ETL tools for Data Warehousing: An empirical study of Open Source Talend Studio versus Microsoft SSIS

Ranjith Katragadda
Unitech Institute of Technology
Auckland, New Zealand

Sreenivas Sremath Tirumala
Unitech Institute of Technology
Auckland, New Zealand

David Nandigam
Unitech Institute of Technology
Auckland, New Zealand

***Abstract*—Relational databases are bound to follow various database integrity rules and constraints that makes the reporting a time consuming process. Data Warehousing has evolved out of the desperate need for easy access to structured storage of quality data that can be used for effective decision making. Data are turned into knowledge and knowledge into plans which are instrumental in profitable business decision making. To serve this purpose, data need to be extracted from various sources, transformed and loaded into the data warehouse which constitute the process of ETL (Extract, Transform and Load). ETL process can be accomplished using various tools both open source and proprietary. In this paper, we provide an empirical study of two ETL tools, an open source Talend Studio and Microsoft SSIS. In spite of the dominance among a vast majority of computer software solutions, open source technologies, as the comparative analysis that this study has undertaken, concludes that open sources tools are yet to evolve in order to be sustainable.**

Index terms: ETL, Talend studio, SSIS, Data Warehousing, open source ETL

I. Introduction

The term ”Data Warehouse” was first coined by Bill Inmon in 1990 [1]. According to him, Data warehouse is subject Oriented, Integrated, Time-Variant and non-volatile collection of data that supports decision making process in an organization.The operational database undergoes several day to day transactions which makes the process of data analysis more and more complex and time consuming. The Data Warehouses provide us generalized and consolidated data in multidimensional view and makes it possible to use Online Analytical Processing (OLAP) tools for interactive and effective analysis of data in multidimensional space.

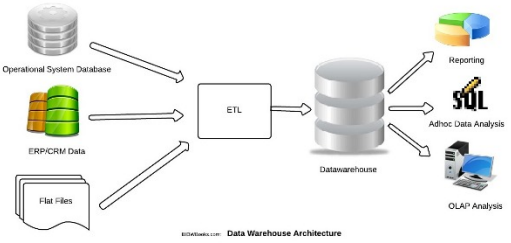

Fig. 1. Data Warehouse Design

The concept of data warehousing has evolved out of the need for easy access to a structured storage of quality data that can be used for effective decision making. The Data Warehouse is a database, isolated from the organization's operational database. Data warehouse contains consolidated historical data which help the organization to understand the various business scenarios by data analysis. Data warehouse helps the executives to organize, understand and use their data for strategic decision making. The data warehouse is an important supplier of information to the business, so it is very important that we model both its physical and logical designs. The physical design determines the performance and functionality of the data warehouse, and the logical design is the view that we present to developers and users to capture business requirements. Efficient transforming and loading of the data into the data warehouse is equally important and is discussed in the next section.

II. The process of ETL

In todays businesses, decision-making processes and daily operations often depend on data that is stored in a variety of data storage systems, formats, and locations. In order to turn this data into useful business information, the data typically needs to be combined, sanitized, standardized, and summarized. For instance, information may need to be converted to a different data type or heterogeneous database servers may store the necessary data using different schemas. Dissimilarities like these must be resolved before the data can be successfully loaded to a target system. After the design and development of

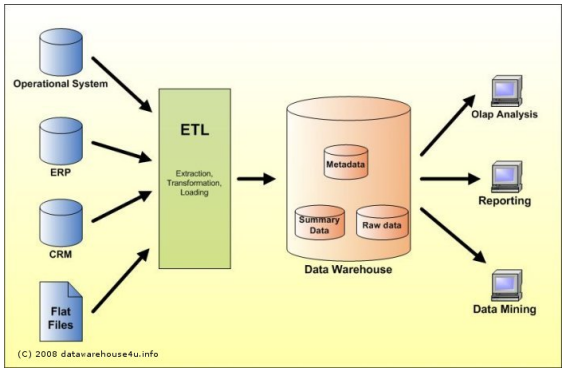

Fig. 2. ETL Process

data warehouse in accordance with the business requirements, the process of consolidating the data into the data warehouse from various sources is to be addressed. Extract Transform Load (ETL) processes are critical in the success of the Data Warehousing projects. The process of extracting data from one system (extract), transforming it in accordance with the design of the data warehouse(transform) and loading it into

data warehouse system (load) constitute ETL. In other words, ETL is the process of extracting data from various data sources, transforms it as per the requirements of the destination data warehouse and successfully loading it into the destination data warehouse (database). In the transformation process data is actually standardized to make it compatible with the destination database along with data cleansing (cleaning) operations. [2]. The process of ETL is detailed in the following sub sections.

