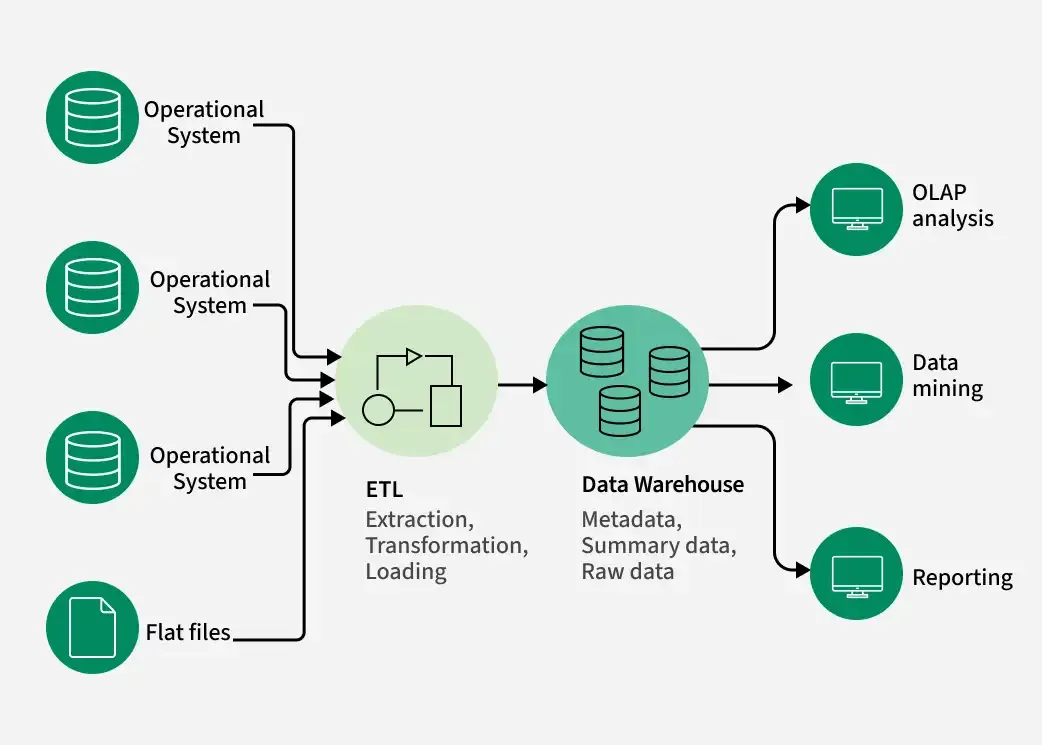
**DATA WAREHOUSE**

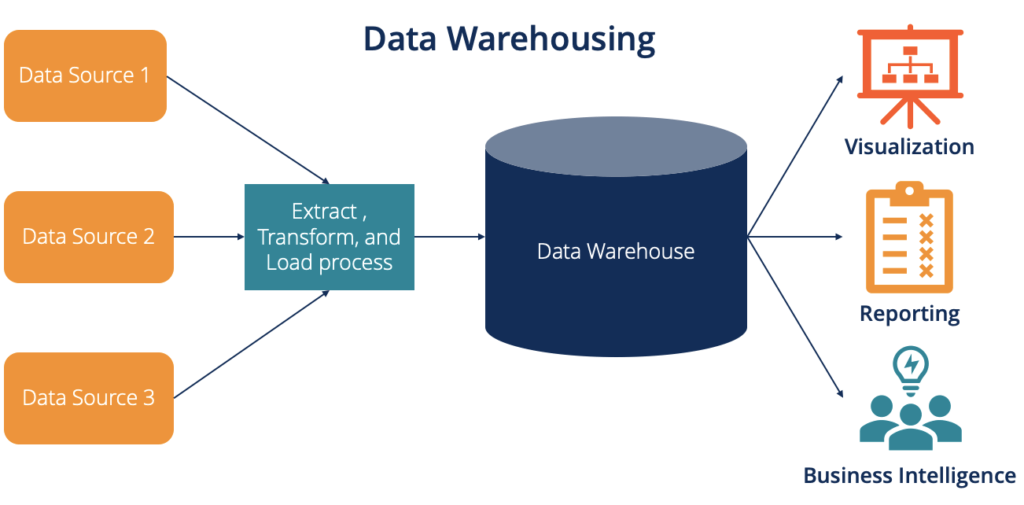
* **Ashwin Harish P**

**What is Data Ware House**

* 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 analyze 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.
* It is optimized for fast querying, reporting, and analytics, not for real-time transactional updates.

