International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4 (2016) pp 2557-2559
 https://dx.doi.org/10.37622/IJAER/11.4.2016.2557-2559

A Technical Comprehensive Survey of ETL Tools

Vaishali A. Kherdekar
Assistant Professor, MIT Art's Commerce & Science College (MIT ACSC), Alandi,
Affiliated to Savitribai Phule Pune University Pune, Maharashtra State, India.

Pravin S. Metkewar
Associate Professor, Symbiosis Institute of Computer Studies and Research(SICSR),
Affiliated to Symbiosis International University(SIU), PUNE-411016, Maharashtra State, India.

Abstract
In modern days ETL tools are vey useful in data integration and data warehousing. Input is given to the datawarehouse through ETL. ETL means Extraction, Transformation and Loading. ETL tools transfer data from one source system to another source system.
As these tools are mainly used in Busineness Intelligence and Data Warehousing, there is lot of space for their progress.There are lot of ETL tools available in the market varying from version to version to stay proficient against other tools. Each and every tool has its own features and limitations. In this paper we have carried out technical survey of existing ETL tools and benchmarking of these tools has been performed by considering certain parametrs including scalability, reusability, interoperability, support to big data, parallelism, usability, flexibility etc. Finally, problems and challenges of ETL tools have been discussed thoroughly and its state of the art is summarized.

Keywords : ETL tools, Data warehouse

Introduction

Now a days data warehouse is used in industry to maintain optimiza model of data for further mining and uage and also for report generation. By using datawarehouse one can maintain historical data and used it in decision support system. To construct data warehouse model, ETL tools is being used. ETL tools act as basis for construction of data warehouse. Input is given to the data warehouse through ETL. ETL stands for Extraction, Transformation and Loading. In extraction phase, data is extracted from various heterogeneous sources in different formats such as flat files, databases, xml files etc; it means extraction of data is achieved with the help of structured and semi-structured databases and files. In transformation phase extracted data is transformed into specific format for data analysis. In loading transformed data is loaded into datawarehouse. Now a days, a large number of ETL tools are available in the market. However, in general, they follow different design and modeling techniques, and uses different internal language.

Sample selection of ETL Tools

In market variety of ETL tools are exists either that are from open source or proprietary tool. We have considered few reputed tools for our study in order to perform benchmarking of these tools that are including Pentaho, Clover ETL, Rapid Miner, Jedox, JasperSoft and Talend tools.
Pentaho ETL tool was established by the Pentaho Corporation, United States. In market it is named as Pentaho Kettle. It is open source and provides services for business intelligence and data integration. The transformations carried out in Pentaho is stored in XML.It is executed in Java.
CloverETL was developed by JavlinInc in 2002. It is Java based data integration tool used for transforming and distributing data and other functionalities of data warehousing. It can be used as either standalone or embedded to server application. It works on different platforms.
Rapidminer tool is specially used for Regression Analysis, Gaussian process And statistical process. It provides a wide range of functionalities and support.
Jaspersoft extract and transform data from multiple systems and loads it into data store for reporting and analysis. It works as ETL job designer having data integration capabilities.
Talend was introduced in 2006. It is also open source Java based tool used for data integration and data analysis process.

Parametric based benchmarking of ETL tools

On the basis of analysis so far, we may put our insight ahead. Usability of ETL tool plays an important role in benchmarking i.e., tool should be easy to use, understand and fast to get used. The component used in ETL tool should be reusable. Transformations should be used in multiple times in multiple jobs. In interoperability ETL tool should be run on any platform. Scalability of ETL tool indicate handling of large volume of data which also include parallelism, partitioning and clustering. In parallelism many files run in parallel thus utilizing multi core h/w architecture. In partitioning data is distributed over parallel streams. In clustering workload is divided over multiple machine. Flexibility of ETL tool provide freedom to developer to use any design flow and should not limit designer by offering only a fixed way of working. Map base ETL tool limiting the freedom to design jobs where as process based tool always provide additional steps if needed due to change in business requirements. By considering these factors; dilution of considered ETL tools is summarized as below.

International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4 (2016) pp 2557-2559
 https://dx.doi.org/10.37622/IJAER/11.4.2016.2557-2559

Table 1: Benchmarking of ETL Tools

Parameters	Pentaho	Talend	Clover ETL	Jedox	Jaspersoft	RapidMiner
Usability	Very good	User friendly	User friendly	User friendly web interface	User friendly	User friendly
Reusability	Yes	Yes	Yes	Yes	Yes	NA
Interoperability	Yes	Yes	Yes	Yes	Yes	Yes
Scalability	Cluster (carte server).	Highly scalable	Scalable	NA	Yes	NA
Flexibility	More flexible	NA	Yes	More flexible	Less flexible	NA
purpose/type	Business intelligence	Data integration, data quality and data management	ETL tool	Business Intelligence	Business Intelligence	Statistical Analysis, DM, predictive analysis
Support to big data	Hadoop, NoSQL, and Analytic databases	Hadoop, Cassandra, MongoDB, Hive, and even PIG	Hadoop HDFS data storage and HIVE	NA	Hadoop and NoSQL Data Sources.	NA
Generate code or language to define their own component	Javascript	Perl or Java Script	CTL Scripting Language	JavaScript, Groovy	Perl or JavaScript	Groovy
Supported Platform	Windows, Unix, Linux	Windows, Unix, Linux	Windows, Unix, Linux, OS X	Windows	Windows, Red Hat, Ubuntu, SuSE and other leading distributions	Windows, Linux
Web service support	No	LDIF	No	SOAP,	LDIF	No
Speed/Performance	Faster	Slower	utilizes multiple CPUs/cores and can run on a cluster of computers to increase performance	NA	NA	NA
Parallelism	Parallel job execution (asynchronous). Parallel mapping execution (e.g. parallel DML).	support component & pipeline parallelism to speed up executation	Pipeline & Parallel job execution help to cope with big data problems.	support component & pipeline parallelism to speed up executation	NA	NA
GUI	Design tool (spoon) based on SWT	Eclipse based GUI	Eclipse based GUI	NA	Eclipse based GUI	NA

