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

Report Generation using Slowly Changing Dimension

V. Arul Kumar[1], L. Akshayaa[2], K. Madhumidha[3], D. Radhika[4], E. Ramya Kamatchi[5]
[1]*Assistant Professor, Dept. of Computer Science and Engg., Sri Eshwar College of Engg., Coimbatore, India*
[2,3,4,5]*Student, Dept. of Computer Science and Engg., Sri Eshwar College of Engg., Coimbatore, India*

***Abstract*: This paper describes the methodology of data warehouse used for analysis and generating employee details with the support of ETL tool. The Employee details are generated as a raw data which are in an unstructured format. The raw data can be converted from unstructured to structure format using some staging activities. To solve this problem, the building of an employee data warehouse seems to be efficient. Here in this paper we explain the concepts of the data warehouse, Talend open studio and online analytical processing (OLAP). Conversion of data in the data warehouse into a multidimensional data cube is used for analyzing. More information about the employee can be calculated with the reduced query time and the reports are generated.**

***Keywords*: Data Warehouse, Talend tool, Data Management, Data Integration, ETL.**

1. Introduction

The goal of our proposed system is to generate structured form of data that can be analyzed easily to increasing readability. This analysis can be used to understand the employee details clearly. Data is gathered about each individual employee from various sources and integrated. After integration of data ETL process is done and then loaded into data warehouse. A data warehousing is the approach for collecting and managing data from various sources to provide meaningful business needs. Here all the details are stored in the database in the specific structured format using ETL tool. It is the process of copying data from various sources into a destination system which differs from the sources. Organizations typically deal with large volumes of data containing valuable information about employee details. But these data are stored in operational databases that are not useful for decision makers. In order to achieve this goal, data integration process is done by using efficient ETL mechanisms. In this new landscape, Talend tool acts as a consolidated repository to collect all the master data from sources and performs efficient ETL process.

2. Database

Database is a systematic collection of data. Databases support storage and manipulation of data. Databases make data management easy. Mostly data represents recordable facts. The data which are represented as a fact. Some database components are: 1) Character 2) Field 3) Record 4) File.

A. Database management system

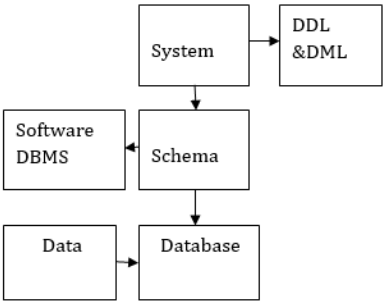

Fig. 1. Data management system

Database Management System is software that has been used to create and interact with the database, and it contains the interrelated data. The information about the particular domain is available in the database management system. It also provides set of languages to perform operation on the interrelated data's and the languages are 1) Data Definition Language 2) Data Manipulation Language 3) Data Control Language. Some 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 Model be used to describe the structure of the database and it will also provide the definition and format of the data that are being stored in the database. It is an abstract model that establishes the elements of data and standardizes how the relation between real world entities explains. For instance, a data model may specify that the data element representing a car be composed of a number of other elements which, in turn, represent the properties of the owner. The different types of data models are 1) High-Level Model 2) Representation Model 3) Low-Level. High-Level Model ensures the requirement of the users and it is not concerned with representation of data but it is a conceptual form of data. Representation model is used to represent the physical structure of the data that are stored in the database. This model is classified as 1) Hierarchical Database Model 2) Relational Database Model 3) Network Database Model. The data in the hierarchical database model is represented by collection of records from various sources and the relationship is represented by links. This model use tree structure to represent the records rather than arbitrary graph.

Database architecture uses the different programming languages to design a particular type of the software for organizations. The design of a DBMS depends on its own architecture. It can be centralized or decentralized or hierarchical. The architecture of a DBMS has two types which are single tier, multi-tier.

