TUGAS MISI 3 PERGUDANGAN DATA
“Perancangan Data Warehouse untuk Analisis Customer Churn pada Industri Telekomunikasi”

KELOMPOK 8 DW RA

ANGGOTA KELOMPOK :	
EKSANTY F SUGMA ISLAMIATY	122450001
RESIDEN NUSANTARA R M	122450080
AISYAH TIARA PRATIWI	121450074
UKASYAH MUNTAHA	122450028
RENDRA EKA PRAYOGA	122450112

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2025

1. Design and Implement Dimension Tables

Dalam tahap awal pembangunan *data warehouse*, perancangan tabel dimensi merupakan langkah penting karena menyimpan informasi deskriptif yang melengkapi data numerik pada tabel fakta. Tabel fakta berisi data seperti total tagihan, biaya bulanan, atau status churn, dan tabel dimensi menjelaskan pelanggan, jenis layanan yang digunakan, metode pembayaran yang dipilih, dan lokasi tempat tinggal. Tabel dimensi digunakan untuk analisis lebih lanjut, seperti mengidentifikasi pelanggan yang paling sering berhenti langganan, pola perilaku pelanggan berdasarkan kontrak, atau pengaruh metode pembayaran terhadap loyalitas.

Agar mudah digunakan oleh berbagai pihak, seperti *data analys*, tim manajer, atau staf layanan pelanggan, struktur tabel dimensi perlu dirancang secara sederhana dan fleksibel untuk analisis dari berbagai sudut pandang, misalnya berdasarkan wilayah (kota atau kode pos), jenis layanan, atau kelompok usia dan lama berlangganan. Penambahan hierarki yang jelas, seperti pada dimensi waktu (hari → bulan → kuartal → tahun), untuk membantu melihat tren churn. Dengan desain yang rapi dan kontekstual, tabel dimensi menjadi fondasi penting dalam mendukung pengambilan keputusan berbasis data.

1.1 Merancang dan Mengimplementasikan Tabel Dimensi

Perancangan dimensi dimulai dengan mengidentifikasi atribut penting yang dapat menjawab kebutuhan bisnis, seperti:

- Siapa pelanggan yang *churn*?
- Apa jenis layanan yang digunakan pelanggan tersebut?
- Di mana pelanggan tinggal?
- Bagaimana metode pembayarannya?

Tiap dimensi dirancang memiliki hierarki untuk mendukung analisis mendalam pada analisis ini digunakan metode drill down untuk melihat data secara lebih rinci dari atas ke bawah.

Tabel dimensi diimplementasikan sebagai tabel dalam basis data, dengan struktur yang stabil dan jelas dengan ciri-ciri diantaranya yaitu setiap dimensi memiliki primary key, tipe data disesuaikan dengan jenis atribut seperti *int* untuk jumlah, dan relasi *many-to-one* dibangun dari tabel fakta ke setiap tabel dimensi.

Berikut adalah tabel dimensi yang dirancang:

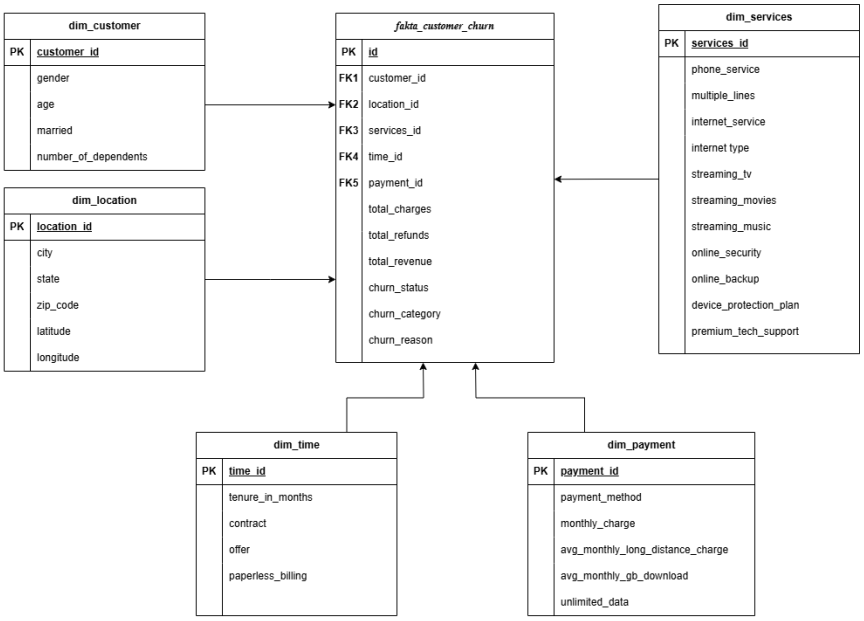
Nama Dimensi	Atribut	Deskripsi	Nilai
Dim_Customer	customer_id, gender, age, marital_status, dependents, tenure	Menyediakan informasi deskriptif tentang pelanggan.	54123, Male, 45, Married, 2, 24 bulan

