ETL Tool Evaluation – A Criteria Framework

Nils Schmidt
University of Texas-Pan American
1201 W. University Drive, Edinburg, TX
(956) 381-UTPA
nilsschmidt@gmx.de

Mario Rosa
University of Texas-Pan American
1201 W. University Drive, Edinburg, TX
(956) 381-UTPA
marioarosa1@gmail.com

Rick Garcia
University of Texas-Pan American
1201 W. University Drive, Edinburg, TX
(956) 381-UTPA
rjgarcia5@utpa.edu

Efrain Molina
University of Texas-Pan American
1201 W. University Drive, Edinburg, TX
(956) 381-UTPA
emoli81@hotmail.com

Ricardo Reyna
University of Texas-Pan American
1201 W. University Drive, Edinburg, TX
(956) 381-UTPA
rreyna11@utpa.edu

John Gonzalez
University of Texas-Pan American
1201 W. University Drive, Edinburg, TX
(956) 381-UTPA
jegonzalez11@gmail.com

Abstract

In this paper Business Intelligence students developed a criteria framework which might help to evaluate today's ETL tools. After setting up the framework they used the framework to compare and evaluate Microsoft SQL Integration Service, an open source leader known as Pentaho Integration Services, and ETI High Performance Corporate Data Integration. The results are shown in a straightforward chart approach. Furthermore, in their finding the students revealed the differences between these tools based on their criteria framework and they made suggestions who should implement which ETL program.

Introduction

Modern businesses face many new challenges because of the vast amount of connectivity brought upon by advances in computer technology. Even small companies can generate large amounts of customer data, sales information, or inventory details on a daily basis. In a world of terabytes upon terabytes of data there is a very real need to make sense of the raw data in order to make informed decisions about their business. Making sense of this raw data can be a monumental task without the right tools. There is a need to remain competitive in the business environment so many professionals turn to business intelligence tools. Business Intelligence is the delivery of accurate, useful information to the necessary personnel in order to facilitate effective decision making (Larson, 2008). They soon realize that developing business intelligence solutions can be very expensive. Business intelligence tools can cost thousands of dollars, but if used correctly they can help generate future revenue for many years to come. There are many different tools for business intelligence such as Data Mining, OLAP, and Data Warehousing, but there is a crucial component that is at the foundation of Business Intelligence. The crucial component is ETL (Extraction, Transformation, and Loading) tools. ETL tools are important for Business Intelligence because your results are only as accurate as the input you feed it. ETL is the one of the most important bases you must cover when evaluating any Business Intelligence tool. The process of using ETL is one the most time consuming portions when trying to develop your business intelligence, but choosing the right ETL tool is a fundamental step in achieving your strategic goals. There are hundreds of software solutions out there today all promising the fastest and most accurate results. Some software packages can cost thousands while others may cost a small fortune. How can one ever be sure they are making the right decisions when investing their company's finances? The truth of the matter is that every business environment is different, but there are general guidelines that can be used to evaluate any set of ETL software.

Objective

The purpose of our research paper is to provide a framework of criteria which might help to evaluate ETL tools. Furthermore we are evaluating three ETL tools based on these criteria which we think are suitable for an ETL tool to have. In this comparison we are aiming to compare current market leaders as well as provide awareness for open source software as an alternative ETL solution. We will be putting three different programs against each other to determine the best overall choice according to a specific set of criteria. We will be evaluating Microsoft Integration Services (MSIS) because of Microsoft's status as a market leader in technology. We will also be evaluating Pentaho Integrations Service (PIS) as the open source solution which is beginning to make a name for itself because of its large toolset (Pentaho Corporation, 2010). The final ETL tool that we will be evaluating is ETI High Performance

Corporate Data Integration (ETI). We chose ETI because of high search rankings when beginning our research; we gathered that it was a fairly strong competitor in the ETL marketplace.

ETL Processes

ETL is a part of Business Intelligence (BI) and usually the starting point for BI. To better understand ETL, let us first understand what its abbreviation means: ETL is better know as Extract, Transform and Load. It is a process which is used to extract data from different systems and transform that data into a structure that can be used to create reports and analyze data. The data then can be loaded into a database from this point on. Today's ETL is more robust and has more to offer end users. For example, it offers data profiling and data quality control, in addition to data cleansing and on-demand data integration service just to name a few modern enhancements. There are several individual steps that make up the ETL process. The first step in the ETL process is Extracting data from the source. In this step, the selected software extracts data from different sources. These sources can be data within the software (internal) or it can be data from outside the software (external). In the second step, ETL transforms the data once the data has been moved into the staging area, where it becomes one platform and one database. At this point, the user can begin to join and unionize tables and filter or sort the specific data to be used. In the third step, ETL data is loaded into the Business Intelligence data warehouse, which is its final destination. This is the location of the loaded data and is typically used to create fact and dimension tables (ETL).

