PAPER • OPEN ACCESS

Extract transformation loading from OLTP to OLAP data using pentaho data integration

To cite this article: R J Salaki *et al* 2016 *IOP Conf. Ser.: Mater. Sci. Eng.* **128** 012020

View the article online for updates and enhancements.

Extract transformation loading from OLTP to OLAP data using pentaho data integration

R J Salaki*, J Waworuntu and I R H T Tangkawarow

Pendidikan Teknologi Informasi dan Komunikasi, UNIMA, Manado, Indonesia

*Corresponding author: reynaldojsalaki@unima.ac.id

Abstract. The design of the data warehouse in this case is expected to solve the problem of evaluation of learning results as well as the relevance of the information received to support decision-making by the leader. Data warehouse design is very important, which is designed to utilize the existing resources of information. GPA (Grade Point Average) data warehouse can be used for the process of evaluation, decision making and even further planning of the study program of PTIK. The diversity of data sources in the course PTIK make decision-making and evaluation process becomes not easier. Pentaho Data Integration is used to integrate data in PTIK easy. CPI data warehouse design with multidimensional database modeling approach using the dimension tables and fact tables.

1. Introduction

The use of information technology that is integrated with the process of work in an institution or corporation has become an absolute necessity today. This is caused by the need of the institution to optimize the ability to analyze the problems encountered which will affect the decision-making process. The availability of complete and accurate data is a measurement of the viability of the institution.

One indication of the success of a program of study is the absence of data GPA of students that can be accessed quickly and accurately. Problems that occur at this time generally lies in the variety of input data, causing delays in data processing. This study is expected to overcome the diversity of data sources by using Pentaho Data Integration-Kettle. This application can integrate the data that is ready to be processed in the data warehouse so it can later be presented with accurate and timely. The integration process is well known in the Business Intelligence as a process Extract Transformation Loading (ETL). ETL process will change the data On-Line Transactional Processing (OLTP) into data On-Line Analytical Processing (OLAP).

This research built data warehouse to display the GPA of students, especially in PTIK Study Program. The Manado State University established in 1955, which was originally the Teachers' Training College Manado. The PTIK study program was officially open in 2010.

The data were collected from the student data and students' GPA. This research is expected to provide precise and accurate data, especially for evaluation, decision making and planning for the development of the study program.

2. Literature review

2.1. 2.1 Business Intelligence

Business Intelligence describes a concept and method of how North to improve the quality of decision business decisions based system based data. BI often be treated equate as briefing books, report and query tools, and information systems executive. BI is a support system based decision making data (Imelda, 2015).

Research Herring, 1999, confirms that the companies without intelligence needs facing performance bad and frustrated at their CI departments. On the other hand company with a form of intelligence requires the identification process benefit from their successful CI programs. But why do not have many companies have a CI department? Reason for this might be difficult to track down intelligence needs the right to a certain strategy (Budi Harijanto, 2013).
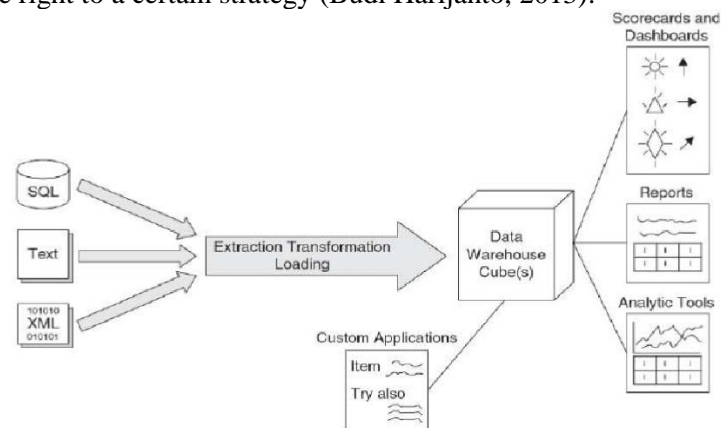
Figure 1. BI Process Steps (Source : Prastuti Sulistyorini, 2010)

2.2 Data Warehouse
The process of data warehousing carried out in three main steps , also known as extraction, transformation and loading (ETL).Extraction program retrieves data from various database Heterogeneous operations based on specific models. Metadata describes the model and the definition of the data source element (Budi Harijanto, 2013).

