Utilization Extract, Transform, Load For Developing Data Warehouse In Education Using Pentaho Data Integration

A D Barahama¹ and R Wardani²

¹,²Department of Electronics and Informatics Engineering Education, Post-Graduate Program, Universitas Negeri Yogyakarta Jalan Colombo 01, Daerah Istimewa Yogyakarta 55281, Indonesia.

Email: ameilianianinanra.2019@student.uny.ac.id

Abstract. The utilization of data warehouses in various fields is an absolute necessity. A data warehouse is a database that contains large amounts of data that aims to help organizations, fields, and institutions specifically for decision making. Data warehouses can produce important information in the future. Loading data from various sources and processed through an ETL (Extract, Transform, Load) process that displays data consistently is the basis for creating a data warehouse architecture. The development of a data warehouse in education will provide significant benefits for the progress of education. Integration of data and processing results stored in the data warehouse can be the basis for evaluating better planning. Development of data warehouse adopt the multidimensional modelling method which consists of four stages: select the business process, declare the grain, select dimensions, and identify facts. This stage produces a data warehouse architecture and influences and contributes to the advanced information technology in education.

1. Introduction
Supporting the achievement of Indonesian education can be measured through 6 indicators: students, teachers, infrastructure, costs, curriculum, and involvement. Elements of society are interrelated and have their respective duties and roles as educators and teachers. The availability of educators must be adequate in each region, one of which is the implementation of technology information. The use of information technology in data integration is a vital need in education. The data integration process can produce information that can help users find appropriate and efficient solutions to be used as a basis for decision making. The availability of data will support success in the field of education. Extract-Transform-Load has the task and is responsible for integrating data into a data warehouse [1]. Process implementation data into information can be used as a basis for decision making and is helpful for knowledge system management in organizations to improve human resources and information technology. [2].

Data availability can be support by the existence of a Data Warehouse which aims to process, store and manage large amounts of data for analysis and decision-making needs. [3] The analytical process stored in the data warehouse will be helpful in the future as a support in obtaining efficient decision making [4] The data warehouse will provide sources of information that work directly in solving problems so that they find the solution [5]. This research will use the Pentaho Data Integration-kettle as a tool that makes it easier to overcome the diversity of data in education to produce a data analysis
process for educators. Data integration through extracts, transformations, loading in the development of
Data Warehouses in education can assist in conducting evaluations, planning and an accurate basis to
support decision making in the future [6].

2. Methods

2.1. Preparing Data source
In developing a data warehouse, the basic thing needed is to prepare data sources. Data source will be
processed with the stages of data cleaning, data selection so that the data is ready to be transformed into
a data warehouse. Preparation of data sources will be used in development of data warehouse.

2.2. Design of data warehouse
The basis of data warehouse development is to have dimensional data modeling. According to Ralph
Kimball and Joe Caserta in their book The Data Warehouse ETL Toolkit (practical techniques for
extracting, cleaning, and delivering data), there are four stages in developing a Data Warehouse.[7]

2.2.1. Select the business process
In the first stage, the thing to do is identify the requirements that must be available, the data sources to
be processed, understand what kind of modeling and requirement are needed.

2.2.2. Declare the grain
The second stage is the process of identifying the problems found. What are the right dimension and the
right way to solve the problem.

2.2.3. Choose the dimension
This stage is the stage to choose the appropriate dimensions so that it can produce a fact table.

2.2.4. Identify the facts
This stage is the stage to choose the dimensions so that it can produce an appropriate fact table. This
stage is the last step of dimensional data modeling, the stage to produce a fact table which will be the
core of dimensional data modeling from data warehouse development.

2.3. Dimensional data modeling
Multidimensional modeling is an old technique for creating simple databases. Dimensional data
modeling will produce several dimensions and fact tables, which will explain the flow of the data
integration process into a data warehouse.

2.4. Extract, Transform, Load – ETL
ETL is the process of extracting data from various sources, transforming it and loading it into a data
warehouse. The ETL process is an important step, the data stored in the data warehouse can be used as
a source to obtain information that can be used as a basis for decision making or as a knowledge
management system within the organization. In the ETL process, the main things to do are extract, clean,
conform and deliver. There are important parts in the ETL process namely: XML data collection, using
an independent DBMS, having a schema, normalized entity/relationship (E/R), and dimensional data
modeling [7]. ETL plays an important role in the data interaction process because it can help determine
the outcome of the problem so that it can get the appropriate retrieval [2]. ETL can support the
Identification of relevant information from its source, perform information extraction, customization,
and integration from various sources, perform cleaning and load it in the big data into a Data Warehouse.
[8]. The purpose of the ETL process is described in figure 1:
Figure 1. The proposed ETL tool [9]

Figure 1. Describes the integration process starting from extraction, transformation where the data is then cleaned, processed in meta data and loaded into the data warehouse.

