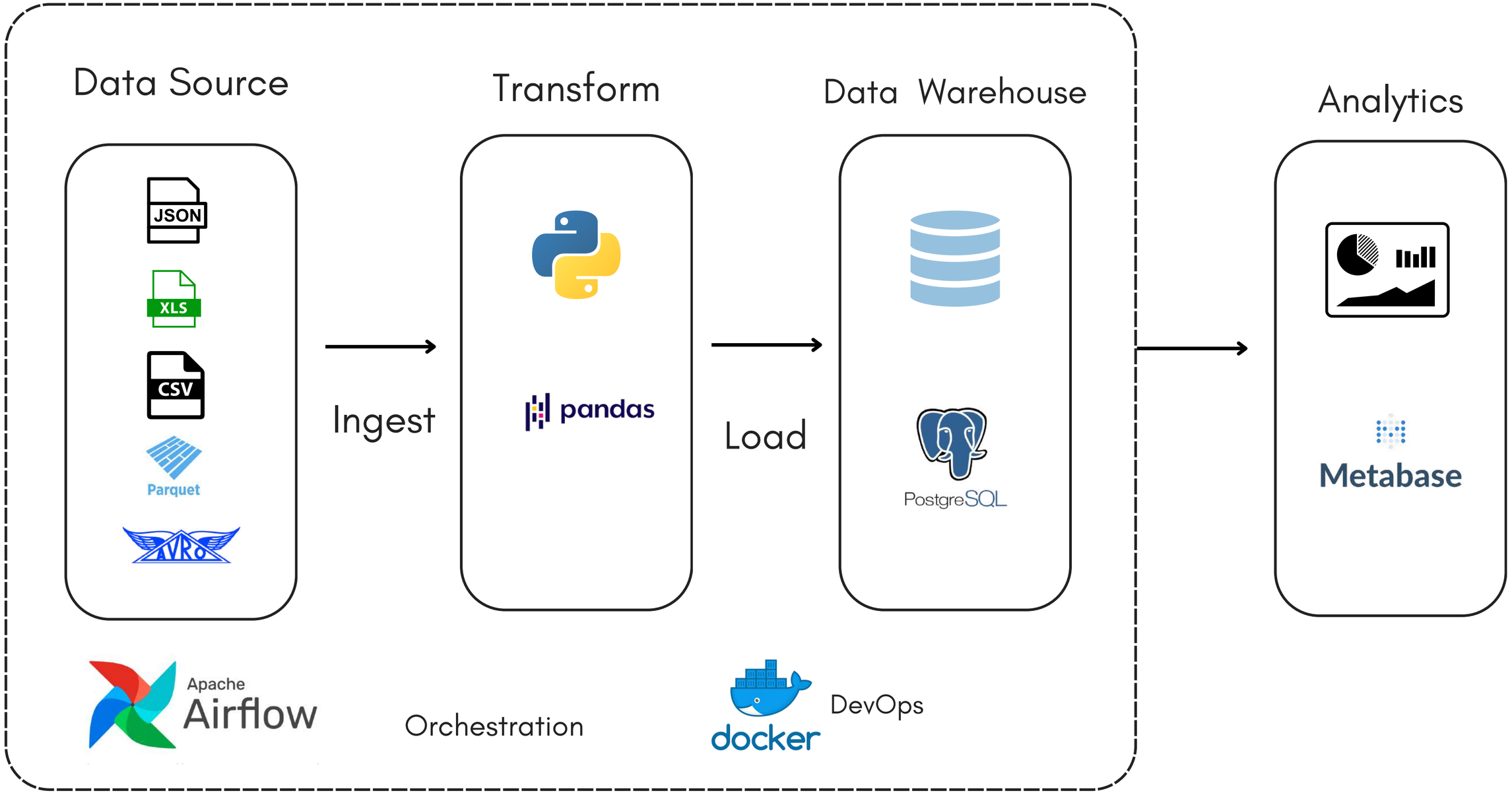
# TOOLS



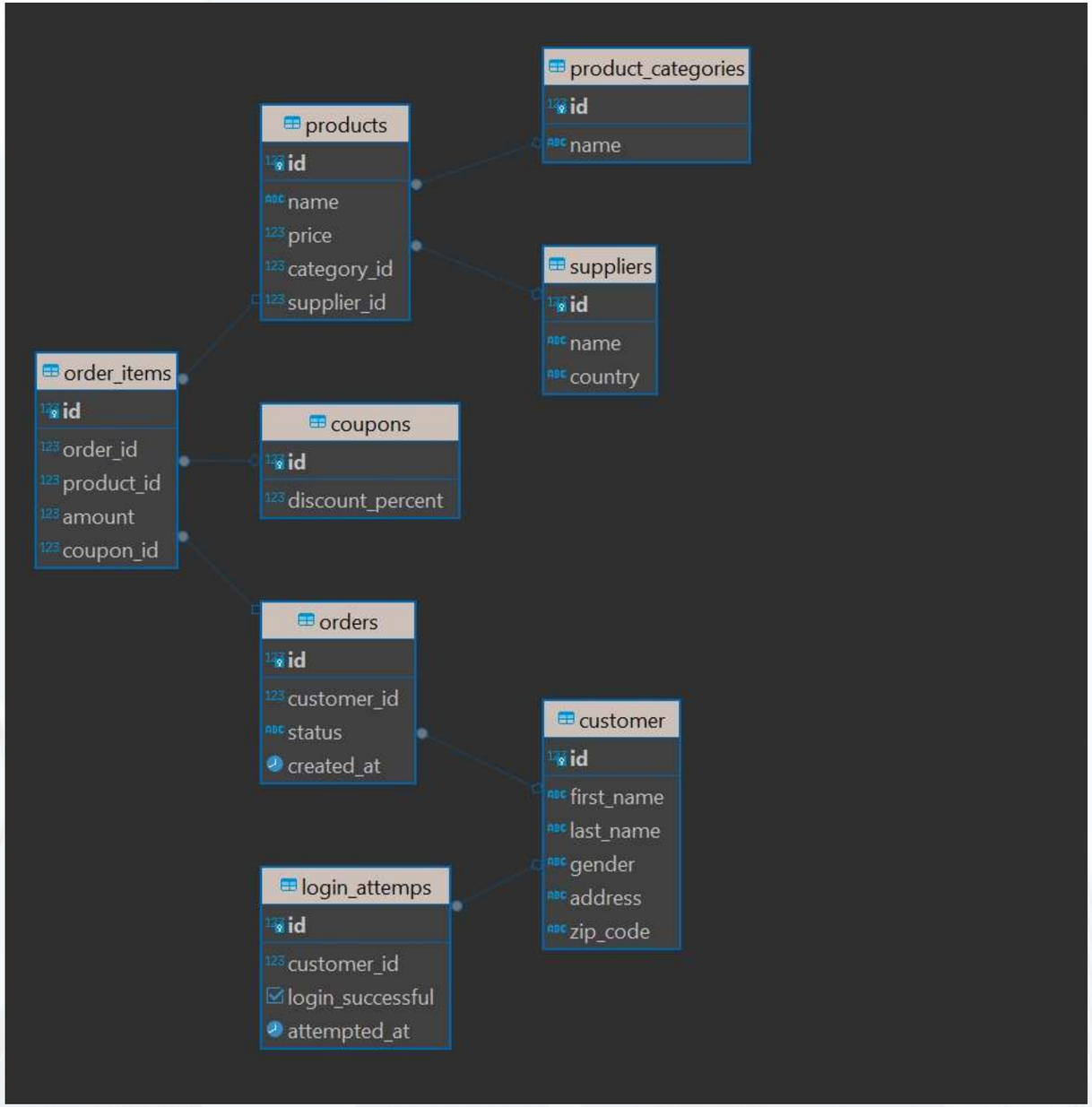


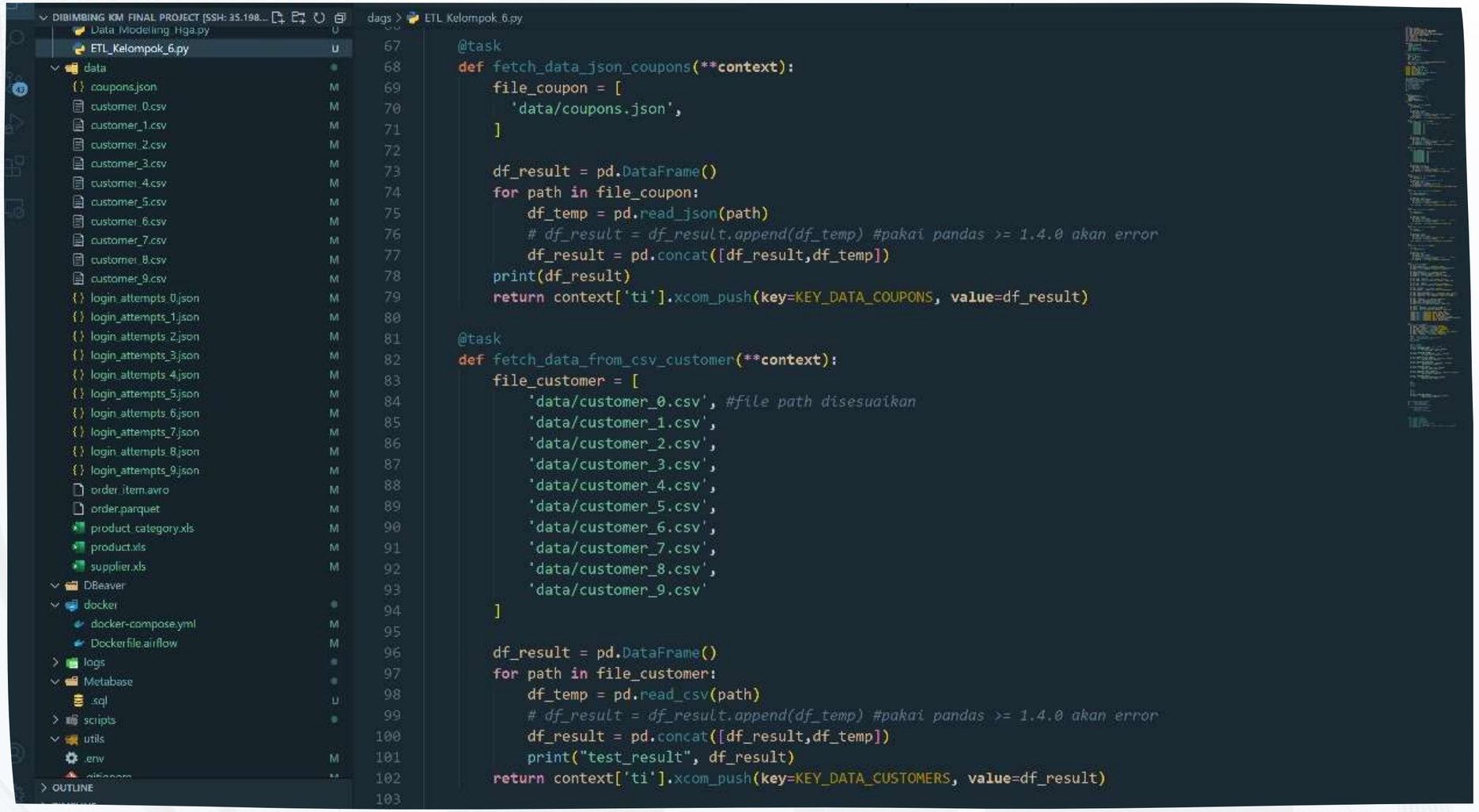
# Data Platform Architecture



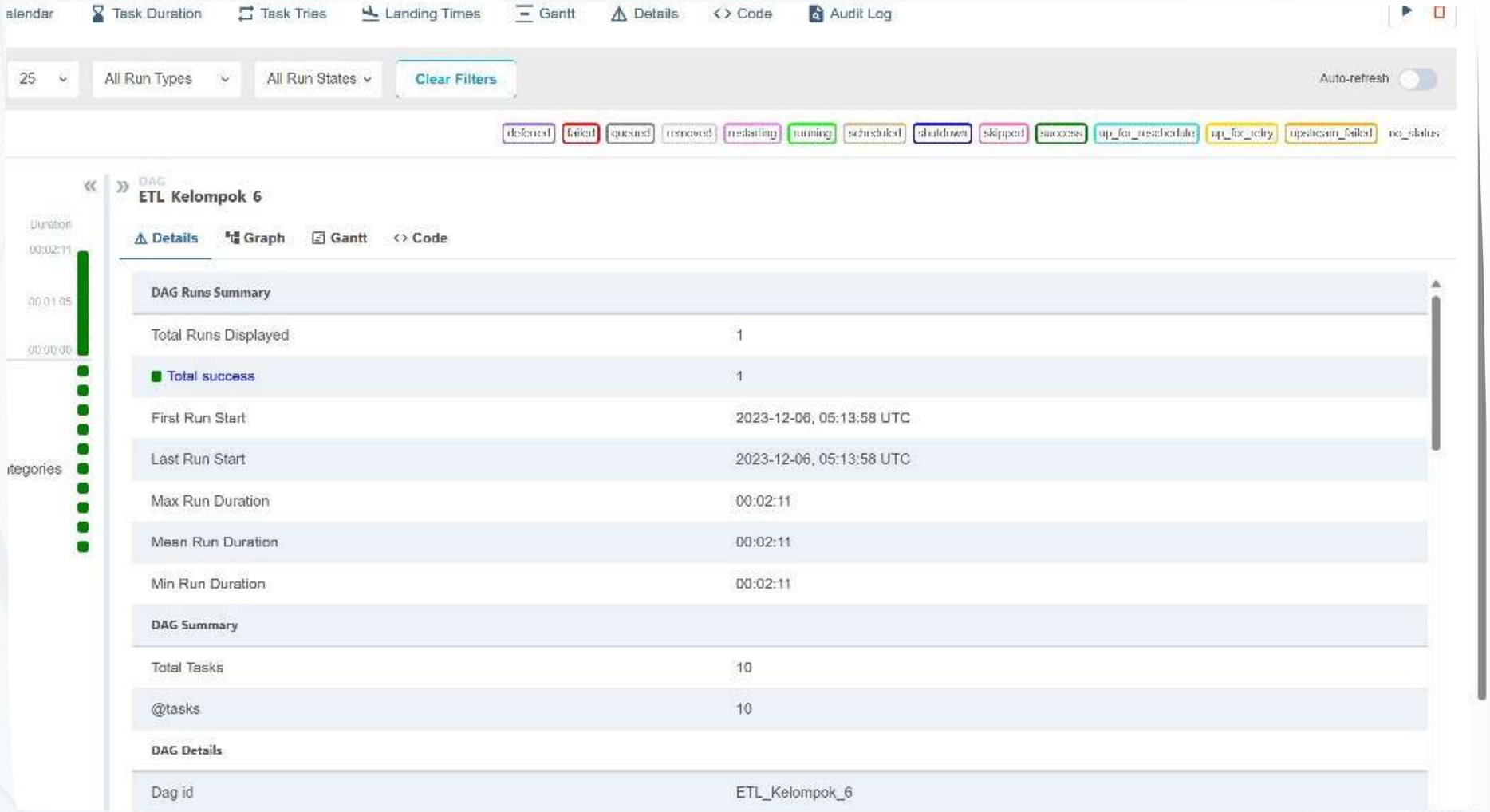
# ERD


product_categories
id
name
products
id
name
price
category_id
supplier_id
suppliers
id
name
country
order_items
id
order_id
product_id
amount
coupon_id
coupons
id
discount_percent
orders
id
customer_id
status
created_at
customer
id
first_name
last_name
gender
address
zip_code
login_attemps
id
customer_id
login_successful
attempted_at

