Data Warehouse – Design Architecture, Applications and Security

Name of Author – Jay Prakash Soja
Designation – Assistant Professor
Name of Department – Department of Computer Engineering /Applications
Name of organization – Ambedkar Institute of Technology, Shakarpur, Delhi (India)

Abstract- Due to rapid change and vast use of Information Technology in the world, the heterogeneous data is being received and processed on large scale for various analytical and operational purposes by various organizations to meet the demand of end users. So it has become important for the organizations to store the data on central place and to use it for various purposes. The term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization. A data warehouses provides us generalized and consolidated data in multidimensional view. A data warehouses also provides us Online Analytical Processing (OLAP) tools. These tools help us in interactive and effective analysis of data in a multidimensional space. Subject-Oriented, Integrated, Time-variant and Non-volatile are some characteristics of data warehouse.

Keywords – Matadata, OLAP, OLTP, ETL, Data Mart,

I. INTRODUCTION

The term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization. Online Transaction Processing (OLTP) captures, stores, and processes data from transactions in real time. Online Analytical Processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems. Data warehouse is the example of OLAP. The data warehouse involves historical processing of information while operational database involves day-to-day processing. OLAP systems are used by knowledge workers such as executives, managers, and analysts but OLTP systems are used by clerks, DBAs, or database professionals. ETL stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse.

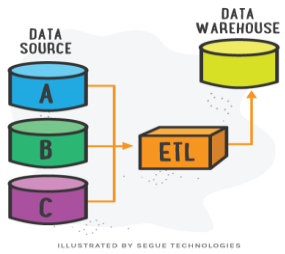


Fig-1 Formation of Data Warehouse

E(Extracted): Data is extracted from External data source.
T(Transform): Data is transformed into the standard format.
L(Load): Data is loaded into data warehouse after transforming it into the standard format.

Functions of Data Warehouse Tools and Utilities

Data Extraction − Involves gathering data from multiple heterogeneous sources.
Data Cleaning − Involves finding and correcting the errors in data.
Data Transformation − Involves converting the data from legacy format to warehouse format.
Data Loading − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
Refreshing − Involves updating from data sources to warehouse.

II Data Warehouse Architecture

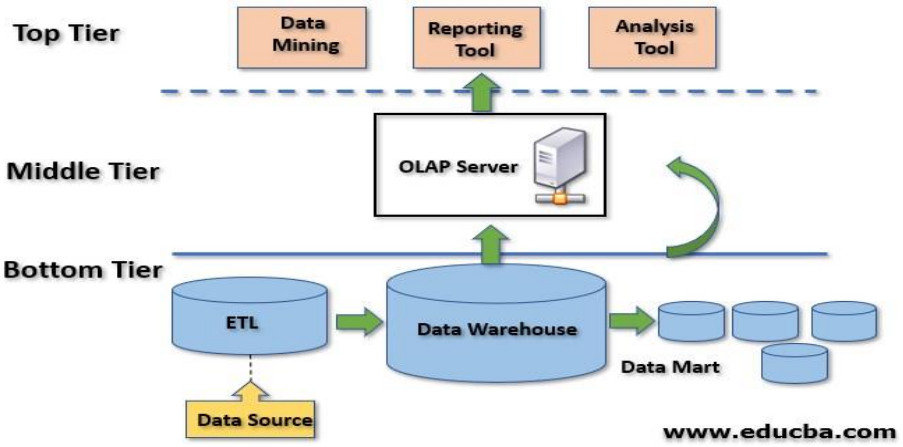


Fig-2 Three-Tier Data Warehouse Architecture

Bottom Tier (Data Warehouse Server) - It consists of Data Warehouse, Data marts (sub sets of data warehouse).
Middle Tier (OLAP Server) - It consists of an OLAP server for fast querying of the data warehouse.
Top Tier (Front end Tools)- It contains front-end tools for displaying results provided by OLAP.

