Journal of Advances in Science and Technology

Vol. III, No. VI, August-2012, ISSN 2230-9659

HOW DOES EXTRACT TRANSFORM LOAD HELP IN DATA WAREHOUSING

www.ignited.in

Journal of Advances in Science and Technology
Vol. III, No. VI, August-2012, ISSN 2230-9659

How Does Extract Transform Load help in Data Warehousing

S. K. Humayun

Research Scholar (Computer Science), CMJ University, Shillong, Meghalaya

Abstract: ***The data warehouse is the centralized repository which stores the data from multiple information sources and then transforms it into the common and multidimensional data model for the efficient querying and analysis. ETL is one of the tools that are used by data warehouse and ETL helps data to perform its operations flexible and consistent and also more efficient. ETL (Extraction, transformation, and loading) takes place in the data warehouse, when the data warehouse is populated for first time and also it takes place regularly at the time of updating the data warehouse. ETL tools provide more value to the data warehousing.***

Index Terms— Data warehouse (DW), ETL (Extract, Transform, Load)

----------♦----------

1. INTRODUCTION TO DATA WAREHOUSING

Data warehouse is the database that is used for data analysis and reporting. Data warehouse is the central repository of data that is created by integrating the data from many disparate sources (Hammergren, 1996). Data warehouse is used to store historical data as well as the current data and it is mainly used for creating the trending reports for the senior management reporting such as quarterly and annual comparisons (Imhoff, 2003). Data warehouse is the integrated and subject-oriented, non-volatile and time variant collection of the data that used in the strategic decision making (Inmon and Bill, 1992).

Data warehouse definition focuses on the data storage. In data warehouse, the main source of data is cleaned, cataloged, transformed and made available for use by business professionals and other managers for market research, decision making support, online analytical processing and data mining (Marakas & O'Brien 2009). Data warehouse is means to retrieve and analyze data, to ETL (extract, transform and load) data and also to manage the dictionary (Kimball, Ralph, and Margy, 2002). These are all considered as the most essential components of the data warehousing system.

Data warehouse is the collection of data which supports the decision making processes (Inmon, 2005). Data warehouse provides various important features as shown in the following table.

Feature	Data Warehouses
Users	Hundreds
Workload	Specific analysis queries
Access	To millions of records, mainly read only mode
Goal	Decision-making support
Data	Summed up, mainly numeric
Data integration	Subject-based
Quality	In terms of consistency
Time coverage	Current and historical data
Updates	Periodical
Model	De-normalized, multidimensional
Optimization	For OLAP access to most of the database

Table: Features of Data Warehouses

Source: Kelly and Sean (1997): Data warehousing in action, Wiley, New York

The multipurpose nature of the data warehouse:

- Data warehouse is enterprise focused.
- Data warehouse must be designed to have the capacity to load enormous amounts of data in very in short time
- Data warehouse design must be as flexible as to change as possible.
- Data warehouse's data must be in the format which supports all the BI (business intelligence) analyses in all technologies.
- Data warehouse must be designed for the optimal data extraction processing by delivery programs.

2. INTRODUCTION TO ETL (EXTRACT LOAD TRANSFORM)

ETL processes helps to extract the data and integrate it and also clean the data from the operational sources and then feed it to the data warehouse layer (Ding et al, 2004). The process of ETL takes place, when the data warehouse is populated for first time and then it takes place every time when the data warehouse is regularly updated. ETL (Extraction, transformation, and loading) mainly consists of 4 separate phases and they are: extraction (or capture), transformation, cleansing (or scrubbing or cleaning), and loading.

(A) EXTRACTION:

Relevant data can be obtained from the sources in extraction phase. Static extraction can be used when the data warehouse needs populating for first time. Incremental extraction can be used to update the data warehouses regularly, and to make the changes that applied to source data of the latest extraction. According to English (1999), the data that wants to be extracted is mostly selected on the basis of the data quality.

(B) CLEANSING:

The cleansing phase is most important in data warehouse system. The cleansing phase helps to improve the data quality and it is normally quite poor in the sources (White, 2005). The most frequent inconsistencies and mistakes that make the data to be dirty are: duplicate data; missing data; inconsistent values which are logically associated; wrong or impossible values; unexpected use of fields; inconsistent values for the individual entity due to typing mistakes; and inconsistent values for single entity due to different practices. These mistakes can be cleaned up in the cleansing phase.

The following figure illustrates the different phases of ETL.