**Eg:** Amazon has a warehouse in which it have all the products and those products are in different categories

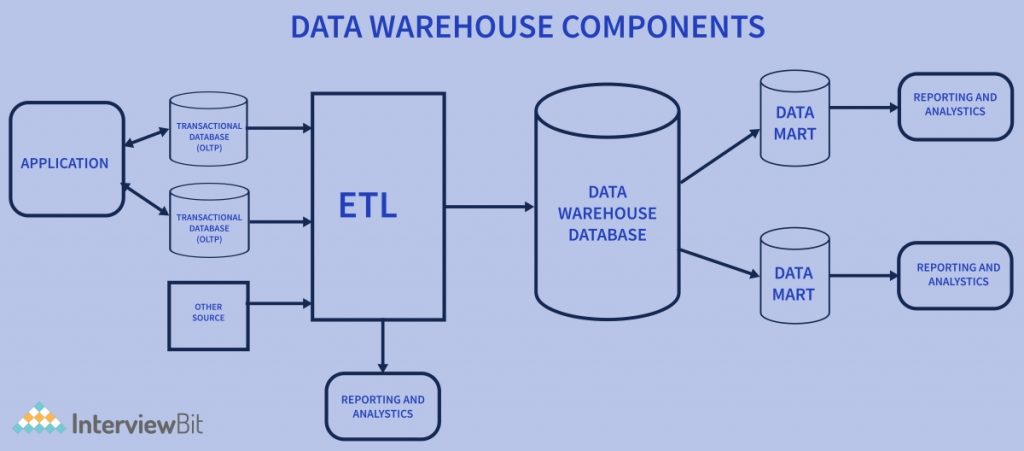




**Need for Data Warehousing**

1. **Handling Large Volumes of Data:** Traditional databases can only store a limited amount of data (MBs to GBs), whereas a data warehouse is designed to handle much larger datasets (TBs), allowing businesses to store and manage massive amounts of historical data.
2. **Enhanced Analytics:** Transactional databases are not optimized for analytical purposes. A data warehouse is built specifically for data analysis, enabling businesses to perform complex queries and gain insights from historical data.
3. **Centralized Data Storage**: A data warehouse acts as a central repository for all organizational data, helping businesses to integrate data from multiple sources and have a unified view of their operations for better decision-making.
4. **Trend Analysis:** By storing historical data, a data warehouse allows businesses to analyze trends over time, enabling them to make strategic decisions based on past performance and predict future outcomes.
5. **Support for Business Intelligence:** Data warehouses support business intelligence tools and reporting systems, providing decision-makers with easy access to critical information, which enhances operational efficiency and supports data-driven strategies.

**Components of Data Warehouse**



1. **Data Sources:** These are the various operational systems, databases, and external data feeds that provide raw data to be stored in the warehouse.
2. **ETL (Extract, Transform, Load) Process:** The ETL Process is responsible for extracting data from different sources, transforming it into a suitable format, and loading it into the data warehouse.
3. **Data Warehouse Database:** This is the central repository where cleaned and transformed data is stored. It is typically organized in a multidimensional format for efficient querying and reporting.
4. **Metadata:** Metadata describes the structure, source, and usage of data within the warehouse, making it easier for users and systems to understand and work with the data.
5. **Data Marts:** These are smaller, more focused data repositories derived from the data warehouse, designed to meet the needs of specific business departments or functions.
6. **OLAP (Online Analytical Processing) Tools:** OLAP Tools allow users to analyze data in multiple dimensions, providing deeper insights and supporting complex analytical queries.
7. **End-User Access Tools:** These are reporting and analysis tools, such as dashboards or Business Intelligence (BI) Tools that enable business users to query the data warehouse and generate reports.

**Key Characteristics**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Feature** | **Description** | | --- | --- | | Subject-Oriented | Organized around key subjects like customers, sales, products, etc., rather than application processes. | | Integrated | Combines data from multiple heterogeneous sources into a unified format. | | Time-Variant | Maintains historical data, enabling trend analysis over time. | | Non-Volatile | Data is read-only; once entered, it’s not updated or deleted (unlike operational databases). | |

**Types of Data Warehouses**

The different types of Data Warehouses are:

1. **Enterprise Data Warehouse (EDW):** A centralized warehouse that stores data from across the organization for analysis and reporting.
2. **Operational Data Store (ODS):** Stores real-time operational data used for day-to-day operations, not for deep analytics.
3. **Data Mart:** A subset of a data warehouse, focusing on a specific business area or department.
4. **Cloud Data Warehouse:** A data warehouse hosted in the cloud, offering scalability and flexibility.
5. **Big Data Warehouse:** Designed to store vast amounts of unstructured and structured data for big data analysis.
6. **Virtual Data Warehouse:** Provides access to data from multiple sources without physically storing it.
7. **Hybrid Data Warehouse:** Combines on-premises and cloud-based storage to offer flexibility.
8. **Real-time Data Warehouse:** Designed to handle real-time data streaming and analysis for immediate insights.
9. **Data Lakehouse:** Combines the capabilities of a traditional data warehouse (structured data and analytics) with a data lake (unstructured and semi-structured data).
10. **Columnar Data Warehouse**: Stores data in columns rather than rows to optimize analytical queries.

**Advantages**

* Better decision-making through historical data analysis
* Consistency and data quality across sources
* Improved performance for complex queries
* Centralized access to business-critical data

**Disadvantages**

* High setup and maintenance cost
* Complexity in design and implementation
* Time-consuming ETL processes
* Not ideal for real-time data analysis