Data Warehouses, Data Marts & Data Lakes (Wk1)

Welcome to your first module! This module provides a gentle but thorough introduction to data warehouse systems, data lakes, and data marts. When you complete this module, you'll be able to identify and compare data warehouse systems, data mart, and data lake architecture, and understand how organizations can benefit from each of these three data storage entities. Optionally, you'll explore the workings of IBM Db2 data warehouse system architecture, view use cases, and understand the key capabilities and integrations available with IBM Db2 Warehouse. Then, you'll learn about three types of data warehouse systems and popular data warehouse system vendors. You will be ready to help your organization assess new data warehouse system offerings when you know the five essential, critical criteria, including total cost of ownership, to evaluate before changing to a new data warehouse system.

Learning Objectives

- Define a data warehouse.
- Identify the benefits of and use cases for data warehouses.
- Recall the three most popular data warehouse platforms.
- Discuss the five mission-critical criteria to evaluate before implementing a new data warehouse system.
- Outline the various types of data warehouse costs.
- Define a data mart, describe how to populate data marts, relate data mart examples, and compare data marts to transactional databases and enterprise data warehouses.
- Describe a data lake, identify the benefits of data lakes, and recap the differences between data lakes and data warehouses.
- Create a Db2 instance on IBM Cloud.
- Locate and explore the Db2 console on IBM Cloud.
- Create credentials for accessing a Db2 instance on IBM Cloud.

Course Introduction

The website "Valuates Reports" projects that the data warehousing market valued at $21.18 billion in 2019 will reach $51.18 billion by 2028.

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**.

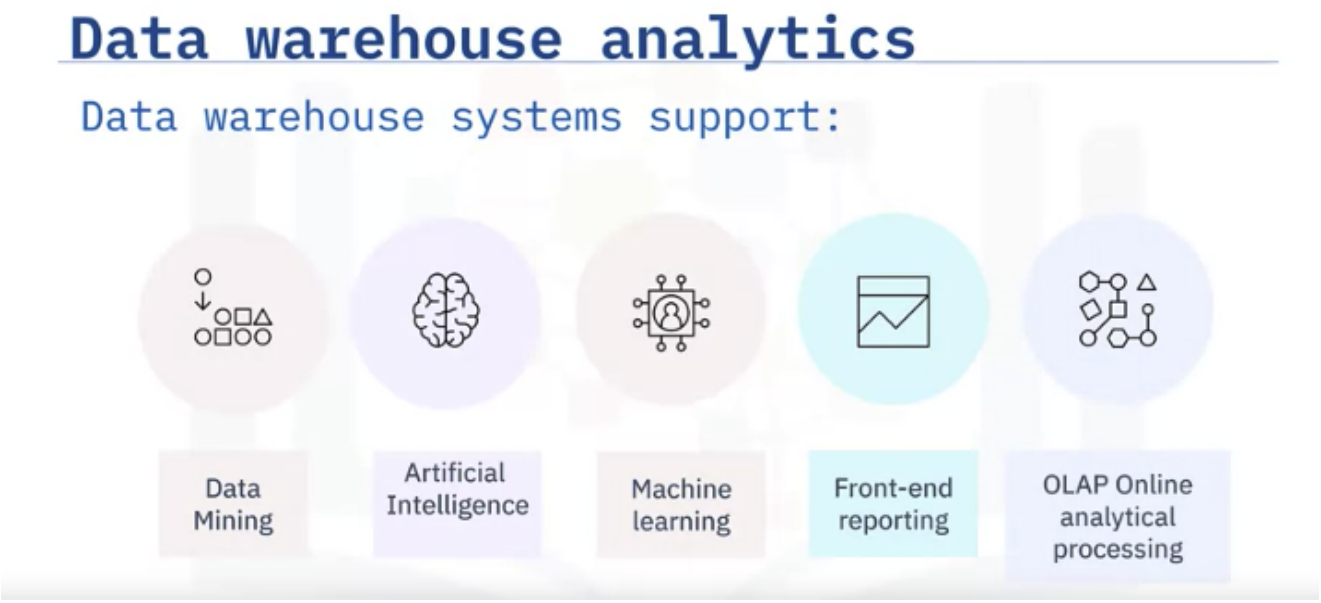
Your final project enables you to demonstrate all the skills you acquired in the first three modules. You'll walk through an enterprise data warehouse system scenario to apply your skills to design and implement a data warehouse schema and create materialized queries. To complete your project, you'll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

This course does not require any prior data warehouse, data analysis, or computer science experience. All you need to get started are basic computer literacy, high-school-level math, and access to a modern web browser such as Chrome or Firefox.

