DATABASES AND DATA WAREHOUSES

Abstract

A database is generally referred as a collection of data organized in a formatted structure for storage, accessibility and retrieval. There are various types of databases which include XML, CSV files, flat files and spreadsheets etc., the main two categories of databases includes data from OLTP (Online Transactional Processing) and OLAP (Online Analytical Processing). Both systems store data in the form of tables, columns, indexes, keys, views and data types. MySQL, Oracle and NoSQL are some query languages used to query the database.

In the other side data warehouses are also the types of database that integrates data from multiple sources and provisions the data for analytical process. Data warehouse is an OLAP database. Data warehouses are mainly designed to enable and support business intelligence mainly data analytics. Data warehouses are often consists of huge volume of historical data. As a whole data warehouse collects data from different sources and stores it into a single repository for analytics and decision making. This chapter presents an overview about databases, types of databases, data warehouses and the differences between databases and data warehouses.

Keywords: Database, OLTP, data warehouse, XML.

Author

Dr. S. Ramalakshmi
Assistant Professor
Department of Computer Science and Applications
Don Bosco College (Arts and Science)
Karaikal, Puducherry.

Copyright © 2024 Authors

I. AN INTRODUCTION TO DATABASES

1. **Data:** Data can be any kind of information or media that is being transferred from one person to another. All the communications are in the form of data. In computers data can be used in many forms like raw text, numbers, characters, images etc. Data may be used as variables in a programming process. In its general sense, data refers to an aggregation of discrete or continuous values that serves as a conduit for conveying information. These values encompass a spectrum of attributes, delineating factors such as quantity, quality, factual representation, statistical metrics, and other diverse entities. Alternatively, data can manifest as symbolic sequences, carrying the potential for subsequent elucidation and interpretation. Within this overarching compilation, the term "datum" designates an elemental entity—a solitary value—within the ensemble of data, signifying a distinct unit of significance. In all the possible communication data is the most crucial part of the computer world.

2. **Database:** A database denotes a meticulously arranged assemblage of structured information or data, typically residing in digital form within a computerized framework. The administration and oversight of a database are orchestrated by a Database Management System (DBMS), which assumes the role of control and management. Many dynamic websites on the internet today are stored in databases [5]. A database is a formatted collection of data and a set of program to access the data example for this kind of programs are Mysql and Oracle. Databases can store data in the form of tables depending upon the type of database. The primary goal of the database is to store a huge amount of data.

   In general databases are used to store information about people, such as customers or users. An operational database system will store much of the data organized and allowing user to access the data [6]. For example if it is an E-Commerce application then the data we access and store in a database system includes customer data, business data and relationship data.

II. TYPES OF DATABASES

The various types of databases provide different functionality to the users. Since the data is a dynamic one, the way it stored also varies a lot. This is the reason that companies designing their own types of databases that satisfy with their requirements.

1. **Hierarchical Databases:** It is developed in 1960’s and it is looks similar to a tree structure. A single node “parent” has one or more objects under it and has no child have more than one parent. This hierarchical database offers high performance because of its easy access and quick query time.

2. **Relational Databases:** Relational databases were designed in 1970. It uses SQL for operations like create, read, update and delete data. In this type of database the data is stored in discrete tables and it can be joined together by using foreign keys.