Dim_Service	internet_service, phone_service, streaming_tv, contract_type	Informasi jenis layanan yang digunakan pelanggan.	Fiber, Yes, Yes, One year
Dim_Payment	payment_method, paperless_billing	Menyediakan informasi tentang cara pembayaran pelanggan.	Credit card, Yes
Dim_Location	city, zip_code, latitude, longitude	Menyediakan lokasi tempat tinggal pelanggan.	Los Angeles, 90001, 34.05, -118.25
Dim_Churninfo	customer_status, churn_category, churn_reason	Memberikan informasi status langganan dan alasan pelanggan berhenti.	Churned, Competitor, Found Better Deal
Dim_Date	date_key, day, month, quarter, year	Menyediakan konteks waktu saat data dicatat.	2024-09-01, 01, September, Q3, 2024

1.2 Merancang dan Mengimplementasikan Tabel Fakta

Dalam *data warehouse*, tabel fakta merupakan komponen utama yang menyimpan data numerik atau kuantitatif dari peristiwa atau kejadian bisnis. Tabel fakta merepresentasikan informasi penting seperti biaya langganan, total pendapatan, jumlah *refund*, hingga status *churn* pelanggan. Data ini bersifat terukur dan dapat dianalisis secara agregat, seperti total *revenue*, rata-rata biaya bulanan, dan perbandingan antara pelanggan *churn* dan aktif.

Sedangkan tabel dimensi berisi data deskriptif yang memberikan konteks bagi data di tabel fakta. Seperti, siapa pelanggan itu, di mana pelanggan tinggal, layanan apa yang digunakan, bagaimana cara pembayarannya, dan kapan pelanggan berhenti langganan. Kombinasi antara fakta dan dimensi digunakan untuk analisis dari berbagai perspektif seperti apakah *churn* lebih tinggi di kota tertentu, pada kelompok usia tertentu, atau pada jenis kontrak tertentu.



Dalam kasus ini digunakan *Star Schema* dengan tabel fakta berada di pusat, dan terhubung langsung ke setiap tabel dimensi. Model ini disederhanakan dan efisien untuk analisis OLAP, serta mudah dipahami oleh pengguna non-teknis. Dengan desain tabel fakta sebagai berikut:

Nama Atribut	Deskripsi
Customer_ID	ID unik pelanggan (foreign key ke Dim_Customer)
Date_Key	Tanggal observasi (foreign key ke Dim_Date)
Monthly_Charge	Biaya bulanan
Total_Charges	Total biaya sejak awal langganan
Total_Revenue	Total pendapatan yang dihasilkan pelanggan
Extra_Data_Charges	Biaya tambahan data
Long_Distance_Charges	Biaya panggilan jarak jauh
Total_Refunds	Total dana yang dikembalikan
Churn_Flag	1 = *churned*, 0 = aktif
ARPU	Pendapatan rata-rata per pelanggan (Total Revenue ÷ Tenure in Months)

Dengan pendekatan *star schema*, sistem ini tidak hanya mendukung analisis teknis oleh *data analyst*, juga dapat digunakan untuk pengambilan keputusan strategis oleh tim bisnis.

1.3 Struktur Aliran Data

a. Sumber data (*Data Source*)

Data terdapat pada file telecom_customer_churn.csv yang berjenis data pelanggan layanan telekomunikasi. Data ini berasal dari sistem CRM, billing, dan layanan teknis perusahaan.

b. Tahap pengolahan (*Processing Layer*)

Tahap pengolahan terdiri atas pra-pemrosesan dan transformasi. Pra-pemrosesan terdiri dari pembersihan data yaitu menangani *missing value* dan format numerik seperti Total Charges, encoding variabel kategorik seperti Gender, Contract, Payment Method, dan lain-lain.