B. Structured query language

Structured Query Language (SQL) is used to access the data that are available in the MySQL database. The set of related information that are stored in the relational database management system are created and operated using Structured Query Language. The benefits of databases are: 1) reduce the duplication of data 2) allow of sharing of data by several users 3) data is accurate and consistent. The OLTP (Online Transaction Processing) system is a source of original data, and it provides the data to warehouse, the system emphasis on fast query processing and maintaining data integrity and its effectiveness is measured by the number of transaction that the system has performed per second. The OLTP system makes use of simple queries to return the records as requested by the user, and it also maintains the current data that are stored in the form of schema in the entity model. The table in the OLTP systems are normalized in order to reduce the redundancy and to avoid the space constraint. It is used to do much small transaction with simple query, used for data entry, financial transaction, customer relationship management and retail sales. Benefits of OLTP system are: 1) It reduces the paper work 2) It handles large data, complex calculation and higher peak loads 3) It provides higher performance. The raw data's are collected from various sources and it is inserted into the database with help of SQL queries. Queries like insert, update and create are used to store the data in the database and queries like select are used to retrieve the data from the database.

3. ETL process

ETL process is responsible for the extraction of data from various heterogeneous data sources, their transformation and the loading the whole transformed data into data warehouse The important factor for successful implementation of data warehouse project depends upon the quality of ETL design process and also based on it maintenance.

A. Extraction

In this step, the desired data is identified and extracted from different sources, including database systems and applications. The size of the extracted data varies from hundreds of kilobyte up to the gigabyte which depends on the source systems and business situations. Here the data is extracted from the OLAP database. Basically the extraction phase is desire to convert the data into a single format convenient for transformation processing.

Three Data Extraction methods:

- Full Extraction
- Partial Extraction- without update notification.
- Partial Extraction- with update notification.

Validations are done During Extraction:

- Reconcile records with the source data.
- Make sure that no spam/unwanted data loaded.
- Data type check.
- Remove all types of duplicate/fragmented data.
- Check even if all the keys are in place or not.

Some predefined structures involve:

- Flat files
- Dump files
- Redo and archive logs
- Transportable table spaces.

B. Transformation

Data transformation takes place in the staging layer. The main purpose of this step is to do some operations on the extracted data and make it a valid processed data and to give it to the loading step to load the data into the data warehouse. Reconcile records with the source data. Make sure that no spam/unwanted data loaded. Data type check. It removes all types of fragmented data and check whether all the keys are in place or not. Data which does not require any transformation is called as direct move or pass through data.

There are two ways to approach ETL transformation:

- Multistage data transformation.
- In-warehouse data transformation

Validations are done during this stage:

- Filtering - Select only particular columns to load.
- Using lookup tables and rules for data standardizations.
- Character Set Conversion and encoding Handling. Conversion of Measurement Units likes currency Conversion, Date Time conversions, numerical conversions, etc.
- Data threshold validation check.
- Data flow validation from staging area to intermediate tables.
- Required fields should not be blank.
- Cleaning.
- Split the columns into multiples and merging multiple columns into a single column.
- Transposing the columns and rows.
- Use lookups to merge the data.
- Using any complicated data validation.

Some other things that are done to the extracted data are:

- Filtering
- Cleaning
- Splitting
- Enriching
- Joining

There are some cases where data do not undergo transformation phase. In such case, those non-transformed data are called as rich data or pass through data. Some types of

transformations are aggregating transformations, joiner transformations and expression transformation etc

C. Load

Loading the data into the target data warehouse database is final step of the ETL process. In a classic Data warehouse, vast amount of data needed to be loaded in a relatively short period. Hence, load process should be optimized for performance. The extracted and transformed data is loaded in this final step. In this step the processed data are being loaded into the end target or the data warehouse as a flat file or in other file formats. This load stage of the ETL process depends on what you expect to do with the data once it's loaded into the data warehouse.

1) Types of Loading:

- *Initial Load* - occupying all the Data Warehouse tables
- *Incremental Load* - implementing ongoing changes as when needed systematically.
- *Full Refresh* - eliminating the contents of one or more tables and reloading with a new data.

