Soal Project Data Warehouse

Note: Jika Anda mengalami masalah dalam mengakses project SSIS yang saya zip, Anda bisa mendapatkan project ini di GitHub saya (https://github.com/Ella-Raputri/MiniProject-Advance/tree/main/DE). Terima kasih.

1. Berdasarkan dengan tabel yang sudah dibuat pada Latihan SSIS, Jelaskan mengenai dengan poin berikut:
   a. Nama tabel Fact yang sudah dibuat di Data Warehouse

      Jawab:

      Tabel yang telah dibuat sebelumnya pada latihan SSIS (Netflix_Join) bukanlah tabel Fact yang sebenarnya, melainkan merupakan *denormalized staging table*. *Denormalized staging table* adalah tabel yang menggabungkan setiap atribut dari semua tabel sehingga dapat memberikan gambaran lengkap setiap show. Meskipun begitu, karena tabel ini merupakan hasil denormalisasi dari semua tabel yang ada, maka data yang ada menjadi redundan.

      Oleh sebab itu, maka diperlukan tabel baru sebagai tabel Fact untuk dataset ini. Untuk query yang saya gunakan untuk membuat tabel Fact dan Dimension dapat dilihat di file table_fact_dim.sql (di folder yang sama dengan dokumen ini). Nama tabel Fact yang saya buat adalah tabel “FactNetflixShow”.

   b. List kolom apa saja yang terdapat pada tabel fact yang sudah dibuat.

      Jawab:

      Tabel Fact adalah tabel yang berisi data kuantitatif atau fakta yang ingin dianalisis. Pada dataset ini, tidak banyak metrik kuantitatif yang bisa digunakan sehingga tabel Fact yang saya buat hanya berisi 4 kolom saja. Kolom-kolom yang ada pada tabel Fact yang saya buat adalah

      - ShowId
      - DurationMinutes
      - DurationSeasons
      - DateAdded

Dengan tabel Fact ini, show pada Netflix dapat dianalisis berdasarkan durasinya dan tanggal penambahan show ini ke dalam Netflix.

c. Tipe data yang terdapat pada setiap kolom

Jawab:

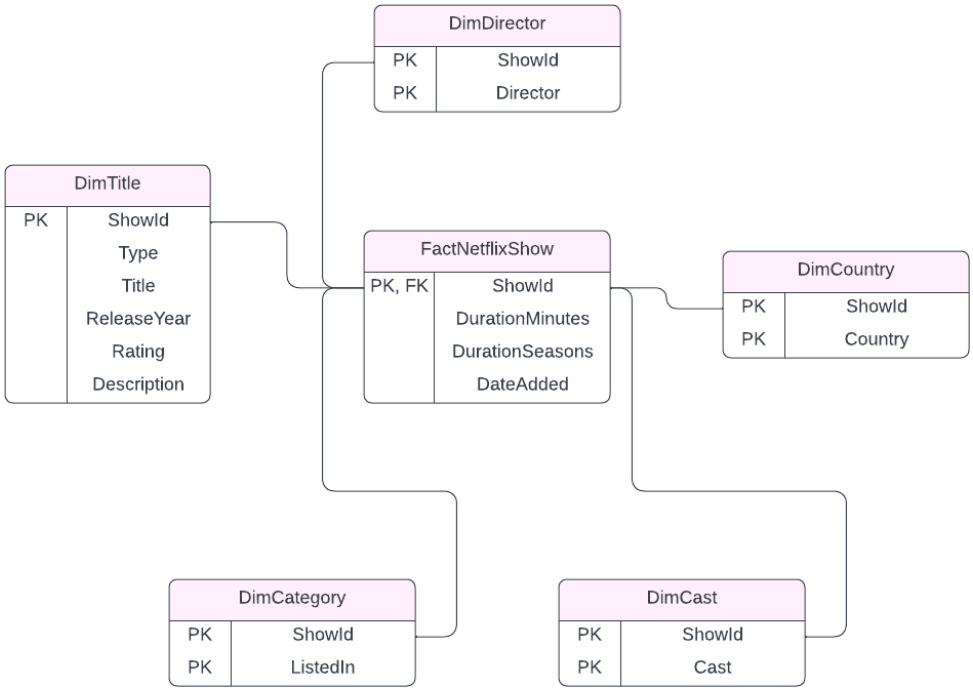
Berikut tipe data dari tiap kolom yang ada pada tabel Fact saya.

