Big Data

Big Data Course

Eng.Mostafa Nabieh

Mostafa Nabieh

وزارة الاتصالات وتكنولوجيا المعلومات

MANNING

وزارة الاتصالات وتقنية المعلومات
MINISTRY OF COMMUNICATIONS AND INFORMATION TECHNOLOGY

AI TALENTS HUB

UDACITY

PLURALSIGHT

Learning Space
YOUR SPACE TO LEARN FUTURE SKILLS

BIG DATA

Mostafa Nabieh

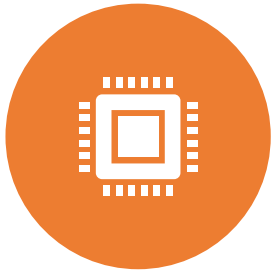
HUAWEI

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

IBM

Mostafa Nabieh

CONTENTS



WHAT IS DATA ENGINEERING?

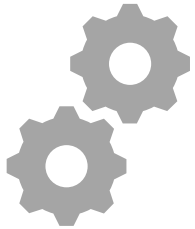
BIG DATA ECOSYSTEM

BIG DATA LIFECYCLE

CAREER OPPORTUNITIES

Data Repository

Data Repository is a general term used to refer to data that has been collected, organized, and isolated



for use in business operations

mined for reporting operations and data analysis

Data Repository

Types of data repositories include:

- Databases
- Data Warehouses
- Big Data Stores

Databases

- Collection of data for input, storage, search, retrieval, and modification of data.
- Set of programs for creating and maintaining the database, and storing, modifying, and extracting information from the database.
- Even though a database and DBMS mean different things the terms are often used interchangeably.
- Factors governing choice of database include:
  - Data type
  - Data structure
  - Querying mechanisms
  - Latency requirements
  - Transaction speeds
  - Intended use of data

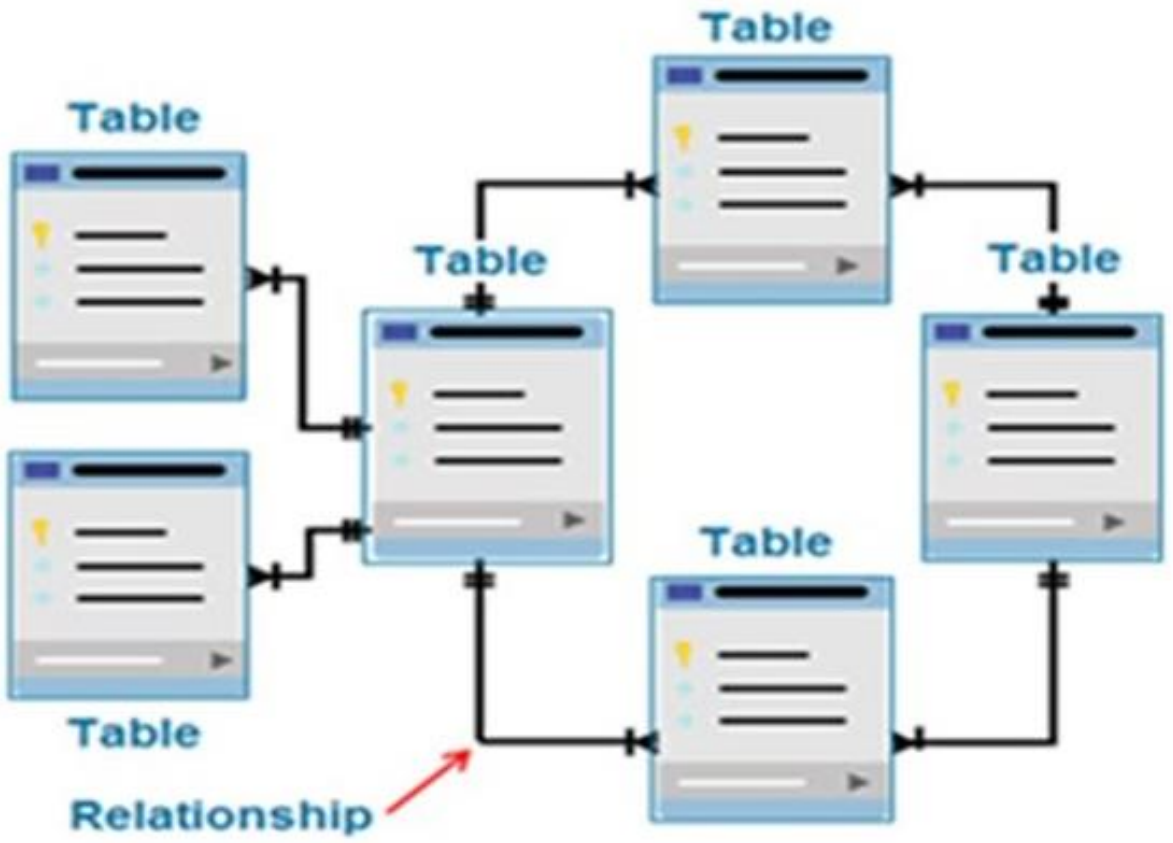
Relational

VS

Non-relational

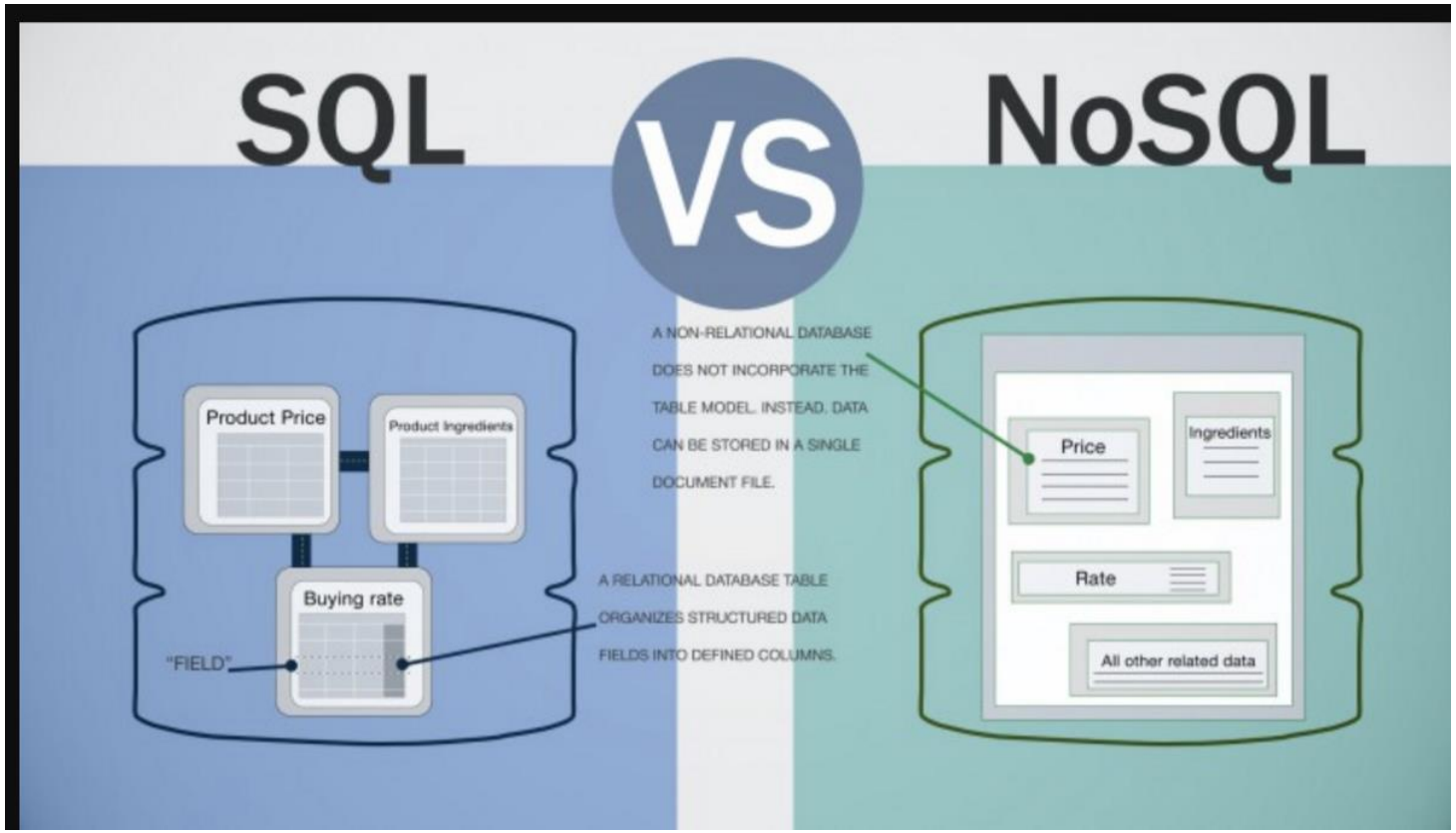
Relational database

- Data is organized into a tabular format with rows and columns
- Well-defined structure and schema
- Optimized for data operations and querying
- Use SQL as the standard querying language



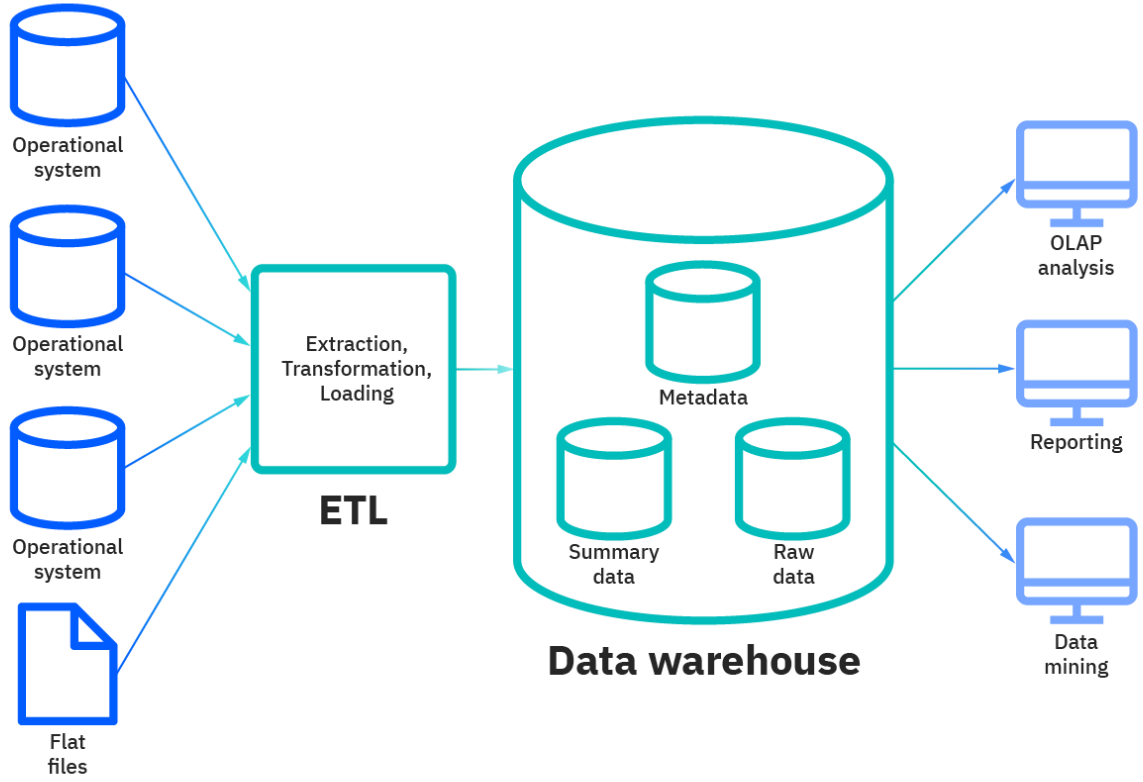
Non-Relational Databases

- Emerged in response to the volume, diversity, and speed at which data is being generated today
- Built for speed, flexibility, and scale
- Data can be stored in a schema-less form
- Widely used for processing big data



