ETL METHODOLOGIES, LIMITATIONS AND FRAMEWORK FOR THE SELECTION AND DEVELOPMENT OF AN ETL TOOL

Mr. Aman partap Singh Pall[1],

Assistant Professor,
School of IT, Apeejay Institute of Management Technical Campus,
Jalandhar

Dr. Jaiteg Singh[2]

Associate Professor,
Chitkara Institute of Engg. & Technology,
Rajpura

Abstract

Extraction-Transformation-Loading (ETL) processes are responsible for all the operations taking place at the warehouse. The ETL is done by specialized tools that deal with the extraction of the data from various database sources and then cleaning it and transforming it to be stored in the data warehouse. These sophisticated tools though available in the market are not developed by any standardized parameters and every organization has specific data and with this the problems associated with the ETL tools are varying according to the organization. The main objective of this paper presents an insight into the different methodologies being used for data integration along with the limitations of the ETL tools. A suggestive framework is also being presented for better efficiency of the data integration tools with the business intelligence.

Keywords: ETL tools, ETL process, data warehouse, Business Intelligence

Introduction

Extraction-Transformation-Loading (ETL) tools are specialized tools that dealwith data warehouse homogeneity, cleaning and loading problems. ETL (Data Integration) and Data Cleaning tools are estimated to cost at least one third of the effort and expenses in the budget of the data warehouse and this may further increase to 80% of the development time in a data warehouse project[1][2].

ETL processes are responsible for the operations taking place at the back stage of data warehouse architecture. In the first phase, the data is extracted from multiple sources called as source data stores. These can be OLTP (Online Transaction Processing, Legacy systems, data from websites, spreadsheet documents, simple text files, images and even video streaming.The second phase sees the extracted data propagated to a special purpose area in the warehouse called as Data Staging Area (DSA). It is here that the transformation, homogenization and the cleansing of the data takes place. The propagated data is strictly checked for the business rules and integrity constraints as well as the schema transformations to ensure the data fit the target data warehouse. The transformations include filtering and other checks. The final step is to load this transformed data into the central data warehouse (DW) with all its counterparts i.e. data marts and data views. Figure-1 provides a diagrammatic view of ETL process. The traditional data warehouse setting required a refreshing of the data by the ETL process during idle or low-loads and needed the operations to be completed in a fixed time-window.

