

PROTOTYPE METADATA PADA DATA WAREHOUSE UNTUK BASIS DATA HETEROGEN

Stefany Yunita Bara'langi

Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Atma Jaya Makassar
Alamat email : fbaralangi@lecturer.uajm.ac.id

ABSTRACT

This research aims to design a prototype for data warehouse as a data storage that has a structure and derived from different data sources (heterogeneous). Metadata is the data description which can be used as a catalog so that users can get complete information about the database is built. Data warehouse is a collection of data that are subject-oriented, integrated, time variant and non-volatile . These properties make the data warehouse included in OLAP (On-Line Analytical Processing) where the data stored and can not be changed at any time (real time). That was the distinguishes between a regular information system. The ability of data warehouses is extraction data from heterogeneous databases. Heterogeneous database is the database that is derived from several different data sources and have different table structures also vary. The author designed a prototype with the ETL (Extract, Transform, Load) and using a star schema. The design of data warehouse was used Kimball method which designed from branch table to centralized (bottom-up approach). The prototype of this metadata can help the decision-making process with the use of data mining techniques.

Keywords: metadata, data warehouse, OLAP, ETL, Kimball

1. PENDAHULUAN

Menurut Peraturan Pemerintah Nomor 82 tahun 2012, teknologi informasi merupakan teknik untuk mengumpulkan, menyiapkan, menyimpan, memproses, mengumumkan, menganalisis, dan menyebarkan informasi. Informasi yang dikelola dapat berupa teks dokumen, gambar, audio, video, kode/symbol, dan lain-lain. Informasi merupakan data yang diolah dan hasilnya memberi manfaat bagi penggunaannya [1]. Data dapat diperoleh dari berbagai sumber.

Salah satu pemanfaatan data dalam bidang pendidikan adalah data tugas akhir mahasiswa. Data tugas akhir tersebut disimpan dalam basis data masing-masing perguruan tinggi yang didefinisikan dengan struktur basis data yang berbeda-beda. Dengan adanya basis data tugas akhir diharapkan dapat membantu mahasiswa dalam menemukan referensi penelitian yang dapat dikembangkan.

Kondisi nyata yang terjadi adalah data tugas akhir dari setiap perguruan tinggi belum terintegrasi dengan perguruan tinggi lainnya. Akibatnya mahasiswa memperoleh

data tugas akhir yang akan dijadikan referensi hanya dari perguruan tinggi asal.

Bagi perguruan tinggi data merupakan aset penting. Dari data inilah maka muncul informasi yang berguna. Data tugas akhir dapat membantu sivitas akademika dalam mengembangkan berbagai penelitian. Kondisi ideal yang diharapkan adalah pemanfaatan teknologi informasi pada perguruan tinggi dengan pengumpulan, penyimpanan, pemrosesan, dan penyebaran informasi tugas akhir yang dapat memperkaya khasanah referensi penelitian sivitas akademika. Oleh karena itu diperlukan teknologi basis data tertentu dalam mengelola integrasi data dari perguruan tinggi yang berbeda-beda.

Data warehouse merupakan teknologi basis data yang dapat mendukung integrasi data dari sumber data yang berbeda dan struktur basis data yang berbeda pula [2]. Kondisi struktur data dan sumber data yang berbeda-beda ini disebut dengan basis data heterogen. Kemampuan data warehouse inilah yang tidak terdapat basis data sistem informasi biasa. Selain itu kemampuan data warehouse untuk melakukan OLAP dimana

data yang tersimpan bersifat *history* sehingga dapat menyimpan data yang berukuran besar.

Berdasarkan rumusan masalah yang dipaparkan maka yang menjadi tujuan penelitian ini adalah menyediakan prototipe data warehouse yang dapat digunakan sebagai media penyimpanan data tugas akhir dari sumber data yang berbeda (heterogen) dengan struktur yang berbeda pula.

Dalam penelitian ini, penulis akan menerapkan konsep ETL untuk ekstraksi data dan *star schema* untuk menggambarkan model dimensional.

