Brief Overview of Data Management

Computer Systems, Data Structures and Data Management (4CM508)

Dr Sam O'Neill

Some Housekeeping…

Timetable

Lecture

- **MS125** @ Monday 9am

Practical

Either:

- **MS214/MS215** @ Friday 11am
- **MS214/MS215** @ Friday 1pm

Your timetable will show a 2-hour block for both 4CM506 and 4CM508 as per last semester.

Basically turn up for the 2-hour block, my tutorial first, Chris' second!

Assessment

Computer Systems, Data Structures and Data Management (4CM508) is a portfolio assessment made up of three components:

1. Assessed labs (25%)
2. In-class test (25%) - Provisionally in week 12 practical
3. Coursework (50%)

What should I have done to date?

- To date, you should have attempted the 9 MCQ Quizzes (you have unlimited attempts).
- There are going to be 12 assessed MCQ quizzes (originally 15)

I will reopen the first 9 tests for the first two weeks of the semester (Deadline 16th February 23:59). **After that they will be closed for good!**

Coursework

- Coursework will be released in a few weeks (provisionally week 4).
- Relational Database task.
- I will give you a complete overview of what is required including submission requirements and the deadline.
- You will be given more than enough time for this.

Semester 2 Outline (Subject to Change)

1. Brief Overview of Data Management
2. Introduction to Databases
3. Relational Databases
4. NOSQL Databases
5. Sorting 2 (Divide and Conquer)
6. Non-comparison Sort
7. Heaps and Priority Queues
8. Binary Search Trees and AVL Trees
9. Graphs and Basic Graph Algorithms
10. Pathfinding Algorithms

Brief Overview of Data Management

These slides are intended as a brief overview.

You should use the linked resources on the slides to explore these topics in more detail.

Agenda

1. Overview of Data Management
2. Data Lifecycle
3. Data Governance
4. Data Quality
5. Types of Data
6. Data Storage
7. Data Security

Question: What is data?

Overview of Data Management

Definition

- **Data management** encompasses the **entire lifecycle of data**, including its acquisition, storage, processing, and utilization.