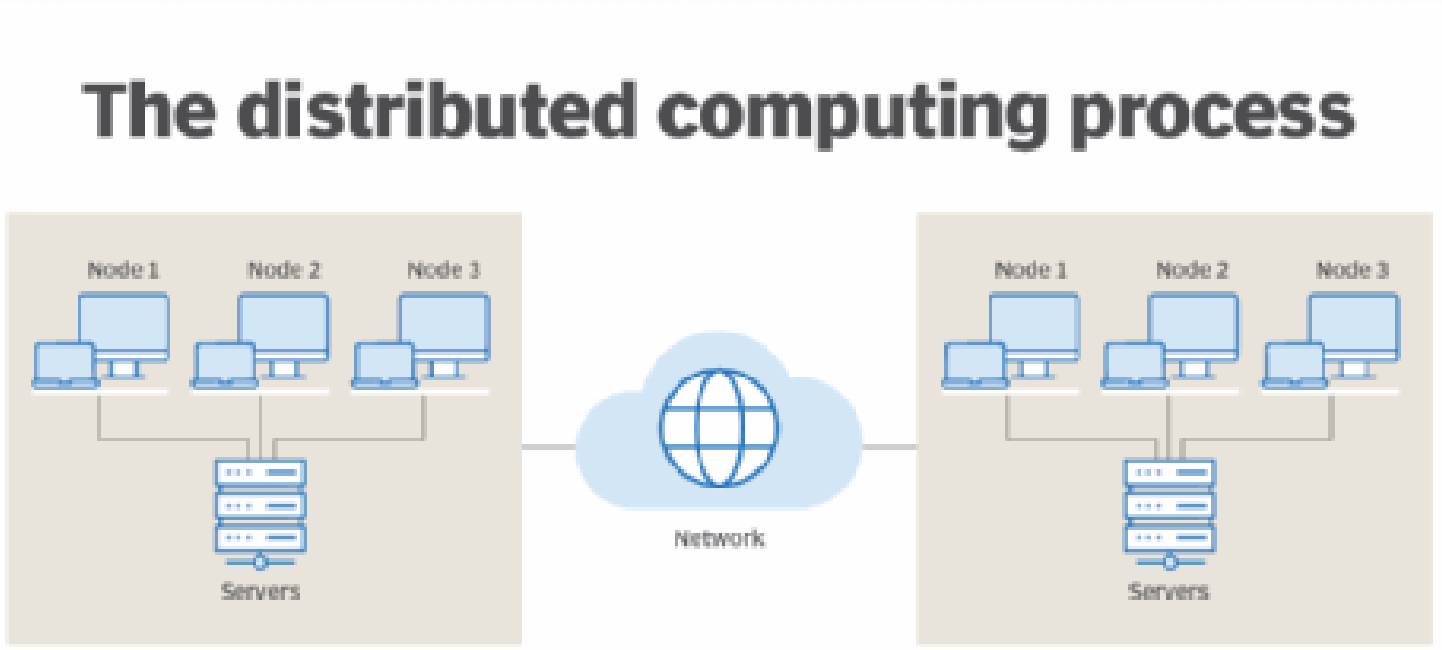
Data Warehouse

- Consolidates data through the extract, transform, and load process, also known as the ETL process, into one comprehensive database for analytics and business intelligence.

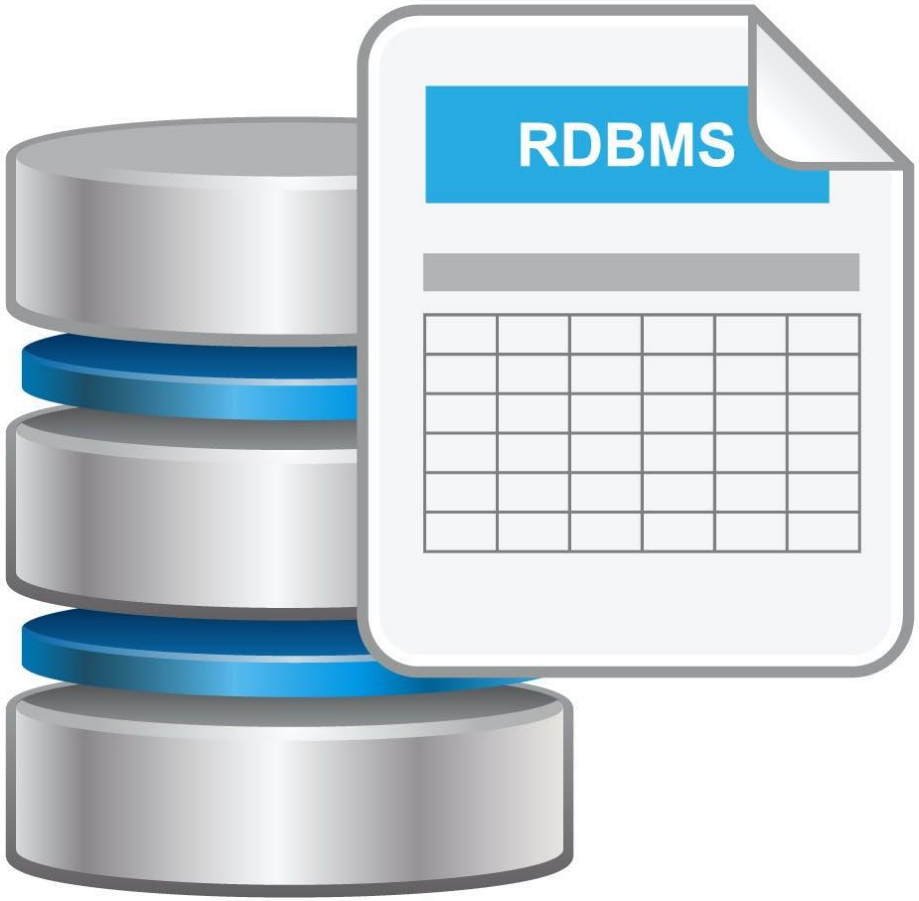


Big Data Stores

- Distributed computational and storage infrastructure to store, scale, and process very large data sets.

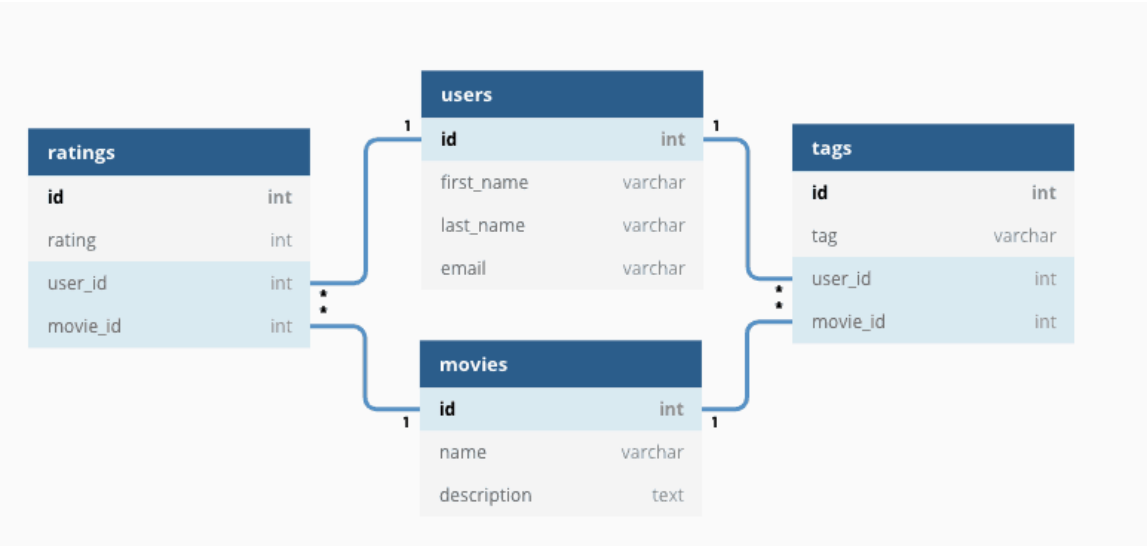


RDBMS



What is a Relational Database?

- Relational databases use structured query language, or SQL, for querying data.
- Similarities between relational databases and spreadsheets: Relational databases build on the organizational principles of flat files such as spreadsheets, with data organized into rows and columns following a well-defined structure and schema.



