Design of Library Data Warehouse Using SnowFlake Scheme Method
Case Study: Library Database of Campus XYZ

Akhmad Dahlan
Department of Informatics Engineering
STMIK AMIKOM Yogyakarta
Yogyakarta, Indonesia
e-mail: allandi@amikom.ac.id

Ferry Wahyu Wibowo
Department of Informatics Engineering
STMIK AMIKOM Yogyakarta
Yogyakarta, Indonesia
e-mail: ferry.w.wibowo@ieee.org

Abstract - The existence of the library as a technical service unit in a campus is very important to provide services to the academic community. With increasing number of library collections, database that is built has to be able to improve the services oriented on providing data warehouse. Especially at managerial level requires complete information, quickly and accurately to support the process and planning, evaluation, and right decision-making. The design of data warehouse is determined by description of proper information requirement, selection of valid data source, design of data warehouse and ETL process to integrate, extract, cleanse, transform and populate it into data warehouse. The snowflake scheme design method has been applied to accommodate dimension tables of database and other dimension sub-table, so it can generate more information that would be used as material to make a decision.

Keywords - campus; data warehouse; ETL; library; snowflake

I. INTRODUCTION
Utilization of data warehouse technology is almost needed by all organizations, including library of university. The library of university is expected to have data warehouse, so the process of data integration could be done easily. Basically, the process that has been done is data summarizing activities stored in data warehouse or due to database system. Supervisors are necessary to obtain data summaries of transaction data occurred on their departments or known as transactional information system (TPS). At higher levels, the information should be presented in such a way to support the decision-making process. Systems at this level are known as decision support systems. At service operational level, to handle transaction process is using transactional database systems. In the database would store all detail information for all events and transactions that have been done. The stored data will always be dynamic as increasing transactions in the database. However, need of information managerial level is result of aggregation or summary of the transactional database. The most important thing in the process of building data warehouse is the process of modeling [1]. Because from the designed data model would be used to determine capability of data warehouse in providing managerial information needs.

Modeling of data warehouse has some kinds of method e.g. star scheme and snowflake scheme methods. The snowflake scheme method is developing method of star scheme namely method of normalization for dimension tables existed in the star scheme [2]. In the star scheme method, a dimension tables don’t have a sub-table or sub-dimension so data those could be employed as material providing information are not much. In a snowflake scheme method, each dimension table may have another sub-dimension tables. It aims to minimize redundancy data. This dimension data which become subject of information to be material in making decision, because in each dimension of data is possible to do derivation more detail. So that, data source that could be processed into information may become huge number and more detail.

Due to the important of managed database, this paper shows how to prepare a data model using snowflake scheme method to build a data warehouse of library of campus XYZ. In additional, it also defines process of extraction, transformation and loading (ETL) of data from transactional database into the data warehouse and how to arrange and analyze it using a reporting system for management of the library.

II. RELATED WORKS

A. Data Warehouse
Data warehouse is a collection of subject-oriented data, integrated data, non-updateable, having time variant, for historically information and able to support decision-making [3]. The data warehouse is built to prevent technical problem and business due to the usage of data and information to take a decision [4]. The data warehouse is also a system that takes and consolidates data periodically from source system to the dimension table or storage data those have been normalized. The data usually are renewed for such time not once transaction in the source system [5].

B. Snowflake Scheme
Snowflake scheme method is a data modeling method for data warehouse developed from star scheme method. The implementation of both has trade-off [6, 7]. The snowflake scheme is mixing of some data modeling of star scheme normalized. In the snowflake scheme, each dimension table
can have another dimension sub-table. This data dimension becomes information subject that is employed as material in taking decision, because in each data dimension is possible to be done breakdown more detail. So that, the data source that could be processed becomes information may a huge number and more detail.

C. ETL

Extraction, transformation, and loading (ETL) are processes of taking and changing data from the source system and put it into data warehouse [8]. In the process of taking a data, the data must be clean in order to obtain a good data quality. Examples there are invalid telephone numbers, not existed book code, some null data, etc. The traditional approaching in the ETL processes retrieves data from data source, puts in a staging area, and then transforms and loads it into data warehouse [5].

III. METHODOLOGY AND RESULTS

A. Define Data Subject of Warehouse

In the planning stage, the main subject in the library database is process of book circulation that student could borrow book(s). This process describes the mechanism of borrowing and returning a library collection books for students.

