BAB II

LANDASAN TEORI

2.1 Data Visualization

Data Visualization dapat membuat pengambilan keputusan lebih atraktif dan mudah untuk dimengerti. *Data visualization* adalah teknologi yang mendukung visualisasi dan interpretasi dari data-data yang ada dan informasi pada beberapa titik di sepanjang rangkaian *data processing*. Visualisasi dapat diinterpretasikan dalam bentuk gambar *digital*, GIS (Geographic Information System), grafik, *virtual reality*, presentasi dimensional, *graphical user interface*, video dan animasi (Turban, 2007).

Ada dua tujuan utama dari data visualisasi menurut Chart (tanpa tahun) dalam *White Paper Princeiples of Data Visualization – What We See in a Visual, yaitu:*

a. *Explain data to solve specific problems*: visualisasi dapat membantu pengguna seperti untuk mengambil keputusan, menjawab sebuah pertanyaan dan menyampaikan informasi pada suatu masalah tertentu.
b. *Explore large data sets for better understanding*: *exploratory visuals* akan memberikan banyak dimensi dari kumpulan data, atau membandingkan set data dengan data lain, sehingga dapat menarik pengguna untuk mengeksplor visual tersebut, timbul pertanyaan-pertanyaan selama proses, dan menjawab setiap pertanyaan yang ada.

2.2 Database

Menurut Connolly dan Begg (2005), *database* adalah sekumpulan data yang saling berhubungan secara logis, dan dibuat untuk memenuhi kebutuhan informasi dari suatu organisasi.

Sedangkan menurut Kristanto (2004), *database* adalah sekumpulan *file-file* yang saling terkait antara satu *file* dengan *file* lain sehingga membentuk satu rangkaian data untuk menginformasikan suatu perusahaan instansi dalam batasan tertentu.

Berdasarkan dari pengertian beberapa ahli di atas maka dapat disimpulkan *database* adalah sekumpulan data yang saling terkait satu dengan lainya dan berguna untuk memberikan kebutuhan informasi dari suatu organisasi atau perusahaan.

2.3 Database Management System (DBMS)

Database Management System atau DBMS merupakan sebuah sistem yang memungkinkan penggunanya untuk, membuat, mengatur, dan mengontrol akses ke dalam *database* (Connolly, 2010).

Berdasarkan Kimball dan Ross (2002), *Database Management System* adalah sistem yang bertujuan untuk menyimpan, mengambil, dan mengubah data secara terstruktur.

Database Management System adalah menajemen data untuk membantu melakukan pengolahan data dengan cepat (Muchtar, 2010).

Berdasarkan pengertian di atas maka dapat disimpulkan *Database Management System* adalah sebuah sistem manajemen data yang berfungsi untuk membantu mengolah data seperti membuat, mengatur, mengontrol akses ke dalam *database* secara cepat dan terstruktur.

2.4 Data Warehouse

Menurut Turban (2007), *data warehouse* adalah sekumpulan data yang dihasilkan untuk mendukung keputusan, dan juga merupakan repositori dari data saat ini dan juga data yang lalu yang berpotensi dan menarik bagi Manajer di seluruh organisasi.

Menurut Darudiato (2008) dalam jurnalnya menyatakan, *data warehouse* adalah sekumpulan informasi yang diperoleh dari *database* operasional yang akan digunakan untuk membuat *business intelligence* yang mendukung aktivitas analisis bisnis dan juga pengambilan keputusan.

Menurut Youssef (2011), *data warehouse* adalah sekumpulan data yang berorientasi subjek, terintegrasi, *time-variant* dan *non-updateable* yang digunakan dalam pengambilan keputusan oleh pihak menejerial.

