2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)

Data Integration in ETL Using TALEND

[1]Sreemathy J,[2] Infant Joseph V, [3]Nisha S, [4]Chaaru Prabha I, [5]Gokula Priya RM
[1,2,3,4,5] Department of Computer Science and Engineering,
[1,2,3,4,5] Sri Eshwar College of Engineering, Coimbatore, India
[1]sreemathy.j@sece.ac.in, [2]infantjoseph.v@gmail.com, [3] nishasece@gmail.com, [4]chaaruprabhai@gmail.com,
[5]priyasaikrish25@gmail.com

Abstract –Data Integration is the process of combining data from different sources to support Data Analytics in organizations. The best definition of data integration is given by IBM, stating "Data Integration is the combination of technical processes and business processes used to combine data from disparate sources into valuable and meaningful information." The important terms here are "combine data… into valuable and meaningful data" where it's about making the data more organized and useful. There are various methods of combining data into an integrated view. This paper describes the various steps involved in integrating data from various sources using the ETL process – The Extract, Transform and Load process, how the Talend Open Studio acting as a Data Integration and ETL tool helps in transforming heterogeneous data into homogeneous data for easy analysis and how all the integrated data is stored in a Data Warehouse to provide Business Intelligence users with suitable data for easy analysis.

Keywords: Data Integration, ETL, Talend, Data Warehouse, Business Intelligence.

I. INTRODUCTION

Every organization is forced to deal with a large amount of data. For easy access these data from different sources are integrated together using ETL or ELT process. The process of Data Integration is the most basic step.

The term Data Integration can be interpreted differently under different contexts. A few common DI approaches are as follows. One can argue that all the following approaches are a type of data integration, the difference being whether the data is physically moved or it is being manipulated.

DATA CONSOLIDATION
Data consolidation is physically bringing the data together from separate systems and consolidating it into one data store. The main aim of Data Consolidation is to reduce the number of data storage locations.

DATA PROPAGATION
Data propagation uses applications to move data from one location to another. Enterprise Application Integration (EAI) and Enterprise Data Replication (EDR) are technologies that support Data Propagation.

DATA VIRTUALIZATION
Through an interface, Data Virtualization provides an unified view of data from disparate sources.

DATA FEDERATION
Data Federation is technically a form of Data Virtualization. Instead of an interface Data Federation uses an virtual database to store the heterogeneous data in an unified model.

DATA WAREHOUSING
This method is included here because, the term "Data Warehousing" generally implies cleaning, reformatting and storing it in a single repository called the Data Warehouse, which basically is nothing but Data Integration.
The ETL technology supports Data Integration. The ETL process is the most common method used and it involves extracting the data onto a staging area, transforming it and loading it onto the Data Warehouse. The Data Warehouse also known as Enterprise Data Warehouse acts as a single large repository of all the integrated data for the client usage, implemented within a company to handle huge amount of data.

DATAWAREHOUSE -DISPARATE TO UNIFIED
As already discussed, a data warehouse is a single huge repository that stores huge amount of data. It can be considered as a consistent database that brings together information/data from different sources.
The data warehouse has a similar background as that of DBMS. The concept of data warehouse was introduced because the traditional database was able to store MB's or GB's of data, but as years passed and the data size exceeded MB or GB.
Hence the data warehouse was introduced that can handle data in Tera bytes. Also, the data warehouse stores historic data which will help an organization to easily analyze the already available data. This quality of a data warehouse being able to store historic data and current data under one roof makes it an integral part of business intelligence.

In a data warehouse the data is arranged into hierarchical groups, known as dimensions and into facts and aggregate facts. This arrangement is known as the star schema. While talking about data warehouse, Data Marts should also be mentioned. Data marts are related to data warehouse.

978-1-7281-5197-7/20/$31.00 ©2020 IEEE

A data mart stores data related to a single subject such as just sales or finance or marketing etc., whereas a data warehouse stores data based on multiple subjects as shown in the figure:1.