Dilution of Pentaho ETL

Pentaho is SWT (Standard Wideget Toolkit) based design tool. It is user friendly ETL tool and platform independent i.e. works on windows, linux, unix. It is Javascript based tool and support parallelism. It's main purpose is business intelligence and provision for big data such as Hadoop, NoSQL, and Analytic databases. From our study it reveals that it does not support web services.

Dilution of Clover ETL

Clover ETL is an Eclipse based GUI. It is also user friendly and works on different platforms such as windows, linux, OS –X etc. It uses it's own scripting language to write the transformations and components known as CTL (Clover Transformation Language). It is mainly used for ETL purpose and provision for big data such as Hadoop HDFS (Hadoop Distributed File System) data storage and HIVE. It supports pipelining and parallel job execution which helps to cope up with big data problems. From our review; it reveals that pentaho does not support web services.

Dilution of Talend ETL

Talend is user friendly and an Eclipse based ETL tool. It uses Perl or Javascript language to define their own component. It is highly scalable and works on windows, unix and linux. It is specially used for data integration, data quality, data management and it's pipelining and parallelism feature speed up it's execution. It supports big data such as Hadoop, Cassandra, MongoDB, Hive, and even PIG (Programming tool-Platform for processing and analyzing large data). It supports web services such as LDIF (LDAP-Light Weight Directory Access Protocol) Data Interchange Format).

Dilution of Jaspersoft ETL

From our observations, it shows that features like usability, reusability support to web services and scripting language used to generate code of Jasper Soft ETL that is similar to Talend ETL. It works on Windows, Red Hat, Ubuntu, SuSE and other leading distributions. It is used for business intelligence purpose and support big data such as Hadoop and NoSQL Data Sources.

Dilution of Jedox

Jedox is having usesr friendly web interface. It works only on windows. It provide Groovy and Javascript language to write the components and these components can be used in multiple jobs in multiple times. It's pipelining and parallelism feature helps to speed up the execution and it is flexible one. It support a SOAP (Simple Object Access Protocol) web service for exchanging structured information.

Dilution of Rapid Miner

It is used for Statistical Analysis, Data Mining, predictive analysis. It is groovy based tool and works on windows and linux.

Conclusion

This paper gives overview of ETL tools by putting technical emphasize on it. Our technical survey reveals that most of the ETL tools are Eclipse based and Eclipse is open source IDE which Support to big data. Few tools support Hadoop, few tools support Hive and few based on supporting PIG. Out of six ETL tools that are selected for technical comprehension ; three tools were gives support to Web services where as three tools does not. Most of the ETL tools support parallel job execution. Almost all tools works on windows operating system. This is the current status of observed and identified tools and brief about each tool above independently because its importance is singularly important.

To brought these details into a single tool is bit challenging due to varied technologies, IDE's, different domains and its corresponding aspects, different operating systems, types of networks and networking techniques. In other words, we may conclude that a unique API would be developed for the purpose of accessing data from any remote location by looking into discussed factors.

Finally, we conclude that there is no such unique tool which combines all these features and hence this is the need of the day.

References

[1] Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos "Conceptual modeling for ETL processes" Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP 2002.

[2] A book Nestor Rodriguez, Kent Lawson, Eddie Molina Data Warehousing Tool Evaluation – ETL Focused.

[3] A book on Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho...By Matt Casters, Roland Bouman, Jos van Dongen

[4] Sachin Naik, P.S. Metkewar "A Technical Comprehensive Survey For User Input Models For Mathematical Expression", IJAER ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 28369-28378

[5] ETL as a Necessity for Business Architectures by Aurelian TITIRISCA *Database Systems Journal* vol. IV, no. 2/2013

[6] Kalpana Rangra Dr. K. L. Bansal " Comparative Study of Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 6, June 2014.

[7] A book on" Data Mining Concepts & Techniques" by Jiawei Han, Micheline Kamber, Jian Pei.

[8] A book on "Data Warehousing in the Real World A practical guide for Building Decision Support Systems" by Sam Anahory, Dennis Murray.

[9] AStefano Rizzi, Alberto Abelló, Jens Lechtenbörge, Juan Trujillo" Research in data warehouse modeling and design: dead or alive?" Proceeding of the 9th ACM international workshop on Data Warehousing and OLAP 2006.