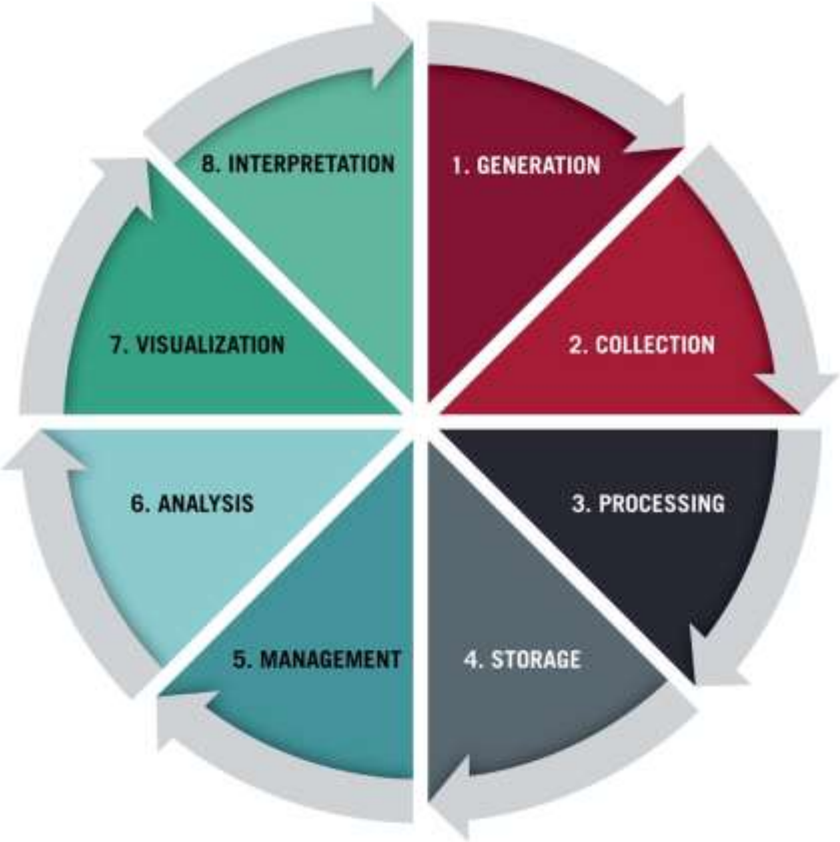
Importance

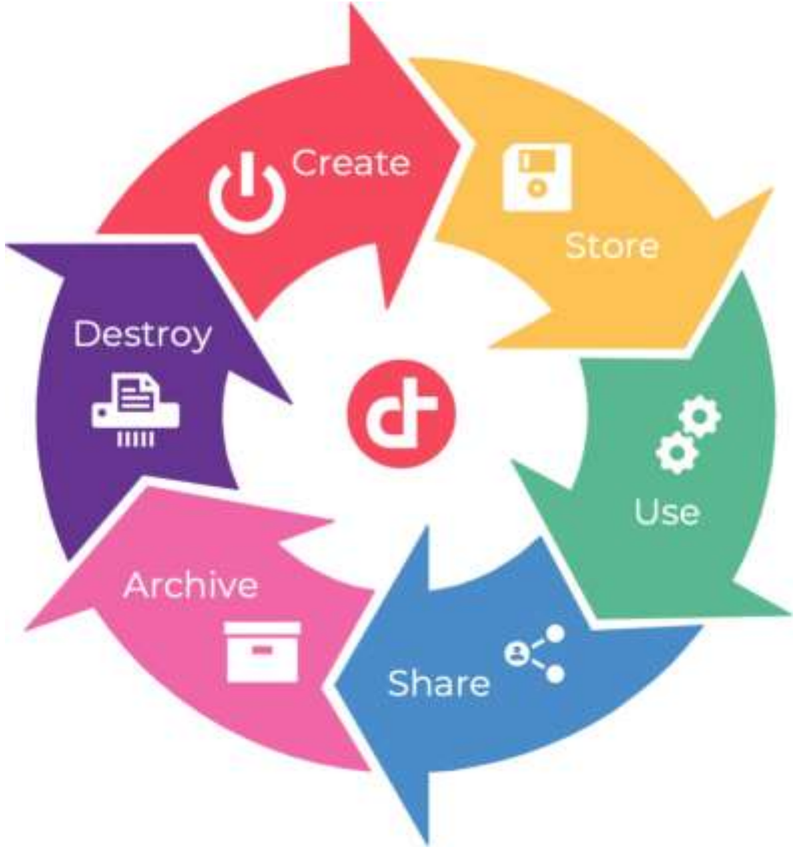
- **Data** is a **valuable asset for organizations** enabling **informed decision-making** and driving **business growth**.
- Ensuring data **quality, security, and compliance** is essential for **maintaining trust** and **meeting regulatory requirements**.

Data Lifecycle

The **data lifecycle** describes the **stages** that data goes through during its **lifespan**.

Here are two common ways to think about it:





Data Governance

- **What?** Comprehensive approach to manage data throughout complete lifecycle.
- **Why?** Aligns data requirements with business strategy.

Key Aspects

- **Availability**: Easy access to data
- **Quality**: High-quality, secure, and accessible data
- **Security**: Prevents unauthorized access and misuse
- **Compliance**: Regulatory requirements

Benefits

- Increased efficiency/reduced costs
- Improved productivity/ decision-making
- Enhanced collaboration
- Enhanced security and privacy
- Ensure compliance

4CM506 will look into specifics of data governance and implementation strategies.

Data Quality

- **What?** Dataset is **accurate**, **complete**, **valid**, **unique** and **fit for purpose**.
- **Why?** Ensures companies make correct data-driven decisions to meet their goals.

- Key Aspects
  - **Accurate**: Correct and reliable.
  - **Complete**: All data is present.
  - **Valid**: Conforms to the syntax of its definition.
  - **Unique**: Without unnecessary duplication.
  - **Fit for Purpose**: The data can be used for its intended purpose.

