Survey on Data Warehousing Strategies for Integrating Business Data

Mrs. Vijayshri D. Vaidya[1], Mr. Abhishek A. Nibe[2], Mr. Avinash B. Anap[3], Mr. Vasimraj S. Tamboli[4], Mr. Ravindra S. Kakade[5]

[1]*Lecturer, Dept. of Computer Technology Dr. Vitthalrao Vikhe Patil Polytechnic, Loni Maharashtra*
[2]*Lecturer, Dept. of Computer Technology Dr. Vitthalrao Vikhe Patil Polytechnic, Loni Maharashtra*
[3]*Lecturer, Dept. of Computer Technology Dr. Vitthalrao Vikhe Patil Polytechnic, Loni Maharashtra*
[4]*Lecturer, Dept. of Computer Technology Dr. Vitthalrao Vikhe Patil Polytechnic, Loni Maharashtra*
[5]*Lecturer, Dept. of Computer Technology Dr. Vitthalrao Vikhe Patil Polytechnic, Loni Maharashtra*

ABSTRACT

A Data Warehousing (DW) is a process for collecting and managing business data from varied sources to provide meaningful business insights. Data warehouse is an information system that contains historical and commutative data from single or multiple sources. A Data warehouse is typically used to connect and analyze business data from heterogeneous (different by nature) sources. It includes different characteristics like subject-oriented, integrated, time-variant and non-volatile. It simplifies reporting and analysis process of the organization. It is also a single version of truth for any company for decision making and forecasting.

Keywords: *business data, historical, heterogeneous sources, integrated, non-volatile*

1. INTRODUCTION

There is continuous expansion of business and business constantly upgrading with technologies. So, it is required to integrate varied sources of business data for dealing with complex business information environment.

A Data Warehousing (DW)[1] is process is used for collecting and managing business data from varied sources to provide meaningful business insights.

Data warehouse is an information system that contains historical and commutative data from single or multiple sources, which will be helpful for business expansions.

2. CHARACTERISTICS OF DATA WAREHOUSING

2.1. Subject-Oriented:

A data warehouse is subject oriented as it offers information regarding a theme instead of companies ongoing operations. These subjects can be sales, marketing, distributions, etc.

A data warehouse never focuses on the ongoing operations. It concentrates on modelling and analysis of data for decision making**.**

It also provides a simple and concise (summarized) view around the specific subject by excluding data which is not helpful to support the decision process.

2.2. Integrated:

In Data Warehouse, integration means the establishment of a common unit of measure for all similar data from the dissimilar database.[4]

A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. and convert all data in single standard format.
This integration helps in effective analysis of data.
Example: There are three different applications labelled A, B and C.

In Application A gender field store logical values like M or F.
In Application B gender field is a numerical value.
In Application C application, gender field stored in the form of a character value.

However, after transformation and cleaning process all this data is stored in common format in the Data Warehouse.

2.3. Time-Variant:

The data collected in a data warehouse with a particular period (time) and offers information from the historical point of view.
Another aspect of time variance is that once data is inserted in the warehouse, it can't be updated or changed.

2.4. Non-volatile:

Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it. Data is read-only and periodically refreshed.
This also helps to analyse historical data and understand what & when happened. Activities like delete, update, and insert which are performed in an operational application environment are omitted in Data warehouse environment.
Only two types of data operations performed in the Data Warehousing are Data loading and Data access.

3. DATA WAREHOUSE ARCHITECTURE

Three-tier architecture of data warehouse is the most widely used architecture. It consists of the Top, Middle and Bottom Tier.

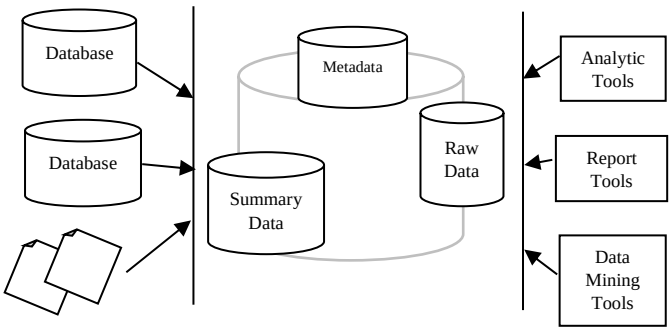

Fig 1. Three-Tiered Data Warehouse Architecture

3.1 Bottom Tier:

The database or data sources of the Datawarehouse servers is the bottom tier.
It is usually a relational database system.

3.2 Middle Tier:

The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model.
This application tier presents an abstracted view of the database.
This layer also acts as a mediator between the end-user and the database.

3.3 Top-Tier:

The top tier is a front-end client layer.

Top tier is the tools and API that user used to get data out from the data warehouse.

The different tools are Query tools, reporting tools, managed query tools, Analysis tools and Data mining tools.

4. ETL PROCESS IN DATA WAREHOUSE

ETL[3][6] is a process in Data Warehousing and it stands for Extract**,** Transform and Load**.** It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then finally, loads it into the Data Warehouse system.

