Week 1 - Data Warehouses, Data Marts, and Data Lakes

Tipo	Anotaciones
Posición	
Etiquetas	
Bibliografía	
Fecha de creación	@June 18, 2022 8:15 PM
Property	

▼ An Introduction to Data Warehouses, Data Marts, and Data Lakes

▼ Course Introduction

A data warehouse is a large repository of data that has been cleaned to a consistent quality. Not all data repositories are used the same way or require the same rigor when choosing what data to store. Data warehouses enable rapid business decision-making through accurate and flexible reporting and data analysis. A data warehouse system is one of the most fundamental business intelligence tools in use today and a tool that successful data engineers must understand. You will see how data warehouses serve as a single source of data truth for an organization's current and historical data.

This course provides insight into the three main operational data stores that organizations use: enterprise data warehouse systems, data lakes, and data marts. You'll first learn about the architecture, features, and benefits of each of

these data stores. You'll focus on the primary data store used by growing organizations—the enterprise data warehouse system. With three platforms to choose from, learn why organizations select a specific data warehouse platform and the decision-making considerations that organizations apply to select a particular vendor.

Next, you'll focus on the data populating the warehouse and its structure. You'll learn how facts, fact tables, dimensions, and dimension tables work to design your data warehouse. You'll gain a practical understanding of the star and snowflake schemas commonly used in today's data warehouses and work with data to create a data warehouse schema. You'll learn how to apply CUBE and ROLLUP functions to speed the retrieval of aggregated data using materialized views.

No course about data warehouse systems would be complete without acquiring data analytics and business intelligence (BI) tools skills. First, you'll learn to identify common data analytics and BI tools and vendors. Then you'll gain job-ready, hands-on experience creating basic and advanced data visualizations using IBM Cognos Analytics.

Your final project enables you to demonstrate all the skills you acquired in the first three modules. You'll walk through an enterprise data warehouse system scenario to apply your skills to design and implement a data warehouse schema and create materialized queries. To complete your project, you'll create data visualizations using IBM Cognos Analytics. Then, you can share your completed project with peers, professional communities, or prospective employers.

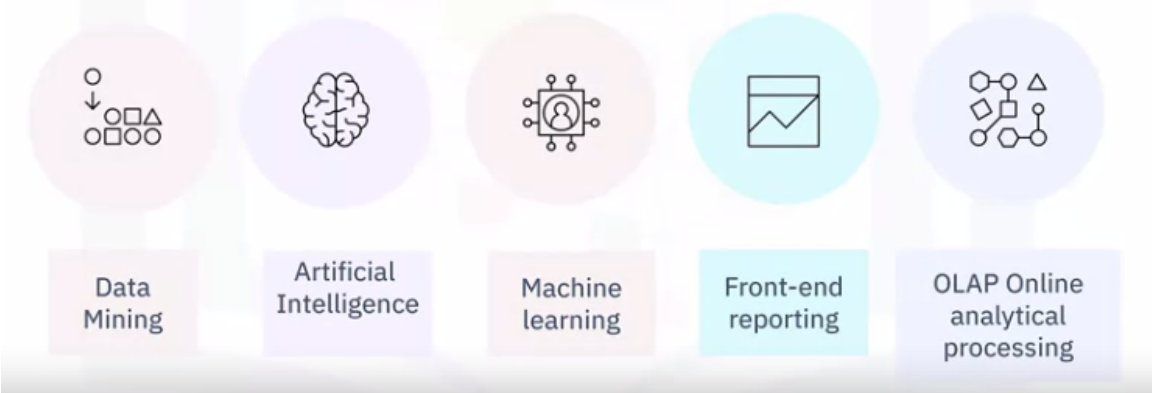
▼ Data Warehouse Overview

What is a data warehouse?

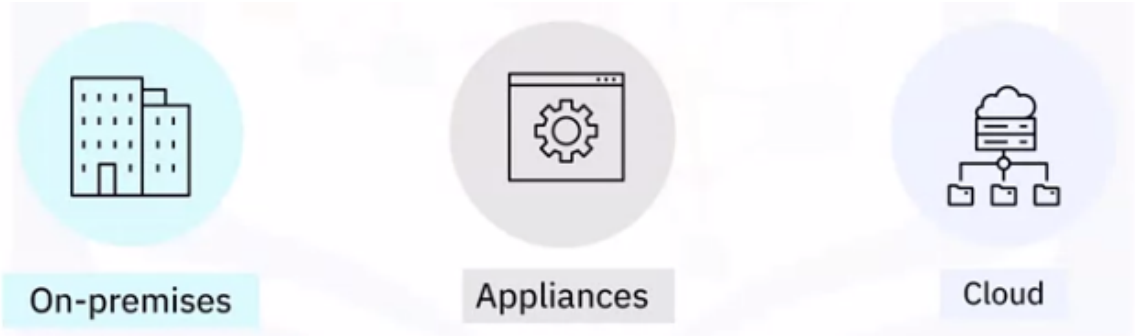
A data warehouse is a system that aggregates data from one or more sources into a single, central, consistent data store to support various data analytics requirements.

Data warehouse analytics

Data warehouse systems support:



Where are data warehouses hosted?



The beginning

Data warehouses have been hosted on-premises within enterprise data centers, initially on mainframes and then on Unix, Windows, and Linux systems.