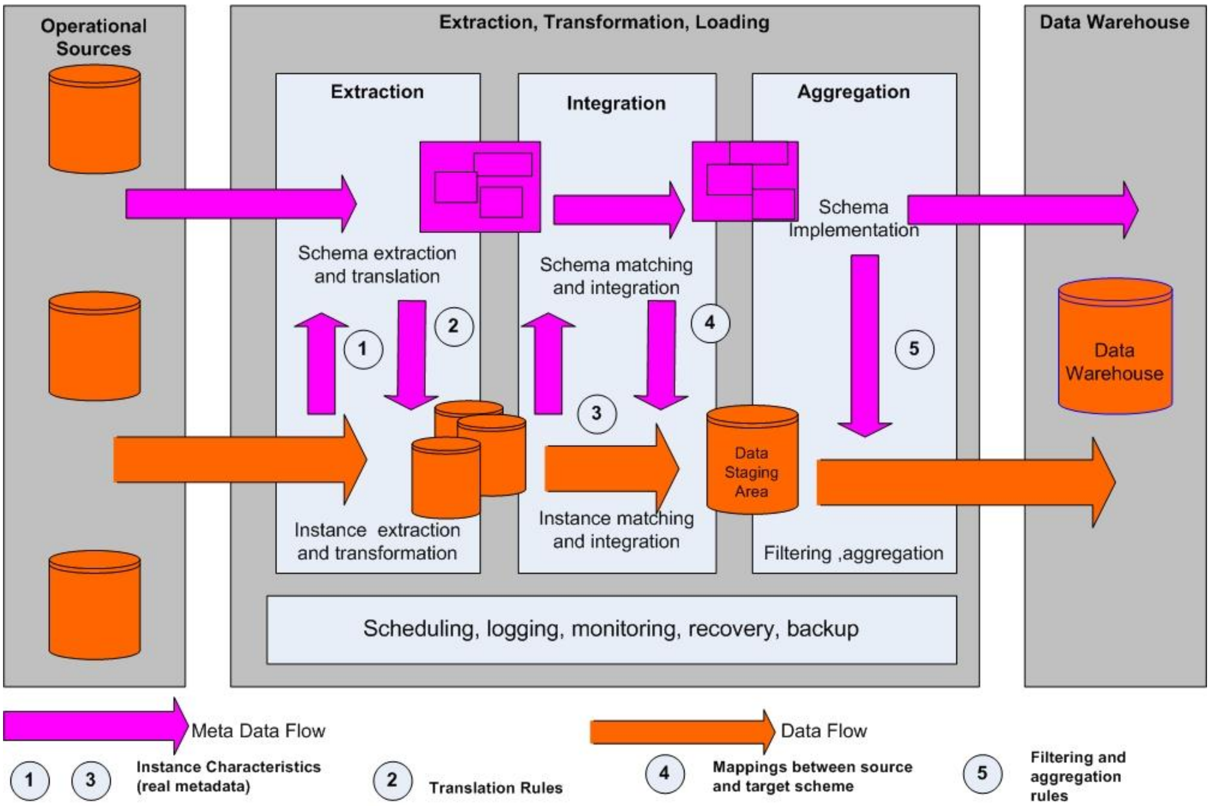
Operational Sources
Extraction, Transformation, Loading
Data Warehouse
Extraction
Integration
Aggregation
Schema extraction and translation
Schema matching and integration
Schema Implementation
1
2
3
4
5
Data Warehouse
Data Staging Area
Instance extraction and transformation
Instance matching and integration
Filtering ,aggregation
Scheduling, logging, monitoring, recovery, backup
Meta Data Flow
Data Flow
1 3 Instance Characteristics (real metadata)
2 Translation Rules
4 Mappings between source and target scheme
5 Filtering and aggregation rules

Figure -1: The ETL process that makes up the data warehouse.

However, nowadays business demands real-time data warehouse refreshing and much more attention is towards the enduring technology advancements [3].

Many organizations prefer in-house development of the ETL tools firstly because the costs of these tools is 55% of the total costs of the data warehouse runtime and secondly because of the complexity of the learning curve of these tools [4].

The data warehouse expenses are expected to cross $14 Billion whereas the projected sales of the ETL tools may not rise more than $300 million thus, it is apparent that the designing, development and deployment of the ad-hoc processes needs re-modeling, re-designing. The most important phase in this remodeling and redesigning is the design flow of data from the source relations towards the target data warehouse relations which is provided by the ETL tools Thus, the Extraction-Transformation-Loading (ETL) tools are pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse.

Data warehousing has evolved during this decade and organizations have realized the need for a process that loads and maintains the data in the data warehouse. Initially, the organizations developed their own custom codes to perform the ETL activity (and are still developing) which is referred to as Hand-Coded ETL process, but soon realized it to be quite lengthy and hard to maintain. Soon companies started developing the ETL tools for an efficient performance.

In the next sections of the paper the Literature Review looks into the different methodologies followed for the ETL. We discuss the limitations of these methodologies in the next section and finally to conclude we provide our implications in the Conclusion and Implication section.

Objectives of the Study

The study aims to fulfill some objectives which are as follows:

1) The main objective of this paper presents an insight into the different methodologies being used for data integration

2) Limitations of the ETL tools.

3) A suggestive framework is also being presented for better efficiency of the data integration tools with the business intelligence.

Research Methodology

The study reviews the existing literature, survey reports, and online reports available and secondary data to summarize the various methodologies being used for the ETL in Business Intelligence. The exploration into this secondary data has initiated us to formulate a recommendation framework for efficient utility and efficiency of the ETL in Business Intelligence and Data Mining.

