Data Warehouses

CSC230 Database Technologies for Analytics

02 November 2021

Contents

1 Data Warehouse **2**

2 What is a data warehouse? **2**

3 Data warehouse architecture **3**

4 Understanding OLAP and OLTP in data warehouses **4**

5 Schemas in data warehouses **5**

6 Data warehouse vs. database, data lake, and data mart **6**

6.1 Data warehouse vs. data lake 6

6.2 Data warehouse vs. data mart 6

6.3 Data warehouse vs. database 7

7 Types of data warehouses **7**

7.1 Cloud data warehouse 7

7.2 Data warehouse software (on-premises/license) 7

7.3 Data warehouse appliance 7

8 Benefits of a data warehouse **8**

9 What Is a Data Warehouse? **8**

9.1 Data Warehouse Defined 9

9.2 Benefits of a Data Warehouse 9
9.3 Data Warehouse Architecture 10
9.4 The Evolution of Data Warehouses—From Data Analytics to AI and Machine Learning 11
9.4.1 Data Warehouses, Data Marts, and Operation Data Stores 13
9.5 What is a Cloud Data Warehouse? 13
9.6 What is a Modern Data Warehouse? 14
9.7 Designing a Data Warehouse 15
9.7.1 The Cloud and the Data Warehouse 16
9.8 Do I Need a Data Lake? 16
9.9 Why Not Run Analytics Against Your OLTP Environment? . . . 16
9.10 Zero-Complexity Deployment: The Autonomous Data Warehouse 17

1 Data Warehouse

https://www.ibm.com/cloud/learn/data-warehouse

A core component of business intelligence, a data warehouse pulls together data from many different sources into a single data repository for sophisticated analytics and decision support.

- business intelligence
- data from many sources
- single repository
- analytics / decision support

2 What is a data warehouse?

A data warehouse, or enterprise data warehouse (EDW), is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI), and machine learning. A data warehouse system enables an organization to run powerful analytics on huge volumes (petabytes and petabytes) of historical data in ways

that a standard database cannot.

Data warehousing systems have been a part of business intelligence (BI) solutions for over three decades, but they have evolved recently with the emergence of new data types and data hosting methods. Traditionally, a data warehouse was hosted on-premises—often on a mainframe computer—and its functionality was focused on extracting data from other sources, cleansing and preparing the data, and loading and maintaining the data in a relational database. More recently, a data warehouse might be hosted on a dedicated appliance or in the cloud, and most data warehouses have added analytics capabilities and data visualization and presentation tools.

- data warehouse also called Enterprise Data Warehouse (EDW)
- single, central, consistent data store
- may contains many petabytes of historical data
- used to be in general-purpose computer in company's offices
- then in a special purpose computer
- now in the cloud
- coupled with software for analysis and visualization of data

3 Data warehouse architecture

Bottom tier: The bottom tier consists of a data warehouse server, usually a relational database system, which collects, cleanses, and transforms data from multiple data sources through a process known as Extract, Transform, and Load (ETL) or a process known as Extract, Load, and Transform (ELT).

Middle tier: The middle tier consists of an OLAP (i.e. online analytical processing) server which enables fast query speeds. Three types of OLAP models can be used in this tier, which are known as ROLAP, MOLAP and HOLAP. The type of OLAP model used is dependent on the type of database system that exists.

Top tier: The top tier is represented by some kind of front-end user interface or reporting tool, which enables end users to conduct ad-hoc data analysis on their business data.

- multiple layers of software and hardware
- OLAP is OnLine Analytical Processing
- ETL (sometimes ELT) is Extract, Transform, and Load

4 Understanding OLAP and OLTP in data warehouses

OLAP (for online analytical processing) is software for performing multidimensional analysis at high speeds on large volumes of data from unified, centralized data store, like a data warehouse. OLTP, or online transactional processing, enables the real-time execution of large numbers of database transactions by large numbers of people, typically over the internet. The main difference between OLAP and OLTP is in the name: OLAP is analytical in nature, and OLTP is transactional.