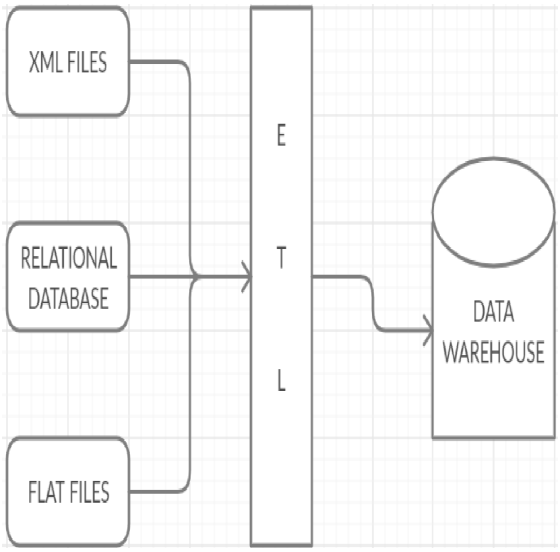

Fig :1-Data Warehouse

The data from disparate sources may pass through an operational data base for cleansing purpose before being loaded onto the data warehouse, to ensure the quality of the data that is being stored.Data Warehousing can be done based on ETL or ELT process. The ETL based Data Warehousing is the process which involves the use of a separate tool to gather, stage, transform and load data from disparate sources onto the warehouse.

The transformed data from the staging layer is integrated and stored in an operational data store (ODS) database. The ELT based Data Warehousing is the process where it eliminates the need for a separate tool for integration. Instead, a staging area is maintained within the Data Warehouse itself.

Every transformation done in the separate tool is done inside the warehouse itself. However, the final result of both processes is the same, that is the data from disparate sources is combined together to form a unified view. And, the unified data is of most use to the Business Intelligence users as shown in fig: 2.

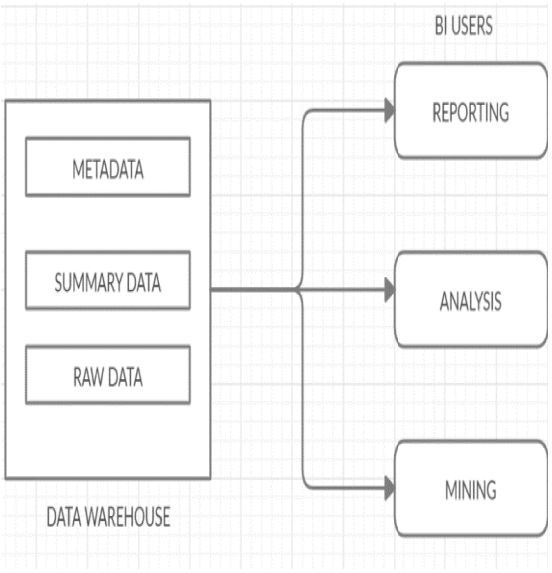

Fig: 2-Business Intelligence users and DW

Storing data in a data warehouse has its own benefits:

1. It maintains historical data even when the source transaction systems do not.
2. Integrating all the disparate data together provides the organizations with a central view across the enterprise.
3. The quality of data being stored in the data warehouse is more when compared to the same data in their originality.
4. Re-Organizing data from different sources into a single storage so that it makes more sense to the business users.
5. It helps in decision making problems.
6. Restructuring the data to improve the performance.

WHAT'S WHAT? -A DATABASE AND ADATAWAREHOUSE

A Database is a repository designed to store information/data related to a specific task that you are currently working on whereas a Data Warehouse is a single storage that stores historical data and any new data can be added to it.

While in a database you can only store information that is happening in the present whereas in a data warehouse you can analyze your business by using the already available data or add new data. The data warehouse is non-volatile and hence does not erase old data when new data is added as in fig:3.

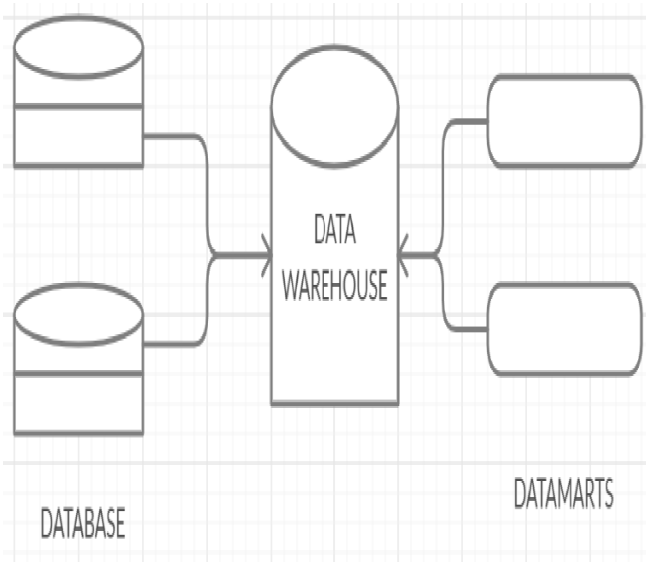

Fig: 3 Database and Data Warehouse

A database is used for Online Transactional Processing (OLTP) but can also be used for other purposes. A data warehouse is used for Online Analytical Processing (OLAP). A database collects data which acts as a source for the data warehouse.