Review of Literature

Vassiliadis et al. [5]highlighted the key problems of the ETL tools to be complexity, usability, maintainability and price. Raman and Hellerstein[6] presented Potter's Wheel an interactive system that cleanses the data by detecting the discrepancies in transformations. The users could see the effects instantaneously.QELT- Query-based ETL methodology was given by Rifaiehand Benharkat[7]for the extraction of data. The system used the SQL queries for the transformations between the source and the target. ETL efficiency, optimality of algorithms for the ETL and its reliability is a major area of research these days and this was well predicted by Vassiliadis et al. [8]. Henry et al. [9] studied the ETL tools comprehensively through some testing procedures but could not reach to any universal criteria for choosing the ETL tools hence concluded that organizations could use these evaluation methods to build their own ETL tool. These commercial tools could generate significant improvements if only they could be made to incorporate UML, XML and EMF modeling technology [10] [11].

The two approaches ETL and E-LT (in terms of Full Pushdown, Target Pushdown and Source Pushdown)provide no performance difference when jobs are loaded into the data warehouse[12]. The commercial ETL tools can be categorized into activities like composition, coupling and swapping [13]. Few efforts are being proposed at the logical and physical level to optimize the ETL processes and the ETL flows [14].Akkaoui and Zimányi[15]presented a conceptual language for ETL modeling processes based on a business de-facto known as Business Process Modeling Notation (BPMN). This too does not solve the problem as the commercial tools are built on J2EE and SQL framework. ETL processes, ETL monitoring, ETL log forms the entire system from the user's point of view and perspective [17]. Incremental loading is much more useful and efficient than full reloading provided the operational data sources does not change [18]. However, the incremental loading is not always supported by the commercial tools. This limitation is overcome through an ETL management functionality for the large ETL processes [19].Jiang et al. [20] presented the construction of data warehouse through ontology-based approach that finishes the processes semantically and the transformations are done in much more efficient manner. Reddy et al. [21] presented a GUI ETL procedure for continuous loading of data in the Active Data Warehouse. The tool prepares the procedures, functions and triggers for the mappings and transformations much efficiently and in much less time.Jian and Bihua[22] analyzed the ETL tools for its openness and development and proposed a three-layer architecture. Pei et al.[23] proposed methods which can schedule and manage the metadata through a framework supporting flexible data transformations. The commercial tools cannot directly load the XML file into the data warehouse. Guohua and Jingting[24] justified this through the analysis of the characteristics of semi-structured data. A metadata driven ETL service model has strong flexibility, extensibility and has the ability to process large scale data. The model had the advantage of designing and sharing the ETL processes processed by open-source or commercial ETL tools [25].Bergamaschi et al. [26]gave an ETL tool that focused on data integration and analysis. The tool implemented a technique that semi-automatically defines mappings between a data warehouse schema, new data source and transformationsusing drill down operations. Muthukumar et al. [27] discussed the key issues related with creation, migration and harvesting Knowledge Repositories and harvesters using open source tools. Chen and Zhao [28]showed that the optimization technique would be improved through the SETL and new transformations would be generated at no extra cost. Sun et al.[29] presented the extraction, transformation and loading of heterogeneous data sources into data warehouse through SETL.SETL has been designed and implemented using PERL subroutine attribute and data partition. SETL can implement ETL job easily and perform ETL job efficiently, and the plug-in design makes SETL with high scalability, and the design that performing one ETL job in one ETL pipeline makes SETL with distribution environment support. Malik et al. [30] discussed a set of considerations which are required for effective workflow management and addition of different components in workflow scheduling layer ETL process may become more modular and efficient. Benchmarking of ETL processes is problematic and standardization of ETL tool work flows is needed [31].

