APTIKOM Journal on Computer Science and Information Technologies
Vol. 4, No. 2, 2019, pp. 61~68
ISSN: 2528-2417
DOI: 10.11591/APTIKOM.J.CSIT.36

Extract transform load (ETL) process in distributed database academic data warehouse

Ardhian Agung Yulianto
Industrial Engineering Department, Engineering Faculty, Andalas University
Kampus Limau Manis Kota Padang, West Sumatera, Indonesia, 25163
Corresponding author, e-mail: ardhian.ay@eng.unand.ac.id

Abstract

While a data warehouse is designed to support the decision-making function, the most time-consuming part is the Extract Transform Load (ETL) process. Case in Academic Data Warehouse, when data source came from the faculty's distributed database, although having a typical database but become not easier to integrate. This paper presents how to an ETL process in distributed database academic data warehouse. Following Data Flow Thread process in the data staging area, a deep analysis performed for identifying all tables in each data sources, including content profiling. Then the cleaning, confirming, and data delivery steps pour the different data source into the data warehouse (DW). Since DW development using bottom-up Kimball's multidimensional approach, we found the three types of extraction activities from data source table: merge, merge-union, and union. Result for cleaning and conforming step set by creating conform dimension on data source analysis, refinement, and hierarchy structure. The final of the ETL step is loading it into integrating dimension and fact tables by a generation of a surrogate key. Those processes are running gradually from each distributed database data sources until it incorporated. This technical activity in distributed database ETL process generally can be adopted widely in other industries which designer must have advance knowledge to structure and content of data source.

Keywords: *data warehouse; distributed database; ETL; multidimensional*

1. Introduction

A simple definition of Extract, Transform, and Load (ETL) could be "the set of processes for getting data from OLTP (On-Line Transaction Processing) systems into a data warehouse." ETL process as a part of Data Staging Area, as a first stage of receiving data derived from heterogeneous of data sources, assuring data quality, and consistency [1]. In a data warehouse (DW), the ETL process is the most time-consuming part, even mentioned that the number could rise to 80% of the total project development time [2]. As the important component of database-oriented decision support system, DW stores the records about activities and events related to a distinct business process. DW itself, defined as "a system that extracts, clean, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for decision making"[3].

The main components of DW are operational data sources, data staging area, data presentation area, and data access tools [4]. Operational data sources support different business area, different technologies, and formats for staging data. In this research, as a case study is the Academic Information System at Andalas University, Padang, Indonesia. Andalas University consists of 15 faculties, and each faculty has several departments with a total of 120 departments. This information system provides all services in the academic process started from course registration, lecturer distribution for every course, grading system, and graduation.

When initiating this information system, the database administrator designed it as a distributed database system. Based on Distributed Database (DDB) definition said that when a DDB is deployed using only one computer, it remains a DDB because it is still possible to deploy it across multiple computers [5]. The general terminology of DDB is the collection of multiple, separate DBMSs, each running on a separate computer, and utilizing some communication facility to coordinate their activities in providing shared access to the enterprise data [6]. The advantages of this configuration are as a reflection of organizational structure, improved local autonomy, availability, and performance. Although in another side also there are disadvantages such as lack of standards, complexity, and integrity control [7].

Received May 15, 2019; Revised June 10, 2019; Accepted June 30, 2019

By this design, the database server is installed on a single machine and consisted of several separate academic databases for faculty. It means that every faculty has its single typical database with identic structures (table and attribute), but there is no guarantee for the same data of the reference tables. This information system is running well as a transactional system and for preparing the enhanced academic business intelligent, we propose a construction of Academic DW include integration data process.

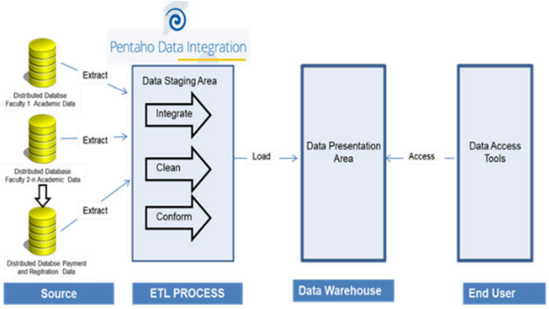

Figure 1. The main components of a proposed academic data warehouse. The ETL activities are represented in data staging area

The Figure 1 is essential for an academic data warehouse to have the truthful and comprehensive integral data. The first objective of this study is the implementation of bottom-up dimensional data warehouse design approach. This system initiated by the completeness of the business requirements provided by the university executives, then through the dimensional and fact tables with measurement data as a representative of multidimensional technique. The other objective in this research is to identify related attributes from each data source then identify them as the master tables or operational tables. In another hand, several activities performed in the ETL process such as extracting, cleaning, conforming tables from and loading them into DW. This process carried on data quality and the data analysis from different faculty's academic database data sources.