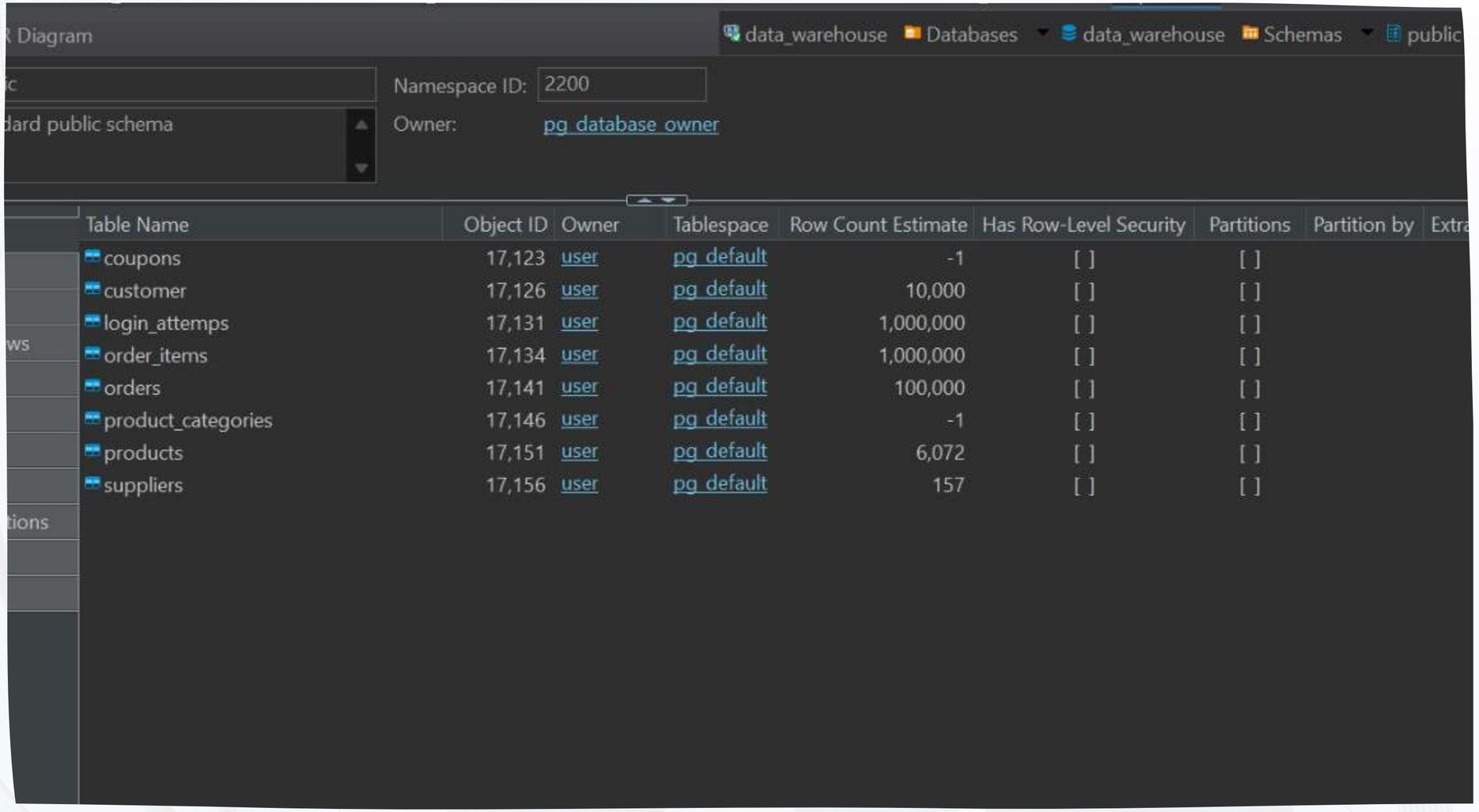
# ETL


DIBIMBING KM FINAL PROJECT [SSH: 35.198...
Data Modelling_Hga.py
ETL_Kelompok_6.py
data
coupons.json
customer_0.csv
customer_1.csv
customer_2.csv
customer_3.csv
customer_4.csv
customer_5.csv
customer_6.csv
customer_7.csv
customer_8.csv
customer_9.csv
login_attempts_0.json
login_attempts_1.json
login_attempts_2.json
login_attempts_3.json
login_attempts_4.json
login_attempts_5.json
login_attempts_6.json
login_attempts_7.json
login_attempts_8.json
login_attempts_9.json
order_item.avro
order.parquet
product_category.xls
product.xls
supplier.xls
DBeaver
docker
docker-compose.yml
Dockerfile.airflow
logs
Metabase
.sql
scripts
utils
.env
OUTLINE
dags > ETL_Kelompok_6.py
@task
def fetch_data_json_coupons(**context):
file_coupon = [
'data/coupons.json',
]
df_result = pd.DataFrame()
for path in file_coupon:
df_temp = pd.read_json(path)
# df_result = df_result.append(df_temp) #pakai pandas >= 1.4.0 akan error
df_result = pd.concat([df_result,df_temp])
print(df_result)
return context['ti'].xcom_push(key=KEY_DATA_COUPONS, value=df_result)
@task
def fetch_data_from_csv_customer(**context):
file_customer = [