A. *Extract*

The first step of ETL process is extraction of data from various data sources that contain the information that need to be transferred to the data warehouse. Some of these sources might be relational, some of them might be just single flat files without any data integrity rules. In the extract process, data is extracted from the source system and is made accessible for further processing. The main objective of the extract step is to extract the required data from the source systems utilizing least possible little resources. Further, the extract process should be designed in such a way that it does not affect the source system in terms of performance, response time or any kind of locking. Data extraction can be performed in several ways such as update notifications, incremental extract, full extract etc. The frequency i.e, number of times an extract to be performed or the time interval between each extract is very crucial in the case of incremental of full extracts as the volumes of the data can be in tens of gigabytes.

Fig. 3. ETL - Extract

B. *Transform*

The most complex part of ETL process is the transformation phase. At this point, all the required data is exported from the possible sources but there is a great chance that data might still look different to the destination schema of the data warehouse. At times data itself need to be formatted to conform to the data types and other constraints of the data warehouse. In some cases data need to be represented differently in order to make it more fruitful.

In the transform process, a set of rules are applied to transform the data from the source to meet the requirement of the target which is a data warehouse. This includes converting any measured data to the same dimension (i.e. conformed dimension as per the requirements of the data warehouse) using the same units so that they can later be joined. The transformation process also involves joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

Fig. 4. ETL - Transform

Another important aspect of the transformation process is data clensing. Cleaning is an important process as it ensures the good quality of the data in the data warehouse. Cleaning should be performed on basic data unification rules, such as

1) Transforming the various identifies into an unique representation. For example sex categories like Male, Female, Unknown, or M, F,null or Man,Woman,Not Available,Not applicable are translated to standard Male, Female, Unknown.
2) Convert null values into standardized Not Available or Not Provided value
3) Convert phone numbers, ZIP codes into a standardized form
4) Validate address fields against each other (State,Country, City,State, City,ZIP code, City,Street).

C. *Load*

Once the data is extracted and transformed according to the requirements of the target data warehouse , the data is assumed to be ready for loading. However, several aspects like how the data to be loaded, the impact and implication of loading process as well as handling such implications are to considered before loading the data into the data warehouse. The process of loading may impact thre processing speed of the server both for loading as well as analysis. can occur. It is also crucial to avoid database crippling while loading the data.

ETL processes can be performed using almost any programming language. Building such program from scratch can be complex which makes it is necessary to use ETL tools. ETL process can be carried out manually or by using a tool of automation. When the ETL process is carried out with a automation tool, data mappings are fed into the tool of automation and the code that performs the mappings is created. Generally mappings are done manually when there are only few procedures to be written. However, it is more efficient to use an automation tool for ETL process. The widely used open source ETL tool is Talend Open Studio and proprietary tool is SQL Server Integration Services.

III. Talend Open Studio

Talend Open Studio is the first open source data integration software released in 2006 after an intense research for over three year [3].It is based Eclipse RCP that primarily supports ETL-oriented implementations and is provided for on-premises deployment and Software as a Service (SaaS) delivery model. Talend Open Studio is used for integrating operational systems as well as an ETL tool for Data Warehousing, Business Intelligence and data migration. The company Talend shatters the traditional proprietary model by supplying open, innovative and powerful software solutions with the flexibility to meet the data integration needs of all types of organizations. Talend Open Studio is the most innovative and powerful open source data integration solution on the market today. Talend Open Studio for data integration helps you get your data to the right place, in the right form, at the right time. As the leading open source ETL solution for data warehousing and business intelligence, Talend Open Studio is

