An Overview and Implementation of Extraction-Transformation-Loading (ETL) Process in Data Warehouse

(Case Study: Department of Agriculture)

Rahmadi Wijaya[1], Bambang Pudjoatmodjo[2]
School of Applied Science
Telkom University
Bandung, Indonesia
[1]rahmadi@telkomuniversity.ac.id, [2]bpudjoatmodjo@telkomuniversity.ac.id

Abstract*-Extraction-transformation-loading (ETL) process in data warehouse development perform data extraction from various resources, transform the data into suitable format and loadit into data warehouse storage. In the ETL process, there is data cleansing process function that handles redundancy, inconsistency and integrity data. ETL process will move data from the source to the integration layer (data store in data warehouse). In the integration layer, the data can be grouped into smaller scope and more specific for the requirement in other repositories called data marts. Reporting program of data warehouse will be associated with a data mart as its data source. In this research, the data warehouse is built to handle the ETL process. The data warehouse build metadata to support the process. The metadata construction for ETL processes will lead to ETL programs with high degree of reusability. The conclusion from this research is the use of dynamic ETL process (using metadata ETL) is required when ETL process is dealing with the operational system that still unstable and likely to change the database schema. Dynamic ETL process is also needed to address the increase requirement for report from the users.

***Keyword*-Data Warehouse, ETL, Integration Layer, Reusability, Metadata.**

I. Inroduction

The key success for a company to survive in the recent time is the ability to analyze, plan and react toward changes in the business environment. This ability will only be fulfill if adequate information is available in the data format and structure is suitable with the decision-making [10].

Department of Agriculture is a government agency who has data that must be analyze, such as financial data, personnel data, project data, the data management equipment, and agricultural data as well as export-import data. In addition, the department of Agriculture has different information systems to process the data in each domain in its divisions, such as Human Resources Management Information System, Project Management Information System, Financial Management Information System, Equipment Management Information System, Agricultural Statistics Database System, Export-Import Database System. In the present time, the Department of Agriculture has extracted data from each domain manually, and then use it as information in decisions making for the department [5].

Data warehouse technology is used to overcome variety of data and information system. Data warehouse development is accomplished in a group, that includes three components; extraction, transformation, and loading (ETL) process, and as a supporting component is data mart and reporting data warehouse, where each component separately constructed according to the conditions at the Department of Agriculture. Building the data warehouse, it is expected that manual process can be automated [8].

The current condition of the Department of Agriculture in analyzing and processing its data is described as follow: Each information system has its own different database. There even is a flat file format (.xls), the data is scattered in each information system, redundancy data is still occurred, in the analysis and decision-making, the Executives of Department of Agriculture had to do a recap of data from each information system.

For improving the performance of the Department, the Department of Agriculture executives must be provided with the accurate and up to date data to generate strategic decisions. This research will discuss about the data warehouse planning with case study at the Department of Agriculture [7].

II. Data Warehouse

A. Definition

Bill In nondefined data warehouseas collection of data in support of management's decision making processes. It has several characteristics [8]:

- Subject Oriented
- Integrated
- Non Volatile
- Time Variant

978-1-4799-7752-9/15/$31.00 ©2015 IEEE

B. Data warehouse Architecture

Data warehouse is an open architecture system, and then it can be architected in many different ways, depending on the specific needs of system requirements. Fig.1. describes an example of data warehouse architecture. The architecture shows the difference between data warehouse and database in operational system. In data warehouse, data are used for specific purposes, usually as analysis tools for applications in Executive Information System (EIS) or Decision Support System (DSS). On the other hand, operational databases are generally used in transactional applications that will make read/ write operation on the database[10].

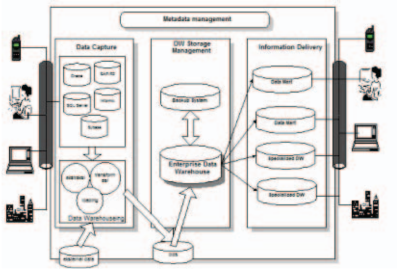

Fig.1. Data warehouse architecture

The architecture in Fig 1 is built upon specific purposes, those are [8] :

1. The data input for data warehouse is not only from internal system. The data input must accommodate external resources.
2. Stored information in data warehouse is able to be specialized into several more specific data warehouse (data mart), so that a number of additional process in the architecture are needed to populate data from data warehouse to some data mart.
3. Application on user layer develops into several models, for example: web-based,
desktop-based, even mobile-based system

C. User Access

User access is a component that defines how a user access the data warehouse. Typically, user uses Business Intelligence Software to define query and analysis in accessing data warehouse. Below are lists of Business Intelligence Software examples:

1. Decision Support System
2. Executive Information Systems
3. OLAP (Online Analytical Processing)
4. Data Mining Tools

III. ETL (EXTRACTION, TRANSFORMATION, LOADING)

During ETL process, data from many sources will be extracted and integrated into data warehouse periodically. Extraction is a process to identified and retrieve all relevant data from the sources. The role of transformation is to cleansing the data and integrated different schema to defined schema in data warehouse. Meanwhile, loading is a process to physically move the data from operational system to data warehouse.

