International Journal of Research in Engineering, Science and Management
Volume-2, Issue-2, February-2019
www.ijresm.com | ISSN (Online): 2581-5792

Reliable Data Integration using Talend

V. Arul Kumar[1], M. Divya Anandhi[2], T. K. Gopika[3], V. Sheela Devi[4], S. Thenmozhi[5]
[1]*Assistant Professor, Dept. of Computer Science and Engg., Sri Eshwar College of Engg., Coimbatore, India*
[2,3,4,5]*UG Student, Dept. of Computer Science and Engg., Sri Eshwar College of Engg., Coimbatore, India*

***Abstract*: This project will store the customers on premises data to cloud based tables. The objective is to validate all the customer requirements using simple range check operations. Customer wants to use an ETL workflow for the migration. To serve this purpose, data needed to be extracted from various sources, transformed and loaded into the data warehouse which is the process called ETL. Data are turned into knowledge and knowledge into plans which are instrumental in profitable business intelligence. Data Warehousing has been evolved out of the desperate need for easy access to structured storage of quality data that can be used for effective decision making. ETL process can be proficient using various tools both open source and proprietary. Hence, transformation and integration of data is implemented using Talend Open Studio. This project will transform unstructured form of data to structured form of data. Then the data are processed using various business logics based on the requirements given which will be transfigured to the intermediate table called staging table. Finally, the staging table is transformed to the target table which can be loaded in the data warehouse for further retrieval. In this paper, we provide an empirical study of the ETL tool, an open source Talend Studio. Even though the dominance among a vast majority of software solutions, open source technologies, as the comparative analysis that this study has been undertaken and concluded that open sources tools are yet to be evolved in order to be sustainable.**

***Keywords*: data integration, talend**

1. Introduction

The term "Data Warehouse" was first introduced by Bill Inmon in 1990. According to Bill Inmon, Data warehouse is defined as subject Oriented, Integrated, Time-Variant and non-volatile collection of data supports decision making process in an organization. The operational database encounters several days to day transactions which makes the process of data analysis more and more complex and time consuming. The Data Warehouse provide us generalized and integrated data in multidimensional view and make it possible to use Online Analytical Processing (OLAP) tools for interactive and effective analysis of data in multidimensional space. The abstracting 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 frameworks 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 structures determine 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.

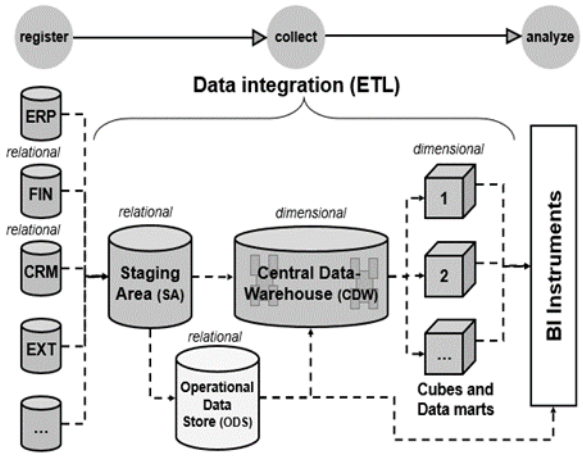

Fig. 1. Data warehouse architecture

2. The process of ETL

In today's businesses, decision-making processes and daily operations often depend on data that are stored in a variety of data storage systems, formats, and locations. In order to turn these data into useful business information, the data needed to be combined, sanitized, standardized, and summarized. For instance, data sources may need to be converted to a different data type or heterogeneous database servers may store the necessary data using the different schema. Dissimilarities like these must be resolved before the information can be successfully loaded to a target system. After the design and development of data warehouse in accordance with the business requirements, the process of consolidating the data into the data warehouse from various sources had to be addressed. Those Extract Transform Load (ETL) processes are critical in success of the Data Warehousing solutions. The process of extracting data from one system (extract), transforming it in accordance with design of the data warehouse(transform) and loading it into data warehouse system (load) combined to form 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 finally 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.

A. *Extract*

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

B. *Transform*

The complex part of ETL process is the transformation phase. In this phase, all the required data is exported from the possible sources but there is a great chance that data might still look different to the target schema of the data warehouse. Although data itself need to be formatted to confirm to the data type and other constraints of the data warehouse.

In the transformation 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 from measured data to the same dimension (i.e. conformed dimension as per the requirements of the data warehouse) using the same unit's so that they can later be joined. The transformation process also involves joining data from several sources, generating aggregate and surrogate keys, sorting, deriving new calculated values, and also applies advanced validation rules.

Another important aspect of the transformation process is data cleansing. Data cleansing 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 various identifies into a 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 (State, Country, City, State, City, ZIP code, City, Street).

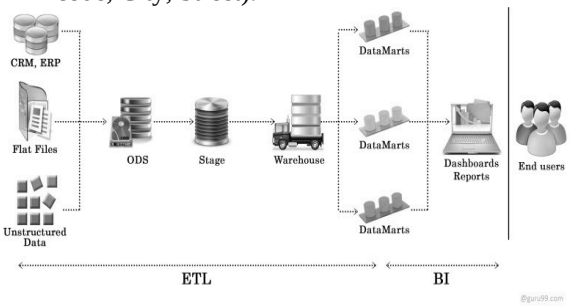

Fig. 2. ETL Workflow

C. *Load*

Once the data are extracted and transformed according to the requirements of the target data warehouse, the data are assumed to be ready to loading. Although, several aspects like how the data to be loaded, the impact and implication of loading process as well as handling such implications are to consider before loading the data into the data warehouse. The process of loading may impact the processing speed of the server both for loading and analysis can occur. It is also very much crucial to avoid database crippling while loading the data. ETL processes can be performed using almost any programming language. Developing 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 automation with help of specified tools. When the ETL process is carried out with an automation tool, data source mappings are fed into the tool of automation and the code that performs the mappings are created. Generally, mappings are done manually when there are only few tasks to be written. Nevertheless, it is more efficient to use an automation tool for ETL process. The mostly used open source ETL tool is Talend Open Studio and proprietary tool is SQL server integrated services.