2. TINJAUAN PUSTAKA

2.1 Data warehouse

Sifat data warehouse adalah [2]:

a. *Subject oriented* (berorientasi subyek)

Berorientasi pada subyek enterprise seperti mahasiswa, pelanggan, pasien, produk, dan lain-lain. Hal ini bertujuan untuk mengambil data yang berguna bagi keputusan.

b. *Integrated*

Data warehouse dibangun dari integrasi basis data yang beragam seperti basis data relational, *on line transaction records*, dan *flat file*. Data warehouse merupakan konsolidasi dan integrasi data dari berbagai sumber data dan menyusun data tersebut dalam format yang lebih berarti untuk keputusan yang akurat.

c. *Time-variant*

Data disimpan untuk menyediakan informasi dalam bentuk data histori (misalnya data empat tahun yang lalu).

d. *Non-volatile*

Data tidak diperbaharui terus menerus (tidak bersifat *real time*). Data warehouse tidak memerlukan pemrosesan transaksi dan recovery.

2.2 ETL (Extract, Transform, Load)

Pembersihan data (*data cleaning*) dan teknik integrasi data disediakan untuk menjamin konsistensi struktur data. Pembersihan data dilakukan untuk menghilangkan *missing value*, *outlier*, *noise*, dan *data redundant*. Data warehouse bukan hanya tempat penyimpanan data tetapi juga sebagai alat bantu untuk mengekstrak (*extract*), merubah (*transform*), dan

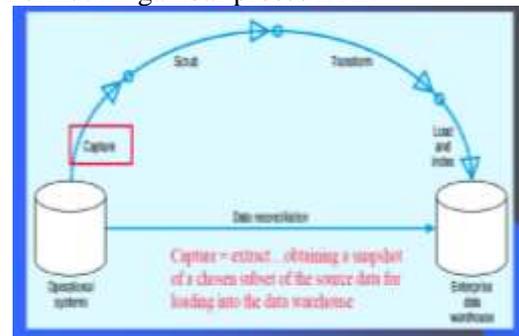
menerima data (*load*) ke penyimpanan serta mengelola dan menerima metadata.

Ekstrak data sebagai langkah awal pengambilan data dari sumber data terbagi atas ekstrak statis dan ekstrak inkremental. Ekstrak statis adalah ekstrak yang berlangsung hanya satu kali diawal. Ekstrak inkremental adalah ekstrak yang terjadi setelah ekstrak terakhir dilakukan. Ekstrak ini dilakukan hanya dalam jangka waktu tertentu.

Transformasi data dapat dilakukan pada level baris (*select*, *join*, dan agregasi data) dan level kolom (satu kolom dan banyak kolom).

Proses *load* terbagi dua yaitu *refresh* dan *update*. *Refresh* adalah mengisi data secara periodik. *Update* adalah hanya mengisi data yang mengalami perubahan.

Berikut ini gambar proses ETL:



Gambar 1 Proses ETL

2.3 Model Dimensional

Model dimensional merupakan pemodelan data khusus untuk data warehouse. Model dimensional memodelkan struktur tabel dari sumber dan struktur tabel yang didefinisikan untuk data warehouse. Komponen model dimensional adalah:

a. Tabel fakta

Tabel fakta berisi data faktual dan data kuantitatif. Selain itu, tabel fakta berisi kunci primer dari setiap tabel fakta. Kunci primer tersebut akan menjadi kunci tamu dalam tabel fakta.

b. Tabel dimensional

Tabel dimensional berisi deskripsi tentang subyek organisasi.

Terdapat tiga skema dalam model dimensional, yaitu:

a. *Star schema*

Star schema merupakan skema yang digunakan untuk data yang belum dinormalisasi.

b. *Snowflake schema*

Snowflake schema merupakan skema yang digunakan untuk data yang sudah dinormalisasi.

c. *Fact constellation schema*

Fact constellation merupakan skema yang digunakan untuk data yang belum dan sudah dinormalisasi.

2.4 Metadata

Metadata merupakan data yang menjelaskan sifat/karakteristik data yang lain [3]. Metadata dapat dianalogikan seperti sebuah katalog buku dimana pengguna mendapatkan informasi mengenai basis data sebuah sistem [4]. Deskripsi data disimpan dalam tabel khusus dan digunakan untuk evaluasi menggunakan perintah query. Metadata dapat berisi penjelasan atribut, penggunaan alias, struktur tabel, dan info keterangan tambahan yang dapat berarti bagi pengguna basis data.

