Laporan Proyek Data Lakehouse - AdventureWorks

1. Manajemen Folder & File

Struktur folder proyek:

- data_lake/adventureworks/files/ : menyimpan file sumber (csv, pdf, txt)
- data_lake/adventureworks/tweets/ : menyimpan tweet mentah & hasil visualisasi
- database staging: warehouse_temp_sensor, market_share_report, external_sentiment
- database warehouse: fact_sales, dim_customer, dim_product, dim_time, dim_location

2. Desain Database

Database Staging:

- Tabel 'warehouse_temp_sensor': menyimpan data suhu dari gudang
- Tabel 'market_share_report': berisi hasil ekstraksi dari file PDF
- Tabel 'external_sentiment': tweet tentang 'adventureworks'

Database Warehouse (Star Schema):

- fact_sales: menyimpan metrik penjualan (misal dari file CSV gabungan)
- dim_customer, dim_product, dim_time, dim_location sebagai dimension tables

3. Jenis File & Analisis

- PDF: Ekstraksi teks laporan pasar (market share)
- CSV: Membaca data sensor suhu dan waktu
- TXT: Tweet teks diproses dan dibersihkan untuk wordcloud serta sentiment

4. Ringkasan Kode Utama

- data_lake_hans_ingest.py: memproses file tweet dan menyimpan dalam berbagai format
- staggingdb.py: mengimpor semua file (csv, pdf, txt) ke dalam database staging
- analysis.py: melakukan analisis data
- tweet_visualizer.py: menghasilkan visualisasi wordcloud

5. Diagram Star Schema

Laporan Proyek Data Lakehouse - AdventureWorks