'data/customer_0.csv', #file path disesuaikan
'data/customer_1.csv',
'data/customer_2.csv',
'data/customer_3.csv',
'data/customer_4.csv',
'data/customer_5.csv',
'data/customer_6.csv',
'data/customer_7.csv',
'data/customer_8.csv',
'data/customer_9.csv'
]
df_result = pd.DataFrame()
for path in file_customer:
df_temp = pd.read_csv(path)
# df_result = df_result.append(df_temp) #pakai pandas >= 1.4.0 akan error
df_result = pd.concat([df_result,df_temp])
print("test_result", df_result)
return context['ti'].xcom_push(key=KEY_DATA_CUSTOMERS, value=df_result)

# ETL

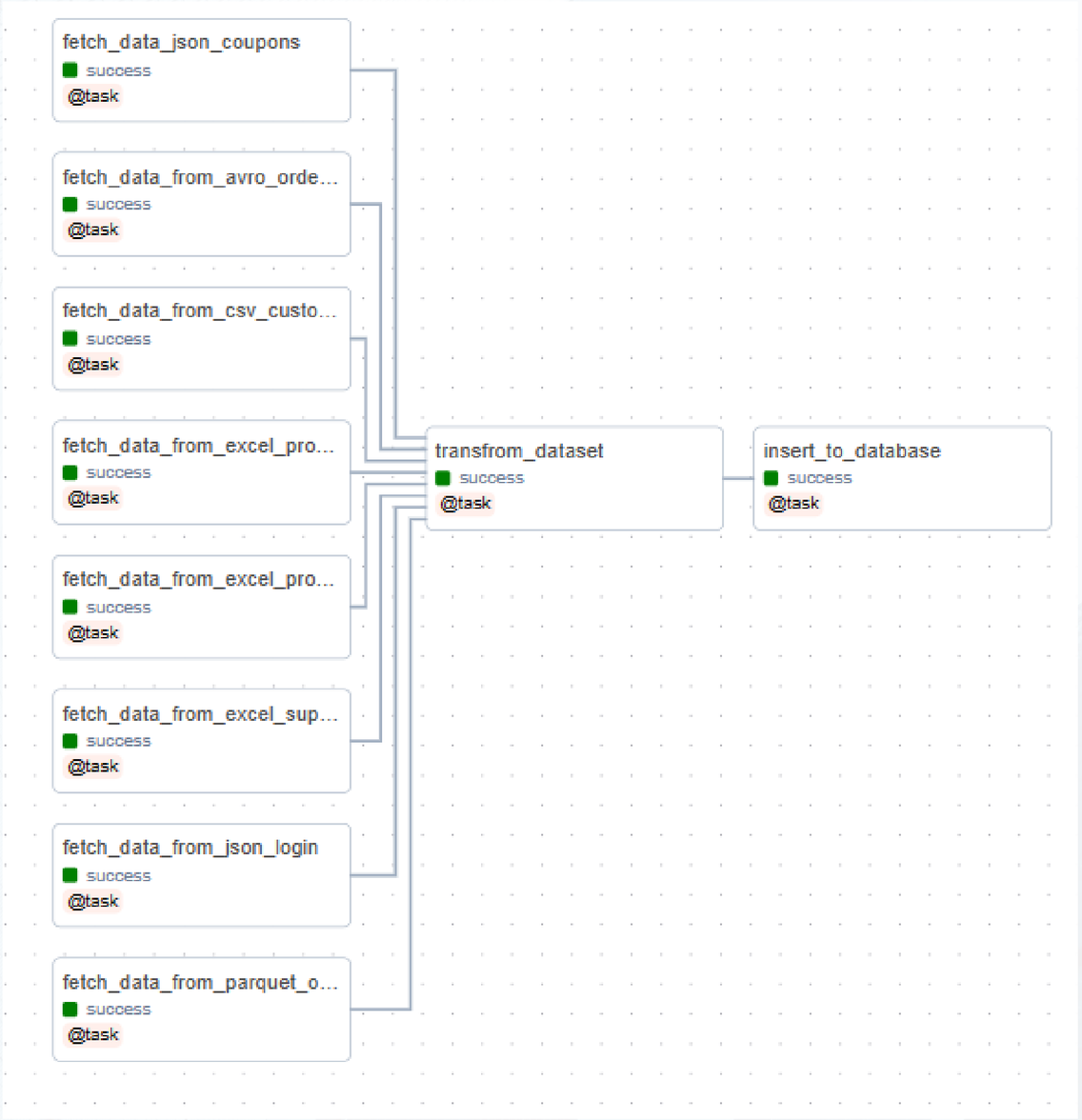

alendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
25
All Run Types
All Run States
Clear Filters
Auto-refresh
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
DAG
ETL Kelompok 6
Details
Graph
Gantt
Code
Duration
00:02:11
00:01:05
00:00:00
tegories
DAG Runs Summary
Total Runs Displayed 1
Total success 1
First Run Start 2023-12-06, 05:13:58 UTC
Last Run Start 2023-12-06, 05:13:58 UTC
Max Run Duration 00:02:11
Mean Run Duration 00:02:11
Min Run Duration 00:02:11
DAG Summary
Total Tasks 10
@tasks 10
DAG Details
Dag id ETL_Kelompok_6