OLAP tools are designed for multidimensional analysis of data in a data warehouse, which contains both historical and transactional data. Common uses of OLAP include data mining and other business intelligence applications, complex analytical calculations, and predictive scenarios, as well as business reporting functions like financial analysis, budgeting, and forecast planning.

OLTP is designed to support transaction-oriented applications by processing recent transactions as quickly and accurately as possible. Common uses of OLTP include ATMs, e-commerce software, credit card payment processing, online bookings, reservation systems, and record-keeping tools.

- OLAP is OnLine Analytical Processing
  - data mining
  - predicting / forecasting
  - budgeting
  - financial analysis
  - using both historical data and transactional data
- OLTP is OnLine Transactional Processing
  - ATMs
  - credit card payment processing
  - online sales
  - bookings and reservations (airlines, hotels, etc.)

5 Schemas in data warehouses

Schemas are ways in which data is organized within a database or data warehouse. There are two main types of schema structures, the star schema and the snowflake schema, which will impact the design of your data model.

Star schema: This schema consists of one fact table which can be joined to a number of denormalized dimension tables. It is considered the simplest and most common type of schema, and its users benefit from its faster speeds while querying.

Snowflake schema: While not as widely adopted, the snowflake schema is another organization structure in data warehouses. In this case, the fact table is connected to a number of normalized dimension tables, and these dimension tables have child tables. Users of a snowflake schema benefit from its low levels of data redundancy, but it comes at a cost to query performance.

- star schema is more popular
  - fact table
  - connected to denormalized dimension tables
  - faster queries
- snowflake schema is also used
  - fact table
  - connected to normalized dimension tables
  - dimension tables connected to other (child) dimension tables
  - less redundancy

6 Data warehouse vs. database, data lake, and data mart

Data warehouse, database, data lake, and data mart are all terms that tend to be used interchangeably. While the terms are similar, important differences exist:

6.1 Data warehouse vs. data lake

A data warehouse gathers raw data from multiple sources into a central repository, structured using predefined schemas designed for data analytics. A data lake is a data warehouse without the predefined schemas. As a result, it enables more types of analytics than a data warehouse. Data lakes are commonly built on big data platforms such as Apache Hadoop.

6.2 Data warehouse vs. data mart

A data mart is a subset of a data warehouse that contains data specific to a particular business line or department. Because they contain a smaller subset of data, data marts enable a department or business line to discover more-focused insights more quickly than possible when working with the broader data warehouse data set.

6.3 Data warehouse vs. database

A database is built primarily for fast queries and transaction processing, not analytics. A database typically serves as the focused data store for a specific application, whereas a data warehouse stores data from any number (or even all) of the applications in your organization.

A database focuses on updating real-time data while a data warehouse has a broader scope, capturing current and historical data for predictive analytics, machine learning, and other advanced types of analysis.

7 Types of data warehouses

7.1 Cloud data warehouse

A cloud data warehouse is a data warehouse specifically built to run in the cloud, and it is offered to customers as a managed service. Cloud-based data warehouses have grown more popular over the last five to seven years as more companies use cloud services and seek to reduce their on-premises data center footprint.

With a cloud data warehouse, the physical data warehouse infrastructure is managed by the cloud company, meaning that the customer doesn't have to make an upfront investment in hardware or software and doesn't have to manage or maintain the data warehouse solution.

7.2 Data warehouse software (on-premises/license)

A business can purchase a data warehouse license and then deploy a data warehouse on their own on-premises infrastructure. Although this is typically more expensive than a cloud data warehouse service, it might be a better choice for government entities, financial institutions, or other organizations that want more control over their data or need to comply with strict security or data privacy standards or regulations.

7.3 Data warehouse appliance

A data warehouse appliance is a pre-integrated bundle of hardware and software—CPUs, storage, operating system, and data warehouse software—that a business can connect to its network and start using as-is. A data warehouse appliance sits somewhere between cloud and on-premises implementations in terms of upfront cost, speed of deployment, ease of scalability, and management control.

8 Benefits of a data warehouse

A data warehouse provides a foundation for the following:

Better data quality: A data warehouse centralizes data from a variety of data sources, such as transactional systems, operational databases, and flat files. It then cleanses it, eliminates duplicates, and standardizes it to create a single source of the truth.