According to pioneer the concept and term of data warehouse, William Inmon, the definition of the data warehouse is a collection of data subject-oriented, integrated, non-volatile, and time-variant in order to support decisions of management (Stephanie Pamela Adithama, 2013).
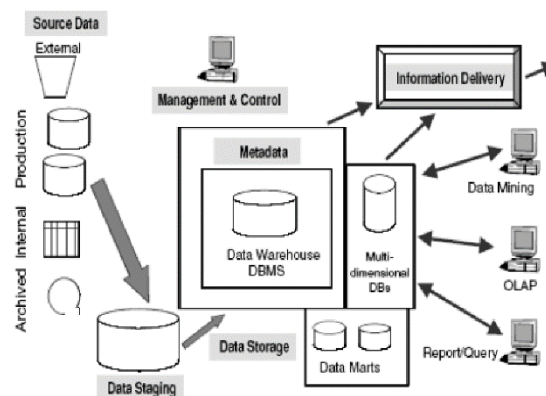
Figure 2. Component of Data Warehouse
(Source : Stephanie Pamela Adithama, 2013)

3. Research Methodology
Design method multidimensional data modelling according Kimball used includes 4 stages (Kimball & Ross, 2002). The fourth stage these are:
1. Select the Business Process
    The first step in the design is to decide what business process(es) to model by combining an understanding of the business requirements with an understanding of the available data.
2. Declare the Grain
    Once the business process has been identified, the data warehouse team faces a serious decision about the granularity. What level of data detail should be made available in the dimensional model? This brings us to an important design tip.

3. Choose the Dimensions
Once the grain of the fact table has been chosen, the date, product, and store dimensions fall out immediately.
4. Identify the Facts
The fourth and final step in the design is to make a careful determination of which facts will appear in the fact table. Again, the grain declaration helps anchor our thinking.

4. Design Data Warehouse

4.1. Data Warehouse
Data Warehouse is a special database that is used as a "data warehouse" or data which has been consolidated from various data sources of existing information systems in an organization/company. According to Kimball, there are some requirements for the data warehouse, some of which are :
  • Data warehouse must make information from a company/institution/organization easily accessible.
  • Data warehouse must display information about the company/institution constantly.
  • Data warehouse must present data that will be used as a basis or guidelines for decision making.
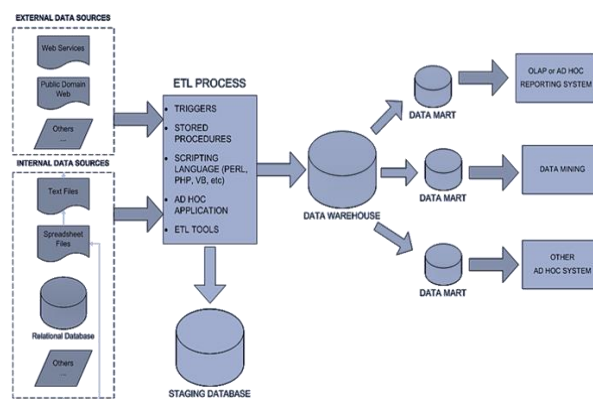

Figure 3. Data Warehouse Architecture
(Source : Kimball, 2002)