Transformasi terdiri dari pembuatan fitur baru seperti rasio pengeluaran vs pendapatan, normalisasi atau standarisasi fitur numerik, enkripsi data untuk melindungi informasi sensitif pelanggan, dan integrasi data untuk menggabungkan data dari berbagai sumber sistem. Lalu data disimpan dalam format terstruktur yaitu CSV untuk analisis.

c. Penyimpanan sementara (*Staging Area*)

Data disimpan dalam DataFrame dalam SQL Table. Tujuannya untuk eksplorasi data, analisis, dan visualisasi.

d. Konsumen data (*Data Customer*)

- Tim *Data Analyst / Scientist* untuk menganalisis *churn* dan segmentasi pelanggan berdasarkan umur, lokasi, layanan yang digunakan.
- Model Machine Learning untuk memprediksi kemungkinan churn berdasarkan pola historis.
- Dashboard / Visualisasi untuk tim manajemen dan marketing agar memahami tren pelanggan.
- Strategi Retensi berdasarkan “Churn Reason” dan kategori churn.

2. Design and implement Fact Table

3. Design and implement indexes

4. Design Storage

5. Design and Implement Partitioned Tables and Views

Dalam konteks perancangan gudang data untuk analisis customer churn di industri telekomunikasi, penggunaan tabel yang dipartisi dan indexed views memainkan peran penting dalam mengelola data pelanggan dalam skala besar serta mempercepat kueri analitik berdasarkan periode waktu dan segmentasi pelanggan. Strategi ini ditujukan untuk mengoptimalkan pemrosesan data historis pelanggan, aktivitas langganan, serta perilaku pemakaian layanan yang dapat berubah secara dinamis.

5.1 Pemilihan Struktur Partisi

A. Partisi Horizontal (Row Partitioning)

Partisi horizontal diterapkan pada tabel fakta utama Fakta_Customer_Activity, berdasarkan atribut churn_year yang diperoleh dari kolom activity_date. Tujuan utama strategi ini antara lain:

- Mempercepat kueri waktu-spesifik, seperti analisis churn bulanan atau tahunan.
- Mempermudah manajemen data historis dengan memisahkan data berdasarkan tahun.
- Mengurangi beban query dengan hanya memproses subset data yang relevan.

Contoh Implementasi Partisi:

```
CREATE PARTITION FUNCTION pf_ChurnYear (INT)
AS RANGE LEFT FOR VALUES (2019, 2020, 2021, 2022, 2023);

CREATE PARTITION SCHEME ps_ChurnYear
AS PARTITION pf_ChurnYear ALL TO ([DW_Customer]);

CREATE CLUSTERED INDEX idx_activity_date
ON Fakta_Customer_Activity (activity_date)
ON ps_ChurnYear (churn_year);
```

B. Strategi Pemeliharaan Partisi

- Split Partition dilakukan setiap awal tahun baru untuk menyambut data periode berikutnya.
- Merge Partition digunakan untuk menggabungkan data lama yang sudah jarang diakses.
- Switch Partition dipakai untuk pemindahan batch data dari zona staging ke zona produksi dengan downtime minimal.

5.2 Perancangan Indexed Views

Untuk mendukung analisis churn berbasis agregasi, digunakan indexed views di zona Gold (Semantic Zone). View ini dirancang untuk menampilkan ringkasan data churn berdasarkan dimensi seperti wilayah pelanggan, waktu, dan jenis layanan.

Contoh View Agregasi:

```
CREATE VIEW vw_Churn_Per_Region_Year
WITH SCHEMABINDING
AS
SELECT
r.region_id,
w.churn_year,
COUNT_BIG(*) AS total_churn
FROM dbo.Fakta_Customer_Activity f
JOIN dbo.Dim_Region r ON f.region_id = r.region_id
```

```
JOIN dbo.Dim_Waktu w ON f.date_id = w.date_id
WHERE f.churn_flag = 1
GROUP BY r.region_id, w.churn_year;

CREATE UNIQUE CLUSTERED INDEX idx_churn_region_year
ON vw_Churn_Per_Region_Year (region_id, churn_year);
```

5.3 Strategi Penyimpanan Data

Arsitektur penyimpanan mengikuti prinsip Medallion Architecture:

- Bronze (Raw Zone): Data mentah dari sistem operasional disimpan di filegroup Staging_Filegroup, tanpa partisi.
- Silver (Clean Zone): Data yang telah dibersihkan dan ditransformasikan disimpan di PRIMARY dan DW_Customer, telah dipartisi dan dikompresi.
- Gold (Semantic Zone): View yang telah diindeks dan dioptimalkan untuk analisis churn disimpan di zona ini.

5.4 Optimasi dan Monitoring

- Kueri terhadap partisi spesifik dilakukan langsung ke partisi target, menghindari full table scan.
- Statistik dan indeks pada indexed views diperbarui secara berkala menggunakan UPDATE STATISTICS.
- Monitoring performa dilakukan dengan memanfaatkan DMV seperti sys.dm_db_partition_stats, sys.dm_db_index_usage_stats, dan sys.indexes.