A Comparative Analysis of Extract, Transformation and Loading (ETL) Process

A Comparative Analysis of Extract, Transformation and Loading (ETL) Process

J P A Runtuwene, I R H T Tangkawarow*, C T M Manoppo and R J Salaki
Universitas Negeri Manado, Tondano 95618, Sulawesi Utara, Indonesia

*irene.tangkawarow@unima.ac.id

Abstract: The current growth of data and information occurs rapidly in varying amount and media. These types of development will eventually produce large number of data better known as the Big Data. Business Intelligence (BI) utilizes large number of data and information for analysis so that one can obtain important information. This type of information can be used to support decision-making process. In practice a process integrating existing data and information into data warehouse is needed. This data integration process is known as Extract, Transformation and Loading (ETL). In practice, many applications have been developed to carry out the ETL process, but selection which applications are more time, cost and power effective and efficient may become a challenge. Therefore, the objective of the study was to provide comparative analysis through comparison between the ETL process using Microsoft SQL Server Integration Service (SSIS) and one using Pentaho Data Integration (PDI).

1. Introduction

Business Intelligence (BI) is concept of utilizing large number of corporate data (Big Data) processed in such a way to produce useful information. Companies used such information to improve decision-making quality based on data-based existing system. In order to run BI, one should have data warehouse that can store the entire data from various data source and different types of data well-known as Big Data [1]. To develop data warehouse, data source should go through a process called Extract Transformation and Loading (ETL).

In the current development of BI, many vendors have developed tools that are able to perform ETL easily. This study will use 2 tools, one is developed by Microsoft called Sql Server Integration Service (SSIS) and the other is developed by Pentaho called Pentaho Data Integration (PDI) or Kettle. In practice, users find it difficult to select which tolls that can run the ETL to get data warehouse effectively and efficiently. In fact, a lot of developers are wasting too much labor, time and cost because they select unsuitable ETL tools.

The objective of the study is to compare between the use of SSIS and PDI. Using Northwind database as the data source, the researchers will observe ETL process in both BI tools mentioned previously. Comparison scenario is determined to facilitate the comparative analysis between the two tools. This research was conducted to give recommendation for BI developer based on the result of comparative study between Microsoft SSIS and PDI. The goal of this research is to compare Microsoft SQL Server Integration Service and Pentaho Data Integration in running ETL process and integrating Northwind database by using Microsoft SQL Server Integration Service and Pentaho Data Integration.

2. Research Methods

The following was the scenario that compares between the ETL process from Microsoft SQL Server Integration Service and Pentaho Data Integration (Kettle). Before the comparative study was conducted, the researchers prepared the specimen/ object for comparison. The specimen was architecture design using Microsoft SQL Server Integration Service and one using Pentaho Data Integration (Kettle). Furthermore, data warehouse that would be used to store the result of integration should be prepared. Figure 1 described the design of both BI tools architecture.

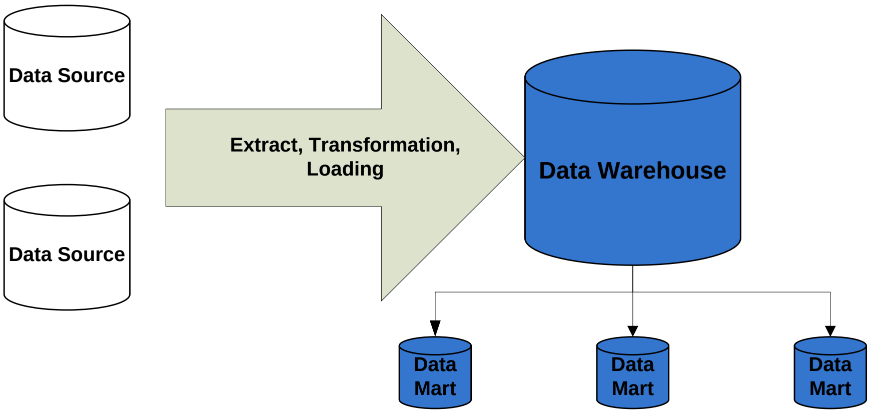

Figure 1. BI Architecture.

Comparative scenario conducted using Microsoft SQL Server Integration Service were as follow:

a. Data Collection on Northwind Database (Extract)
System data input was in the form of Northwind Inc data sales. This data were stored in SQL Server 2008 database.

b. Data Transformation Process
The data from this source was extracted into the same data format, SQL Server 2008. Before data were stored into the data warehouse, some adjustment was conducted in order that the data meet the requirement of the data warehouse designed previously.

c. Loading Data Processes into Data Warehouse
The following procedure was to store/ import (loading) the data into data warehouse using data import feature in SQL Server 2008.