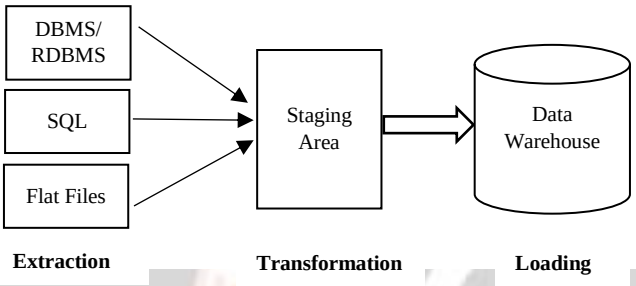

Fig 2. ETL process in Data Warehousing

4.1 Extraction:

The first step of the ETL process is extraction.

In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML and flat files into the staging area.

The data cannot be loaded in data warehouse; therefore, this is one of the most important steps of ETL process.

4.2 Transformation:

The second step of the ETL [6] process is transformation.

In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format.

It may involve following processes/tasks:

Filtering – loading only certain attributes into the data warehouse.

Cleaning – filling up the NULL values and missing values.

Joining – joining multiple attributes into one.

Splitting – splitting a single attribute into multiple attributes.

Sorting – sorting tuples on the basis of some attribute (generally key-attribute).

4.3 Loading:

The third and final step of the ETL process is loading.

In this step, the transformed data is finally loaded into the data warehouse.

5. DATA WAREHOUSE MODELS

5.1 Enterprise Data Warehouse (EDW):

Enterprise Data Warehouse is a centralized warehouse,[5] which aggregates the information automatically.

It offers a unified approach for organizing and representing data.
It also provides the ability to classify data according to the subject and give access accordingly to users.
It provides decision support service across the enterprise.

5.2 Data Marts:

A data mart[2] is a subset of the data warehouse.
It is specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.
Due to large amount of data, a single warehouse can become overburdened.
So, to prevent the warehouse from becoming impossible to navigate, subdivisions created, called as Data Marts.
These data marts divide the information saved in the warehouse into categories or specific groups of users.
In a simple word Data mart is a subsidiary of a data warehouse.

5.3 Virtual Warehouse:

A virtual warehouse is essentially a separate business database, which contains only required data for operation system.
The data found in a virtual warehouse is usually copied from multiple sources throughout an operation system.
Virtual warehouse is used to search the data quickly and without accessing the entire system.
It speeds up the overall access process.

ADVANTAGES OF A DATA WAREHOUSE

1. Delivers enhanced business intelligence
By having access to information from various sources in a single platform, decision makers will no longer need to rely on limited data.

2. Saves times
A data warehouse standardizes, preserves, and stores data from different sources, and integration of all the data in one place.
So, all critical data is available to all users simultaneously.

3. Enhances data quality and consistency
A data warehouse converts data from multiple sources into a consistent format.
The data from different sources can be filtered, sorted, cleaned. This will lead to more accurate data, which will become the basis for solid decisions.

4. Generates a high Return on Investment (ROI)
Companies experience higher revenues and cost savings than those that haven't invested in a data warehouse.

5. Provides competitive advantage
Data warehouses helps to get a holistic (as a whole not parts) view of their current standing and evaluate opportunities and risks, thus providing companies with a competitive advantage.

6. Improves the decision-making process
Data warehousing provides better insights (detailed understanding) to decision makers by maintaining a related database of current and historical data.

7. Enables organizations to forecast with confidence
With advanced features of Data warehouse, organization can forecast their line of action easily.

8. Streamlines (well organized) the flow of information

Data warehousing facilitates the flow of information through all related or non-related parties.

APPLICATIONS OF DATA WAREHOUSE

1. Airline:
Analysis of crew assignments, flight data, flight routes, fairs.

2. Banking:
Analysis of customer data, transactions, loans, accounts, KYC.

3. Healthcare:
Generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.
4. Public sector:
In the public sector, data warehouse is used for intelligence gathering. It helps government agencies to maintain and analyze tax records, health policy records, for every individual.

5. Investment and Insurance sector:
In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.

6. Retail chain:
In retail chains, Data warehouse is widely used for distribution and marketing. It also helps to track items, customer buying pattern, promotions and also used for determining pricing policy.

7. Telecommunication:
A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.

8. Hospitality Industry:
This Industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

CONCLUSION

Data warehouse can integrate the data required for business analysis from various sources using ETL process and creates different data warehouse models for quick and easy decision making in business areas.

REFERENCES

[1] Jarke, M., M. Lenzerini, Y. Vaasssiliou, P.Vassiliadis, "Fundamentals of Data Warehouse," 2000.

[2] R. Kimball, L. Reeves, M. Ross and W. Thornthwaite. " The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses", John Wiley & Sons, 1998.

[3] R. Kimball and J. Caserta. "The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data", John Wiley & Sons, 2004.

[4] T. Manjunath, S. Ravindra, and G. Ravikumar, “Analysis of data quality aspects in data warehouse systems,” International Journal of Computer Science and Information Technologies, Vol. 2, No. 1, 2010, pp. 477485.

[5] B. Pinar, A Comparison of Data Warehouse Design Models, Master Thesis, Atilim University, Jan. 2005.

[6] W. Eckerson and C. White, "Evaluating ETL and Data Integration Platforms", TDWI REPORT SERIES, 101communications LLC, 2003