A comparative Review of Extraction, Transformation and Loading Tools

Amanpartap Singh PALL, Dr. Jaiteg Singh KHAIRA
School of Information Technology, APJIMTC, Jalandhar
Chitkara Institute of Engineering and Technology, Rajpura
amanpall@hotmail.com, jaitegkhaira@yahoo.co.in

Business today forces the enterprises to run different but coexisting information systems. However, data warehousing enterprises have a dilemma of choosing the right ETL process and the right ETL tool for their organization as one wrong step or choice may lead to a series of losses both monetarily and by time, not to mention the amount of laborious work that the workers would put in. The organization can choose from a variety of ETL tools but without exploring or the knowledge of their features this would again result in a bad decision making process. In this paper, we have tried to present a comparative review of some of the leading ETL tools just to acquaint the users with its features and drawbacks.

***Keywords**: ETL tools, tool comparison*

1 Objective

The purpose of our research paper is to provide a comparative review of the various features of the leading ETL tools (Table 2). Furthermore, we are evaluating these ETL tools based on some criteria which we think are suitable for an ETL tool to have. Through this comparison we have tried to provide an upfront knowledge to the users as to what could be the alternatives amongst the market leaders. No doubt there are other tools also available, but we have chosen only the market leaders as per Gartner Report 2010. We have carefully chosen the market leaders, the challengers e.g. the Microsoft SQL Server IS and some open source products like Pentaho and CloverETL.

Research Method

To conduct our research we needed to create a framework of criteria (Table 1) which would allow us to compare the ETL tools against each other. We conducted research in different survey reports by leading groups, articles, journals and books and websites. Upon analysing these Reports, whitepapers, journals, website we successfully revealed the criteria and its categories which will allow us to compare.

Introduction

Data integration involves practices, architectural techniques and tools for achieving consistent access to and delivery of data across a wide range of subject areas and structure types in an enterprise. Data integration capabilities are at the heart of the information-centric infrastructure and will power the frictionless sharing of data across all organizational and system boundaries. There is no doubt that the investment in data integration is increasing day by day and has started to be a part of the budget that the organization sets aside each year. Specifically, market demand is becoming more diversified, as buyers procure tools with intent to support multiple use-cases. The traditional focus of activity has been in support of business intelligence initiatives. While this remains the most significant use-case driving demand, many others have emerged. Further, synchronization of data between

operational applications and across enterprise boundaries (between trading partners or between on-premises and cloud-based applications) also represent areas of growth. These requirements have generally been met via point-to-point interfaces supported by data integration tools. With the on-going evolution of the data integration tools market, separate and distinct submarkets continue to converge, both at the vendor level and the technology level. This is being driven by buyers' demands. Specifically, organizations increasingly acknowledge a diversity of data integration problem types that are supported by equally diverse architectural styles and patterns for data delivery. It is also being driven by vendors' actions — specifically, vendors in individual data integration submarkets organically expanding their capabilities into neighbouring areas, and acquisition activity bringing vendors from multiple submarkets together. The result is a progressively maturing market for complete data integration tools that address a range of different data integration styles based on common design tooling, metadata and runtime architecture.

Categories of ETL tools

ETL tools may be categorized into two broad categories

a) **Hand-Coded ETL Process**: ETL tools that are in-house developed in Perl, COBOL, C, and PL/SQL to e extract data from multiple source files, transform the data, and load the target databases. The programs written using this method were lengthy and hard to document. The ETL developer has to use different programming languages to perform the ETL task, such as; Perl scripts for extracting the data from source systems, performing transformations, and SQL Loader and PL/SQL Bulk procedures were used to load the data in target warehouse. Hand-Coded ETL tools have the advantage that the metadata created can be managed directly and they give the flexibility to the developer to manipulate to new needs and of-course unit testing is much easier. However, its limitations are also there, to cater to the continuous changes in the high volumes of data generated through various sources the programs need to be modified frequently which causes a burden on the overall project. And moreover, changes done at the metadata needs tables to be modified as well which in hand-coded is stored separately, so changes are done manually. And lastly the hand-coded ETL are generally slow in execution as they are single threaded whereas the modern tool-based tools are multiple threaded and run on high speed engines.

b)Tool-Based ETL: Since, the hand-coded tools involve overheads and are slow in execution hence many vendors developed these tools to be purchased by the organizations. These ETL tools started from simple extractions on mainframes to target database and now-a-days they are available in full GUI's with added functionalities and performances. These are the ETL tools of today that provide transformation features, support multiple input or output database or flat files, multi-dimensional designs, surrogate key generation, various transformation functions and native database or O/S utility. They have internal metadata repositories that may be different from the data warehouse metadata repository. They eliminate the overhead of developing and maintaining the complex routines and transformations in ETL workflows. Also, these tools are providing user friendly GUI's which enables the developer to work without under going through training. These tools also have the features such as monitoring, scheduling, bulk loading, incremental aggregation, etc.

