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

Integration of Employee Details using Talend

S. Aruna[1], R. Bhavya[2], C. Rudhra[3], S. Monakatherine[4], G. Monica[5]
[1,2,3,4]*Student, Dept. of Computer Science and Engg., Sri Eshwar College of Engineering, Coimbatore, India*
[5]*Assistant Professor, 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, generating reports and related tools for support of those technologies, which are used to generate reports. The company is a global market leader in providing employee consultancy services across various regions in the globe. The company holds 10% of the salary of an employee as a commission. This project wants to upgrade the existing version 1.0(Beta) to an upgraded version 2.0(Lamda). 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. The company members can view data about a particular employee with reduced query time, thus generating reports.**

***Keywords*: data warehouse, talend tool, data integration, ETL.**

1. Introduction

The goal of our proposed system is to generate consolidated reports for data in the form of flat files that can be analyzed easily. This analysis can be used to understand the progress of employee details. At present, we are using Beta version for maintaining the employee information regarding individual details like salary information, etc. Here all the details are stored in the database in SQL format. There is no data marts to categorize the data present and to deliver the information as requested by the company in a consolidated manner. Information is one of the most factors to an organization success that is needed during decision making. Organizations typically deal with large volumes of data containing valuable information about employee details, stock details, administration details and others. 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.

to perform analysis and to generate a report for business and they are Pentaho 2) Talend. The Pentaho and Talend tool differ by the following properties:

A. Pentaho

- Job and transformation are the building block of Pentaho.
- Logs are produced at the transformation and job levels.
- Jobs are continued from where it is left off when an issue occurred.
- Jobs are executed directly from the spoon as well as pan.
- JavaScript code cannot be reused
- Pentaho stores the processed files as either xml file or it stores in a database.
- Runs on various platform like Windows, Linux, Unix.
- Pentaho analytics platform, Pentaho report generator are the product of pentaho.

B. Talend open studio

- Project, jobs, components are the building block of talend open studio. Logs are produced at the project level.
- Jobs are restarted from the beginning when an issue occur.
- Jobs are executed via tool interface and talend scheduler.
- Java code can be made visible and modifiable by the user.
- Talend open studio stores the file in the file level.
- It can run in various platform like windows, Linux.
- Big Data, MDM, Data quality are the product of talend open studio.

2. Database

A Database is a single repository that contains the collection of logically related and similar data. It contains the data that are organized in a way that can be easily accessed, managed, updated. The information that are necessary for the decision making process in the business are maintained in the database and those information will serve as many application as possible. Both retrieval and modification of data can be done on the database based on the operation performed. Some of the database components are:1) Character 2) Field 3) Record 4) File. Database Management System is a 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 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 Model is used to describe the structure of the database and it will also provide the definition and format of the data that is been stored in the database. 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 is 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 source and the relationship is represented by links. This model use tree structure to represent the records rather than arbitrary graph. The data representation in network database model is similar to hierarchical database model but here the link is used to represent the association between two records that is been stored in the database. In relational database model the data are organized as tables with rows and columns. An attribute is a unique name used to identify each column. Row in the table is used to represent the relationship between the set of values that is stored in the table. The database system architecture have different layers and they are: 1) Centralized and Client- Server system 2) Server System Architecture 3) Parallel System 4) Distributed System 5)Network Types. Centralized and Client-Server System will run on a single computer system and have a common bus to which number of controller are connected. In Server System Architecture the server will act as both transaction server and data server. The Parallel System database will have multiple processor and multiple disk are connected by an interconnection network. In Distributed System the data's are spread over multiple machines and those machines are interconnected by network. The Network Type database make use of two types of network like LAN (Local Area Network) and WAN (Wide Area Network).

Structured Query Language (SQL) is used to access the data that is 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 database are :1) reduce the duplication of data 2) allow of sharing of data by several users 3)data are 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 make use of simple queries to return the records as requested by the user and it also maintain 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 many small transactions with simple query, used for data entry, financial transaction, customer relationship management and retail sales. The database size of OLTP is 100 MB to 1 GB. 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 source and it is inserted into the database with the 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 is the process in data warehousing that deals with extracting data from different source systems and placing the processed data in the data warehouse. ETL is stands for Extract, Transform and Load. These three steps are important in data warehousing.