2. Research Method

On distributed database data sources, the ETL process becomes more complicated because there is no single control mechanism over the contents of the data of each database. The gradual and careful process to identify which table should be merged or not, to profile the table's attribute and analyze data in tables are implemented in detail in this work. Knowing the data structure will help in cleaning and conforming dimension and making fact table. In the case of an academic data warehouse, all steps arrange based on the business process; grain declared, dimension and data measurement in fact tables as the steps for the building data warehouse.

In the building of the data warehouse, we adapted Kimball's bottom-up approach in developing a multidimensional model [8]. Based on Kimball's four-step dimensional design process, the model design activities are built by the following steps:

a. Select the business processes: to be modeled, taking into account business requirements and available data sources.
b. Declare the grain: defining an individual fact table representation which related to business process measurements.
c. Choose the dimensions: determining the sets of attributes describing fact table measurements. Typical dimensions are time and date.
d. Identify the numeric facts: define what measure to include in fact tables. Facts must be consistent with the declared grain.

The important advantage of this approach is consists of data marts as a representation of the business unit, which have a very quick initial set up and effortless to integrate between another data marts. For consideration to enterprise scale DW, this method tranquil in extend from existing DW and also provide reporting capability.

After that, in the ETL process, there are prominent four staging steps while data is staged (written to the disk) in parallel with the data being transferred to the next stage, as describe in Figure 2.

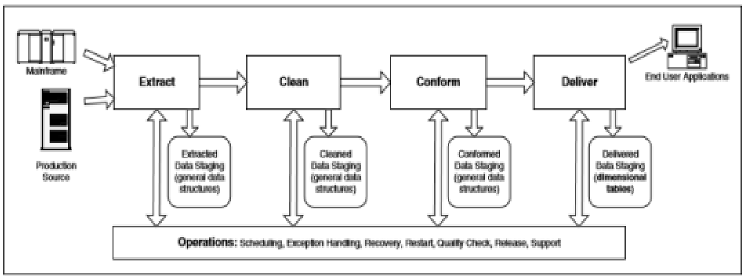
Mainframe
Production Source
Extract
Clean
Conform
Deliver
End User Applications
Extracted Data Staging (general data structures)
Cleaned Data Staging (general data structures)
Conformed Data Staging (general data structures)
Delivered Data Staging (dimensional tables)
Operations: Scheduling, Exception Handling, Recovery, Restart, Quality Check, Release, Support

Figure 2. Four steps data flow thread in data staging area as etl activities [3]

Research about ETL performance already carried out by researchers. Igor Mekterrovic et al. [9] proposed a robust ETL procedure with three agents: changer, cleaner, and loader. This research focuses to handle horizontal data segmentation and versioning when the same fact tables used by various users and must not influence the other. Vishal Gour et al. [10] proposed another improvement ETL performance by query cache technique to reduce the response time of ETL. E. Tute et al. [11] describe an approach for modeling of ETL-processes to serve as a kind of interface between regular IT-management, DW, and users in the case in the clinical data warehouse. Abid Ahmad et al. [12] notify the use of the distributed database in development of DW. This study explains more on architecture and change detection component. Sonali Vyas et al. [13] define the various ETL process and testing techniques to facilitate for selection of the best ETL techniques and testing.

Examining the mentioned study, it is clear the research gap in this ETL process is advance analyzing technique of distributed database data source. Also, the specific case in the academic area was applied with the strategy in each ETL steps after analyzing the academic data.

3. Results and Discussion

3.1. Designing Multidimensional Model

Follow Kimball's four-step dimensional design for building an academic data warehouse shows in Table 1.