The ETL tools today can further be classified into four subcategories

(i) **Pure ETL tools**: These products are independent of the database and the Business Intelligence tool with which it will be used. The companies do not rely on any other product for the functionalities offered by them and they also allow migration to different database without changing the integration process.
(ii) **Data base integrated**: These products are supplied as an option when you buy the database software and some of the functionality is built into the database and not available separately in the ETL tool itself.
(iii) **Business Intelligence Integrated**: These are the products from the same supplier as the BI software .In many cases these re separate products and the supplier will claim that they can be used independently of the BI tool.
(iv) **Niche Product**: These are the products that don't fit well into any of the above mentioned groups, but still have considerable ETL functionality in them.

Research Method

The ETL tools selected for the comparative review are only the market leaders although there are so many tools, but selection of all of them would be done in a step by step manner since only peers should be compared with each other. To compare these tools a criteria or basis on which these are compared should be a universal one. To come to a universal one we reviewed various journals, articles, books and more importantly the reports that are generated from time to time to remove any bias.

Table 1. Various criteria for the comparative review

Sr. No	Criteria
1	Sales in
2	Standalone or Integrated
3	Platforms
4	Version
5	Engine based or code generated
6	SaaS
7	Ease-of-use
8	Reusability
9	Debugging
10	Corrections to syntax errors
11	Compiler/ Validate
12	Separate Modules
13	Data mechanisms
14	joined tables as source
15	address information support
16	native connections
17	real time connections
18	Scheduler
19	Pivoting/de-pivoting
20	SMP
21	MPP
22	Grid

23	Partitioning
24	CWM support
25	Integration batch-real time
26	Package / enterprise applications

Source: The Data Integration & ETL Product Survey 2013, Passionned Group

The comparative review has been done only for these tools:

Table 2. The products\tools for which the above criteria has been reviewed

Sr. No	Organization	Product Name	Current Version
1	IBM	Information Server	8.1
2	Informatica	PowerCenter	9.5
3	Talend	Talend Open Studio for Data Integration	5.2
4	Oracle	Data Integrator	11.1.1.5
5	Microsoft	SQL Server Integrated Services	10
6	SAS	Data Integration Studio	V4.4
7	Kettle	Pentaho Data Integration	4.1
8	CloverETL	CloverETL	3.1.2

Comparative Review of the ETL tools

Table 3. Comparative review of leading ETL tools on different criteria

Criteria	IBM Information Server	Informatica PowerCenter	Talend Open Studio	Oracle Data Integrator	SQL Server Integration Services	SAS Data Integration Studio	Pentaho	Clover ETL
Sales in	1996	1996	2007	1999	1997	1996	2006	2005
Standalone or Integrated	Standalone	Standalone	Standalone	Standalone	Standalone	Standalone	Standalone	Standalone
Platforms	6	5	7	6	1	8	4	7
Version	8.1	9.5	5.2	11.1.1.5	10	v4.21	3.2	2.9.2
Engine based or code generated	Both	engine based	code generated	code generated	Both	code generated	engine based	Engine based
SaaS	Yes	yes	no	yes	-	No	not standalone	no
Ease-of-use	high in logical orders	yes	yes	highly user friendly	Highly	highly	no	no
Reusability	Yes	yes	yes	yes	Yes	yes	yes	yes
Debugging	Yes	yes	yes	yes	Yes	yes	no	no
Corrections	Yes	half	yes	yes	-	yes	no	yes

Criteria	IBM Information Server	Informatica PowerCenter	Talend Open Studio	Oracle Data Integrator	SQL Server Integration Services	SAS Data Integration Studio	Pentaho	Clover ETL
to syntax and field names								
Compiler/ Validate	Yes	half	yes	half	Yes	yes	yes	yes
Separate Modules	No	yes	yes	no		yes	no	yes
Data mechanisms	logging+triggers	logging	message queuing + triggers	message queuing +logging+ triggers	message queuing + logging + triggers	message queuing	no	message queuing+ triggers
joined tables as source	Yes	no	yes	yes	No	yes	no	no
address information support	All	all	all	all			third party	all
native connections	41	50	35	22	4	18	20	7
real time connections	2	6	3	3	2	3	3	3

