2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)

Data Validation in ETL Using TALEND

Sreemathy J
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India
sreemathy.j@sece.ac.in

Priyadharshini S
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India
priya24sundaram@gmail.com

Radha K
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India
radhakrishnarajbe@gmail.com

Sangeerna K
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India
ksangeerna@gmail.com

Nivetha G
Department of Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India
nivethaganesh0510@gmail.com

***Abstract* - This paper describes the methodology of a data warehouse used for analysis, validation using the flat files from various sources. A flat file is a file that stores data in plain text format. A flat file includes a record per line of a table. The columns in a record are delimited by a comma separated fields. Flat file database does not contain multiple tables. Extract Transform Load (ETL), is the process in data warehousing that deals with extracting data from different source system and placing the processed data in the data warehouse. In this paper we are going to explain about etl process, talend tool and validation of the data using talend tool.**

Keywords: Data warehouse, ETL, Talend and data validation.

I. INTRODUCTION

The goal of our proposed system is to generate consolidated file that can be analyzed easily. This analysis helps us to remove all the duplicates and maintain the correctness of files in the database. Many Organization deals with huge amount of data containing various information about employee details, salary information and others. But these data may contain unnecessary or duplicate files. To achieve consolidated files, data integration process is done by using efficient ETL mechanisms. In this new landscape, this tool is used to collect all the files from sources and performs efficient ETL process to achieve validated files.

II. DATAWAREHOUSE

Data warehouse is a large set of consistent data from various sources that is been stored in a single large repository and made available for the clients or users for their own business intelligence reports and analysis. Fig: 1 shows the function of Data Warehouse. It provides to induce and centralized data to us in multidimensional view. It also provides Online Analytical Processing (OLAP) tools. In multidimensional space we use multiple functions like Data Extraction, Data Cleaning, and Data Transformation. Data loading is the functions of data warehouse tools and utilities. Its environment consists of a data store, data mart and the metadata. The main function of a data store is to feed the data into a data warehouse for the purpose of business analysis. Data Mart is a subset of a data warehouse where the data can be accessed quickly with less processing time. The key factors to develop a data warehouse are Data redundancy and Type of end users. The concept of data warehouse has been grown tremendously from the integration of different technologies. We need to understand and analyze the business needs to design an effective and efficient data warehouse. They are widely used in the following fields like financial services, Banking services, Consumer goods, Retail sectors, controlled manufacturing.

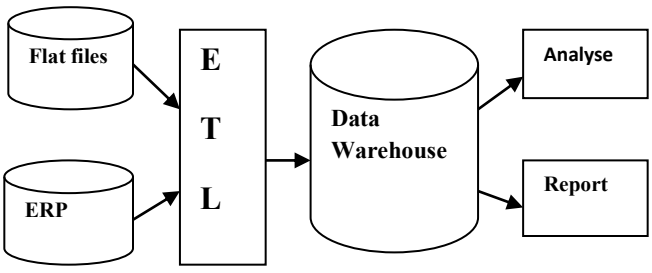

Fig: 1-Funtion of Data Warehouse

III. DATABASE

A Database is a collection of information, and it can be easily stored, accessed, updated, modified, retrieved, and managed. Data is organized into tables, rows, columns and it is easy to find relevant information. It is a single repository that contains the collection of logically related and similar data. Both retrieval and modification of data can be done on the database based on the operation performed. DBMS (Database Management System) is software for managing and creating databases. Database Management System provides a systematic way to create, update, retrieve, store and manage data. A DBMS makes an end user to create, store, read, update and delete data in a database. It is software that has been used to create and interact with the database, and it contains the interrelated data.

