DOI: http://dx.doi.org/10.26483/ijarcs.v9i4.6848

ISSN No. 0976-5697

Volume 9, No. 4, July – August 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A COMPARATIVE STUDY OF ETL TOOLS

Deepti Durgawat
Research Scholar,
JRN Rajasthan Vidyapeeth University, Udaipur (Rajasthan),
India

Manish Shrimali
Director, Department of Computer Science & I.T.
JRN Rajasthan Vidyapeeth University, Udaipur (Rajasthan),
India

Abstract–ETLstands for Extract, Transform and Load.It is an integral component of data warehouse application.Efficiency of any data warehouse application is mainly depended upon the effective implementation of this ETL process for acquiring strategic information. ETL tool combines theses three database functions to automate the process of retrieving data from one or more databases, then processes and place it in a target database. So, it is important to choose right ETL tool for organizations according to the requirements.In this study, a comparative review of some of the leading ETL tools has been discussed just to acquaint with their features and utility.

Keywords–ETL tools, data integration, data warehouse, business intelligence

I. Introduction

In the era of digitalization, information is processed for making intelligent business decisions. Most of the organizations are now employing own data warehouses for making quick and effective decisions to keep them ahead.Decision making based on data warehouse and business intelligence forces the enterprises to use different but coexisting information systems.Choosing the appropriate data warehouse application with right ETL toolfor implementing pertinent ETL process saves time and money, both.

The ETL (Extract, Transform, and Load) is the process of extracting data from one or multiple data sources, then transformed to required standards of data warehouse before loading into it. ETL is an essential component of any data warehouse and analytics. There is wide variety of ETL tools available for different requirements and applications.The objective of this research paper is to explore various features, pros and cons of the leading ETL tools. Furthermore, evaluation of these ETL tools isbased on certain criteria for comparison. This is carried out in this study to provide knowledge to help users in choosing the right ETL tool as per their requirements.

Data integration is synchronization of data that involves practices, tools and techniques for achieving consistent sharing of data across a wide range of subject areas including business intelligence, data standards and structure type in an organization. Data integration is becoming more important for information-centric infrastructures that require systematic investment for frictionless access to and delivery of diverse databeyond limitations of enterprise and system boundaries.ETL and data integration tools acts as an interface to solve the problem of diversity of data integration by addressing range of data patterns and architectural styles resulting in cohesive database.

Based on development styles, ETL tools may be mainly classified as in-house hand coded ETL process and tools based ETL process.Before the evolution of ETL tools, in-house development for implementing ETL processeswere used to meet specific requirements of organization.This task is challenging and cumbersome as it requires complex coding involving multiple resources.Also, In-house implementation of ETL processesinvolves overheads and demands continuous tedious efforts to meet the challenging requirements generated by high volume of dynamic data.Another category is off-the-shelf tool based ETL processes with updated functionalities and capabilities. These tools are easy to use with full GUI support to data profiling, monitoring, debugging, scheduling and aggregating ETL processes.They have independent internal metadata repositories to cater requirements of data warehousing and business intelligence. It relieves the overhead of developing, implementing and maintaining the complex ETL routines and workflows.

II. Review of Literature

Literature review provides a bird's eye view of the research work by other researchers in the domain of ETL tools and data integration platforms. It is essential to avoid the possibility of unnecessary duplication of work and efforts. A brief summary of researches related to present study has been presented in this section. The review of literature indicates that despite very diverse theoretical perspectives and applied approaches, many researchers have identified to encapsulate the unique characteristics of ETL tools.

The objective of this study is to explore the ETL tools from the available literature to address the core aspects and characteristics of ETL and data integration platforms in the perspective of business intelligence and data warehousing. Various relevant publications including conference and seminar proceedings and organizational reports were part of literature review and qualitative analysis. This initial level of study is the starting block for the further research work in ETL domain. The snowballing technique was used for this purpose. In this technique, the data have been searched by following references in the bibliographies of various related articles. Keywords like ETL, data integration platform, survey, comparison along with ETL tools name were used to find diverse and relevant literature for this review.