Faster, business insights: Data from disparate sources limit the ability of decision makers to set business strategies with confidence. Data warehouses enable data integration, allowing business users to leverage all of a company's data into each business decision.

Smarter decision-making: A data warehouse supports large-scale BI functions such as data mining (finding unseen patterns and relationships in data), artificial intelligence, and machine learning—tools data professionals and business leaders can use to get hard evidence for making smarter decisions in virtually every area of the organization, from business processes to financial management and inventory management

Gaining and growing competitive advantage: All of the above combine to help an organization finding more opportunities in data, more quickly than is possible from disparate data stores.

- subject-oriented
- integrated (consistancy among data from different sources)
- non-volatile (stable over time)
- time variant (changes over time)

---

9 What Is a Data Warehouse?

https://www.oracle.com/database/what-is-a-data-warehouse/

9.1 Data Warehouse Defined

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making. Over time, it builds a historical record that can be invaluable to data scientists and business analysts. Because of these capabilities, a data warehouse can be considered an organization's "single source of truth."

A typical data warehouse often includes the following elements:

- A relational database to store and manage data
- An extraction, loading, and transformation (ELT) solution for preparing the data for analysis
- Statistical analysis, reporting, and data mining capabilities
- Client analysis tools for visualizing and presenting data to business users
- Other, more sophisticated analytical applications that generate actionable information by applying data science and artificial intelligence (AI) algorithms, or graph and spatial features that enable more kinds of analysis of data at scale

- designed to support analytics
- includes relational databases, statistical analysis
- data derived from wide range of sources

9.2 Benefits of a Data Warehouse

Data warehouses offer the overarching and unique benefit of allowing organizations to analyze large amounts of variant data and extract significant value from it, as well as to keep a historical record.

Four unique characteristics (described by computer scientist William Inmon, who is considered the father of the data warehouse) allow data warehouses to deliver this overarching benefit. According to this definition, data warehouses are:

Subject-oriented. They can analyze data about a particular subject or functional area (such as sales).

Integrated. Data warehouses create consistency among different data types from disparate sources.

Nonvolatile. Once data is in a data warehouse, it's stable and doesn't change.

Time-variant. Data warehouse analysis looks at change over time.

A well-designed data warehouse will perform queries very quickly, deliver high data throughput, and provide enough flexibility for end users to "slice and dice" or reduce the volume of data for closer examination to meet a variety of demands—whether at a high level or at a very fine, detailed level. The data warehouse serves as the functional foundation for middleware BI environments that provide end users with reports, dashboards, and other interfaces.

9.3 Data Warehouse Architecture

The architecture of a data warehouse is determined by the organization's specific needs. Common architectures include:

Simple. All data warehouses share a basic design in which metadata, summary data, and raw data are stored within the central repository of the warehouse. The repository is fed by data sources on one end and accessed by end users for analysis, reporting, and mining on the other end.

Simple with a staging area. Operational data must be cleaned and processed before being put in the warehouse. Although this can be done programmatically, many data warehouses add a staging area for data before it enters the warehouse, to simplify data preparation.

Hub and spoke. Adding data marts between the central repository and end users allows an organization to customize its data warehouse to serve various lines of business. When the data is ready for use, it is moved to the appropriate data mart.

Sandboxes. Sandboxes are private, secure, safe areas that allow companies to quickly and informally explore new datasets or ways of analyzing data without having to conform to or comply with the formal rules and protocol of the data warehouse.

- multiple architectures from which to choose
- common ones include
  - simple—central repository
  - simple and staging area—central repository plus staging area for data cleaning
  - hub and spoke—central repository plus data marts (for specific uses)
  - sandbox—less structured, less formal

9.4 The Evolution of Data Warehouses—From Data Analytics to AI and Machine Learning

When data warehouses first came onto the scene in the late 1980s, their purpose was to help data flow from operational systems into decision-support systems (DSSs). These early data warehouses required an enormous amount of redundancy. Most organizations had multiple DSS environments that served their various users. Although the DSS environments used much of the same data, the gathering, cleaning, and integration of the data was often replicated for each environment.