Data Warehouse Overview

- 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

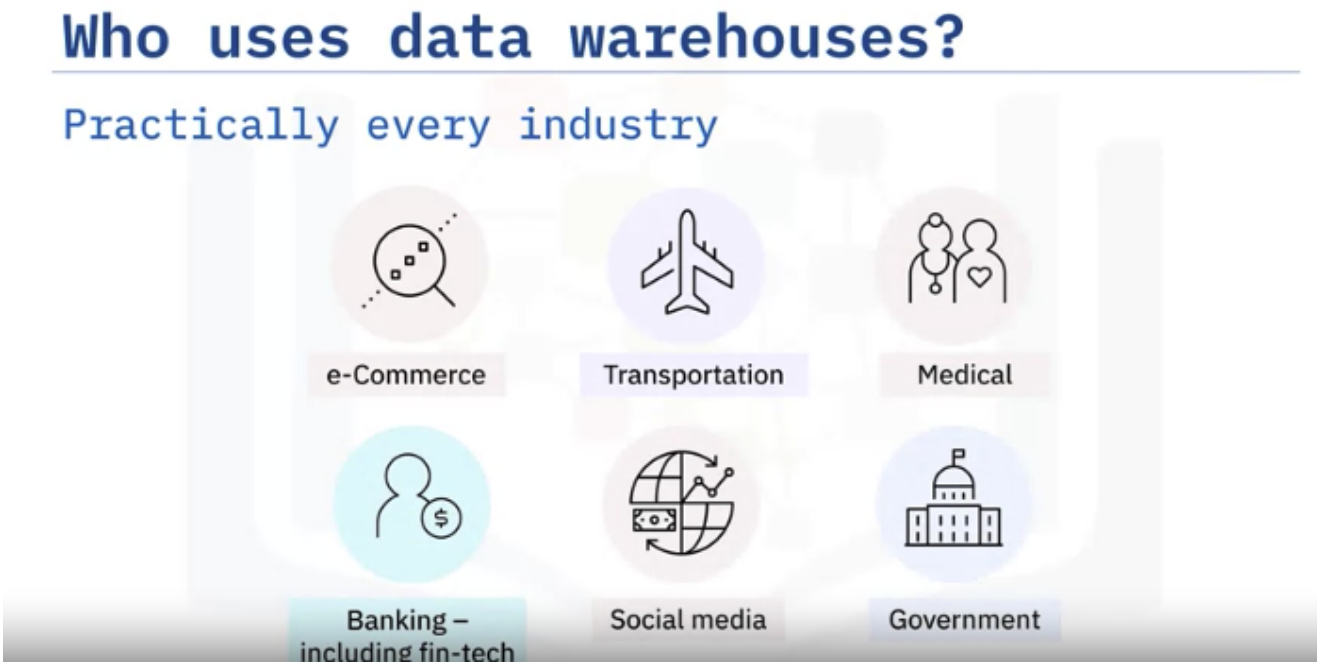
Warehouse Systems Support



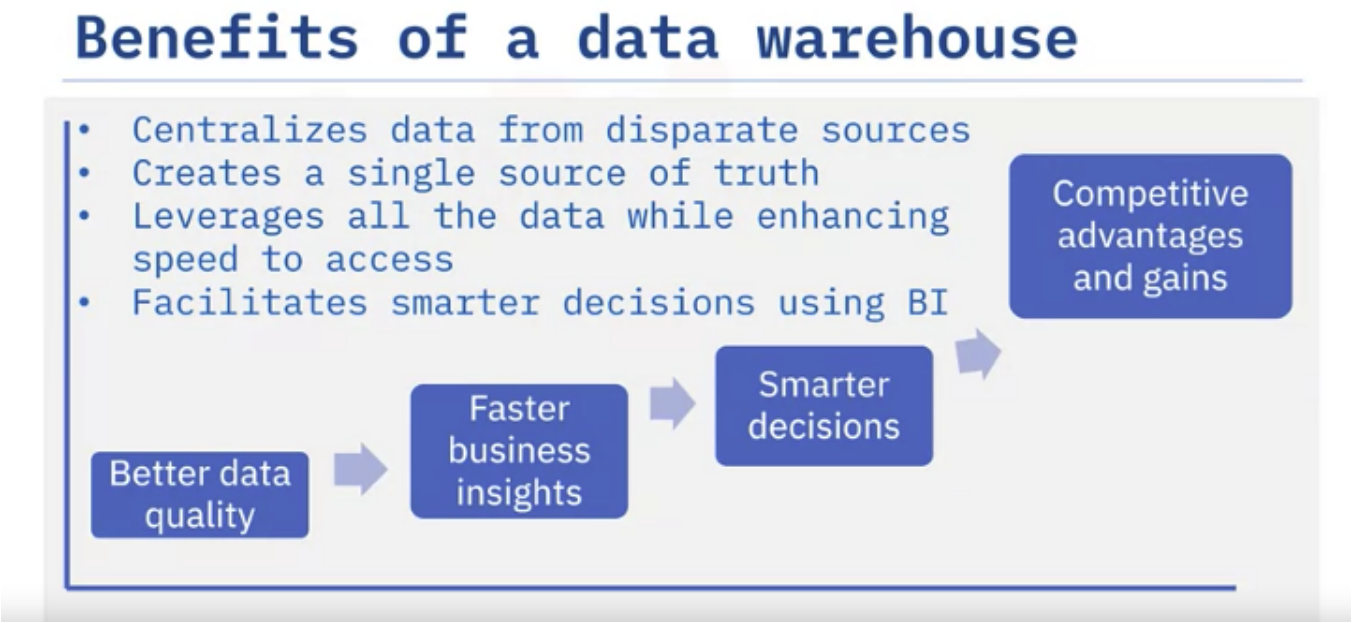
Background - Hosting

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

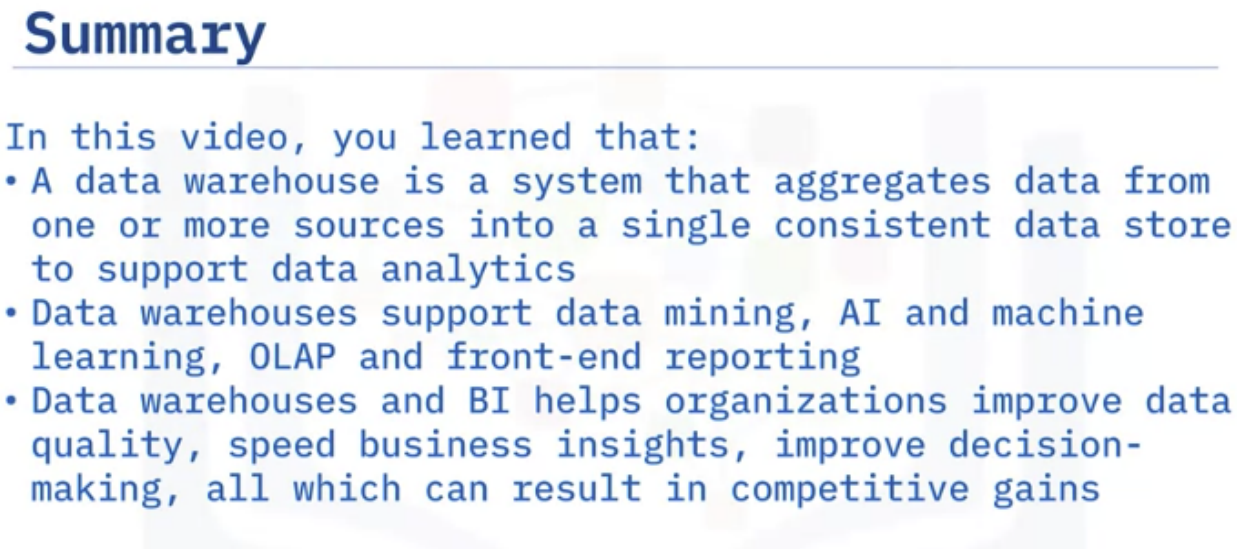
Pervasive Data Warehouse - Industry Usage



