TECHNICAL TEST – DATA ENGINEER

TRIPUTRA AGRO PERSADA

1. Buat akun **Snowflake Free Trial**
2. Download dataset dari Kaggle

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Tables yang digunakan:

- orders.csv
- order_items.csv
- customers.csv
- products.csv
- payments.csv

I

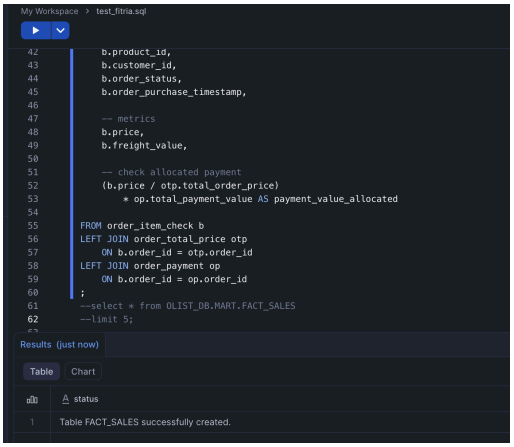
3. Upload CSV ke Snowflake stage
4. Create raw tables:
   - raw_orders
   - raw_order_items
   - raw_customers
   - raw_products
   - raw_payments

Task 1

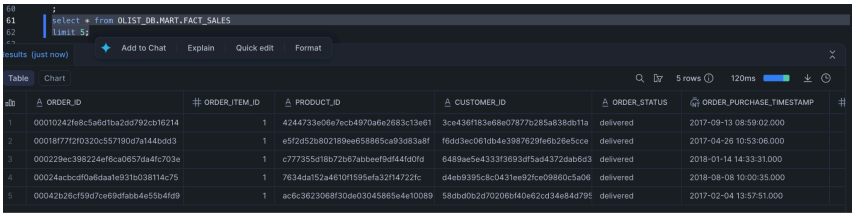
Buat:

- 1 fact table: fact_sales

Jawab : Create Table Fact_sales



limit 5 fact tables



penjelasan : fact_sale merepresentasikan aktivitas penjualan (sales) pada level order item, di mana 1 baris = 1 produk yang terjual dalam satu order.

Fact table ini menjadi sumber utama untuk analisis revenue, volume penjualan, dan performa customer maupun produk.

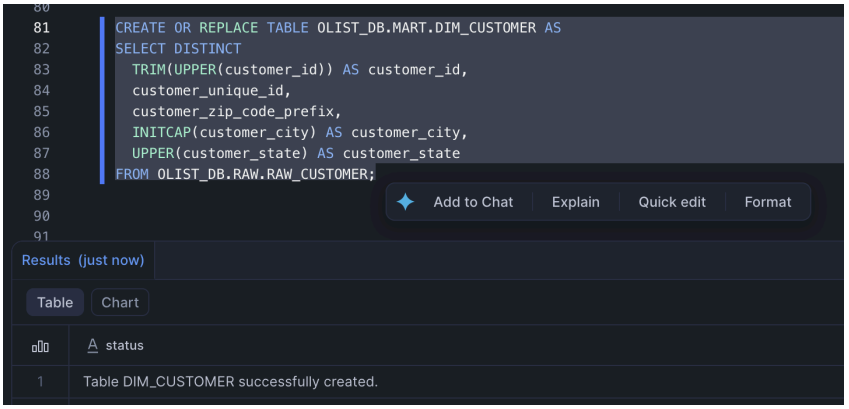
Sumber yang saya gunakan meliputi :

- RAW_ORDER_ITEM (pondasi), karena tedapat relasi order terhadap product, terdapat harga per item beserta biaya pengirimannya.
- RAW_ORDER (informasi order), karena terdapat relasi terhadap customer, status order dan waktu pembelian
- RAW Customer (informasi customer), table ini digunakan untuk memberikan detail customer yang melakukan pemesanan dan akan menjadi relasi untuk pembuatan dim_customer
- RAW PAYMENT (informasi finansial), karena payment berada di level order (1 row per order), sedangkan fact_sale berada di level order item (1 row per product per oder), dilakukan pendekatan.
  allocated_payment = (item_price / total_order_price) × total_payment_value

