**DATA WAREHOUSE**

**WHAT IS DATA WAREHOUSE?**

A data warehouse is a centralized system used for storing and managing large volumes of data from various sources. It is designed to help businesses analyse historical data and make informed decisions. Data from different operational systems is collected, cleaned, and stored in a structured way, enabling efficient querying and reporting.

* **Goal is to produce statistical results that may help in decision-making.**
* **Ensures fast data retrieval even with the vast datasets.**

A key in a keyhole

AI-generated content may be incorrect.

**KEY CHARACTERISTICS OF A DATA WAREHOUSE:**

Several fundamental characteristics distinguish a data warehouse from other data management systems:

* **Subject-Oriented:** Data in a warehouse is organized around key business subjects, such as customers, products, sales, and suppliers. This contrasts with operational databases that are application-oriented, focusing on the data needed for specific business processes. By focusing on subjects, the data warehouse provides a holistic view relevant for analysis. For instance, all information related to a customer, including their purchase history, demographics, and interactions, would be consolidated under the "customer" subject.
* **Integrated:** Data from diverse source systems, which may have inconsistent formats, structures, and naming conventions, is cleansed, transformed, and integrated into a unified and consistent format within the data warehouse. This integration process ensures data quality and allows for meaningful comparisons and analyses across different business areas. For example, customer names might be stored differently in a CRM system and an order processing system. The data warehouse would resolve these inconsistencies and store a standardized customer name.
* **Time-Variant:** Data in a data warehouse is historical, capturing data over an extended period. This allows for trend analysis, comparisons over time, and forecasting. Each data record is typically associated with a timestamp, enabling users to track changes and analyse data from different points in the past. Unlike operational databases that usually maintain only the current state of data, a data warehouse retains a historical record, enabling insights like year-over-year sales growth or changes in customer behavior over time.
* **Non-Volatile:** Once data is loaded into the data warehouse, it is generally not updated or deleted. The data remains persistent, providing a stable and reliable source for analysis. New data is typically added in batches during periodic refresh cycles. This non-volatility ensures the integrity of historical data and supports consistent reporting and analysis.

**DATA WAREHOUSE ARCHITECTURE**

**Architecture of a Data Warehouse:**

A typical data warehouse architecture involves several key components working together:

1. **Source Systems:** These are the various operational systems and external data sources that feed data into the data warehouse. Examples include CRM systems, ERP systems, marketing automation platforms, point-of-sale systems, and external market research data. These systems are designed for transactional processing and often have diverse data structures.
2. **ETL (Extract, Transform, Load) Process:** This is a crucial set of processes responsible for preparing data for the data warehouse.
   * **Extraction:** Data is extracted from the various source systems. This might involve selecting relevant data, handling different file formats, and connecting to various database systems.
   * **Transformation:** The extracted data is cleaned, standardized, and transformed to fit the data warehouse schema. This includes tasks like data cleansing (handling missing values, correcting errors), data integration (merging data from different sources), data transformation (converting data types, calculating new values), and data filtering.
   * **Loading:** The transformed data is loaded into the data warehouse. This typically involves bulk loading processes to efficiently add large volumes of data.
3. **Data Warehouse Database:** This is the core of the data warehouse, a relational database management system (RDBMS) or a specialized data warehouse appliance optimized for analytical queries.
4. **Data Marts:** Subject-oriented subsets of the data warehouse designed to serve the specific needs of business units or user groups (e.g., a marketing data mart, a sales data mart). They provide faster access to relevant data for specific analytical needs.
5. **Metadata Repository:** This component stores information about the data in the data warehouse, including its source, format, structure, transformation rules, and usage. Metadata is crucial for understanding the data warehouse, managing it effectively, and enabling users to find and interpret data correctly. It acts as a "data dictionary" for the data warehouse.
6. **Reporting Tools:** These are the applications used by end-users to access and analyse the data in the data warehouse. They enable users to create reports, dashboards, perform ad-hoc queries, and conduct sophisticated data analysis to gain insights and support decision-making. Examples include SQL query tools, reporting platforms, OLAP (Online Analytical Processing) tools, and data visualization software.

A diagram of data storage

AI-generated content may be incorrect.

**WHY DO WE NEED DATA WAREHOUSING**

In today's data-rich environment, data warehousing is essential for organizations seeking to transform raw, disparate operational data into a unified and historically rich resource optimized for analytical purposes. By integrating and cleansing data from various sources, a data warehouse provides a consistent "single version of the truth," enabling comprehensive historical analysis, trend identification, and ultimately, more informed and strategic business decisions. This separation of analytical processing from transactional systems ensures optimal performance for both, while the enhanced data quality and business intelligence capabilities offered by a data warehouse provide a significant competitive advantage in the modern business landscape.

A diagram of data storage

AI-generated content may be incorrect.