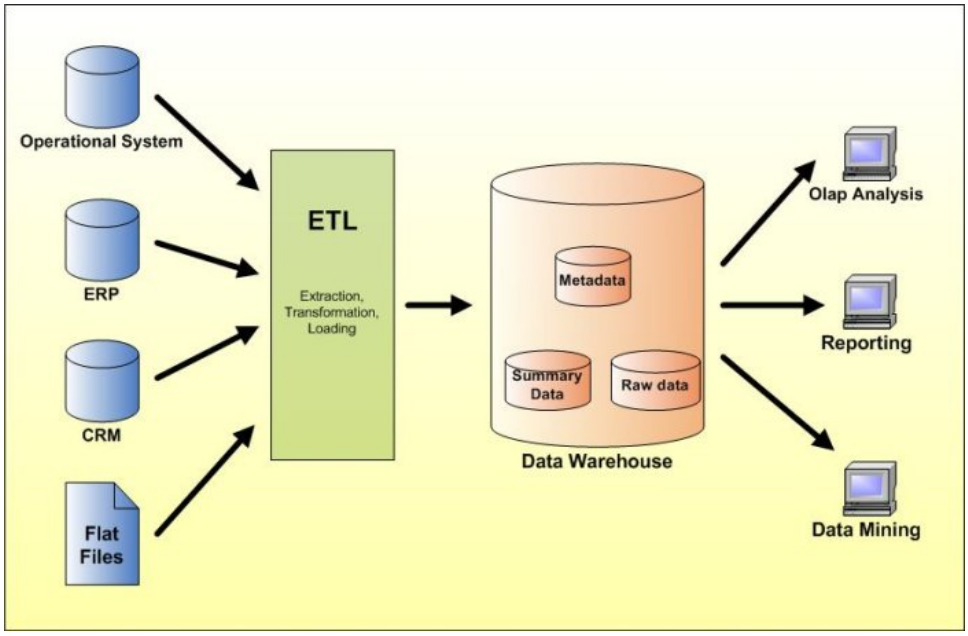
BAB III

TINJAUAN PUSTAKA

3.1 *Data Warehouse*

Data warehouse merupakan suatu bentuk penyimpanan data yang menerima data dari berbagai sumber berbeda dan disatukan sedemikian rupa dengan tujuan mendukung proses-proses analisa seperti *aggregation* dengan lebih cepat dan efisien. Konsep ini berlawanan dengan konsep penyimpanan data-data berupa transaksi, karena untuk menyimpan data-data transaksi, diperlukan *integrity checking* untuk memastikan data tersebut tercatat dengan benar semua atau tidak sama sekali.

Apabila diperlukan, sebuah *data warehouse* dapat dipartisi atau disegmentasi lebih jauh untuk menyesuaikan dengan kebutuhan per divisi dari suatu perusahaan. *Subset* dari *data warehouse* yang terspesialisasi ini dikenal dengan istilah *data mart*. Tujuan segmentasi tersebut adalah untuk mempercepat proses pengolahan data yang ingin dilakukan masing-masing divisi, dan terkadang dibutuhkan pemisahan data supaya data milik suatu divisi tidak bisa diakses oleh divisi lain.



Gambar 3.1. *Data Warehouse Model* [2]

3.2 *Multi-Dimensional Data Model*

Multi-Dimensional Data Model adalah konsep struktur data yang terdiri dari 2 atau lebih data-data (umumnya dalam bentuk *database table*) yang saling berhubungan satu sama lain. Pemodelan data seperti ini dibutuhkan dalam organisasi atau bisnis yang mengonsumsi banyak data yang hubungannya kompleks dalam jumlah besar dengan kebutuhan pemrosesan seperti aggregasi yang efisien.

3.2.1 Facts

Fact merupakan sebutan untuk data atau informasi yang berkaitan dengan kejadian-kejadian yang terjadi di dalam proses bisnis yang menghubungkan antara *measure* dan dimensi [3]. Penentuan *fact table* dilakukan dengan menentukan proses bisnis atau kejadian apa yang ingin direpresentasikan oleh data yang bersangkutan. Untuk model bisnis komersial, umumnya *fact table* mencakup

informasi-informasi seperti penjualan, pembelian, komplain dari pelanggan, dan lain sebagainya.