B. Define Data Source and Information Requirement

This stage defines all data sources those would be implemented in the data warehouse and defines information requirements those have to be yielded by data warehouse through reporting analysis. The data source that would be processed to establish data warehouse obtained from library transactional database based on such periods. In this case, the obtained data are library transactional data from year 2003 until January 2013. The relation tables of library database shown in Figure 1.

Based on observation of the library, some needed important information is as follows,

a. Data analysis of books which are borrowed by students

b. Student data of most frequently borrowed books based on a multidimensional tables

c. Information on the type of books which are most frequently borrowed

d. Analysis of time series or circulation process trends

e. Information on the name of authors of the books which are most frequently borrowed

f. Information on the name of publishers of the books which are most frequently borrowed

g. Analysis of circulation related with the multidimensional tables

C. Design a Snowflake Scheme and ETL Processes

In this stage is to make dimensional data model that forms snowflake scheme and enhanced making a design of ETL processes. Which became the main subject in the library warehouse is the circulation of borrowing books for students that would be the center of scheme. In this scheme would be implemented in the form of fact table named Fact_StudentLibrary. While other data would be implemented into dimension and sub-dimension tables those all tables related to the fact table. The snowflake scheme design model is shown in Figure 2.

D. Doing ETL Processes In Data Warehouse Database

In the processes of ETL from the library database to the data warehouse database, there are a few things must be done. Generally divided into 2 stages i.e.

1) Eliminating Noise Data

Before the ETL processes are done, it is necessary to eliminate the noise data which exist in the source database or on line transaction processing (OLTP) database. This is done to maintain the integrity of data which would be processed. After doing analysis process of database, it found a lot of noise data existed in the source database. For example there were 58 rows in the table of dbo_Karang which was inconsistent. It could be seen from ISBN data in table of dbo_Karang, although ISBN data didn’t exist in the reference table namely table of

Figure 1. The Relation Tables of Library Database.

Figure 2. The Snowflake Scheme of Data Warehouse.
dbo_Buku. The eliminating process proof of noise data is shown in Figure 3.

![Figure 3. Testing on Data Consistency.](image)

2) Data Population Process

The process of extracting data from operational database of library to data warehouse database using a population technique. This process consists of techniques to populate data into dimension and fact tables. The design of data population process from table in the operational database of dimension and fact tables can be shown in Figure 4. The design of dimension tables describe flows of dimension table population for book table, author table, publisher table, source table, type table, and student table done directly from the OLTP database.

![Figure 4. Mapping of Data Population Process.](image)

Once the mapping process for all the tables were successfully done, then the data transformation services (DTS) process would be run until all data were distributed in accordance with tables mapping which was done. The DTS process indication was successful is shown in Figure 5.

![Figure 5. Indication of Successful DTS Process.](image)

The next process is the process of periodic loading at each dimension table. The periodic loading process was conducted periodically based on the time required in accordance with developing data in OLTP database. These processes have been done using structured query language (SQL) transact commands in database of data warehouse. The transact SQL commands shown as follows,

```sql
/* 1- Populasi Data dim_jenisbuku */
insert into dim_jenisbuku
select * from DBPerpus.dbo.dbo_JENIS
where DBPerpus.dbo.dbo_JENIS.no_jenis <> dim_jenisbuku.no_jenis

/* 2- Populasi Data dim_penerbit */
insert into dim_penerbit
select * from DBPerpus.dbo.dbo_PENERBIT
where DBPerpus.dbo.dbo_PENERBIT.kd_terbit <> dim_penerbit.kd_penerbit

/* 3- Populasi Data dim_Pengarang */
insert into dim_pengarang
select * from DBPerpus.dbo.dbo_PENGARANG
where DBPerpus.dbo.dbo_PENGARANG.kd_karang <> dim_pengarang.kd_pengarang

/* 4- Populasi Data dim_mahasiswa*/
insert into dim_mahasiswa
select * from DBPerpus.dbo.dbo_mhs
where DBPerpus.dbo.dbo_mhs.NPM <> dim_mahasiswa.NPM

/* 5- Populasi Data dim_Sumber */
insert into dim_sumber
select * from DBPerpus.dbo.dbo_SUMBER
where DBPerpus.dbo.dbo_SUMBER.kd_sum <> dim_sumber.kd_sum

/* 6- Populasi Data dim_buku*/
insert into dim_buku
select [isbn],[judul],[jenis]
```
Looking at this report, the library management can determine author’s names, which the books are a best seller books. By showing in Figure 6.