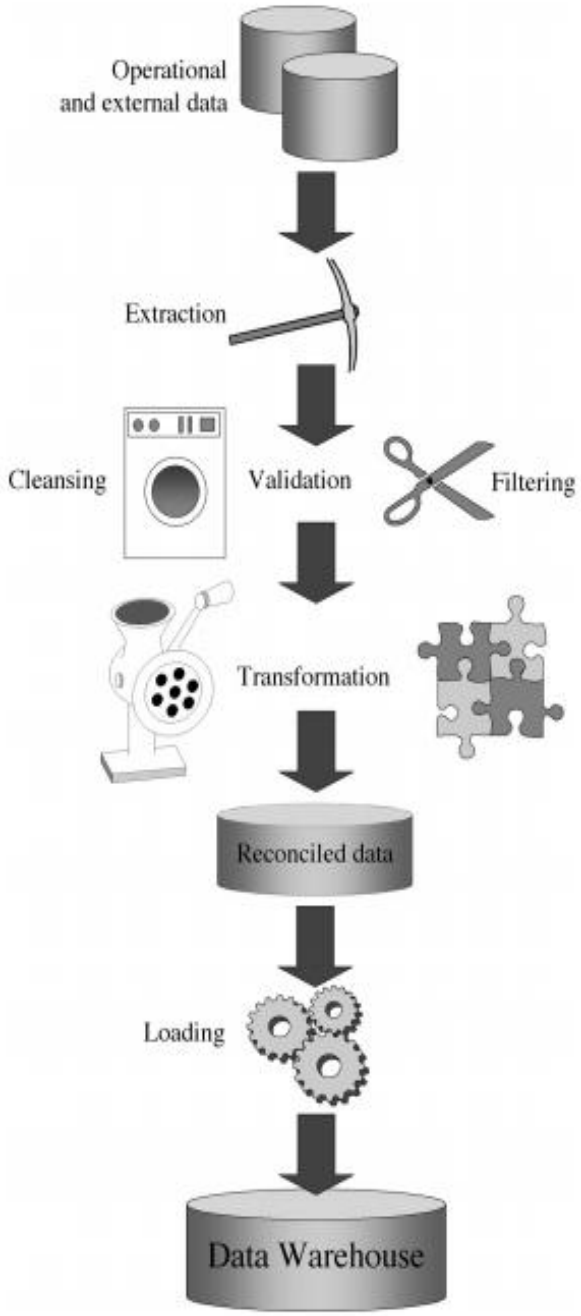
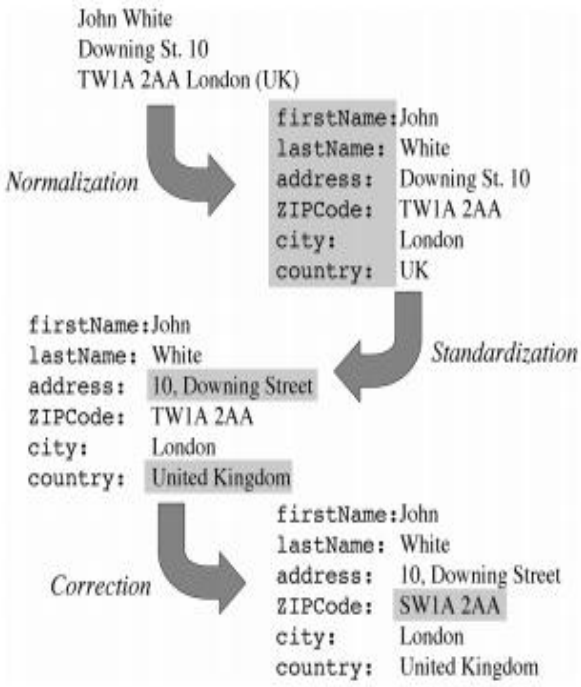


Figure: ETL (Extraction, Transformation and Loading)

Source: English, L.P. (1999): Improving data warehouse and business information quality, Wiley, New Jersey

(C) TRANSFORMATION:

Transformation is the core process of the reconciliation phase. Transformation converts data from operational source format into the particular data warehouse format. Transformation and cleansing processes are closely connected with the ETL tools. The following is the example for cleansing and transforming the some data.



The above example of customer data shows that: the field based structure is extracted from some loose text and then few values are standardized so it removes some abbreviations and those values are logically associated and then it can be rectified when needed.

(D) LOADING:

Loading is the last step that should be taken in the data warehouse (Weir et al, 2003) . Loading can be done in two ways:

Refresh: Older data is completely replaced that means the data in the data warehouse is completely rewritten. Refresh is used in the combination with static extraction and this is to initially populate the data warehouse.

Update: Some changes can be applied to source data and then it is added to the data warehouse. Update can be done without modifying or deleting the preexisting data. The update technique will be used in the combination with incremental extraction and this is to update the data warehouses regularly.

(E) ETL PROCESS

ETL (Extract, transform, and load) has three stage processes (Ballou and Pazer, 2003). The ETL tools were created in order to improve and facilitate the data warehousing. The following figure illustrates the process of ETL (Extract, transform, and load).