Copyright © 2024 Authors

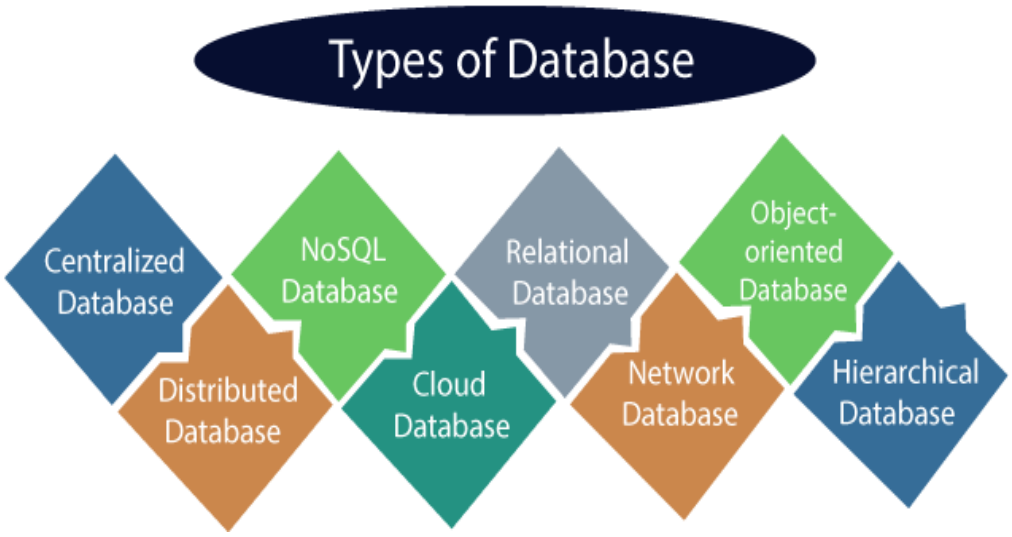

Figure 1: Types of Databases [1]

3. **Non Relational Databases:** These are commonly referred to as NoSQL databases. These databases are complex because of modern web application.

4. **Centralized Databases:** In this type of database that stores data at a centralized database system. It allows users to access the stored data from different locations through several applications. This type of databases contains the authentication process to let user's access data securely.

5. **Distributed Databases:** In distributed OS the data is distributed among different database systems of an organization. These DB systems are connected via communication links, by using this links the end user can access the data easily. The two types of distributed data bases are,
   - Homogeneous Databases
   - Heterogeneous Databases

6. **Cloud Databases:** Here the data is stored in a virtual environment and executes over the cloud computing platform. It provides users with various cloud computing services for accessing the database. These are varieties of cloud platform some among them are,
   > AWS
   > MS Azure
   > Google cloud SQL> Science Soft

7. **Network Databases:** It is the database that typically follows data model. Here the representation of data is in the form of nodes connected via links between them.

8. **Object Oriented Databases:** Within these databases, objects are stored and supervised on the storage disk of a dedicated database server. What sets these databases apart is their distinctive ability to maintain enduring associations between these objects, aligned with the principles of object-oriented programming – a prevalent and widely embraced programming paradigm.

Copyright © 2024 Authors

9. **NoSQL Databases:** It is an alternative system to traditional SQL databases. A NoSQL database uses a data model that has a different structure than the rows and columns table structure used with RDBMS [7].

10. **Database Software:** Software that is used to create, maintain and edit database files is called as database software. This software is also handles data storage, backup, reporting and security mechanisms

    The database software also enables easier record or file creation, data entry, editing the data, updating and reporting.

    Database software enables the data management in a simpler form by allowing users to save their data in a structured format. GUI (Graphical User Interface) is used to create and manage data. For example MS Access software permits users to create, maintain and edit the database using its GUI controls and other features without the need of queries. Users can also create their customized databases by using database software.

11. **Database Management System:** Databases requires software known as DBMS, this software serves as an interface between the database and its end users. This management system allows the user to store and retrieve the data and managing the structure of the database. Database management software also manages the security and access controls of the database.

12. **Important features of the DBMS**
    - Data integrity and security
    - Concurrency control
    - Data modeling
    - Data storage & retrieval
    - Backup and recovery

13. **Database languages:** The Database Management System has proper languages to deal with queries and updates, it can be used to read, store and update the data in the database. There are four types of data languages that are listed below.
    - Data Definition Language
    - Data Manipulation Language
    - Data Control Language
    - Transactional Control Language

14. **Challenges:** In recent days databases need to support for complex queries and there is a need for immediate responses so as a result a wide variety of methods must be employed to improve the performance of the database. Some common challenges databases faces are,

    - **Safeguarding Data Integrity:** Guaranteeing the protection and confidentiality of data.
    - **Meeting Growing Requirements:** Adapting to increasing needs and requests.
    - **Sustaining the Framework of the Database:** Managing and up keeping the underlying structure of the database.

Copyright © 2024 Authors

- **Expanding Scalability Boundaries:** Eliminating constraints on the system's capacity to scale.
- **Addressing Surge in Data Magnitude:** Handling substantial surges in the volume of data.
- **Ensuring Compliance with Data Regulations:** Securing adherence to regulations regarding the location and storage of data.

III. AN INTRODUCTION TO DATA WAREHOUSES

A data warehouse is an integrated and centralized repository of data from various sources. Its core objective is to support for business intelligence, data analytics, and decision making and reporting. Data warehouses play a main role in maintaining a large volume of data to gain valuable insights.