○ Dimension tables:

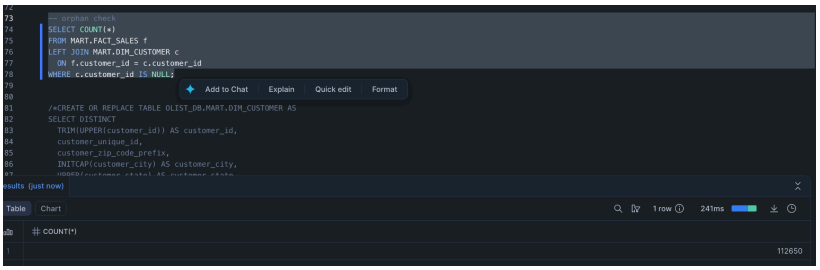
§ dim_customers

Jawab :

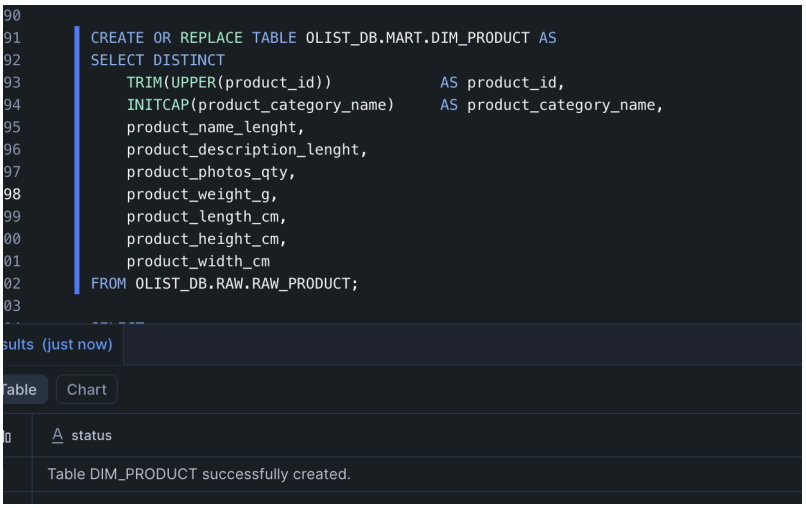


Penjelasan : pada dim_customer, dilakukan data cleansing pada script diatas dan dilakukan pengecekan jumlah terhadap fact_sales untuk