3.2.2 Measure

Measure merupakan data-data numerik yang akan diolah untuk dijadikan tolak ukur dari performa suatu bisnis. Umumnya data-data ini akan diaggregasi untuk mendapatkan gambaran besar dari performa suatu entitas dalam bisnis [4]. Dalam model bisnis komersial, beberapa contoh dari data yang tergolong *measure* adalah harga jual, harga beli, banyak produk terjual, jumlah transaksi, dan lain sebagainya.

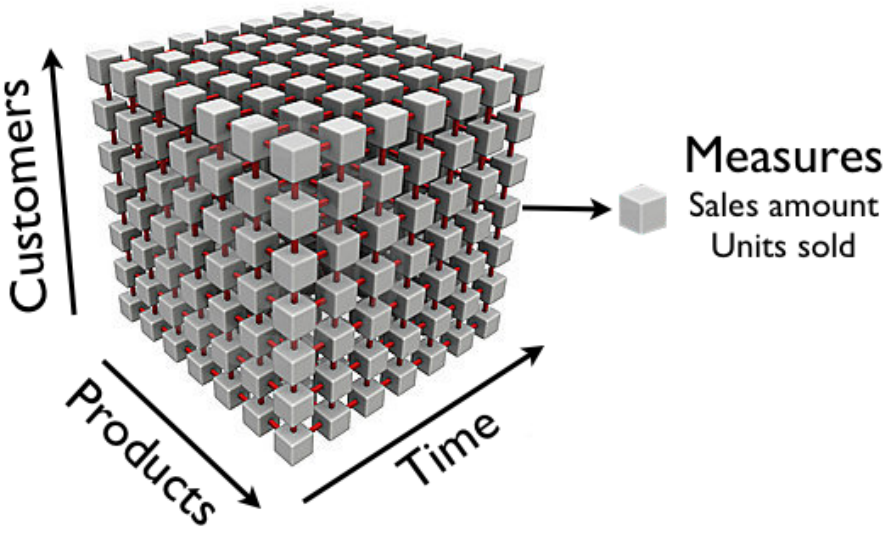
3.2.3 Dimensi

Dimensi merupakan istilah yang merujuk pada data atau informasi dasar dari entitas yang terlibat di dalam proses bisnis suatu perusahaan. Informasi dasar tersebut antara lain apa, siapa, di mana, kapan, mengapa, dan bagaimana suatu kejadian di dalam proses bisnis terjadi [5]. Data-data yang tersimpan di dalam dimensi umumnya jarang berubah baik dari segi struktur maupun nilai. Umumnya dalam proses perancangan *data warehouse*, di dalam setiap dimensi ditambahkan sebuah nomor yang tidak memiliki hubungan dengan data dalam dimensi tersebut untuk berperan sebagai *primary key*. Nomor ini berupa *counter* sekuensial yang dikenal dengan istilah *surrogate key*.

3.2.4 Data Cube

Data cube merupakan model atau susunan data yang tersusun atas 2 atau lebih dimensi dan 1 atau lebih *facts*. Susunan data seperti ini bertujuan merepresentasikan hubungan dan keterkaitan antara dimensi dan *fact*.

Pembentukan *data cube* ini akan dilanjutkan dengan pembuatan visualisasi berdasarkan *data cube* untuk menyajikan data dalam bentuk yang lebih informatif dan mudah dipahami.