2.5. Result of data warehouse

Experiment results from the initial stages of preparing data source, design of data warehouse, star schema until the completion of the extract, transform, load stages will produce a data warehouse that is ready to be used.

3. Result and discussion

3.1. Data source

In developing the data warehouse, data source comes from raw data in Microsoft Excel format containing educational data, namely teacher and educator data, then cleaning the selection data and adding data types until the data is ready to be integrated through the extract, transform, load process. The following illustrates the change of raw data into data that is ready to be integrated into the Extract, Transform, Load (ETL) process.

Figure 2. Modified data source table

Figure 2. Describes the raw data table that has been processed and cleaned to prepare the data for the ETL process.

3.2. Design of data warehouse

The data warehouse produces from the completion of the multidimensional modeling stage. A data warehouse is a specific database that generates from various sources and information systems of
organizations, fields of companies, and institution [12]. A data warehouse makes as a multidimensional database that contains large amounts of data that can be organized through several dimensions and fact tables and becomes the basis for an organization, company in the decision-making process[13]. A data warehouse is an organized collection of data that aims to provide the right solution so that it helps in the decision making [14]. According to William H., Inmon Data warehouse is a collection of subject-oriented, integrated, time-variant, and non-volatile data to support the decision-making. [15]. The four main features of a data warehouse are:

1. Subject Oriented: A data warehouse is organized around the main subject concentrating on the day-to-day operations and transaction processing of an organization, data Warehouse focuses on modeling and analyzing data for decision-making.
2. Data warehouses are built by integrating several heterogeneous sources, such as relational databases, files, and online transaction records.
3. Time Variant: Data is stored to provide information from a historical perspective (e.g., the last 5-10 years)
4. Nonvolatile: A data warehouse is always a physically separate data mart that is the transformation from the application data found in the operational environment. The data warehouse is not updated in real-time but is refreshed from the operating system regularly continually and incrementally to be integrated with previous data.

Figure 3. General Data Warehouse [12]

Figure 3. Describes the subject-oriented data warehouse architecture and comes from different sources integrated into a single unit through ETL (extract, transform, load) so that it can achieve the goal of producing a data warehouse.

3.2.1 Architectural design data warehouse for teacher in education.
Extraction, transformation, and loading (ETL) and creates an improved and consistent appearance materialized from a collection of data sources that will be used as the basis for the data warehouse architecture. [8] The construction of a data warehouse for educators in education source from the educational database, which is the process through Microsoft Excel and illustration of data will be processed using Pentaho Data Integration tools. The results of the extraction that depicted out produce an architecture that can explain as follows:
Figure 4. Describes the data warehouse architecture for educators in education, data sources originated from information systems are then utilized and integrated through extraction transformation and loading to produce a special database in the form of a data warehouse which is a necessity to assist stakeholders in institutions education for decision making.

3.3 Dimensional data modeling

Extraction, transformation, and loading (ETL) results create dimensions that will be related to each other to produce a fact table, these dimensions are generated using Pentaho data interaction tools through excel input data sources and generate databases stored in SQL yog, and have a surrogate key that is a key containing an integer value.

3.4 Extract-Transform-Load (ETL)

Extract-Transform-Load (ETL) process that has been used in computing for integration of data from multiple sources. This process refers to the development of a data warehouse that extracts data from outside sources, transforms it according to operational needs, and then loads it in a data mart or data warehouse[10]. The three phases of this process are described as follows:

1. The first process of this phase is the extraction of data available in formats such as xls, cvs and txt. the procedure of extracting data from different sources is done by reading the data and purifying the information received to be used as guidelines and metadata.[11]
2. The second phase is data cleaning carried out according to a predetermined scheme, involving normalization of data, deleting data, adding data, sorting, and grouping them according to need.
3. The third phase aims to load the data into a data mart or data warehouse that is ready for use. Composing the data into the objective database. ETL process move data starting with one
database then onto the next, to shape data marts and data warehouse furthermore to change over databases starting with one organization or institution [11]

Utilization of extract, transform, load in data integration to develop teacher data warehouse using Pentaho data integration aims to facilitate data processing can be used in the future for analysis, evaluation, and decision-making purposes. The following illustrates the integration of teacher data through the ETL process.

![Figure 6](image6.png)

Figure 6. Output dimension teacher from ETL process Pentaho data integration tools.

Figure 6. Describes the ETL process from several tables that are determined to be processed to produce teacher dimensions, connected to the stream lookup, selected with select values, a database query is carried out which will produce the teacher dimensions. The results of the process carried out can be seen through the data preview shown in Figure 7.

![Figure 7](image7.png)

Figure 7. Preview data qualification dimension

Figure 7. Preview data from one dimension that is closely related to the data for educators, a data contains the qualification dimension. Explains the resulting dimension table, in this case the data selected to be displayed is one of the dimensions related to the teacher dimension, namely the qualification dimension.