Source: The Data Integration & ETL Product Survey 2013, Passionned Group

Table 4. Comparative review of leading ETL tools on different criteria

Criteria	IBM Information Server	Informatica PowerCenter	Talend Open Studio	Oracle Data Integrator	SQL Server Integration Services	SAS Data Integration Studio	Pentaho	CloverETL
Scheduler	yes	yes	yes	yes	yes	yes	yes	yes
Pivoting/de-pivoting	yes	yes	yes	yes	yes	yes	yes	yes
SMP	yes	yes	yes	yes	yes	yes	yes	yes
MPP	no	no	yes	no	no	yes	yes	yes
Grid	yes	yes	yes	yes	no	yes	yes	yes
Partitioning	yes	yes	no	no	yes	yes	yes	yes
CWM support	half	yes	yes	yes	no	yes	no	yes
Integration batch-real time	yes	yes	yes	yes	no	yes	half	yes
Package / enterprise applications	8	7	9	8	1	5	2	0

Source: The Data Integration & ETL Product Survey 2013, Passionned Group

Description and Comparison

Here, we have done the comparison on the criteria mentioned above and an overall description of the tools, this would bring out the clear picture as to which tool is best in which type of criteria while the criteria comparison would suggest which tool is the best in a particular criteria. The following are the description and the comparison:

1. Criteria-wise description and comparison

As can be seen from Table 1 companies whose sales started in 1996 like Information Center by IBM to some of the newer companies like Talend Open Studio have been included for the comparative analysis. Though all the ETL tools given above are standalone tools however they have they have their own scoring points. The comparative analysis by each criterion is given below:

Platforms: This criteria signifies how many platforms are supported by the ETL product e.g Windows (all versions have been counted as one), Linux, Solaris etc. As can be seen Microsoft SQL Server has the least platform support i.e. Windows while SAS Data Integrator scores here by providing support to 8 different types of platforms which is indeed a plus point.

Engine Based or Code generated: While Both Information Center and SQL Server are both engine based and code generated all others products are either code generated or engine based.

SaaS: This criterion has been included to see whether or not the product is available as software as a service and it was found that while Information Center, PowerCentre and Oracle provide this facility others do not. This generally means these products can be a part of the cloud computing the latest facility being provided by organizations. This is a major plus point for these products and perhaps one of the reasons that they are so widely being used today.

Ease-of-Use: Ease of use includes how easy is it to use the product, how quickly can it be learnt, number of training days required for the developer and the user to learn the product , screen element designs , GUI interface , and most importantly does it work the way ETL tool should work. Oracle Data Integrator has been found to be the most users friendly followed by SQL Server and SAS Integrator. However, that does not mean that others are not user friendly it's just that these three tools and Oracle Data Integrator in particular conforms to the above said criterion more than the others.

Reusability: How are the components reused whether they are parameter driven, does it support user defined functions to be available to other programs. All the above tools are very much conforming to this criterion.

Debugging: Apart from the Pentaho and CloverETL all others provide a good debugging facility either step by step or row by row.

Corrections to syntax and field names: Pentaho, SQL Server and Informatica does not provide any automatic suggestions if there is an error in syntax or field names whereas this is available in all other tools.

Compiler\validate: How easy it is to locate errors and if any are they highlighted in the code at a click. This facility is available with every tool.

Separate Modules: Usually the tool is made up of at least two modules the real time module and the batch module. Now can they be bought separately? Informatica, Talend, SAS and CloverETL has got this provision whereas this is not the case with Oracle Data Integrator, IBM Information center, SQL Server and Pentaho.

Data Mechanism: The data changes when its extracted and transformed .So the question is how is it recognized i.e. how is

the changed data recognized. IBM information Centre uses triggers and the logs and journal entries to recognize the changed data while Informatica does it with only the logs and journals. Talend and CloverETL do it with the message queuing and database triggers the Oracle Data Integrator and SQL Server leaves no options to neglect such changes as it incorporates all the three techniques. Pentaho stands out in this one as it does not provide this facility.

Joined tables as source: Can you join two tables in a graphical manner letting the database execute the join as opposed to letting the ETL tool join the tables. Informatica PowerCenter, SQL Server Integration, Pentaho and CloverETL does not provide this which is a major drawback.

Address information support: All types of address information are supported by all of the above tools.

Native connections: How much and which native connections does the ETL tool support? (ODBC, OLE DB and flat files excluded).Informatica PowerCenter provides the maximum native connections to the various database sources thus extraction from these sources becomes much more efficient. The IBM Information Centre and Talend Open Studio is not lacking as it follows the Informatica very closely. SQL Server lags here as it can only provide only four types of native connections.