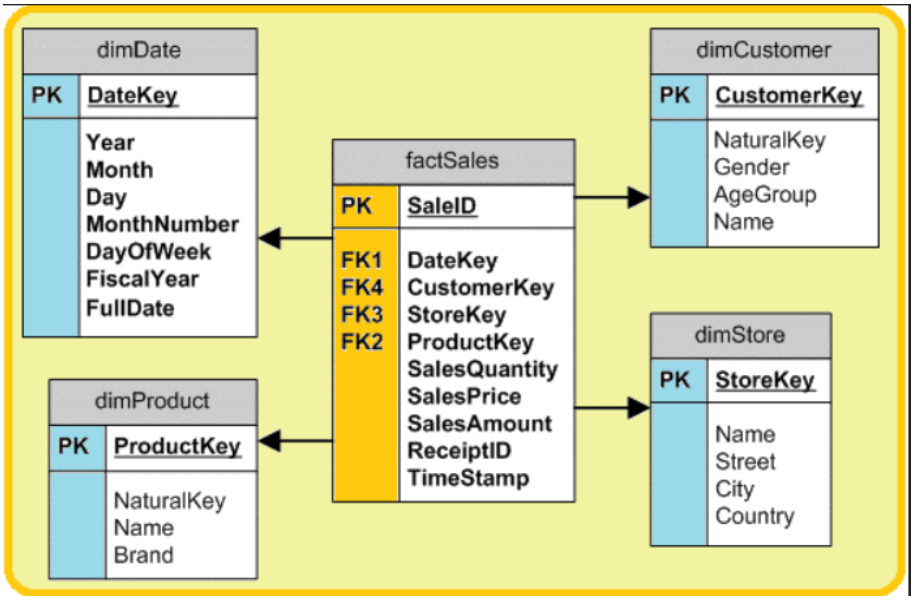
Gambar 3.2. *Data Cube Model* [4]

3.3 *Data Warehouse Schema*

Data warehouse schema merupakan rancangan struktur dari data-data yang akan disimpan dalam suatu *data warehouse*. Secara umum, skema yang digunakan dalam sebuah *data warehouse* ada 3, yakni *Star Schema, Snowflake Schema,* dan *Fact Constellation*.

3.3.1 *Star Schema*

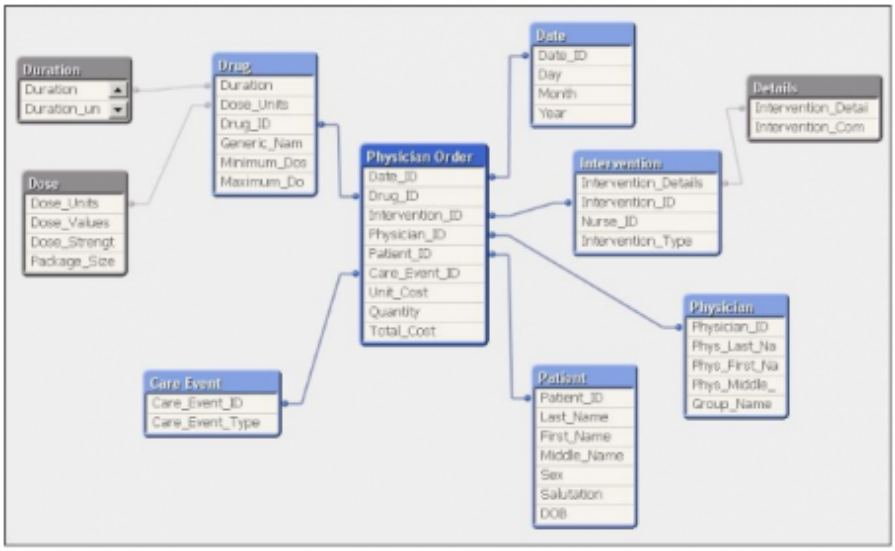
Star Schema menggambarkan relasi 1 tingkat antara tabel-tabel Dimensi dan 1 tabel *Fact* yang berada di pusat skema. *Star Schema* cenderung digunakan ketika analisa yang ingin dilakukan masih dalam tingkat yang sederhana dan tabel-tabel dimensi tidak ternormalisasi. Dari segi performa dan kecepatan pemrosesan, skema ini lebih baik dibandingkan skema satunya yakni *Snowflake schema*. [6]



Gambar 3.3. *Star Schema* [3]