III Data Warehouse Design - A data-warehouse is a heterogeneous collection of different data sources organised under a unified schema. There are 2 approaches for constructing data-warehouse – Top down and Bottom Up.

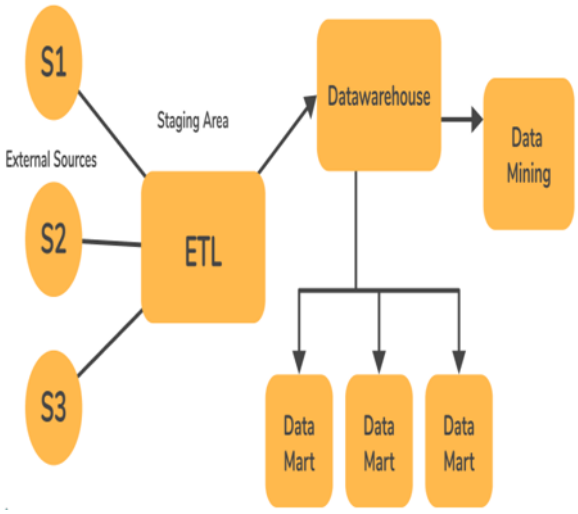


Fig-3 **Top Down Approach**

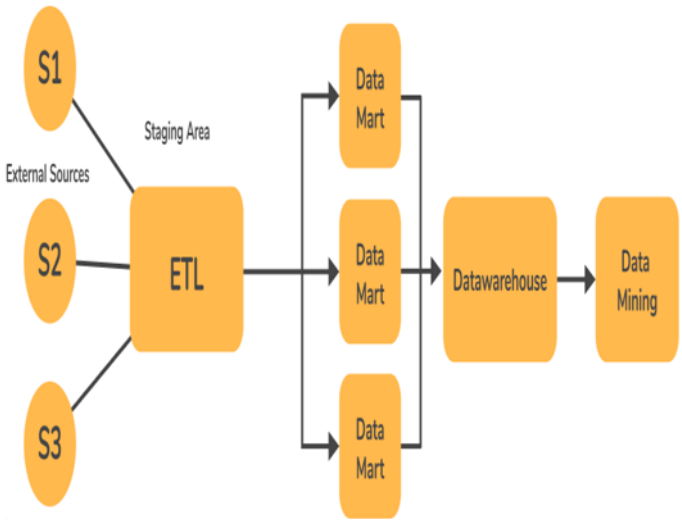


Fig-4 **Bottom Up Approach**

The Top-down approach stores the data first in the data warehouse, followed by the data marts. On the other hand, The Bottom-Up approach stores the data first in data marts followed by the data warehouse. The top-down approach is best used when - the problem is complex and needs to be broken down into smaller, manageable parts.
External Sources – External source is a source from where data is collected irrespective of the type of data. Data can be structured, semi structured and unstructured as well.
Stage Area – Since the data, extracted from the external sources does not follow a particular format, so there is a need to validate this data to load into data warehouse. For this purpose, it is recommended to use **ETL** tool.
Data Marts – Data mart is also a part of storage component. It stores the information of a particular function of an organisation which is handled by single authority. There can be as many number of data marts in an organisation depending upon the functions.

Data Mining –
The practice of analysing the big data present in data warehouse is data mining. It is used to find the hidden patterns that are present in the database or in data warehouse with the help of algorithm of data mining.

IV Applications of Data Warehouse - Every organization, no matter in what industry it works in or how big or small it is, requires a data warehouse to connect its disparate sources for anticipating, analysis, reporting, business intelligence, and facilitating robust decision-making. These services are also required at a reasonable cost and optimal value. Some major applications of Data warehouse are as follows-
E-commerce: E-commerce platforms need to gather key marketing metrics from marketing tools and use that to approach their customers in a better way. This is where data warehouses help. Replicating data, tracking & visualizing Key Performance Indicators (KPIs) such as conversion rates, churn rates, and return on ad spends, safe storage, etc. help companies perform better. In recent times, amazon redshift is the most popular warehouse being used for marketing analytics, because of its user-friendly UI and flexibility.