Relational Database

- Ideal for the optimized storage, retrieval, and processing of data for large volumes of data
- Each table has a unique set of rows and columns
- Relationships can be defined between tables
- Fields can be restricted to specific data types and values
- Can retrieve millions of records in seconds using SQL for querying data
- Security architecture of relational databases provides greater access control and governance

Examples of RDBMS

- Relational Databases can be:
- Open-source with internal support Open-source with commercial support
- Commercial closed-source

- Cloud-Based Relational Databases, or Database-as-a-Service

Advantages of Relational Databases

- **Create meaningful information** by joining tables
- **Flexibility** to make changes while the database is in use
- **Minimize data redundancy** by allowing relationships to be defined between tables
- Offer export and import options that provide **ease of backup and disaster recovery**
- Are **ACID compliant**, ensuring accuracy and reliability in database transactions

Relational Databases are well suited for

Online Transaction Processing (OLTP) application Can support transaction-oriented tasks that run at high rates and

- Accommodate large numbers of users
- Manage small amounts of data
- Support frequent queries and fast response times

Data Warehouses

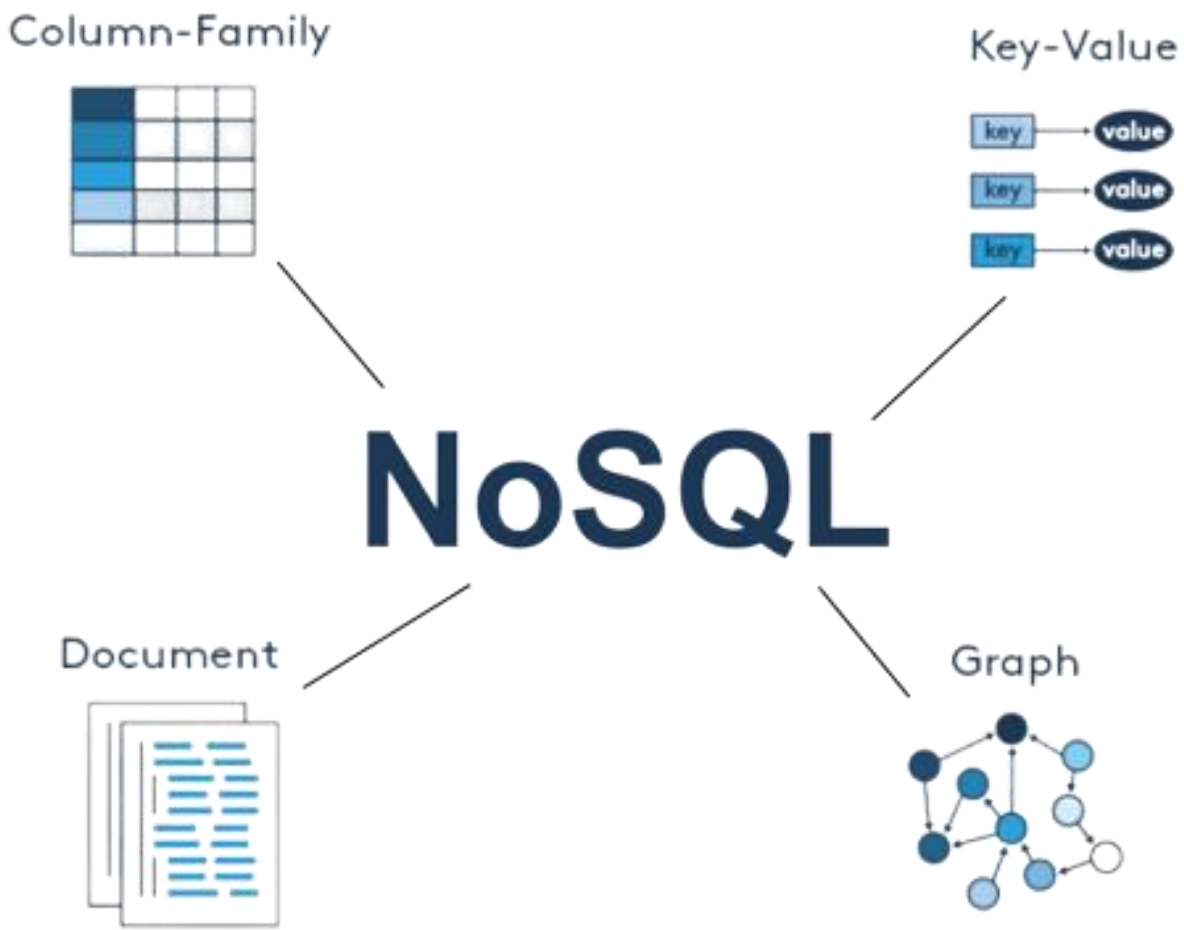
- Can be optimized for online analytical processing (OLAP)

IoT Solutions

- Provide the speed and ability to collect and process data from edge devices

Limitations of RDBMS

- Does not work well with semi-structured and unstructured data
- Migration between two RDBMS's is possible only when the source and destination tables have identical schemas and data types
- Entering a value greater than the defined length of a data field results in a loss of information


Column-Family
Key-Value
key
value
key
value
key
value
NoSQL
Document
Graph


NoSQL

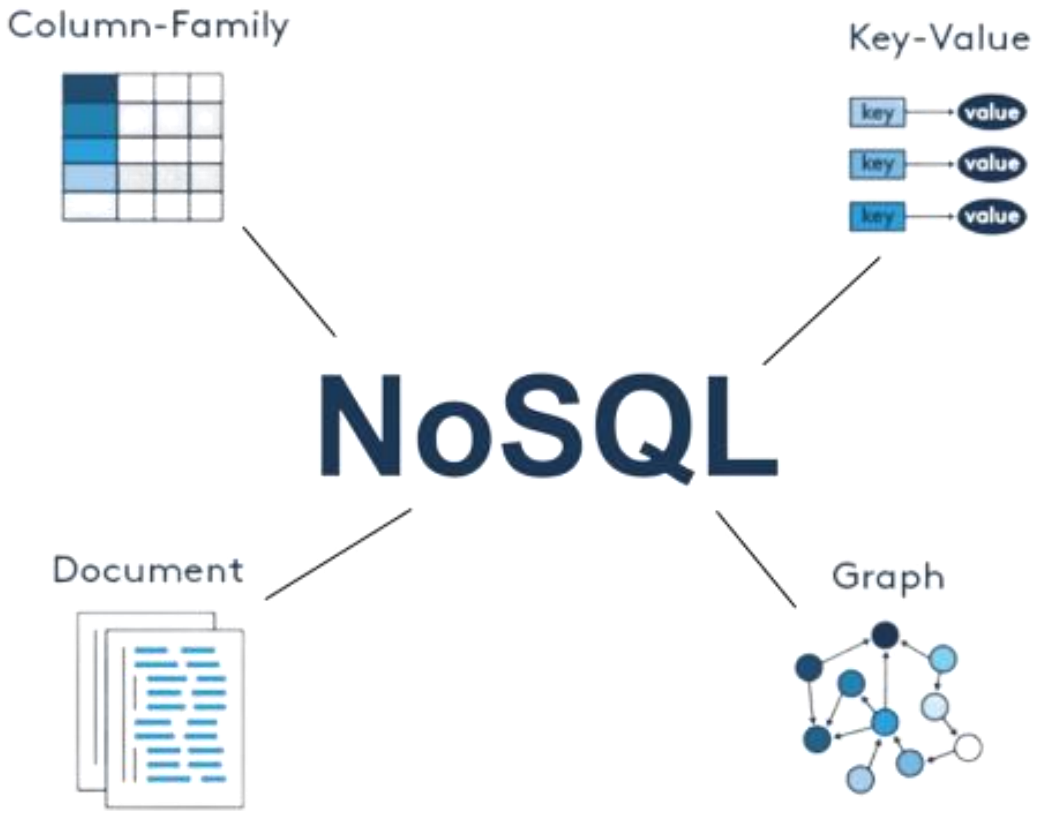
What is a NoSQL database?

NoSQL (not only SQL) or Non-SQL is a non-relational database design that provides flexible schemas for the storage and retrieval of data

- Gained greater popularity due to the emergence of cloud computing, big data, and high-volume web and mobile applications
- Chosen for their attributes around scale, performance, and ease of use
- Built for specific data models
- Has flexible schemas that allow programmers to create and manage modern applications
- Do not use a traditional row/column/table database design with fixed schemas
- Do not, typically, use the structured query language (or SQL) to query data

Types of NoSQL

- Based on the model being used for storing data, there are four common types of NoSQL databases:
    - Key-value store
    - Document Based
    - Column Based
    - Graph Based



Key-value store

Based on the model being used for storing data, there are four common types of NoSQL databases:

Key-value store:

- Data in a key-value database is stored as a collection of key-value pairs.
- A key represents an attribute of the data and is a unique identifier.
- Both keys and values can be anything from simple integers or strings to complex JSON documents.
- Great for storing user session data, user preferences, real-time recommendations, targeted advertising, in-memory data caching.

Key-value store

Not a great fit if you want to:

- Query data on specific data value
- Need relationships between data values
- Need multiple unique keys

Phone directory

Key	Value
Paul	(091) 9786453778
Greg	(091) 9686154559
Marco	(091) 9868564334

MAC table

Key	Value
10.94.214.172	3c:22:fb:86:c1:b1
10.94.214.173	00:0a:95:9d:68:16
10.94.214.174	3c:1b:fb:45:c4:b1

Document-Based

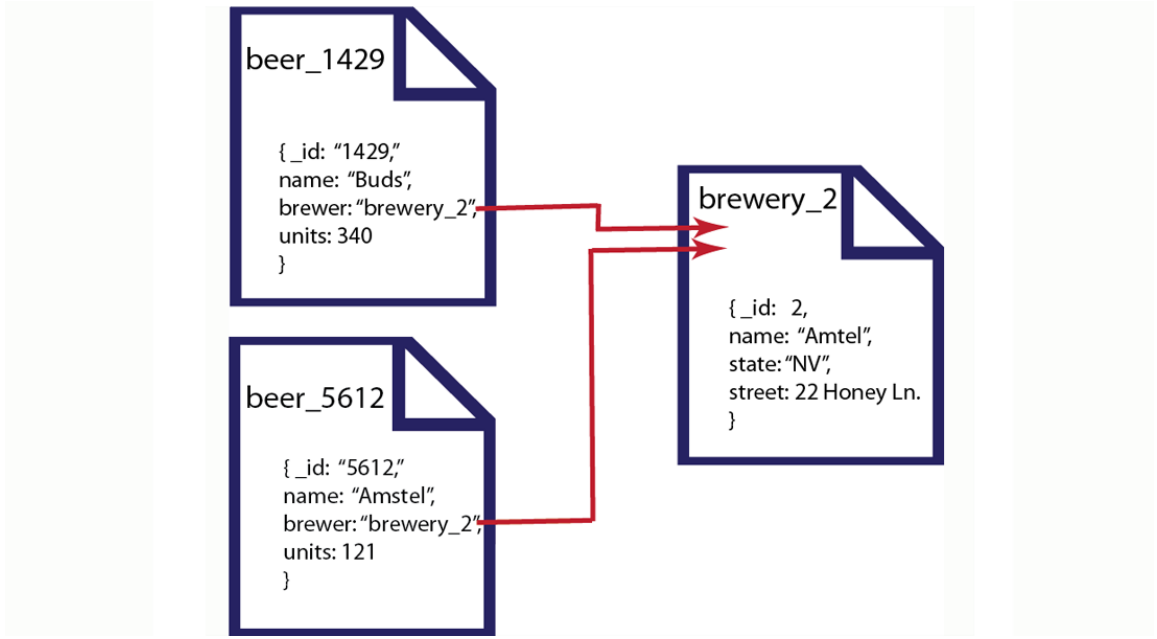
- Document databases store each record and its associated data within a single document.
- They enable flexible indexing, powerful ad hoc queries, and analytics over collections of documents.
- Preferred for eCommerce platforms, medical records storage, CRM platforms, and analytics platforms.