# ETL



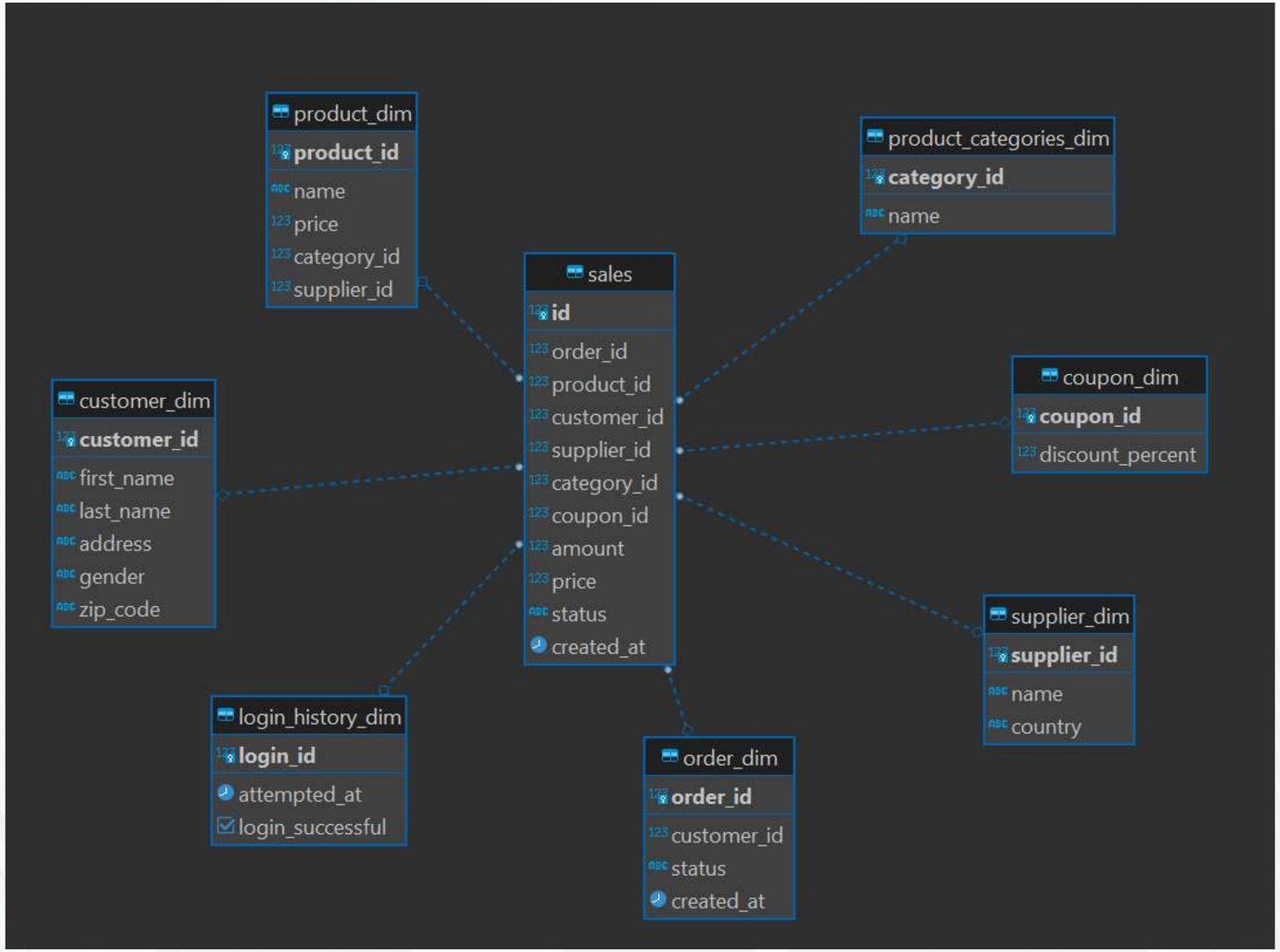

| Table Name | Object ID | Owner | Tablespace | Row Count Estimate | Has Row-Level Security | Partitions | Partition by |
|---|---|---|---|---|---|---|---|
| coupons | 17,123 | user | pg_default | -1 | [ ] | [ ] | |
| customer | 17,126 | user | pg_default | 10,000 | [ ] | [ ] | |
| login_attemps | 17,131 | user | pg_default | 1,000,000 | [ ] | [ ] | |
| order_items | 17,134 | user | pg_default | 1,000,000 | [ ] | [ ] | |
| orders | 17,141 | user | pg_default | 100,000 | [ ] | [ ] | |
| product_categories | 17,146 | user | pg_default | -1 | [ ] | [ ] | |
| products | 17,151 | user | pg_default | 6,072 | [ ] | [ ] | |
| suppliers | 17,156 | user | pg_default | 157 | [ ] | [ ] | |

# Graph


fetch_data_json_coupons
success
@task
fetch_data_from_avro_orde...
success
@task
fetch_data_from_csv_custo...
success
@task
fetch_data_from_excel_pro...
success
@task
fetch_data_from_excel_pro...
success
@task
fetch_data_from_excel_sup...
success
@task
fetch_data_from_json_login
success
@task
fetch_data_from_parquet_o...
success
@task
transfrom_dataset
success
@task
insert_to_database
success
@task

# Data Modelling

## Star Scheme



# Data Modelling

## Dim Table

```sql
CREATE TABLE "Fact and Dim Tables".customer_dim (
  customer_id integer PRIMARY KEY,
  first_name varchar,
  last_name varchar,
  address varchar,
  gender varchar,
  zip_code varchar
);

-- Assuming "id" is actually "customer_id" in the customers table
-- Insert data into "Fact and Dim Tables".customer_dim from the customers table
INSERT INTO "Fact and Dim Tables".customer_dim (customer_id, first_name, last_name, address, gender, zip_code)
SELECT
  id as customer_id,
  first_name,
  last_name,
  address,
  gender,
  zip_code
FROM
  "Data Modelling Metabase".customer;

CREATE TABLE "Fact and Dim Tables".product_dim (
  product_id integer PRIMARY KEY,
  name varchar,  -- Corrected column name from "names" to "name"
  price float,
  category_id integer,
  supplier_id integer
);
```

# Data Modelling

## Dim Fact

```sql
CREATE TABLE "Fact and Dim Tables".sales (
id serial PRIMARY KEY,
order_id integer,
product_id integer,
customer_id VARCHAR,  -- Assuming customer_id is a string, adjust the type accordingly
supplier_id integer,
category_id integer,
coupon_id varchar (255),
amount integer,
price float,
status text,
created_at date
);

-- Insert data into the fact.sales table using a SELECT statement
INSERT INTO "Fact and Dim Tables".sales (
  order_id,
  product_id,
  customer_id,
  supplier_id,
  category_id,
  coupon_id,
  amount,
  price,
  status,
  created_at
)
SELECT
  o.id as order_id,
  oi.product_id,
  c.id as customer_id,
  p.supplier_id,
  p.category_id,
  oi.coupon_id,
  oi.amount,
  p.price,
  o.status,
  o.created_at as date
FROM
  "Data Modelling Metabase".customer c
  JOIN orders o ON c.id = o.customer_id
  JOIN order_items oi ON o.id = oi.order_id
  JOIN products p ON oi.product_id = p.id
  JOIN suppliers s ON p.supplier_id = s.id
  JOIN login_attemps l ON c.id = l.customer_id;

```

# Data Visualization

## Coupon Effectiveness Analysis:

Objective: Evaluate the effectiveness of coupon campaigns
Data Sources :

- Coupons for coupon information.
- Order_items for coupon usage in orders.

Analytics Queries :

- Calculate the impact of coupons on total revenue.
- Assess the popularity of different types of coupons.
- Measure the conversion rate of orders with and without coupons.
- Display coupon performance metrics on the dashboard

# Data Visualization

## Customer Segmentation

Objective : Identify different customer segments based on their purchasing behavior.

Data Sources :

- Customers for customer information.
- Orders for order history.