2.5 Basis Data Heterogen

Basis data adalah kumpulan informasi yang disimpan di dalam komputer secara sistematis sehingga dapat diperiksa menggunakan suatu program komputer untuk memperoleh informasi dari basis data tersebut. Basis data beragam dikelompokkan menjadi:

a. *Syntactic heterogeneity*

Heterogen dalam kelas ini meliputi perangkat keras (bandwith, CPU, memori), perangkat lunak (sistem operasi, protokol, keamanan), dan antarmuka pengguna (form HTML atau SQL).

b. *Structural heterogeneity*

Heterogen dalam kelas ini meliputi model data (relasional, XML, data berorientasi obyek), dan skema (atribut, relasi, normalisasi).

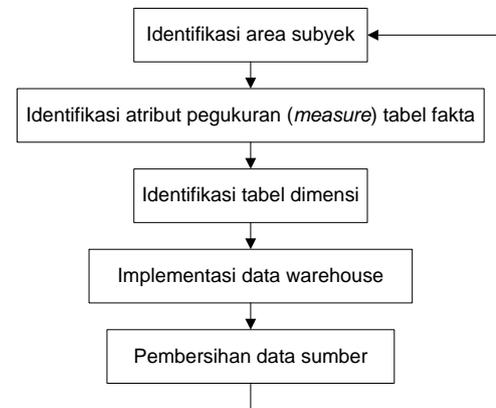
c. *Semantic heterogeneity*

Heterogen dalam kelas ini meliputi penamaan (sinonim, homonim), konflik data (data yang duplikat), dan identitas data.

2.6 Pendekatan Kimball

Pendekatan Kimball merupakan metode perancangan data warehouse yang diawali dengan identifikasi dan perancangan dari tabel dimensi/cabang ke tabel fakta [5]. Pendekatan ini dikenal juga dengan metode

bottom-up. Pendekatan Kimball memiliki keunggulan dapat digunakan pada data yang berdimensi dan cepat menghasilkan prototipe data warehouse. Akan tetapi perlu proses monitoring untuk memastikan kekonsistenan dimensi dari tabel cabang. Berikut ini adalah tahapan pendekatan Kimball [5]:



Gambar 2 Pendekatan Kimball

3. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan Kimball dengan sumber data, teknik pengumpulan data, dan tahapan sebagai berikut:

3.1 Sumber data

Sumber data diambil berasal dari basis data karya ilmiah mahasiswa Program Studi Teknik Informatika Universitas Hasanuddin (Unhas) dan basis data tugas akhir Fakultas Teknologi Informasi Universitas Atma Jaya Makassar (FTI UAJM).

3.2 Teknik pengumpulan data

Teknik pengumpulan data yang digunakan untuk mendapatkan informasi yang dibutuhkan dalam penelitian adalah:

1. Ekstraksi data dari basis data tugas akhir.
2. Studi literature

3.3 Tahapan

Tahap-tahap penelitian yang dilakukan oleh penulis adalah sebagai berikut:

1. Penulis terlebih dahulu melakukan pengumpulan data yang dilakukan dengan penggunaan dokumen, ekstrak data dari basis data dan studi literatur.

- Melakukan sinkronisasi untuk tabel dan field dari data yang diekspor.
- Mengimpor data hasil sinkronisasi tabel dan field di langkah ke-2 dan menyimpan hasil sinkronisasi pada data warehouse.

4. HASIL DAN PEMBAHASAN

4.1 Perancangan

Perancangan data warehouse menggunakan metode Kimball. Metode Kimball dipilih penulis karena proses perancangan diawali dengan identifikasi tabel cabang yang dalam penelitian ini dimulai dari tabel dimensi selanjutnya tabel fakta sebagai metadata data warehouse dibangun. Perancangan diawali dengan proses ETL dimana penulis melakukan *extract, transform, load*. Ekstrak dilakukan dengan *capture* data dari sumber data, yaitu basis data tugas akhir mahasiswa Program Studi Teknik Informatika Universitas Hasanuddin dan Fakultas Teknologi Informasi. Berikut ini struktur basis data Unhas:

Tabel 1 Struktur Tabel Unhas

No	Field	Karakter	Keterangan
1	npm	Varchar (10)	Nomor Pokok Mahasiswa
2	Judul	Varchar (50)	Judul tugas akhir
3	Penulis	Varchar (100)	Nama penulis
4	Karil	Varchar (10)	Karya ilmiah (tabel 2)
5	Kata_kunci	Varchar (50)	Kata kunci abstrak
6	Jenis	Varchar (10)	Jenis tugas akhir
7	Kode	Varchar (20)	Kode penomoran tugas akhir di perpustakaan
8	Jurusan	Varchar (50)	Nama jurusan