As data warehouses became more efficient, they evolved from information stores that supported traditional BI platforms into broad analytics infrastructures that support a wide variety of applications, such as operational analytics and performance management.

Data warehouse iterations have progressed over time to deliver incremental additional value to the enterprise with enterprise data warehouse (EDW).

Step	Capability	Business Value
1	Transactional reporting	Provides relational information to create snapshots of business performance
2	Slice and dice, ad hoc query, BI tools	Expands capabilities for deeper insights and more robust analysis
3	Predicting future performance (data mining)	Develops visualizations and forward-looking business intelligence
4	Tactical analysis (spatial, statistics)	Offers "what-if" scenarios to inform practical decisions based on more comprehensive analysis
5	Stores many months or years of data	Stores data for only weeks or months

Supporting each of these five steps has required an increasing variety of datasets. The last three steps in particular create the imperative for an even broader range of data and analytics capabilities.

Today, AI and machine learning are transforming almost every industry, service, and enterprise asset—and data warehouses are no exception. The expansion of big data and the application of new digital technologies are driving change in data warehouse requirements and capabilities.

The autonomous data warehouse is the latest step in this evolution, offering enterprises the ability to extract even greater value from their data while lowering costs and improving data warehouse reliability and performance.

Find out more about autonomous data warehouses and get started with your own autonomous data warehouse.

- databases have become more autonomous
- lower costs
- improved reliability
- better performance

9.4.1 Data Warehouses, Data Marts, and Operation Data Stores

Though they perform similar roles, data warehouses are different from data marts and operation data stores (ODSs). A data mart performs the same functions as a data warehouse but within a much more limited scope—usually a single department or line of business. This makes data marts easier to establish than data warehouses. However, they tend to introduce inconsistency because it can be difficult to uniformly manage and control data across numerous data marts.

ODSs support only daily operations, so their view of historical data is very limited. Although they work very well as sources of current data and are often used as such by data warehouses, they do not support historically rich queries.

9.5 What is a Cloud Data Warehouse?

A cloud data warehouse uses the cloud to ingest and store data from disparate data sources.

The original data warehouses were built with on-premises servers. These on-premises data warehouses continue to have many advantages today. In many cases, they can offer improved governance, security, data sovereignty, and better latency. However, on-premises data warehouses are not as elastic and they require complex forecasting to determine how to scale the data warehouse for future needs. Managing these data warehouses can also be very complex.

On the other hand, some of the advantages of cloud data warehouses include:

- Elastic, scale-out support for large or variable compute or storage requirements
- Ease of use
- Ease of management
- Cost savings

The best cloud data warehouses are fully managed and self-driving, ensuring that even beginners can create and use a data warehouse with only a few clicks. An easy way to start your migration to a cloud data warehouse is to run your cloud data warehouse on-premises, behind your data center firewall which complies with data sovereignty and security requirements.

In addition, most cloud data warehouses follow a pay-as-you-go model, which brings added cost savings to customers.

- uses cloud to store data from different sources
- easier to scale
- easier to use
- pay as you go / more economical

9.6 What is a Modern Data Warehouse?

Whether they're part of IT, data engineering, business analytics, or data science teams, different users across the organization have different needs for a data warehouse.

A modern data architecture addresses those different needs by providing a way to manage all data types, workloads, and analysis. It consists of architecture patterns with necessary components integrated to work together in alignment with industry best practices. The modern data warehouse includes:

- A converged database that simplifies management of all data types and provides different ways to use data
- Self-service data ingestion and transformation services
- Support for SQL, machine learning, graph, and spatial processing
- Multiple analytics options that make it easy to use data without moving it
- Automated management for simple provisioning, scaling, and administration

A modern data warehouse can efficiently streamline data workflows in a way that other warehouses can't. This means that everyone, from analysts and data engineers to data scientists and IT teams, can perform their jobs more effectively and pursue the innovative work that moves the organization forward, without countless delays and complexity.

