Week 2

Structures of Data

- Structured data, that is data which is well organized in formats that can be stored in databases.
- Semi-structured data, that is data which is partially organized and partially free-form.
- Unstructured data, that is data which can not be organized conventionally into rows and columns.

Types of Data and Understanding different types of File Formats

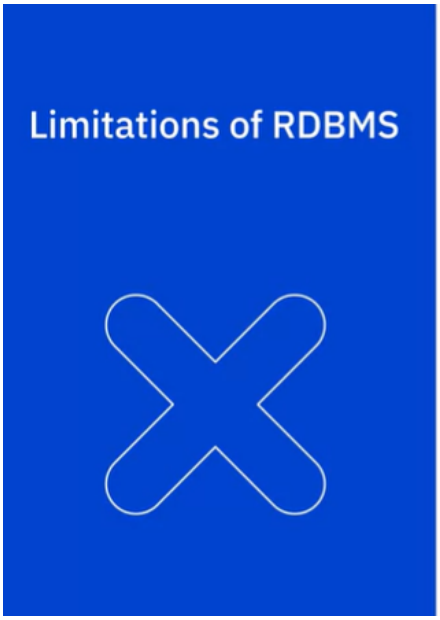
Sources of Data

- Relational DB (SQL Server, Oracle, MySQL
- Flat File(CSV, TSV), Spreadsheet (XLS, XLSX< Google)
- XML(Tags, can support complex data structures)
- API and Web Services (
- Web Scraping (Beautifulsoup, Pandas)
- Data Streams and feeds (GPS Car, IoT,) Like Stock → Kafka, Apache Spark
- RSS Feed (Really Simple Syndication) → Online forums and news where it has been refreshed

Languages for Data Professionals

- Querying languages, such as SQL, are used for accessing and manipulating data from databases.
- Programming languages such as Python, R, and Java, for developing applications and controlling application behavior.
- Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, for automating repetitive operational tasks.