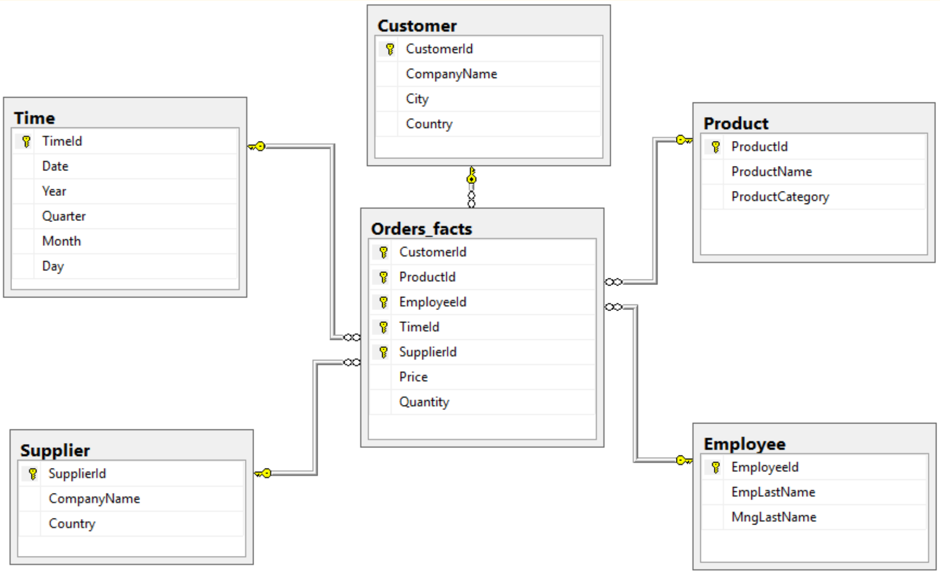

Figure 2. Star Schema Northwind DB.

Comparative scenario conducted using Pentaho Data Integration were as follow:

a. Data Collection on Northwind Database (Extract)
System data input was in the form of Northwind Inc data sales. This data were stored in SQL Server 2008 database.
b. Data Transformation Process
The data from this source was extracted into My SQL data format. Before data were stored into the data warehouse, some adjustment was conducted in order that the data meet the requirement of the data warehouse designed previously.
c. Loading Data Processes Into Data Warehouse
The following procedure was to store/ import (loading) the data into data warehouse using data import feature in Pentaho Data Integration.

The following was a star schema for Extract, Transform and Loading process using Microsoft SQL Server Integration Service and Pentaho Data Integration.

2.1. Comparative criteria

Somya [2] mentioned 5 criteria for comparative analysis, namely:

2.1.1. Comparison from developer perspective. Comparison from developer perspective would compare implementation of both architectures. The perspective would compare what process researchers should take to apply each of the architectures. The finding was the more user-friendly type of architecture.

2.1.2. Comparison from user perspective. Comparison from user perspective would compare how users add and integrate data into the data warehouse. The finding was which type of architecture of which integration process was more user-friendly.

2.1.3. Comparison from system development perspective. The type of comparison would contrast how each type of architecture be developed in the future. The result was the type of architecture of which system was easier to develop.

2.1.4. Comparison from system scalability. It would compare support for various data sizes that were integrated into the data warehouse.

2.1.5. Comparison from system stability. It would compare how stable system was in supporting data integration process, more particularly data integration of various different data.

This study focused on the comparison between the developer, user and system development perspectives.

2.2. Previous related studies

The following was list of previous studies related to data warehouse and business intelligence.

1. A journal entitled Microsoft SSIS and Pentaho Kettle: A Comparative Study for Three-Tier Data Warehouses [3], described the differences between Microsoft SSIS and Pentaho Kettle based on the three-tier data warehouse architectures. It discussed differences taking place at the 3 layers, bottom tier, middle tier and top tier. The study did not emphasize on the process or explain general observations of both existing BI tools.
2. A study entitled *Perbandingan Traditional Business Intelligence dan Service-oriented Business Intelligence (SoBI) untuk Integrasi Data Akademik dan Keuangan (Studi Kasus: Universitas Kristen Satya Wacana)* [2] or Comparison between Traditional Business Intelligence and Service-oriented Business Intelligence (SoBI) for Integration of Academic and Financial Data (Case Study: Satya Wacana Christian University) compared the use of

SoBI and traditional BI for academic and financial data of SWCU . The purpose of the study was to produce a graphical dashboard.