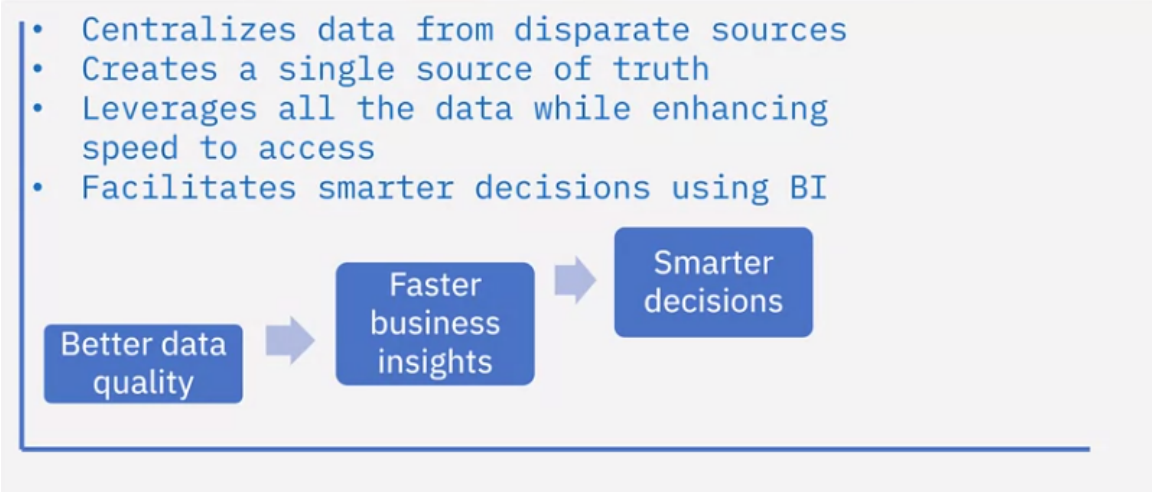
2000s

Data warehouse appliances emerged with the growth of more extensive data volumes in the 2000s. These appliances consisted of a pre-integrated bundle of specialized hardware and optimized data warehousing software that reduced large-scale data warehousing management overhead.

2010 - present

In the last decade or so, with exponential amounts of data being generated and stored in the cloud, Cloud Data Warehouses, frequently called CDWs, have gained popularity, where organizations don't purchase hardware or install warehousing software. Instead, organizations access data warehouses as a scalable, pay-as-you-go service.