Pada tabel 1 didefinisikan delapan field pada tabel Unhas, yaitu npm, judul, penulis, kategori, metode, kata kunci, jenis, kode, dan jurusan. npm merupakan atribut kunci primer pada tabel Unhas. Semua atribut bertipe varchar. Field kategori terdiri dari dua nilai, yaitu:

Tabel 2 Metadata Field Karil

No	Nilai	Keterangan
1	S	Skripsi
2	T	Tesis
3	D	Disertasi
4	J	Jurnal
5	A	Artikel
6	L	Lain-lain

Berikut ini struktur tabel FTI_UAJM:

Tabel 3 Struktur Tabel FTI_UAJM

Field	Karakter	Keterangan
<i>Title</i>	Varchar	Judul tugas akhir
<i>Author</i>	Varchar	Penulis
Tahun	Varchar	Tahun terbit
<i>Academic_Department</i>	Varchar	Jurusan
Kota	Varchar	Kota dimana tugas akhir diterbitkan
Universitas	Varchar	Nama universitas
Level_kualifikasi	Varchar	Kualifikasi S1 (skripsi) atau S2 (tesis)
DDC	Varchar (18)	<i>Dewey Decimal Classification</i>
Detail	Varchar	Detail tugas akhir

Pada tabel 3 didefinisikan sembilan field, yaitu *title*, *author*, *tahun*, *academic_department*, *kota*, *universitas*, *level_kategori*, *DDC*, dan *detail*. DDC merupakan sistem pengelompokan koleksi perpustakaan seperti berdasarkan bidang ilmu.

Tabel 4 Metadata Data Warehouse

Field	Type	Panjang	Keterangan
No	Int	20	Nomor
Judul	Varchar	1000	Judul Skripsi
Metode	Varchar	254	Metode yang digunakan
Jenis	Varchar	254	Jenis tugas akhir
Lokasi	Varchar	254	Lokasi Penelitian

Pada tabel 2 didefinisikan struktur data warehouse yang berisi hasil ekstrak data dari tabel tugas akhir Unhas dan FTI UAJM. Atribut dari tabel Unhas dan FTI UAJM ditransformasikan (*transform*) menggunakan *field level*. Hal ini terjadi karena struktur tabel

dari kedua perguruan tinggi berbeda sehingga perlu dilakukan sinkronisasi penamaan *field* terlebih dahulu. Pembersihan data yaitu pada *missing value* dimana terdapat beberapa sel data yang kosong. Pembersihan data dilakukan dengan mengisi nilai default 0 pada *missing value* tersebut.

Setiap *field* memiliki nilai yang didefinisikan secara berbeda seperti pada tabel-tabel berikut ini:

Tabel 3 Field Jenis

Nilai	Keterangan
Skripsi	Jenis tugas akhir Unhas dan FTI_UAJM
Tesis	Jenis tugas akhir Unhas

Tabel 4 Field Lokasi

Nilai	Keterangan
(nama lokasi)	Nama lokasi penelitian
U	<i>Unknown</i> = jika lokasi tidak diketahui

Tabel 5 Field Metode

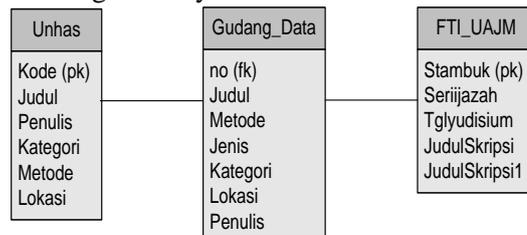
Nilai	Keterangan
(nama metode)	Nama lokasi penelitian
U	<i>Unknown</i> = jika metode tidak dicantumkan

Field No didefinisikan menggunakan format xx01 dimana dua digit pertama xx merupakan keterangan lokasi tabel asal seperti di bawah ini:

Tabel 6 Field No

Nilai	Keterangan
01	Unhas
02	FTI_UAJM

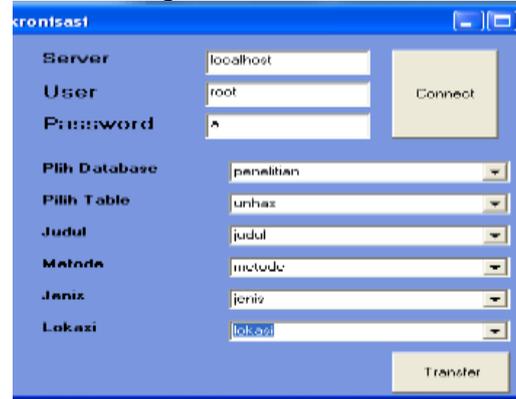
Pemodelan data warehouse menggunakan model *star schema* karena terdapat data yang belum dinormalisasi sehingga proses penelusuran basis data lebih cepat. Berikut ini adalah gambarnya:



Gambar 3 Star Schema

4.2 Hasil

Setelah melakukan perancangan struktur tabel data warehouse maka penulis merancang tampilan antarmuka dimana pengguna dapat melakukan sinkronisasi (*load*) tabel tanpa harus akses ke basis data. Tampilan antarmuka form sinkronisasi dapat dilihat dalam gambar 4 berikut:



Gambar 4 Tampilan Antar Muka

Pada gambar 4 terdapat *edit box* untuk mengisi nama server, user, dan password. Setelah mengisi tiga komponen tersebut maka pengguna dapat terhubung ke basis data dengan menekan tombol Connect. Pada saat tombol Connect ditekan, proses *load* sebagai bagian dari ETL terjadi. Jika akses berhasil maka akan tampil semua nama basis data. Pengguna kemudian memilih salah satu nama basis data dan nama tabel yang bersesuaian dengan basis data tersebut akan ditampilkan. Proses ETL akan bekerja pada saat pengguna menekan tombol Transfer. Berikut ini adalah potongan sintaks program yang melakukan proses ETL:

```
Private Sub Command1_Click()
Set con = New ADODB.Connection
con.CursorLocation = adUseClient
con.ConnectionString = "DRIVER={MySQL ODBC 3.51 Driver};SERVER=127.0.0.1;DATABASE=penelitian;UID=root;PWD=a;port="
con.Open
Set rs = con.Execute("show databases")
Do While Not rs.EOF
    Cdb.AddItem rs.Fields(0)
    rs.MoveNext
Loop
```

4.3 Pembahasan Hasil

Data warehouse dirancang sebagai teknologi yang dapat membantu pengguna dalam melakukan sinkronisasi basis data heterogen. Hasil yang dicapai pada penelitian ini disimulasikan pada dua tabel yang memiliki struktur berbeda dan berasal dari basis data yang berbeda. Penulis menyediakan tampilan sinkronisasi yang fleksibel bagi pengguna untuk memilih basis data, tabel, dan field-field yang akan terlibat dalam data warehouse.

Data warehouse yang dihasilkan dapat disinkronisasi dengan cepat karena penulis memilih pemodelan star schema karena masih terdapat data yang belum dinormalisasi. Jika data sudah ternormalisasi maka dapat digunakan pemodelan *snowflake schema*. Berikut ini adalah tabel konfigurasi:

Tabel 7 Konfigurasi Prototipe

Fitur	Kemampuan	Alasan
ETL	ETL	Proses awal dari data warehouse
OLAP	Drill down, Slice, Dice	Kemampuan OLAP disediakan untuk kemampuan <i>mining</i> selanjutnya
Pemodelan	<i>Star schema</i>	Data belum dinormalisasi

5. KESIMPULAN

Kesimpulan dari penelitian ini adalah:

- Prototipe metadata data warehouse yang dihasilkan dapat digunakan sebagai media penyimpanan data tugas akhir dari dua perguruan tinggi.
- Pemanfaatan teknologi data warehouse dapat membantu sivitas akademika untuk memperoleh informasi tentang data tugas akhir dari perguruan tinggi yang berbeda.

6. DAFTAR PUSTAKA

- [1] Peraturan Pemerintah Republik Indonesia Nomor 82 tahun 2012 tentang Penyelenggaraan Sistem dan Transaksi Elektronik. [Online]. Diakses dari www.serdos.dikti.go.id [26 November 2012]
- [2] Han, Jiawei., Kamber, Micheline. 2008. *Data mining: Concepts and Techniques*. Morgan Kaufmann Publishers, California.
- [3] Kadir, Abdul. Konsep Basis Data.
- [4] Ramakrishnan., Gehrke. 2003. Database Management System (3rd edition). McGraw Hill, New York.
- [5] Subhan, Muhammad. Pengantar Data Warehouse [Online]. Diakses dari www.ilmukomputer.org [Diakses pada 20 Juni 2014]