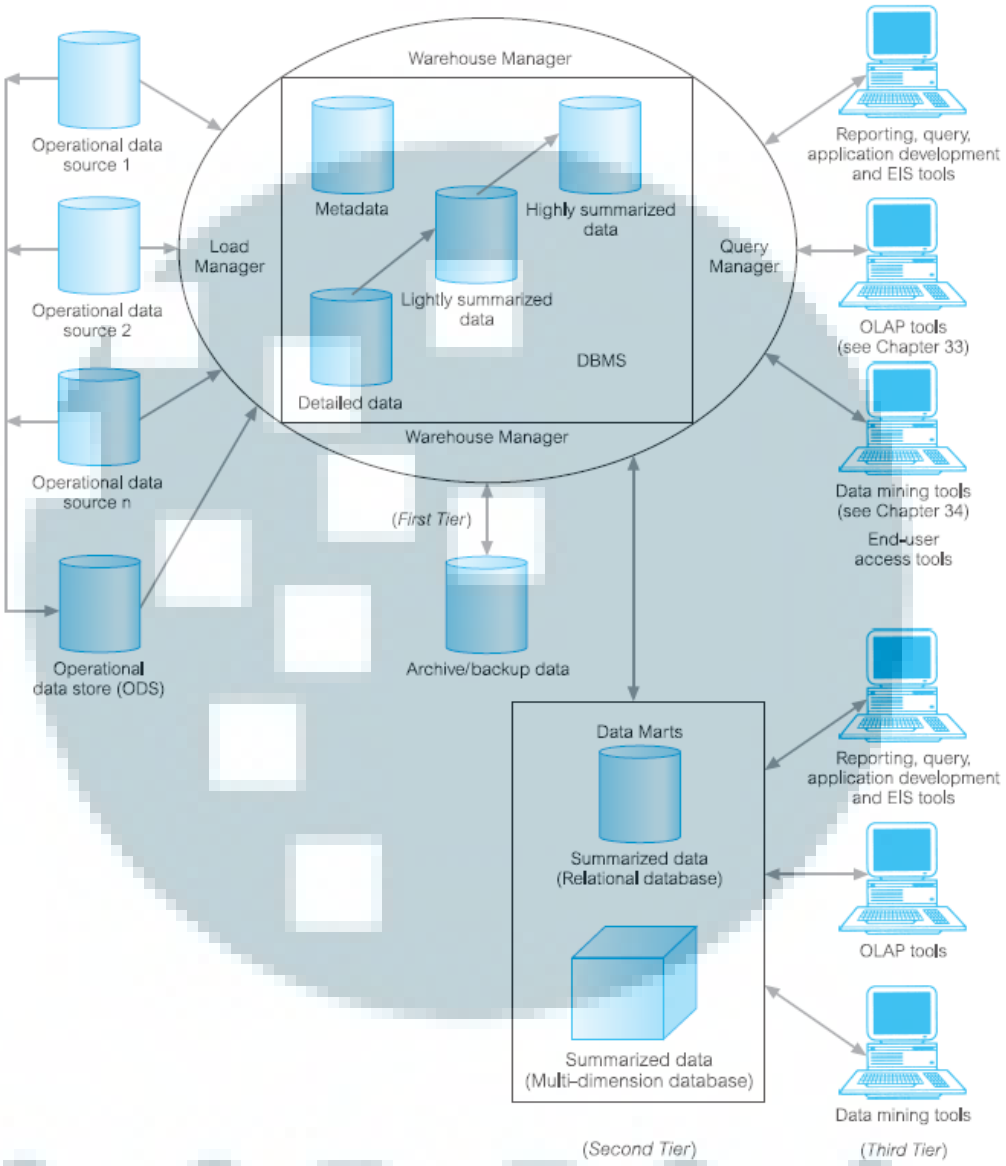
Berdasarkan definisi oleh para ahli di atas, dapat disimpulkan *data warehouse* merupakan sekumpulan data yang berorientasi subjek, terintergrasi, *time-variant*, dan *non-updateable* yang diperoleh dari basis data operasional yang berguna untuk membuat *business intelligence* untuk mendukung pengambilan keputusan oleh pihak manajemen atau pengambil keputusan.