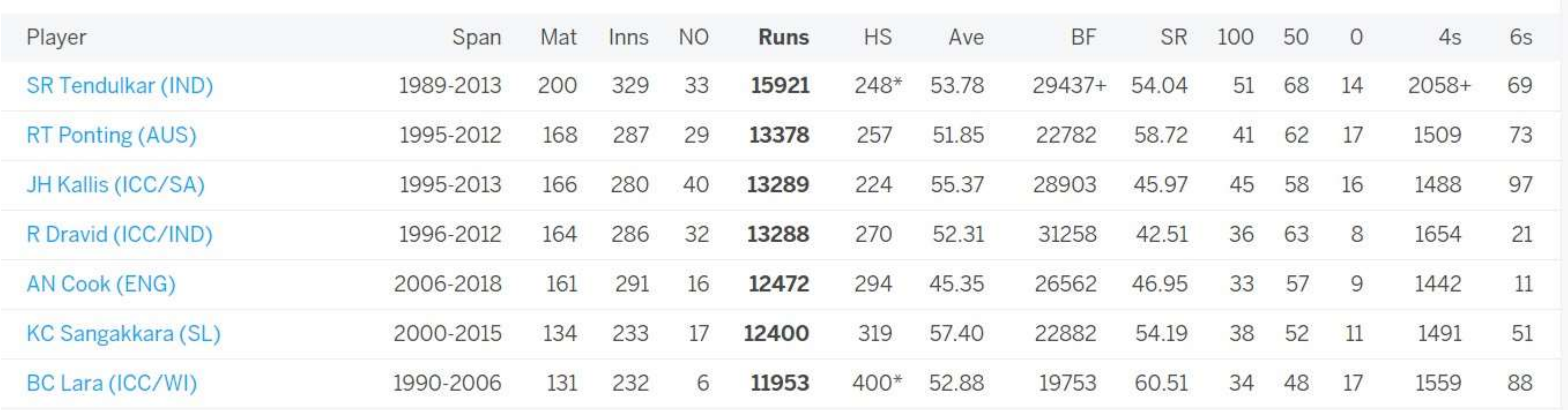
- Benefits
  - **Improved decision-making**: Better business decisions.
  - **Increased operational efficiency**: Reduces errors and costs.
  - **Enhanced customer satisfaction**: Accurate/timely/relevant data.
  - **Regulatory compliance**: Ensure compliance with regulations.

Types of Data

1. Structured Data
2. Unstructured Data
3. Semi-Structured Data

Structured Data

- Organised in a **predefined format** with clear data types and relationships:
- Typically in **tabular form** (tables, rows and columns)



Player	Span	Mat	Inns	NO	**Runs**	HS	Ave	BF	SR	100	50	0	4s	6s
SR Tendulkar (IND)	1989-2013	200	329	33	**15921**	248*	53.78	29437+	54.04	51	68	14	2058+	69
RT Ponting (AUS)	1995-2012	168	287	29	**13378**	257	51.85	22782	58.72	41	62	17	1509	73
JH Kallis (ICC/SA)	1995-2013	166	280	40	**13289**	224	55.37	28903	45.97	45	58	16	1488	97
R Dravid (ICC/IND)	1996-2012	164	286	32	**13288**	270	52.31	31258	42.51	36	63	8	1654	21
AN Cook (ENG)	2006-2018	161	291	16	**12472**	294	45.35	26562	46.95	33	57	9	1442	11
KC Sangakkara (SL)	2000-2015	134	233	17	**12400**	319	57.40	22882	54.19	38	52	11	1491	51
BC Lara (ICC/WI)	1990-2006	131	232	6	**11953**	400*	52.88	19753	60.51	34	48	17	1559	88

What is Structured Data? - AWS

Structured Data

Examples

- **Databases**: MySQL, PostgreSQL, Microsoft SQL Server, SQLite, Oracle Database
- **Spreadsheets**: Excel, Google Sheets
- **Structured Text Files**: CSV, XML, JSON (if formatted in a structured way)

Advantages

Easy to:

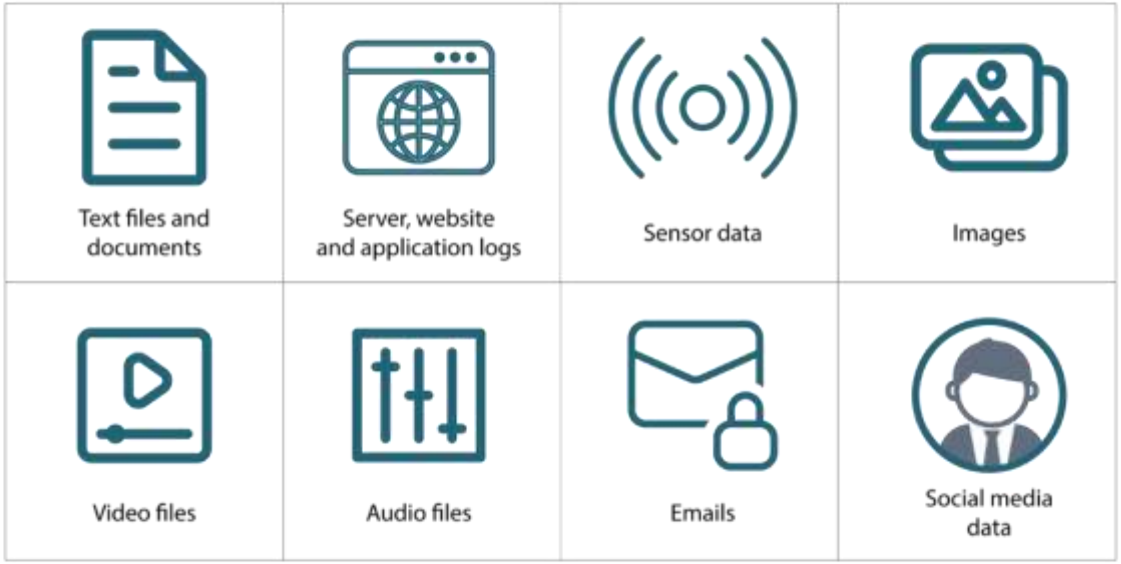
- Organise
- Clean
- Search
- Analyse

Disadvantages

- Data must fit the prescribed model, i.e. Not flexible

Unstructured Data

Unstructured data refers to data that lacks a predefined structure, such as:



Making it more challenging to analyze and process compared to structured data.

[Structured Data 'vs' Unstructured Data - AWS]