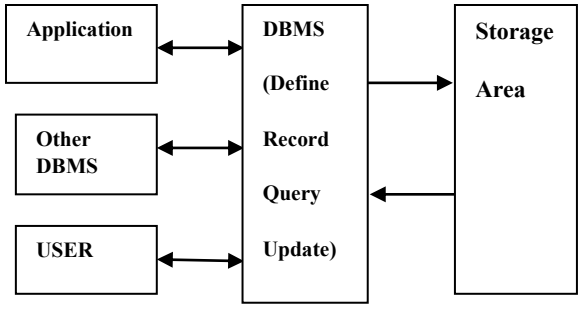

Fig: 2-View of Database

It also provides a set of languages to perform operation on the interrelated data's and the languages are Data Definition Language, Data Manipulation Language, and Data Control Language. Some of the DDL commands are, create, alter, drop, grant and revoke and DML commands are update, delete, insert and TCL command are commit, rollback, save point, set transaction. Data models define how the logical structure is modelled for a database. Data Models are entities to introduce an abstraction in a DBMS. Data models define how they are processed and how data is connected to each other and stored inside the system. Data models are of different types they are High Level Model, Representation Model, and Low Level. High Level model guarantee the requirement of the users and it is conceptual form of data but it is not concerned with representation of data. Representation model is used to represent the physical structure of the data that is stored in the database. This model is classified as Hierarchical Database Model, Relational Database Model, and Network Database Model. Hierarchical Database Model organizes the data in tree structure. A Relational Database represents are simple two-dimensional tables for all data in the database called relations. In the relational table, each row is called as tuple, which represents an attribute (fields).Network Database Model was developed as an alternate to the hierarchical database. This model takes the advantage on the hierarchical model by providing multiple paths among segments. The database system architecture has different layers, and they are Centralized and Client Server system, Server System Architecture, Parallel System, Distributed System, and Network Types. Centralized and Client-server System will run on a single computer system. The centralized database system has a single processor, which is associated with the data storage devices. With the help of a computer network, data can be accessed from the multiple sites, and it is maintained at the central site. Client/Server architecture has two logical components namely server and client. Clients are small workstations whereas a server is a larger workstation, mainframe computer or system mini range computer system. Parallel database system architecture consists of data storage disk and multiple Central Processing Units (CPUs) in parallel. SQL is used for storing, accessing and manipulating database. SQL is abbreviated as Structured Query Language. SQL can retrieve data from a database. Here for our premise we use MySQL database. MySQL database is an open source relational database management system works on structured query language. MySQL can be used in data warehousing to improve its performance. MySQL deals with different types of table. They are Fact Table, Target Table, Dimension, and Normalization Table. Normalization is a key factor of Data Warehouse to improve its performance. The main advantage of using normalization is to reduce the storage of data warehouses. As MySQL is stable and reliable, it provides data security, works with larger database, customizable.

IV. *ETL PROCES*

Extraction is the process of fetching the data from source system. The fetched data is transformed depending upon the client request. After the transformation the data is loaded into the database. ETL is expanded as extract, transform, and load. ETL refers to a brief process.

A. Extract- The first step of an ETL process is extracting the data from the source system(s). This portrays the most important step of ETL that is extracting data correctly for the next subsequent steps. ETL process combine data from different source systems such as flat files, xml files, etc. Each different sources may use a different data format. Some of the common source formats include relational databases, XML, JSON and flat files, but may also include non-relational database structure. The extracted data is loaded to the destination database is another way of performing ETL where there is no intermediate data storage is required. In general, extraction is used to convert the common data format into transformable data.

An important part of the extracting the data starts from validation to check whether the data fetched from the sources has the correct or expected values. The wrong data is preferably sent back to the main system for further validation to identify the incorrect records. There are many ways to perform extraction, and some of them are, Full Extraction, Physical Extraction, and Incremental Extraction.

There is no additional information is necessary. An example of full extraction may be a SQL statement is used for accessing the complete source table. Physical extraction depends on the logical extraction method and the capabilities and restrictions on the source, and the extracted data can be physically extracted by two ways. The data can be extracted from online source system or from offline. An offline structure will already exist, or it is generated by an extraction method. The methods of physical extraction are, Online Extraction and Offline Extraction.

In incremental extraction, the data which is changed well-defined will be extracted. This is the last step of extraction. For finding this change we must identify all the changed information since this specific time event. Many ETL functions do not use any change-capture techniques. The tables are extracted from the main source systems to the staging area, and these tables are compared with a extracted table from the source system to identify the change. This does not have significant impact on the source systems, but it place a burden on the data warehouse processes, particularly if the data volumes are large.

B. Transform- Transformation is the process of performing cleansing and aggregation of extracted data that may need for analysis. ETL transformation can be achieved using two methods are Multistage data transformation – The classic extract, transform, load process. Extracted data is moved to a staging area where transformations occur before loading the data into the warehouse. In-warehouse data transformation – the process flow changes to something more like ELT. Data is extracted and loaded into the intermediate phase, and transformations are done there. There are different ETL transformation types. They are, Basic transformations are Cleaning- removing the unwanted data formats from the table. Key restructuring- performing key relationship. Advanced transformations are Filtering- Selecting particular rows or columns. Joining- Joining data from different sources. Splitting- Splitting one column into multiple columns. Data validation: Complex or simple validations – example, the row is rejected if the first two columns are empty or null for processing. Aggregation: Elements are aggregated from various source tables and databases. Integration: Giving each unique data element one common name with one common definition. For the same data element Data integration checks different names and values.