2.5 Data Mart

Menurut Connolly dan Begg (2005), *data mart* merupakan bagian kecil dari *data warehouse* yang berkaitan dengan tingkat departemen atau fungsi bisnis tertentu dalam sebuah perusahaan. Berikut ini adalah karakteristik yang membedakan antara *data mart* dan *data warehouse*:

a. *Data mart* berfokus hanya pada kebutuhan pada tingkat departemen atau fungsi bisnis.
b. *Data mart* biasanya tidak mengandung data operasional yang rinci seperti pada *data warehouse.*
c. *Data mart* hanya memiliki sedikit informasi dibandingkan dengan *data warehouse*, *data mart* lebih mudah dipahami dan dinavigasi.

Data yang ada di dalam *data warehouse* dapat dibagi sesuai dengan kebutuhan dalam informasi. Inilah yang disebut *data mart*. *Data mart* memiliki karakteristik yang sama dengan *data warehouse*, perbedaannya hanya terdapat pada jumlah data yang digunakan. Dalam *data mart*, data yang digunakan berasal dari satu bagian atau departemen saja, sedangkan *data warehouse* data yang digunakan berasal dari seluruh bagian perusahaan. Berikut ini adalah gambaran arsitektur *data warehouse* dan *data mart.*



Gambar 2. 1 *Typical Data Warehouse and Data Mart Architecture*

Sumber: *Database System: A Practical Approach to Design, Implementation, and Management*

Perbedaan antara *data warehouse* dengan *data mart* dapat dilihat dalam tabel berikut.

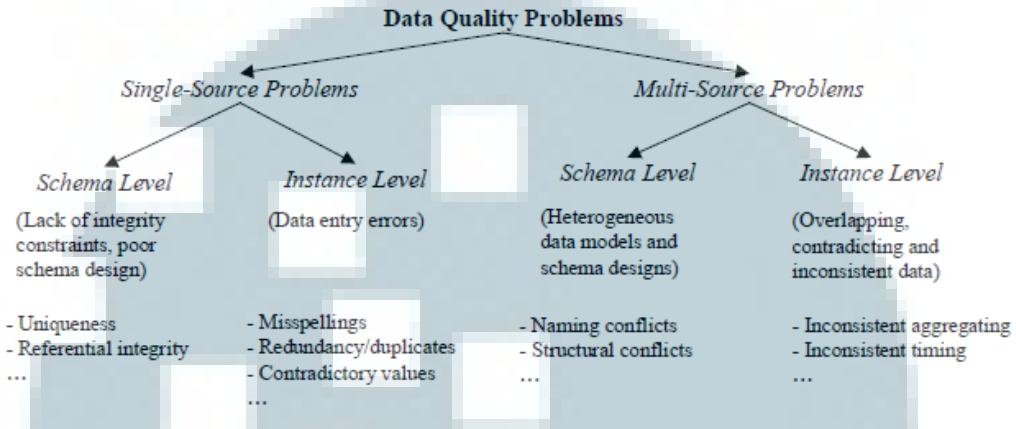
Tabel 2. 1 Perbedaan *Data Warehouse* dan *Data Mart*

	Data Warehouse	*Data Mart*
Ruang lingkup (*scope*)	Perusahaan/*Enterprise*	*Departmen*t
Subjek (*subject*)	*Multiple*	*Single*
Sumber Data (*Data Source*)	Banyak	Sedikit
Ukuran Data	100 GB – 1TB	<100 GB
Waktu Implementasi	Berbulan- bulan bahkan bertahun tahun	Beberapa bulan

2.6 Data Cleansing

Berdasrkan Rahm (2000) *Data cleaning* atau juga yang disebut dengan *data cleansing* atau *scrubbing* adalah mendeteksi, menghilangkan *error* dan ketidakkonsistenan pada data untuk meningkatkan kualitas dari data tersebut. Masalah kualitas data akan muncul pada saat mengintegrasikan satu sumber data atau dari banyak sumber data. Masalah yang sering terjadi adalah kesalahan ejaan penulisan, informasi yang tidak lengkap, dan data yang tidak valid. Semakin banyak sumber data yang akan diintegrasikan maka kebutuhan *data cleansing* akan semakin tinggi. *Data cleansing* dilakukan sebelum data dari operasional *database* dimasukkan ke dalam *data warehouse*.