The final result of each dimension that has been generated is to combine them to produce a fact table. The fact table contains a collection of dimensions that will produce a fact table and display dimensional data modeling. The following is the result of data integration that combines the existing dimensions into a fact table:
Figure 8. Describes the dimensions associated with stream_lookup and performs metadata on the attributes generated from the dimensions using select values will generate fact_tables. will display data in the form of preview data which can be seen in figure 9.

Figure 9. The result - transformasi fact tabel

4. Conclusion
Thereby, the building of a data warehouse requires the execution of the Extraction-Transformation-Load (ETL) process. Integration of data from different sources forms a data warehouse that can be used as a core source in obtaining accurate information in making decisions for the advancement of education. The data warehouse offers a summary of operational data that is simple and easy for management to understand if needed. The process carried out in building a data warehouse for teachers and educators in the field of education using Pentaho data integration tools facilitates data processing and analysis. The source of data obtained from raw data is then processed following the stages of extract, transform, load to produce several dimensions and fact tables which are the key to the development of this data warehouse. Thus, the development of this data warehouse will make it easier for decision-makers to
determine policies, facilitate the development of DSS and EIS applications, support a better analysis process for decision-makers for the advancement of education.

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 Dinas Pendidikan Provinsi Sulawesi Utara.

References
[1] P. S. Diouf, A. Boly, and S. Ndiaye, “Variety of data in the ETL processes in the cloud: State of the art,” 2018 IEEE Int. Conf. Innov. Res. Dev. ICIRD 2018, no. May, pp. 1–5, 2018, doi: 10.1109/ICIRD.2018.8376308.
[2] P. S. Diouf, A. Boly, and S. Ndiaye, “Performance of the ETL processes in terms of volume and velocity in the cloud: State of the art,” 4th IEEE Int. Conf. Eng. Technol. Appl. Sci. ICETAS 2017, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICETAS.2017.8277875.
[3] O. Moscoso-Zea, J. Paredes-Gualtor, and S. Luján-Mora, “A Holistic View of Data Warehousing in Education,” IEEE Access, vol. 6, pp. 64659–64673, 2018, doi: 10.1109/ACCESS.2018.2876753.
[4] G. Blaži, P. Poš, and D. Jakši, “Data Warehouse Architecture Classification,” pp. 1491–1495, 2017.
[5] I. I. Kholod and M. S. Efimova, “Using ETL Tools for Developing a Virtual Data Warehouse,” pp. 351–354, 2016.
[6] A. D. Barahama and R. Wardani, “Data analysis and data warehouse design based on Pentaho data integration (kettle) to support the determination of student learning achievement,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1098, no. 5, p. 052089, 2021, doi: 10.1088/1757-899x/1098/5/052089.
[7] Ralph Kimball & Joe Caserta, The Data Warehouse ETL Toolkit (Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data). 2004.
[8] S. Bergamaschi, F. Guerra, M. Orsini, C. Sartori, and M. Vincini, “A semantic approach to ETL technologies,” Data Knowl. Eng., vol. 70, no. 8, pp. 717–731, 2011, doi: 10.1016/j.datak.2011.03.003.
[9] S. Souissi and M. Benayed, “GENUS: An ETL tool treating the Big Data Variety,” Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA, vol. 0, 2016, doi: 10.1109/AICCSA.2016.7945615.
[10] J. Chakraborty, A. Padki, and S. K. Bansal, “Semantic ETL-State-of-the-Art and Open Research Challenges,” Proc. - IEEE 11th Int. Conf. Semant. Comput. ICSC 2017, pp. 413–418, 2017, doi: 10.1109/ICSC.2017.94.
[11] L. Fu, “A Recommendation System Using OLAP Approach,” pp. 622–625, 2016, doi: 10.1109/WI.2016.108.
[12] R. J. Salaki, J. Waworuntu, and I. R. H. T. Tangkawarow, “Extract transformation loading from OLTP to OLAP data using pentaho data integration,” IOP Conf. Ser. Mater. Sci. Eng., vol. 128, no. 1, 2016, doi: 10.1088/1757-899X/128/1/012020.
[13] G. Garani, A. V. Chernov, I. K. Savvas, and M. A. Butakova, “A Data Warehouse Approach for Business Intelligence,” Proc. - 2019 IEEE 28th Int. Conf. Enabling Technol. Infrastruct. Collab. Enterp. WETICE 2019, pp. 70–75, 2019, doi: 10.1109/WETICE.2019.00022.
[14] A. Lamani, B. Erraha, M. Elkyal, and A. Sair, “Data mining techniques application for prediction in OLAP cube,” Int. J. Electr. Comput. Eng., vol. 9, no. 3, pp. 2094–2102, 2019, doi: 10.11591/ijece.v9i3.pp2094-2102.
[15] J. P. Jiawei Han, Micheline Kamber, Data mining: Data mining concepts and techniques, 3rd ed. San Fransisco, 2014.