4.2. Architectural Design Data Warehouse PTIK.
Source of data to be processed is taken from academic databases PTIK Study Program. Such data contain all the academic data of the students in the study program PTIK that would normally be in upload for each semester to be displayed on the Academic Information System Manado State University. Sorting data is done to sort out what data is to be used without interfering with operational data while in use. The staging process is also done to facilitate the ETL process later, because the data are taken really only the data required for the Data Warehouse GPA student. Here is the data warehouse architecture design PTIK Study Program :
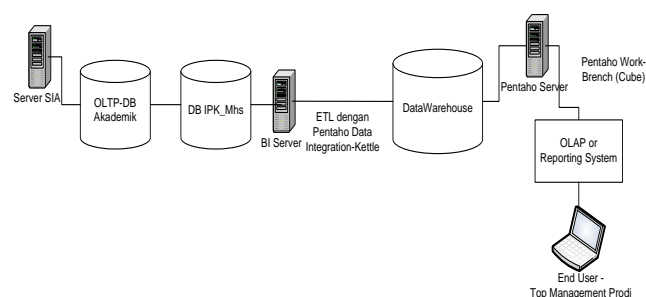

Figure 4. Physical Data Warehouse Architecture PTIK Study Program

4.3. Dimensional Data Modeling

Modeling multidimensional database will consist of fact tables (fact tables) and dimension tables (dimension tables) are interlinked. A fact table contains various value aggregations on which the measurement (measure) as well as some of the key that corresponds to the dimension tables that will be the viewpoint of the measure.

The structure of the fact table and the dimension table has a design scheme that can improve performance and ease of translation.

OLAP systems schema is the basis for doing data warehousing. Two schema most commonly used by the various OLAP engine are the star schema (Star Schema) and scheme snowflakes (Snowflake Schema). In this research multidimensional depiction of data using the Star Schema.

Dimension table is a table that contains the data that show the results of a review of various points of view. Table dimensions will construct cube. Table dimensions are available :

Table dim_mhs

The dimension tables contain data on students who are limited only students' register number (nim), the name of the place and date of birth, gender and year of admission.

Table 1. Students Dimension Table

dim_mhs	
PK	**sk_mhs**
	nim nama_mhs tempat_lahir tanggal_lahir jk thn_masuk kd_prodi

Table dim_dosen

The dimension tables contain data lecturers only limited to the name of the lecturer, and lecturer code. For lecturers code is NIDN.

Table 2. Lecturer Dimension Table

dim_dosen	
PK	**sk_dosen**
	kd_dosen nip nama_dosen Gelar_dosen TempatLahir TanggalLahir JK No_KTP Kd_Prodi Kd_MK Nama_MK

Table dim_mk

The dimension tables contain data subjects, the weight of credits from the courses and in the semester how mk (subjects) exists.

Table 3. Subjects Dimension Table

dim_mk	
PK	**sk_mk**
	kd_mk nama_mk sks semester GanjilGenap Kd_Dosen

Table dim_waktu

The dimension tables contain data such as time of day, quarter, semester, month, year and date.

Table 4. Time Dimension Table

dim_waktu	
PK	**sk_waktu**
	hari
	kuartal
	semester
	bulan
	tahun
	tanggal

table dim_nilai

The dimension tables contain data on the value of each course students each academic year and the first semester and the second semester.

Table 5. Value Dimension Table

dim_nilai	
PK	**sk_nilai**
	nim
	tahun_akademik
	ganjilgenap
	kd_mk
	kelas
	nilai_akhir (0-4)
	grade(A-E)

Fact table is a table containing the facts of business, generally a table the details of transactions that have occurred (Mulyana, 2015) . Fact Table designed in a data warehouse is taken relating to the evaluation of data from PTIK Study Program. Fact_ipk only one fact table is actually what is enough GPA display data from year to year. But considered necessary also for information about the value of each student.

Table fact_nilai

This table contains all the data values PTIK students in each course that is already in the contract.