Limitation of RDMS

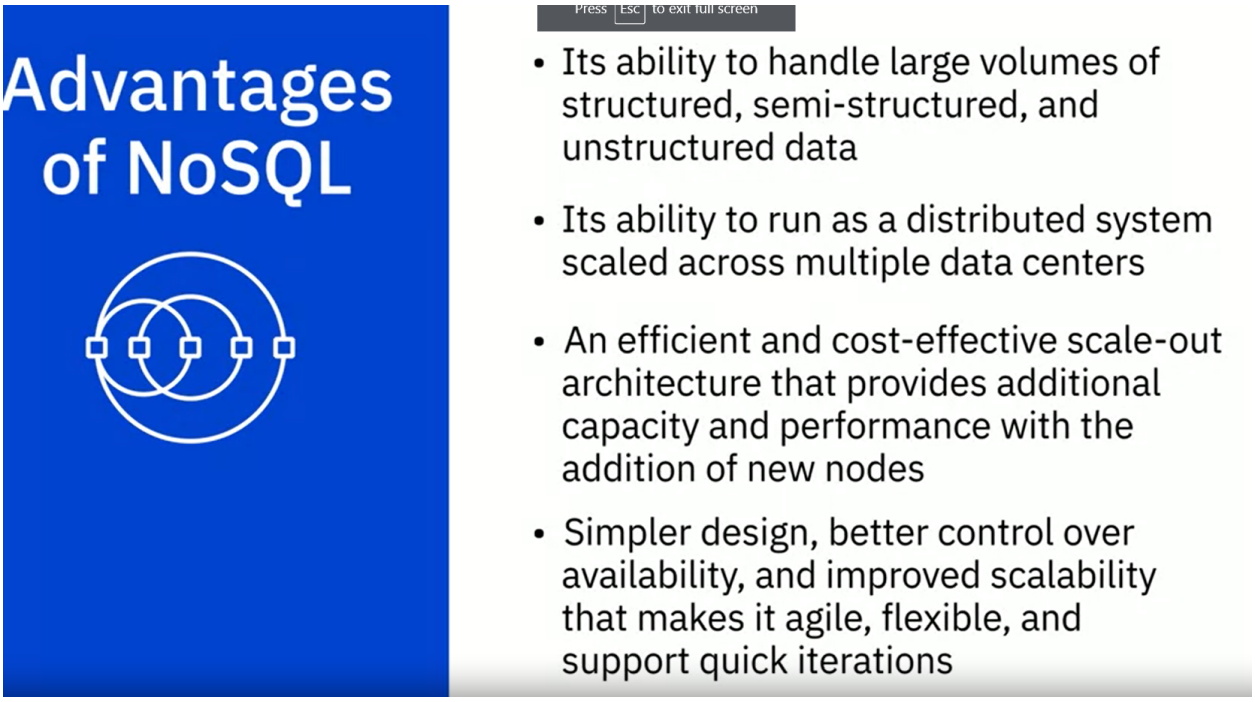


Limitations of RDBMS:

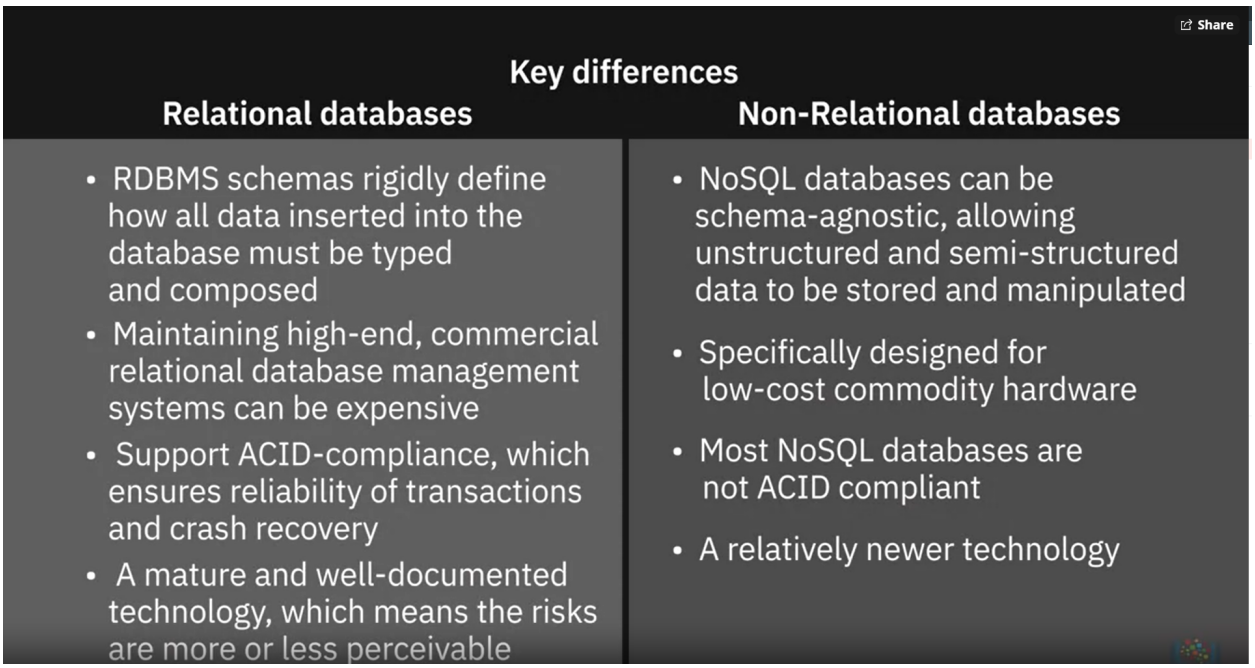
- Does not work well with semi-structured and unstructured data
- Migration between two RDBMS's is possible only when the source and destination tables have identical schemas and data types

NO SQL DB (Not Only SQL)

- Key-Value Storing (Redis, Memcached, DynamoDB)
- Document-Based (MongoDB, DocumentDB,CouchDB,Cloudant)
- Column-Based (Cassandra, Apache Hbase)
- Graph -Based (Neo4J, CosmosDB)



Difference Between SQL and NoSQL



What does ACID Mean?

