Scientific Journal Impact Factor (SJIF) : 4.364
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

DATA PROCESSING THROUGH DATA WAREHOUSE AND DATA MINING

G. Raghavendra[1] and K.Thanweer Basha[2]
[1,2]*Department of MCA, Sree Vidyanikethan Institute of Management*

Abstract— This paper exposes the content of data processing in data warehouse with data mining tool. Data changes are possible to examine the patterns and trends by using tool. Here we had taken one of the data integration tool i.e Informatica. Data warehousing is integrated data with multiple databases. A Data warehouse is an integrated, time-variant, subject oriented, non-volatile collection of data. The process of design warehouse using ETL tools, like Ab Initio Software, Amazon Redshift, Analytix DS, CodeFetures, DATAllegro, Holistic Data Management, Informatica, ParAccel and Teradata. Data Mining is to identify patterns and solve problems through data analysis. Data mining help to generate reports. Mining is the biggest task to analysis large data sets. It is time taken process. To Process the mining functions using tools, like RapidMiner, Weka, R-Programming, Orange, Knime, NLTK.

Keywords— Data Processing, Data Warehouse, Data Mining, Informatica, Identigy Patterns and Report Generation

I. INTRODUCTION

Data warehouse and data mining are important to analysis the data. It helps to take decision on manager level. Present data analysis is most important factor because huge amount of data is stored day to day. The top most companies are facing problems to analyzing data. Here we focus on tools, one is Informatica to design warehouse and second one is mining process.

The [7]author simple thought behind writing all the essential ingredients of Informatica, starting from to extraction, installation to working on client. Author explains Informatica PowerCenter tool helps with integration of data from almost any business in almost any format .Author cover all the aspects of the Informatica PowerCenter tool. Informatica is widely-used tool across the globe for various data integration process.

II. ETL TOOLS

ETL tool is used to integrate the data and process design of data warehouse. ETL means is Extract, Transform and Load. Extract is the data process to collect data different databases. Transform is the process of converting existing data based on user requirement. Load is the process to written the data into the target.

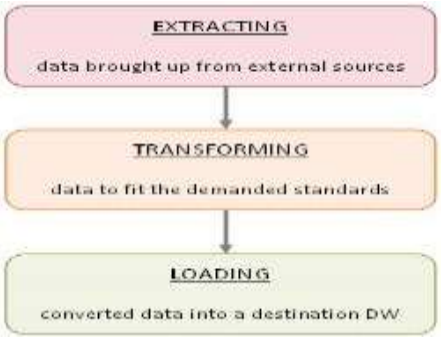

Figure 1. ETL Process

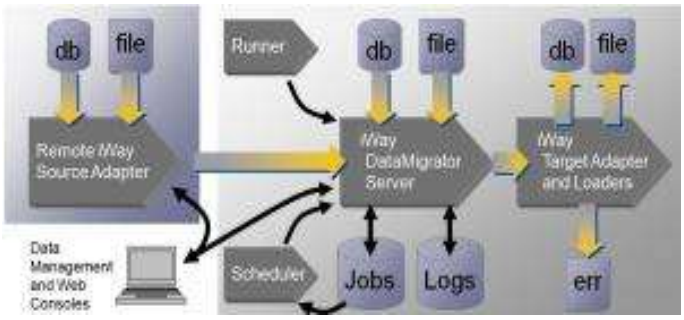

Figure 2. ETL Data flow

III. DATA WAREHOUSE

Data warehousing is the electronic storage data of a large amount of information by a business and users. Warehoused data must be stored which is secure, reliable, easy to retrieve and easy to manage. It is subject oriented, Integrated, non-violated, time variant it helps to decision making to the managers. A good data warehousing system can also make it easier and flexible to users for different departments within a company to access each other's data. Effective data storage with help database and management are also what make things like making travel reservations and using different tools. Data warehousing is huge amount of data extracting from different sources, cleaning the data and storing it in the warehouse. Example is data warehouse of a company stores all the relevant information of projects and employees. Data warehousing emphasizes the capture of data using tools from diverse sources for useful analysis and access, but does not generally start from the point-of-view of the end user who may need access to specialized, sometimes local databases. Data warehouse implementation is depending on the storage device and database. Here we used Informatica tool to design data warehouse.

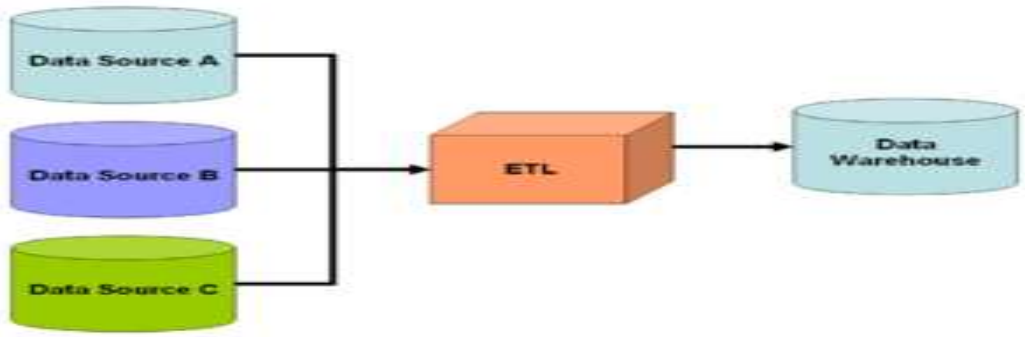

Figure 3. Data Warehouse Architecture

IV. INFORMATICA POWER CENTER

Informatica is a software development company offers data integration tools and products. Most popular tool like ETL, data masking, data quality, data replica, data virtualization, master data management. Business organizations setting up their data warehouse and maintaining data reports require an ETL tool. Among other product Informatica powercenter provides wide range of product editions. Informatica PowerCenter is a data integration tool based on ETL architecture. Applications of Informatica are data migration, Application integration, datawarehousing, Middleware.