Analytics Queries :

- Aggregate total spending per customer.
- Categorize customers into segments (e.g., high spenders, frequent shoppers, one-time buyers).
- Visualize customer segments on the dashboard.

# Data Visualization

## Order Fulfillment Metrics

Objective : Monitor order fulfillment and track order status

Data Sources :

- Orders for order information

Analytics Queries :

- Track the status distribution of orders ( returned).
- Analyze return rate for each supplier.
- Identify bottlenecks in the fulfillment process.
- Visualize order status trends on the dashboard.

# Data Visualization

## Product Performance Analysis

Objective : Evaluate the performance of products in terms of sales and popularity.

Data Sources :

- fact.products for product information.
- fact.order_items for order details.

Analytics Queries :

- Calculate total revenue generated by each product.
- Identify best-selling products.
- Analyze product sales trends over time.
- Display top-performing products on the dashboard.

# Data Visualization

## Supplier Performance Metrics

Objective : Assess the performance of suppliers in terms of product quality and delivery.

Data Sources :

- fact.suppliers for supplier information.
- fact.products for product-supplier relationships

Analytics Queries :

- Evaluate supplier reliability based on order fulfillment.
- Monitor product return rates by supplier.
- Identify top-performing suppliers.
- Provide supplier performance scores on the dashboard.

# Data Visualization

## Login Activity Monitoring

Objective : Monitor and analyze customer login activity.

Data Sources :

- fact.login_attempt_history for login attempts.

Analytics Queries :

- Count successful and unsuccessful login attempts.
- Identify login patterns (e.g., peak login hours).
- Highlight any unusual login activity.
- Provide insights into account security on the dashboard

# Data Visualization

Coupons | Customer | Orders | Products | Supplier | Login

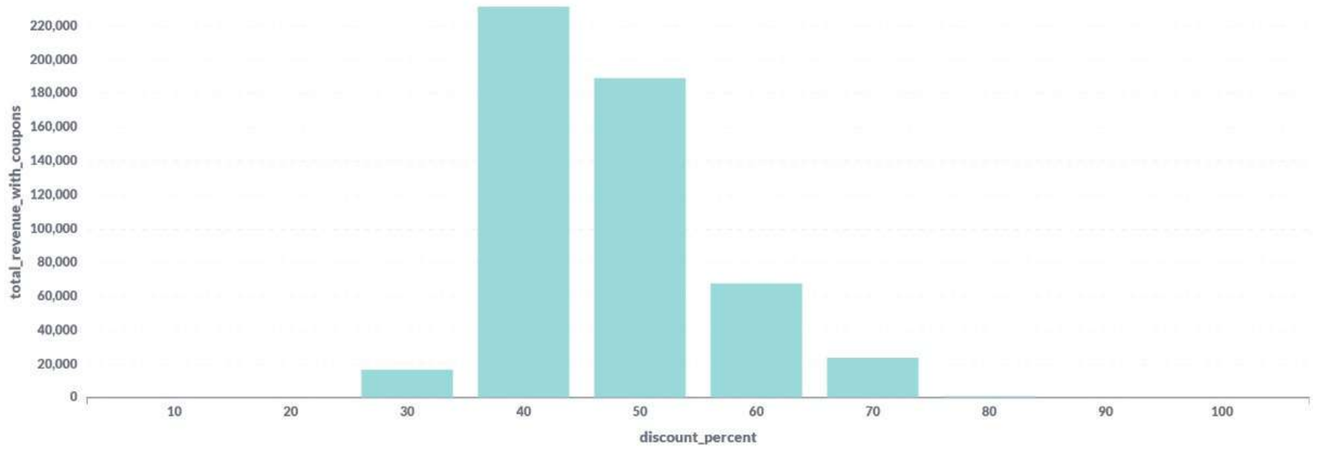
**The Conversion Rate of Orders with and without Coupons**

| | |
|---|---|
| orders_with_coupons | 237 |
| total_orders | 99,995 |
| conversion_rate | 0.24% |