Benefits of a Data Warehouse

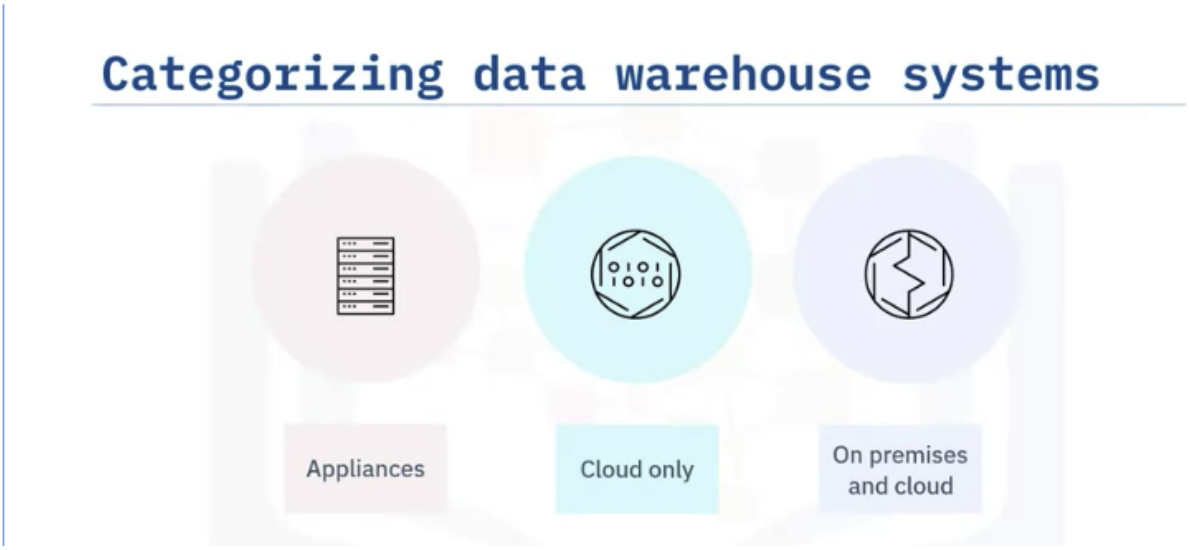


Quick Summary



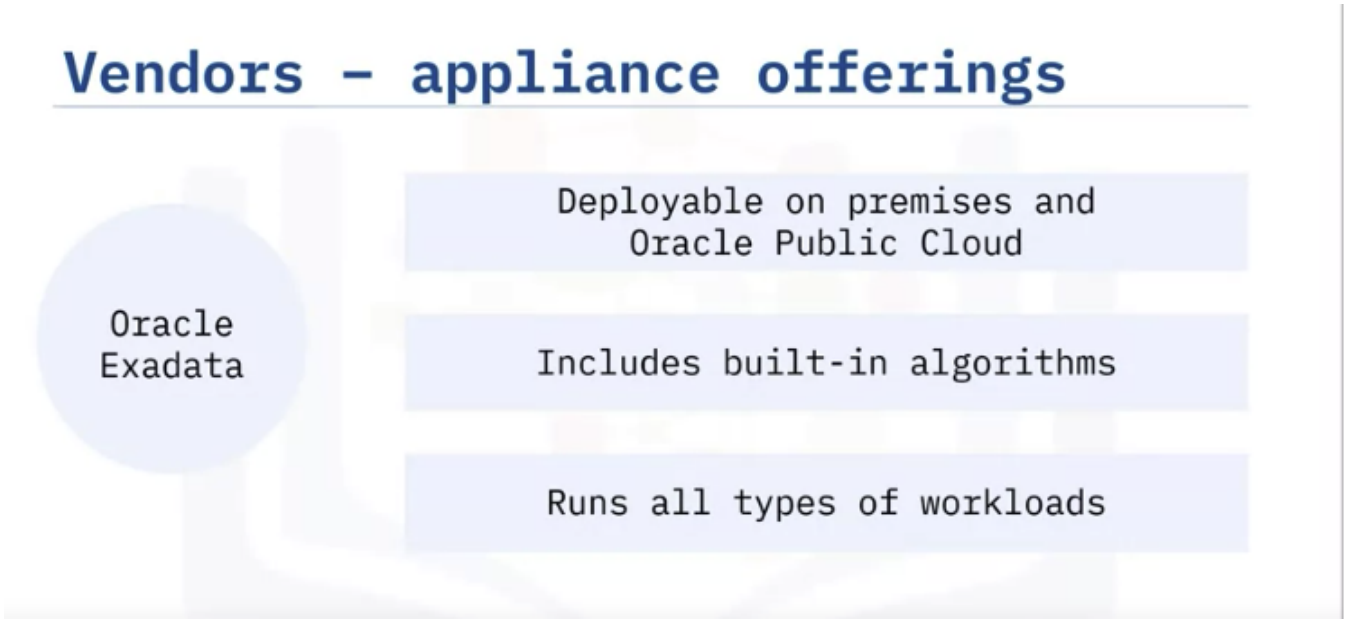
Popular Data Warehouse Systems

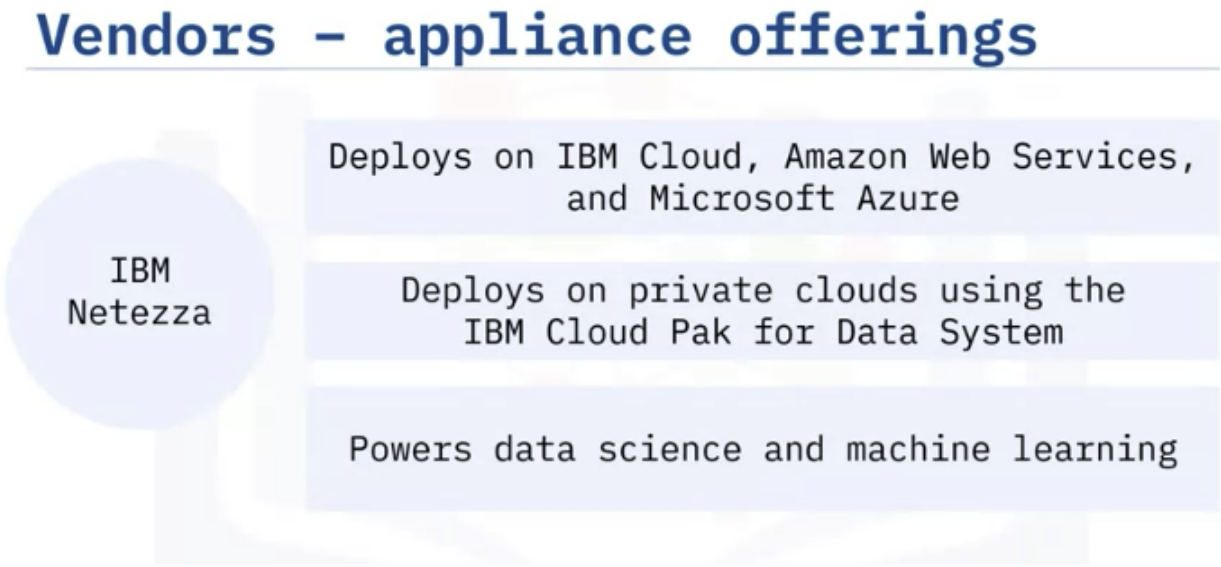
Categorizing Data Warehouse Systems



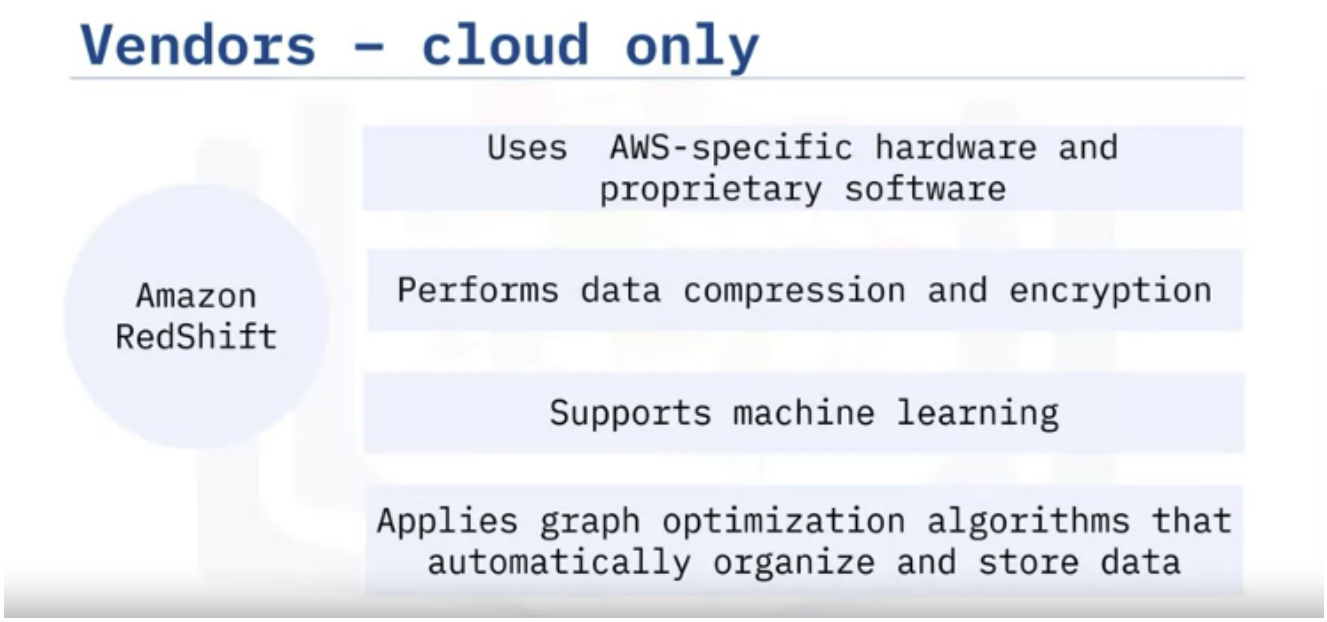
- Most data warehouse systems are supported via one or more of three platforms.
    - First are appliances, which are pre-integrated bundles of hardware and software that provide high performance for workloads and low maintenance overhead.
    - Other vendors support cloud deployments only, offering 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.

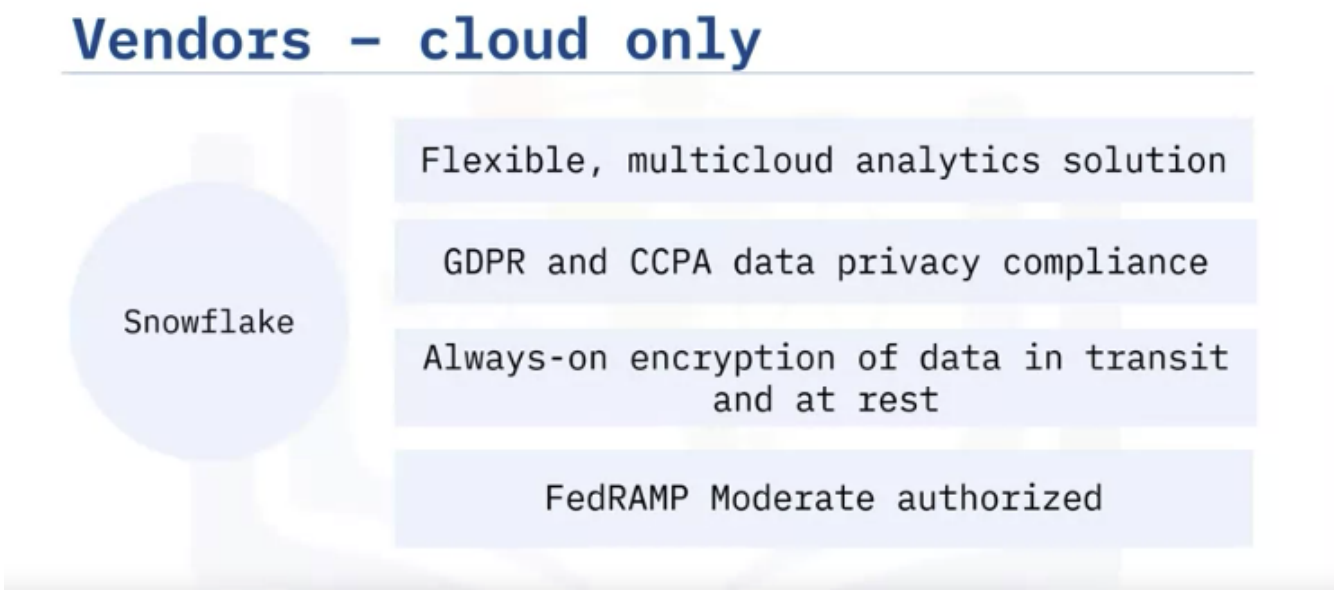
Popular Solutions - Vendors (Appliance Category)