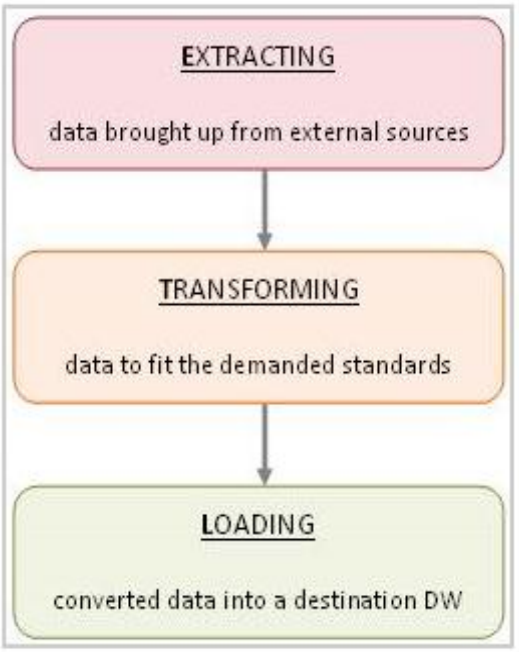


Figure: ETL Process

Source: ETL Tools (2012): ETL, retrieved on 21st December 2012 from http://www.etltools.net/

The ETL (Extract, transform, and load) process may consist of the steps such as: initiation; audit reports; build reference data; validate; extract from sources; load into stages tables; archive; transform; publish; and clean up. The purpose of using the extract, transform, and load tools is mainly to save the time and also to make the entire process more reliable.

(F) ETL TOOL FUNCTIONALITIES

The selection of hardware platform and database is must. In addition to these, the selection of ETL (Extract, transform, and load) tool is highly recommended (Shankaranarayan, 2003). ETL tools have some characteristics which makes the data warehouse to be more flexible and stable and also consistent.

Functional capability: ETL tool include both the 'cleansing' piece and 'transformation' piece. Generally, the typical ETL tools will be either geared towards having the strong cleansing capabilities or having the strong transformation capabilities, but they will be seldom strong in both.

Ability to read directly from your data source: ETL tools makes organization to read the data directly from the data source. For each and every organization or concern, there is some different set of

the data sources. It is essential to make sure that the ETL tools should connect directly to the source data.

Metadata support: The ETL (Extract, transform, and load) tool plays a main role in the metadata and this is because it helps to map the source data to destination, which is an important process of metadata. It is very essential to select the ETL tool which works with the overall metadata strategy.

ETL tools are provided by providers in both commercial ETL tools and open-sources (freeware) ETL tools.

Some of the commercial ETL tools are:

- ODI (Oracle Data Integrator)
- IBM Infosphere DataStage
- SSIS (Microsoft SQL Server Integration Services)
- OWB (Oracle Warehouse Builder)
- BODI (Business Objects Data Integrator) Some of the Freeware, open source ETL tools are:
- CloverETL
- Talend Integrator Suite
- Jasper ETL

CONCLUSION

It can be understood that the main purpose of data warehouse is to store the data. This study concludes that, data warehouse helps to support strategic modeling, planning, and forecasting at business organization level. Data warehouse helps to fulfill the need by providing the knowledge of uncertainty and growth for the business organization. It is concluded that, data warehousing strategies are reflecting the changes in both external and internal business environment. ETL tools helps data warehouse to be effective in various process. Data warehouses are playing an integral role in development, growth and success of the business organization.

REFERENCES

1. Marakas G & O'Brien J (2009): Introduction to Information Systems, McGraw-Hill Companies.
2. Inmon, Bill (1992). Building the Data Warehouse. Wiley
3. Imhoff (2003): Mastering Data Warehouse Design : Relational and Dimensional Techniques, Wiley.
4. Inmon W (2005): Building the Data Warehouse, 2005, John Wiley and Sons, New York.
5. English, L.P. (1999): Improving data warehouse and business information quality, Wiley, New Jersey
6. Kimball, Ralph, Margy (2002): The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, John Wiley and Sons, Inc., Chichester, 2002
7. White C. (2005): Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise, TDWI Whitepaper. URL: http://www.tdwi.org/research/display.aspx?ID=7908. Published Nov, 2005.
8. Hammergren, T.C. (1996): Data Warehousing; Building the Corporate Knowledge Base, Thomson Learning, London.
9. Ballou D P and Pazer H (2003), Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts, IEEE Transactions on Knowledge and Data Engineering , vol. 15, no. 1.
10. Shankaranarayan G, Ziad M and Wang R Y (2003), Managing data quality in dynamic decision environments: an information product approach. Journal of Data Management, vol 14, no. 4, pp. 14-32.
11. Weir R, Peng T and Jon K (2003), Best Practice for Implementing a Data warehouse: A Review for Strategic Alignment, DMDW
12. Ding L, Finin T, Joshi A, Pan R, Cost R S, Peng Y, Reddivari P, Doshi V and Sachs J (2004), Swoogle: a search and metadata engine for the semantic web, In Proceedings of the thirteenth ACM international conference on Information and knowledge management, p 652-659.