Retail: Data warehouses can be used by retailers to easily identify products with high demand and the fastest selling demand. The data can then be used to react to a rise or fall in consumer demand quickly, which can ultimately be used to gain a competitive advantage. Reverse ETL is a popular concept that leverages data from warehouses and helps target audiences better. They are the mediators between wholesalers and end customers, and that's why it is necessary for them to maintain the records of both parties. For helping them store data in an organized manner, the application of data warehousing comes into the frame.

AI/ML: With many companies embracing AI for their data journey, it's critical to get a reliable data warehouse now. AI enables data maturity, which is intertwined with the flexibility, scalability, and agility that a warehouse offers. On the other hand, machine learning is used on the data after the data has been replicated and transformed in the warehouse, to help newer business models emerge and advance digital disruption.

Agritech: Data storage is a must when it comes to the new age of farming. Data related to crop yield, weather conditions, pesticides, crop inventory and so much more demands a data warehouse. With advanced analytics, engineers and business analysts are able to figure out inefficiencies in the ecosystem, such as problems in the soil quality, unnecessary use of pesticides, etc. and iron them out.

Sustainability and climate action: Climate data requires a versatile data infrastructure, with cloud-first models for warehouses. To bring out sustainability insights, the data architecture must be able to integrate raw data from multiple sources and make it easy for end-users for making predictions and effective decisions related to climate change.

Manufacturing & Supply chain: Being one of the industries involving a higher number of intermediaries, the supply chain industry needs data warehouses to limit the number of data silos and ultimately human error. Data warehouses can help in inventory management (which items are low in count and what is the cost of each step in the life cycle), all the data related to vendors, logistics (for example: timestamp data related to product delivery), and ultimately serving the customer better.

Healthcare: With constant advancement in healthcare, the data captured by machines is huge. To digitally improve hospital infrastructure, reduce wait time, and make processes more efficient, data warehouses are making data work constantly in this field. Getting personalized healthcare can be possible with a single platform (such as having one place for all diagnostics, tests, prescriptions, and follow-ups). All the clinical, financial, and employee data are stored in the warehouse, and analysis is run to derive valuable insights to strategize resources in the best way possible.

Banking & Finance: Data security is critical for the BFSI sector, and data warehouses solve that problem by vouching for industry-standard security compliances. The warehouses can be used to get updates about customer deposits, loans, funds, deposits, etc., and a better understanding of the performance of different branches. The right solution helps the financing industry analyze customer expenses that enable them to outline better strategies to maximize profits at both ends.

Financial Auditing: With access to real-time financial data, warehouses ensure decisions related to the business's current financial performance can be reached quickly. Data warehouses enable the collection of data on a daily basis and information can then be regularly used to identify any discrepancies in financial reporting & audits.

Pharmaceuticals: As data warehouses make data more accessible, it's now being used for making better strategic decisions and identifying & developing customer buying trends in pharmaceuticals. This results in better customer targeting, pre-call analysis as well as post-call assessments, helping the pharma industry at scale.

(V) SECURITY ISSUES

Data security focuses mainly on three issues: confidentiality, integrity, and availability. The objective of a data warehouse is to make large amounts of data easily accessible to the users, hence allowing the users to extract information about the business as a whole. But we know that there could be some security restrictions applied on the data that can be an obstacle for accessing the information. If the analyst has a restricted view of data, then it is impossible to capture a complete picture of the trends within the business. We should consider the following possibilities during the design phase -

a) Whether the new data sources will require new security and/or audit restrictions to be implemented?
b) Whether the new users added who have restricted access to data that is already generally available?

The following activities get affected by security measures −

(i) User access
(ii) Data load
(iii) Data movement
(iv) Query generation

User Access

We need to first classify the data and then classify the users on the basis of the data they can access. In other words, the users are classified according to the data they can access.