- ShowId, bertipe NVARCHAR
- DurationMinutes, bertipe INT
- DurationSeasons, bertipe INT
- DateAdded, bertipe DATETIME

d. Gambar Table Fact yang sudah dibuat

Jawab:

Hubungan antara tabel Fact dan tabel Dimension yang saya buat dapat dikatakan menggunakan Star Schema karena tabel FactNetflixShow dikelilingi oleh tabel-tabel Dimension lainnya. Berikut model gambar tabel Fact dan Dimension yang telah saya buat.



2. Kalian merupakan seorang data Engineer yang sedang bekerja di Perusahaan Netflix, dan tabel yang sudah kalian buat akan dipakai oleh dashboard yang digunakan oleh Board of Director Perusahaan. Berikut adalah kondisi mengenai penggunaan dashboard:
   a. Dashboard akan dipakai untuk keperluan meeting monthly. Namun anggota Board of Director kerap melihat dashboard setiap harinya untuk melihat update dari list film yang ditayangkan. Berdasarkan kondisi tersebut, jelaskan frekuensi refresh data yang akan digunakan pada tabel fact yang sudah dibuat beserta dengan alasannya.

      Jawab:

      Karena anggota Board of Director sering melihat dashboard setiap harinya, maka frekuensi refresh data yang ideal adalah setiap hari (*daily*). Hal ini dilakukan agar anggota Board of Directors bisa melihat perubahan yang ada pada perusahaan Netflix setiap harinya dan data yang ada dapat digunakan sebagai referensi untuk mengambil keputusan. Refresh data harian dapat menjaga data agar tetap *up-to-date* sesuai situasi yang ada sehingga keputusan bisnis yang tepat bisa diambil. Selain itu, untuk perusahaan besar seperti Netflix, penambahan atau pengurangan data kerap sekali terjadi sehingga refresh harian itu perlu agar tetap memastikan data yang ditampilkan di dashboard sesuai dengan kondisi yang ada.

   b. Setelah perundingan dengan tim aplikasi, diketahui bahwa sumber data akan ditambahkan status audit untuk menandai bahwa pada row data tersebut terjadi perubahan. Berdasarkan dengan informasi tersebut. Tentukan apakah metode yang dipakai pada saat transfer data menggunakan Incremental Load atau Full Refresh beserta dengan alasannya

      Jawab:

      Metode yang baik untuk melakukan transfer data berdasarkan informasi tersebut (setelah kita memiliki kolom status audit) adalah Incremental Load. Hal ini disebabkan karena dengan adanya kolom status audit yang menandakan ada tidaknya perubahan pada row, kita bisa menjadi tahu row mana yang berubah dan yang mana tidak. Oleh karena itu, kita hanya perlu mentransfer data yang diperbarui saja sejak transfer data terakhir sehingga waktu dan resource yang digunakan bisa lebih efisien. Full refresh kurang cocok dalam situasi ini karena full refresh biasanya digunakan jika ingin melakukan transformasi besar-besaran pada tabel atau melakukan perubahan yang signifikan pada tabel dengan alasan ketidakonsistenan data yang lama atau ingin menghapus data lama.

3. Perusahaan Netflix mempunyai 3 Business Unit yang berbeda, Divisi Operasional yang berhubungan dengan penyiaran film, Divisi Sales yang berhubungan dengan penjualan subscription, dan Divisi Legal yang mengurusi lisensi dari film yang beredar. Setiap divisi mempunyai data mart yang berisi tabel untuk menunjang kebutuhan tiap divisi. Dari kondisi tersebut, tentukan:
   a. Pada Data Mart mana tabel yang kalian buat berada, beserta dengan alasannya

      Jawab:

      Tabel yang saya buat itu berada pada Data Mart Divisi Operasional. Hal ini disebabkan karena Divisi Operasional memiliki tugas utama untuk menangani aktivitas penyiaran film sehingga data seperti nama film, tanggal ditambahkan ke Netflix, durasi film, dan lain sebagainya itu diperlukan. Tabel-tabel Dimension yang dibuat juga dapat membantu Divisi Operasional untuk melakukan analisis terhadap jumlah atau tren film yang ada dari tahun ke tahun. Selain itu, tabel yang saya buat dapat dikatakan kurang relevan untuk Divisi Sales dan Divisi Legal karena Divisi Sales lebih mementingkan data penjualan, sedangkan Divisi Legal lebih membutuhkan data tentang lisensi tiap film.

   b. Selain tabel yang sudah kalian buat, terdapat dua tabel lain bernama FactSubscriptionSales dan FactMovieLicense. Tentukan pada data mart mana kedua tabel tersebut berada.