ACID Explained: Atomic, Consistent, Isolated & Durable

I don't think it's an overstatement to say that data is pretty important. Data is especially important for modern organizations. In fact, The Economist went so far as to say that data has surpassed

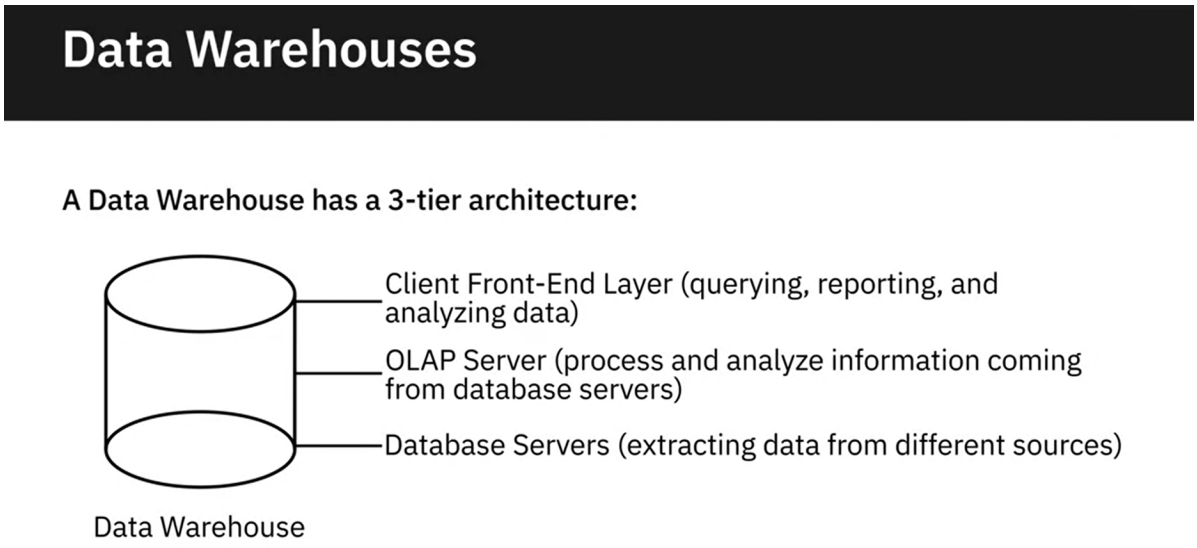
https://www.bmc.com/blogs/acid-atomic-consistent-isolated-durable/

Data Warehouses, Data Marts, and Data Lakes

Data Warehouse →

Single Source of Truth, Analysis Ready (Teradata, Oracle Exadata, IBM Db2, Netezza, Amazon Redshift, Google Big Query, Cloudera, Snowflake)

- Analysis Ready (CRM, HR)
- Has 3 tier architecture



- Moving to Cloud (Lower Cost, Pay as you go, faster disaster recovery, Limitless storage, and compute ready)

Data Marts →

- A subsection of Data Warehouse
- Get data to a specific department (Finance, Analytics)

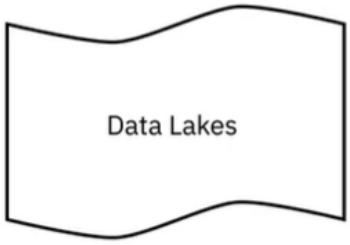
Data Marts

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 →

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

However, not a place for data dumped without Governance. [Amazon S3, Apache Hadoop]

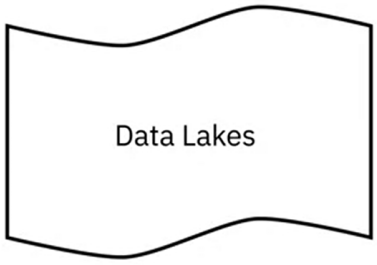
→ The story of Hadoop is quite a fascinating story

The Friendship That Made Google Huge

One day in March of 2000, six of Google 's best engineers gathered in a makeshift war room. The company was in the midst of an unprecedented emergency. In October, its core systems, which

https://www.newyorker.com/magazine/2018/12/10/the-friendship-that-made-google-huge