Document-Based

Not a great fit if you want to:

- Query data on specific data value
- Need relationships between data values
- Need multiple unique keys

MongoDB DocumentDB CouchDB Cloudant



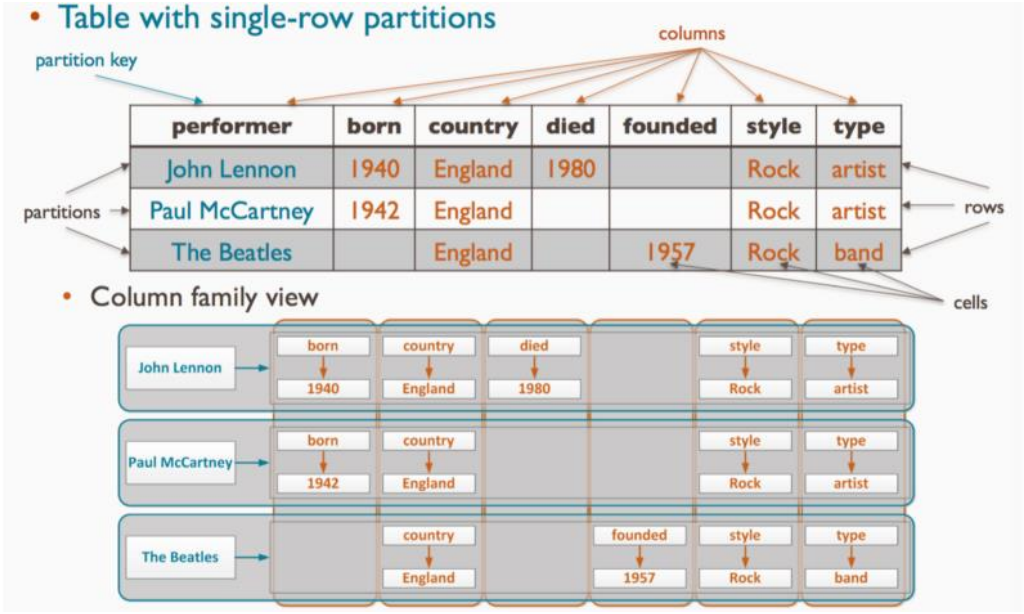
Column-Based

- Data is stored in cells grouped as columns of data instead of rows.
- A logical grouping of columns is referred to as a column family.
- All cells corresponding to a column are saved as a continuous disk entry, making access and search easier and faster.
- Great for systems that require heavy write requests, storing time-series data, weather data, and IoT data.

Column-Based

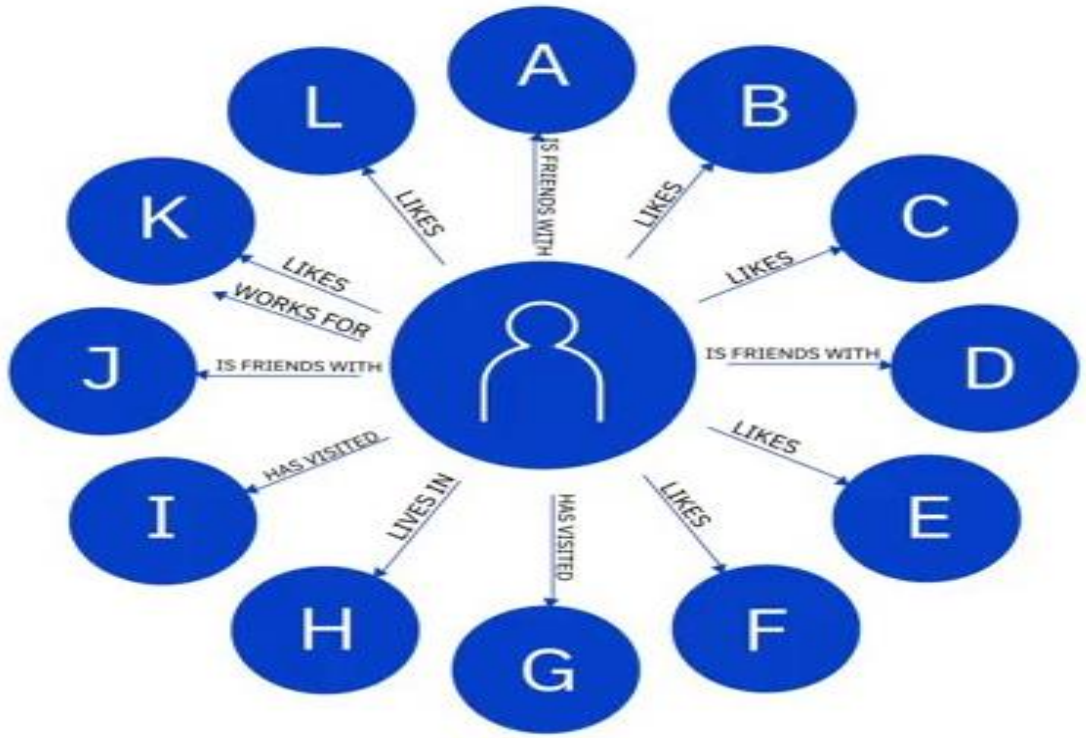
Not a great fit if you want to:

- Run complex queries
- Change querying patterns frequently



Graph-Based

- Graph-based databases use a graphical model to represent and store data.
- Useful for visualizing, analyzing, and finding connections between different pieces of data.



An excellent choice for working with connected data.

Graph-Based

Not a great fit if you want to:

- Process high volumes of transactions

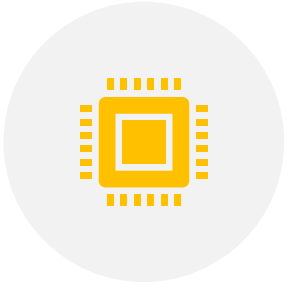
Neo4J

CosmosDB

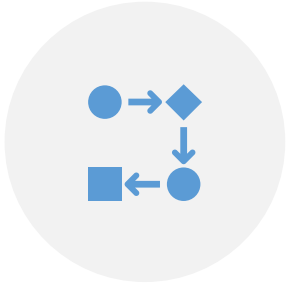
Advantages of NoSQL

Its ability to handle large volumes of structured, semi-structured, and unstructured data

Its ability to run as a distributed system scaled across multiple data centers



An efficient and cost-effective scale-out architecture that provides additional capacity and performance with the addition of new nodes



Simpler design, better control over the availability, and improved scalability that makes it agile, flexible, and supports quick iterations

SQL

Relational Data Model

Pros
- Easy to use and setup.
- Universal, compatible with many tools.
- Good at high-performance workloads.
- Good at structure data.

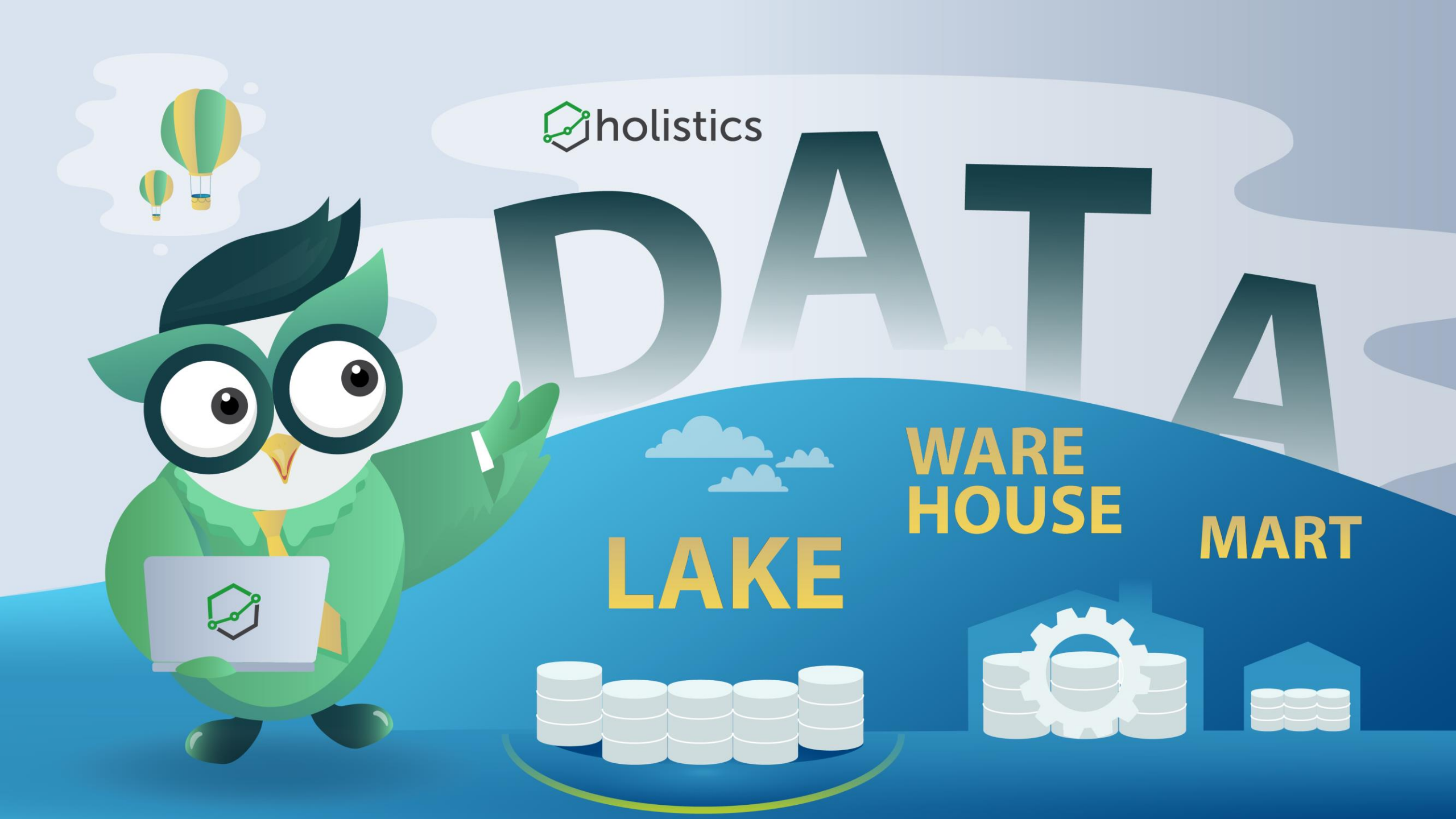
Cons
- Time consuming to understand and design the structure of the database.
- Can be difficult to scale.