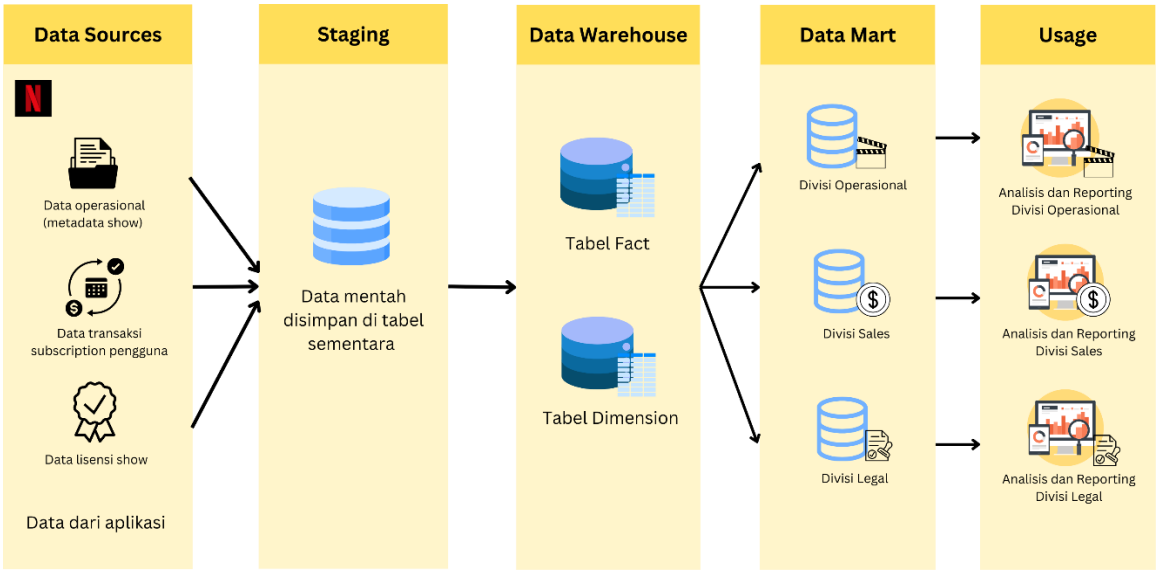
      Jawab:

      Tabel FactSubscriptionSales berada di Data Mart Divisi Sales, sedangkan tabel FactMovieLicense berada di Data Mart Divisi Legal. Divisi Sales memiliki tugas utama berhubungan dengan penjualan sehingga tabel Fact mengenai penjualan subscription tentunya penting bagi mereka. Sementara itu, Divisi Legal mengurusi hal-hal berhubungan dengan lisensi film sehingga memiliki tabel Fact tentang lisensi film merupakan hal yang sangat berguna bagi mereka.

   c. Buatlah Flow data dari Aplikasi sampai ke Business Unit.

      Jawab:

      Berikut gambar flow data dari aplikasi sampai penggunaan di tiap Business Unit.



Pada awalnya, data yang didapat dari sistem aplikasi Netflix dikumpulkan. Kemudian, data-data tersebut disimpan sementara di tabel staging. Pada tahap ini, juga akan dilakukan proses ETL (Extract, Transform, Load), yaitu Extract untuk mengambil data dari sumber-sumber data, Transform untuk membersihkan dan menformat data, dan Load untuk memuat data ke dalam Data Warehouse.

Setelah proses Load, data sekarang ada di data warehouse. Data warehouse merupakan repositori utama untuk menyimpan data secara terstruktur. Dalam case ini, data disimpan dalam model dimensional menggunakan tabel Fact dan Dimension. Lalu, dibuatlah data mart untuk setiap divisi.

Data mart merupakan subset dari data warehouse yang digunakan untuk memenuhi kebutuhan setiap divisi. Lalu data dari data mart akan digunakan oleh tiap divisi atau Business Unit untuk keperluan mereka masing-masing. Dengan data mart khusus untuk setiap divisi, Divisi Operasional dapat menganalisis tren film yang ada, Divisi Sales memantau pertumbuhan transaksi subscription, dan Divisi Legal memastikan lisensi film.