Data analysis and data warehouse design based on Pentaho data integration (kettle) to support the determination of student learning achievement

A D Barahama* and R Wardani
Electronics and Informatics Engineering Education, Post-Graduate Program, Yogyakarta State University Jalan Colombo 01, Daerah Istimewa Yogyakarta 55281, Indonesia

*ameilianidianira.2019@student.uny.ac.id

Abstract. Data analysis has become a vital requirement in various fields. The results of data analysis can be used as a data warehouse design that is expected to help in solving problems, evaluating learning outcomes and supporting decision making. In the field of education, student achievement is a reference for achieving quality learning. This success can be seen from the achievement of student learning completeness that is the basis for evaluating and making decisions. The data warehouse design can be used as a basis in knowing student learning progress that can be seen from the value obtained. The diversity of data makes managing values difficult and delayed, data warehouse design using Pentaho can help and simplify data integration. The results of the analysis and design of the data warehouse will be presented in a multidimensional form that can be seen through dimension tables and fact tables.

1. Introduction
Data analysis is a very important part and has become a very vital requirement in various fields. In the digital age, information exchange takes place so quickly, data processing for generating information is a very crucial thing to do. Technological success, causing swelling of data so that it will cause difficulties and require time long time in its management. Data processing problems are also experienced by institutions education, one of them is school. in the field of education, student achievement becomes a reference from the achievement of quality learning, success can be seen from the achievement of completeness student learning. which is the basis for evaluating and making decisions.

Achievement of students can be seen from the progress of student grades obtained in each subject. to obtain information on the value of each student it is necessary to value data analysis and design data warehouse that can be accessed quickly and accurately. The results of data analysis and data warehouse design can overtake the delay in obtaining information. the diversity of data sources will be overcome 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. Data Warehouse is a special database that is used as a "data warehouse" or data which has been consolidated from various information systems data sources in an organization or company [1]. Data integration and in particular data warehouses are central to these efforts. Data integration and in particular data warehouses are central to these efforts. In this context, the database system will be
processed and then organized to integrate different data into a common layout that efficiently supports complex analysis [2] data warehouse can provide a data transaction process ETL (extract-transform-load) that is specifically structured for integrating data from multiple heterogeneous information sources and transforming them into a multidimensional representation for decision support applications [3] with the data warehouse, the school will be helped to evaluate, plan and make decisions related to student achievement in school.

2. Methods
Design method multidimensional data modeling according Kimball used includes 4 stages [4].

2.1. Select the business process
Business processes are operational activities carried out by your organization, such as take orders, process insurance claims, register students for classes, or take a picture of each account every month. Business process events produce or capture performance metrics that translate into facts in the fact table.

2.2. Declare the grain
After the business processes are identified, the data warehouse team faces a serious decision problem about the details. What level of data detail should be available in the dimension model. Step important in the design of dimensions is to assign representatives to this Consistency fact table enforce uniformity on all design dimensions that are essential for performance BI application and ease of use.

2.3. Identify the dimensions
Dimensions provide the context of "who, what, where, when, why and how" around business process events. Dimension table contains descriptive attributes used by BI applications to filter and group facts. Taking into account the fact table, all possible dimensions that can be identified.

2.4. Identify the facts
Facts are measures that are generated from business process events and are almost always numerical. Single fact table rows have a one-to-one relationship with measurement events such as which is explained by fact table items. So the fourth step is to determine the fact table related to physical events that can be observed.

3. Results and discussion

3.1. Data warehouse
This phase involves the building of the data warehouse, which includes the identification of the purpose of the warehouse, the identification of the data sources and determination of the different formats for those sources, the identification of the dimensions and measures, the determination of the granularity of the data warehouse and the actual design of the data warehouse. For the study, the Kimball Data Warehouse Model is used, as this advocates a bottom-up approach in building data warehouse, which is suitable for independent repositories such as the implementation of Institute’s information systems [5].

Data warehouses exist to facilitate decision support in organizations. Support system decisions help users with strategic analysis and decision making. According to Kimball, there are several requirements for data warehouses, some of which are:

- Data warehouse must be able to easily make information from companies/institutions/organizations accessed.
- Data warehouse must display information about the company/institution constantly.
- The data warehouse must present data that will be used as a basis or guidance for decision making.
3.2. Architectural design data warehouse school academic value

Source of data that will be processed comes from the results of data analysis of student grades at school. Data contains the total number of students in school, grade levels and majors of interest, and grades from each subject obtained, from each semester. This data will then process and sorting is done to facilitate the ETL process. The following is a data warehouse design for Academic Value in school.

Figure 1. Data Warehouse architecture [6].

Figure 1. Describe the architecture of data warehouse data then processed in the ETL system to produce dimension table and fact table or OLAP. the results of the data warehouse can be presented on the pentaho BI server.

Figure 2. Physical data warehouse architecture academic value in school.

Figure 2. Describe the architecture of school data warehouse academic values. Value data derived from the school's academic information system is processed by OLTP (Online Transaction Processing) to produce a database that will be processed on a BI server or by ETL the process of producing a data warehouse so that it can be presented on a Pentaho BI server via OLAP (Online Analytical Processing) or reporting system.
3.3. Dimensional data modeling

Dimensional data modeling is a collection of concepts and techniques used to design and store data, has a unique structure and is very often known as a star schema [7]. In this dimensional data modeling consists of fact tables and dimensional tables related to each other. The fact table contains a combined key from the dimension table on which to base measurement, each dimension is stored in one dimension of the table and each input given a unique identification. The fact table structure and dimension tables provide levels of performance and easy to understand. Dimension table or dimension table contains attributes or fields that provide special information. During its development, the arrangement of fact tables and dimension tables This has a standard design or schema because it is proven to improve performance and ease of translation into OLAP systems. OLAP (On-line Analytical Processing) servers can be relational systems or multidimensional systems. A relational OLAP is an extended relational system which aims to map operations on dimensional data to standard relational operations [8]. In dimensional data modeling consists of student dimensions, time dimensions, grades dimensions, and subjects dimensions.