1. Characteristics of Data warehouses

- **Subject Oriented:** Data placed in the warehouse are organized in a specific subjects or business areas instead of a particular transaction data. The major subject areas include sales, customer data, inventory, finance and marketing.

- **Time Variant:** Data warehouses stores historical data and it allowing users to analyze the trends and the continuous changes in the business environment. This nature helps to identify the patterns and makes decision based on historical views.

- **Data Integration:** Data warehouses collect data from different sources like operational systems, spread sheets, flat files and more so the integration process ensures that information from various systems can be combined and analyzed together.

- **Non Volatile:** Once the data warehouse is loaded with the data it becomes a historical record and it is not altered and updated directly. It provides consistency in data and a reliable environment for analysis.

- **Data Transformation and Loading:** The process involves in data warehouses are Extract, Transformation and Loading (ETL). The data is extracted from various sources then it is transformed to the warehouse data model and at the end cleaned to achieve the accuracy.

- **OLAP (Online Analytical Processing):** Data warehouses always uses OLAP cubes or multidimensional databases for fast data analysis. This helps users to explore data from different dimensions easily.

- **Data Marts:** It is a subset of data warehouses mainly designed to achieve specific business units. It focuses on particular subject area and provides a more specialized view of data to support the need of a specific group of people.

Copyright © 2024 Authors

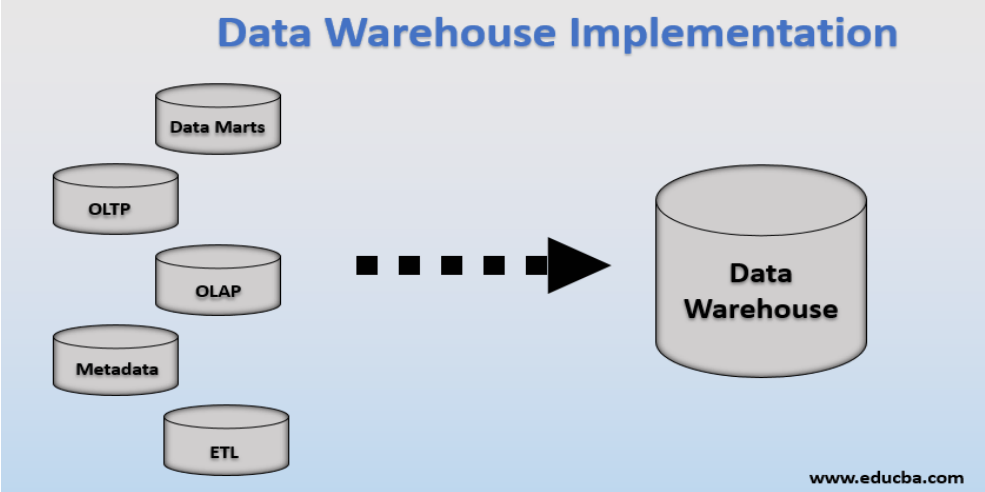

Figure 2: Data Warehouse Implementation [2]

2. **Components of Data Warehouses:** The Data Warehouse operates by utilizing an RDBMS server as a nucleus, serving as a pivotal repository for data. This central repository is encompassed by crucial components inherent to Data Warehousing, collectively fostering an operational, controllable, and approachable environment. [9].

- **Data Warehouse Database:** The central database is the basic of data warehouse environment and this database is implemented by RDBMS technology. To obtain the RDBMS technology for data warehousing some alternative approaches must be followed those are,

  - Relational database must be implemented by using parallel technology for the purpose of scalability.
  - Novel index mechanisms are deployed to overcome the bypass relational table scan and to improve the speed.
  - Using multi dimensional databases to overcome the limitations of relational data warehouse models.