Table 6. Value Fact Table

fact_nilai	
FK1	sk_mhs
FK2	sk_dosen
FK3	sk_waktu

Here is a picture of a star schema GPA study programs and student grades.
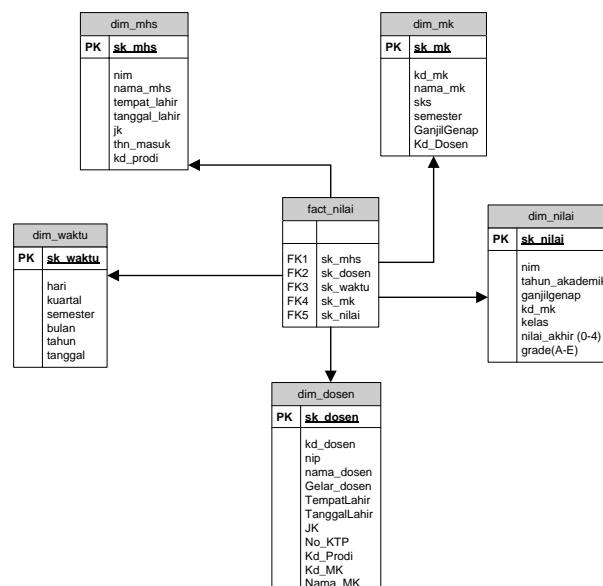
Figure 5. Star Schema Data Warehouse GPA PTIK Study Program

5. Experiment Results

5.1. Pentaho Data Integration

Pentaho Data Integration (PDI) or Kettle ETL utility is open source under Pentaho Corp.Amerika. This project was originally an initiative of Matt Casters, a programmer and consultant Business Intelligence (BI) from Belgium who has managed projects for enterprise BI big.

Currently Kettle ETL is a utility that is very popular and one of the best on the market. Some advantages are as follows :

- Have a collection of data processing modules that quite a lot. More than 100 modules or step.
- Have a module that facilitates the design of the data warehouse model as Slowly Changing Dimensions Dimension and Junk.
- Performance and scalability are well-proven.
- Can be developed with a variety of additional plugins.
- Utility Kettle to be used in the integration of this data using the Spoon.

5.2. Data Integration for Dimension Table

By using PDI Kettle, diversity of data available on PTIK study program can be integrated into the database with the database platform used is MySQL. Tables of this dimension that will form the Data Warehouse GPA for PTIK Study Program.

Dimension table Students (Dim_Mhs)

Data source of dim_mhs table is derived from a list of names of students who enrolled in the first semester of academic year 2014 (1). Where the total registered students is 1293. However, this data is only as names and NIM are stored in Excel files. While the data warehouse requires a complete student data, such as place of birth, date of birth, gender, year in, etc. Therefore, data collaboration with master data that is used in PDPT (Database PT) needs to be done. The PDPT the data is the data Ms.Access. Merging the data source from Excel files and data on Ms.Acces can be done using this PDI. Here is an overview transformation design to use PDI-Kettle.
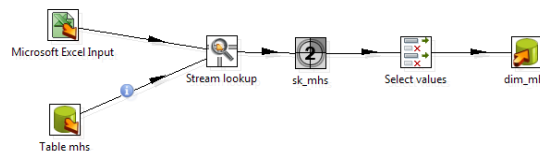

Figure 6. The transformation to a table dim_mhs

Once the transformation is done resulting student dimension tables as follows :

Figure 7. Student Dimension Table

Dimension table Lecturer (dim_dosen)

Data lecturer there in PTIK Study Program (Data Source) is the data stored in Excel files. Merging data is done to obtain data on the complete faculty.
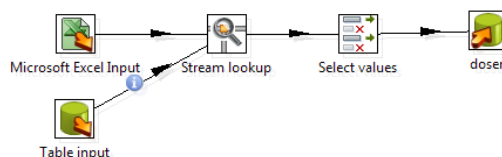

Figure 8. Transformation for table dim_dosen

Once the transformation is done resulting lecturer dimension tables as follows :