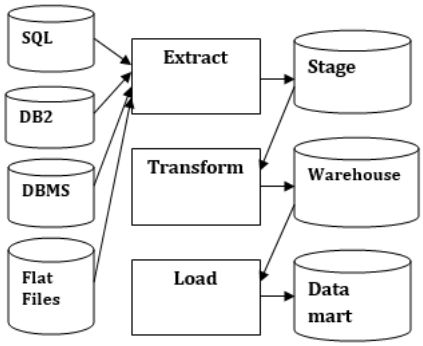

Fig. 2. ETL Stage

4. Data warehouse

Data warehousing is used for collection of data from different sources. It is constructed for integration of data from multiple sources that support analytical reporting, structured and decision making. It has data cleaning, data integration, and data consolidations.

A. Using data warehouse information

There are some resolving support technologies that make use of the available data in a data warehouse. They can able to gather some data, analyze the data and make decisions based on the information present in the warehouse. The information gathered in a data warehouse can also be used in any of the following domains.

B. Tuning production strategies

Tuning production strategies can be calibrate by altering the products and managing the product portfolios by comparing the sales quarterly or yearly.

C. Customer Analysis

It is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.

D. Operations Analysis

In operations analysis Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to evaluate business operations.

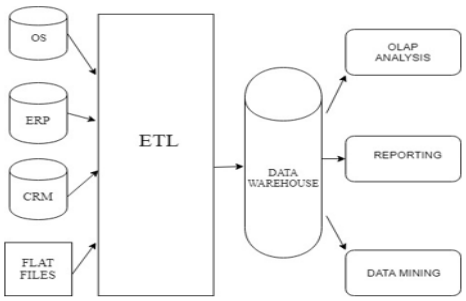

Fig. 3. Data warehouse architecture

The data warehousing uses the concept of OLAP (Online Analytical and Processing). Hence, it is needed for solving business problems like market analysis etc. Which requires query-centric database schemas that are array oriented and multidimensional thus, it could have the large collection of historical data that consist of transactional data from different sources? These transactional data is used for querying and reporting using OLAP techniques. OLAP tools are based on the multidimensional databases and allow a user to analyze the data in using elaborate, multidimensional, complex views. These tools provide an intuitive way to view corporate data. The operations of OLAP include roll up, roll down, slicing, dicing and pivot. The benefits of data warehousing includes locating the right information, presentation of information, testing of hypothesis, discovery of information and sharing the analysis. Data warehousing results in useful insights and helps in developing breakthrough ideas for reengineering processes.

5. Database design methodology

A multi-dimensional database (MDB) is one type of database which is used for data warehouses and online analytical processing applications. An OLAP application that accesses data from such database is known as a MOLAP (multi-dimensional OLAP) application. In this section, we describe the design of relational database schemas that reflect the multidimensional views of data. ER diagrams and normalization techniques are mostly used for the database design in OLTP projects. But the database designs suggested by ER diagrams are inappropriate for decision supports system where efficiency in querying and in loading data is significant.

Most data warehouses uses star schema to represent the Multi-dimensional data model the database consists of a single fact table and many numbers of dimension tables. Each record in the fact table includes a pointer (foreign key-that uses a generated key for efficiency) to each of the dimensions that provide multidimensional coordinates and stores numeric measures for those coordinates. Each dimension table consists

of columns that correspond to dimension's attributes. This figure represents the star schema with attributes.

A. *Star schema*

The star schema is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts. The star schema consists of one or more fact tables referencing any number of dimension tables. The star schema is a smooth style of data mart schema and is the preach which is widely used to develop data warehouses and dimensional data marts. The star schema consists of one or more fact tables preferring many numbers of dimension tables.