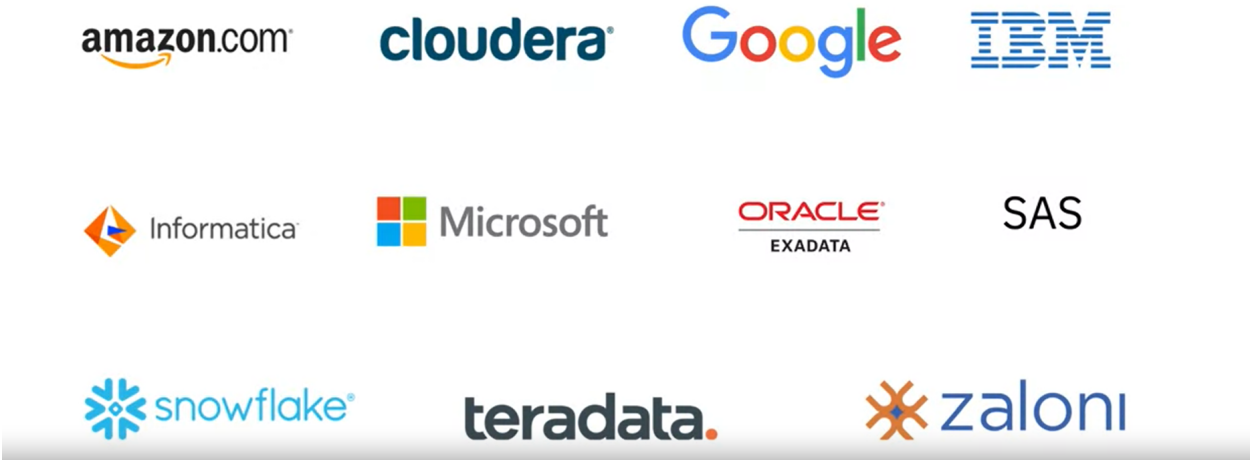
Data Lakes



Benefits:

- Ability to store all types of data (unstructured, semi-structured and structured data)
- Agility to scale based on storage capacity (growing from terabytes to petabytes)
- Saving time in defining structures, schemas, and transformations (data is imported in its original format)
- Ability to repurpose data in several different ways and wide-ranging use cases

Data Lakes



Considerations for choice

- [ ] Use case
- [ ] Schema of the data [Storing Structured, Semi or unstructured]
- [ ] Performance Requirement
- [ ] Data at rest of streaming Data (Data in motion)
- [ ] Volume of Data
- [ ] Storage Requirements [Updated frequently, Access frequently]
- [ ] Organizations standard [which are we allowed to use

- [ ] Capacities required to handle
- [ ] Type of access [Short interval or long queries]
- [ ] Purpose
  - [ ] Transactional

- [ ] Analytical
- [ ] Archival
- [ ] Data Warehousing

- [ ] Compatible with the current style, language
- [ ] Security Features
- [ ] Scalability

- [ ] Think about the skills you have or want to foster in the organization
- [ ] Cost of various solution
- [ ] Hosting Platform [Amazon, Microsoft, Google]

Data Integration Platforms

Data Pipeline for the whole process:

Short Summary

A Data Repository is a general term that refers to data that has been collected, organized, and isolated so that it can be used for reporting, analytics, and also for archival purposes.

The different types of Data Repositories include:

- Databases, which can be relational or non-relational, each following a set of organizational principles, the types of data they can store, and the tools that can be used to query, organize, and retrieve data.
- Data Warehouses, that consolidate incoming data into one comprehensive store house.
- Data Marts, that are essentially sub-sections of a data warehouse, built to isolate data for a particular business function or use case.

- Data Lakes, that serve as storage repositories for large amounts of structured, semi-structured, and unstructured data in their native format.
- Big Data Stores, that provide distributed computational and storage infrastructure to store, scale, and process very large data sets.

The ETL, or Extract Transform and Load, Process is an automated process that converts raw data into analysis-ready data by:

- Extracting data from source locations.
- Transforming raw data by cleaning, enriching, standardizing, and validating it.
- Loading the processed data into a destination system or data repository.

The ELT, or Extract Load and Transfer, Process is a variation of the ETL Process. In this process, extracted data is loaded into the target system before the transformations are applied. This process is ideal for Data Lakes and working with Big Data.

Data Pipeline, sometimes used interchangeably with ETL and ELT, encompasses the entire journey of moving data from its source to a destination data lake or application, using the ETL or ELT process.

Data Integration Platforms combine disparate sources of data, physically or logically, to provide a unified view of the data for analytics purposes.