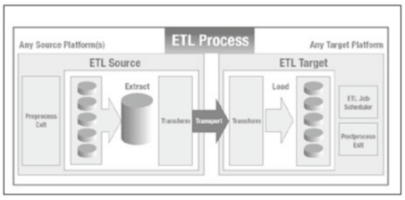

Fig.2.ETL Process

A. ETL Concept

It is necessary to define ETL scope by analyzing each target table (its dimension and facts) at the beginning of ETL architecture process. It is necessary to fetch the behavior of each target table; where its source is and what kind of business process that depend on it [8].

1. Metadata

In the early development of data warehouse, integration is formed by making specific ETL program for the structure of source database and data warehouse database.

As the time passed by, it is found that those specific ETL programs essentially are doing the same process. Many block programs are reusable for another ETL process. At this point, ETL tools that can do automated data integration to data warehouse are developed.

To put it briefly, metadata is data about data. Specially, metadata ETL is data about ETL process. Defining metadata ETL is necessary to build ETL program with high reusability [8].

2. Extraction

Extraction identifies all relevant sources, then extract the data as efficiently as possible. This process is going through file or database, using various criteria in selecting data, finding proper data, then transport the data into file or other database.

Change data capture (CDC) is an important element in extraction analysis. Almost all transaction in data fact have timestamps. However, not all dimension data in source system has timestamps due to its tendency to not depend on an event. Therefore, it is very difficult to implement CDC in dimension data[8].

3. Transformation

Transformation is a manipulation process towards data from source system to another format in data warehouse or data mart to make it a meaningful information. Transformation functions that can be used are list below[8]:

4. Loading

Loading will move transformed data to data warehouse. There are two loading strategies into integration layer; loading strategy for dimension table and loading strategy for fact table.

978-1-4799-7752-9/15/$31.00 ©2015 IEEE

B. Loading Strategy for Dimension Table

There are three strategies for loading dimension table. Each strategy addresses data dimension changing differently, depends on the way it handle the old dimension data. In all strategies, if there is no record with the same natural key, then new record will be added. Natural key is a dimension attribute that uniquely differentiates dimension record (it is not surrogate key). If there is record with the same natural key, then:

- Strategy 1: Data history is not stored. If the input value has occurred in dimension table based on natural key value, then the record will be updated.
- Strategy 2: Critical columns are important columns in dimension table that should be reserved. If in input data is found one or more columns in critical column category, and the value in input data is different with the matching column in dimension table (based on natural key value), then the matching record will be expired and new record with new surrogate key will be added. If there is no similarity in critical column, then it will be treated as in strategy 1, the matching record will be updated.
- Strategy 3: Its strategy in handling value changes in critical column is quite similar with strategy 2. The difference is, in strategy 3, for each critical column there are different columns that placed in each record to store current data value and n previous data value. When there is change in critical column value, all value will be shifted and the last value (the oldest one) will be deleted, and the newest value is added

C. Loading Strategy for Fact Table

Loading for fact table will only comprises of adding new data. Each new record will be connected with its dimension based on natural key value. Matching surrogate key in dimension table is taken from natural key information of record input [8].

IV. Case Study

In this research, Executive Information System (EIS), usually Executive Support System (ESS) will be implemented in Department of agriculture. The ministry has many divisions that operate in different fields. EIS is used to integrated data from many scattered database source in production division. EIS will be used to enhance efficiency and effectiveness of decision making of the executives [10].

Department of agriculture does not have integrated database among its divisions. The database is managed separately and locally in each division, and it has uninformed data format. It leads to difficulty in global decision-making for the executives, the high-level officers. The dispersed data will trouble the executive in doing analysis to extract the data into desired information [3].

In addition, database is generally used to store detailed data. It is not sufficient with data requirement from the executives. They usually need data summary, presented in graphical view as diagram to help them in making a decision [10].