Limitations of the ETL tools

There are a variety of ETL tools available in the market suffering from a general problem of interoperatability of the API and the proprietary metadata formats. This makes the functionalities of the ETL tools difficult to combine [32]. A Meta model ARKTOS capable of modeling and executing practical ETL scenarios and capturing common tasks of data cleaning, scheduling and transformation of data [5]. The commercial and data quality tools are classified by three perspectives. These three perspectives are data quality problems, generic functionalities for extracting data from data sources and for every data quality problem identified a detail must be kept as to which ones are addresses and which ones have been left [33]. The strengths and weaknesses of the hand coded ETL process and Tool-based ETL process were studied by Zode and the factors based on which the choice can be made between the two were discussed. Each one has its pros and cons and the criteria for selecting the tool still remains a topic of research. Setting up of criteria is not an easy task and this is generally based on the classification and categorization of the operations that are specifically marked for the organizations [34]. Such classification and categorization of the ETL operations in accordance to some built-in operations of some commercial tools was done by Vassiliadis. [35]Stressed on the fact that the current generation of ETL tools provides little support for systematically capturing business requirements and translating these into optimized designs that meet the correctness and quality requirements. The next generation of BI solutions will impose even more challenging requirements (near real-time execution, integration of structured and unstructured data, and more flexible flow of data between the operational applications and analytic applications), resulting in even more complexity in integration flow design. Hence, it is important to create automated or semi-automated techniques that will help practitioners to deal with this complexity.

Conclusions and Implications of the Study

The literature review highlights the fact that the data warehousing industry needs an ETL tool that is less complex and meets the criteria set by the organization for the cleansing, transformation and loading of the data from the sources to the target destination. The organizations are practically forced to move into in-house development of these ETL tools because the price of the tool-based ETL tools is out of the reach of the small and mid-size organizations. The open-source ETL tools are too generic to serve the purpose of these organizations and tweaking them implicitly means developing the tool from the scratch. The organization hasn't got that much time and budget to go in for hit and trial methods. It has to develop a tool that too based on certain business rules, constraints and criteria. A framework is required which can serve as a benchmark for the organizations to care for these parameters while developing the ETL tool. Here we present a framework for selection and developing an ETL tool that could be much easier to use and would integrate well with the business intelligence tool used at the organization.

Framework for Selection and Development of ETL Tool

Immense research has been done to figure out the problems associated with the different techniques deployed to populate the data warehouse. The framework suggested here would serve as a benchmark for the parameters to be incorporated while selecting or developing an ETL tool.

a) GUI Support: It is important for the ETL tool to have a Graphical User Interface as it provides the user the ease of use. The options shown on the screen should not be vague and should be understandable by the user. A tool with GUI interface is the first choice among the users and integrates well with other business intelligence tools in the organization.

b) Support for Incremental Update: The organizations does not provide an idle time to carry on the backup tasks or the loading of the data from every source available. Hence, an incremental update i.e. backing up from the last updated checkpoint saves the wastage of the resources and eases up the task to a great extent.

c) Inclusion of common tasks: The tool that is being selected or developed needs to accommodate some common extraction, transformation and loading tasks even though they might not be that much used in your organization. These tasks include data extraction from multiple sources, data aggregation, cleansing of data, reorganization and sorting of data and load operations. These activities may sound basic but they are important nevertheless.

d) Inclusion of common functions: As the basic operations need to be there, similarly basic functions should not be missed out. These include calculations, accessing the data from multiple operational data, mapping the source data to the format used in the organization, filtering the data for the target databases, performing table lookups, deriving values and changing the dimension support

e) Legacy data support: The most important task here is how to keep a support for the legacy data in the ETL tool. With the years gone by the data and its format has changed drastically but it is our legacy data that would help us predict the future. The task at hand is a complex one, since the data has to conform into a format wherein our business model software would work.

f) Real-time clickstream support for data warehousing: Web is fast becoming a channel for resource. It is a medium to reach to the potential customer. Clickstream data is the data generated by the clicks of the users on the websites. They are well used to study modeling behavior and valuable customer insights. The storage of these clickstream data in the data warehouse would integrate well with the business intelligence analysis used in organization.