Research Method

To conduct our research we needed to create a framework of criteria which would allow us to compare the ETL tools against each other. Our criteria are found in the functionalities, requirements and characteristics of each ETL tool. To reveal the importance in functionalities, requirements, and characteristics of each ETL tool and to be able to compare them against each other we conducted research in different articles, journals and books. After analyzing these journals we successfully revealed the criteria and its categories which will allow us to compare and rate ETL program against each other. All these criteria were adopted and categorized in our criteria framework. Originally we had a five categories, but we removed the category named "Connectivity" but because the ETL tools did not differ in terms of connectivity. The categories of the framework are shown in Table 1.

Table 1

Categories
Price
Functionality
Ease of Use
Architecture

Price

We created the category "Price" during the framework building process because the Oracle Expert Ian Abramson weighted costs an important factor when rating ETL tools (Abramson). Furthermore, we included in this category all criteria which are related to costs such as the criteria "License Cost", "OS Costs", "Support Costs" and "Hardware cost".

The criterion "License Cost" was implemented to compare the ETL tools based on costs which occur by buying one license. Furthermore, the criterion "OS Costs" was introduced to compare the tools based on operating system costs. Next, the criterion “Support Costs” was setup to gain information about the difference in costs for additional service support. Finally, the criterion “Hardware cost” was implemented to reveal the difference in cost to buy the hardware which is needed to run the program. All mentioned price criteria are compared in dollar values.

Functionality

The category “Functionality” was setup to compare and evaluate the functionalities of our ETL tools. According to the author Mark Madsen from Information Management Direct the criteria in this category basically determines if you are able to process your data (Madsen, 2008). After determine the functionality category as an important category we started to determine each criteria. Research was basically based on journals from the author Mark Madsen who provides us with many important ETL criteria. After conducting the research we decided to implement the following criteria “Basic processing support”, “Performance”, “Transformations”, “Cleansing”, “on-demand support”, “Secure Packages”, “ETL reporting”, “Scheduling”, “Metadata”, “Rollback”, “Connectivity”, “Calculation”, “Data Warehouse support”, “Aggregation”, and “Reorganization”. Possible values for all criteria in this category can only be “yes” or “no” indicating that the criteria is supported or not is not supported.

The criterion “Basic processing support” was adopted form the Clickstream Data Warehouse book where an ETL tool should be capable of importing the data and being capable of processing this data sequentially (Lombard, Sweiger, Madsen, & Jimmy Langston, 2002). Furthermore, this criterion includes robust SQL support (Lombard, Sweiger, Madsen, & Jimmy Langston, 2002). The criteria “Performance” should allow us the reveal the performance of the tool. Tools matching this criterion must be able to handle large amount of data and must be faster than other tools. The criterion “Transformations” was adopted because transformation appears to be the most important part of the ETL process. If this criterion is matches the tool is able to handle prebuilt transformations but this criterion here does not reveal any information about the kind of transformation performed. The different kinds of transformation are handled in the criteria “Cleansing”, “Calculation”, and “Aggregation”. Furthermore, according to Mark Madsen “Cleansing” should include a cleanup and a way to synchronize data automatically (Madsen, 2008). If the criterion “Calculation” is matched, the tool will be able to calculate all basic mathematical and statistical functions. According to the author Mark Madsen all these mentioned transformation criteria are important and the ETL system should be able to handle them (Madsen, 2008). “On-demand support” is met when the system is able to deliver data ad-hoc and in real time. Usually “on-demand support” is given in very powerful systems. The criterion “Secure Packages” should express the ability of an ETL tool having certain features to secure the data (Abramson). If the “ETL reporting” criterion is matched the tool is able to provide a report on every process done. “Scheduling” provides the tool with the ability to plan tasks and run the process automatically. The criterion “Metadata” provides the ETL tool with the ability to integrate and handle data marts across business units (Mimno, Myers & Holum, 2001). If “Rollback” is matched the tool is able to undo the last transaction. Being able of satisfying the criterion “Connectivity” requires the software of being able to handle multiple storage formats such as flat file and different kind f databases (Lombard, Sweiger, Madsen, & Jimmy Langston, 2002). Furthermore, we decided that the criterion is matched if the following database support is given: ADO, ADO.NET, CACHE, EXCEL, FILE, FLATFILE, FTP, HTTP, MSMQ, MSOLAP100, MULTIFILE, MULTIFLATFILE, OLEDB, ODBC, SMOServer, SMTP, SQLMOBILE, WMI, ORACLE, SAPBI, TERADATA, and COBOL LEGACY. The setup criterion “Data Warehouse support” reflects the possibility of the software to load the data directly in the data warehouse. Finally, if the criterion “Reorganization” is matched if the ETL software is able to organize the data after the data is loaded and transformed into the database (Madsen, 2008).