Data Classification

The following two approaches can be used to classify the data −

(i) Data can be classified according to its sensitivity. Highly-sensitive data is classified as highly restricted and less-sensitive data is classified as less restrictive.

(ii) Data can also be classified according to the job function. This restriction allows only specific users to view particular data. Here we restrict the users to view only that part of the data in which they are interested and are responsible for.

User classification

The following approaches can be used to classify the users −

(i) Users can be classified as per the hierarchy of users in an organization, i.e., users can be classified by departments, sections, groups, and so on.

(ii) Users can also be classified according to their role, with people grouped across departments based on their role.

Classification on the basis of Department

Let's have an example of a data warehouse where the users are from sales and marketing department. We can have security by top-to-down company view, with access centered on the different departments. But there could be some restrictions on users at different levels. This structure is shown in the Fig - 5

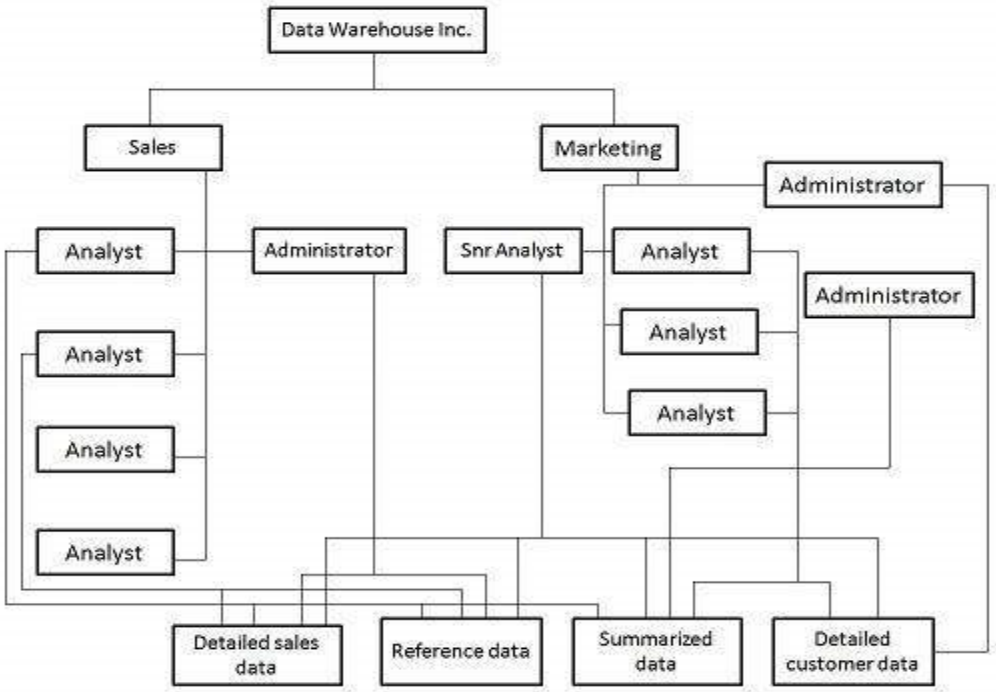
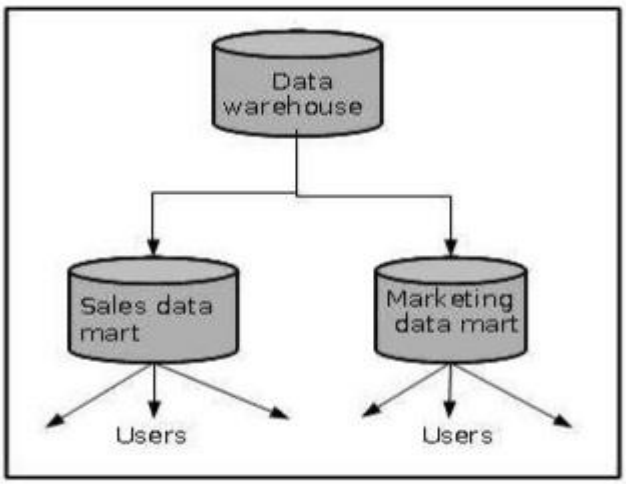


Fig-5 Classification based on the department

But if each department accesses different data, then we should design the security access for each department separately. This can be achieved by departmental data marts. Since these data marts are separated from the data warehouse, we can enforce separate security restrictions on each data mart. This approach is shown in the following Fig-6.



