**LAPORAN UAS**

**ETL DATA WAREHOUSE**

**“PENJUALAN PRODUK RETAIL”**

A logo of a building

Description automatically generated

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**

### 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

1. **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)

### 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

### Desain dan Implementasi Proses ETL

#### Extract

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

#### Transform

1. **Select Values**: Memilih kolom relevan untuk masing-masing tabel.
2. **Sort Rows** + **Unique Rows**: Menghilangkan duplikat untuk tabel dimensi.
3. **Add Sequence**: Menambahkan ID unik untuk setiap dimensi (misal region\_id).
4. **Database Lookup**: Mengambil region\_id, product\_id, dan customer\_id dari tabel dimensi agar bisa digunakan di fact\_sales.

#### Load

**Table Output**:

1. dim\_customer: 793 baris unik
2. dim\_product: 1861 baris unik
3. dim\_region: 600 baris unik
4. fact\_sales: 9800 baris (setelah lookup ID dan pembersihan data)

### Permasalahan dan Penyelesaian

#### Error 1: Field region\_id not found

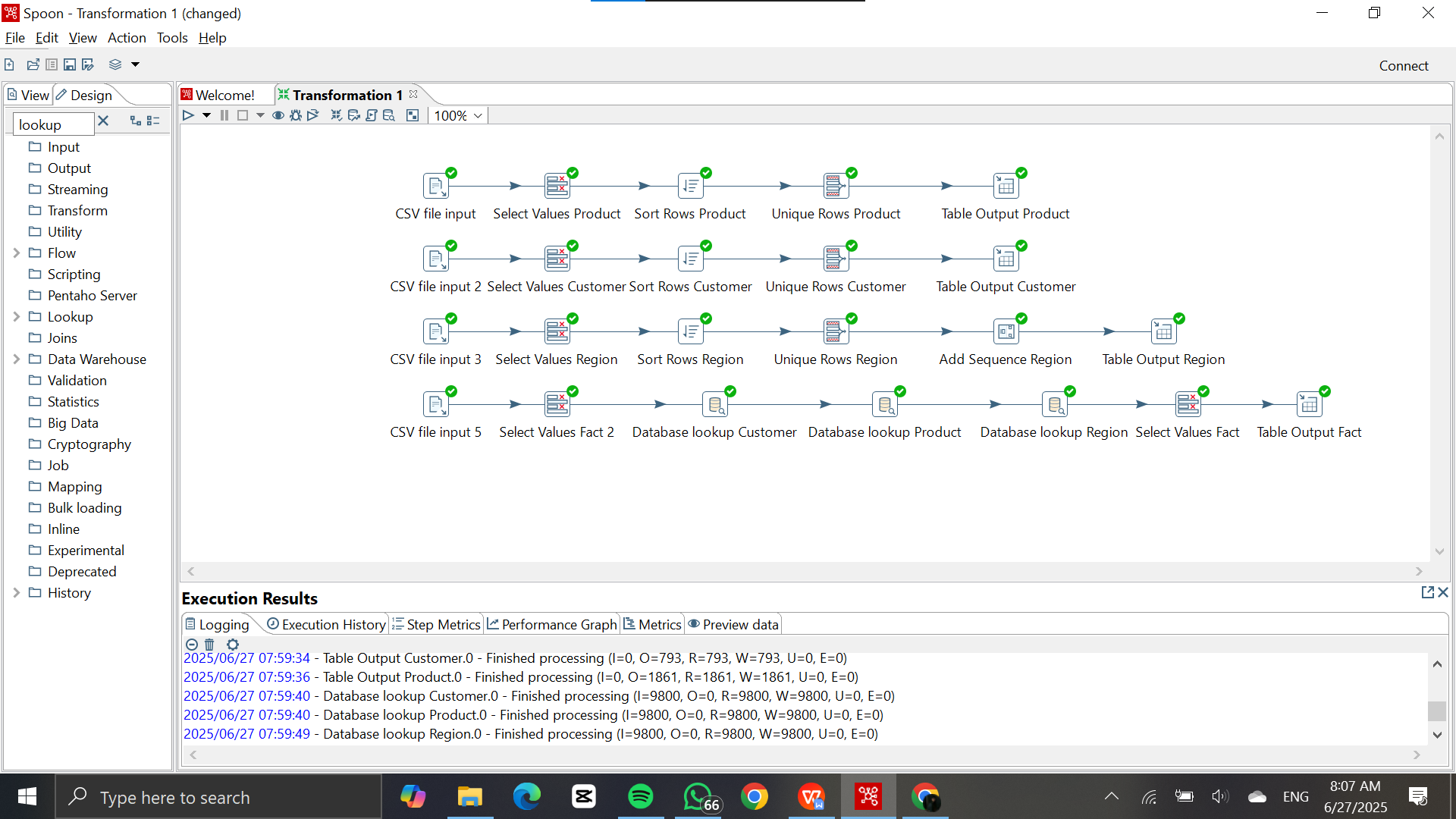
Solusi: Menambahkan step Database Lookup setelah dim\_region untuk mencari region\_id berdasarkan region\_name.

#### 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).

### Dokumentasi Pipeline ETL



Pipeline ETL terdiri dari:

1. 4 alur CSV file input ke masing-masing entitas.
2. Transformasi Select Values, Sort Rows, Unique Rows.
3. Lookup ke tabel dimensi sebelum mengisi tabel fakta.
4. Output ke database MySQL dw\_sales.

### Hasil dan Analisis

#### Hasil Akhir:

dim\_customer: 793 record

dim\_product: 1861 record

dim\_region: 600 record

fact\_sales: 9800 record berhasil masuk

### 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.