3. A Journal of which title was A Service-oriented Architecture for Business Intelligence [4] discussed comparison between service-oriented concepts in BI and the traditional BI concept. The literature study concluded that with the concept of service-oriented in BI, the process of receiving information and technology integration became more effective compared to doing so with the traditional BI concept.
4. A previous study on data warehouse and data mining was a study entitled *Penggunaan Data Warehouse dan Data Mining untuk Data Akademik Sebuah Studi Kasus Pada Universitas Nasional*, the use of Data Warehouse and Data Mining for Academic Data: A Case Study in National University [5]. It examined the extraction of operational data into data warehouse, and then continued with data analysis using data mining techniques. The finding was data warehouse and web-based information reporting application. In addition, the results of the data mining were pattern of student characteristics taking a particular interest.
5. Extract transformation from OLTP to OLAP data using pentaho data integration was a research on bottom tier level which explained in detail ETL process using Pentaho Data Integration [6].
6. Vertical Information System: A Case Study of Civil Servant Teachers' Data in Manado [7] was a study using Pentaho tool in analyzing and designing data warehouse on data related to teachers in Manado
7. A Report on Implementation of Business Intelligence in the Ministry of Finance, Capital Market and Financial Institution Supervisory Agency in 2007 determined important aspects to understand prior to preparing for and applying Business Intelligence (BI) system within the BPPM and Financial Institutions [8].
8. Developing Business Intelligence Application Design for Distribution Information Systems in Pertamina Lubricant LLC using Pentaho (6) resulted in BI applications using Pentaho [9]. The first phase was to determine input. The input was derived from data that had been recorded in the distribution information system of Pertamina Lubricant SR III LLC, namely transactional data, master data, and reference data. The second phase was system design where input data source was integrated in one data warehous through ETL (Extract, Transformation, Load) process, making cube or data mart logically that consisted of various dimensions of data and fact tables. OLAP was used as the method for data analysis.
9. A study entitled Data Warehouse Design and Processing in the Library at STIMIK AMIKOM Yogyakarta [10] used SQL Server Integration Service (SSIS), SQL Server 2005 tool. The study consisted of 7 stages. A study entitled Big Data, Data Analysis, and Development of Librarian Competence was a library research [1].
10. White paper entitled Microsoft SSIS and Pentaho Kettle: A Comparative Study for Three-Tier Data Warehouses [3] described the differences between Microsoft SSIS and Pentaho Kettle based on three-tier data warehouse architecture. It discussed the differences in 3 layers, bottom tier, middle tier and top tier. It did not emphasize on the process or explain general observations towards both existing BI tools.

3. Results and Discussion

3.1. Comparative scenario using Microsoft SQL Server Integration Service

In the section, the Northwind database was transferred into data warehouse using Microsoft SQL Server Integration Service. The steps were as follow:

a. Selecting data from the Northwind Database (Extract)
b. Data Transformation
c. Loading Data into the Data Warehouse

Figure 3 described the transformation design developed in the study.