Table 1. Four steps data modeling process for academic data warehouse

Step	Output
1. Identify the business process	Business Process • Analysis of New Student • Analysis of New Student Registration By Semester Period • Analysis of Student Grade By Semester • Analysis of Graduation Student • Analysis of Student Payment
2. Declare the grain	Measurement and their associated granularity • New student information (every year) • Student registration information (every semester) • Grade Distribution (every semester) • Graduation information (every graduation period) • Student Payment information (every month, semester, year)
3. Identify dimension	Dimension and their associated role. _Tahun_ (Year), _JenisKelamin_ (Gender), Semester, _Agama_ (Religion), _Mahasiswa_* (Student), _MahasiswaRegistrasi_ (Student Registration), _Program Studi_ *(Uni), _Program Studi DIKTI_* (National), _Model_*, _Jenjang_ *(degree), _Fakultas_ (Faculty), _PendidikanOrtu_ (Parent's Education Background), _PenghasilanOrtu_ (Parent's Monthly Wage), _Mata Kuliah_ (Course Name), _Dosen_ (Lecturer), Smta (Senior High School)*, _Pembayaran_ (Payment), _Nama Bank_ (Name of Bank).
4. Identify the facts	Facts and their associated fact tables • New student fact table| factless | student_count • Student Registration fact table | factless | student_count • Student grade fact table | snapshot – current GPA | GPA_average • Graduation fact table | event | student_count, GPA_biggest • Student Payment fact table | Event | student_payment_sum, status

[a] The Entity that needs to merge into a single dimension

In integration and cross-process work, making conformed dimensions help to identify and map every individual fact to dimension [14]. If we got the right dimension, we could focus on it in each single database data sources, data shows in table 2.

Table 2. Bus dimension conformance matrix.

Aspect of Analysis \ Dimension	Student				
	New Student	New Semester Registration	Grade/Academic Process	Graduation	Payment Transaction
Time					
Time					X
Tahun (Year)	X			X	X
Semester / Period		X	X		X
	X	X	X	X	X
		X	X		
Prodi			X	X	X
Program Studi (Univ)	X	X	X	X	
rogram Studi DIKTI (National)	X	X	X	X	
Model	X	X	X	X	X
Jenjang (Degree)	X	X	X	X	
	X	X	X	X	
	X				
School					
Nama Sekolah (School of Origin)	X				
Kota (City)	X				
Provinsi (Province)	X				
Negara (Country)	X				
Type Sekolah (School Type)	X			X	
	X	X			
	X				
	X				
	X				
			X	X	
			X		
					X
					X

3.2. ETL Process

ETL Tool using in this research is Pentaho Data Integration (PDI) v6.1 as a part of Pentaho, a open source commercial business intelligence tools. The most common use of PDI is to perform ETL and also powerful for obtaining and manipulating data [15]. In Pentaho Business Analytics Server develop several plugins such as Saiku Analytics that provides interactive analysis with tables and graphs.

3.2.1. Extraction

In the case of distributed database, extraction step initiated by data source identification, including investigate the table structure and description in each faculty database. We divided the table into three types of tables:

a. Master table with merge strategy;

b. Master table with merge-union strategy;

c. Transaction table with union strategy.

Merge strategy means that the same table in different databases needs to merge into the single table with identical data. Merge action usually happened at the table which represent the global policy from university such as entrance type, model of program study, grade of education. Merge-union strategy means that some tables will carry with two consecutive actions: union and merge. These table in the different database needs to merge first and then to be union. In the consequence of distributed database is high possibility for inserting individual data in each database which not influence to the other. So, for integration reason, value of these table need to merge for the same code, then after that apply union action to unify different value. For example for the staff table involved in this strategy because a staff although belong to one faculty, but she/he can teach in different faculty. Her/his data stored in more than one database, so in target integration table, the same value must be merge then the unique value can be unified. The last table strategy is a union strategy. It means a data consolidation into a larger amount of row in target table from the data sources that have an exclusive value. Identification table from five database data source as a pilot project is described in Table 3.

Second: integration of heterogeneous data source. Each single database must be connected and then can be merged/union based on the identification and analyzing of the source. This point also related to connectivity to access the database. To create a connection, we will need to know the connection settings. At least as the following [16]: Type of database connection, hostname (Domain name or IP address of the database server); Database name (The schema or other database identifiers); Port Username and password to access the data source**.** Figure 3 shows integration transformation process in PDI.

Third: data content analysis. In this analysis, we focused on NULL Values. Dealing with NULL is another solution in this step. We define two serial solutions; the first is by using IFNULL() expression in SQL script both in creating dimension table or fact tables. The second solution is adding a new column in the master table to handle NULL data or Not Available in a transaction table. This solution need database administration intervention and the privilege in decide it.