1) synchronize data across heterogeneous sources and targets.
2) Easy to use - Start productive work right away with an intuitive interface rich in modelling tools,
3) Job-building components, and more than 450 data connectors, including the Cloud.
4) User friendly and comprehensive IDE
5) Can generate Java Code from the developed packages
6) The Java generated code can be modified to achieve greater control and flexibility.
7) Talend Open Studio for Data Integration is free to download and use.
8) Successfully applied to Real world application [4]

Talend Studio is commercially implemented in Buffalo Studios a social and mobile online casino gaming company. Since, there are quite a few online and offline developers working for Buffalo Studios, their main challenge was to satisfy their developers who were not use to restricted licensing. Further, the cost for implementing a proprietary software for such an environment is very high. As a part of business process, Buffalo Studios receives data from multiple sources and its has became mandatory to use ETL process for data analysis and data-driven decision making. One of the biggest bottle necks for that was the ETL layer. Every time a new game was launched, data ware house instrumentation was required around it to measure the success and usage of features. It was extremely unmanageable in the ETL pipe line as that required going and modifying ETL code and modifying database schemas. More often the functionality did not get instrumented and data would make its way into the raw logs. Buffalo Studios realized the need to make changes to their ETL system. This made them turn to Talend and is been successfully running since then.

Talend provides a very cost-effective way for ETL process, data quality and master data management initiative without the need for any significant investment. This kind of approach quite suitable for the current iterative and incremental project environment further reducing the business risk by allowing small but valuable piece of business functionality delivered in continues shorter time frames. As Talend is committed to open source, the effort the company is undertaking to sustain their support for open source products is appreciable. For instance, Talends data integration tool transparently supports the use of Hadoop clusters and of the Hive data warehousing environment.

However there are some limitations to the community edition of Talend Open Studio. It is developed as a product for individual use only and so it is not possible to have more than one user (not just one user at a time but just one user per system). This causes a practical implementation problem as it might be needed to have multiple users using the same computer at different times or when user loses password. Furtherm the free version doesn't support automation of tasks like scheduling, routing data etc. Another major drawback is lack of any commercial support.

IV. SQL Server Integration Services (SSIS)

SQL Server Integration Services (SSIS) is one the Business Intelligence tools (BI) developed by Microsoft corporation to ease and automate the ETL process. the [5]. While ETL processing is common in data warehousing applications, SSIS is by no means limited to just data warehousing. For example we can automate the SQL Server Maintenance Plan by creating an SSIS package. SSIS provides a standardize mechanism to accomplish the needs of both individual researchers and top management as well as commercial consultants. The first version of SSIS was released with SQL Server 2005 replacing the Data Transformation Services (DTS) which was available with SQL Server 7.0 and SQL Server 2000. As show in Fig IV-A SSIS holds the record for fastest ETL process ever

A commercial implementation of SSIS is done by Innovapost [6]. Innovapost needed to collect data every half hour from multiple Canada Post mail-sorting machines. Canada Post, which serves 15 million addresses each day wanted to implement data ware house technology to optimize processing and restructure routes by analysing the reports generated from the data available from the data ware house, thus improving both the quality and cost-effectiveness of its services. But in achieve this, Canada Post needs data to be reliable as well as quickly available. To meet these requirements, Innovapost was asked by Canada Post to develop a mechanism to extract, transform, and load data every half hour, 24 hours a day and seven days a week and produce reports.

Innovapost built an application based on Microsoft SQL Server 2008 Integration Services. The application turns raw files from multiple sources into clean and dependable data, quickly and reliably. Reports on the data are often available to sorting facilities just 35 minutes later, helping Canada Post make business decisions faster. Innovapost managed to provide Canada post with dependable, clean, timely and reliable data which aided the company in decision making process to improve its business and cutting in costs.

A. Benefits of using Microsoft SSIS

Microsoft SQL Server Integration Services is very low in cost compared to the famous Informatica Power Center and almost offers everything you need to build your ETL solution. The following are some of the advantages of using Microsoft SSIS for ETL process.