No SQL

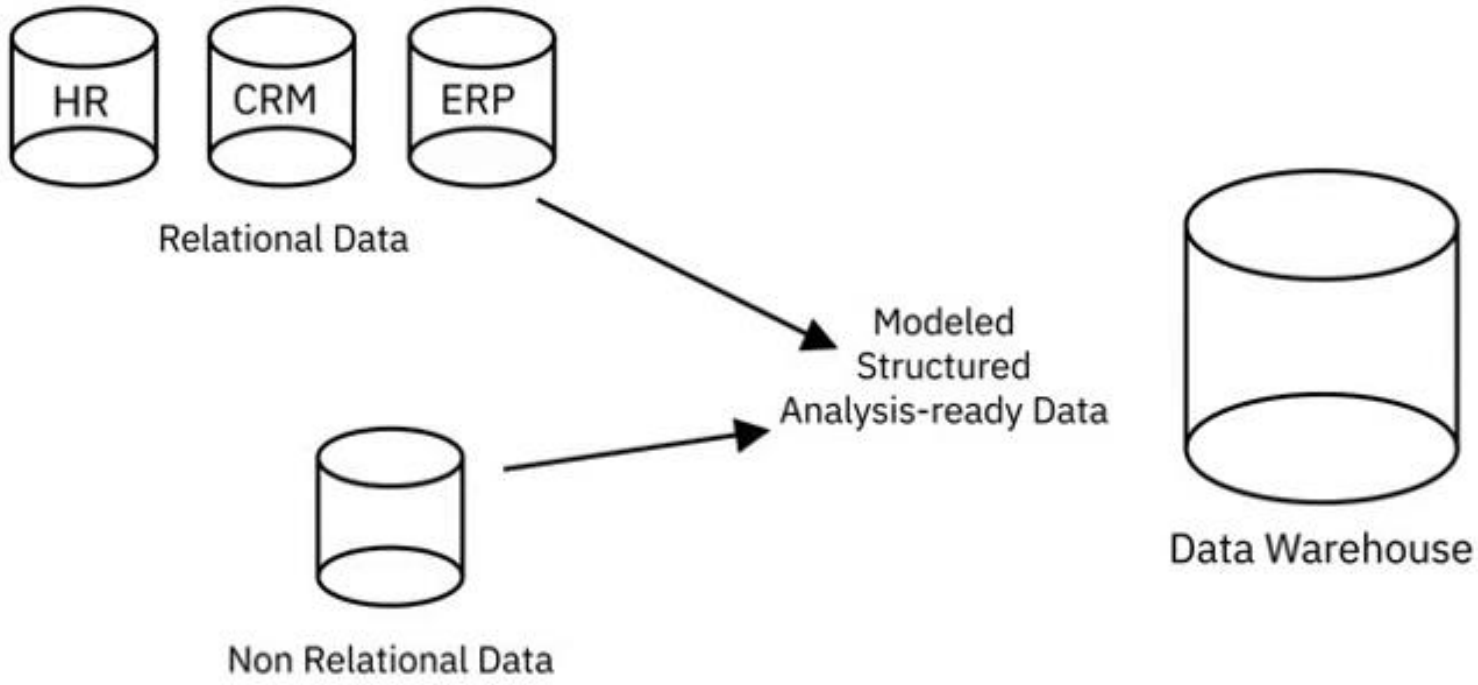
Document Data Model

Pros
- No investment to design model.
- Rapid development cycles.
- In general faster than SQL.
- Runs well on the cloud.

Cons
- Unsuited for interconnected data.
- Technology still maturing.
- Can have slower response time.

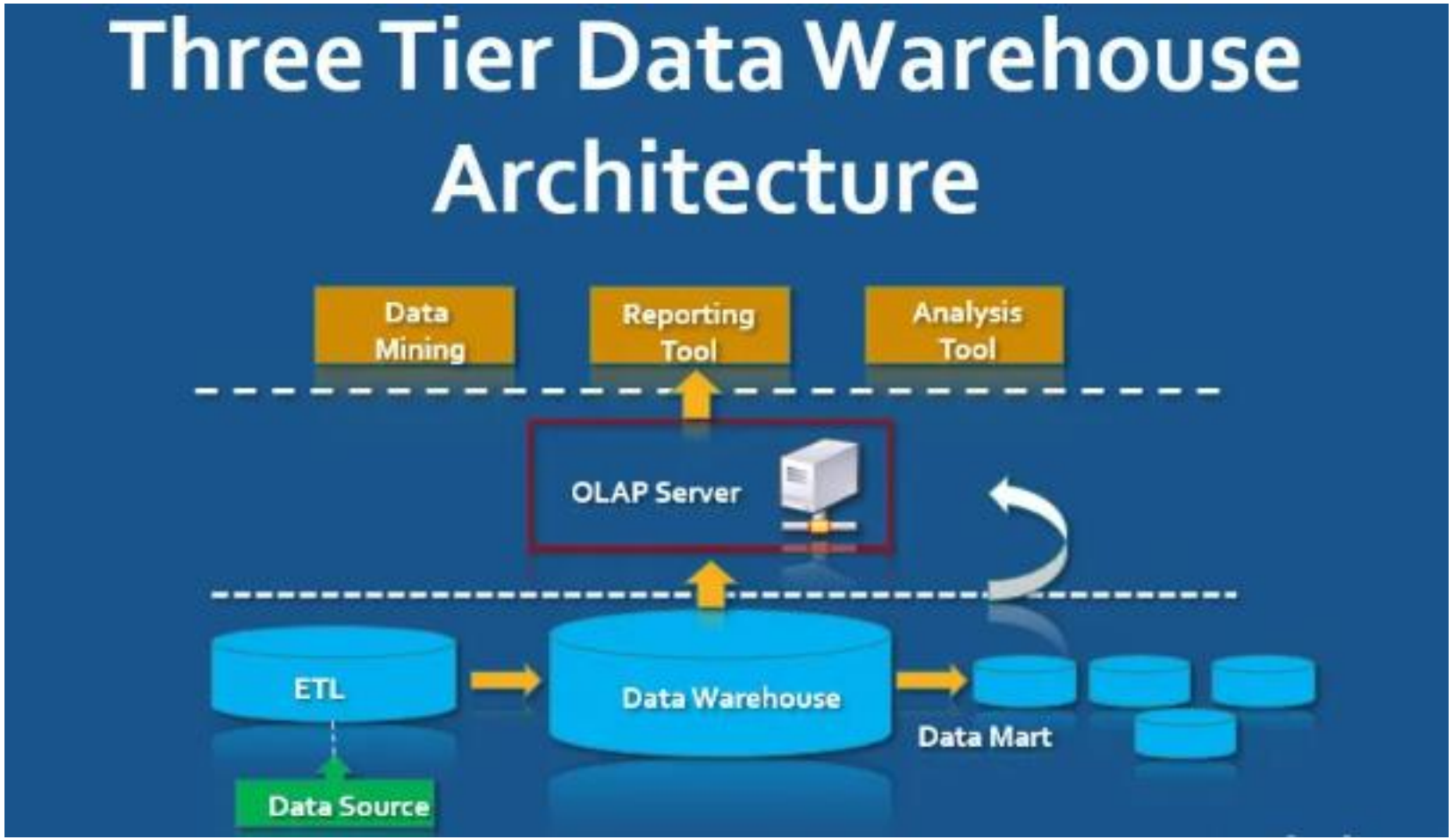

holistics
DATA
LAKE
WARE
HOUSE
MART

Data Warehouses



- Relational data from transactional systems and operational databases
- Non-relational data

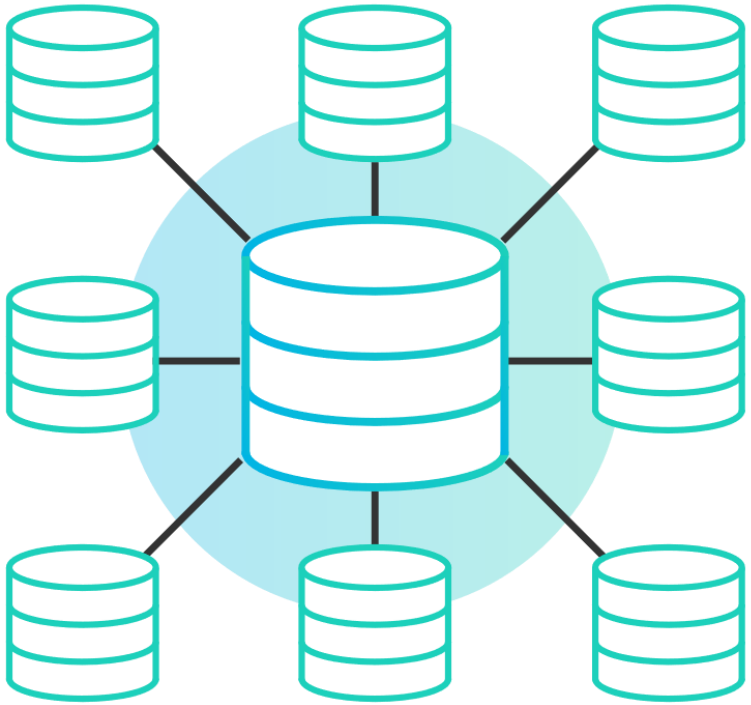
Data Warehouses



Data Warehouses

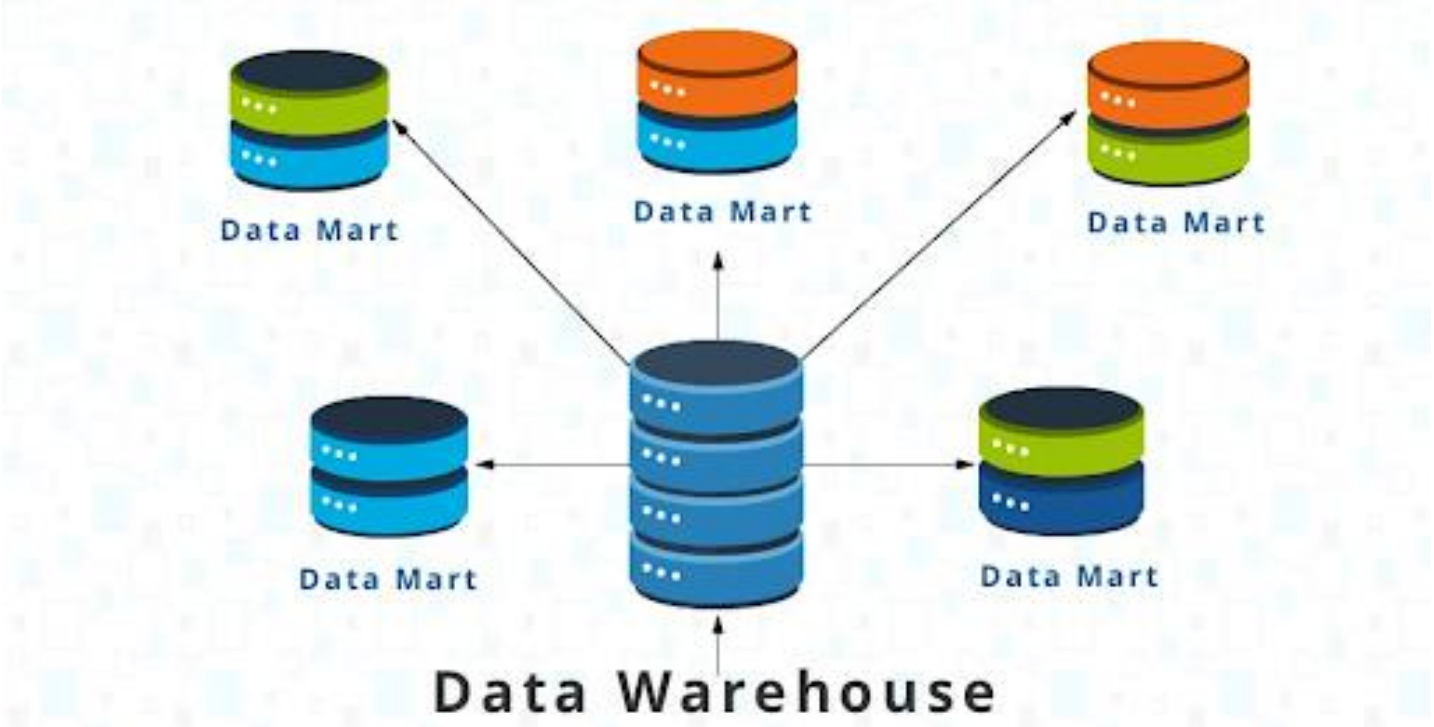
Benefits of cloud-based data warehouses:

- Lower costs
- Limitless storage and computing capabilities
- Scale on a pay-as-you-go basis
- Faster disaster recovery



Data Warehouses

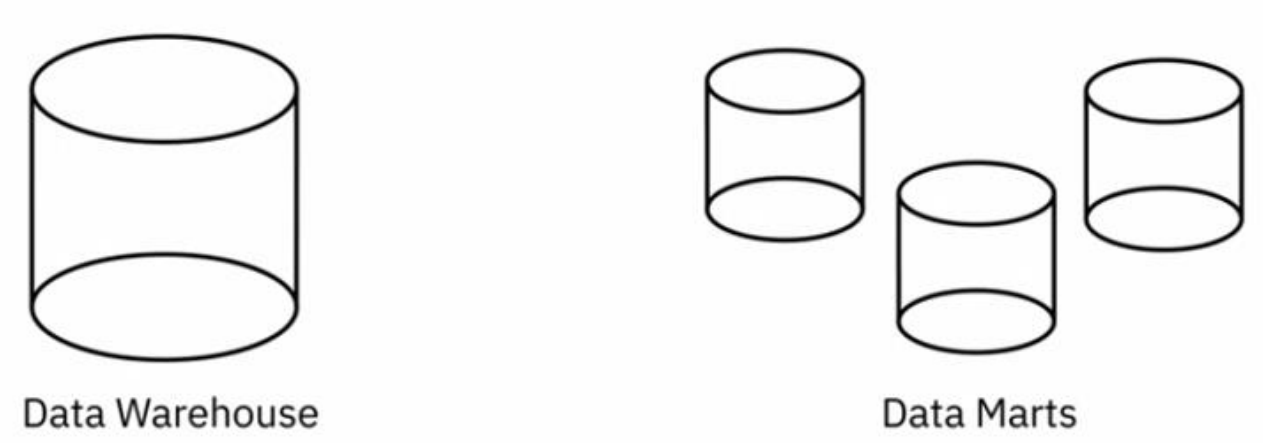
Data Mart


Data Mart
Data Mart
Data Mart
Data Mart
Data Mart
Data Warehouse

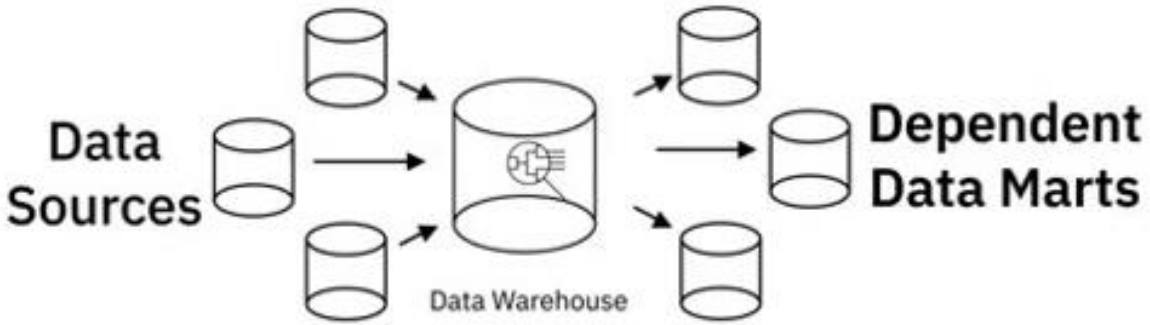
Data Marts

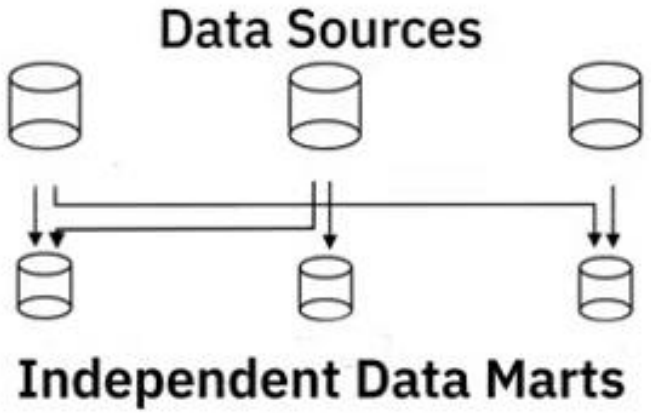
A data mart is a sub-section of the data warehouse, built specifically for a particular business function, purpose, or community of users.

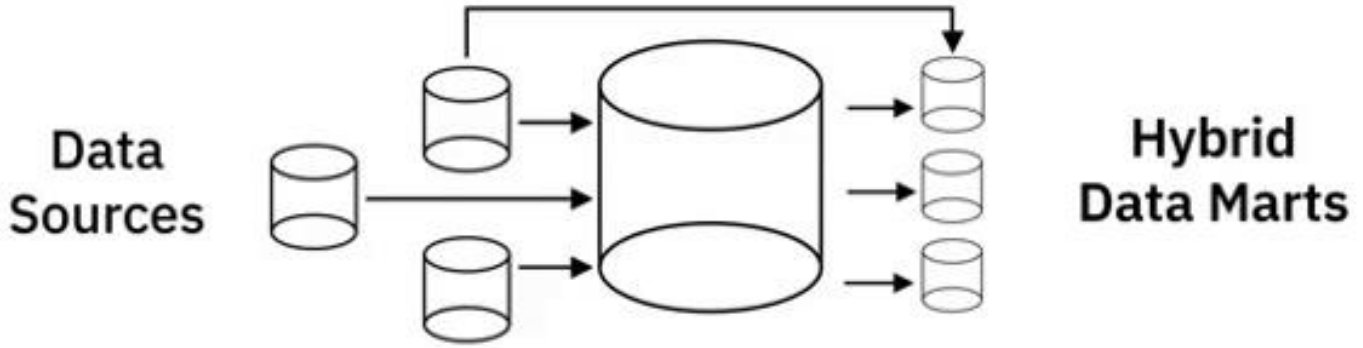
- Dependent
- Independent
- Hybrid



Data Mart


Data
Sources
Data Warehouse
Dependent
Data Marts


Data Sources
Independent Data Marts


Data
Sources
Hybrid
Data Marts

Data Marts

The purpose of a Data Mart is to:

- Provide data to users that is most relevant to them when they need it
- Accelerate business processes
- Provide a cost and time efficient way in which data-driven decisions can be taken
- Improve end-user response time
- Provide secure access and control

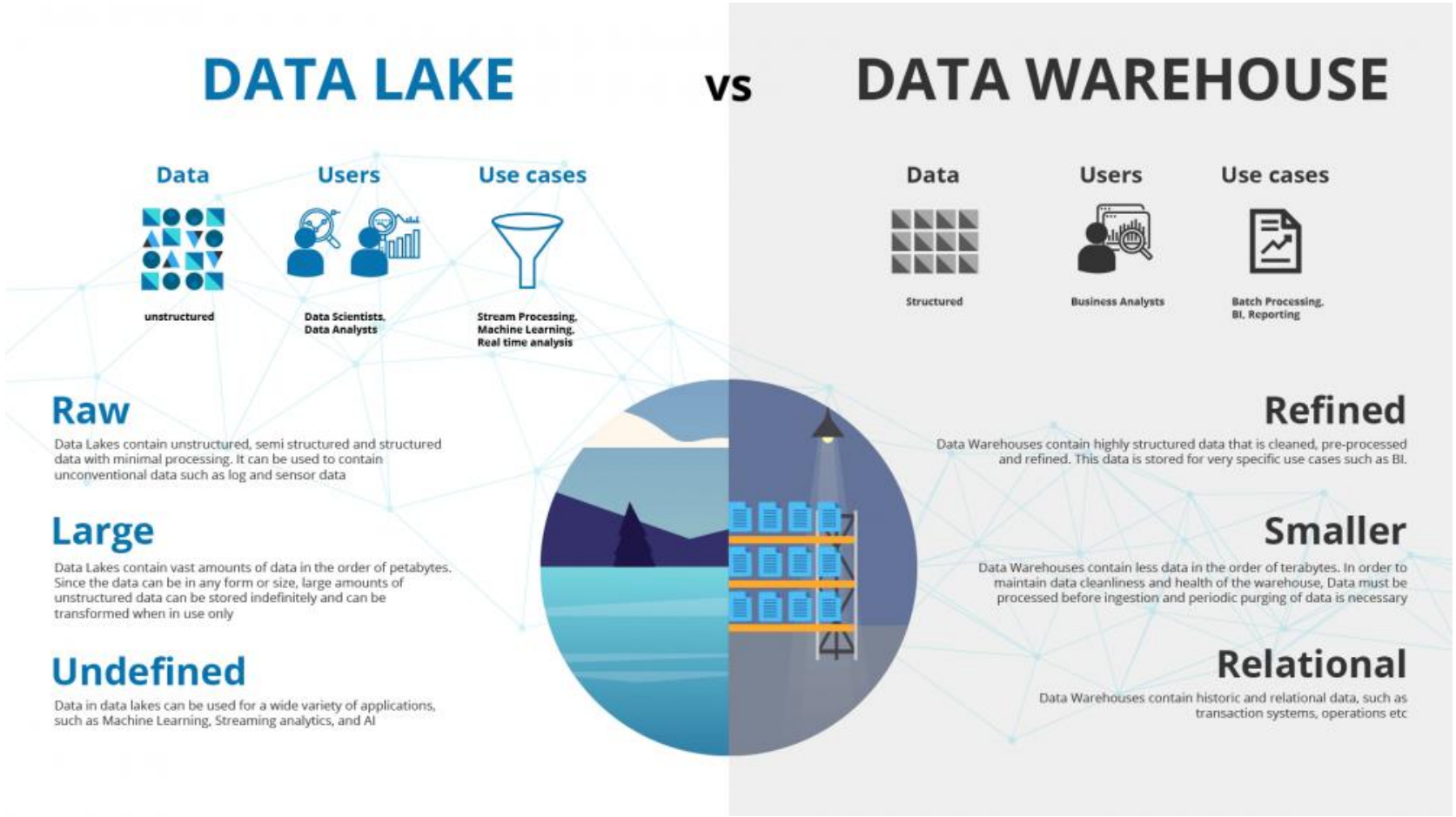
Data Lakes

- Store large amounts of structured, semi-structured, and unstructured data in their native format
- Data can be loaded without defining the structure and schema of data
- Exist as a repository of raw data straight from the source, to be transformed based on the use case
- Data is classified, protected, and governed
- A reference architecture that combines multiple technologies