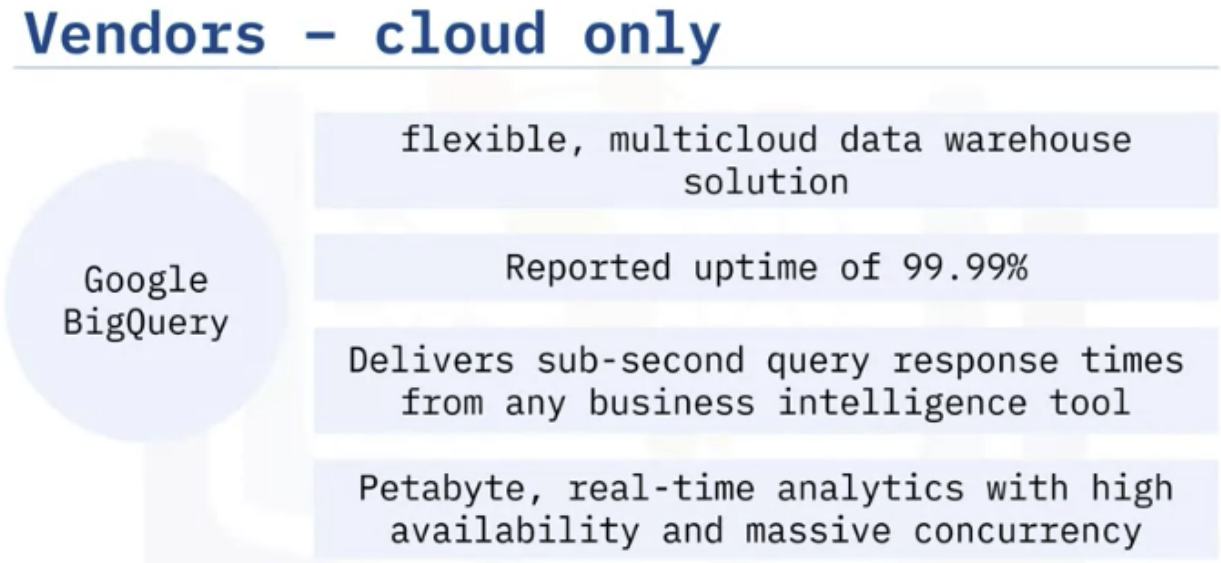




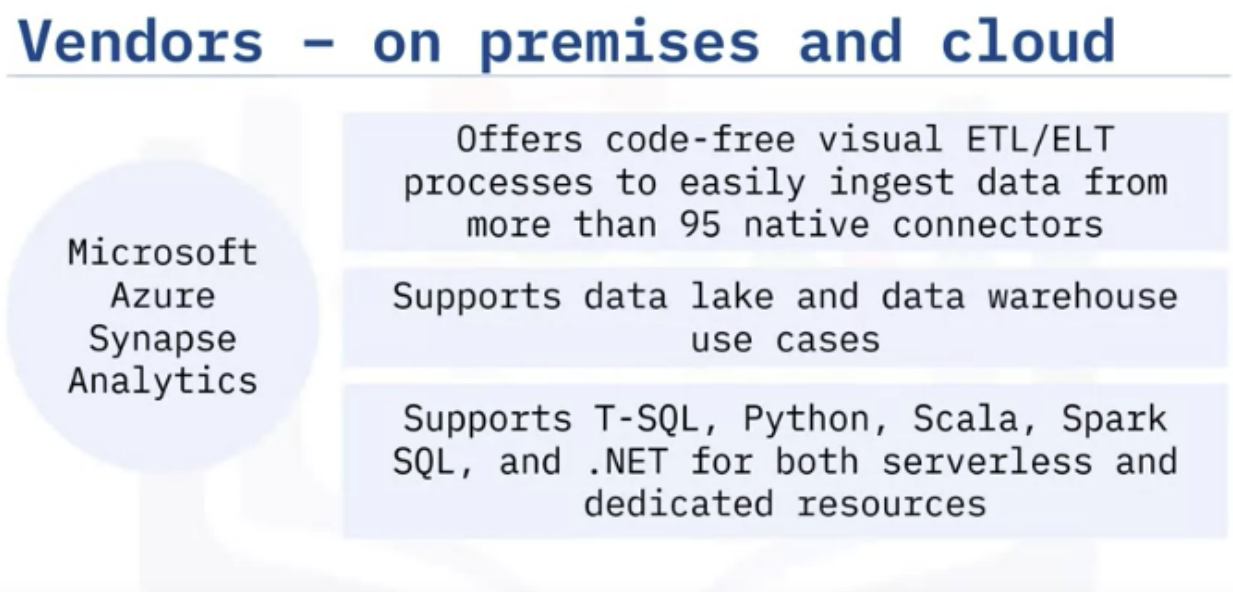
Vendors - Cloud Only







Vendors - On premises & Cloud



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

Vendors – on premises and cloud

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

Vendors – on premises and cloud

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

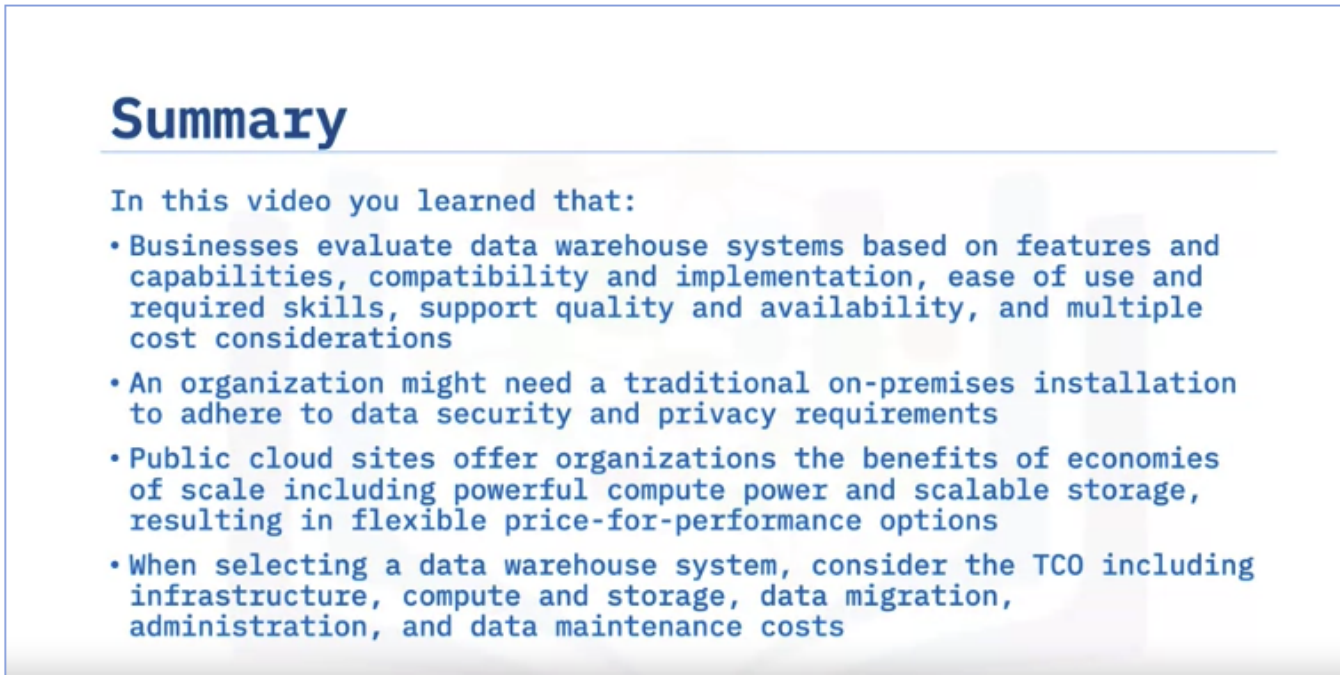
Vendors – on premises and cloud

Oracle Autonomous Data Warehouse	Available in both Oracle Public Cloud and on premises
	Supports multimodel data and multiple workloads
	Built to eliminate manual data management
	Autonomously secures data and performs threat detection

Selecting a Data Warehouse System (Summary)