Table 3. Identification of data sources

No	Table Name	Update Strategy	Load Frequency	Initial Row Count*					Description
				DB1	DB2	DB3	DB4	DB5	
1	jenjang_akademik_ref	Merge	On demand	10	10	10	10	10	Degree of Study Program in University
2	model_ref	Merge	On demand	3	3	3	3	3	Model of Study Program
3	s_jalur_ref	Merge	On demand	51	51	51	51	51	Entrance University Type
4	agama_ref	Merge	On demand	7	7	7	7	7	Religion
5	s_pendidikan_ref	Merge	On demand	9	9	9	9	9	Level of Parents's Education Background
6	s_penghasilan_ref	Merge	On demand	7	7	7	7	7	Level of Parents's Wage
7	Kota	Merge	On demand	780	780	780	780	780	City
8	Propinsi	Merge	On demand	80	80	80	80	80	Province
9	Negara	Merge	On demand	233	233	233	233	233	Country
10	s_nilai_matakuliah_ref	Merge	On demand	11	11	11	11	11	Grade of Point
11	Dosen	Union – merge	On demand	270	214	319	454	408	Name of staff/lecturer
12	program_studi_dikti	Union – merge	On demand	492	490	491	491	497	National Study Program
13	s_smta	Union - merge	On demand	12207	12208	12206	12207	12207	Name of Senior High School
14	Fakultas	Union	On demand	1	1	1	1	1	Name of Faculty (active faculty in each DB)
15	program_studi	Union	On demand	21	7	20	26	15	University Study Program
16	program_studi_akreditasi	Union	On demand	21	7	20	26	15	University Study Program's Level Accreditation
17	s_semester_prodi	Union	On demand	391	112	434	132	296	Semester and University Study Program
18	Mahasiswa	Union	yearly	3739	4861	4463	6480	5597	Student in Faculty
19	mahasiswa_orangtua	Union	yearly	3722	4279	4446	6406	5587	Student's Parents Data
20	mahasiswa_registrasi	Union	semester	19892	24906	22657	29568	30682	Student Registration in Semester Period
21	s_kurikulum	Union	On demand	29	21	45	23	28	Name of Curriculum
22	s_mata_kuliah_kurikulum	Union	On demand	1488	1115	2011	682	1854	Name of course
23	s_krs	Union	semester	22891	27197	26243	27466	33715	Status Enrollment in Semester Period
24	s_krs_detil	Union	semester	133477	196063	168589	145497	229478	Subject and Grade
25	s_v_ip	Union	semester	32470	52057	46527	28700	54709	GPA per Semester Periode

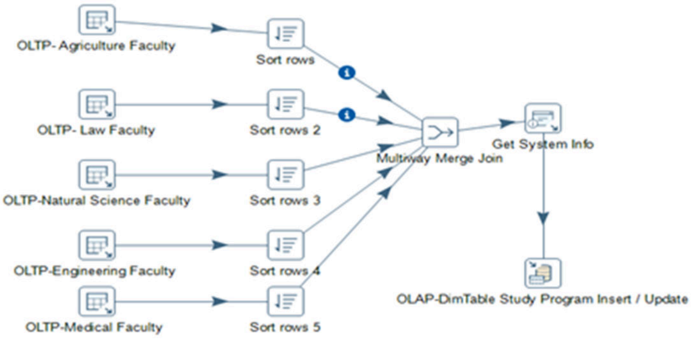

Figure 3. Pentaho data integration with transformation

3.2.2. Cleaning and Conforming

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data to improve the quality of data. The data cleansing approaches involve three steps include: data analysis, data refinement, and data verification [17]. There are single source problems and multisource problems. Table 4 describes one table (s_semester_nama) derived from each database with different and inconsistent value.

Table 4. Multisource problem in table s_semester_nama

No	Faculty Name	Semester Name Code					
		1	2	3	4	5	6
1	Agriculture	Ganjil	Genap	SP			
2	Law	Ganjil	Genap			Pendek	KKN
3	Medical	Ganjil	Genap	SP			
4	Natural Sc	Ganjil	Genap	Pendek	KKN		
5	Engineering	Ganjil	Genap	Pendek	KKN		

General data and value rule reasonability [3]. This condition, after doing re-analyze, that ETL Team can confirm, join, and integrate as a simplification.

Create conform dimension. Conforming activities here refers to 2 conditions: hierarchy and conform dimension. A hierarchy within a dimension, then, is one of many potential drill paths. Dimensions are not limited to a single hierarchy. It is common to find several hierarchies within a dimension [14]. A conformed dimension means the same thing with every possible fact table to which it can be joined [3] as describe in Figure 4.

