LAPORAN UAS

ETL DATA WAREHOUSE

“PENJUALAN PRODUK RETAIL”

Disusun Oleh:

DANICA NASYWA PUTRINIAR	**2341760122**
KARTIKA TRI JULIANA	**2341760116**
PUTERA BHAGASWARA R.	**2341760136**
QUUENADHYNAR AZARINE	**2341760109**
YUSRA YUSUF	**2341760044**

JURUSAN TEKNOLOGI INFORMASI
D4 SISTEM INFORMASI BISNIS
POLITEKNIK NEGERI MALANG
2025

1. **Studi Kasus: Penjualan Produk Retail**

Data yang digunakan adalah file train.csv berisi transaksi penjualan yang mencakup atribut berikut:

- Customer Name
- Product Name
- Region
- Sales
- Quantity
- Profit
- Order Date

2. **Pemilihan dan Persiapan Data**

Dari file train.csv, dibuat 4 kategori data:

- Data Pelanggan (Customer)
- Data Produk (Product)
- Data Wilayah (Region)
- Data Penjualan (Fact Sales)

3. **Skema Bintang (Star Schema)**

- Tabel Fakta: fact_sales
  product_id (FK ke dim_product)
  customer_id (FK ke dim_customer)
  region_id (FK ke dim_region)
- Tabel Dimensi:
  dim_customer: customer_id, customer_name
  dim_product: product_id, product_name
  dim_region: region_id, region_name

4. **Desain dan Implementasi Proses ETL**

➢ **Extract**

a. Menggunakan komponen CSV file input untuk membaca file train.csv.
b. Header dibaca satu kali, lalu didistribusikan ke jalur masing-masing (Customer, Product, Region, Fact).

➢ **Transform**

a. Select Values: Memilih kolom relevan untuk masing-masing tabel.
b. Sort Rows + Unique Rows: Menghilangkan duplikat untuk tabel dimensi.

c. Add Sequence: Menambahkan ID unik untuk setiap dimensi (misal region_id).

d. Database Lookup: Mengambil region_id, product_id, dan customer_id dari tabel dimensi agar bisa digunakan di fact_sales.

➢ **Load**

Table Output:

a. dim_customer: 793 baris unik

b. dim_product: 1861 baris unik

c. dim_region: 600 baris unik

d. fact_sales: 9800 baris (setelah lookup ID dan pembersihan data)

5. Permasalahan dan Penyelesaian

a. Error 1: Field region_id not found

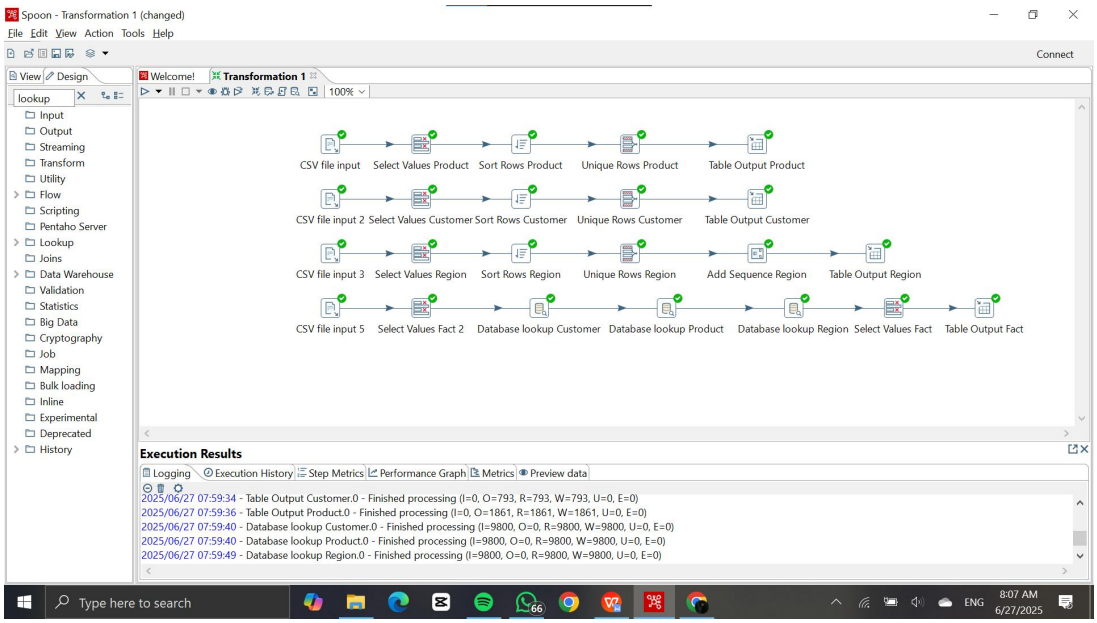
Solusi: Menambahkan step Database Lookup setelah dim_region untuk mencari region_id berdasarkan region_name.

b. Error 2: Data truncation product_id

Solusi: Kemungkinan product_id terlalu panjang untuk tipe data di database.

Solusi dilakukan dengan: memastikan tipe kolom product_id di tabel fact_sales sama (misal VARCHAR(255) atau INT jika pakai ID numerik dari sequence).

6. Dokumentasi Pipeline ETL



Pipeline ETL terdiri dari:

a. 4 alur CSV file input ke masing-masing entitas.

b. Transformasi Select Values, Sort Rows, Unique Rows.

c. Lookup ke tabel dimensi sebelum mengisi tabel fakta.

d. Output ke database MySQL dw_sales.

7. Hasil dan Analisis

- Hasil Akhir:

  dim_customer: 793 record

  dim_product: 1861 record

  dim_region: 600 record

  fact_sales: 9800 record berhasil masuk

8. Evaluasi KPI (Key Performance Indicators)

KPI	Hasil
Waktu Proses ETL	± 2-3 detik per load
Konsitensi Data	Duplikat berhasil dihapus (unique rows)
Kualitas Data	Kolom yang dibutuhkan terisi lengkap
Akurasi Foreign Key	Sudah tervalidasi melalui Database Lookup
Eror Rate	0 (setelah perbaikan region_id & product_id)

9. Kesimpulan

Proyek ETL berhasil mengintegrasikan data mentah dari file CSV ke dalam struktur data warehouse berbasis skema bintang. Data berhasil dipisahkan ke dalam dimensi dan fakta yang dapat di-query secara efisien untuk kebutuhan analisis lebih lanjut.