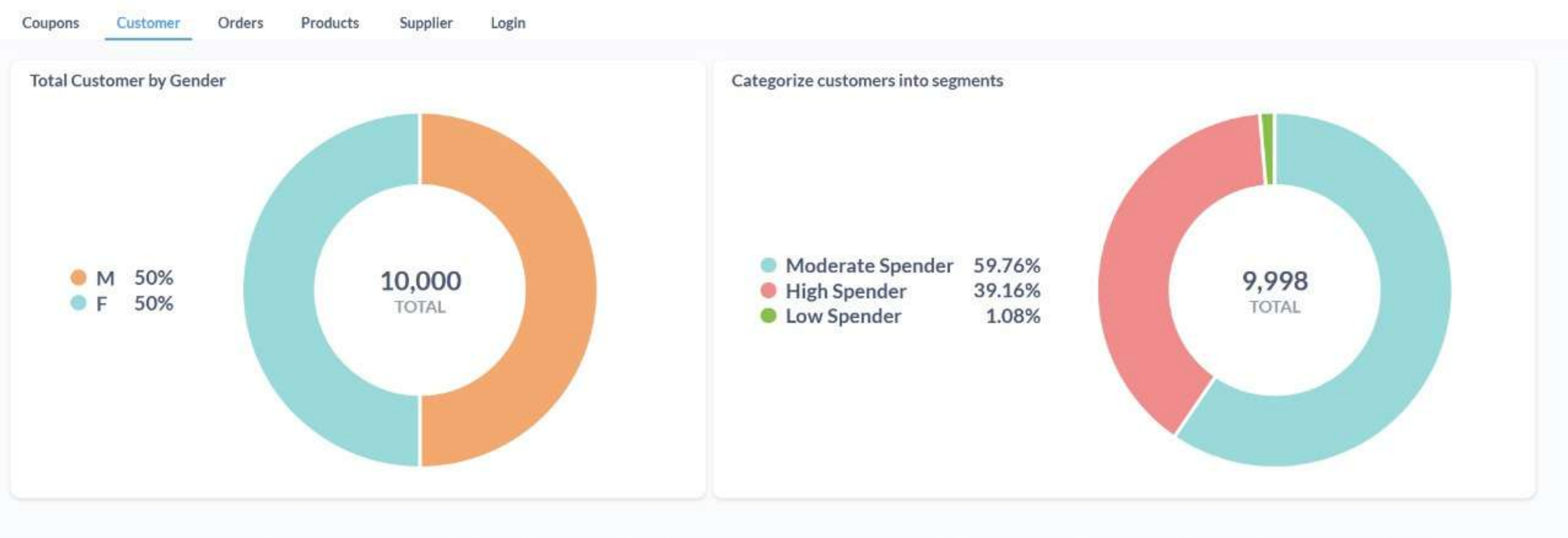
**Total Revenue in 2023**

| order_year | total_revenue |
|---|---|
| 2023 | 4,652,094,680 |

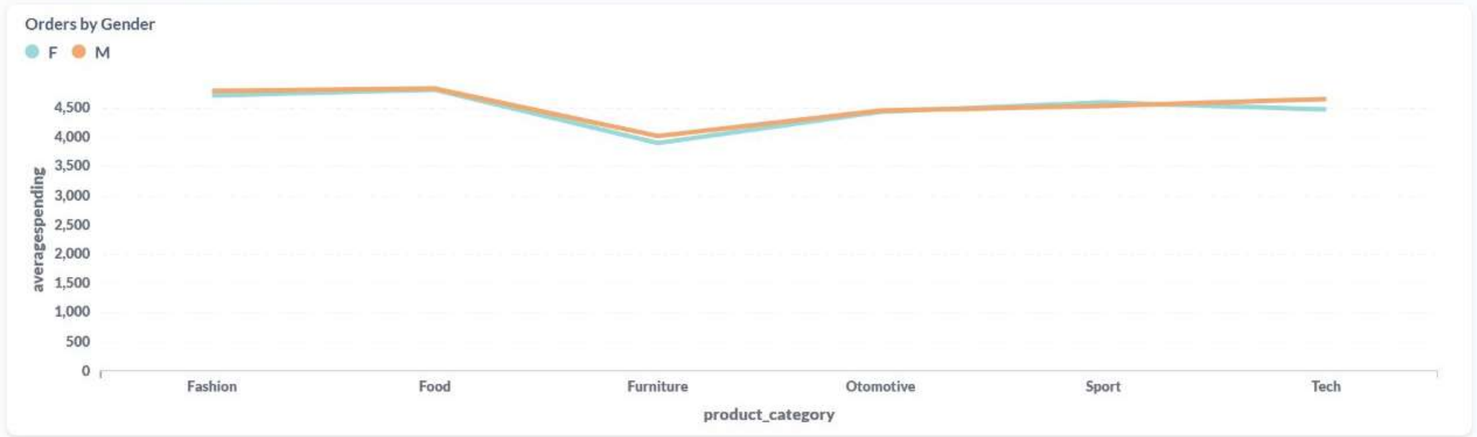
**Coupon performance metrics**

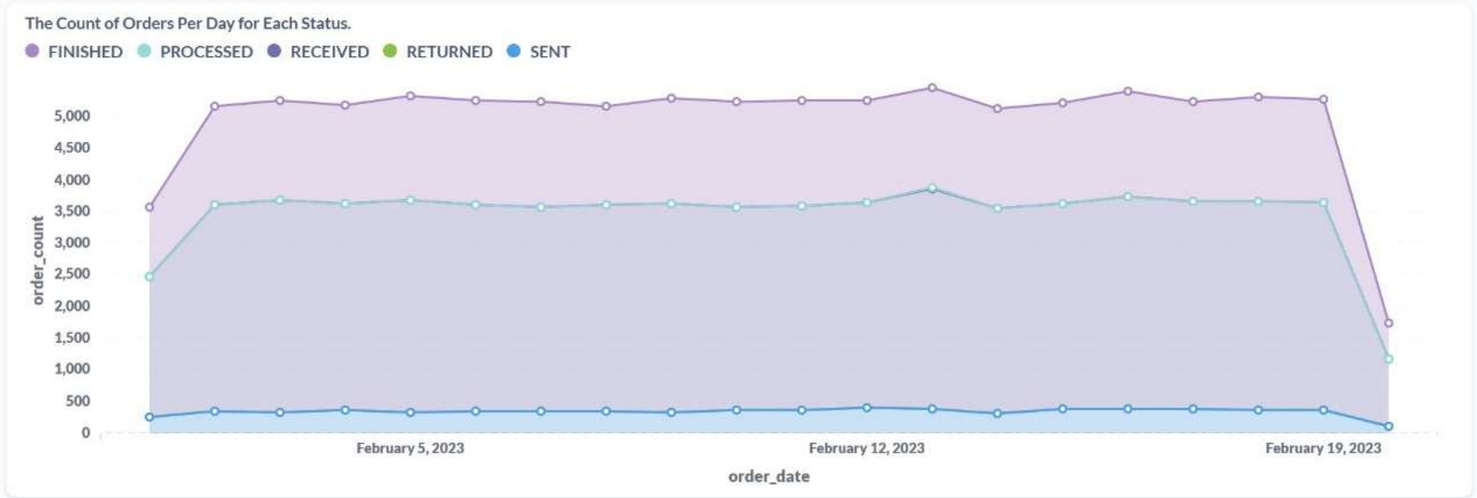


# Data Visualization



# Data Visualization



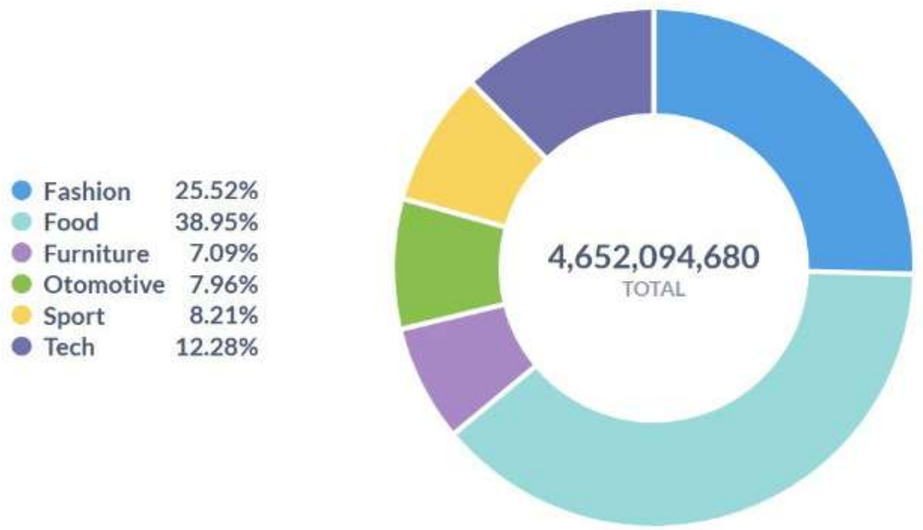



# Data Visualization

Sales by Product Price

| product_price | total_amount_sold |
|---|---|
| 5 | 2,877,546 |
| 10 | 5,764,018 |
| 15 | 8,820,388 |
| 25 | 11,734,081 |
| 50 | 8,744,353 |
| 75 | 6,155,477 |
| 100 | 2,965,645 |
| 1,000 | 2,958,966 |