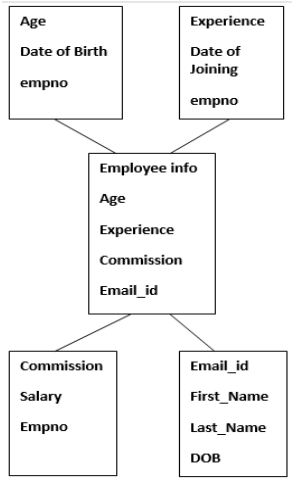

Fig. 4. Star Schema

A star schema database has various small numbers of tables and clear join paths, queries run faster than they did against an Online Transaction Process (OLTP) system. Small single-table queries, which are of dimension tables, are almost instantaneous. A star schema has some referential integrity building in when data is loaded. The main disadvantage of the star schema is that data integrity is not enforced since it is in a highly denormalized state. One-off inserts and updates can result in data anomalies in which normalized schemas are designed to avoid.

B. *Data flow diagram*

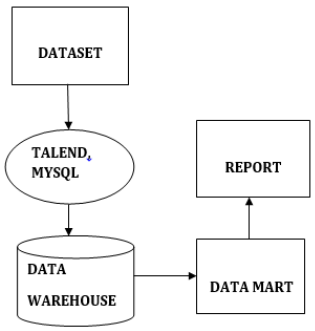

Fig. 5. Schema Chart

6. Research issues

The main problem with talend is that the feature which is used for scheduling is a basic one where it is available only with the enterprise editions and not with open studio distribution. Due to this issue, users working with talend open studio are not able to make use of the scheduling feature. Another drawback is that only users who have already subscribed to the real time big data package has the availability to use the spark streaming and machine learning in talend. In the management of data warehouse, the major issue is that the required data is not being captured by the source systems. But that data might be the essential one for data warehouse purpose. For example the date of registration for the property may be not used in source system but it may be very important for analysis purpose. The above issues require further investigation and must be resolved.

7. Conclusion

The dashboard creation for student non academics is to show the performance of the student in their extracurricular activities like number of papers presented, journals published, prizes won in sports events, etc. in a bar chart. This helps the students to enhance their talents in non-academic activities in case of low performance. This data integration project involves the data mining from various sources. It is difficult for a student to analyze their non-academics performance by themselves. Hence this project will be helpful in such scenarios. Enhancement of the student's participation in different events rather than academics will be the practical outcome.

References

[1] Sun Wei, Zhang Zhongneng, "ETL Architecture Research", Microcomputer Application, vol. 21, no. 3, pp 13-15, 2005.
[2] M. Vieira, H. Madeira, Detection of malicious transactions in DBMS, 2006.
[3] Boon Keong Seah, Nor Ezam Selan," Design and implementation of data warehouse with data model using survey-based services data", 20 October 2014.
[4] Wang Zhijuan, Wei Hongchang,Wu Xuefang," A Data Warehouse Design Method"31 December 2012.
[5] J. You, T.Dillon, J. Liu, "An integration of data mining and data warehousing for hierarchical multimedia information retrieval", 07 August 2002.
[6] Donglan Liu,Lei Ma,Xin Liu,Hao Yu, "Research on key issues of data integration technology in electric power system in big data environment", 21 December 2017.
[7] Ayad Hameed Mousa, Norshuhada Shiratuddin, "Data Warehouse and Data Virtualization Comparative Study", 12 September 2016.
[8] V. Arulkumar. "An Intelligent Technique for Uniquely Recognising Face and Finger Image Using Learning Vector Quantisation (LVQ)-based Template Key Generation," International Journal of Biomedical Engineering and Technology 26, no. 3/4 (February 2, 2018): 237-49.
[9] R. Arora, P. Pahwa, S. Bansal, "Alliance Rules of Data Warehouse Cleansing", IEEE International Conference on Signal Processing Systems Singapore, pp. 743-747, May 2009.
[10] A. Jeeva, C. Selvan, and A. Anitha, "Secure Token Based Storage System toPreserve the Sensitive Data Using ProxyRe-Encryption Technique", International Journal of Computer Science and Mobile Computing.
[11] https://help.talend.com/reader/wDRBNUuxk629sNcI0dNYaA/3Lyn4CR4M5Q2uOD8FWmOwg
[12] https://help.talend.com/reader/tXRG~nTonRYUwbOJscDgxw/~K8zTBzM7FdvgXmYD5XTmQ