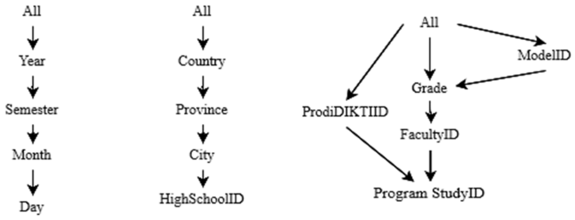

Figure 4. (a) hierarchy of date, (b) high school location, and (c) program study

3.2.3. Delivery/Loading

Surrogate generated [18]. The surrogate key as a nothing integers key which does not have any business meaning and only uses for data analysis. So, we generate this surrogate key in every dimensional table and fact table. This technique, besides the best practice in the data warehouse also is needed to implement slowly changing dimensions.

Star schema dimension tables. These activities include the denormalizing data as a star schema table characterized as being “not normalized.” And then deliver the data to the target dimension table.

Slowly Changing Dimension (SCD). This a step for accommodating the history by add 2 (two) new columns in OLTP master table (created date time & update date time) and column valued by the last update date time in DW. All of the current and historical data over time will be stored in this field.

Loading Fact Table. One of the most important aspects of preparing fact data for loading is performing the dimension lookups with taking correct surrogate key [1].

Working with Online Analytical Processing (OLAP) Data. Considered as integrate database DW target, this activity related in working with multidimensional data (OLAP Cube). Creating a cube in data presentation area using the Pentaho Schema Workbench tools.

3.3. ETL Testing

The main objective of ETL testing is to verify and relieve data defects and general errors that occur earlier before processing it for analytical and reporting. Several tests can be conducted based on their function. Here we performed these scenarios: data completeness validation, meta data testing, and incremental ETL testing [19]. As a sample we picked one target table named dim_program_studi_dikti in OLAP DW database.

The goal of data completeness validation is to measure that all data is loaded as expected. Meta data testing is to verify that the table definition conform to the model and application design. Mapping document include data type check, data length, index/constraint check between source and target table reveals this validation. One of incremental ETL testing is a duplication validation for measuring unwanted duplication existence. Table 5 and Table 6 illustrate the ETL testing.

Table 5. Data completeness and incremental ETL testing

Test Case name	Query	Expected result	Actual Result
Check number of records present in source table	*Select count(*) from sia_teknik.program_studi_dikti*	497	497, Passed.
Check number of records present in target table after data is loaded	*Select count(*) from sia_dwh.dim program_studi_dikti*	497	497, Passed.
Check the records present in source table which are not in target table	*Select prodidiktiKode from sia_teknik.program_studi_dikti* *Minus* *Select prodidiktiKode from sia_dwh.dim program_studi_dikti*	None	No records found, Passed.
Check if any duplicate records in target tables	*Select * from sia_dwh.dim program_studi_dikti where prodidiktiKode in (select sk from sia_dwh.dim program_studi_dikti group by prodidiktiKode Having Count(*) >1*	None	No records found, Passed

Table 6. Meta data testing by mapping document.

Source table	Source Column	Data Type	Target Table	Target Column	Data Type	Not Null	Unique	Transform a-tion
			Dim_program_studi _dikti	sk_prodi _dikti	Int	Yes	Yes	Create surrogate key
Program_ studi_dikti	Prodidikti _Kode	Varchar (10)	Dim_program_studi _dikti	Prodidikti_Ko de	Int	Yes	Yes	
Program_ studi_dikti	Prodidikti _Nama	Varchar (255)	Dim_program_studi _dikti	Prodidikti_Na ma		Yes	No	
			Dim_program_studi _dikti	Last_update	datetime	Yes	No	Create last update for historical

3.4. Discussion

ETL process as a part of main component DW development should give more attention for DW manager. To guide DW development congregate the business requirements of institution, the Kimball’s bottom up approach put into process. The single functional area is set off instead of enterprise scale. For distributed data source, the applied of data staging provide benefits before it loaded to OLAP DB. All of ETL steps covered in data staging area to ensure data transformation from source to target database running on the track. This option also can realize widely in another circumstance which concerned in data quality issues.