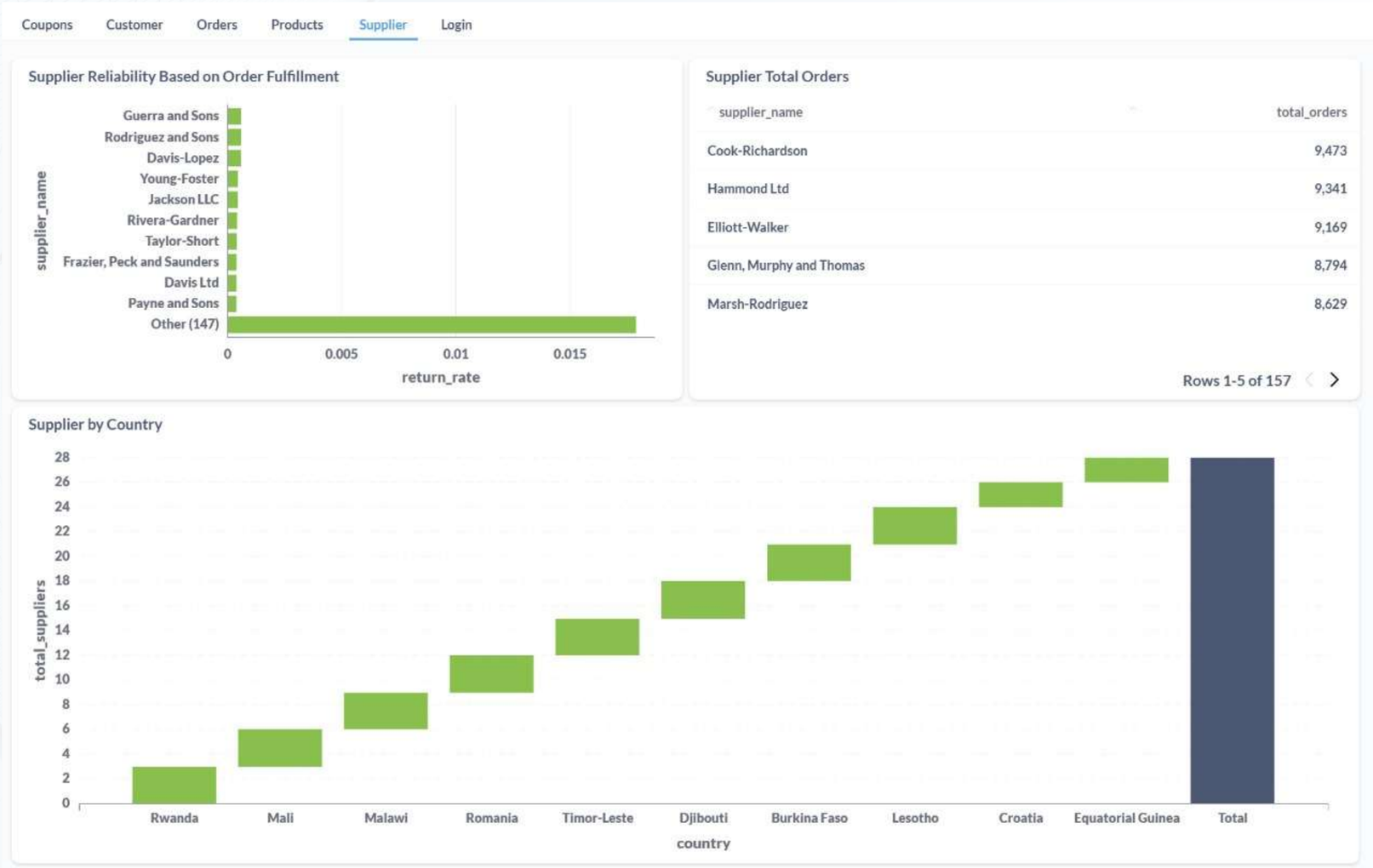
Sales by Product Category



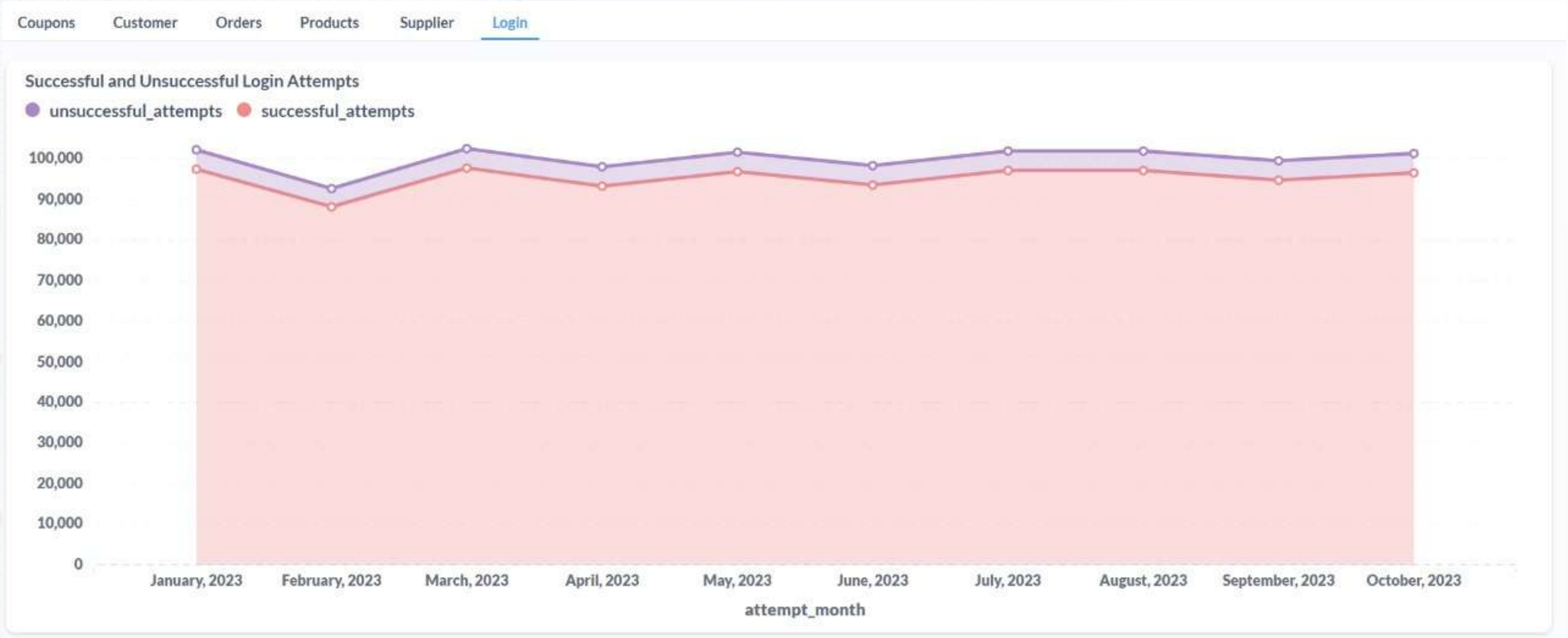

**4,652,094,680**

Total Revenue in the Last Year

# Data Visualization



# Data Visualization



# Sesi Tanya Jawab

## 10 Menit

# TERIMA KASIH

*Atas Waktu dan Kesediaan Anda Untuk*
*Mendengarkan Presentasi dari Kami.*