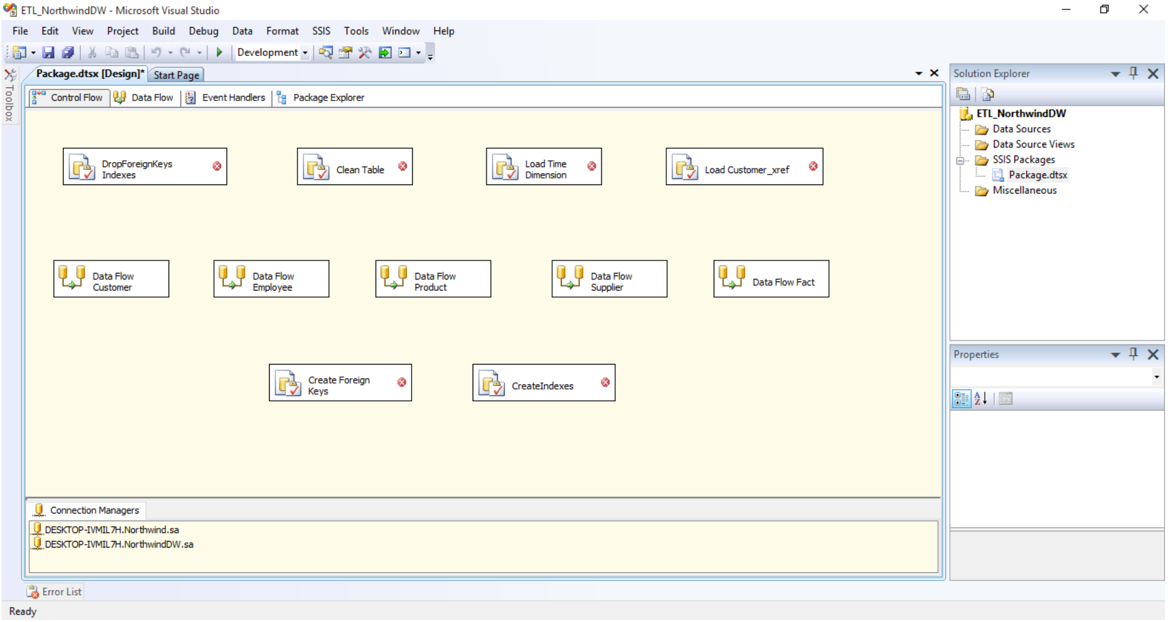

Figure 3. ETL with SSIS.

Error! Reference source not found. had 6 Execute SQL Task, namely DropForeignKeys Indexes, Clean Table, Load Time Dimension, Load Customer_xref, Create Foreign Keys, and CreateIndexes. In addition, there were 5 Data Flow Task, Data Flow Customer, Data Flow Employee, Data Flow Product, Data Flow Supplier, and Data Flow Fact.

*3.2. Comparative scenario using Pentaho Data Integ*ration

In the section, the Northwind database was transferred into data warehouse using Pentaho Data Integration. The steps were as follow:

a. Selecting data from the Northwind Database (Extract)

b. Data Transformation

c. Loading Data into the Data Warehouse

Figure 4 described how new file transformation using list of content was developed.

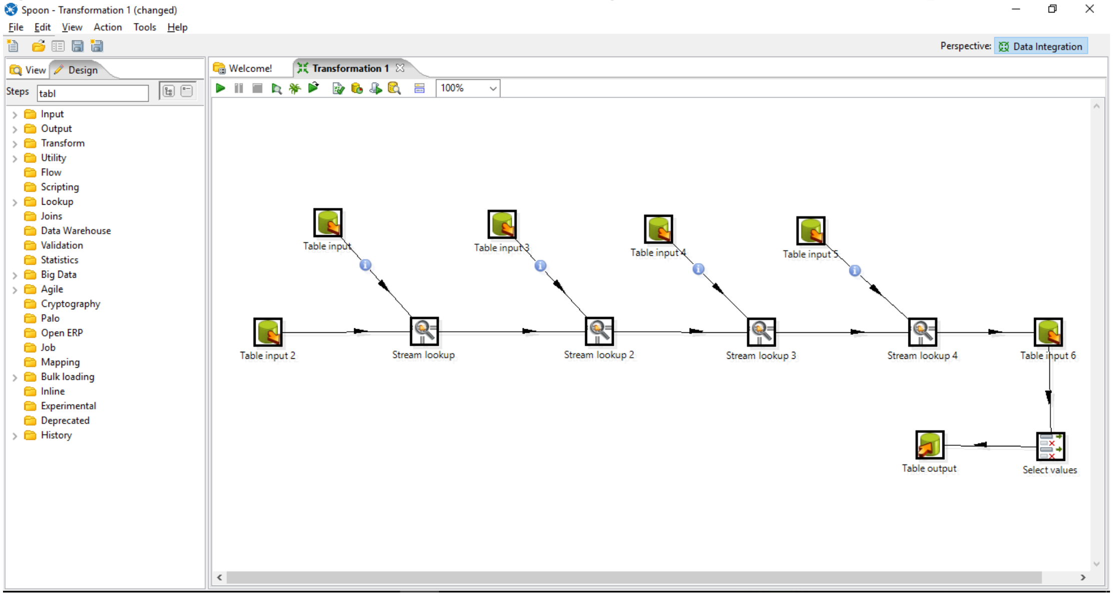

Figure 4. ETL with PDI-KETTLE.

Based on the experiments related to the implementation of ETL using SQL Server Integration (SSIS) 2008 and Pentaho Data Integration (PDI) Kettle, the following section discussed several comparative criteria used to distinguish the difference between ETL process using SSIS and one using PDI.