- addresses different needs of different jobs
- simplifies management of all data types
- self-service data services
- support for SQL, spatial processing, machine learning, graph
- multiple analytic options
- automated mangagement
- streamlined workflow (greater efficiency)

9.7 Designing a Data Warehouse

When an organization sets out to design a data warehouse, it must begin by defining its specific business requirements, agreeing on the scope, and drafting a conceptual design. The organization can then create both the logical and physical design for the data warehouse. The logical design involves the relationships between the objects, and the physical design involves the best way to store and retrieve the objects. The physical design also incorporates transportation, backup, and recovery processes.

Any data warehouse design must address the following:

- Specific data content
- Relationships within and between groups of data
- The systems environment that will support the data warehouse
- The types of data transformations required
- Data refresh frequency

A primary factor in the design is the needs of the end users. Most end users are interested in performing analysis and looking at data in aggregate, instead of as individual transactions. However, often end users don't really know what they want until a specific need arises. Thus, the planning process should include enough exploration to anticipate needs. Finally, the data warehouse design should allow room for expansion and evolution to keep pace with the evolving needs of end users.

9.7.1 The Cloud and the Data Warehouse

Data warehouses in the cloud offer the same characteristics and benefits of on-premises data warehouses but with the added benefits of cloud computing—such as flexibility, scalability, agility, security, and reduced costs. Cloud data warehouses allow enterprises to focus solely on extracting value from their data rather than having to build and manage the hardware and software infrastructure to support the data warehouse.

9.8 Do I Need a Data Lake?

Organizations use both data lakes and data warehouses for large volumes of data from various sources. The choice of when to use one or the other depends on what the organization intends to do with the data. The following describes how each is best used:

Data lakes store an abundance of disparate, unfiltered data to be used later for a particular purpose. Data from line-of-business applications, mobile apps, social media, IoT devices, and more is captured as raw data in a data lake. The structure, integrity, selection, and format of the various datasets is derived at the time of analysis by the person doing the analysis. When organizations need low-cost storage for unformatted, unstructured data from multiple sources that they intend to use for some purpose in the future, a data lake might be the right choice.

Data warehouses are specifically intended to analyze data. Analytical processing within a data warehouse is performed on data that has been readied for analysis—gathered, contextualized, and transformed—with the purpose of generating analysis-based insights. Data warehouses are also adept at handling large quantities of data from various sources. When organizations need advanced data analytics or analysis that draws on historical data from multiple sources across their enterprise, a data warehouse is likely the right choice.

9.9 Why Not Run Analytics Against Your OLTP Environment?

Data warehouses are relational environments that are used for data analysis, particularly of historical data. Organizations use data warehouses to discover patterns and relationships in their data that develop over time.

In contrast, transactional environments are used to process transactions on an ongoing basis and are commonly used for order entry and financial and retail transactions. They do not build on historical data; in fact, in OLTP environ-

ments, historical data is often archived or simply deleted to improve performance.

Data warehouses and OLTP systems differ significantly.

	Data Warehouse	OLTP System
Workload	Accommodates ad hoc queries and data analysis	Supports only predefined operation
s		
Data modifications	Automatically updates on a regular basis	Updates by end users issuing individual statements
Schema design	Uses partially denormalized schemas to optimize performance	Uses fully normalized schemas to guarantee data consistency
Data scanning	Encompasses thousands to millions of rows	Accesses only a handful of records at a time
Historical data	Stores many months or years of data	Stores data for only weeks or months

9.10 Zero-Complexity Deployment: The Autonomous Data Warehouse

The most recent iteration of the data warehouse is the autonomous data warehouse, which relies on AI and machine learning to eliminate manual tasks and simplify setup, deployment, and data management. An as-a-service autonomous data warehouse in the cloud requires no human-performed database administration, hardware configuration or management, or software installation.

Creating the data warehouse, backing up, patching and upgrading the database, and expanding or reducing the database are all performed automatically—with the same flexibility, scalability, agility, and reduced costs that cloud platforms offer. The autonomous data warehouse removes complexity, speeds deployment, and frees up resources so organizations can focus on activities that add value to the business.