

Pertemuan 3

Persiapan dan Kualitas Data untuk Dashboard

3.1 OLTP vs OLAP: Mengapa Dashboard Tidak Membaca Database Produksi Secara Langsung

Sistem transaksional (OLTP — Online Transaction Processing) dirancang untuk menangani volume transaksi tinggi dengan latensi rendah per transaksi tunggal (mis. sistem kasir mencatat satu penjualan dalam hitungan milidetik). Struktur datanya sangat ternormalisasi untuk menghindari duplikasi dan menjaga integritas transaksi. Sebaliknya, sistem analitis (OLAP — Online Analytical Processing) dirancang untuk membaca dan mengagregasi jutaan baris data sekaligus guna mendukung analisis (mis. "total penjualan per bulan per kategori produk selama 3 tahun terakhir"). Menjalankan query analitis berat langsung pada database OLTP produksi berisiko dua hal serius: menurunkan performa sistem transaksional yang sedang melayani pelanggan secara real-time, dan struktur data OLTP yang ternormalisasi membuat query analitis menjadi kompleks dan lambat. Oleh karena itu, praktik standar industri adalah memisahkan data untuk kebutuhan analitis ke dalam data warehouse atau data mart yang terpisah dari sistem produksi.

3.2 Data Warehouse, Data Mart, dan Star Schema

Data warehouse adalah repositori data terpusat yang mengumpulkan data dari berbagai sistem sumber (penjualan, keuangan, SDM, dsb.), disusun khusus untuk mendukung pelaporan dan analisis lintas fungsi. Data mart adalah subset data warehouse yang berfokus pada satu area fungsional tertentu (mis. data mart penjualan). Pendekatan pemodelan yang paling umum untuk data warehouse/mart adalah star schema, terdiri atas: fact table (tabel yang menyimpan ukuran/metrik numerik pada tingkat granularitas tertentu, mis. tabel Penjualan berisi kolom jumlah unit dan nilai transaksi per baris transaksi) yang dikelilingi oleh beberapa dimension table (tabel yang menyimpan atribut deskriptif untuk memberi konteks pada fact table, mis. dimensi Produk berisi nama produk dan kategori, dimensi Waktu berisi tanggal/bulan/tahun, dimensi Cabang berisi lokasi dan wilayah). Sebagai ilustrasi konkret untuk Studi Kasus Retailindo: Fact_Penjualan (kolom: ID_Transaksi, ID_Produk, ID_Cabang, ID_Waktu, Jumlah_Unit, Nilai_Penjualan) berelasi one-to-many dengan Dim_Produk, Dim_Cabang, dan Dim_Waktu. Struktur ini memungkinkan agregasi cepat (mis. "total penjualan per kategori produk per bulan") karena tools BI dapat memanfaatkan relasi yang sudah terdefinisi dengan jelas, alih-alih melakukan join kompleks pada struktur ternormalisasi penuh.

3.3 Proses ETL/ELT: Tahapan dan Contoh Transformasi

ETL (Extract, Transform, Load) adalah proses memindahkan data dari sumber ke tujuan analitis melalui tiga tahap: Extract (mengambil data mentah dari sumber, mis. database transaksional, file Excel, API), Transform (membersihkan dan mengubah bentuk data agar siap dianalisis), dan Load (memuat data hasil transformasi ke tujuan, mis. data warehouse atau langsung ke model data tools BI). ELT (Extract, Load, Transform) adalah varian modern di mana transformasi dilakukan setelah data dimuat ke platform tujuan yang memiliki kapasitas komputasi besar (umum pada platform cloud data warehouse).