A data warehouse has high performance regarding analytical queries than a database as the data added to a warehouse is de-normalized to reduce response time as opposite to a database where the data is normalized to reduce redundant data.

One reason that the data warehouse can store huge amount of data than a database is because it contains information in multidimensional tables whereas database contain information in two dimensional tables. The data warehouse contains layers of columns and rows to store all the data.

The most important application of maintaining a data warehouse in an organization is that it is useful for easy analysis of the accumulated historic data in business intelligence and to make accurate decisions for the company. But the database is not of that much use in business intelligence.

DATA INTEGRATION IN ETL

As already said, Data Integration is the process of combining data from different sources into a single view mainly for reporting, analysis and business intelligence. The purpose of doing data integration is to convert raw data into integrated data. The integrated data is data without redundancy or duplicate records.
Data Integration is done with the help of various integration tools which is where the Talend tool comes in. The Talend open studio provides various components that can be dragged and dropped that helps in data integration.
Data Integration has become an integral part of today's IT industries. However, it comes with its own consequences. Some of the most common problems are,

1. Understanding the process
2. ETL mapping of data from different sources
3. Processing the heterogeneous data according to the client
4. Large volume of data from the sources
5. Final output of the process

The main reason for doing data integration is to save time and improve efficiency. Ultimately, doing data integration helps in easy analysis of heterogeneous data for the purpose of business intelligence.

ETL PROCESS

ETL expands to extract, transform and load. As the name suggests it is the process of extracting data from different sources and transforming or editing it according to the request of the client. And after the transformation process the data is loaded onto a data warehouse.

Extract-In general the process of extraction of data from different sources is termed as the extract part of the ETL process. The extracted data are of different formats like XML, Relational databases, Flat files etc., The data are extracted into a staging area and not directly into the data warehouse as rollback of the extracted data will become challenging if corrupted data is copied to the data warehouse. Therefore, this stage is an opportunity to validate data before it is sent to data warehouse. During the validation process the incorrect data found are preferably sent back to the source system to identify and correct the wrong values. A logical data map is needed before extracting and loading the data-physically. This map describes the relationship between the source and target data.

Types of Extraction:
1. Full Extraction
2. Partial Extraction-without update notification
3. Partial Extraction-with update notification

Transform-In this process the extracted data is converted into the required format of the user or the client. So, the data is not in its original format. Therefore, for processing it requires some sort of cleansing, mapping and transformations .Here few customization steps or operations like aggregation can be performed.

ETL Transformation Methods:

1. Multistage data transformation
2. Warehouse data transformation

The data that does not require any change is called as direct move or pass through data. In this stage, customized user defined actions can be performed on the data.

The main transformations performed during this stage are done for validating the data that is to be loaded to the warehouse. The main transformations performed are,

1. Cleaning
2. Filtering
3. Data Standardization
4. Data flow validation
5. Data threshold validation check
6. Row and Column transposing
7. Joining
8. Splitting
9. Sorting

C. Load – The load process is writing the data into the destination or the final database. Loading is the final stage of ETL process. Here a huge volume of data is to be loaded into the final database.

Also, while this process is ongoing the Data Warehouse admins needs to monitor the process completely and also they can reduce the load based on the server performance as in fig:4.