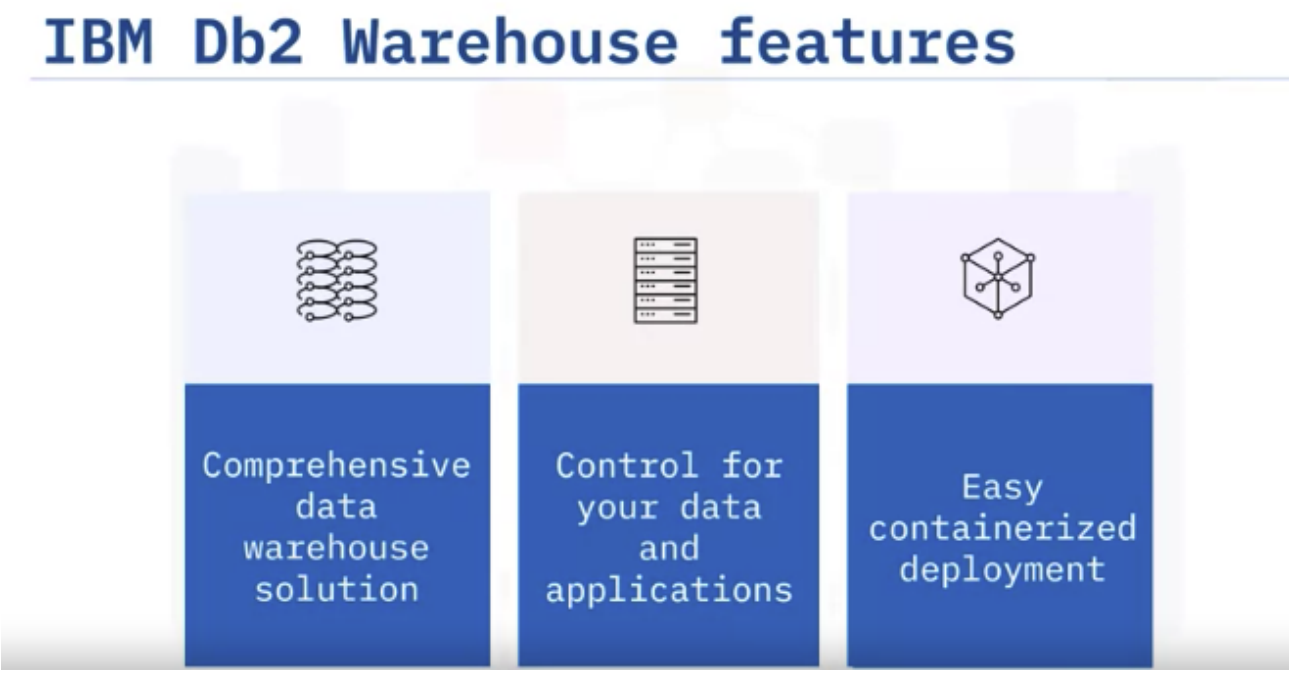
Selecting a Data Warehouse System

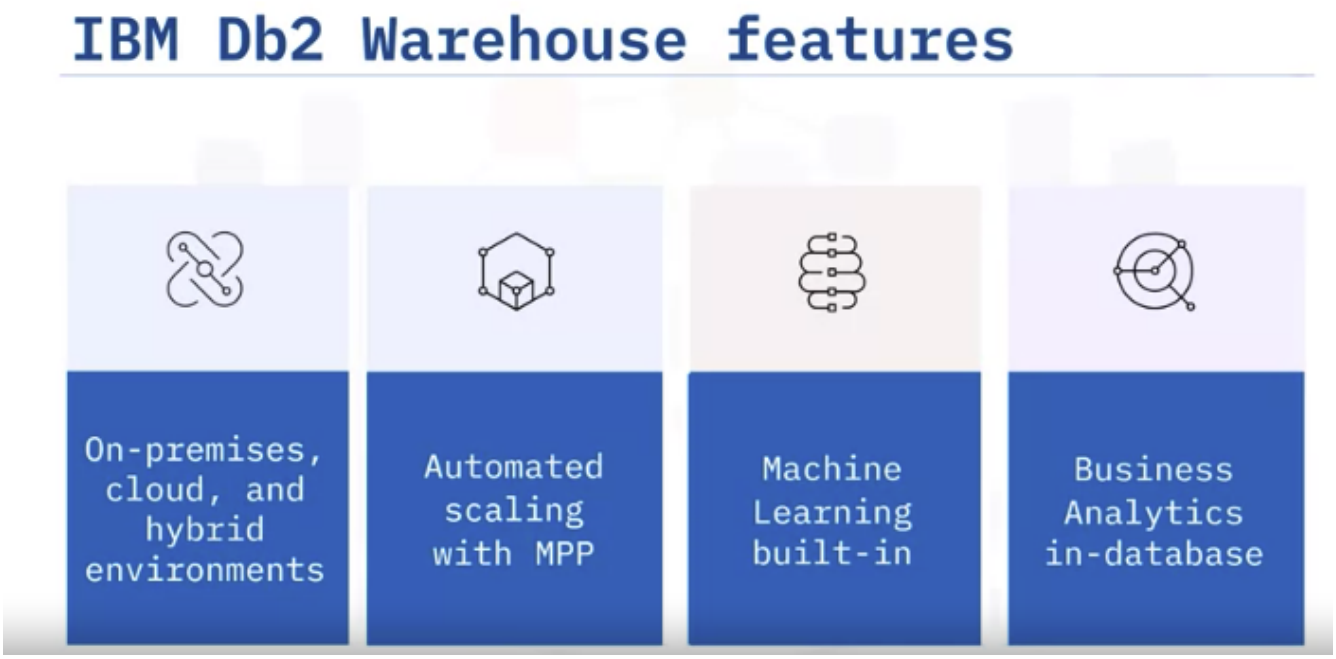
Notes Discuss



IBM Db2 Warehouse (Optional)

Features





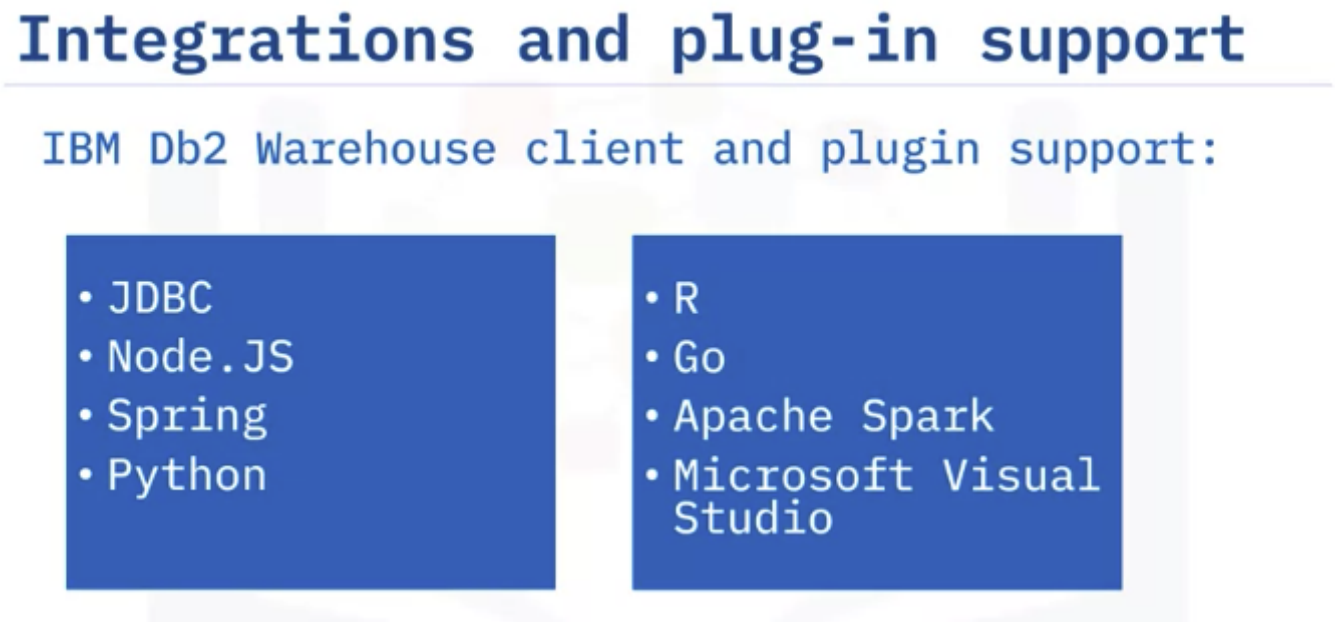
IBM Db2 Warehouse features

Use Cases

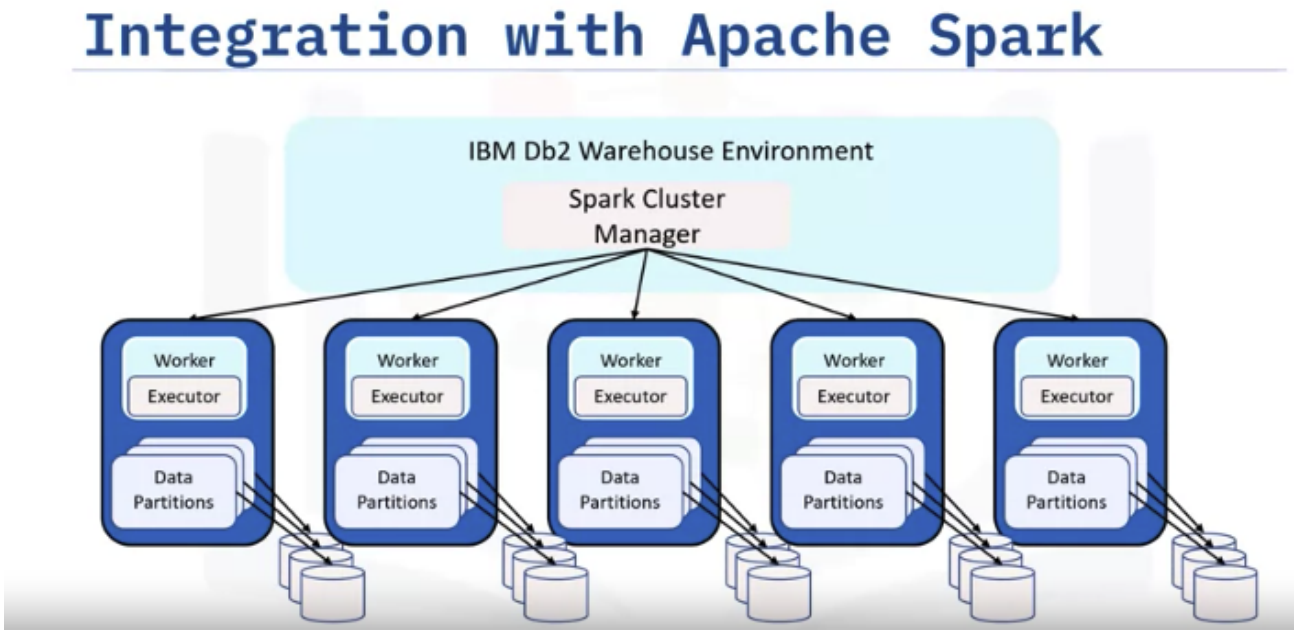
IBM Db2 Warehouse use cases

- High-scalability requirements
- Cloud, on-premises, or hybrid hosting
- Consolidation and integration of data silos
- Accelerated development of data marts
- Management of sensitive or regulated data
- Storage of cold SQL data