C. Load- Loading commonly depends on large amount of data to be loaded into the database. Uses are layering of logical data, creating a tool for search, developing an algorithm to detect duplication, and Running a real-time system. There are two methods of loading a data into a warehouse; Full load- entire data is dumped at a time, when a data source is loaded into the database. Incremental load- data between the target and source is dumped at regular intervals. The lastly extracted date is noted so that the records of these dates are loaded. There are two types of loading. They are, Streaming incremental load and Batch incremental load The full load is comparatively straightforward. When the time you start incremental loads, things get more complex. The three most common problems are Ordering, Schema evolution, and Monitor ability.

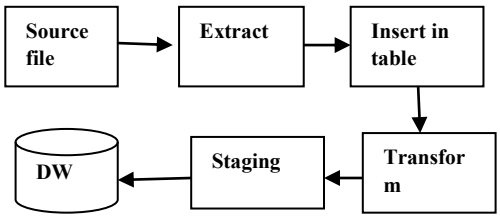

Fig: 3-ETL Process

V. TALEND OPEN STUDIO

Talend is an Open Source project for data integration and Big Data. It is a developer tool and job designer based on eclipse. To create and run ETL or ETL Jobs, Drag and Drop components and connect them. It automatically creates the Java code for the job instead of typing a single line of code.

RDBMS, Excel, SaaS Big Data ecosystem are the multiple options to connect with Data Sources and it connects with apps and technologies like SAP, CRM, Drop box and many more.

Talend Open Studio has the following characteristics. They are, It is cost-effective and highly versatile, In the latest version, they have enhanced its support for the Internet of Things (IoT), User friendly and adaptable, Combine data across different data sources and targets, Provide deeper control of data lineage, more visibility and control, Enables faster updates to MDM data models.

Some important benefits of Talend Open Studio are, It provides all features needed for data integration and converting jobs to Java code automatically and much more, It is a free tool; hence, there are big cost savings,

Multiple organizations have adopted TOS for Data integration, which shows very high trust factor in this tool. The Talend community for Data Integration is very active.

Talend Open Studio for Data Integration includes, Data migration from one database to another, Regular file exchanges between systems, and Data synchronization. It is the most inventive and effectual data integration solution on the today's world market. Business modelling, Real-time debugging and robust execution, Graphical development, Metadata-driven design and execution are the main features and benefits of talend

VI. DATA VALIDATION PROCESS IN ETL

Data validation (ref 1) is a used for checking accuracy and quality of your data, typically performed prior to processing and importing. It can be considered as a form of data cleansing. Data validation make sure that your data is complete (no blank or null values), unique (contains distinct values that are not duplicated), and the range of values is consistent with what you expect. Data validation is a part of processes such as ETL (Extract, Transform and Load) (ref 3) where you extract a data from a source to a target database, and you can join it with other data for analysis. Data validation ensures that when you perform analysis, and your results are accurate. Flat files from various flat files are extracted. Depending upon the client requirements, the fields are chosen from the incoming files. These files are initially inserted into the intermediate table. Some type of transformations according to the client is done in the intermediate table. After all transformations are finished, the data are moved into an area called staging. From this stage the data is moved into the database for storage. Once the data is moved into the data mart, the staging table should be deleted.

VII. CONCLUSION

This data validation process involves in validating the files from various sources. It is difficult for an organization to analyse by them, so in such scenarios our project helps to provide a valid file that can be used for easy analysis. Using ETL process and with the help of ETL Tools like Talend we can perform analysis, report generation, data validation, duplicate deduction in an organized manner. Extraction of data can be done from various file formats. The files can be converted into a common format required by the client. Different transformation techniques can be applied to the data. Finally, the target table or file is loaded into the database.

REFERENCES

1. G. Furlow, "The case for building a data warehouse," in *IT Professional*, vol. 3, no. 4, pp. 31-34, July-Aug. 2001.
2. J.Quevedo, J.Pascual, S.Espin, J.Roquet,IFAC, Data Validation and Reconstruction for Performance Enhancement and Maintenance of Water Networks, Volume, 2016.
3. Y. Guo, C. Ten and P. Jirutitijaroen, "Online Data Validation for Distribution Operations Against Cybertampering," in *IEEE Transactions on Power Systems*, 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. Simitsis, P. Vassiliadis and T. Sellis, "State-space optimization of ETL workflows," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 10, pp. 1404-1419, Oct. 2005.
6. P. S. Diouf, A. Boly and S. Ndiaye, "Variety of data in the ETL processes in the cloud: State of the art," *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, 2018, pp. 1-5.