Ease of Use

To determine information about the usability of the ETL Tools we created the category “Ease of Use”. According to Mark Madison it is difficult to establish criteria here because every user has different preferences how a program should work (Madsen, 2008). After research we established the following criteria for comparison in this category “Completeness of the GUI”, “Custom Code”, “Integrated Toolset”, “Debugging support”, and “Source Control”. Possible values for all criteria in this category can only be “yes” or “no” indicating that the criteria is supported or not is not supported.

The criterion “Completeness of the GUI” should reflex in our evaluation that a good visual interface is given. Furthermore, according to Mark Madsen the better the visual interface the easier the extracts are to write (Madsen, 2008). The next criterion “Custom Code” will allow the user to enter source code or to highly customize the process. The criterion “Integrated Toolset” is matched if the user has the possibility to purchase addition tools or add-ons for the product when the ETL tools are not integrated in one program (Madsen, 2008). If the criterion “Debugging support” is matched the user can set breakpoints to easier analyses errors (Madsen, 2008). The criterion “Source Control” should make it easy for the user to easily select and integrate different sources.

Architecture

The criteria category “Architecture” was setup to gather information about the hardware and operating system (OS) support by the software in terms of platform, backup and performance. According to the book from the Clickstream Data Warehouse each ETL product supports different architecture (Lombard, Sweiger, Madsen, & Jimmy Langston, 2002). In this category we implemented the criteria “Platform independent”, “Expandable”,” Recovery”, and “Backup”. Possible values for all criteria in this category can only be “yes” or “no” indicating that the criteria is supported or not is not supported.

The book about the Clickstream Data Warehouse states that it is important to know what platform is supported. Furthermore, it might happen that the ETL tool supports only a low end platform such as Windows NT (Lombard, Sweiger, Madsen, & Jimmy Langston, 2002). Therefore we implemented the criterion “Platform independent” to show if the ETL tool can be installed on multiple platforms. The criterion “Expandable” was chosen to see if it is possible to easily modify and upgrade the hardware architecture. If the criterion “Recovery” or “Backup” is matched, the ETL tool is able to perform a Recovery or a Backup on the architecture.

After setting up the criterion framework consistent of our categories and the criteria we decided not weight the categories and its criteria. We decided against a weight scale because you might need just little functionality or you might value a lower price or better appearance more than many functions. It depends on you requirements or preferences for the program which program might be best. Therefore, just the criteria and the match with the program are shown in our evaluation in the next paragraphs.

Tool Description

Tool 1 - Microsoft Integration Service

Microsoft Integration Service seems to be one of the most robust systems and it is evident in the pricing of the software. It is meant to handle large enterprises and comes with some hefty support costs as well as operating systems because it requires a Microsoft Server operating system.

Microsoft Integration Service (MSIS) appears to be a good equipped ETL tool according to its functionalities. Furthermore, MSIS supports nearly all functional criteria's from our criteria framework except the "Cleansing" criteria. According the MSIS website it appears that MSIS supports next to the criteria's "Basic processing support" also advanced ETL criteria's from our framework such as "Transformation", "Real-Time support", "Secure Packages", and "Scheduling". Furthermore, it appears that MSIS supports based on our criteria framework the criteria's "on-demand support", and "ETL reporting". Microsoft says about MSIS that MSIS is the fastest ETL Tool on the market. For this reason MSIS satisfied the criteria "Performance" (Microsoft, 2009). The "Metadata" criterion for MSIS appears to be one of the strongest points for the software. MSIS allows for many connectivity options to many databases as described in the criteria description. It matches from our criteria advanced calculations when combine aggregate data and allows the tool to recover from a catastrophe with the rollback.