Contoh konkret transformasi yang umum dilakukan pada tahap Transform, menggunakan ilustrasi dataset transaksi penjualan yang "kotor":

Masalah pada Data Mentah	Contoh	Transformasi yang Dilakukan
Format tanggal tidak konsisten	"05/03/2025", "5 Maret 2025", "2025-03-05" muncul bercampur	Standardisasi seluruh kolom tanggal ke format ISO (YYYY-MM-DD)
Duplikasi baris	Transaksi yang sama tercatat dua kali akibat kegagalan sinkronisasi sistem	Deduplikasi berdasarkan kombinasi kunci unik (ID transaksi)

Nilai hilang (missing value)	Kolom kategori produk kosong pada sebagian baris	Diisi dengan nilai default "Tidak Terklasifikasi" atau dihapus jika proporsinya kecil dan tidak sistematis
Inkonsistensi penamaan kategori	"Elektronik", "elektronik", "ELEKTRONIK" dianggap tiga kategori berbeda oleh sistem	Standardisasi huruf besar/kecil dan pemetaan ke daftar kategori baku (master data)

3.4 Dimensi Kualitas Data dan Dampaknya terhadap Kepercayaan Pengguna

- Akurasi — data mencerminkan realitas sebenarnya (mis. angka penjualan sesuai dengan bukti transaksi fisik).
- Kelengkapan (completeness) — tidak ada data penting yang hilang secara sistematis.
- Konsistensi — nilai yang sama direpresentasikan secara seragam di seluruh sistem (mis. nama cabang yang sama tidak dieja berbeda-beda).
- Aktualitas (timeliness) — data tersedia cukup cepat untuk relevan dengan keputusan yang akan diambil (dashboard operasional harian menjadi tidak berguna jika datanya baru diperbarui seminggu sekali).
- Validitas — data mengikuti aturan/format yang telah ditentukan (mis. kolom persentase tidak berisi nilai di luar rentang 0–100).

Dampak buruknya kualitas data terhadap dashboard bersifat mendasar: sekali pengguna menemukan satu angka pada dashboard yang tidak sesuai dengan yang mereka ketahui secara langsung (mis. manajer cabang tahu penjualan hari ini lebih tinggi dari yang ditampilkan dashboard), kepercayaan terhadap keseluruhan dashboard dapat runtuh secara permanen ("data trust gap") — bahkan jika kesalahan tersebut kecil dan terisolasi. Oleh karena itu, validasi kualitas data harus dilakukan sebelum dashboard dirilis ke pengguna, bukan ditemukan oleh pengguna sendiri setelah go-live.

3.5 Ilustrasi Konseptual Power Query

Power Query (tersedia di Excel dan Power BI) adalah alat transformasi data berbasis antarmuka visual yang mencatat setiap langkah transformasi sebagai skrip (bahasa M) yang dapat diulang otomatis setiap kali data diperbarui (refresh). Alur kerja umum: (1) hubungkan ke sumber data, (2) tinjau pratinjau data dan identifikasi masalah kualitas, (3) terapkan transformasi langkah demi langkah (mis. hapus kolom tidak relevan, ubah tipe data, pisahkan kolom gabungan, gabungkan dua tabel dengan Merge Queries, tumpuk beberapa tabel dengan struktur sama menggunakan Append Queries), (4) muat hasil ke model data. Keunggulan pendekatan ini dibandingkan pembersihan manual di Excel adalah reproducibility — proses yang sama otomatis terulang saat data baru masuk, mengurangi risiko human error yang melekat pada pembersihan manual berulang.

Rangkuman

Dashboard yang andal memerlukan fondasi data yang tepat: dipisahkan dari sistem transaksional (OLTP) melalui data warehouse/mart bermodel star schema, melalui proses ETL yang menangani masalah kualitas data secara sistematis dan reproducible. Kepercayaan pengguna terhadap dashboard sangat rentan terhadap kesalahan data yang terdeteksi langsung oleh pengguna, sehingga validasi kualitas data adalah tahap yang tidak boleh dilewatkan.

Latihan / Pertanyaan Reflektif

1. Diberikan potongan dataset transaksi dengan minimal 3 jenis masalah kualitas data (dapat disiapkan dosen), identifikasi seluruh masalah dan usulkan langkah transformasi untuk masing-masing.
2. Mengapa star schema lebih disukai dibandingkan struktur data yang sepenuhnya ternormalisasi (3NF) untuk kebutuhan pelaporan BI, meskipun star schema mengandung duplikasi data pada dimension table?