g) Support for an e-Business environment: E-Business is the trend these days and integration of e-business with data warehouse at the metadata level with BI tools, ERP applications, CRM applications is a must, keeping in mind the future needs and scope.

h) Platform independence and scalability: The organization would evolve and so would be the Information Technology. The operating system used today and the applications used today would become obsolete in no time. Also, the enterprise is bound to grow hence, the independence of the ETL tool towards the platform and the scalability would only save our pocket in the future.

i) Conform to the business rules and API: The ETL tool being developed should conform to the business rules being implied in the organization. So set forth these rules in the tool there and then, negligence at the developing phase of the ETL tool poses cascading effects on the applications in the other business intelligence tools.

j) Basic functionality support: The ETL tool should not miss out on the basic functionalities. These may vary from organization to organization. These can be

-Multi-threaded processing
-compile source code
-concurrent processing of multiple source data streams
-Built-in transformation objects.
-Transformation objects
-target data models
-Metadata exchange
-Support of metadata standards, including OLE DB for OLAP

References

[1] Shilakes, C., &Tylman, J. (1998). Enterprise Information Portals. Enterprise Software Team.

[2] H. Galhardas, D. Florescu, D. Shasha and E. Simon.Ajax: An Extensible Data Cleaning Tool. In Proc. ACM SIGMOD (Dallas, Texas, 2000), 590. At http://www.eti.com/

[3] Vassiliadis, P., &Simitsis, A. (2009). Extraction, Transformation, And Loading. *Encyclopedia of Database Systems*, *32*.

[4] Kimball, R., Ross Margy,Thornthwaite Warren, Mundy Joy & Becher Bob. The Data Warehouse Lifecycle Toolkit.John Wiley and Sons, 1998.

[5] Vassiliadis, P., Vagena, Z., Skiadopoulos, S., Karayannidis, N., and Sellis, T. (2001). ARKTOS: Towards The Modeling, Design, Control And Execution Of ETL Processes. *Information Systems*, *26*(8), 537-561.

[6] Raman, V., &Hellerstein, J. M. (2001). Potter's Wheel: An Interactive Data Cleaning System. In *Proceedings of the international conference on Very Large Data Bases* (381-390).

[7] Rifaieh, R., &Benharkat, N. A. (2002). Query-Based Data Warehousing Tool. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP* (35-42). ACM.

[8] Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., and Skiadopoulos, S. (2005). A Generic And Customizable Framework For The Design Of ETL Scenarios. *Information Systems*, *30*(7), 492-525.

[9] Henry, S., Hoon, S., Hwang, M., Lee, D., and DeVore, M. D. (2005). Engineering Trade Study: Extract, Transform, Load Tools For Data Migration. In *Systems and Information Engineering Design Symposium, 2005 IEEE*(1-8). IEEE.

[10] Agrawal, H., Chafle, G., Goyal, S., Mittal, S., and Mukherjea, S. (2008). An Enhanced Extract-Transform-Load System For Migrating Data In Telecom Billing. In IEEE *24th International Conference onData Engineering, 2008. ICDE 2008. IEEE* 1277-1286.

[11] Morris, H., Liao, H., Sriram, P., Srinivasan, S., Lau, P., Shan, J., and Wisnesky, R. (2008). Bringing Business Objects into Extract-Transform-Load (ETL) Technology. In *e-Business Engineering, 2008. ICEBE'08. IEEE International Conference on* (709-714). IEEE.

[12] Ranjan, V. (2009). *A Comparative Study Between ETL (Extract, Transform, Load) And ELT (Extract, Load And Transform) Approach For Loading Data Into Data Warehouse*. viewed 2010-03-05, http://www. ecst. csuchico. edu/~ juliano/csci693/Presentations/2009w/Materials/Ranjan/Ranjan. pdf.

[13] Vassiliadis, P., Simitsis, A., &Baikousi, E. (2009). A Taxonomy Of ETL Activities. In *Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP*(25-32). ACM

[14] Castellanos, M., Simitsis, A., Wilkinson, K., and Dayal, U. (2009). Automating The Loading Of Business Process Data Warehouses. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* : 612-623. ACM.