A. Extraction

Data Extraction is a process in data warehouse where the data from various heterogeneous or homogeneous sources are collected. During extraction the required data is been identified and gathered together and allowed to transform to get the desired data. The extraction is one of the step that consumes larger time than transformation and loading. There are several ways to perform extraction. Some are:

1) update notification
2) incremental extraction
3) full extraction

Update notification is an easiest way to collect data since if the source system provide notification about the changes modified in the data and this information in the notification can be updated. The next method is incremental extract which means an occurrence of an update is not been notified but the modified records could be identified and extract of those records could be retrieved. Full extract - some systems are not able to identify which data has been changed at all, so a full extract is the only way one can get the data out of the system. The full extract requires keeping a copy of the last extract in the same format in order to be able to identify changes. Full extract handles deletions as well. And other physical extraction methods are (1) Online Extraction and (2) Offline Extraction. In online extraction, data collection directly deals with the source system to access the data directly from the source system or connected to the intermediate system to access those preconfigured data. In offline extraction, the data is not directly taken from the source but it undergoes staging process. These data do have a predefined structure. Some of the predefined structures involve:

1) flat files
2) dump files
3) Redo and archive logs
4) 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. In this phase, the extracted data is cross checked for data quality. Some the data quality paradigms are whether:

1) The data is correct
2) The data is consistent
3) The data is complete

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

1) Filtering
2) Cleaning
3) Splitting
4) Enriching
5) Joining

There are some cases where data does 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 aggregate transformations, joiner transformations and expression transformation etc.

C. Load

The extracted and transformed data is loaded in this final step. In this step the processed data is been loaded into the end target or the data warehouse as a flat file or in other file formats. Based on the requirements of the organization the extracted data is loaded into the data warehouse or it is been extracted periodically like weekly or daily or monthly basis. To make the loaded data files efficient, the data is been indexed which will be much more efficient and easily understandable. These processed data that undergone extraction and transformation are loaded into the analytical database like OLAP. Some load processes physically insert each record as a new row into the target data warehouse's table using a SQL insert statement. While other load processes include a massive bulk insert of data utilizing a bulk load routine. The SQL insert is a slower routine for imports of data, but does allow for integrity checking with every record. The bulk load routine may be faster for loads of large amounts of data, but, does not allow for integrity check upon load of each individual record.

4. Talend open studio

Talend Open Studio is an open source graphical development environment for creating and deploying custom integrations between systems. It is an open architecture for data integration, data profiling, big data, cloud integration and more. It comes with over 600 pre-built connectors that make it quick and easy to connect databases, transfer file, load data, move, copy and rename files. It allows each component to define complex processes. It provides a comprehensive suite of open source (and commercial) integration products. This includes.,

- Data Integration (ETL, ELT)
- Data Quality
- Master Data Management (MDM)
- Enterprise Service Bus (ESB)
- Business Process Management (BPM)
- Big Data

Benefits of this solution:

- Business modeling
- Graphical development
- Metadata-driven design and execution
- Real-time debugging
- Robust execution

Provided as a packaged, out of the box, ready-to-install platform, Talend Open Studio meet data integration requirements of all organizations–irrespective of their size or level of data integration expertise.

A. Talend data integration

Talend data integration software tool has an open, scalable architecture. It allows faster response to business requests. Talend tool is used to develop and deploy data integration process faster than manual coding. we can easily integrate all the data with other data warehouse or synchronize data between systems. Data integration combines data extracted from different sources and provides users with a unified view of those data. It is used to manage various ETL jobs and empower users with simple, self-service data preparation.

5. Data warehouse

Data warehouse is basically the relational database hosted on cloud or an enterprise mainframe server. It collects data from varied, heterogeneous sources for data consolidation, analysis and reporting at different aggregate levels thereby supporting decision making for business users and provides business insights.

The data warehouse environment consists of data store, data mart and the metadata. The main function of data store is to feed the data into data warehouse for the purpose of business analysis.

A. Data Mart

It is a subset of data warehouse where the data can be accessed quickly with less processing time. Metadata contains the information about the warehouse rather than the information kept within the warehouse.