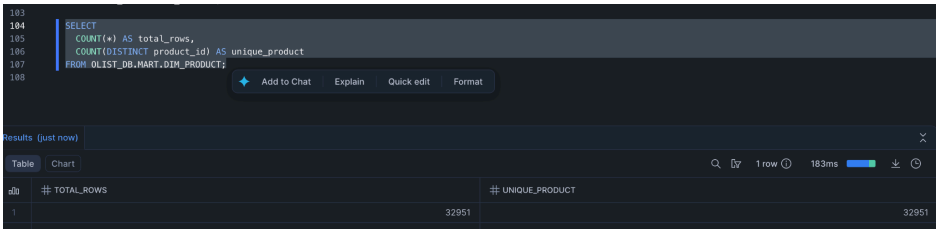
menghindari data yang tidak sesuai



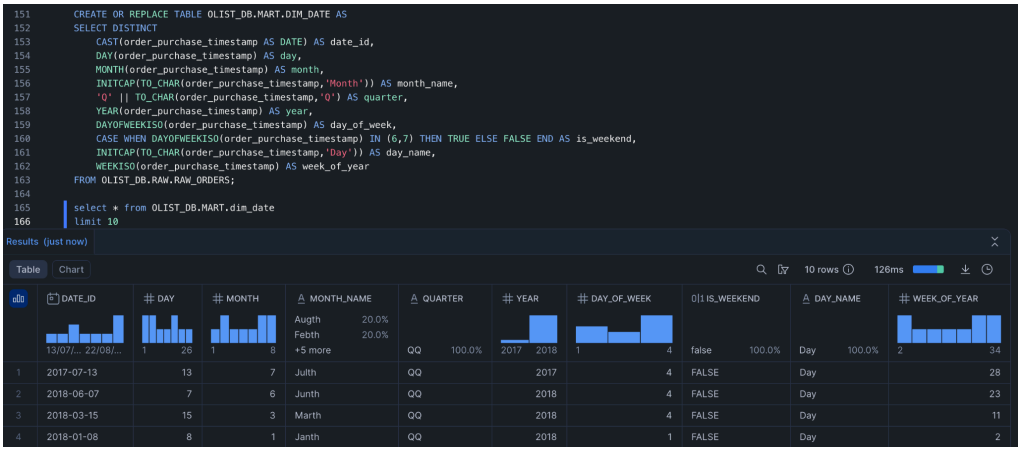
§ dim_products



Penjelasan : Dilakukan cleansing data dan cross check agak tidak ada duplicat



§ dim_date
jawab :

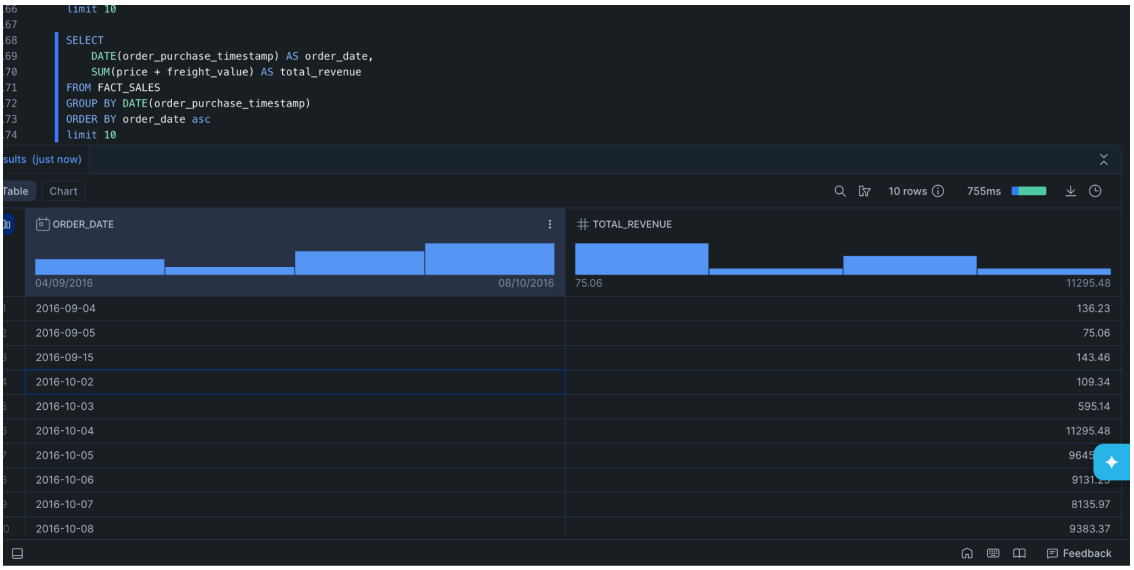


Task 2

Buat SQL query untuk menjawab:

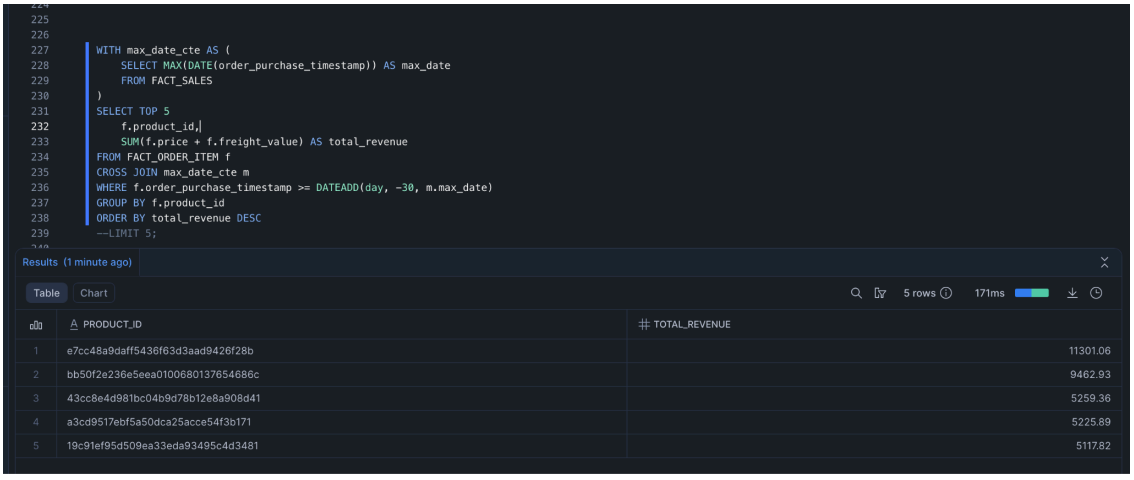
1. Total revenue per day

Jawab :



2. Top 5 products by revenue in last 30 days

Jawab :



Penjelasan : disini untuk revenue 30 hari terhari, menggunakan 30 hari terakhir dari Maxdate

3. Monthly active customers

Jawab

```
ORDER BY total_revenue DESC

SELECT
    YEAR(order_purchase_timestamp) AS year,
    MONTH(order_purchase_timestamp) AS month,
    COUNT(DISTINCT customer_id) AS active_customers
FROM FACT_SALES
GROUP BY YEAR(order_purchase_timestamp), MONTH(order_purchase_timestamp)
ORDER BY year, month;
```

Results (just now) | Add to Chat | Explain | Quick edit | Format

Table | Chart | 24 rows | 191ms

	YEAR	MONTH	ACTIVE_CUSTOMERS
	2016 – 2018	1 – 12	1
1	2016	9	
2	2016	10	
3	2016	12	
4	2017	1	
5	2017	2	
6	2017	3	
7	2017	4	
8	2017	5	

4. Average order value per customer

Jawab :

```
SELECT
    f.customer_id,
    AVG(f.price + f.freight_value) AS avg_order_value
FROM FACT_SALES f

GROUP BY f.customer_id
ORDER BY avg_order_value DESc
limit 10
;

*I to generate
```

Results (2 minutes ago)

Table | Chart | 10 rows | 146ms

	CUSTOMER_ID	AVG_ORDER_VALUE
	1afc82cd60e303ef09b4ef9837c9505c 10.0% 24bbf5fd2f2e1b359ee7de94defc4a15 10.0% +8 more	4163.51 – 6929.31
1	c6e2731c5b391845f6800c97401a43a9	6929.31000000
2	f48d464a0baaea338cb25f816991ab1f	6922.21000000
3	3fd6777bbce08a352fddd04e4a7cc8f6	6726.66000000
4	df55c14d1476a9a3467f131269c2477f	4950.34000000
5	24bbf5fd2f2e1b359ee7de94defc4a15	4764.34000000
6	3d979689f636322c62418b6346b1c6d2	4681.78000000
7	1afc82cd60e303ef09b4ef9837c9505c	4513.32000
8	926b6a6fb8b6081e00b335edaf578d35	4194.76000000

5. Data anomaly:
    - price <= 0
    - quantity <= 0
    - missing customer_id

    Jawab :

```
SELECT
    SUM(CASE WHEN price <= 0 THEN 1 ELSE 0 END) AS price_invalid,
    SUM(CASE WHEN customer_id IS NULL THEN 1 ELSE 0 END) AS missing_customer
FROM FACT_SALES;
```