3.4. Fact table

In this fact table modeling consists of several dimension tables and one fact table that is a table of values, The following are pictures of academic schemes and student grades.

![Star schema data warehouse academic](image)

Figure 3. Star schema data warehouse academic.

The application used to see the results of the design and application of Business intelligence Pentaho based is Pentaho BI server. Pentaho BI server is a web application is a work to do a variety of work related to the problem Business Intelligence. Pentaho BI server is a web based java server application that is on there are solutions and integration of components of Pentaho products such as Pentaho data integration (Kettle), Pentaho data integration (PDI) or the ETL kettle utility are open source at under Pentaho Corp. America. ETL process is the important process to built data warehouse. A data warehouse is needed to store all transaction data (OLTP) before it analyzes into On Line Analytical Processing (OLAP) [9]. ETL (Extraction, Transformation, Load) system: The system in charge extract data from internal transactional databases and other sources (such as external data), change it to accommodate the data warehouse scheme, initially loading the data and periodically refresh data. Design of ETL systems is generally the most task time consuming data warehouse development [10].
3.4.1. Dimension table_student. The data source for the dim_siswa table comes from the list of students who registered according to the year Academically, the data consists of the student's name, student number, year of entry, majors, and grades for each subject every term. The data is then processed using Pentaho Data Integration (kettle). The following is the result of the dim_siswa transformation.

![Figure 4](image1.png)

Figure 4. The transformation to a table dim_siswa.

Figure 4. Describe the process of transformation student table produce student dimension.

3.4.2. Dimension table_value. In the transformation table there is normalization, for the value of each subject that is generate dimension value.

![Figure 5](image2.png)

Figure 5. The transformation to a table dim_nilai.

Figure 5. Describe the process of transformation value table to produce value dimension.

3.4.3. Table fact value (fact_value)
The fact table combines several dimension tables consisting of: dim_siswa, dim_mapel, dim_nilai, dim_kelas, dim_jurusan, dim_katagori, dim_waktu previously designed.

![Figure 6](image3.png)

Figure 6. Transformation ETL for table fact_value.

Figure 6. Describe the process of transformation dimension table to produce fact tables (fact Value).
Table fact_value after transformation is executed is as follows:

![Table](image)

Figure 7. Table data fact_value.

Figure 7. describe the results of the transformation of the fact table presented in the data warehouse.

4. Conclusion

Data warehouse is a technology developed to overcome problems in data so that its use is more directed, efficient and useful. The application of data warehouse in the field of education provides convenience and good impact in the management and integration of data. The results of data analysis and data warehouse design using Pentaho data integration (kettle) makes it easy for the school because it produces information that can be seen from a more organized table, from the data warehouse design the school is more easily made to determine student learning achievement. Then the results of the data warehouse design are ready to be processed use Pentaho BI server which will provide reports and information for the school so that it can be helped to plan, evaluate learning outcomes, and support decision-making.

Acknowledgment

The authors thank Dr. Ratna Wardani M.T in Yogyakarta State University for advice in writing of this manuscript. This study was supported in part by SMA Negeri 9 Manado.

References

[1] Salaki R J, Waworuntu J, and Tangkawarow I R H T 2016 Extract transformation loading from OLTP to OLAP data using pentaho data integration IOP Conf. Ser. Mater. Sci. Eng. 128(1)
[2] Prasser F, Spengler H, Bild R, Eicher J, and Kuhn K A 2019 Privacy-enhancing ETL-processes for biomedical data Int. J. Med. Inform. 126 November 2018 pp. 72–81
[3] Subramanian G H and Wang K 2019 Systems Dynamics-Based Modeling of Data Warehouse Quality J. Comput. Inf. Syst. 59(4) pp 384–391
[4] Kimball R & Ross M2013 The Data Warehouse Toolkit (The Definitive Guide to Dimensional Modeling)
[5] Lapura E V F, Fernandez J K J, Pagatpat M J K, and Dinawanao D D 2018 Development of a University Financial Data Warehouse and its Visualization Tool Procedia Comput. Sci. 135 pp 587–595
[6] Kimball R and Caserta J2004 The Data Warehouse ETL Toolkit (Parctical Techniques for Extracting, Cleaning, Conforming, and Delivering Data
[7] Sulaiman N S and Yahaya J H 2013 Development of Dashboard Visualization for Cardiovascular Disease based on Star Scheme Procedia Technol. 11 no. Iccii pp 455–462
[8] Oliva S Z and Felipe J C 2018 Optimizing Public Healthcare Management Through a Data Warehousing Analytical Framework IFAC-PapersOnLine 51(27) pp 407–412
[9] Runtuwene J P A and Tangkawarow I R H T 2017 Vertical information system: A case study of
civil servant teachers data in Manado city Indones. J. Electr. Eng. Comput. Sci. 6(1) pp 42–49

[10] Deshpande P M and Ramasamy K 2005 Data warehousing, multi-dimensional data models and OLAP Encycl. Database Technol. Appl. pp 134–139