Terdapat masalah utama dalam *data cleansing* yang dapat dilihat pada skema berikut:



Gambar 2. 2 Klasifikasi Masalah Kualitas Data Dalam Sumber Data

Sumber: *Data Cleaning: Problems and Current Approaches*

Secara umum, *data cleansing* melibatkan beberapa fase yaitu:

a. *Data analysis*: berguna untuk mendeteksi berbagai *error* dan ketidakkonsistenan yang akan dihilangkan.
b. *Definition of transformation workflow and mapping rule*: tergantung kepada jumlah sumber data, dengan tingkat dari heterogenitas dan kekotoran data, besarnya data transformasi dan langkah pembersihan mungkin harus dijalankan.
c. *Verification*: ketepatan dan keefektifan dari *transformation workflow* dan *transformation definition* harus dites dan dievaluasi.
d. *Transformation*: mengeksekusi *transformation step* dengan menjalankan ETL *workflow* untuk me-*load* dan memperbarui *data warehouse*.

e. *Backflow of cleaned data*: setelah *error* dihilangkan, data yang telah bersih harus menggantikan data kotor pada sumber data asli untuk memberikan data yang telah ditingkatkan kepada aplikasi dan untuk menghindarkan berulangnya pembersihan data ke depannya.

2.7 ETL

Extraction, Transformation, and Load (ETL) adalah salah satu proses pada *data warehousing*. Proses dari ETL adalah mengumpulkan data dari berbagai macam sumber. ETL adalah proses untuk mengolah data menjadi data yang bersih sesuai dengan ketentuan *data warehouse*. Proses ETL pada umumnya terdiri dari berbagai macam aktivitas dan membutuhkan waktu serta memori yang cukup besar (Dharayani, 2015).

Menurut Vassiliadis, et al (2002) ETL adalah bagian dari *software* yang bertugas untuk mengekstrak data dari berbagai jenis sumber data, melakukan pembersihan data, kostumisasi dan memasukkannya ke dalam *data warehouse.*

Proses ETL juga memakan waktu hampir 80% untuk satu proyek *data warehouse*. Tahapan-tahapan yang akan dilalui oleh ETL *designer* dalam membuat sebuah *data warehouse* adalah sebagai berikut:

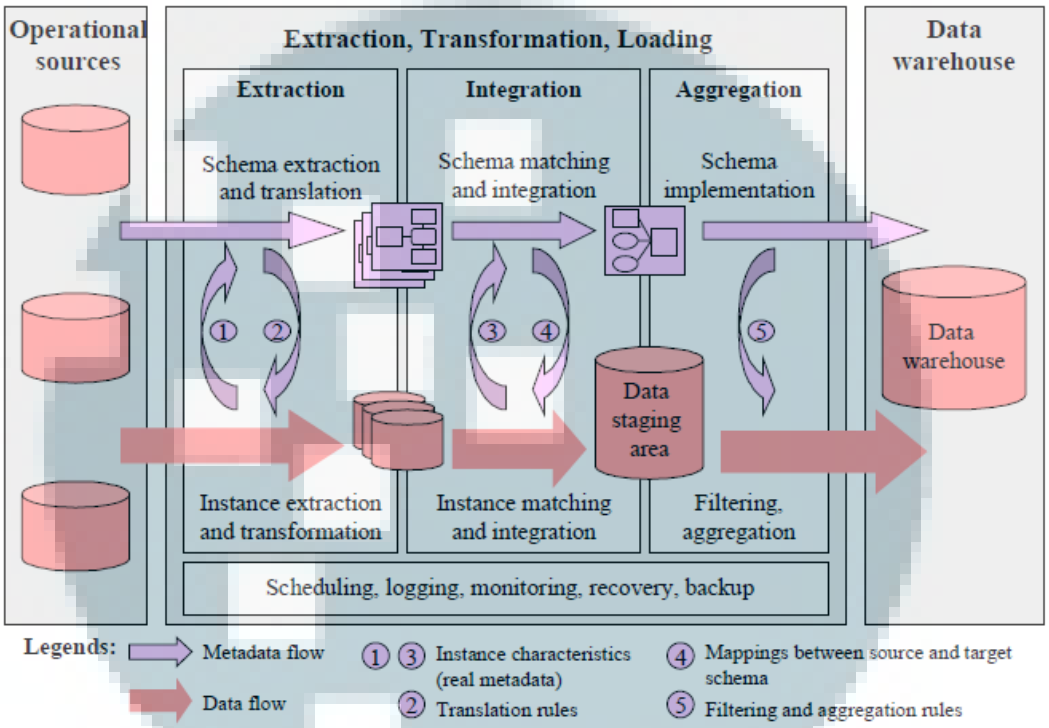
a. *Identification of the proper data stores*: mengidentifikasi sumber data apa saja yang berhubungan dengan *data warehouse* yang akan dibuat.
b. *Candidates and active candidates for involved data stores*: selama periode analisis desainer bisa menentukan lebih dari satu sumber data yang menjadi kandidat dari *data warehouse* yang akan dibuat.