Can be deployed using

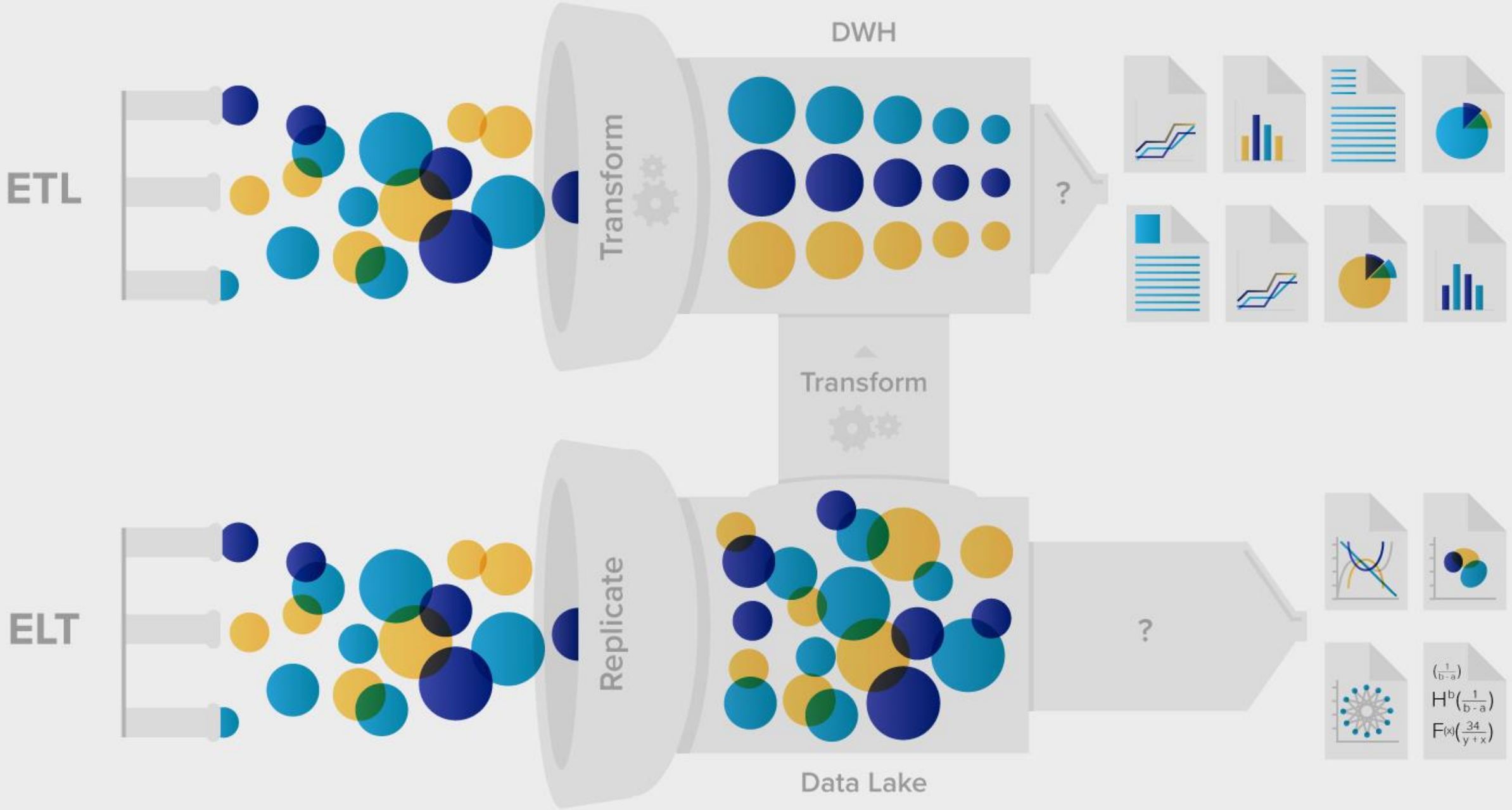
- Cloud Object Storage, such as Amazon S3
- Large-scale distributed systems such as Apache Hadoop
- Relational Database Management Systems, as well as NoSQL data repositories

Data Lakes

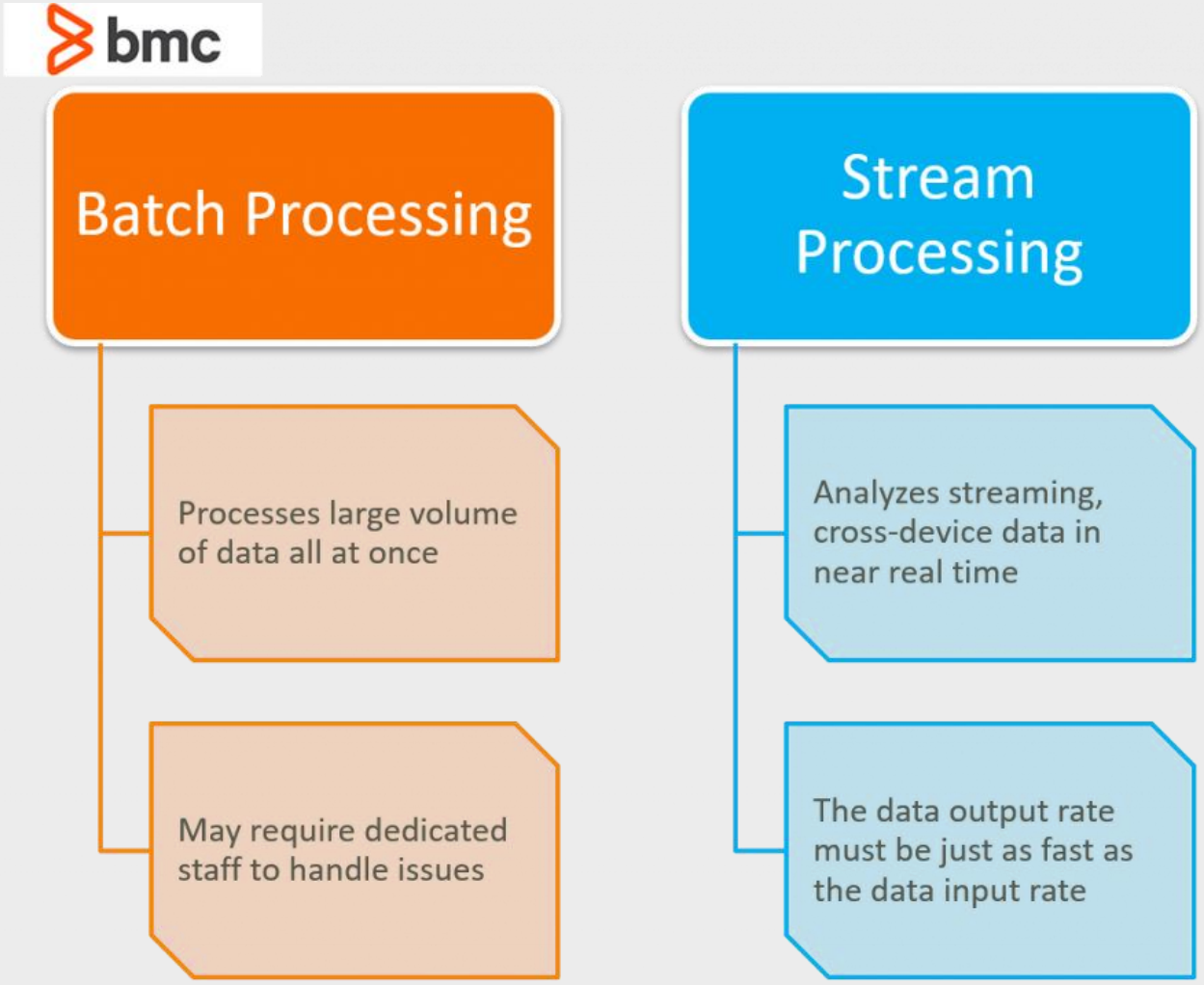

DATA LAKE
vs
DATA WAREHOUSE
Data
Users
Use cases
unstructured
Data Scientists,
Data Analysts
Stream Processing,
Machine Learning,
Real time analysis
Raw
Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data
Large
Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only
Undefined
Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI
Data
Users
Use cases
Structured
Business Analysts
Batch Processing,
BI, Reporting
Refined
Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.
Smaller
Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary
Relational
Data Warehouses contain historic and relational data, such as transaction systems, operations etc

Data Lakes


amazon.com
cloudera
Google
IBM
Informatica
Microsoft
ORACLE
EXADATA
SAS
snowflake
teradata.
zaloni