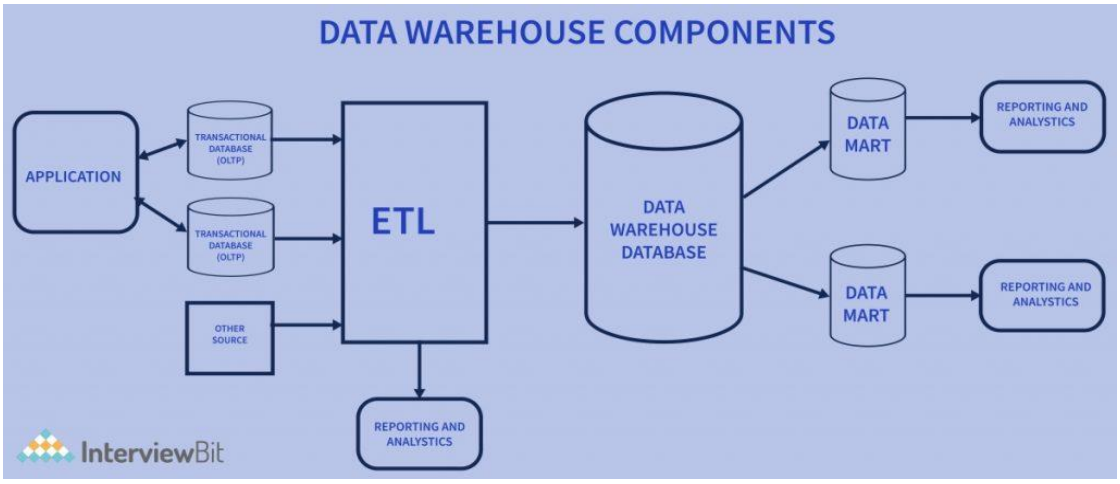

Figure 3: Data Warehouse Components [3]

Copyright © 2024 Authors

- **Extract, Transformation and Loading (ETL) Tools:** These tools are used to perform all the conversions, summarizations and the changes needed to transform the data in the same format in the data warehouse. These tools are called ETL tools. Their main functionalities are,

  - **Filtering Out Redundant Data:** Removing superfluous information in operational databases before transferring it to the data warehouse.
  - **Standardizing Terminology and Concepts:** Substituting shared descriptions and labels for data gathered from diverse origins.
  - **Generating Summarized Insights:** Computing condensed overviews or summaries of data.
  - **Handling Missing Information:** Substituting predetermined default values for absent data points as necessary.

- **Metadata:** Metadata refers to information about data, effectively constituting the data that delineates the data warehouse. This metadata assumes a critical role in formulating, sustaining, and supervising the data warehouse. Within the framework of Data Warehouse Architecture, metadata assumes a pivotal function, outlining the origin, utility, attributes, and attributes of the data housed within the warehouse. Furthermore, it prescribes the methods by which data can be altered and manipulated. This metadata maintains a closely intertwined relationship with the data warehouse.

3. **Query Tools:** One of the main objectives of data warehousing is to provide information to businesses for making better decisions. Query tools allow users to interact with the database system. The tool fall into four main categories that are,

   - Query & Reporting tools
   - Application development tools
   - Data mining tools
   - OLAP tools

4. **Data Warehouse Implementation:** Implementing a data warehouse include series of steps and processes that are listed below,

   - **Requirements Gathering and Planning:** First we need to identify and understand the requirements of the data warehouse this includes determines the scope of an data warehouse projects , finding the proper data sources and data volume and the analytical needs [9].

   - **Data Source Identification:** Next identifying the various sources of data needs to be integrated in the Data warehouse. Determine the process of extraction of data from the source system. This process includes batch processing, real time streaming and the other methods also.

   - **Data Extraction:** Extract the data from the identified source system by using ETL tools or any other data integration mechanisms then transform the data into the data warehouse. Data cleaning will be done to handle the inconsistencies or errors in the data.

Copyright © 2024 Authors

- **Data Modeling:** In this step the data warehouse schema is designed based on the business requirements and the analytical need. Frequently star and snowflake schemas are used which supports querying and reporting.

