Menganalisis Kebutuhan Bisnis dan Teknis untuk Merancang Data Warehouse (DW) di Industri **Asuransi**

Disusun Oleh:

Kelompok 4
RA

Evan Aprianto	121450024
Kiwit Novitasari	121450126
Ibrahin Al-Kahfi	122450100
Meira Listyaningrum	122450011
Salwa Farhanatussaidah	122450055

PROGRAM STUDI SAINS DATA

FAKULTAS SAINS

INSTITUT TEKNOLOGI SUMATERA

2025

1. **Profil Industri dan Masalah Bisnis**

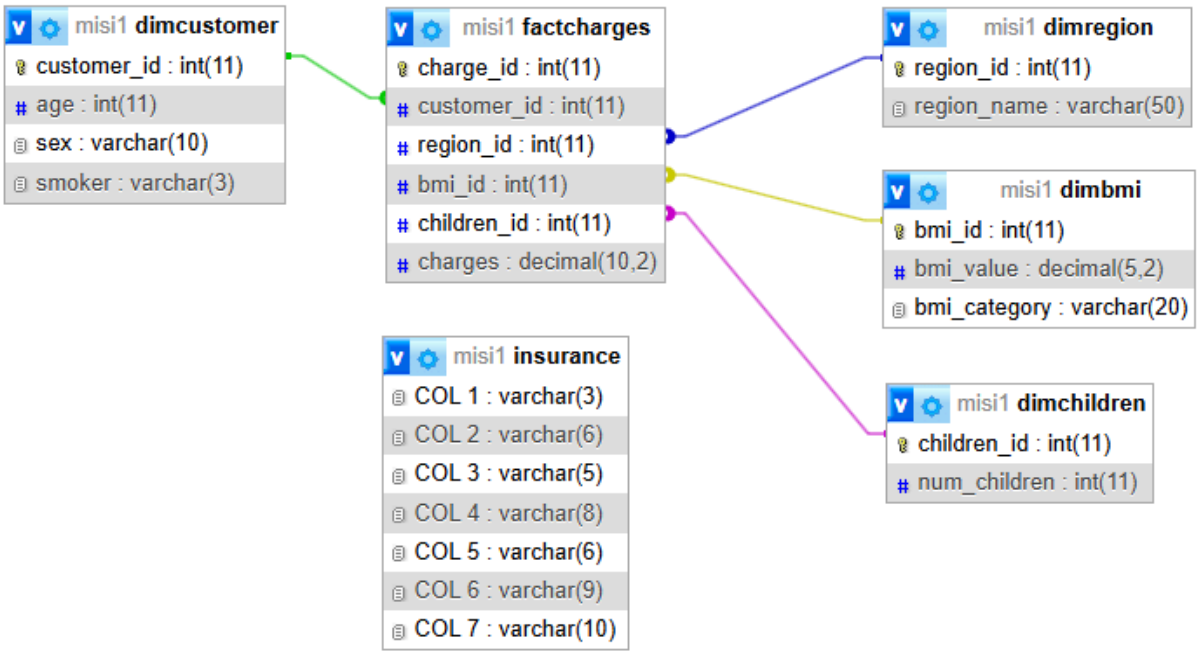
   Industri asuransi kesehatan memanfaatkan data pelanggan untuk menilai risiko dan menentukan premi. Namun, dalam persaingan yang semakin ketat, perusahaan asuransi menghadapi tantangan untuk mengelola dan menganalisis data pelanggan secara efisien. Masalah bisnis utama adalah bagaimana mengintegrasikan data pelanggan dalam bentuk yang terstruktur (data warehouse) untuk mendukung analisis biaya, segmentasi risiko, dan pengambilan keputusan berbasis data.

2. **Daftar Stakeholder & Tujuan Bisnis**

Stakeholder	Tujuan Bisnis
Manajer Keuangan	Mengetahui distribusi dan rata-rata biaya klaim di tiap wilayah
Analis Data	Membuat segmentasi pelanggan berdasarkan risiko (usia, perokok, BMI)
Manajer Operasional	Memantau jumlah anak tanggungan yang berpengaruh pada beban biaya
Eksekutif Perusahaan	Menentukan strategi premi dan kebijakan berdasarkan profil pelanggan
Tim Pemasaran	Menyesuaikan penawaran produk ke wilayah dan kelompok pelanggan tertentu

3. **Fakta & Dimensi**

Dalam sistem data warehouse untuk industri asuransi kesehatan, skema *star schema* digunakan untuk mempermudah analisis data terkait biaya dan profil pelanggan. Skema ini terdiri dari satu **tabel fakta** dan beberapa **tabel dimensi** yang saling terhubung melalui kunci asing (*foreign key*).