[15] El Akkaoui, Z., &Zimányi, E. (2009). Defining ETL Worfklows Using BPMN And BPEL. In *Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP*(41-48). ACM.

[16] Ying-lan, F., & Bing, H. (2009). Design And Implementation Of ETL Management Tool. In *Knowledge Acquisition and Modeling, 2009. KAM'09. Second International Symposium on* (Vol. 1,446-449). IEEE.

[18] Jörg, T., &Dessloch, S. (2009). Formalizing ETL Jobs For Incremental Loading Of Data Warehouses. *Business Tech. and Web*, 57-64.

[19] Albrecht, A. (2009). METL: Managing and Integrating ETL Processes. In *VLDB PhD workshop.*

[20] Jiang, L., Cai, H., & Xu, B. (2010, November). A Domain Ontology Approach In The ETL Process Of Data Warehousing. In *e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on* (30-35). IEEE.

[21] Reddy, V. M., Jena, S. K., and Rao, M. N. (2010). Active Datawarehouse Loading By GUI Based ETL Procedure.

[22] Jian, L., &Bihua, X. (2010). ETL Tool Research And Implementation Based On Drilling Data Warehouse. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on* (Vol. 6, 2567-2569). IEEE.

[23] Pei, Y., Xu, J., & Wang, Q. (2010). One CWM-based Data Transformation Method in ETL Process. In *Database Technology and Applications (DBTA), 2010 2nd International Workshop on* (1-4). IEEE.

[24] Guohua, Y., &Jingting, W. (2010). The Design And Implementation Of XML Semi-Structured Data Extraction And Loading Into The Data Warehouse. In *Information Technology and Applications (IFITA), 2010 International Forum on* (Vol. 3, 30-33). IEEE.

[25] Xu, L., Liao, J., Zhao, R., & Wu, B. (2011). A Paas Based Metadata-Driven Etl Framework. In *Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on* (477-481). IEEE.

[26] Bergamaschi, S., Guerra, F., Orsini, M., Sartori, C., and Vincini, M. (2011). A Semantic Approach To ETL Technologies. *Data and Knowledge Engineering*, *70*(8): 717-731.

[27] Muthukumar, P., Suresh, P., ShaliniPunithavathani, S., and Nafeesa Begum, J. (2012). A Realistic Approach For The Deployment Of National Knowledge Repositories By Leveraging ETL Tools. In *Recent Trends In Information Technology (ICRTIT), 2012 International Conference on*(542-547). IEEE

[28] Chen, Z., & Zhao, T. (2012, November). A new tool for ETL process. In *Image Analysis and Signal Processing (IASP), 2012 International Conference on*(1-5). IEEE.

[29] Sun, K., &Lan, Y. (2012). SETL: A Scalable And High Performance ETL System. In *System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2012 3rd International Conference on* (Vol. 1, 6-9). IEEE

[30] Malik, S. R., Shamim, A., Bibi, Z., Khan, S. U., &Gorsi, S. A. (2013). Revised Framework for ETL Workflow Management for Efficient Business Decision-Making. International Journal of Computer Theory & Engineering, 5(3).

[31] Anitha, J., &Babu, M. P. (2014). ETL Work Flow for Extract Transform Loading. IJCSMC, Vol. 3, Issue. 6, June 2014, pg.610 – 617.

[32] Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems And Current Approaches. *IEEE Data Engineering Bulletin*, *23*(4), 3-13.

[33] Barateiro, J., &Galhardas, H. (2005). A Survey Of Data Quality Tools. *Datenbank-Spektrum*, *14:* 15-21.

[34]Zode, M. The Evolution Of ETL. Available at
http://hosteddocs.ittoolbox.com/mz071807b.pdf

[35] Dayal, U., Castellanos, M., Simitsis, A., and Wilkinson, K. (2009). Data Integration Flows For Business Intelligence. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (1-11). ACM.