Benefits of a data warehouse

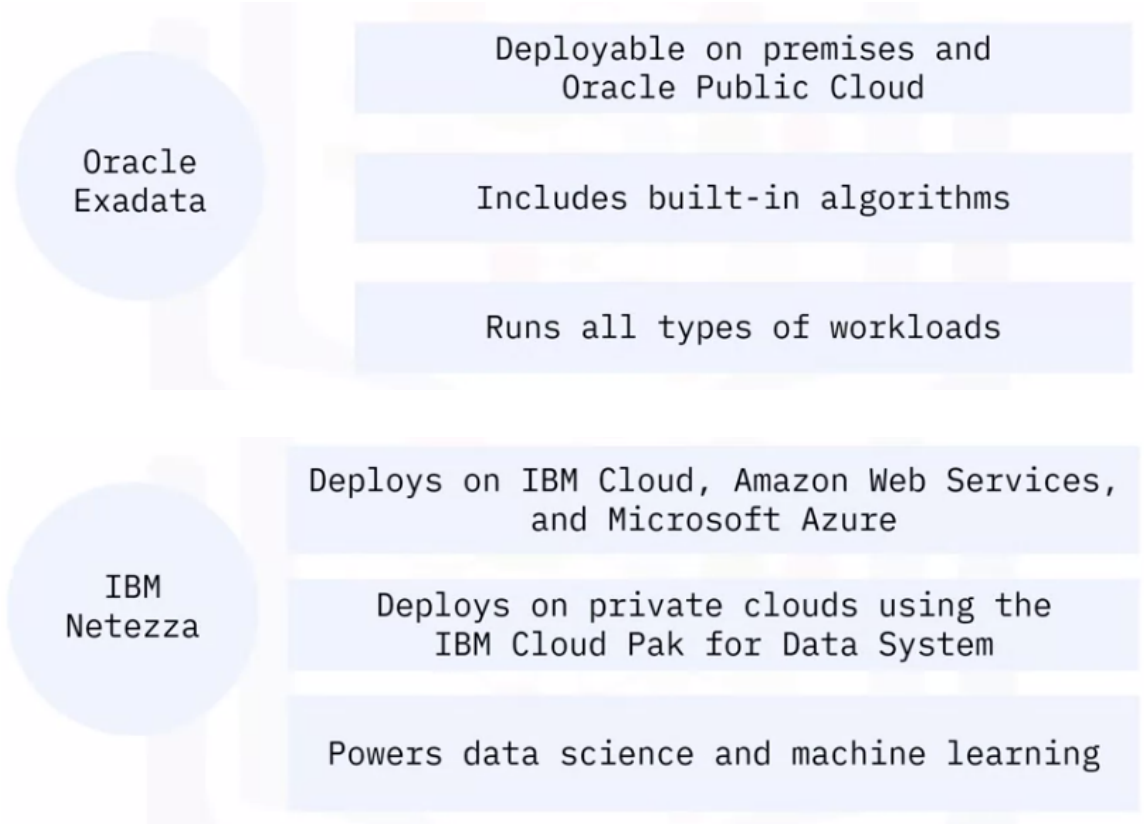


▼ Popular Data Warehouse Systems

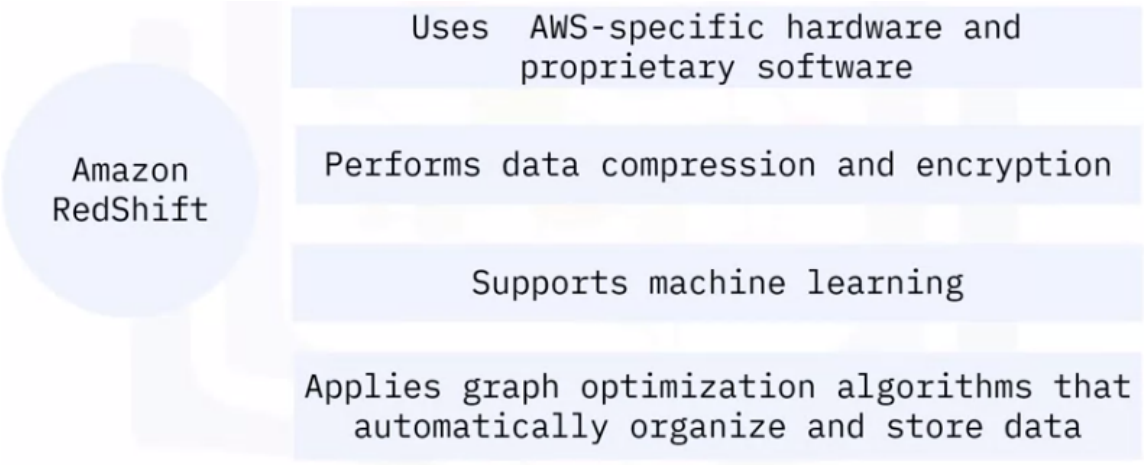
Categorizing data warehouse systems

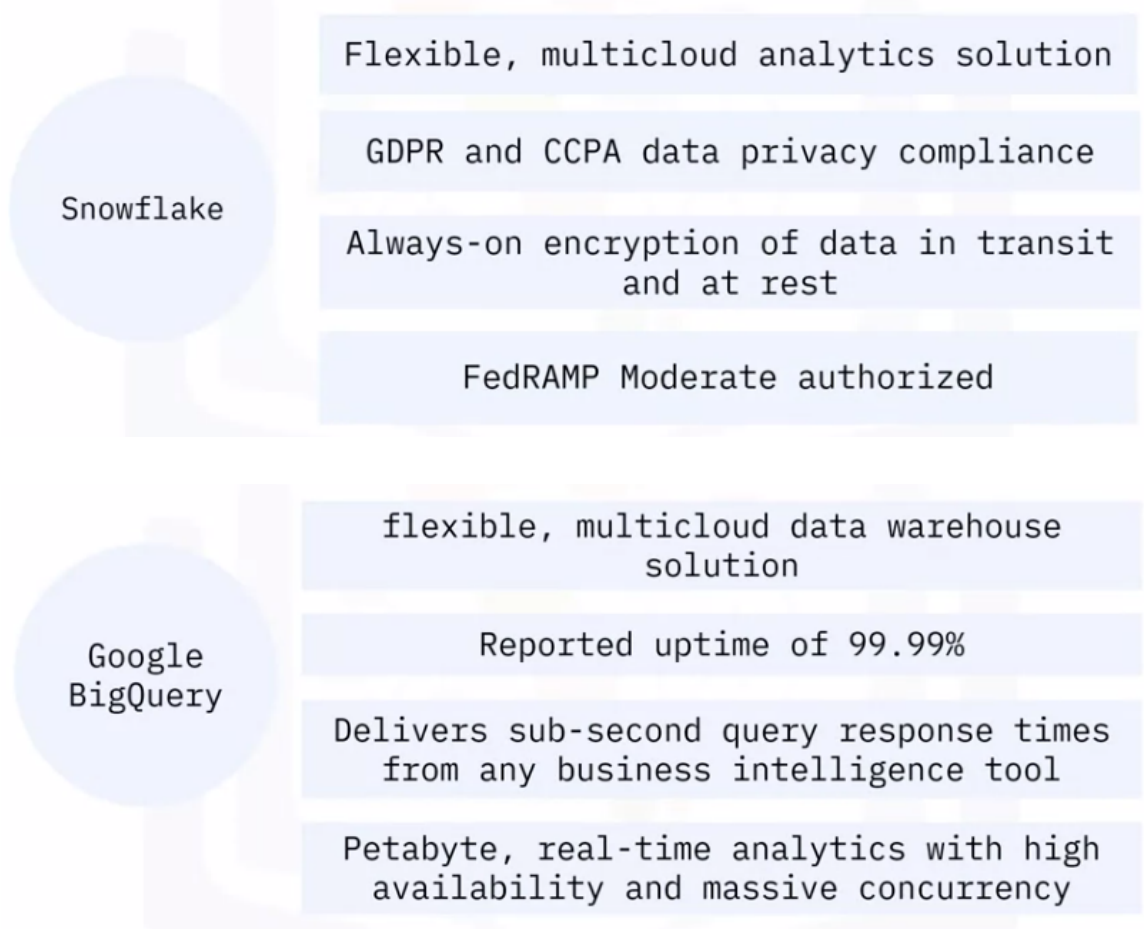
- **Appliances -** Pre-integrated bundles of hardware and software that provide high performance for workloads and low maintenance overhead.
- Other vendors support **cloud** deployments only, offering the benefits of cloud scalability and pay-per-use economics, and in many cases, deliver their data warehouses as fully managed services.
- Some warehouse offerings have traditionally been available as software installed only within **on-premises** environments, but in recent years, most of these vendors now offer cloud-deployed data warehouse systems.