Types of Loading:
1).Initial load
2).Incremental load
3).Full refresh

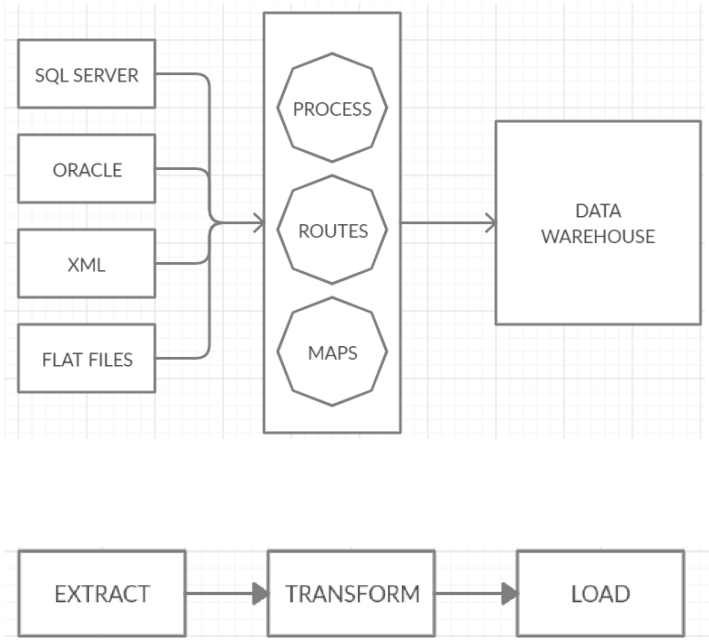

Fig: 4-ETL Architecture and Process

TALEND OPEN STUDIO

The open source data integration software is first provided by Talend. One of its main-products is the Talend open studio. Talend is used between operational system to provide integration and it is also used for migration, business intelligence, ETL (Extract, Transform, Load) and for data warehousing. The fully editable java code can be produced by Talend.

The business intelligence and database are independent. It should not have any special server because all jobs can run embedded. It is a GUI (Graphical User Interface) environment. The first version of that software was released in 2006 by after three years of intense research and development investment. It is an Open Source project for data integration based on Eclipse RCP that primarily supports ETL-oriented implementations and is provided for on-premises deployment as well as in a software-as-a-service (SaaS) delivery model. Talend should offer a completely new vision, and it utilizes technology, as well as business model in it.

The company 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.

Nowadays Talend Open Studio is the most open, innovative and powerful data integration solution. It can also provide integration suite, on demand, open profiler, data quality. It has many features there are business modeling, real-time debugging, robust execution, graphical development and metadata-driven design and execution.

BUSINESS INTELLIGENCE

Business Intelligence is the process of analyzing accumulated data for profitable business actions. Simply, it is the data analysis of Business Information. Business Intelligence is a process mainly done for making accurate business decisions. It analyses data from the past and present to make present and futuristic decisions. It gathers stores and analyses historical data for decision making. Data Warehousing, Data Integration are all examples of Business Intelligence. They are all carried out to provide a unified source of data for the users of business intelligence. An employee working with business intelligence is known as a business analyst. The main purpose of business intelligence is its role in making strategic business planning.

The data is organized using a framework called the Business Intelligence Architecture. This underlying architecture plays an important role in the process of business intelligence as they affect implementation and development decisions.

CONCLUSION

The process of data integration is the main and the most important step in the process of integrating data from different sources. It makes the difficult process of analyzing disparate data into a much more easy process. Data Integration, ETL process, Data Warehousing and Business Intelligence are all related to each other. The data is extracted onto a staging table, integrated using the ETL tools and loaded onto a Data Warehouse, from which business intelligence analysts uses the data to make business decisions.

REFERENCES

1. Y.Arens, CY.Chee, CN.Hsu, "Eetriving and integrating data from multiple information sources," vol. 3, no. 4, pp. 31- 34, July-Aug. 2001.
2. L.Munoz, J.Mazon, J.Trujillo, “ETL Process Modeling Conceptual for Data Warehouse:A Systematical Mapping Study,in June 2011, Vol, 2016.
3. Simitsis, P.Vassiliadis, T.Sellis, “in IEEE State-state Optimization of ETL workflows, vol. 29, no. 2, pp. 550-560, March 2014.
4. L. Munoz, J. Mazon and J. Trujillo, "ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study," in IEEE Latin America Transactions, vol. 9, no. 3, pp. 358-363, June 2011.
5. R Elmasri and S Navathe **“**Fundamentals of Database Systems”, vol. 4, no.7, pp.321-329,Nov 2012.
6. P. S. Diouf, A. Boly and S. Ndiaye, "Variety of data in the ETL processes in the cloud migration and validation : State of the art," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, 2018, pp. 1-5.