III.Methodology

There are many data integration tools available, but based on Gartner report, products like Microsoft SQL Server IS, TalendOpen Studio, SAS Data Integration Studio etc. have been chosen from the market leaders and the challengers quadrant.A framework of criteria and categories for comparing the ETL tools is prepared (Table1) that is based on the analysis of various survey reports, articles, websites, whitepapers, books and journals. Following table summarizes comparison based on selected criteria:

Table 1: Comparison of leading ETL tools

S.No.	Criteria	IBM InfoSphere	Informatica Platform	Talend Open Studio	Oracle Data Integrator	Microsoft SSIS	SAS Data Integration Studio
1	Platforms	6	5	7	6	1	8
2	Engine based or code generated	Both	Engine based	Code generated	Code generated	Both	Code generated
3	SaaSapplication	✓	✓	✗	✓	✗	✗
4	Ease-of-use	High in logical orders	Easy GUI, requires appropriate training	Does have GUI as add-on	Highly User Friendly	Most easy GUI, requires little training	Most easy GUI, requires little training
5	Join tables using graphical tool	✓	✗	✓	✓	✗	✓
6	Auto correction / suggestion	✓	Partial	✓	✓	Partial	✓
7	Compiler / Validate	✓	Partial	✓	Partial	✓	✓
8	Separate Modules for real-time or batch	✗	✓	✓	✗	✗	✓
9	Data change recognition mechanisms	Logging + triggers	Logging	Message queuing + triggers	Message queuing + logging + triggers	Message queuing + logging + triggers	Message queuing
10	Native connections	41	50	35	22	4	18
11	Real time connections	2	6	3	3	2	3
12	MPP	✗	✗	✓	✗	✗	✓
13	Partitioning	✓	✓	No	No	✓	✓

1. Criteria-wise comparisonanddescription

Comparison of selected ETL tools for this study is summarized in Table 1. It can be observed that it includes standaloneETL integration tools from well established companies like IBM to newer companies like Talend.Some of the criteria are satisfied by all the selected tools. Such as, all the selected tools supports reusability feature by providing facility of user defined functions and debugging features. Also, the tools taken for comparison have the basic capability of scheduling jobs for handling interdependencies. Symmetric multiprocessing is available in all the selected ETL tools. These tools provide transformation methods for normalize and denormalize data.

The platform criteria signify count of different platforms supported by ETL products.It is observed that SAS data integration studio supports more number of platforms in comparison to other tools. It is observed that IBM InfoSphere and Microsoft SSISare engine based as well as code generated, while others are either engine based or code generated.The SaaS (software as a Service) criterion informs about facilities of being supported by cloud-resident data sources. Different reviews and literature suggests that Oracle Data Integrator is most easy to use in terms of minimum training period required for the developer besides providing user friendly environment.Next, Microsoft SSIS and SAS Data Integration Studio follow in terms of Ease-of-Use criterion.Some of the selected tools provide facility to join tables from database in graphical manner. It is another dimension of Ease-of-Usecriteria that uses GUI for implementing ETL processes that eases user's efforts.In case of any syntax or field name error, SSIS and Informatica does not provide any autocorrect feature or suggestions. However, all tools provide the location of errors during compilation or validation.Informatica, Talend and SAS provides separate modules for real time and batch processing. One of the important features is data recognition mechanism of changes in data after extraction and transformation. ETL tools use triggers, message queuing, logs and journals or different combination of these mechanisms to recognize changes in data. Informatica and InfoSphere provide more native connections to the database resources. Due to this, extraction of data becomes more efficient. Informatica is also ahead in facilitating effective queuing mechanism by providing more number of real time connections. Talend and SAS have the functionality of MPP system by having their own internal and external memory and databases resulting in high performance. Partitioning is another feature that makes it possible to partition database to allocate machine or processor for processing the data.

2. Tool-wisecomparison and description

IBM InfoSphere is easy to use, flexible and provides common metadata platform oriented towards market demand. IBM offers many ETL and data integration products including different variants of IBM InfoSpherefor a strong satisfied customer base. However, Processing power requirements are increased with growth of data.