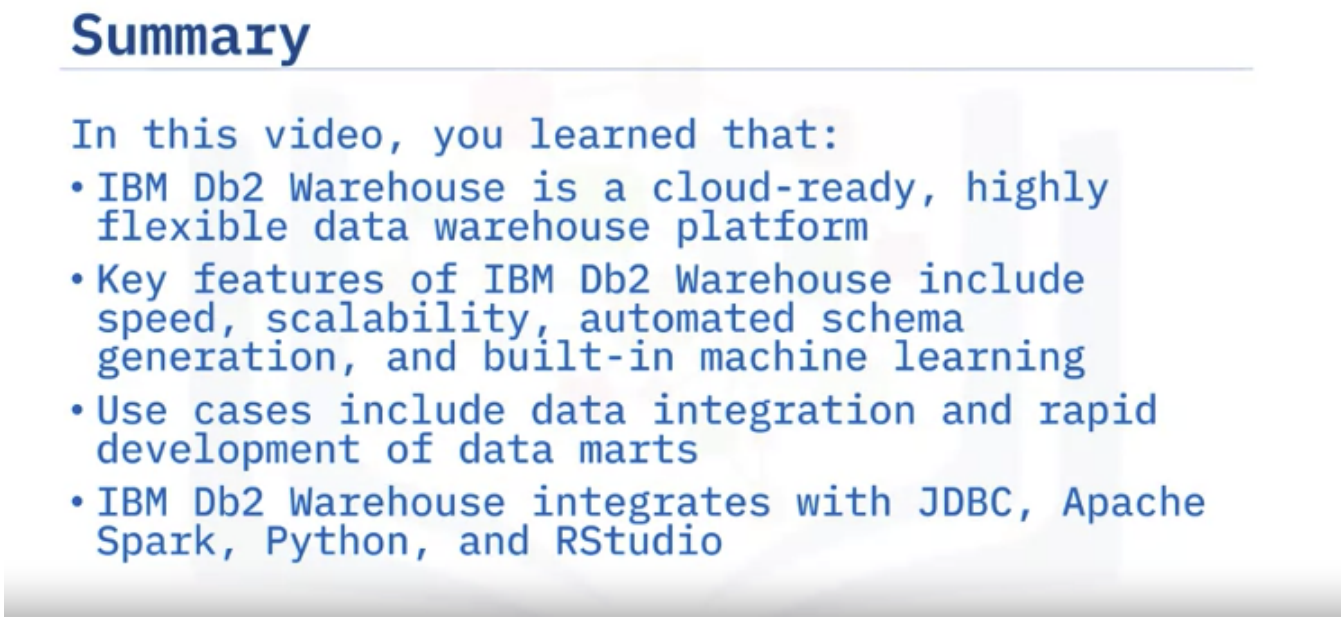
Integrations & Plug-In Support



Sample Integration Use Cases

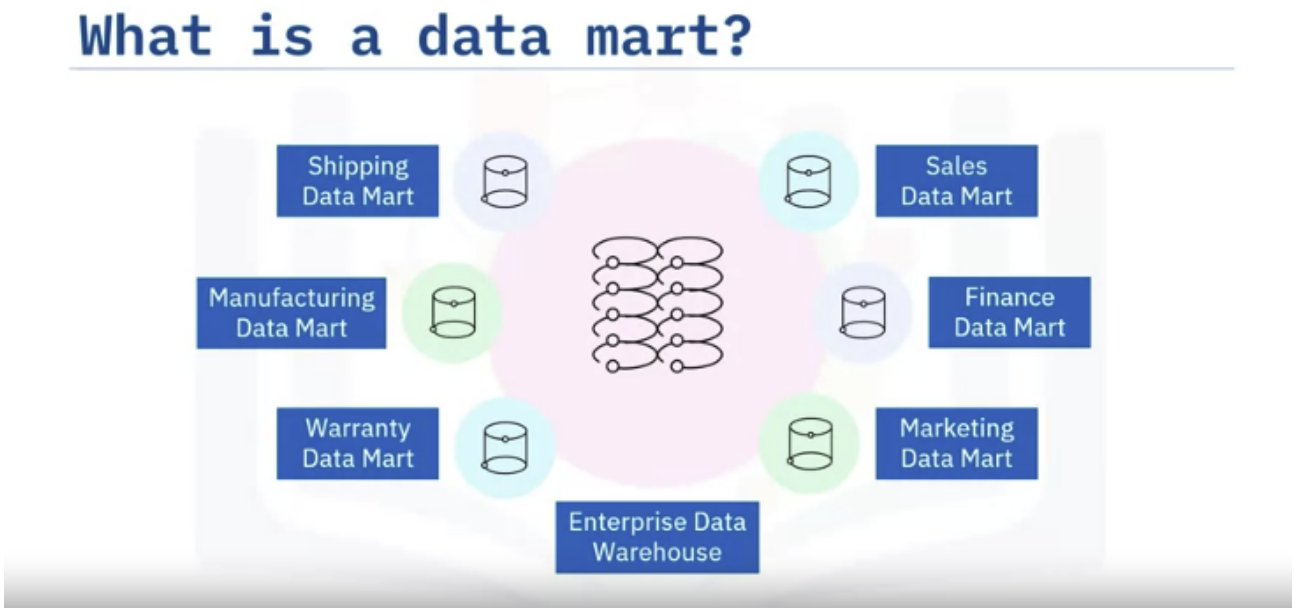


Summary

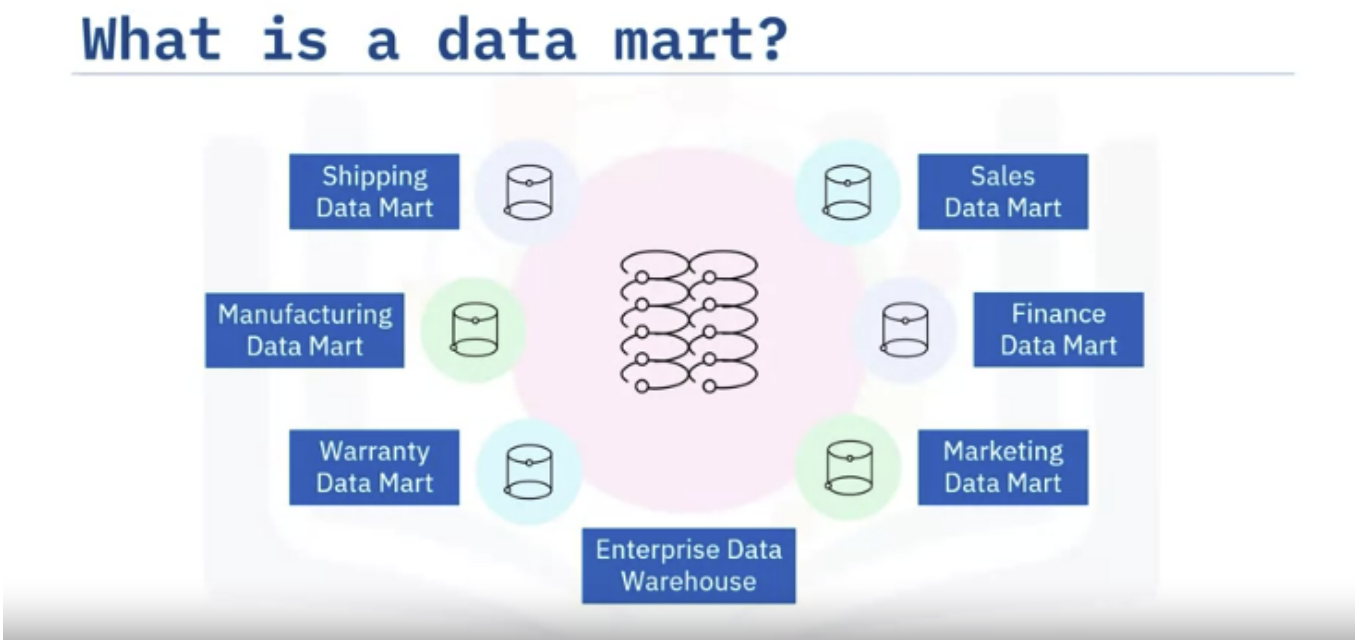


Data Marts Overview

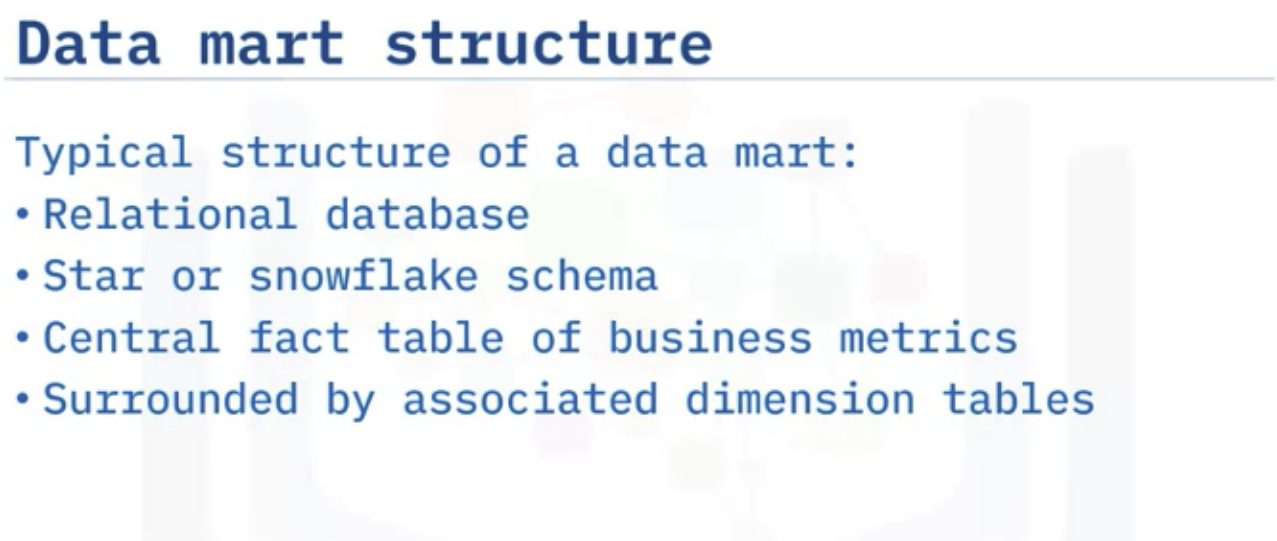
- A **data mart** is an isolated part of the larger enterprise data warehouse that is specifically built to serve a particular business function, purpose, or community of users.