Figure 9. Table data from dim_dosen

Table Fact Value (fact_nilai)

Fact tables are created in the transformation by combining the dimension tables that dim_dosen, dim_mk, dim_mhs, dim_nilai, dim_waktu who had previously designed.
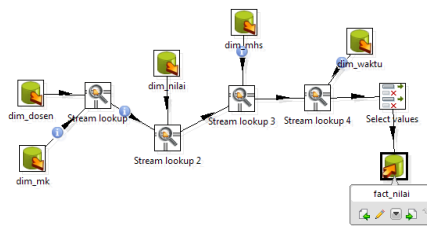

Figure 10. Transformation ETL for table fact_nilai

Table fact_nilai after transformation is executed is as follows :

sk_mhs	sk_mk	sk_dosen	sk_waktu	sk_nilai
21	65	18	20160102	1
21	67	22	20160102	2
21	66	32	20160102	3
21	(NULL)	(NULL)	20160102	4
21	(NULL)	(NULL)	20160102	5
21	(NULL)	(NULL)	20160102	6
(NULL)	66	32	20160102	7
(NULL)	(NULL)	(NULL)	20160102	8
(NULL)	67	22	20160102	9
(NULL)	66	32	20160102	10
203	44	29	20160102	11
203	78	7	20160102	12
203	(NULL)	(NULL)	20160102	13
203	54	17	20160102	14
203	64	4	20160102	15
203	56	3	20160102	16
203	80	20	20160102	17
203	(NULL)	(NULL)	20160102	18
203	81	19	20160102	19
203	66	32	20160102	20
203	43	24	20160102	21

Figure 11. Table Data fact_nilai

6. Conclusion

The diversity of data into an existing data source in Prodi PTIK can be overcome by using Pentaho Data Integration (PDI)-Kettle. Data warehouse design results generated GPA makes the data can be viewed in tabular form a more orderly so easily processed. This is the data that is ready to be processed into the Pentaho schema workbench and then the dashboard can be presented using Pentaho Business Intelligence (BI) Server.

References

[1]   Budi Harijanto, Gunawan Budiprasetyo. Perancangan Aplikasi Business Intelligence Hasil Proses Belajar Mengajar (Studi Kasus Program Studi Manajemen Informatika). Jurnal ELTEK, Vol.11 No.1, April 2013.
[2]   "CIBIA Courseware", 2015, Multimatics, Jakarta.
[3]   Imelda. Business Intelligence. Majalah Ilmiah UNIKOM Vol.11 No.1 2015.
[4]   JRP, Mulyana., 2014, "Pentaho: Solusi Open Source untuk Membangun Data Warehouse", ANDI, Yogyakarta.
[5]   Kimball, Ralph., Ross, Margy., 2002, "The Data Warehouse Toolkit, Second Edition", John Wiley & Sons, Inc, Canada.

[6]   Kimball, Ralph., Caserta, Joe., 2004, "The Data Warehouse ETL Toolkit", John Wiley & Sons, Inc, Canada.[online].

[7]   Meta Suzanam, Jemakmun, Suyanto. Analisis dan Perancangan Data Warehouse Rumah Sakit Umum di Daerah Palembang Bari. Jurnal Ilmiah dan Teknik Informatika Universitas Binadarma Vol1.No.1 November, 2013.

[8]   Prastuti Sulistyorini. Business Intelligence dan Manfaatnya Bagi Organisasi. Majalah Ilmiah IC Tech Vol.5 No.2 Mei,2010.

[9]   Stephanie Pamela Adithama, Irya Wisnubhadra, Benyamin L. Sinaga. Analisis dan Desain Real-Time Business Intelligence untuk subjek Kegiatan Akademik pada Universitas menggunakan Change Data Capture. SENTIKA Yogyakarta, 9 Maret 2013.

[10]   Thia Feris, "Pentaho Knowledge Based", [online], (http://pentaho-en.phi-integration.com)