Vendors - appliance offerings

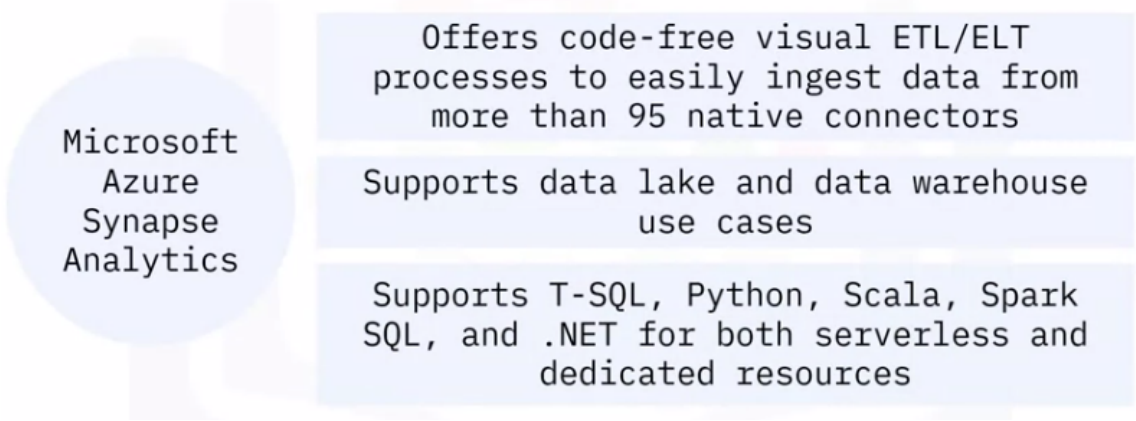


Vendors - cloud only

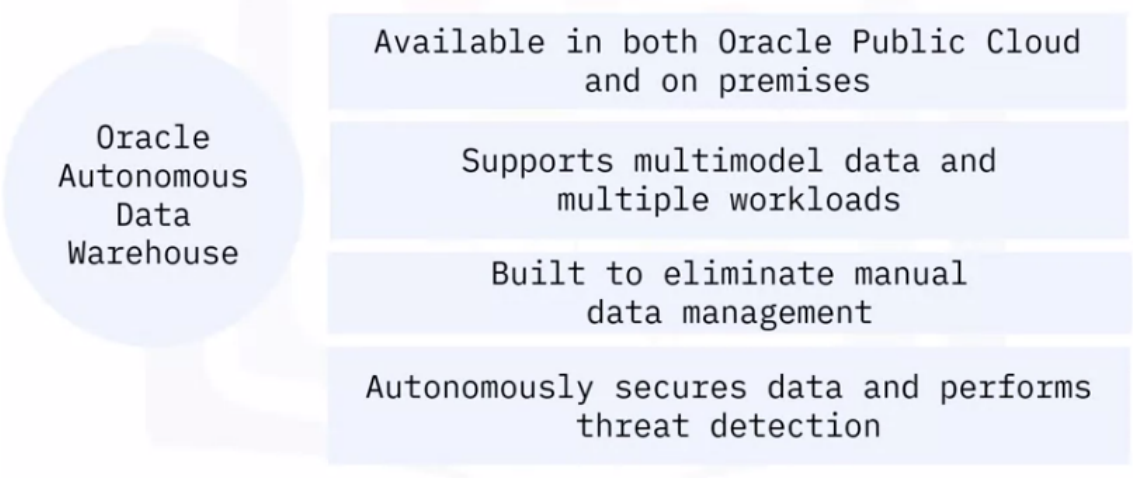




Vendors - on premises and cloud

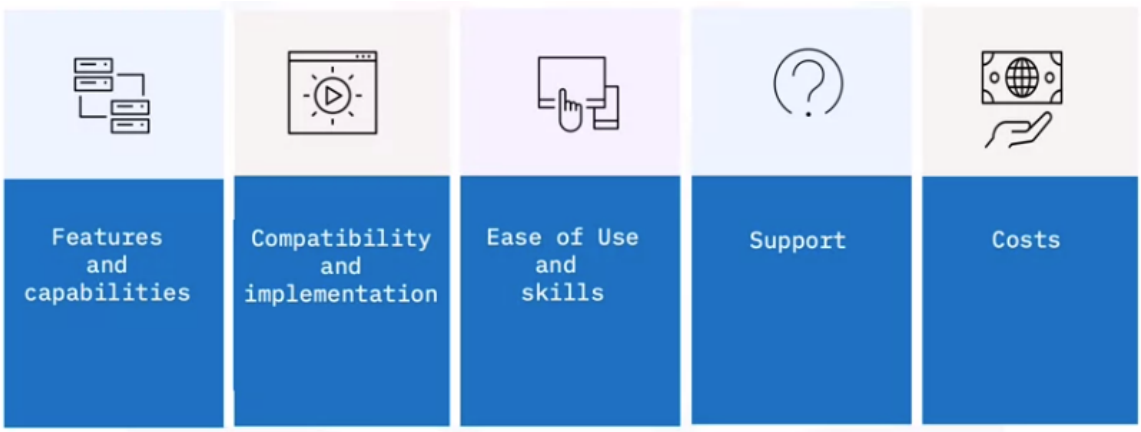


Teradata Vantage	Multicloud data platform for enterprise analytics that unifies everything
	Supports mixed workloads with high query concurrency using workload management and adaptive optimization
	Provides a single point of contact for operational task services
IBM Db2 Warehouse	Widely recognized for its scalability, MPP capabilities, petaflop speeds
	Rich security features with 99.99% service uptime
	Designed as a containerized, scale-out data warehousing solution
	Move workloads with minimal or no changes required
Vertica	Multicloud support for AWS, Google, Microsoft Azure, and on-premises Linux hardware
	Fast multi-GB data transfer rates
	Scalable, elastic compute and storage
	Eon Mode provides notable fault tolerance for volatile cloud environments



▼ Selecting a Data Warehouse System

Evaluating data warehouse systems



Features and capabilities

To select a location, organizations must balance multiple demands related to data ingestion, storage, and access. For some organizations, securing their data is their highest priority, requiring a mandatory on-premises solution. Multi-location businesses that grapple with data privacy requirements such as CCPA or GDPR need on-premises or geo-specific data warehouse locations. Every organization balances security and data privacy requirements with the need for speed that delivers critical, profit-producing business insights.