3.3. Comparison using developer perspective

Implementation of both tools with 1 similar architectural database (Northwind star) requires developers to carry out several steps. Table 1 described the implementation of ETI using SSIS and PDI.

Table 1. ETL Implementation using SSIS and PDI.

SQL Server Integration Service	Pentaho Data Integration
1. Identifying data from data source,	1. Identifying data from data source,
2. Extracting data manually,	2. Developing database as data warehouse,
3. Transforming data manually,	3. Developing dimension table (extract and transform),
4. Loading data manually.	4. Loading data using application.

Table 2. Loading Data with SSIS and PDI.

SQL Server Integration Service	Pentaho Data Integration
1. Selecting “Import Data” menu,	1. Selecting which type of data to add,
2. Selecting data source, a Microsoft excel file,	2. Loading data
3. Selecting destination for data storage,	
4. Loading data.	

3.4. Comparison using user perspective

Comparative analysis using the user perspective may be seen from loading data to data warehouse. Table 2 described process of loading data with SSIS and PDI.

3.5. Comparative analysis based on system development system

Having finished the experiment, one may see the difference both architecture had. PDI simplified the data integration process and was able to integrate data from multiple sources. Data sources with different databases may be integrated using PDI easily. In addition, the nature of the reusable web service made it possible to reuse the pre-made web service for future reuse [2].

SSIS may also integrate data but the process was more complicated as the system development should understand the structure of database as data source and operate or work with SQL Server. SSIS requires complex integration process for data integration from data source from different database.

4. Conclusion

Having compared ETL process using both SSIS and PDI, the conclusions: based on the experiment, data integration using SSIS was more complicated and took longer time compared to doing so using PDI. Using SSIS, data integration process began with pretty complicated ETL (Extract Transform Loading) process, while data integration process using PDI was simpler with web service; SSIS supports data source from different formats, such as excel (.xls/.xlsx), Dbase *File* (.dbf), *Text File* (.txt), MS Access *Database* (.mdf) and others, even though the ETL should be conducted manually. PDI has also supported data source that directly came from database of or manual one wih various data source file for instance Traditional Business Intelligence; PDI is more effective in solving technology integration system developer may encounter in the future.

Acknowledgments

We would like to thank Lembaga Penelitian dan Pengabdian Masyarakat (LPPM), Universitas Negeri Manado for supporting the research.

References

[1] A P Narendra 2016 Big Data, Data Analyst, and Improving the Competence of Librarian *Rec. Libr. J.* **1** 2 83–93

[2] R Somya 2012 *Perbandingan Traditional Business Intelligence dan Service-oriented Business Intelligence (SoBI)* Salatiga

[3] Grecol M L 2012 *Microsoft SSIS and Pentaho Kettle: A Comparative Study for Three-Tier Data Warehouses*

[4] Wu L, Barash G, and Bartolini C 2007 A service-oriented architecture for business intelligence Service-Oriented Computing and Applications, 2007. SOCA'07. IEEE International Conference on 279-285 IEEE

[5] A Azimah and Y G Sucahyo 2007 Penggunaan Data Warehouse dan Data Mining untuk Data Akademik *J. Sist. Inf. MTI UI* **3** 2 1–7

[6] R J Salaki, J Waworuntu, and I R H T Tangkawarow 2016 Extract transformation loading from OLTP to OLAP data using pentaho data integration *IOP Conf. Ser. Mater. Sci. Eng.* **128** 12020

[7] J P A Runtuwene and I R H T Tangkawarow 2017 Vertical information system: A case study of civil servant teachers data in Manado city *Indones. J. Electr. Eng. Comput. Sci.* **6** 1 42–49

[8] Tim Studi Implementasi Business Intelligence 2007 *Laporan tim studi tentang implementasi business intelligence (Departemen Keuangan Republik Indonesia Badan Pengawas Pasar Modal Dan Lembaga Keuangan)*

[9] R W Witjaksono, M Wiyogo, and P N Wicaksono 2007 Perancangan Aplikasi Business Intelligence Pada Sistem Informasi Distribusi Pt Pertamina Lubricant Menggunakan Pentaho *J. Rekayasa Sist. dan Ind.* **2** 2 12–18

[10] A Armadyah 2009 *Perancangan Dan Pembuatan Data Warehouse Pada Perpustakaan STMIK AKAKOM Yogyakarta* Tesis. Perpust. MTI 39–52