IV. DISCUSSIONS

Analysis of the data warehouse has been done by making a reporting in accordance with the business requirement. The reporting has been built using transact SQL commands. The few reports those can be made as follows:

A. Report of The Number of Borrowing Per Year

This report provides information to the management of the library due to the number of borrowing books per year as shown in Figure 6.

```sql
-- /* REPORT OF THE NUMBER OF BORROWING PER YEAR */
SELECT dim_time.Year,
COUNT(kd_pinjam) AS jml_peminjam
FROM fact_StudentLibrary JOIN dim_time
ON fact_StudentLibrary.time_key = dim_time.DateKey
GROUP BY dim_time.Year ORDER BY dim_time.Year DESC
```

B. Report of Best Seller Books

This report provides information of the book titles and author’s names, which the books are a best seller books. By looking at this report, the library management can determine the policy during the process of providing books. The SQL commands for this shown as follows,

```sql
/* REPORT OF BEST SELLER BOOKS */
SELECT
dim_buku.isbn,
dim_buku.judul,
dim_pengarang.namadepan + dim_pengarang.namabekal AS Pengarang,
COUNT(dim_buku.isbn) AS jml_dipinjam
FROM dim_buku
INNER JOIN
fact_StudentLibrary ON dim_buku.isbn = fact_StudentLibrary.isbn
INNER JOIN
dim_pengarang ON fact_StudentLibrary.kd_pengarang = dim_pengarang.kd_pengarang
GROUP BY dim_buku.isbn, dim_buku.judul,
dim_pengarang.namadepan,
dim_pengarang.namabekal
ORDER BY jml_dipinjam DESC
```

C. Report of Days At Most Borrowers

This report provides information on the busiest days or days at most borrowing process. From this information, the library management can anticipate the moments where the most process of circulation, so the librarian can serve them better. For example the determination of the time schedule of librarian shift, so the librarian should be in place, if there are other activities requested not at the moment when a lot of circulation process. The SQL commands for this shown as follows,

```sql
/* REPORT OF BUSIEST DAYS */
SELECT dim_date.DayOfWeek AS hari,
COUNT(fact_StudentLibrary.kd_pinjam) AS jml_pinjam
FROM fact_StudentLibrary INNER JOIN
dim_time ON fact_StudentLibrary.time_key = dim_time.Idtime_dimmsesion
GROUP BY dim_time.DayOfWeek
ORDER BY jml_pinjam DESC
```

D. Report of Authors of The Best-Seller Books

This report gives information of names of the author of best-selling where books authored often borrowed by library patrons. The SQL commands for this shown as follows,

```sql
/*REPORT OF BEST SELLING AUTHORS */
SELECT
dim_pengarang.namadepan + dim_pengarang.namabekal AS pengarang,
COUNT(fact_StudentLibrary.kd_pinjam) AS jml_pinjam
FROM dim_buku
INNER JOIN
fact_StudentLibrary ON dim_buku.isbn = fact_StudentLibrary.isbn
```

This report provides information of names of publisher who publishes books and often borrowed by library patrons. The SQL commands for this shown as follows,

```sql
/* REPORT OF BEST SELLING PUBLISHER */
SELECT dim_penerbit.namapenerbit as [Nama Penerbit],
       COUNT(fact_StudentLibrary.kd_pinjam) AS [Jumlah Peminjaman]
FROM fact_StudentLibrary INNER JOIN dim_penerbit ON fact_StudentLibrary.kd_penerbit = dim_penerbit.kd_penerbit
GROUP BY dim_penerbit.namapenerbit
ORDER BY [Jumlah Peminjaman] DESC
```

Figure of report each publisher shown in Figure 7.

V. CONCLUSIONS

The process of building a data warehouse with the snowflake scheme method through several continued steps namely designing the scheme, performing extract, transform, and loading (ETL) processes, and making a cube-data dimensional. Preparation of data warehouse at the stage of ETL was constrained because the field of foreign key is inconsistent. Many foreign keys were null. This is because at the beginning of the system application, the data integrity wasn’t maximized. To ensure consistency of the foreign key, it needs to remove the noise data because the integrity of data in the database used as a source to compose data warehouse is critically important. A data modeling and making of actual transactional database system will ease in the data warehouse preparation.

ACKNOWLEDGMENT

We thank STMIK AMIKOM Yogyakarta which have given us chance for presenting our research paper.

REFERENCES