Results (just now) | Add to Chat | Explain | Quick edit | Format

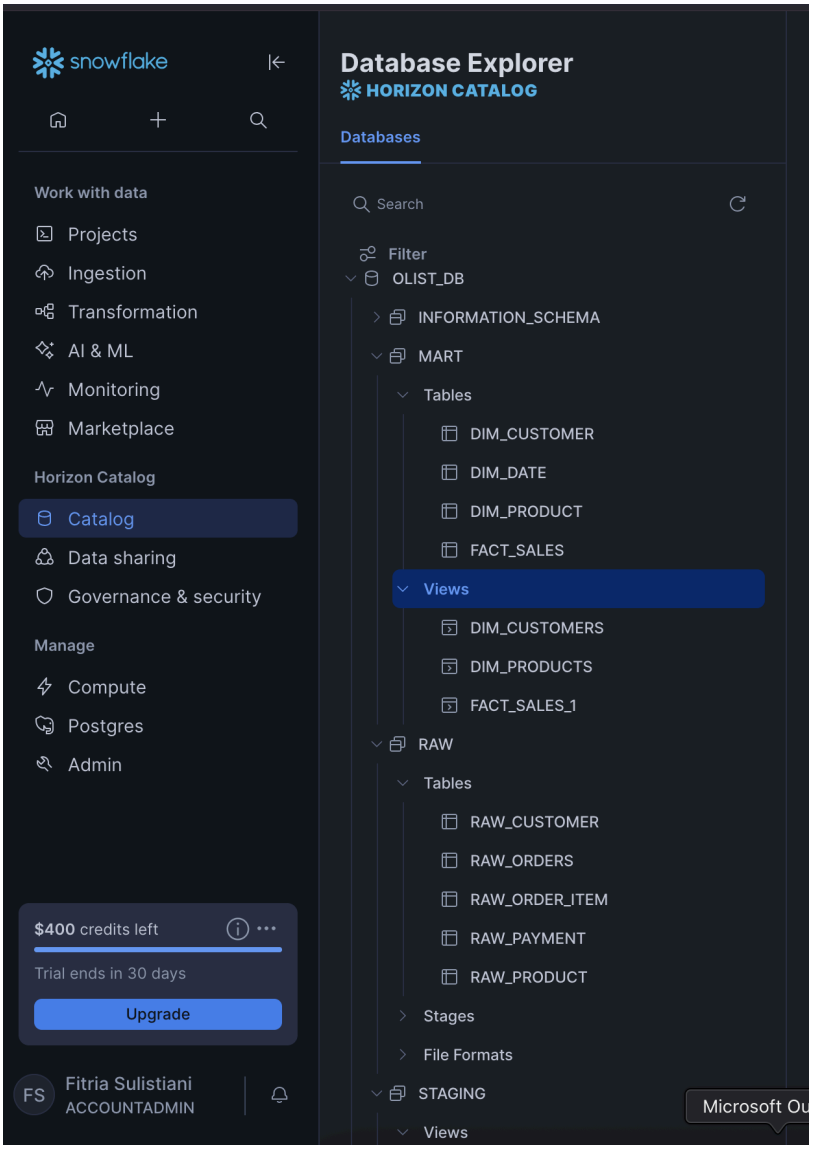
Table | Chart | 1 row | 481ms

	PRICE_INVALID	MISSING_CUSTOMER
1	0	0

Task 3

Buat dbt project dengan struktur:
models/
 staging/
  stg_orders.sql

  stg_customers.sql
 stg_products.sql
 marts/
  dim_customers.sql
  dim_products.sql
  fact_sales.sql
jawab



penjelasan: untuk dbt saya gagal dalam koneksi dilocal, sehingga saya coba bikin alternatif disnowflake