3.3.2 *Snowflake Schema*

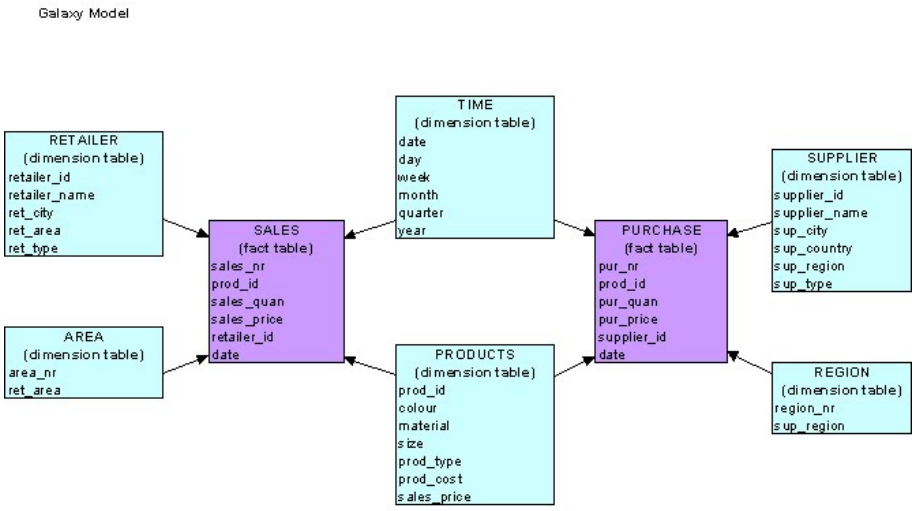
Snowflake Schema merupakan skema yang lebih kompleks, dengan relasi antara tabel-tabel Dimensi yang bisa lebih dari 1 tingkat dalam bentuk *lookup table*. *Snowflake Schema* digunakan ketika analisa yang ingin dilakukan terhadap data lebih rumit dan kompleks, seperti keperluan untuk segmentasi atau *filter* berdasarkan beberapa aspek sekaligus dan tabel-tabel dimensi yang terlibat sudah ternormalisasi. Akan tetapi, terlepas dari kompleksitasnya, skema ini lebih efisien dalam penggunaan memory dibandingkan *Star Schema*. [6]



Gambar 3.4. *Snowflake Schema* [6]

3.3.3 *Fact Constellation*

Fact Constellation merupakan skema yang melibatkan beberapa *fact table* yang terhubung secara tidak langsung melalui 1 atau lebih dimensi. *Fact Constellation* dapat digambarkan seperti kumpulan dari beberapa *Star/Snowflake Schema* dalam satu skema. Penggunaan *Fact Constellation* biasanya diperlukan ketika terdapat lebih dari 1 informasi yang diperlukan, tetapi informasi tersebut disimpan di dalam *fact table* yang berbeda.



Gambar 3.5. *Fact Constellation* [7]

3.4 *Extract Transform Load*

Proses *Extract, Transform, Load* atau dikenal dengan ETL merupakan serangkaian proses yang diberlakukan terhadap data-data dari berbagai sumber untuk mengambil datanya (*Extract*), menambahkan atau membentuk data menjadi lebih terstandarisasi dan efisien (*Transform*), kemudian menyimpannya ke dalam *Data Warehouse* (*Load*). [8]

Gambar 3.6. *ETL Process* [8]

3.4.1 Pentaho Data Integration (PDI)

Pentaho Data Integration (PDI) atau dikenal dengan nama *Kettle* merupakan salah satu *tool ETL open-source* yang dikembangkan oleh Pentaho yang banyak digunakan ETL Developer. Tingginya popularitas dari *tool* ini dikarenakan kemudahan dalam penggunaannya (*drag-and-drop*) serta proses pembelajarannya yang cukup singkat. Meskipun mudah dan cukup sederhana untuk digunakan, Pentaho Data Integration memiliki beragam fitur yang dapat digunakan untuk menerima data dari berbagai sumber data dan mengolahnya dengan berbagai cara untuk kemudian dikeluarkan dalam berbagai bentuk [9]. Sifat *open source* dari *tool* ini juga menjadi nilai tambah bagi *ETL Developer*. Selama masa kerja magang, penulis menggunakan *tool* Pentaho Data Integration untuk memenuhi permintaan klien.