3. Talend open studio

Talend Open Studio is the first open source data integration software released in 2006 after an intense research for over years. It is based on Eclipse RCP that primarily supports ETL-oriented implementations that are 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, Data Processing, Business Intelligence and data migration. The company breaks the traditional proprietary model by supplying open, innovative, attractive and powerful software solutions with the flexibility to meet the data integration needs of all the environments. Today, Talend Open Studio is the most innovative and powerful open source data integration solution on the market. Talend Open Studio for data integration helps you get your data to the target required place, in the specified form, at the right time. As the

leading the open source ETL solution for data warehousing, business intelligence and Talend Open Studio is

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

Talend Open Studio 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 suitable for the current iterative and incremental project environment further reducing the business risk by allowing little but this valuable piece of business functionality delivered in continues shorter time frames. As Talend is said to be committed to open source, the effort the company is undertake to sustain their support for open source products is appreciable. For instance, Talend data integration tool transparently supports the use of Hadoop clusters and of the Hive data warehousing environment.

Although there are some limitations to the community edition of Talend Open Studio. It is developed as a product for individual usage only and so it has been not possible to have more than one user (not just one user at a time but just one user per system). This creates 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. Further the free version doesn't support automation of tasks like scheduling, routing data etc. Another major drawback is lack of any commercial support.

4. Project methodology

The data gathered from various sources are analyzed and converted into intermediate tables by staging process. When there are no files, the process must exit with the appropriate message. The source file properties must be parameterized.

The customer must have the control in choosing the execution of a particular file. For example, if there are 4 employee files for the customer emp1, emp2, emp3, emp4. The customer can choose to execute emp3 alone. Also, the customer can execute multiple files that include comma separated values. For example, if customer mentions emp2, emp4 - both the files should be processed.

The data in the source files are performed with check operations like validations of employee text files for the given schema, duplicate employee IDs, any invalid characters in the Employee names, salary ranges, etc.

Integration functional architecture is a structural model that identifies data integration functions, interactions and corresponding IT needs. The entire architecture has been described by isolating specific functionalities in functional blocks.

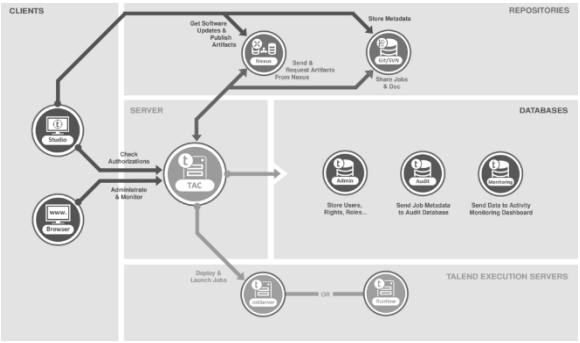

Fig. 3. Talend data integration functional architecture

Several functional blocks are defined:

- The Clients block includes one or more Talend Studio(s) and Web browsers that could be on the same or on different machines. From the Talend Open Studio, you can carry out the data integration processes regardless of the level of data and process complexity. Talend Studio allows you to work on any project for authorized processes. From the browser, you connect to the remotely based Talend Administration Center through a secured HTTP protocol.
- The Server block includes web-based application server and Talend Administration Center, which enables the management and administration of all the projects. Administration metadata (user accounts, access rights and project authorization for instance) is stored in the Administration database. Data of project items (Jobs, Business Models and Routines for example) is stored in the SVN or Git server.
- The Repositories block includes the SVN or Git server and the data repository. The SVN or Git server is used to organize all project items like Jobs and Business Models shared between different end-users. That is accessible from the Talend Open Studio to develop project items, from Talend Administration Center to publish, deploy and monitor the project items.
- The repository is used to store.
- Software Updates available for download.
- Jobs that are published from the Talend Studio and are ready to be deployed and executed.
- The Talend Execution Server blocks include one or more execution servers, which will be deployed inside your information system.
- Talend Jobs is deployed to the Job servers through the Talend Administration Center's Job Conductor to be executed on a scheduled time, date, or event.

The Databases block includes the Administration, the Audit and the Monitoring databases. The Talend Administration database is used to manage user accounts, access rights and project authorization, and so on. The Audit database is used to be evaluated the different aspects of the Jobs implemented in projects developed in Talend Open Studio with the aim of providing solid quantitative and qualitative factors for process-oriented decision support.

5. 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 capability of support, documentation and large scale implementations make it less suitable for commercial application especially with financial and cloud based systems whereas 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.

References

[1] "Simply easy learning," Retrieved from Simply Easy Learning by tutorials point.
[2] A. d. n. Ruiter, "approaches-to-extracting-and-transforming-data."
[3] "Talend products", https://www.talend.com/products bottom.
[4] https://help.talend.com/reader/wDRBNUuxk629sNcI0dNYaA/3Lyn4CR4M5Q2uOD8FWmOwg
[5] https://help.talend.com/reader/tXRG~nTonRYUwbOJscDgxw/~K8zTBzM7FdvgXmYD5XTmQ
[6] J. You, T. Dillon, J. Liu, "An integration of data mining and data warehousing for hierarchical multimedia information retrieval", 07 August 2002.
[7] Donglan Liu, Lei Ma, Xin Liu, and Hao Yu," Research on key issues of data integration technology in electric power system in big data environment", December 2017.