Informatica PowerCenter is used for data integration. Informatica PowerCenter contains four client tools.

1. PowerCenter Designer
2. PowerCenter Workflow Manager
3. PowerCenter Workflow Monitor
4. PowerCenter Repository Manager

Informatica PowerCenter has more components to transform data into ETL tool. The Transformation are effective role execute the data source to target. The basic example of common transformation rules are as show below. They are total seven steps.

4.1 Defining the source metadata.
4.2 Designing the target metadata.
4.3 Designing the mapping.
4.4 Creating the sessions for each mapping
4.5 Creation of source to target connection
4.6 Creation of workflow
4.7 Execution of workflow

4.1 Defining the source metadata:

Source database is most important to the creation of source metadata. Database products are used to handle the huge amount of data, which one is to transform into Informatica tool, it create and establish a connection, then to create metadata into Informatica. Informatica maintain metadata in the form of source qualifier format. Source qualifier format is base information to process the data into Informatica design component.

4.2 Designing the target metadata:

Target metadata is stored or load task completed data. Finally it generates data warehouse. Data warehouse collected depending on the database or any other products.

4.3 Designing the Mapping:

Mapping is one of the most important components in the Informatica design. Mapping is a collection of sources metadata and target metadata linked together by a set of transformations and conditions. This transformation consists of a set of objects to link and condition based on the warehouse design, which defines the data flow and how the data is loaded into the targets. Links are shows effective on the target definition.

4.4 Creating the sessions for each mapping:

Session property is a set of instructions with proper condition. Sessions are instructs Informatica how and when to move the data from source to targets. A session property is a task to execute data through source to target. Finally it design data warehouse.

4.5 Creating source to target connection:

Informatica is integrated sources with different database. Here create a new connection independently source and target. The control of data flow and process is mainly divided by using connection tab. Connection tab help to connect different databases.

4.6 Creation of Workflow:

A mapping creates or configures a set of transformations using session. A workflow is a set of instructions that tell the Informatica server how to execute the tasks. A workflow is an object that represents a set of tasks in Informatica. A session is a set of instructions to move data from sources to target.

4.7 Execute the workflow:

All the properties and rules are set in the transformation, then start the workflow and execute the task. Data processing all the condition are satisfied by the given Informatica transformation. Data load the data into target.

V. DATA MINING

Data mining is the process of analyzing data existing sources. Data mining is a Analysis process using analytical tools. Data mining is the process of finding correlations or patterns of fields in the large database. Mining is analysis and find the hidden information. Analyzing is not simple process, huge amount of processing depending on the effective algorithms'. Data mining techniques are the result of a long process of research and product development. Mining process is depending on mining tasks, as shown below.

Data Mining Tasks:

5.1 Predictive tasks: Classification, Regression, Time Series Analysis, Prediction.
5.2 Descriptive: Clustering, summarization, Association rules, Sequence Discovery.

5.1.1 Classification: Classification maps data into predefined groups or classes. The classes are determined before examining the data. Example is an airport security screening station.

5.1.2 Regression: Regression is used to map a data item to a real valued prediction variable. This is involves the learning of the function that does this mapping. Example is saving data, retirement saving amount analyses and interest rate calculation.

5.1.3 Time series analysis: The value of an attribute is examined as it varies our time. The values usually are obtained as every spaced time points. Example is daily, weekly, hourly data collection.

5.1.4 Prediction: Data states based on past and current data. Prediction can be viewed as a typed of classification. The different is that prediction is predicting a future state rather than a current state. Example is flood prediction data, water level data, rain water amount data.

5.2.1 Clustering: Cluster is a similar to classification except that the groups are no predefined, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. A special type of clustering is called segmentation.

5.2.2 Summarization: summarization maps data into subsets with associated simple descriptions. Summarization is also called characterization or generalization. It extracts or derivers representative information about the database.

5.2.3 Association Rules: Link analysis alternatively referred to as affinity analysis or association. An association rules is a model that identifies specific types of data association. An association rule is a model that identifies specific types of data association.

5.2.4 Sequence Discovery: Sequential analysis or sequence discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of action.

VI. CONCLUSION

Data processing through tool and maintenance is complex task. Data hide, format and existing state are not common to all the databases. The companies are facing problem in integration and analysis of data. Different tools are available but their function and modulation are unique respect to their company polices. Here we had taken a tool Informatica and mining tasks. Data Control and processing in Informatica explained step wise. Informatica is best tool for data integration. Data mining tasks are suitable for all type of data analysis processes.

REFERENCES

[1] https://www.educba.com/10-popular-data-warehouse-tools/
[2] https://thenewstack.io/six-of-the-best-open-source-data-mining-tools/
[3] http://www.investopedia.com/terms/d/data-warehousing.asp
[4] http://www.guru99.com/introduction-informatica.html
[5] https://www.edureka.co/blog/what-is-informatica/
[6] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[7] Learning Informatica powerCenter 9.X - Rahul Malewar.