Relevant problems that will possibly be rose in implementing EIS in department of agriculture, related to the amount of division that it has, are [10]:

1. Due to the lack of standardize format of data, incompatibility between database will likely occur.
2. Extracting information. From various data in data warehouse, EIS should be able to extract raw data from various source available, than providing useful information for the executives, usually to predict trend in five years. It is necessary to determine request format in extracting information, then match it with the data in data warehouse. For example, let the request be general information from all division, and the data format for information request is: sum of expenses in the current month, customer segmentation, operational expenses per month, and cost risks allocation.
3. Keep data accuracy. It should be guaranteed that the accessed data is the required one.
4. Choosing the right time to produce and store data to data warehouse. For example, determining data collecting frequencies, it can be done once a month, twice a month, or in any other intervals.

Determining expired time in keeping the data. The amount of stored data can reach enormous number; therefore the expired period in data storage should be determined. It can save memory capacity and enhance EIS speed performance.

A. Data warehouse Analysis in department of agriculture

The construction of data warehouse in department of agriculture is just initial phase for the development of integrated data warehouse in department of agriculture.

The purpose of building the data warehouse, therefore, is to bring about the need in general analysis for the department of agriculture executives. This research will build dynamic system that adaptive to new requirements from the executives by building the data warehouse.

In the recent times, the executives of department of agriculture are accessing report data through Center of Data and Information of Department of Agriculture (*Pusat Data dan Informasi Dinas Pertanian*, PUSDATIN). Even though they have privilege to directly access the data in operational system under their division, the data is transactional data, therefore it is too detailed and can't be used as analysis reference [7].

B. System Requirements for Data Warehouse in Department of Agriculture

Undertaking the process, extracting operational data and then presenting it as reports, will upsurge troubles if done manually. Combining data that involves many sources is a complex process, thus vulnerable to fault. The complexity of the process is increased if it contains many redundant data [7].

The weakness of the current system are [7]:

1. Access

   The executives are accessing report through Center of Data and Information of Department of Agriculture

978-1-4799-7752-9/15/$31.00 ©2015 IEEE

(*Pusat Data dan Informasi Dinas Pertanian*, PUSDATIN). Although they have direct access to operational system, data in the system is not suitable enough to use as analysis reference.

2. Time
   If a transaction is occurred in operational system, the executives can't get the report about it in a short time. The executives will endure a long bureaucracy to get the report.
3. Format
   Report format is usually defined by the executives at the commencement of report request. After receiving report in one format, the executives must accomplish the same exact procedures to get information in another format. Thus, it is hard for the executives to view information from various perspectives, such as spreadsheets or chart.
4. Integrity
   Even though desired information has been obtained, accuracy of the data is still questionable. Manual process involving enormous redundant data is susceptible to fault.

Data warehouse offers privilege for the executives to get the desired reports in a relatively short amount of time. Using report application that compatible with data warehouse, the executives can see the information in various format.

Data warehouse for Department of Agriculture will have three domains; production, finance and human resource. The data warehouse is built to facilitate accessing information of important data in business process department of agriculture by the executives for strategic decision making. The system loads data from the three domain sources, extracts them into single database and shows the result as reports to its user.

Data is get through operational system that have various data and source database format. The operational system uses Database Management System (DBMS), as well as spreadsheet (Microsoft Excel). Extraction process is scheduled by data warehouse administrator. Before that, the schedule is arranged by the executives, operational system operator and data warehouse administrator. In addition, the administrator has responsibility to maintain the database in data warehouse, such as backup, recovery, and perform extraction process (incidentally) if there is a failure in extraction process.

Default schedule for extraction process is once a month for all data operational system. If there is a change in requirement, administrator can alter the schedule to comply with the change.

This administrator's privilege, user doesn't have this privilege. User is connected to system via OLAP tools as a user interface, as previously stated [8].

The data warehouse uses DBMS SQL Server 2000 for its database. It also uses standard data type from DBMS, such as char, varchar, number, and datetime. For the feature, it uses standard DDL (Data Definition Language) feature to define database and table, and foreign key.

Data Transformation Service (DTS) with ActiveX Script Task in VBScript is used for ETL process.

Data source will be extracted, transformed and loaded to integration layer. Then, the data will get further extraction into several data mart. Report application is connected to data mart with requisite data [2].

This paper is focused solely on process in flow 1 (ETL process from data source to integration layer). Extraction processes to data mart and data utilization in data mart are beyond the scope of this paper.

Receiving extracted data from operational system, data warehouse system will do simple transformation process, including cleaning and integrating process.

The system has dictionary for possible data value and data correction for data value failure.

Integration process is also defined in dictionary as information value reference for another data, if there is an occurrence [2].