Organizations will also want to consider features and capabilities related to architecture and structure:

- Is the organization ready to commit to a vendor-specific architecture?
- Does the organization need multi-cloud installation such as multiple data warehouses in multiple locations?
- Does the solution scale to meet anticipated future needs?
- What data types are supported and what types of data does the organization ingest? If your organization currently analyzes dark data or is planning for the implementation of using semi-structured and unstructured data, you'll want a data warehouse system that supports these data types.
- An organization that processes big data needs a system that supports both batch and streaming data.

Ease of implementation

Capabilities that affectthe **ease of implementation** include:

- Data governance
- Data migration
- Data transformation capabilities
- With the data warehouse system in place, how easily can the organization optimize and reoptimize system performance as needs change?
- User management. With more organizations implementing a zero-trust security policy because of expensive data breaches, implementing programs that manage and validate system users is mandatory.
- Notifications and reports are essential for organizations to correct errors and mitigate risks before minor issues become larger problems.

Ease of use and skills:

Does your organization's staff have the skills needed to implement a specific data warehousing vendor's technology, and if not, how quickly and easily can they gain those skills?

Complex, large data warehouse deployments can require additional work from your implementation partner, so their expertise also greatly matters.

Do the technology and engineering staff who architect, deploy, and administer front-end querying, reporting, and visualization tools have the skills needed to configure your new system quickly?

Support considerations

Support is essential and can become frustrating and expensive if not well planned for.

- You might find that by using a single vendor, you can leverage one highly accountable, responsible source, potentially saving you time, money, and frustration.
- You'll also want to verify the availability of service level agreements for uptime, security, scalability, and other data warehouse system issues. Validate the vendor's support hours and channels, such as by phone, email, chat, or text.
- Does the vendor offer self-service solutions and an active rich user community?

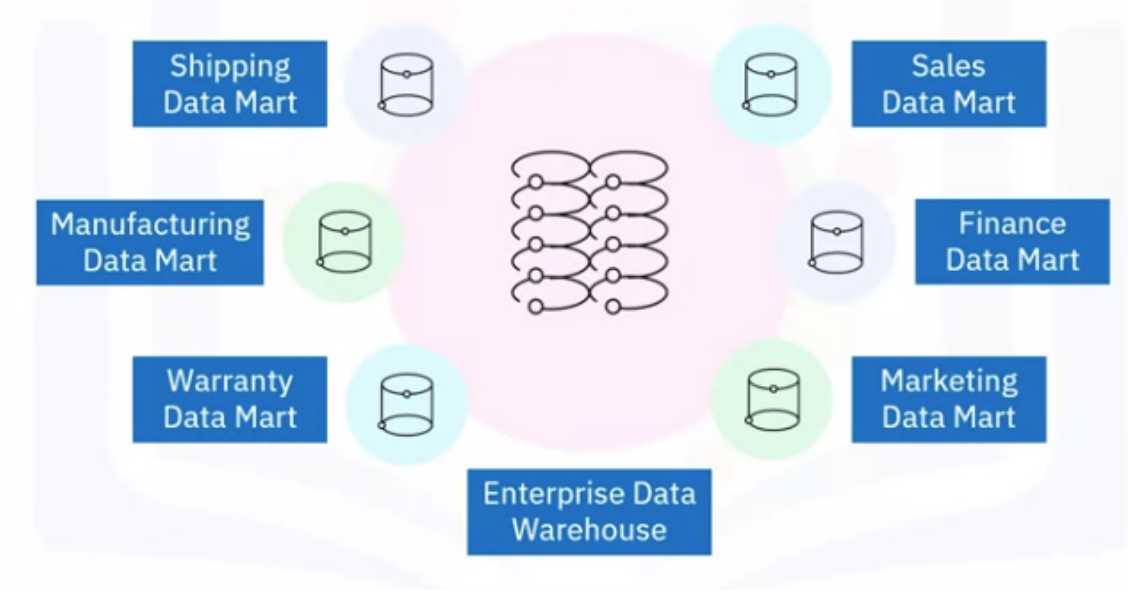
Costs

When calculating costs for a data warehouse system, consider more than the initial costs. Consider the total cost of ownership, or TCO, for running systems for several years. TCO includes:

- Infrastructure such as compute and storage costs – whether on-premises or on cloud.
- Software licensing, or in case of cloud offerings, their subscription or usage costs.
- Data migration and integration costs for moving data into the warehouse and pruning and purging as required.
- Administration costs for personnel to manage the systems and to train them.
- Recurring support and maintenance costs paid to the warehousing vendor or implementation partner.

▼ Data Marts Overview

What is a data mart?



A data mart is an isolated part of the larger enterprise data warehouse that is specifically built to serve a particular business function, purpose, or community of users. For example, the sales and finance departments in a company may have access to dedicated data marts that supply the data required for their quarterly sales reports and projections. The marketing team may use data marts to analyze customer behavior data, and the shipping, manufacturing and warranty departments may have their own data marts.