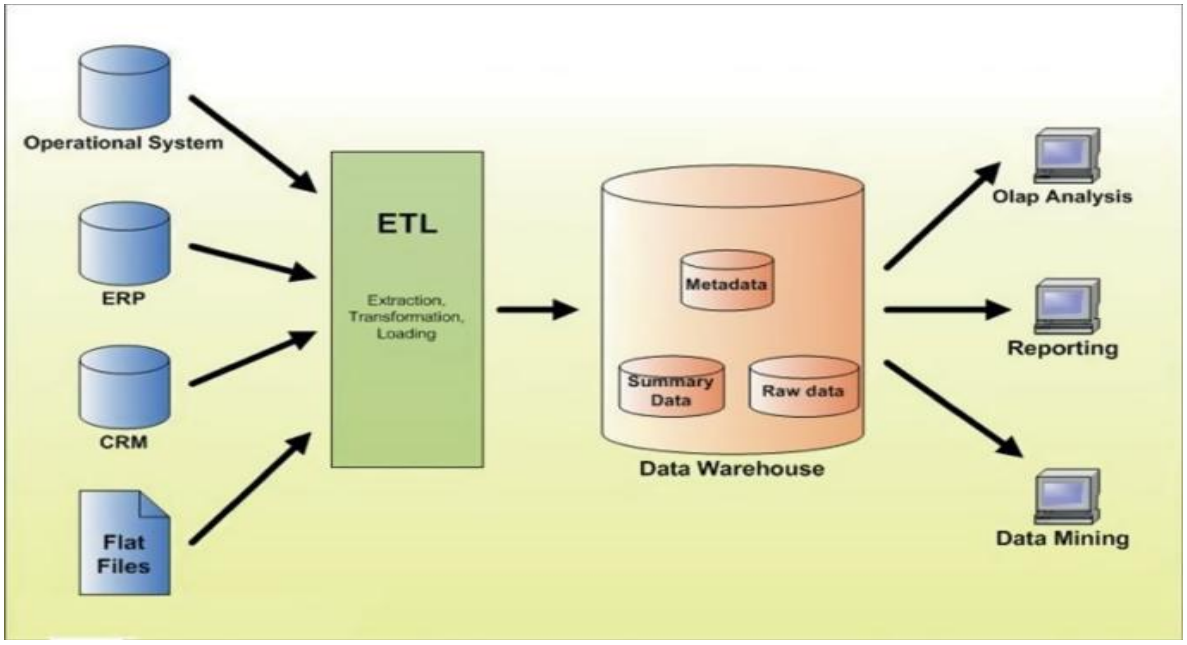

Figure 4: Steps of Data Warehouse Implantation [4]

- **Data Aggregation and Indexing:** In this step appropriate indexing strategies is implemented to improve the query performance.

- **Security and Access Control:** Implement the security measures to control the access to the data warehouse and to assure the data privacy. Access control mechanism is implemented to define roles and permissions for different user groups based on their role and analytical needs.

- **Testing and Validation:** Testing the data warehouse entirely to ensure data accuracy and consistency, validating the data warehouse meets the particular business requirements

- **Deployment and Rollout:** Once the data warehouse has been validated and the client requirements are tested, the warehouse system and the operations may be rolled out for the user's community to use.

5. Database Vs Data Warehouse

Database	Data Warehouse
Supports operational process	Supports analysis, performance and reporting.
Maintain the data	Explore the information
Real time data	Historical data
Data is updated based on the	Data update is a scheduled process

Copyright © 2024 Authors

transactions	
100 MB to GB	100 GB to TB
ER based	Star or Snowflake schema
Application based	Subject based
Detailed information	Consolidated data
Flat relational model.	Multi dimensional model.

IV. CONCLUSION

The above chapter discusses in detail about the databases, types and characteristics of database, database software and database management systems. Data warehouses is the another core concept which is explained in detail including the characteristics and components of data warehouses, implementation steps in data warehouses and a comparison is done at the end between databases and data warehouses. At the end a database is suited for operational and real time data management applications in the other hand data warehouses is designed for analytical processing, provides a consolidated view on the historical data to support business intelligence and effective decision making process.

REFERENCES

[1] www.javatpoint.com
[2] www.educba.com
[3] www.InterviewBit.com
[4] www.springerlink.com
[5] "Database System Concepts" by Abraham Silberschatz, Henry F.Korth, and S.Sudharshan, McGraw-Hill Education, Edition-2019.
[6] "Database Design for Mere Mortals: A Hands –On Guide to Relational Database Design by Michael J.Hernandez, Addidon- Wesley Professional, Edition – 2013.
[7] "The Art of SQL" by Stephane Faroult and Peter Robson, O'Reilly Media,Edition – 2006.
[8] "Building the Data Warehouse" by W.H.Inmon, Publisher: Wiley, Edition – 2005.
[9] "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling" by Ralph Kimball and Margy Ross, Publisher :Wiley, Edition – 2013.

Copyright © 2024 Authors