Microsoft Integration Service (MSIS) appears to be fundamentally easy to use according to its ease of usefulness. Furthermore, MSIS supports nearly all criteria's from our criteria framework except the generation of custom code. According to the MSIS website it appears that MSIS supports most of the criteria listed under our criteria framework; such as the "Completeness of the GUI," "Integrated Toolsets," "Debugging Support," and "Source Control." With the combination of the GUI, Integrated toolsets, debugging support, and source control make MSIS one of the most easy to use ETL tool on the market.

Microsoft Integration Service (MSIS) appears to have some limitations according to its architecture criteria (Microsoft, 2009). Furthermore, this limits also the ability to easily upgrade the system because a change in hardware needs most of the time also a change in license. MSIS supports only platforms which are based on Microsoft Operating Systems (Microsoft, 2009). The ETL tool appears that they offer the backup and recovery feature through the "MSDB Database" (Microsoft, 2009).

Tool 2 - Pentaho Integration Service

Pentaho Integration Service is very inexpensive for any organization to utilize because of its open source nature. There is a commercial version available, but we were unable to find a price for that. Support costs are usually avoided if you utilize the online community.

Pentaho Integration Service (PIS) appears to support just new advanced features next to the criteria "Basic processing support" from our criteria framework. Furthermore, PIS appears to have some limitations such as limitation to the file size / performance in its free open source edition which might not exist in the commercial Edition. The "Calculation" portion of the software is limited to basic calculations like simple arithmetic and simple data joins. This was one of the other main functions that can clearly hurt this software tool. For being a free program, PIS supports still more advanced criteria's such as "Transformation", "Cleansing", and "Scheduling" (Pentaho Corporation, 2010).

firm. We have outlined our criteria for evaluating potential ETL tools, but ultimately, the decision will have to be weighed by your organization and what variables you consider to be deciding factors.

Limitations

The research was conducted by analyzing the vendor’s website and not by testing and benchmarking the programs with real world data. We were not able to install the software on a specific set of hardware or validate any sort of performance claims made by the manufacturers. We were not able to evaluate the tools in a “hands on” manner and had to make decisions solely on the criteria we deemed vital when trying to select the right ETL tool for a general organization. We based our criteria off of other market researchers because we do not have a strong foundation in business intelligence as of yet. Also, when doing research on an ETL tools we found out that ETL was only a small portion of a typical BI suite offered by the vendors. Getting exact pricing was difficult because many solutions are tailored to fit an organization’s specific needs via quotes. It seems that vendors do not openly list their prices in an effort to remain competitive. Testing these programs would have required us to purchase the software at a cost of more than $30,000. All in all, we made the best criteria possible by analyzing other reports and gathering all the necessary information via the vendor sources.

Works Cited

Abramson, I. (n.d.). Rating ETL tools. Retrieved 10 01, 2010, from SearchOracle.com: http://searchoracle.techtarget.com/answer/Rating-ETL-tools

ETI. (2008). PRODUCTS. Retrieved October 08, 2010, from ETI High Performance Data Integration: http://www.eti.com/products/index.html

ETL Process Image Retrieved October 07, 2010,http://www.computerworld.com/s/article/89534/QuickStudy_ETL?taxonomyId=9&pageNumber

ETL Tool Survey 2010 what is ETL? Retrieved October 07, 2010, http://www.etltool.com/what-is-etl.htm

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

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

Madsen, M. (2008, October). Criteria for ETL Product Selection. Retrieved 10 02, 2010, from InfoManagement Direct: http://www.information-management.com/infodirect/20041001/1011217-1.html?pg=1

Microsoft. (2009). SQL Server 2008: Integration Service. Retrieved October 08, 2010, from Integration Service: http://www.microsoft.com/sqlserver/2008/en/us/integration.aspx

Mimno, Myers & Holum. (2001). Data Warehousing, Corporate Portal & e-Business Intelligence Applications. Retrieved October 5, 2010, from Mimno, Myers & Holum: http://www.mimno.com/extracting-data.html

Pentaho Corporation. (2010). Pentaho Data Integration. Retrieved 10 08, 2010, from Pentaho: http://www.pentaho.com/news/releases/20100210_pentaho_and_swissport_cuts_costs_of_flying.php

Pentaho Corporation. (2010). Pentaho Data Integration. Retrieved 10 08, 2010, from Pentaho: http://www.pentaho.com/products/data_integration/