Informatica Platformis a high performance feature rich enterprise data integration platform providing database-neutral solution based on solid technology. It has the capability of real-time data integration with compatibility with many different data sources, including SQL as well as non-SQL databases for ETL workloads. It is one of the consistent, mature and leading data integration tool in the market. However, with a challenging learning curve, smaller organizations feel it costlier with respect to their requirements and budget.

Talend Open Studiois GUI based open-source ETL data integration tool with an approach for supporting code generation and writing customized queries. The Talend platform has data quality features and compatible with pre-built integrations for data sources available on-premise and in cloud. Open source version of Talend fulfils the requirements of smaller organization, but it is not a complete business intelligent suite. Talend's paid Data Management Platform is preferred by larger enterprise for additional tools and featured for productivity and data governance.

Oracle Data Integrator (ODI) is a comprehensive data integration solution considered as one of the leaders in the ETL markets with a support to ELT workloads. It is best choice for current users of Oracle applications, as ODI creates a single data management ecosystem for all Oracle data warehousing applications and tools.

Microsoft SQL Server IntegrationServices (SSIS)provides real-time standardized data integration. It has message-based capabilities with relatively low in cost, easy to use and faster implementation features. It has advantages of larger distribution channel and excellent support. However, it is best suited for Microsoftapplications, tools and environments.

SAS Data Integration Studiois a powerful data integration tool with lots of multi-management features. It provides great support to all categories of companies. It is very flexible with a capability of working with different operating systems and data sources.

IV. Conclusion and Future Work

Enterprises of all sizes now have access to ever-increasing vast amount of data. All this available huge data is of no use if it not efficiently processed and analysed to reveal the invaluable data-driven insights. The study provides overview and summary of features for some of the selected ETL tools. It leads to conclusion that ETL and data integration tools are necessary for enterprises using Business Intelligence (BI) and effective decision making. The utilization of ETL tools is based on requirements of organizations.

Oracle Data Integrator and IBM InfoSphere are more suitable for larger organizations.Some of the criteria are excluded for comparison, such as costing of ETL product. This criterion is related to the requirements of enterprise.However, the study has certain limitations, as the selected ETL products are not tested in real world environment. The research has been conducted based on secondary data collected from journals, articles, reports, vendor's website and documents. In future, support of upcoming technologies like cloud computing, big data analytics, IoT, machine learning will play important role in ETL tool development.This research can be extended by adding weights to the criterion and testing the product in real world environment.

V. References

[1] Madsen, M. (October, 2004). "Criteria for ETL Product Selection". InfoManagement.

[2] Larson, B. (2008). "Delivering BusinessIntelligence with Microsoft SQLServer". New York:McGraw-HillOsborne Media.

[3] Levin Jonathan (2008),"Open SourceETL tools vsCommercial ETL tools". Retrievedfrom http://www.jonathanlevin.co.uk/2008/03/open-source-etl-tools-vs-commerical-etl.html

[4] Mark A. Beyer, Eric Thoo, EhtishamZaidi, Rick Greenwald (2016), Magic Quadrant for Data Integration Tools, Retrived from https://www. gartner.com/doc/reprints

[5] Nils Schmidt, Mario Rosa, Rick Garcia, Efrain Molina, Ricardo Reyna and John Gonzale, "ETL Tool Evaluation – A Criteria Framework", In Proceedings of the SWDSI (Southwest Decision Sciences Institute) 2011 conference.

[6] Zode, M. (2007), "The Evolution of ETL – From Hand-coded ETL to Tool-based ETL".Cognizant Technology Solutions.

[7] Talend (2017), "Talend Open Studio for Data Integration User Guide", Adapted for v6.5.0M1

[8] T. A. Majchrzak, T. Jansen and H. Kuchen (March, 2011), "Efficiency evaluation of open source ETL tools", In Proceedings of the ACM Symposium on Applied Computing, pp. 287-294.