Data Validation for Nexus Solution using Talend

G. Monica[1], S. Vignesh[2], N. Istheyak[3], R. Prakash[4], Mathan Sangar[5]
[1]*Asst. Prof., Dept. of Computer Science and Engineering, Sri Eshwar College of Engg., Coimbatore, India*
[2,3,4,5]*Student, Dept. of Computer Science and Engineering, Sri Eshwar College of Engg., Coimbatore, India*

***Abstract*: Extraction of datasets from various sources and transforming the raw data into a structured form. Transforming a large volume of data is a tedious process and it consumes more time. An ETL tool is used for the transformation and validation process which improves efficiency and reduces querying time. Talend is an open source ETL tool which makes the data integration and data validation process much more simple.**

***Keywords*: data warehouse, Talend tool, data integration, data validation.**

1. Introduction

The objective is to transform the raw data into a structured scheme and generate consolidated reports.

This analysis can be used for validation purpose which eliminates invalid data in the selected files. The migration of data can also be achieved easily through this process. Organizations typically deal with large volumes of data containing valuable information about employee details, salary information, and others. As of now the data is in unstructured form and stored in different regions. Before getting into the validation process, the datasets from other sources should be extracted and loaded into the staging area. This can be achieved with the help of the ETL tool. Talend is an open source data integration platform. It provides various software and services for data integration, data management, enterprise application integration, data quality, cloud storage, and Big Data. Talend tool is used to collect all the files from various sources and performs efficient ETL mechanisms and validate datasets.

2. ETL Process

ETL stands for extract, transform, and the load is a three-stage process in data warehousing. It enables integration and analysis of the data stored in various sources. Once collected from multiple sources (extraction), it is reformatted and cleansed for operational needs (transformation). Finally, the data is loaded into either a target database, data warehouse or a data mart to be analyzed.

Extraction:

Extraction is the first process in the data warehouse. The purpose of the extraction process is to reach to the source systems and collect the data that has to be transformed. The extraction is one of the steps that consume larger time than transformation and loading. The complexity of the extraction process may vary and it depends on the source files. The extraction can be achieved in several methods. They are:

A. *Logical Extraction*

It is divided into two types:1) Full Extraction: The data is extracted completely from the source. Data can be extracted only once through this method. It handles deletion as well. 2) Incremental Extraction: The changes in the source files need to be updated in the staging table before uploading into the target system. This can be achieved using incremental extraction were changes in the dataset will be notified and re-updated.

B. *Physical Extraction*

It is divided into two types:

1) *Online Extraction:*

Data collection directly deals with the source system to access the data.

2) *Offline Extraction:*

The data is not directly extracted from the source instead it undergoes the staging process. These datasets do have its own predefined structure.

Transformation:

Data transformation takes place in the staging layer. The validation process is executed in this phase. It applies a set of rules to transform the data from the source to the target. This includes converting any measured data to the conformed dimension using the same units so that they can later be joined. It requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving newly calculated values, and applying advanced validation rules. In this phase, the extracted data is cross-checked for data quality.

Some basic ETL transformations are Cleaning, De-duplication, Format revision, and Key restructuring.

Some advanced ETL transformations are Derivation, Filtering, Aggregation, Joining, Splitting, Filtering, Data validation and Integration.

Load:

The last step of the ETL process is the loading phase. The information from data sources is loaded and stored in the form of tables. The two types of tables in the database structure, they are fact tables and dimensions tables. Once the fact and dimension tables are loaded, the aggregates are created. The processed data is been loaded into the end target or the data warehouse as a flat file or in other file formats. Loading is divided into two types:

1) *Incremental Load*

The processed datasets are loaded into the target system and whenever the source data is updated the changes are reflected back in the target systems.

2) *Full Load*

The processed data is loaded into the target system for the first time. This process of loading can be performed only once.