Data warehouse has two approaches. They are

- Top down approach
- Bottom up approach

In bottom up approach, initially the data marts are created and are integrated to form a data warehouse, which is been used for reporting.

Whereas in top down design approach, a data warehouse is built initially and from that the data marts are build.

The key factors to develop a data warehouse are

1) Scope of the data warehouse
2) Data redundancy and
3) Type of end- user.

The architecture is made up of following interconnected parts. They are source system layer, source data transport layer, data quality control and data profiling layer, metadata management layer, data integration layer, data processing layer and end user reporting layer. Thus the architecture enhances availability of business intelligence data and also improves the

effectiveness of decision making.

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. Transactional data are used for querying and reporting using OLAP techniques. OLAP tools are based on the multidimensional database.

6. Database design methodology

A multidimensional database(MDB) is a type of database which is optimized for data warehouse and online analytical processing (OLAP) applications. An OLAP application that accesses data from such database is known as a MOLAP (multidimensional 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 support systems where efficiency in querying and in loading data (including incremental loads) are significant.

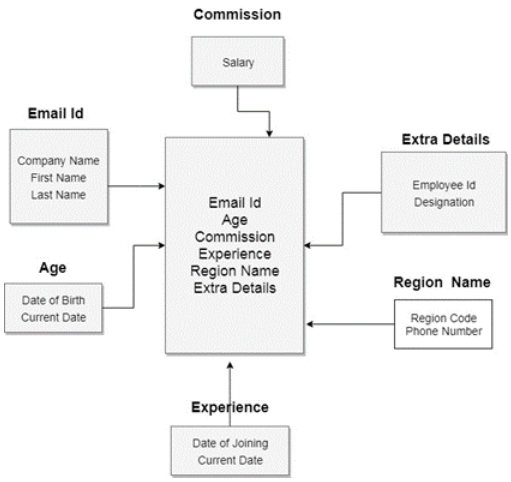

Fig. 1. Fact Table

The star schema is the simplest way of data mart schema and it is most widely used to develop dimensional data marts and data warehouse. The star schema consists of more number of fact tables referencing any number of dimension tables.

Star schema database has a small number of tables and clear join paths, queries run faster than they do against an OLTP system. Small single-table queries, usually of dimension tables, are almost instantaneous. A star schema has referential integrity built in when data is loaded.

The main disadvantage of the star schema is that data integrity is not enforced as well since it is in a highly de-normalized state. One-off inserts and updates can result in data anomalies which normalized schemas are designed.

7. Experiment and results

The table creation for employee is to view the details of the employee in the company and to auto generate the details like age, email id, commission, experience, etc. from the details given by the employee. This helps the company to make the process easy.

This data integration project involves the auto generation of salary of an employee and holding 10 percentage of their salary as commission. Before, the implementation of this project the organization found it difficult to generate the salary. Hence this project will be helpful in such scenarios.

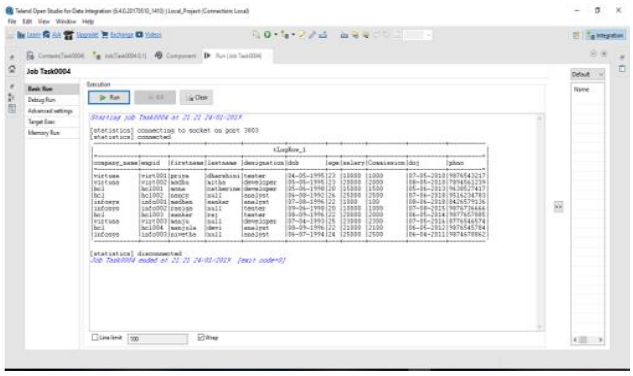

Fig. 2. Result

8. Conclusion

This paper presented the implementation of integration of employee details using talend.

References

[1] https://help.talend.com
[2] www.talend.com
[3] https://community.talend.com
[4] https://www.talend.com
[5] https://www.1keydata.com/datawarehousing/datawarehouse.html
[6] https://bekwam.blogspot.com
[7] https://www.dataintegration.info
[8] https://www.tutorialspoint.com
[9] https://intellipaat.com
[10] https://www.guru99.com