Gambar 3.7. Pentaho Data Integration [10]

3.4.2 *Job*

Job merupakan istilah untuk susunan alur untuk proses-proses transformasi yang harus berjalan dengan urutan tertentu. Proses-proses yang dapat dijalankan di dalam sebuah *job* meliputi proses validasi, *transformation*, *file management*. dan utilitas sederhana yang tidak terlalu berkaitan dengan pengolahan data secara masif.

3.4.3 *Transformation*

Transformation merupakan istilah untuk serangkaian proses yang dijalankan secara paralel. Proses-proses yang dapat dijalankan dalam

transformation cenderung terkait dengan pemrosesan dan pengolahan data secara masif, misalnya pengambilan atau pengeluaran data, pemecahan, penggabungan, penambahan atau pengurangan *field*, dan transformasi data umum seperti proses kalkulasi atau mapping. Oleh karena itu proses-proses ini cenderung dijalankan secara paralel untuk mengoptimasi *runtime*.

3.5 Sumber Data

Sumber Data merupakan produk atau teknologi, umumnya berupa perangkat lunak, yang menghasilkan data yang dapat diproses oleh *ETL tool*. *ETL tool* yang baik akan dapat melakukan pemrosesan data dari berbagai jenis sumber. Sumber-sumber data yang dibahas dalam laporan ini merupakan sumber data yang lazim digunakan di dunia industri dan merupakan sumber data yang digunakan oleh penulis selama masa kerja magang.

3.5.1 Microsoft Excel

Microsoft Excel merupakan aplikasi *spreadsheet* milik perusahaan Microsoft yang lazim digunakan untuk keperluan penyimpanan data dan kalkulasi sederhana dalam bentuk tabel. Sebagai salah satu perangkat lunak besutan Microsoft, Excel banyak digunakan oleh perusahaan-perusahaan yang menggunakan ekosistem Microsoft dalam proses bisnisnya.

Gambar 3.8. Microsoft Excel [11]

3.5.2 MongoDB

MongoDB merupakan salah satu basis data NoSQL *open source* yang paling banyak digunakan para pengembang aplikasi. *Use case* dari penggunaan MongoDB umumnya meliputi katalog dan *realtime analysis* dari data-data perusahaan. [12] Penggunaan MongoDB dalam laporan ini adalah sebagai basis data salah satu dari banyak aplikasi yang digunakan dalam proses bisnis.

Gambar 3.9. MongoDB [12]

3.5.3 MySQL

MySQL merupakan *relational database* sederhana yang banyak digunakan. Tingginya *userbase* dan popularitas dari RDBMS ini tidak lain karena kemudahan dan kompatibilitasnya dengan berbagai platform untuk menyesuaikan dengan kebutuhan pengguna. Penggunaan MySQL dalam laporan ini adalah sebagai basis data salah satu dari banyak aplikasi yang digunakan dalam proses bisnis.

Gambar 3.10. MySQL [13]

3.5.4 Microsoft SQL Server

Microsoft SQL Server merupakan *relational database* yang dikembangkan oleh perusahaan Microsoft yang memiliki keterkaitan erat dengan *tool-tool* yang dapat digunakan untuk pemrosesan data lebih lanjut, seperti SQL Server Analytic Services, SQL Server Data Tools, Microsoft Power BI, dan Microsoft Dynamics AX. Oleh karena itu, banyak perusahaan-perusahaan yang menggunakan sistem operasi Windows dan menggunakan perangkat lunak berbasis Microsoft cenderung menggunakan Microsoft SQL Server untuk bisa memiliki ekosistem yang sama pada seluruh tahapan bisnis mereka. Dalam laporan ini, Microsoft SQL Server digunakan sebagai basis data untuk menampung *data warehouse*, dan *tools* dari SQL Server ini juga digunakan untuk pembentukan *data mart* dan *data cube*.

Gambar 3.11. Microsoft SQL Server [14]