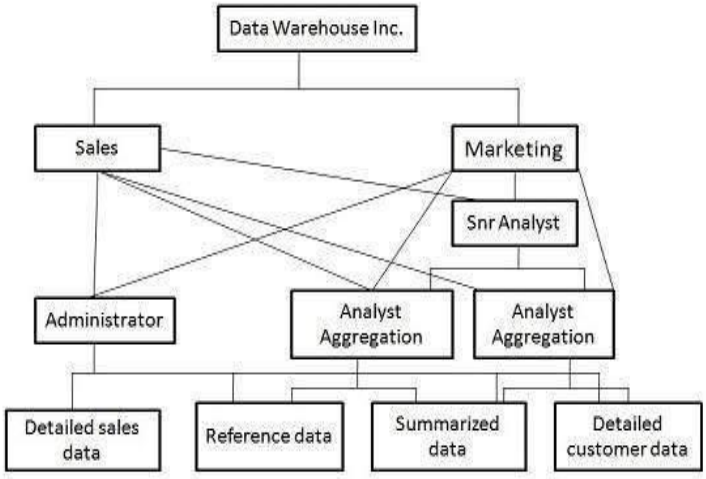


Fig-6 Classification based on different data in each deptt **Fig-7 Classification based on Role**

Classification Based on Role

If the data is generally available to all the departments, then it is useful to follow the role access hierarchy. In other words, if the data is generally accessed by all the departments, then apply security restrictions as per the role of the user. The role access hierarchy is shown in the Fig-7.

Audit Requirements

Auditing is a subset of security, a costly activity. Auditing can cause heavy overheads on the system. To complete an audit in time, we require more hardware and therefore, it is recommended that wherever possible, auditing should be switched off. Audit requirements can be categorized as follows −

a) Connections
b) Disconnections
c) Data access
d) Data change

Network Requirements

Network security is as important as other securities. We cannot ignore the network security requirement. We need to consider the following issues −

[1] Is it necessary to encrypt data before transferring it to the data warehouse?

[2] Are there restrictions on which network routes the data can take?

These restrictions need to be considered carefully. Following are the points to remember −

a) The process of encryption and decryption will increase overheads. It would require more processing power and processing time.

b) The cost of encryption can be high if the system is already a loaded system because the encryption is borne by the source system.

Documentation

The audit and security requirements need to be properly documented. This will be treated as a part of justification. This document can contain all the information gathered from −

a. Data classification
b. User classification
c. Network requirements
d. Data movement and storage requirements
e. All auditable actions

Conclusion

Databases play a critical role in almost all areas where computers are used. Today, there are many challenges in the data mining system. A great example of data warehousing that everyone can relate to is what Facebook does. Data mining is widely used in fraud detection contexts, as an aid in marketing campaigns, and even supermarkets use it. Data warehouse provides us generalized and consolidated data in a multidimensional view. Several types of analytical software are available: statistical, machine learning, and neural networks.

REFERENCES

[1] The Data Warehouse Toolkit by Ralf Kimbell
[2] Agile Data Warehouse design by Jim Stagnitto
[3] Building the data warehouse by Bill Inmon
[4] Data Mining and Data Warehouse by Prateek Bhatia
[5] Data Warehousing Fundamentals for IT Professionals by Paulraj Ponniah
[6] Hacking for Beginners by Ramon Nastase
[7] The Modern Data Warehouse in Azure by Matt How
[8] Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance by Christopher Adamson