Real time connections: How many and which type of message queuing products can the tool connect to? Here also the informatica PowerCenter takes the cake providing the maximum connections.

Scheduler: whether or not there is an ability in the tool to schedule jobs based on interdependencies. Or in other words is the scheduler capable of handling dependencies. All the tools these days support scheduling functionality because it is regarded as a basic necessity for the ETL tool to schedule jobs. The tools taken for comparison does not lack behind in this criteria as they all support it.

Pivoting/de-pivoting: Is it possible to transform denormalised data, putting data in the column names, into rows and the other way around, transform (highly) normalised data to de-normalised data, putting data in the columns. Again this facility is available within each tool.

SMP: Is Symmetric Multiprocessing supported? Standard in Windows NT and UNIX. The processors in SMP systems share their internal and external memory. SMP is available in all the above mentioned tools.

MPP: Every processor in a MPP system has its own internal and external memory and database, allowing high performance to be achieved. These databases should be synchronized. From the ETL tools taken up for the review only Talend, SAS, Pentaho and CloverETL possess this functionality.

Grid: Can an ETL process run on a ‘grid’ of computers or servers? Only SQL Server Integrated Services fails to provide the grid facility whereas all others do.

Partitioning: Is it possible to partition based on, for example, product codes, to determine on which machine or processor the data has to be processed? Talend and Oracle does not let the partition to take place, all others does.

CWM support: Is the ETL tool CWM-compliant, in other words does it support the Common Warehouse Meta Model? If you are looking for a common warehouse meta model then IBM Information Server, SQL Server Integrated Services and Pentaho do not provide you this facility.

Integration batch-real time: Is it possible to define within the ETL tool process flows moving and transforming data in real-time and in batch? While the SQL Server Integrated Services does not provide this and the Pentaho provides this up to a certain extend all others provide full integration batch-real time support.

Package / enterprise applications: How many packages / enterprise applications can the tool read meta data from with one click of the mouse (for example SAP, Siebel, Peoplesoft, JD Edwards, Baan). Talend can read 9 which the maximum followed by IBM Information Server and Oracle Data Integrator which can read 8, followed closely by Informatica PowerCenter which can read 7 and poorest of all is the CloverETL and which can read 0 packages applications.

2. Tool-wise description

IBM Information Server

It provides a great flexibility and is directed towards the market with a vision in mind with common metadata platform. The Information Server provides high level of satisfaction from clients and a variety of initiatives. Though it is easy to use but it becomes very heavy because of the data involved is in GBs and the version 8.x requires a lot of processing power.

Informatica PowerCenter

Informatics PowerCenter offers a so solid technology, straightforward learning curve, ability to address real-time data integration schemes and is highly specialized in ETL and Data Integration. It has a consistent track record with most substantial size and resources on the market of data integration tools vendors.

Talend

Talend is an open-source data integration tool but not a full BI suite. It uses a code-generating approach. Uses a GUI. It has data quality features: from its own GUI, writing more customised SQL queries and Java.

Microsoft SQL Server Integration Services

SQL Server Integration Services (SSIS) provides ease and speed of implementation with standardized data integration, real-time, message-based capabilities which are relatively low cost and provide an excellent support and distribution model. However, it does not support non-Windows environments.

Oracle Data Integrator

There is no doubt as to why it is being regarded as one of the leaders in the ETL markets; it's because of its tight connection to all Oracle data warehousing applications and the tendency to integrate all tools into one application and one environment.

SAS Data Integrator

SAS Data Integrator provides great support and most of all very powerful data integration tool with lots of multi-management features. It is great support for the business-class companies as well for those medium and minor ones. It can work on many operating systems and gather data through number of sources – very flexible.

Pentaho(Kettle)

Pentaho is a commercial open-source BI suite that has a product called Kettle for data integration. It uses an innovative meta-driven approach and has a strong and very easy-to-use GUI. It has a stand-alone java engine that processes the jobs and tasks for moving data between many different databases and files.

CloverETL

CloverETL provides data integration**,** Workflow automation through job flows. It can transform the data at ease. It is a visual tool that replaces everyday scripting and provides full control of the data flows and processes.