1) SSIS enables to build ETL solutions with very minimum background knowledge
2) Supports the ETL tasks as well as provides the ability to write custom code using Script task
3) Uses parallel computational logic thus supporting efficient mechanism to handle large amounts of data transfer
4) Very easy to install and configure
5) Offers broad documentation and support, provides best practices to data warehouse
6) Provides relatively low cost, excellent support and distribution model
7) Integration with other Microsoft software

As show in Fig IV-A SSIS holds the record for fastest ETL process ever

Fig. 5. SSIS World Record

The main limitation of SSIS is the lack of ability to support non windows operating systems. SSIS is more suitable for enterprise level ETL solutions and may not be cost efficient for small businesses.

V. Discussion

Talend is a relatively new company .Unlike most of the companies in software industry, it delivers many open source products which users can freely modify so that they could suit their needs in the best possible way. The fact Talend pays a lot of attention to open source market, doesn't mean it works only with less important customers. In fact, there are worldwide known names on the list of its customers.

On the other hand, SSIS is well known for its multiple capabilities. Its main area of focus is bulk/batch-oriented data delivery. As a part of an SQL Server family, SSIS is an answer to actual market trends and needs. Implementation of SSIS is fast and easy in Microsoft Environment, that's what customers point out as one of the more significant advantages of the product as 90% of computers run Microsoft Products. Further, its integration with most commonly used user friendly Microsoft Office makes it the market leader.

SSIS has multiple capabilities but limited functionalities in integration with other environments whereas Talend has advanced integration capabilities with other environments. Compared with other commercial ETL tools, SSIS is very cost effective with a wide range of data management capabilities. Talend is quite suitable for small implementations but for large implementation SSIS is way ahead. SSIS is from Microsoft, a well established company with both expertise and experience in data handling where as Talend is relatively new to market.

The popularity of Microsoft is another added advantage. Most of the customers still consider it far before other vendors of software and choose its offer even if the others are the same good or almost the same good. This results in, is great support - not only the one provided by Microsoft itself, but also by the third parties. Additionally, there is a lot of documentation prepared by other users which is easily accessible.

Even though the Microsoft SSIS appears to be a real combine of different capabilities and functions, its functionality might be restricted to some users. The biggest problem seems to be lack of support for different integration styles which isn't exactly what today's customers expect from products offered by vendors like Microsoft. Another limitation of SQL Server Integration Services is its strict alignment to Microsoft Windows environment. It is not possible to run the application under any other operating system which is a significant issue for the ones who realize the restriction of Windows, especially when it comes to efficiency in comparison to other operating systems. Eventually, Microsoft SQL Server Integration Services is thought to work with other products, but the integration between two different applications is not trouble-free. It is more time consuming and needs more time and effort which results in reduced efforts which should be focused on other aspects.

A. Support and Documentation

Support and Documentation	TOI	SSIS
Community-based: forums, Bug tracker...	✓	✓
Enterprise grade support with SLAs		✓
User Guide, Reference Guide	✓	✓
Mission Critical Support (optional)		✓

Fig. 6. Support and Documentation

Talend open Studio do not offer Service Level Agreement(SLA) and Mission critical support. On the other hand SQL Server Integration Services offers both SLAs and Mission critical support. When a situation arises, you are not by yourself when using SSIS. There is much more support and better user guide and reference guide documentation for SSIS (V-A.

As shown in the Fig V-A, Talend do not offer Parallelization, Data Viewer, Wizards, Dynamic Schema, Impact Analysis and Data Lineage implementation.

Latest development in ETL tool is parallel processing implementation. This has enabled a number of methods to improve overall ETL process performance while dealing with large volumes of data. Unlike SSIS, Talend do not offer parallelization which is a latest development in ETL process which emphasizes the fact that TOI fails to keep up to the market trends.

Not providing Dynamic schema support implementation is one of the limitations of Talend worth mentioning in

Implementation	TOI	SSIS
Job Designer	✓	✓
Components	✓	✓
ETL support	✓	✓
ELT support	✓	✓
Versioning	✓	✓
Parallelization		✓
Data Viewer		✓
Wizards		✓
Dynamic Schema		✓
Impact Analysis		✓
Data Lineage		✓

Fig. 7. Implementation

comparison with SSIS. Captures information from source to destination along with the various processes and rules involved and shows how the data is used. This knowledge about what data is available, quality, correctness and completeness leads to a mature data governance process. There is a limitation in data governance process if Talend in comparison to SSIS as it do not implement data lineage in ETL process.