Task 4

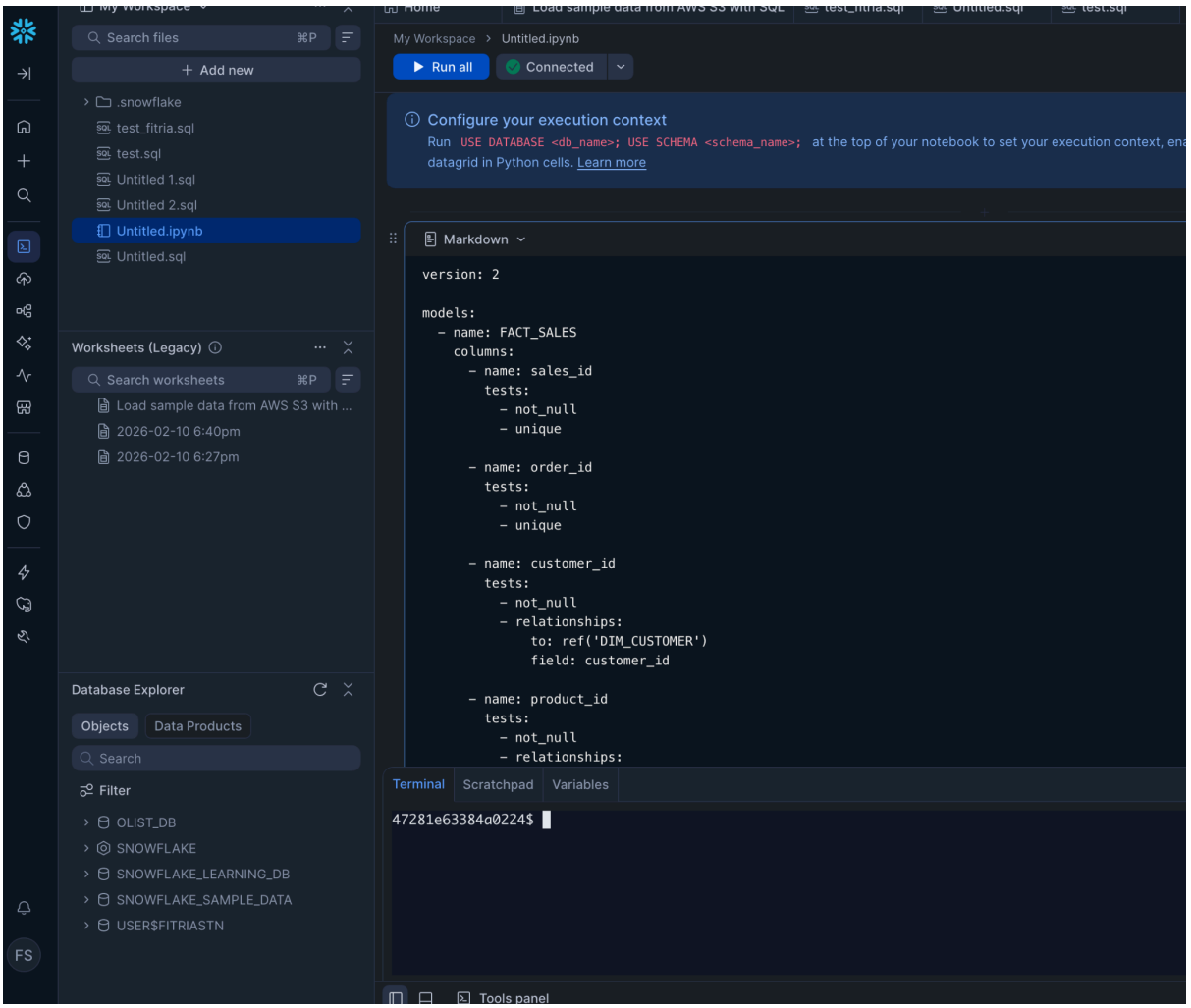
Buat file schema.yml berisi tests berikut:

- not_null pada primary key
- unique pada order_id
- accepted_values pada order_status
- relationships:
  - fact_sales.customer_id → dim_customers.customer_id
  - fact_sales.product_id → dim_products.product_id

Jawab:

```
models:
  - name: FACT_SALES
    columns:
      - name: sales_id
        tests:
          - not_null
          - unique

      - name: order_id
        tests:
          - not_null
          - unique

      - name: customer_id
        tests:
          - not_null
          - relationships:
              to: ref('DIM_CUSTOMER')
              field: customer_id

      - name: product_id
        tests:
          - not_null
          - relationships:
              to: ref('DIM_PRODUCT')
              field: product_id
```

```
    - name: order_status
      tests:
        - not_null
        - accepted_values:
            values: ['pending', 'completed', 'canceled', 'returned']

- name: DIM_CUSTOMER
  columns:
    - name: customer_id
      tests:
        - not_null
        - unique

- name: DIM_PRODUCT
  columns:
    - name: product_id
      tests:
        - not_null
        - unique
```



penjelasan : lanjutan dari task 3, saya masih gagal dalam mebuat koneksi

DELIVERABLE

GitHub repository berisi:

1. SQL scripts
2. dbt project
3. Screenshot :
    - dbt run
    - dbt test
    - Snowflake tables