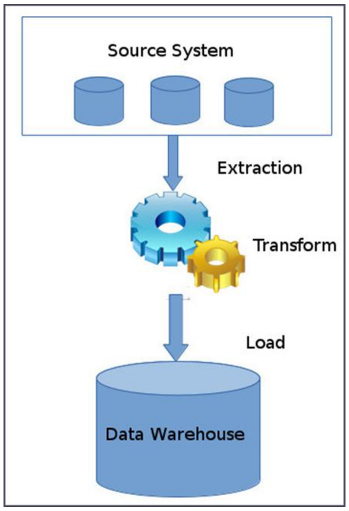

Fig. 1. Process

3. Data warehousing

A data warehouse is a technique of integrating data from various sources to provide analytical reports, structured queries, and decision making. It is a database of information aimed for query and analysis. Data warehousing focuses on cleaning of data, integration of data, and consolidation of data. The data in the data warehouse can be used effectively using certain decision support technologies which can collect, analyze data and make informed decisions. A data warehouse also provides current and historical information. Data warehouse consists of the data mart, metadata and data support. A data mart is a data warehouse that constitutes a sole subject. Whereas metadata is a description of other data.

The data warehousing uses OLAP (Online Analytical Processing). OLAP allows users to extract selective data and query them for analysis. This analysis process involves the collection of data from various sources and their storage in the data warehouse to cleanse and sort them into data cubes. An OLAP cube consists of different categories of data which are taken from dimensional tables. There are different OLAP analytical operations like Roll-up, Slice, Dice, Drill-down, and pivot.

Some of the advantages of using a data warehouse include processing of large complex data efficiently, providing a permanent storage space for the data which allows users to generate reports, security, a creation of metadata. When a data warehouse is properly utilized it provides different advantages to your business.

4. Talend

Talend is the first source of open source data integration software. Its main product is Talend Open Studio. The first version was launched in 2006. It is the Open Source tool for data integration, based on Eclipse RCP that supports ETL-oriented process and is provided for on-premises deployment as well as in a software-as-a-service (SaaS) delivery model. Talend Open Studio is mainly used for integration between the data, as well as ETL (Extract, Transform, Load) for Business Intelligence and Data Warehousing, and for migration. Talend provides a completely new vision, reflected in the way it utilizes technology, as well as in its business model. The company shares the traditional model by providing open source, innovative and powerful software solutions with the flexibility to meet the data integration needs of all types of organizations.

Talend offers a completely new vision, reflected in the way it utilizes technology, and in its business model. This tool is supplying open, innovative and powerful software solutions with the flexibility to meet the data integration needs of all types of organizations. This is the first open source for data quality solution. Talend Data Quality is the graphical data quality management environment that processes data, like address, phone numbers, spellings, synonyms, and abbreviations. Most of the organizations get data from multiple places and are store them separately. Now if the organization has to do decision making, it has to take data from different sources, put it in a unified view and then analyze it to get a result. This process is called as Data Integration. A major challenge with Talend is that the feature which is used for scheduling is a very basic one where it is available only with enterprise editions and not with open studio distribution. Because of this issue, users working with Talend open studio are not able to make use of the scheduling feature.

5. Data Validation

Data validation is used for checking the accuracy and quality of source data before using, importing or otherwise processing data. Different types of validation can be performed such as verifying the proper experience details of the employee, catching invalid characters in the employee table. Data validation also called as data cleansing, which can be achieved through mapping the datasets. When moving and merging the data, the important is to make sure data from different sources and repositories will conform to business rules and not become corrupted due to inconsistencies in type or context. The goal is to create data that is consistent, accurate and complete so to prevent data loss and errors during a move. In data warehousing, data validation is often performed prior to the ETL (Extraction Transformation Load) process. A data validation test is performed so the analyst can get insight into the scope or nature of data conflicts. It is a general term and can be performed on any type of data.

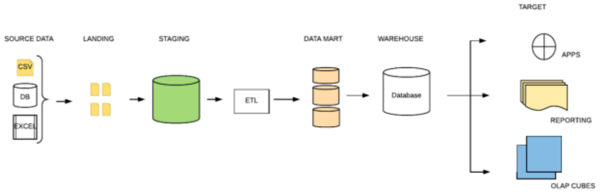

Fig. 2. Workflow

6. Conclusion

The Data validation process is carried out by extracting datasets from various sources. It is difficult to perform data validation when the dataset is in huge volume, still, it can be achieved with the help of an ETL tool which handles a large amount of data in a structured manner. Comparatively, are a large amount of time can be saved with Talend open studio. Our project helps in validating the datasets by removing the duplicates and loading the resultant files in a cloud environment.

References

[1] https://www.1keydata.com
[2] https://www.tutorialspoint.com/sap_bods/etl_introduction.htm
[3] https://help.talend.com
[4] https://help.talend.com/reader/I5~nujSNXadpT4WYWRHrjw/root