Data source advance knowledge by its structure and function must have in this instance. It takes many technical parts. Fortunately, the Pentaho Data Integration as open source BI tool having capability to deal with it. Testing conducted to verify the accuracy of data loading against the signed off business requirement and rule. In general DW testing view, the extensibility aspect is confirmed by designing of data mart to take up again in next data marts. The low time elapsed for DW process will increase the DW performance compare the transactional. The design of multidimensional provides at easy in making drill down and drill across of data in reporting. For analysis of new students, by multidimensional view we obtained data. Dimension data from gender, program study, faculty, year, the origin of senior high school data are the examples.

4. Conclusion

The heterogeneous of data source can be overcome by designing the ETL Process as following ETL Data Flow: Extract, Clean, Conform, and Delivery/Load. Extract step perform data source analyze and content profiling so that proven where table should be master table or transaction table. In distributed database system with the goal for integration data, three activities have been accomplished in each database: merge, merge-union, and union. Result for cleaning and conforming step set by creating conform dimension on data source analysis, refinement and hierarchy structure. The final of ETL step is loading it into integrate dimension and fact tables by generation of surrogate key. In case of distributed database data source, ETL designer must have advance knowledge about data structure and content in each database and then easy to apply ETL process completely. Several ETL testing is held as proven in the correctness of data transformation from data source to target. This procedure generally can be employed widely for implementation of DW in another industry. Pentaho Data Integration (PDI) as open source ETL tools can process all of the steps.

References

[1] Matt Casters, R. Bouman, J.V. Dongen. Pentaho Kettle Solutions: Building Open Source ETL Solution with Pentaho Data Integration. Indianapolis: Wiley Publishing, Inc. 2010:8,234.

[2] W.H.Inmon. Building the Data Warehouse-Third Edition. New York: John Wiley & Sons Inc. 2002:283.

[3] Kimball, Ralph., Caserta, Joe. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data. Indianapolis: Wiley Publishing. Inc. 2004:23,18.

[4] Kimball, Ralph., Ross, Margy. The Data Warehouse Toolkit: The Completed Guide to Dimensional Modeling. New York: John Wiley & Sons Inc. 2002:7.

[5] Saeed K Rahimi, F.S Haug. Distributed Database Management System–A Practical Approach. New Jersey: John Wiley & Sons Inc. 2010:1.

[6] M.T Ozsu, P. Valduriez. Principles of Distributed Database–Third Edition. New York: Springer. 2011:3.

[7] T.Connolly, C. Begg. Database System. A Practical Approach to Design, Implementation and Management. Fourth Edition. Essex: Pearson Education. 2005: 695.

[8] Kimball, Ralph., Ross, Margy. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition. Indianapolis: John Wiley & Sons Inc. 2013:38.

[9] Igor Mekterovic, Ljiljana Brkic, and Mirta Baranovic. *Improving the ETL process of higher education information system data warehouse*. Proceedings of the 9th WSEAS International Conference on Applied Informatics and Communications (AIC'09). Moscow.2009: 265-270.

[10] Vishal Gour, S.S. Sarangdevot, G.S. Tanwar, A. Sharma. Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse. *Int. Journal on Comp. Sci. and Eng.* 2010; 2(3):786-789

[11] Abid Ahmad, Muhammad Zubair. *Using Distributed Database Technology to simplify the ETL Component of Data Warehouse*. Proceedings of WSEAS International Conference on Applied Computer Science (ACS'10). Iwate. 2010; 61-65.

[12] Tute E, Steiner J. Modeling of ETL-Processes and Processed Information in Clinical Data Warehousing. *Stud Health Technol Inform*. 2018; 248 204-211. PMID: 29726438.

[13] Sonali Vyas & Pragya Vaishnav. A comparative study of various ETL process and their testing techniques in data warehouse, *Journal of Statistics and Management Systems*. 2017; 20(4): 753-763.

[14] C. Adamson. Mastering Data Warehouse Aggregates. Solutions for Star Schema Performance. Indianapolis: Wiley Publishing Inc. 2006:20.

[15] W.D Back, N. Goodman,J Hyde. Mondrian in Action. Open Source Business Analytics. New York: Manning Publications Co. 2014:195.

[16] A. Meadows, A.S. Pulvirenti, M.C. Roldan. Pentaho Data Integration Cookbook. Birmingham: Packt Publishing, 2013:11.

[17] Rahm, E., H. H. Do, Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull*. 2000; 23(4): 313.

[18] R. Bouman, J.V. Dongen. Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL. Indianapolis: Wiley Publishing, Inc. 2009:160.

[19] https://www.datagaps.com/concepts/etl-testing, accessed at May 12, 2019.