Use Cases



Data Mart Structure



- The typical structure of a data mart is as follows:
    - It is a 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

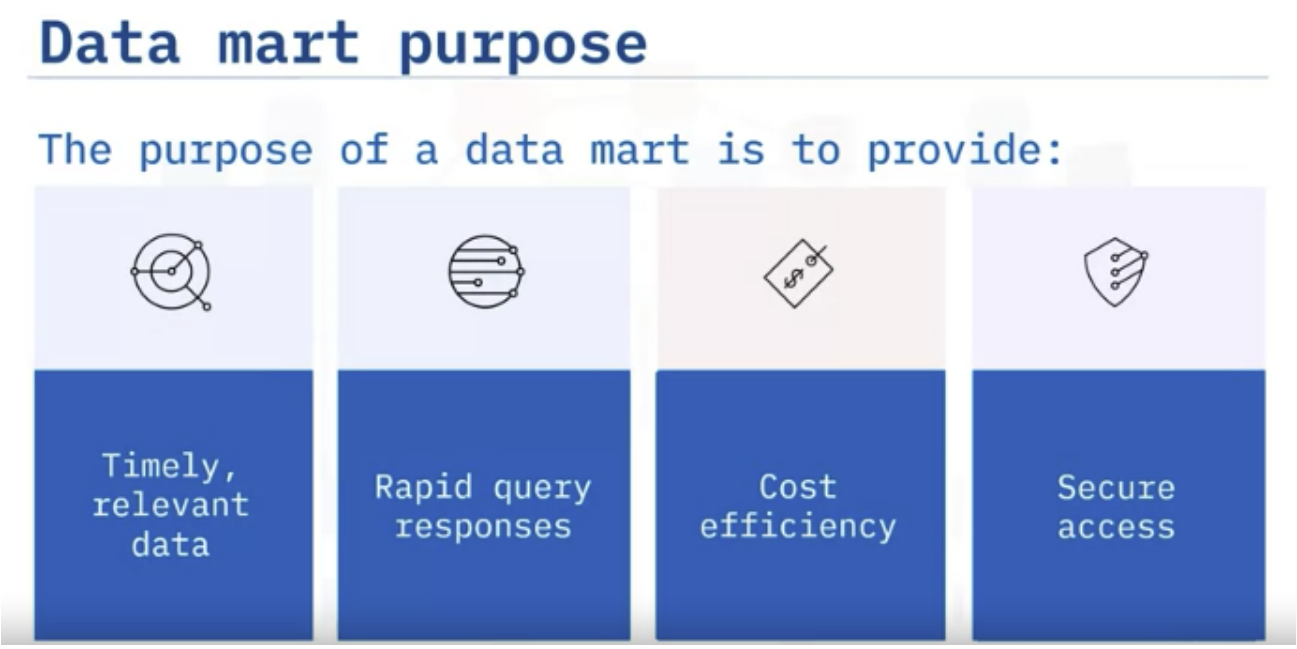
Data Marts	Databases
OLAP systems – read intensive	OLTP systems – write intensive
Use Txn DBs or warehouses as data sources	Use operational applications as sources of data
Contain clean, validated analytical data	Contain raw, unprocessed transactional data
Accumulate history for trend analysis	May not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

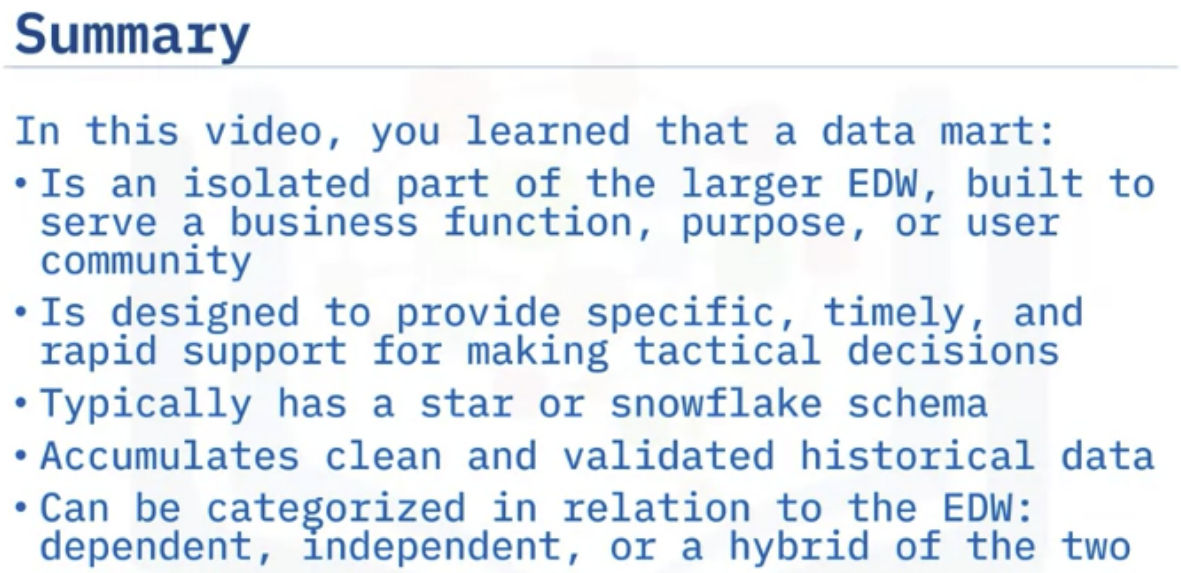
Types of Data Marts

- 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, while **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(s) Objectives



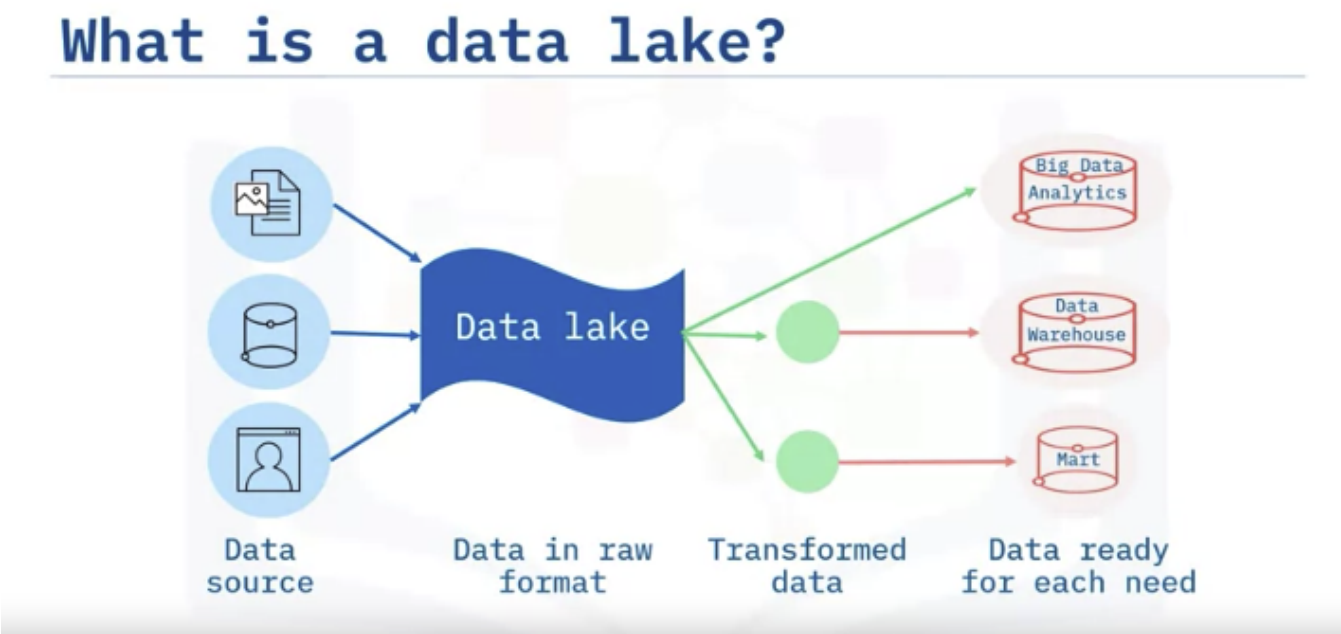
Summary



- EDW - Enterprise Data Warehouse

Data Lakes Overview

Data Lake Defined / Visual



- 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.