Conclusion

The research has given us a conclusion that without ETL products\ tools the analytical reports are not possible and ETL in itself is the basic foundation for any Business Intelligence (BI) tool used by the organization. Since, ETL involves three-part work that of extracting data from the source, transforming the data into a unified format and lastly loading this unified data into a data warehouse. In our comparison of ETL software tools we find that Oracle

Data Integrator and IB Information Server are the ones which satisfy needs of large enterprises and other tools mentioned here have their own aspects to be implemented. Most of the upcoming tools are slow and have very few functions to satisfy the needs of the larger organizations but, they can't be ruled out for semi-large or small enterprises where the only compromise would be on speed and still one would get lots of other functionalities. Some of the aspects we have left out and have not included in our criteria one being the price. Since , every organization can only decide upon buying the product by evaluating the functionalities and the benefits expected to be reaped from it .Also not all products guarantee to provide all functionalities and its always a compromise as to what should work for one may not work for the other firm, so, although we have given our review and evaluation on these tools yet , the enterprise can decide by attaching a weight to the criteria and external variables that the enterprise feels are important and then decide which tool to go with.

Limitations and Future Scope

The research has been done by analysing reports, articles, journals and gathering information from the vendor websites. The research was not conducted by testing the products with real world data. We were not able to evaluate the tools in a "hands on" manner and so the criterion followed to evaluate the tools was based on the reports that we had gone through. We have based our criteria of evaluation of other market researchers because we don't have a strong foundation in BI as yet and moreover the cost involved to do so goes into thousands of dollars. Also, ETL is only a part of the BI suite offered by the vendors. Price of the product was also not included because most of the organizations have not revealed its pricing to remain competitive and many solutions are tailored to fit an organization's specific needs via quotes. Many other criterions has been left out because of space constraints.

Future scope involves giving weight to the criterions and then measuring and analysing the features of all the ETL products rather than just the leaders.

References

[1] Lombard, H., Sweiger, M., Madsen, M., & Jimmy Langston, J. (2002). Clickstream Data Warehousing. John Wiley & Sons.

[2] Madsen, M. (2004, October). Criteria for ETL Product Selection. Retrieved 11 06, 2013, from InfoManagement Direct: http://www.information-management.com/infodirect/20041001/1011217-1.html?pg=1

[3] Larson, B. (2008). Delivering Business Intelligence with Microsoft SQL Server. New York:McGraw-Hill Osborne Media.

[4] Levin Jonathan (2008) Open Source ETL tools vs Commerical ETL tools retrieved on 10-6-2013 from http://www.jonathanlevin.co.uk/2008/03/open-source-etl-tools-vs-commerical-etl.html

[5] Friedman, T., Beyer, M. A., & Bitterer, A. (2008). Magic quadrant for data integration tools. Gartner RAS Core Research Note G, 207435

[6] Friedman, T., Beyer, M. A., & Bitterer, A. (2008). Magic quadrant for data integration tools. Gartner RAS Core Research Note G, 207435

[7] Gartner (2010) Friedman , Mark A. Beyer, Eric Thoo, Magic Quadrant for Data Integration Tools retrieved on 12 06 2013 from http://www.virtualtechtour.com/assets/GARTNER_DI_MQ_2010_magic_quadrant_for_data_inte_207435.pdf.

[8] Microsoft (2012) SQL Server Integration Services retrieved on 12 06 2013 from

http://msdn.microsoft.com/en-us/library/ms141026.aspx
CloverETL
http://www.cloveretl.com/products
retrieved on 11 06 2013.

[9] IBM Information Server IBM InfoSphere Information Fast Track Your Information Server for Linux, Unix and Windows retrieved on 11 06 2013 http://www-01.ibm.com/software/in/data/integration/info_server/

[10] Oracle Data Integrator retrieved on 12 06 2013 from http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html

[11] Oracle Data Integrator retrieved on 12 06 2013 from http://www.oracle.com/us/products/middleware/data-integration/dataservice-integrator-ds-168223.pdf

[12] Passionned Group (2013) The Data Integration & ETL Product Survey 2013

[13] Pentaho Corporation (2013) Pentaho Data Integration retrieved on 12 06 2013 from http://www.pentaho.com/press-room/releases/20100210_pentaho_and_swissport_cuts_costs_of_flying/

[14] Zode, M. The Evolution of ETL. Retrieved on 6/06/2013from http://hosteddocs.ittoolbox.com/mz071807b.pdf

[15] Pentaho Corporation (2013) Pentaho Data Integration retrieved on 12 06 2013 from http://www.pentaho.com/explore/pentaho-data-integration/

[16] Pentaho Pentaho Data Integration (Kettle) retrieved on 12 06 2013 from http://kettle.pentaho.com/

[17] SAS Data Integration Studio SAS Products retrieved on 12 06 2013 from http://support.sas.com/software/products/etls/

[18] Talend Open Studio Talend Open Studio retrieved on 12 06 2013 from http://www.talend.com/products/talend-open-studio