What are data marts used for?

- Data marts are designed to provide specific support for making tactical decisions.
- Data marts are focused only on the most relevant data, which saves end users the time and effort that would otherwise be spent searching the data warehouse for insights.

Data mart structure

Relational database with a star, or more often a snowflake schema, which means it contains a central fact table consisting of the business metrics relevant to a

business process, which is surrounded by a related hierarchy of dimension tables that provide context for the facts.

Data repository comparisons

Data Marts	Databases
OLAP systems – read intensive	OLTP systems – write intensive
Use Txn DBs or warehouses as data sources	Use operational applications as sources of data
Contain clean, validated analytical data	Contain raw, unprocessed transactional data
Accumulate history for trend analysis	May not always store history

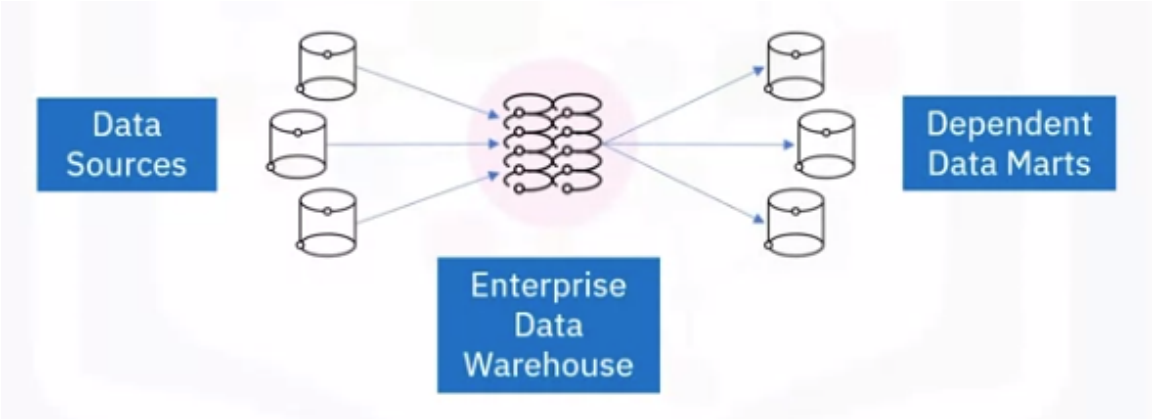
Data Marts	Data Warehouses
Small data warehouses with tactical scope	Large repositories with broad, strategic scope
Lean and fast	Large and slow

Types of data marts

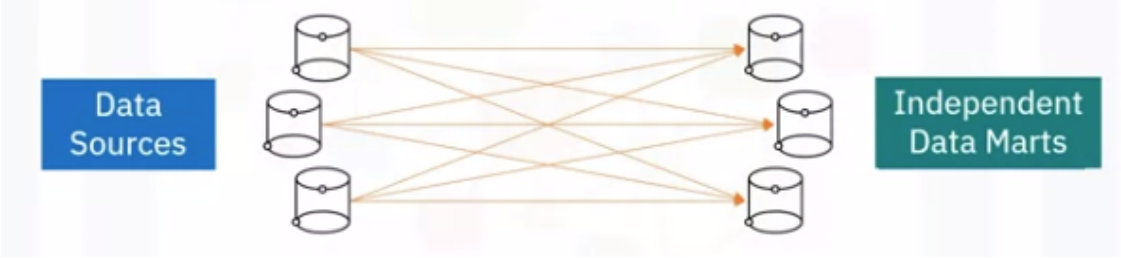
There are three basic types of data marts—dependent, independent, and hybrid.

The difference between these three kinds of data marts depends on their relationship with the data warehouse and the sources used for supplying each of them with data.

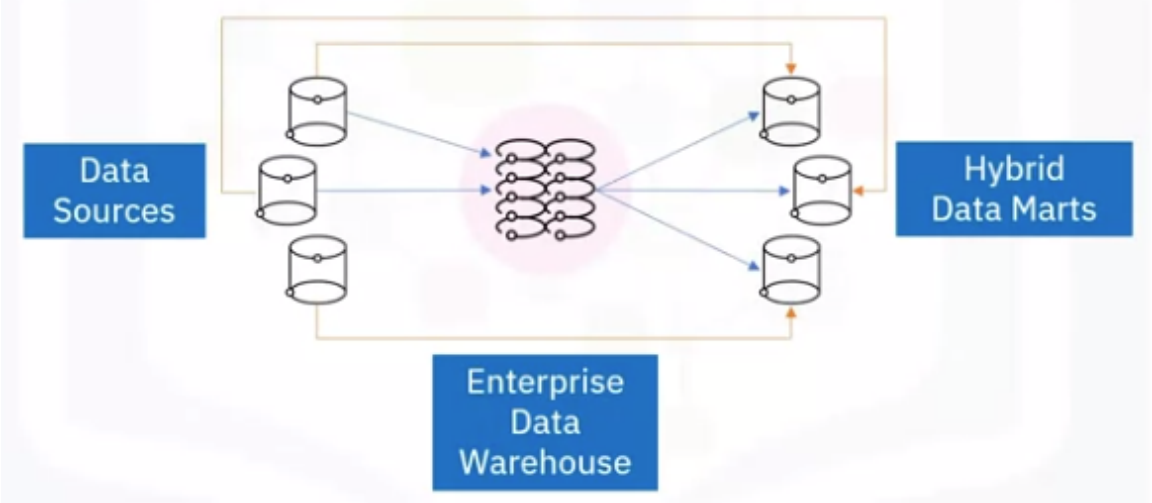
Dependent data marts draw data from the enterprise data warehouse.