C. System Analysis for Department of Agriculture Data Warehouse

In order to be prepared for further development as integrated data warehouse at Department of Agriculture, this initial system is developed by considering dynamic growth of requirement. Requirement changing is accommodated without altering the structure of system as well as the programs that system used. It will only change the system configuration.

This system has a weakness in integration process. The process is not done thoroughly for all data in the system. The reason for this condition is, despite the characteristic of source data (operational system has only part of data).

There is also a same information that is stored as different subject in this system. Thus, further integration process can be developed in the system without disturbing entire system.

D. ETL Process Flow

Process flow to design ETL can be described as follow:

a. Identified tables from source data to load required data for metric and dimension that has been set based on user requirements.
b. Data is moved to staging area, so ETL process that will be done doesn't disturb transaction process OLTP that still run.
   This research uses SQL Server as the staging area, with the name of the database is "pertanian".
c. The stored data in fact and dimension table is transformed into multidimensional data that can be viewed from any viewpoint, so it can be accessed by client for analysis purpose.

E. User Requirements Analysis

At this stage, database in Department of Agriculture is divided into 3 domains; production, finance and human resource. Fig.3 depicts ETL process for this stage.

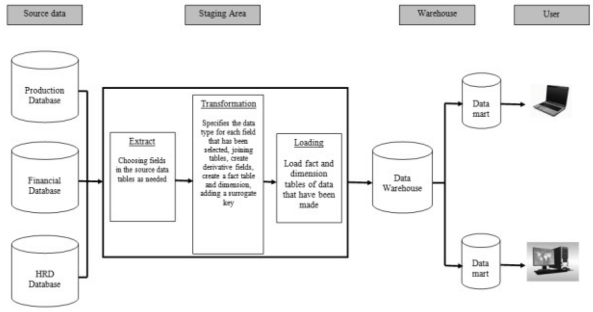
Source data
Staging Area
Warehouse
User
Production Database
Financial Database
HRD Database
Extract
Choosing fields in the source data tables as needed
Transformation
Specifies the data type for each field that has been selected, joining tables, create derivative fields, create a fact table and dimension, adding a surrogate key
Loading
Load fact and dimension tables of data that have been made
Data Warehouse
Data mart
Data mart

Fig. 3. ETL Process

V. CONCLUSION

In order to be prepared for further development as integrated data warehouse at Department of Agriculture, this initial system is developed by considering dynamic growth of requirement. Requirement changing is accommodated without altering the structure of system as well as the programs that system used. It will only change the system configuration.

ETL process for data warehouse at Department of Agriculture is arranged to be accomplished on the fly, it will not involve temporary table as an intermediary. Data is extracted, transformed, and loaded directly into database. This result in faster ETL process, but on the other hand, it will consume resource in operational system as well as data warehouse system. Therefore, ETL process is scheduled when the operational system is not busy with its transaction. Transformation process does simple process, that is correcting data value using dictionary table. Having been transformed, data is loaded directly to database in data warehouse.

REFERENCES

[1]. Fathansyah, Ir., 2002, Buku Teks Ilmu Komputer Basis Data, Informatika, Bandung

[2]. Fowler, Martin, 2004, UML Distilled Edisi 3 Panduan Singkat Bahasa Pemodelan Objek Standar, 2005, Andi, Yogyakarta

[3]. http://www.cert.or.id/~budi/courses/ec7010/dikmenjur-2004/supawi-report.pdf

[4]. http://www.iwaysoftware.com/products/images/etl_chart_s m4.gif

[5]. http://www.mcrit.com/ASSEMBLING/assemb_central/WhatESS.htmhttp://www.ptct.com/EIS.html

[6]. http://www.utminers.utep.edu/mmahmood/cis5311dtmba/slides/chapter02.ppt

[7]. Ibrahim, Nugroho Setyabudhi, Takariyana Heni A., 2004, Perancangan Data Warehouse Pada Pusat Data dan Informasi Pertanian, Tesis Magister Manajemen Informasi Universitas Bina Nusantara, Jakarta

[8]. Inmon, W.H., 2002, Building The Data Warehouse, Third Edition, John Wiley and Sons, Inc., New York

[9]. Pressman, Roger S., Ph.D., 2002, Rekayasa Perangkat Lunak Pendekatan Praktisi (Buku II), Andi, Yogyakarta

[10]. Turban, Efram, Jay E. Aronson, and Ting Peng Liang, 2005, Decision Support Systems and Intelligent Systems (Sistem Pendukung Keputusan dan Sistem Cerdas), Edisi 7 Jilid 1, Andi, Yogyakarta