- FactCharges
  Tabel pusat yang berisi data biaya klaim atau premi (charges) dan menjadi dasar perhitungan metrik seperti total klaim dan rata-rata premi. Tabel ini terhubung ke tabel dimensi pelanggan, wilayah, BMI, dan jumlah anak.
- DimCustomer
  Menyimpan data pribadi pelanggan seperti usia, jenis kelamin, dan status merokok. Berguna untuk segmentasi risiko berdasarkan demografi dan gaya hidup.
- DimRegion
  Berisi informasi wilayah tempat tinggal pelanggan. Membantu analisis distribusi klaim berdasarkan lokasi geografis.
- DimChildren
  Mencatat jumlah anak tanggungan pelanggan. Informasi ini memengaruhi besaran premi dan risiko klaim.
- DimBMI
  Memuat nilai dan kategori indeks massa tubuh (BMI). Berguna untuk mengelompokkan risiko kesehatan pelanggan.

4. Sumber Data & Metadata

Deskripsi: Dataset berasal dari kaggle yaitu, sebuah dataset publik yang berisi informasi pelanggan asuransi kesehatan. Secara umum, dataset ini berisi informasi mengenai pelanggan asuransi kesehatan di Amerika Serikat, termasuk data demografis dan gaya hidup yang berkaitan dengan penentuan biaya klaim atau premi asuransi. Tujuan utama dari dataset ini adalah untuk membantu dalam analisis faktor-faktor yang mempengaruhi besarnya biaya asuransi yang harus dibayarkan oleh pelanggan.

COL 1	COL 2	COL 3	COL 4	COL 5	COL 6	COL 7
age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.70	0	no	northwest	21984.4706
32	male	28.88	0	no	northwest	3866.8552
31	female	25.74	0	no	southeast	3756.6216
46	female	33.44	1	no	southeast	8240.5896
37	female	27.74	3	no	northwest	7281.5056
37	male	29.83	2	no	northeast	6406.4107
60	female	25.84	0	no	northwest	28923.1369
25	male	26.22	0	no	northeast	2721.3208
62	female	26.29	0	yes	southeast	27808.7251
23	male	34.4	0	no	southwest	1826.843
56	female	39.82	0	no	southeast	11090.7178
27	male	42.13	0	yes	southeast	39611.7577
19	male	24.6	1	no	southwest	1837.237
52	female	30.78	1	no	northeast	10797.3362
23	male	23.84	0	no	northeast	2395.17155
56	male	40.3	0	no	southwest	10602.385
30	male	35.3	0	yes	southwest	36837.467
60	female	36.00	0	no	northeast	13228.8469
30	female	32.4	1	no	southwest	4149.736
18	male	34.1	0	no	southeast	1137.011
34	female	31.92	1	yes	northeast	37701.8768

Gambar 1. Dataset Insecure

Dataset ini terdiri dari 1.338 baris data, di mana setiap baris merepresentasikan satu pelanggan. Data yang tersedia cukup bersih karena tidak mengandung nilai kosong, sehingga bisa langsung digunakan untuk keperluan analisis dan pengolahan lebih lanjut. Adapun atribut-atribut yang terdapat dalam dataset meliputi:

- Age : Usia pelanggan, dinyatakan dalam satuan tahun.
- Sex : Jenis kelamin pelanggan, dengan nilai 'male' atau 'female'.
- BMI : Indeks massa tubuh pelanggan, yang mencerminkan rasio berat badan terhadap tinggi badan.
- Children : Jumlah anak yang menjadi tanggungan pelanggan.
- Smoker : Status merokok pelanggan, dengan nilai 'yes' untuk perokok dan 'no' untuk bukan perokok.

- Region : Wilayah geografis tempat tinggal pelanggan, seperti southeast, northwest, northeast, dan southwest.
- Charges : Biaya asuransi tahunan yang dikenakan kepada pelanggan, dinyatakan dalam satuan dolar AS.

Daftar Pustaka:

Kimball, R., & Ross, M. (2016). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (4th ed.). Wiley.

Dutta, S. & Rathi, P. (2019). "The Role of Data Analytics in Healthcare Insurance." *International Journal of Health & Medical Sciences*, 2(4), 9-15.

Batini, C., & Scannapieco, M. (2016). *Data Quality: Concepts, Methodologies and Techniques*. Springer.

Golfarelli, M., & Rizzi, S. (2009). "Conceptual Data Warehouse Design." *ACM Computing Surveys (CSUR)*, 41(4), 1-39.

Link Dataset:
https://www.kaggle.com/datasets/mirichoi0218/insurance?utm_source=chatgpt.com&select=insurance.csv