Independent data marts bypass the data warehouse and are created directly from sources, which may include internal operational systems or external data from vendors or other sources outside the enterprise.

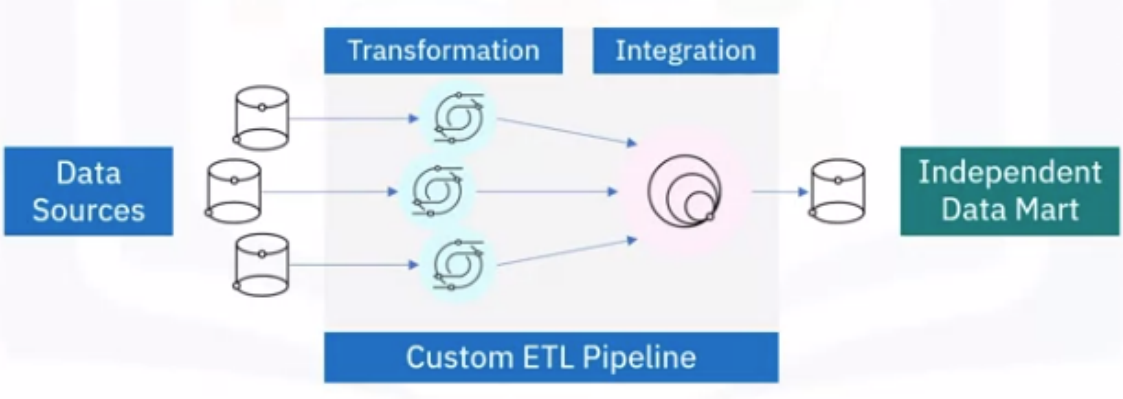


Hybrid data marts only depend partially on the enterprise data warehouse. They combine inputs from data warehouses with data from operational systems and other systems external to the warehouse.



Dependent data marts offer analytical capabilities within a restricted area of the enterprise data warehouse. Thus, they inherit the security that comes with the enterprise data warehouse. And since dependent data marts pull data directly from the data warehouse, where data has already been cleaned and transformed, they tend to have simpler data pipelines than independent data marts.

Independent data marts differ from dependent data marts because they require custom extract, transform and load data pipelines to carry out the transformation and integration processes on the source data since it is coming directly from operational systems and external sources, and independent data marts may also require separate security measures.



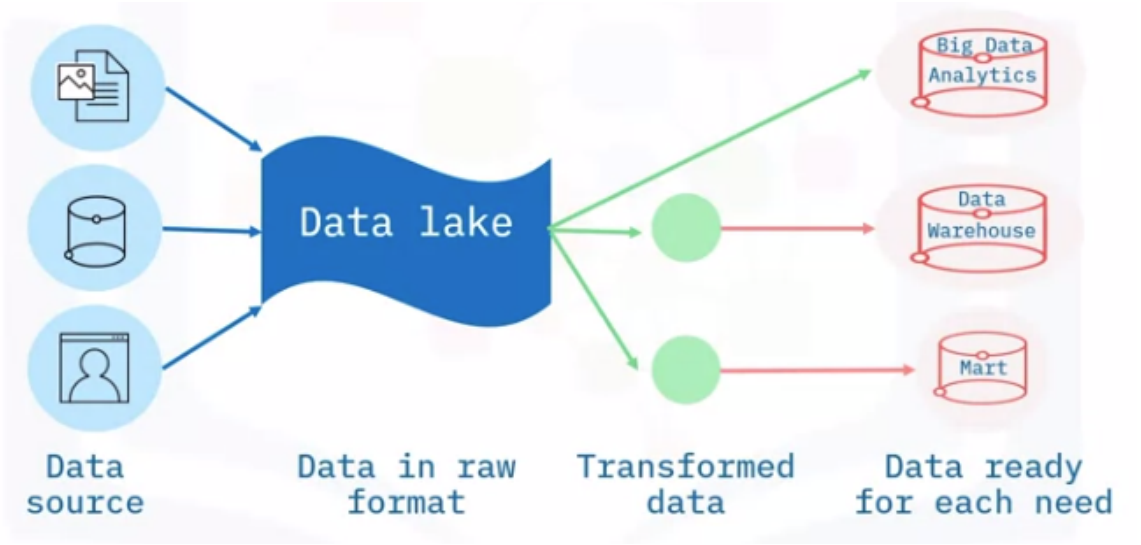
Data mart purpose

- Provide end-users with relevant data when they need it.
- Accelerate business processes by providing efficient query response times.
- Provide a cost-efficient method for informing data-driven decisions.
- Ensure secure access and control over your data.

▼ Data Lakes Overview

What is a data lake?

A data lake is a storage repository that can store large amounts of structured, semi-structured, and unstructured data in their native format, classified and tagged with metadata. While a data warehouse stores data processed for a specific need, a data lake is a pool of raw data where each data element is given a unique identifier and is tagged with metatags for further use. You would opt for a data lake if you generate, or have access to, large amounts of data on an ongoing basis but don’t want to be restricted to specific or pre-defined use cases.



Data lakes are sometimes also used as a staging area for transforming data prior to loading into a data warehouse or a data mart.