c. *Attribute mapping between the providers and customer*: tugas sulit yang akan ditemui desainer adalah bagaimana cara menghubungkan atribut (*mapping*) dari sumber data yang beraneka ragam ke dalam satu *data warehouse*. Dalam langkah ini akan melibatkan orang-orang yang berwenang atas sumber data tersebut untuk mengetahui hal-hal implisit yang tersembunyi dari sumber data tersebut seperti kode, *rules*, dan *values*.

d. *Annotating the diagram with runtime constraint*: bagaimana *maping* dari semua sumber data hingga ke dalam *data warehouse* dijalankan. Berikut ini beberapa parameter yang dibutuhkan untuk menjalankan transformasi yaitu:

- *Time/Event based scheduling*: menentukan frekuensi proses ETL dijalankan.
- *Monitoring*: informasi *online* mengenai *progress/status* dari proses yang dijalankan, sehingga *administrator* dapat memantau tahap apa saja yang sedang dijalankan, kapan mulai, berapa lama durasinya dan sebagainya.
- *Logging*: informasi *offline* yang menunjukan hasil dari keseluruhan proses yang terjadi apakah ada *error* atau tidak.
- *Exception handling*: seberapa besar tingkat *error* yang masih bisa ditoleransi atau diterima.
- *Eror handling*: *crash recovery* dan kemampuan untuk memulai dan memberhentikan secara manual dan juga hal yang dibutuhkan untuk menangani *error* yang terjadi saat proses berlangsung.

Berikut ini adalah gambar mengenai proses ETL dari sumber data asli hingga ke dalam *data warehouse*:



Gambar 2. 3 Proses ETL

Sumber : *Data Cleaning: Problems and Current Approaches*

2.8 Spoon Kettle

Menurut wibisono (2012), Pentaho Kettle adalah sebuah perangkat lunak *open source* yang dikeluarkan oleh *Pentaho*. Perangkat lunak ini dapat digunakan sebagai *tools* untuk mengintegrasikan data. *Pentaho Kettle* menyediakan fasilitas untuk melakukan Proses ETL (Extraction, Transformation dan Loading). Terdapat tiga komponen utama dalam Pentahoo Kettle yaitu Spoon, Pan dan Kitchen. Spoon merupakan *user interface* untuk membuat *job* dan *transformation.* Pan adalah *tools* yang berfungsi membaca, merubah dan menulis data, dan *Kitchen* adalah program yang mengeksekusi *Job.*

2.9 Business Intelligence

Menurut Cui, et al (2007), *business intelligence* adalah sebuah metode atau cara untuk meningkatkan kinerja bisnis dengan memberikan dukungan yang kuat untuk eksekutif perusahaan dalam melakukan pengambilan keputusan dengan memberikan informasi yang dapat ditindaklanjuti.

Business intelligence mengkombinasikan antara data operasional dengan peralatan analisis untuk menampilkan informasi yang komplek dan kompetitif kepada pengguna atau pembuat keputusan (Negash, 2004).

Berdasarkan penjelasan di atas dapat disimpulkan *business intelligence* adalah suatu cara untuk meningkatkan data-data seperti data operasional yang dikombinasikan dengan alat penunjang analisis data dan membuat sebuah *report* atau tampilan informasi yang dapat digunakan oleh pembuat keputusan.