Data Modeling dalam Data Engineering

Data modeling adalah proses mendesain struktur data agar dapat digunakan secara optimal dalam sistem database, warehouse, atau pipeline data.

Ini adalah fondasi penting dalam Data Engineering karena menentukan bagaimana data disimpan, diakses, dan diproses.

1. Pengertian Data Modeling

Data modeling adalah teknik untuk merepresentasikan hubungan antara berbagai entitas dalam suatu sistem informasi. Model ini membantu memastikan

bahwa data dapat digunakan secara efisien untuk analisis, pelaporan, dan pemrosesan lebih lanjut.

2. Jenis-Jenis Data Model

A. Conceptual Data Model (Model Konseptual)

- Representasi tingkat tinggi tentang bagaimana entitas dalam sistem saling berhubungan.
- Tidak spesifik terhadap database tertentu.
- Biasanya dibuat oleh Business Analyst atau Data Architect.
- Contoh: ERD (Entity-Relationship Diagram).

B. Logical Data Model (Model Logis)

- Menerjemahkan model konseptual ke dalam format yang lebih detail.
- Menunjukkan tipe data, primary key, foreign key, constraints, dan normalisasi.
- Masih bersifat independen dari implementasi database tertentu.

C. Physical Data Model (Model Fisik)

- Implementasi konkret ke dalam sistem database tertentu seperti PostgreSQL, Redshift, Snowflake, BigQuery.
- Menyesuaikan dengan indexing, partitioning, clustering, dan optimasi query.
- Spesifik terhadap teknologi database yang digunakan.

3. Teknik Data Modeling

A. Normalization vs Denormalization

- Normalization: Mengurangi duplikasi data dengan membagi tabel menjadi beberapa relasi.
- Denormalization: Menggabungkan tabel untuk mempercepat query.

B. Star Schema vs Snowflake Schema

Star Schema:

- Satu fact table dikelilingi oleh beberapa dimension table.
- Lebih cepat untuk query analitik.

Snowflake Schema:

- Pengembangan dari Star Schema dengan dimensi lebih terpecah.
- Lebih hemat penyimpanan, tapi query bisa lebih lambat karena butuh join lebih banyak.

4. Best Practices dalam Data Modeling

- Gunakan indeks dengan bijak untuk mempercepat query.
- Pilih tipe data yang optimal untuk efisiensi penyimpanan.
- Gunakan partitioning untuk menangani dataset besar.
- Gunakan caching jika perlu untuk mengurangi beban query pada database OLAP.

5. Data Modeling dalam ETL

Sebagai Data Engineer, Anda akan menerapkan data modeling dalam pipeline Extract, Transform, Load (ETL):

- Extract: Mengambil data dari berbagai sumber seperti API, CSV, Parquet, Database OLTP.
- Transform: Membersihkan dan mengubah data sebelum masuk ke warehouse.
- Load: Memasukkan data ke Data Warehouse (DWH) dalam format yang sesuai dengan Star atau Snowflake Schema.

6. Tools untuk Data Modeling

- Database: PostgreSQL, MySQL, Amazon Redshift, Snowflake, Google BigQuery.
- Data Modeling Tools: dbdiagram.io, ER/Studio, Lucidchart, PowerDesigner.
- ETL Tools: Apache Airflow, dbt (data build tool), Fivetran, Talend.
- Cloud Data Platforms: AWS (Glue, Redshift), GCP (BigQuery), Azure Synapse.

Kesimpulan

Data modeling adalah kunci utama dalam pekerjaan Data Engineer:

- Pahami struktur data dari Conceptual -> Logical -> Physical.
- Terapkan strategi Normalization atau Denormalization sesuai kebutuhan.

- Gunakan schema terbaik (Star/Snowflake) untuk data warehouse.
- Optimalkan performa database dengan indexing, partitioning, dan tipe data yang tepat.

Langkah Selanjutnya untuk Menjadi Data Engineer:

1. Kuasai SQL & Database Management (PostgreSQL, Redshift, Snowflake).
2. Latih diri dengan membuat schema untuk dataset besar.
3. Bangun pipeline ETL sederhana menggunakan Python & Apache Airflow.
4. Pelajari Cloud Data Warehousing (AWS/GCP/Azure).
5. Mulai implementasi proyek nyata dengan Data Modeling!