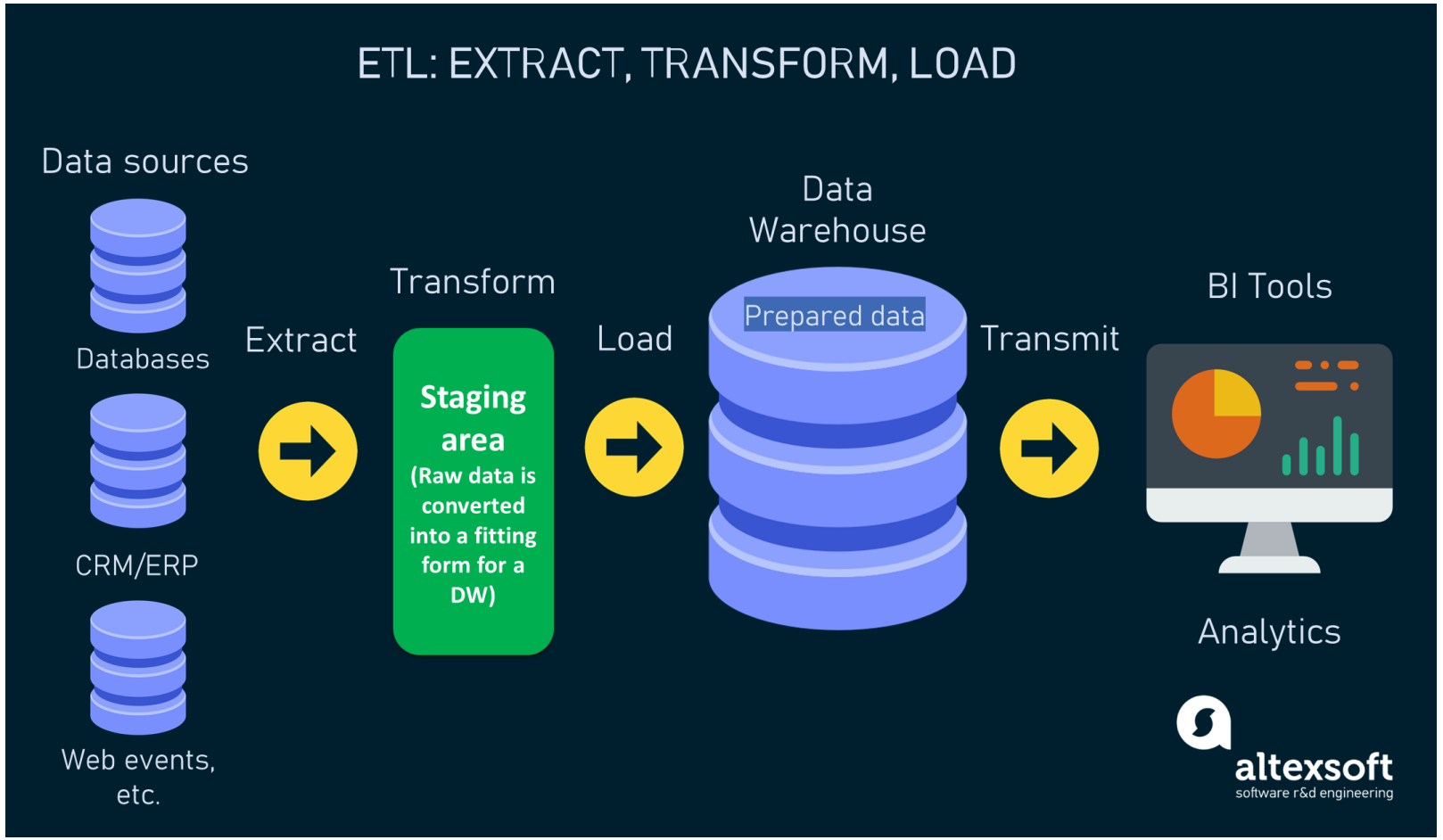

ETL
Transform
DWH
?
Transform
ELT
Replicate
?
Data Lake

Batch and Stream Processing

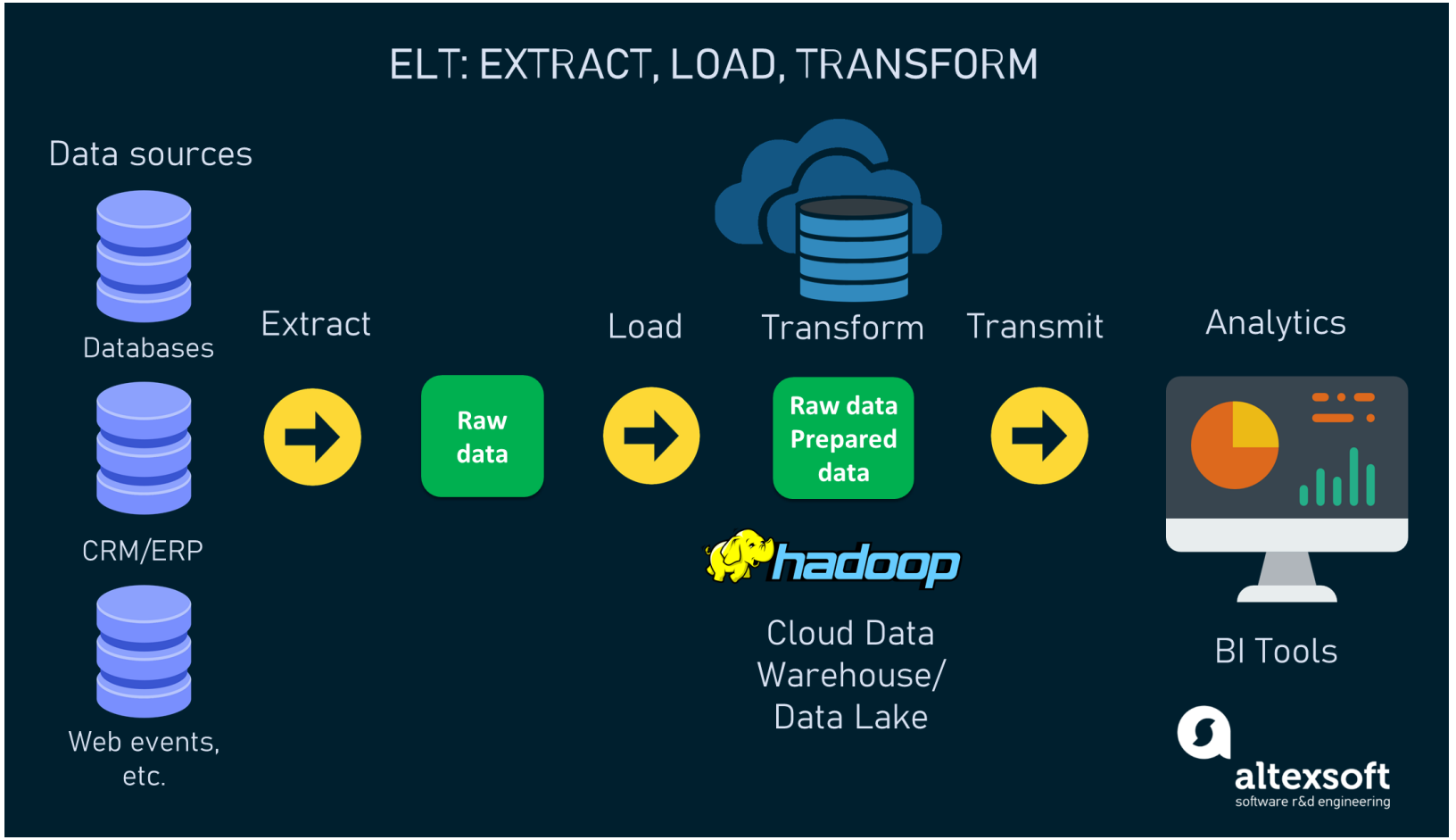


Benefits and drawbacks of common data processing types

ETL


ETL: EXTRACT, TRANSFORM, LOAD
Data sources
Databases
CRM/ERP
Web events, etc.
Extract
Transform
Staging area
(Raw data is converted into a fitting form for a DW)
Load
Data Warehouse
Prepared data
Transmit
BI Tools
Analytics
altexsoft
software r&d engineering

ELT


ELT: EXTRACT, LOAD, TRANSFORM
Data sources
Databases
CRM/ERP
Web events, etc.
Extract
Raw data
Load
Transform
Raw data Prepared data
Cloud Data Warehouse/ Data Lake
Transmit
Analytics
BI Tools
hadoop
altexsoft
software r&d engineering

Data Pipelines

- Encompasses the entire journey of moving data from one system to another, including the ETL process
- Can be used for both batch and streaming data
- Supports both long-running batch queries and smaller interactive queries
- Typically loads data into a data lake but can also load data into a variety of target destinations including other applications and visualization tools

Data Pipelines

Google Dataflow

Apache Airflow

- Can be used for both batch and streaming data
- Supports both long-running batch queries and smaller interactive queries
- Typically loads data into a data lake but can also load data into a variety of target destinations including other applications and visualization tools

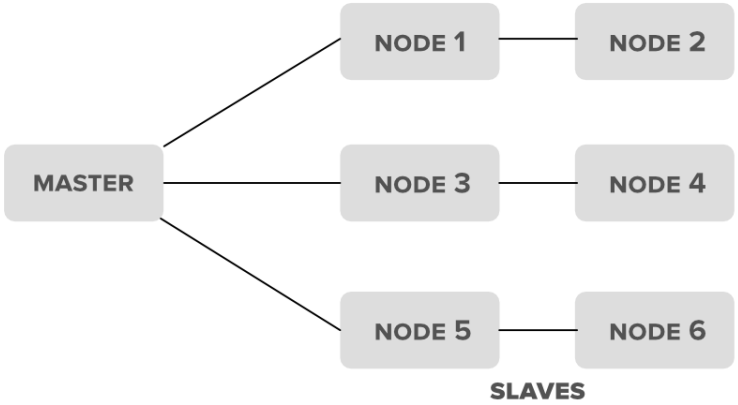
beam


Big Data Tools
Data Storage & Processing
Data Analysis
Data Warehouse
HIVE
No sequel Database
cassandra
Realtime Processing
APACHE Spark
Messaging System

Big Data Processing Tools

Big Data processing technologies provide ways to work with large sets of structured, semi-structured, and unstructured data so that value can be derived from big data.

Hadoop



- Distributed storage and processing of large datasets across clusters of computers.
- Hadoop provides a reliable, scalable, and cost-effective solution for storing data with no format requirements.

Benefits include:

- Better real-time data-driven decisions: Incorporates emerging data formats not traditionally used in data warehouses
- Improved data access and analysis: Provides real-time, self-service access to stakeholders
- Data offload and consolidation: Optimizes and streamlines costs by consolidating data, including cold data, across the organization

Hadoop

Hadoop Distributed File System, or HDFS, is a storage system for big data that runs on multiple commodity hardware connected through a network.

- Provides scalable and reliable big data storage by partitioning files over multiple nodes
- Splits large files across multiple computers, allowing parallel access to them
- Replicates file blocks on different nodes to prevent data loss

Hadoop

Benefits that come from using HDFS include:

- Fast recovery from hardware failures, because HDFS is built to detect faults and automatically recover.
- Access to streaming data, because HDFS supports high data throughput rates.
- Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers, in a single cluster.
- Portability, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems.

Hive

- Open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase.
- Queries have high latency → Not suitable for applications that need fast response times
- Read-based → Not suitable for transaction processing that involves a high percentage of write operations.
- Hive is better suited for →
  - Data warehousing tasks such as ETL, reporting, and data analysis
  - Easy access to data via SQL

Spark

- Spark is a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications.
  - Interactive Analytics
  - Streams Processing
  - Machine Learning
  - Data Integration
  - ETL
- Has in-memory processing which significantly increases speed of computations
- Provides interfaces for major programming languages such as Java, Scala, Python, R, and SQL
- Can run using its standalone clustering technology
- Can also run-on top of other infrastructures, such as Hadoop
- Can access data in a large variety of data sources, including HDFS and Hive
- Processes streaming data fast
- Performs complex analytics in real-time

VOLUME

Huge amount of data

VERACITY

Inconsistencies and uncertainty in data

VARIETY

Different formats of data from various sources

VELOCITY

High speed of accumulation of data

VALUE

Extract useful data


Thank
you