Support and Documentation	TOI	SSIS
Community-based: forums, Bug tracker...	✓	✓
Enterprise grade support with SLAs		✓
User Guide, Reference Guide	✓	✓
Mission Critical Support (optional)		✓

Fig. 8. Data Quality

The data quality process includes data cleansing, data validation, data manipulation, data quality tests, data refining, data filtering and tuning. It is a crucial area to be maintained in order to keep the data warehouse trustworthy for the business users. ETL plays a major role in data cleansing and data quality process as it helps automate most of the tasks. As shown in Fig V-A, Talend do not support most of the data quality tasks mentioned which clearly emphasizes that data quality is poor in Talend in comparison to SSIS.

VI. Conclusion

Both Talend open Studios and SSIS has their own strengths and weaknesses but SSIS has the upper hand due to its maturity and stability, good for enterprise-scale deployments, great support, Speed of implementation, Relevant data integration functions and ease of use . Though Talend is free its in capability of support, documentation and large scale implementations make it less suitable for commercial application especially with finanical and cloud based systems where as Microsoft is more reliable.

Considering the market presence, reliability, usability and support and all other advantages stated in earlier sections, Microsoft SSIS is far ahead compared to open source ETL tool Talend.

VII. Future Work

Microsoft proposed a new concept of Modern Data Warehouse. Traditional data warehouses are under pressure from the growing weight of explosive volumes of data, the expansive variety of data types, and the real-time processing velocity of how data is being used to grow and operate the business. These changes are so seismic that Gartner reports, Data warehousing has reached the most significant tipping point since its establishment. The modern enterprise needs a logical architecture that can smoothly scale to meet these volume demands with real-time processing power and the ability to manage any data type to rapidly connect the business to valuable insights.

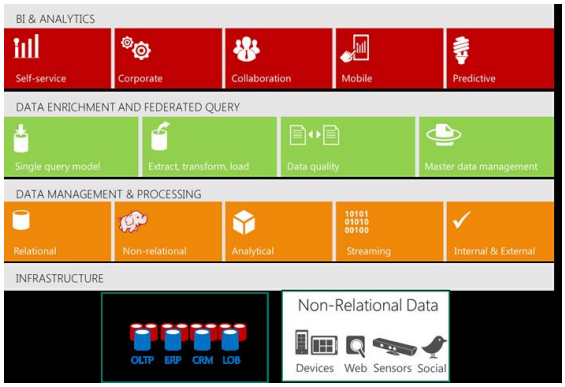

Fig. 9. Modern Data Warehouse [7]

This means that the traditional data warehouse needs to evolve into a modern data warehouse [7]. The modern data warehouse lives up to the promise of business intelligence from all data for business that is growing explosively, changing data types and sources and processing in real-time, with a more robust ability to deliver the right data at the right time. A modern data warehouse delivers a comprehensive logical data and analytics platform with a complete suite of fully supported, solutions and technologies that can meet the needs of even the most sophisticated and demanding modern enterpriseon-premises, in the cloud, or within any hybrid scenario [7] . It would be quite interesting to see the implementation of ETL tools on Modern Data Warehouses.

References

[1] "Simply easy learning," Retrieved from Simply Easy Learning by tutorialspoint.

[2] A. d. n. Ruiter, "approaches-to-extracting-and-transforming-data." [Online]. Available: http://blogs.msdn.com/:

[3] "Talend products," d.). Products. Retrieved from Talend.com. [Online]. Available: https://www.talend.com/products bottom

[4] "customer testimony on talend," d.). customer-reference. Retrieved from Talend. [Online]. Available: http://www.talend.com/resources/customer-reference

[5] Microsoft, "Ssis technical report," Microsoft, SSIS. Retrieved from microsoft.com, Tech. Rep., 2012. [Online]. Available: http://technet.microsoft.com/en-US/

[6] ——, “Casestudies,” Microsoft, CaseStudies. Retrieved from Microsoft.com, Tech. Rep., 2012. [Online]. Available: http://www.microsoft.com/casestudies

[7] ——, “The microsift modern data warehouse,” The Microsift Modern Data Warehouse. Retrieved from Microsoft.com. [Online]. Available: http://www.download.microsoft.com