- Store large amounts of structured data in their native format
- Data can be loaded without defining the structure or schema of data
- Use cases do not need to be know in advance
- Exist as a repository of raw data straight from the source
- A reference architecture that combines multiple technologies
- Can be deployed using:
  - Cloud object storage
  - Large-scale distributed systems
  - Relational database management systems
  - NoSQL data repositories

Data lake benefits

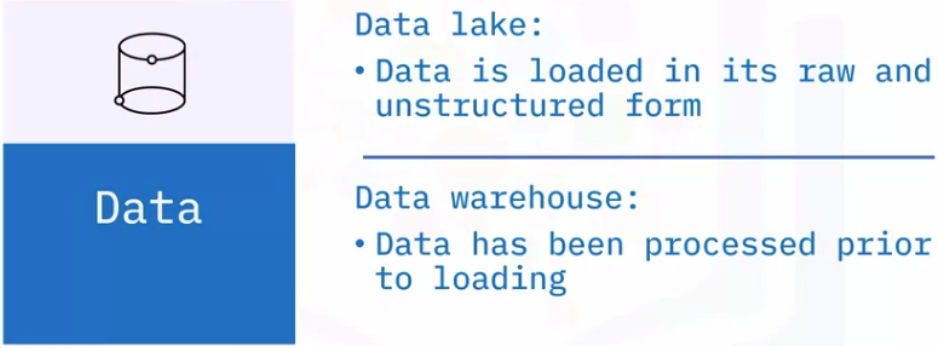
- Handles all types of data:
  - Unstructured
  - Semi-structure
  - Structure

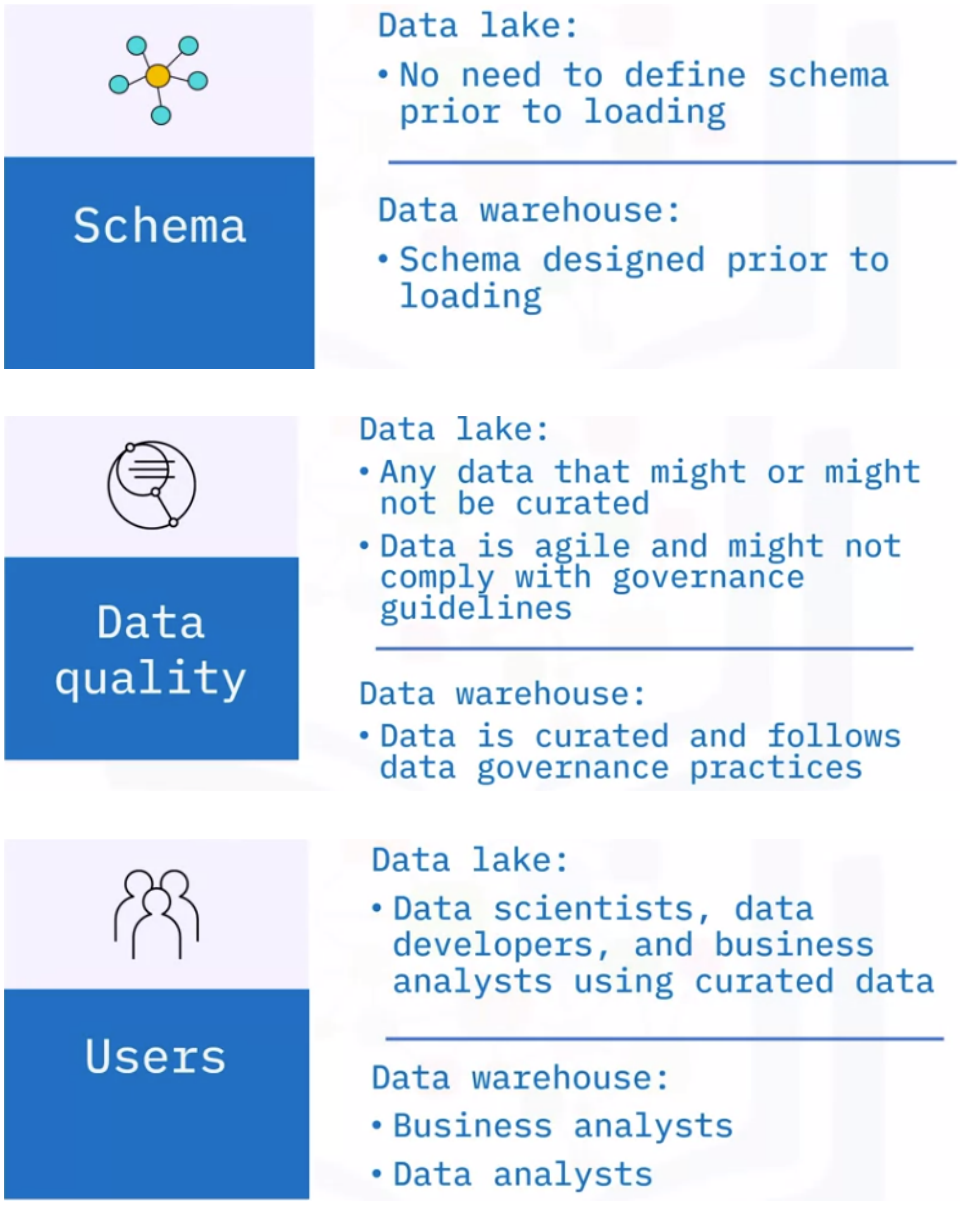
- Can make use of scalable storage capacity—from terabytes to petabytes of data.
- By retaining data in its original format, data lakes save organizations time that would have been used to define structures, create schemas, and transform the data.
- The ability to access data in its original format enables fast, flexible reuse of the data for a wide range of current and future use cases.

Vendors for data lakes

Depending on the requirements, a typical organization will require both a data warehouse and a data lake as they serve different needs.

Data lakes versus data warehouses





Hands-on Lab: Create Db2 service instance and Get started with Db2 console

Hands-on Lab: Create Db2 Service Credentials