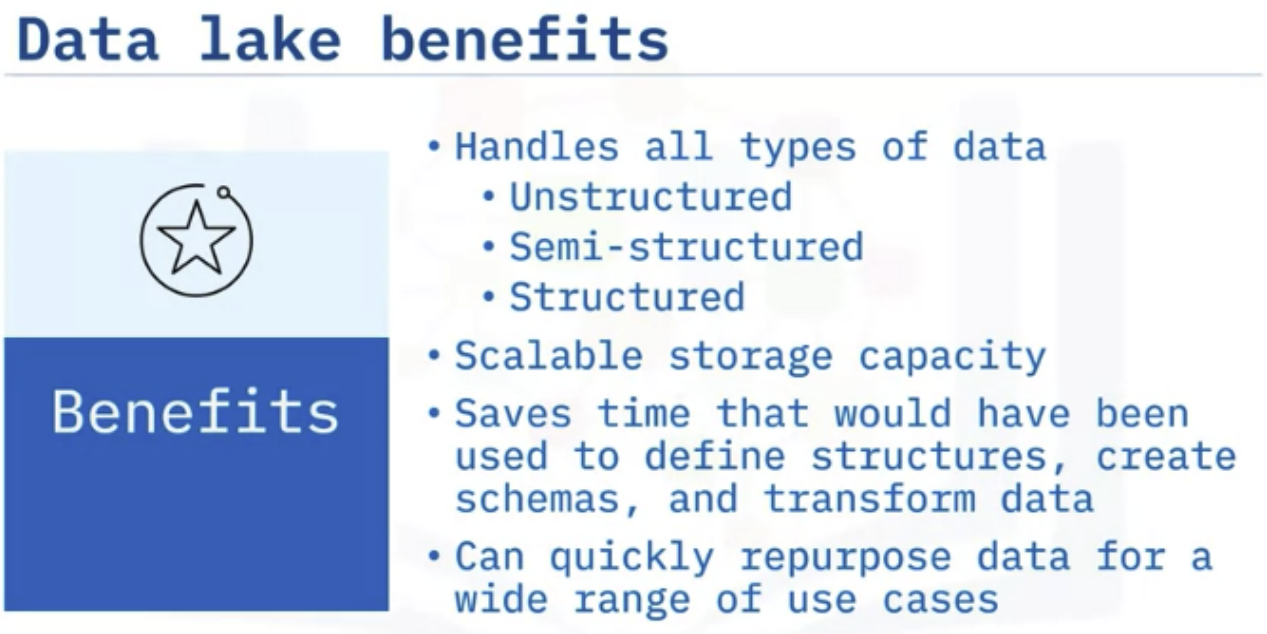
What is a data lake?

- Store large amounts of structured, semi-structured, and unstructured data in their native format
- Data can be loaded without defining the structure or schema of data
- Use cases do not need to be known in advance
- Exist as a repository of raw data straight from the source

What is a data lake?

- 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

Benefits of a Data Lake

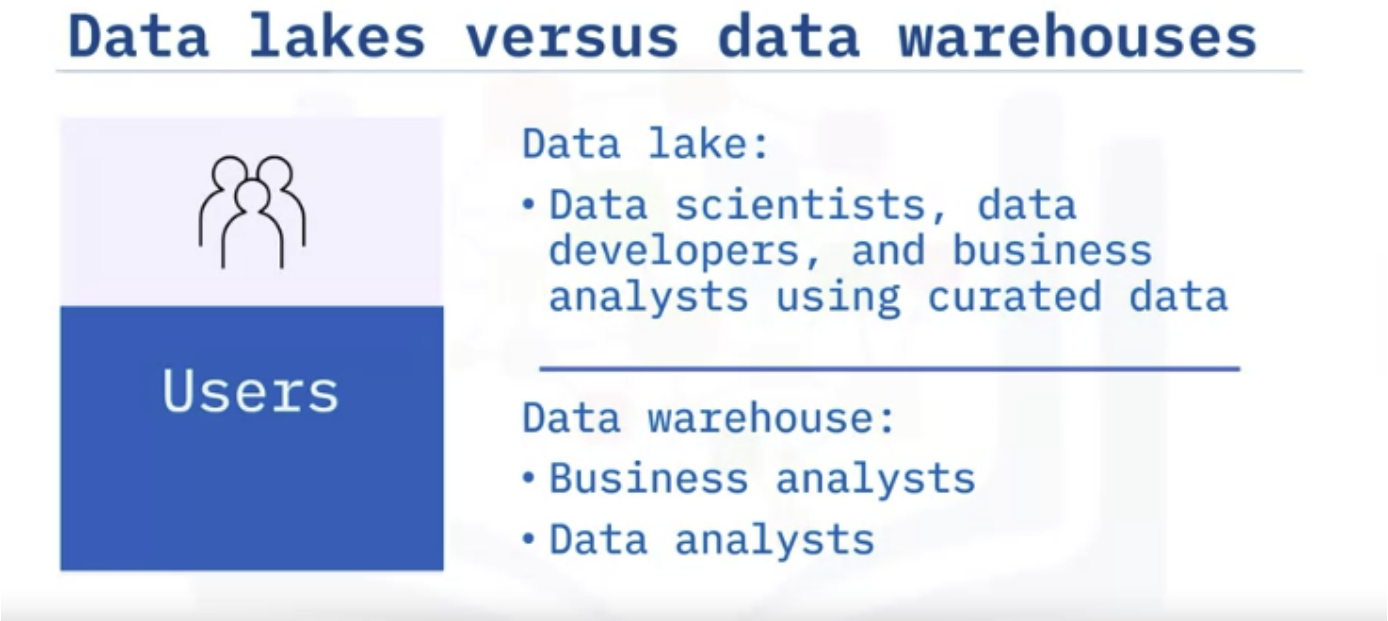


Data Lakes vs Data Warehouses

- When it comes to data, in a **data lake**, data is integrated in its raw and unstructured form.
- A data warehouse is different.
    - Here all data has already been processed and conformed to standards prior to loading to the warehouse.
- Talking about schema, when using data lakes, you do not need to define the structure and schema of the data before loading into the data lake.
- A data warehouse on the other hand requires strict conformance to schema and therefore a schema needs to be designed and implemented prior to loading the data

Data Quality

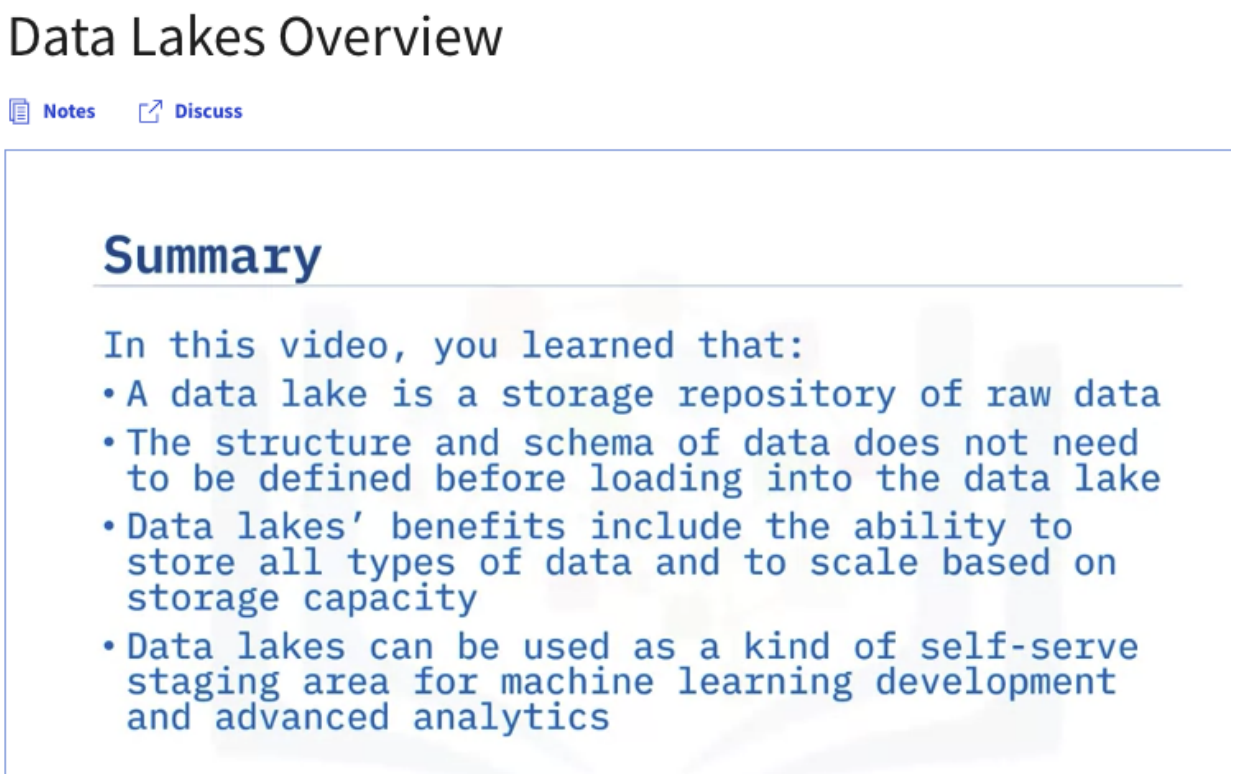
User Differences (General Environments Used)



- **Data curation** is the organization and integration of data collected from various sources.

- It involves annotation, publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation

Quick Summary



Data Lakehouses Explained

- let's take the best of both data lakes and data warehouses and combine them into a new technology called the data lake house.
- We get the flexibility and we get the cost-effectiveness of a data lake and we get the performance and structure of a data warehouse. We'll talk more specifically about the architecture of a data lake house in a future video, but from a value point of view, the lake house lets us store data from the exploding number of new sources in a low-cost way and then leverages built-in data management and governance leaders to allow us to power both business intelligence and high performance machine learning workloads quickly.