Unstructured Data

Examples

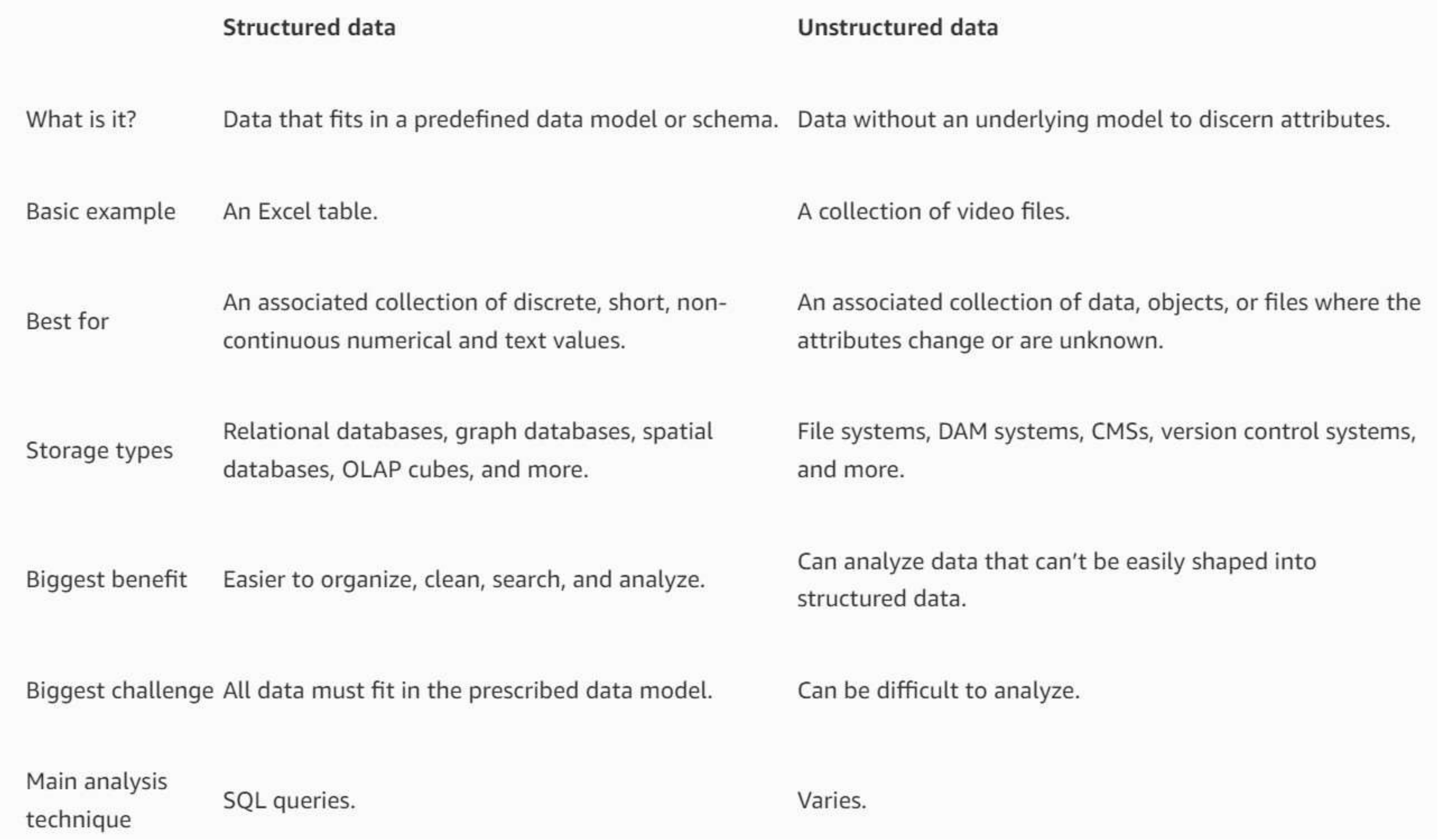
- **Text**: Emails, social media posts, articles
- **Media**: Images, videos, audio files
- **Documents**: PDFs, Word documents
- **Sensor Data**: IoT data streams

Advantages

- Flexibility in capturing diverse data types
- No predefined schema, allowing for easy data capture and storage

Disadvantages

- Difficult to analyze and process without prior structuring
- Lack of consistency and organization can lead to challenges



	Structured data	Unstructured data
What is it?	Data that fits in a predefined data model or schema.	Data without an underlying model to discern attributes.
Basic example	An Excel table.	A collection of video files.
Best for	An associated collection of discrete, short, non-continuous numerical and text values.	An associated collection of data, objects, or files where the attributes change or are unknown.
Storage types	Relational databases, graph databases, spatial databases, OLAP cubes, and more.	File systems, DAM systems, CMSs, version control systems, and more.
Biggest benefit	Easier to organize, clean, search, and analyze.	Can analyze data that can't be easily shaped into structured data.
Biggest challenge	All data must fit in the prescribed data model.	Can be difficult to analyze.
Main analysis technique	SQL queries.	Varies.

Semi-Structured Data

- Semi-structured data sits between structured and unstructured data.
- Understanding semi-structured data is important as it offers flexibility in data representation
- Commonly encountered in various modern data sources.

Semi-structured Data

Examples

- **XML Files**: HTML, XHTML
- **JSON Files**: Configuration files, log files
- **NoSQL Databases**: MongoDB, CouchDB
- **Graph Databases**: Neo4j

Advantages

- Combines flexibility of unstructured data with a level of organization
- Can handle diverse data types without a strict schema

Disadvantages

- May require effort in data normalization/structuring for analysis
- Not as easily queried as structured data

Data Storage

- Databases
- Data Warehouses
- Data Lakes

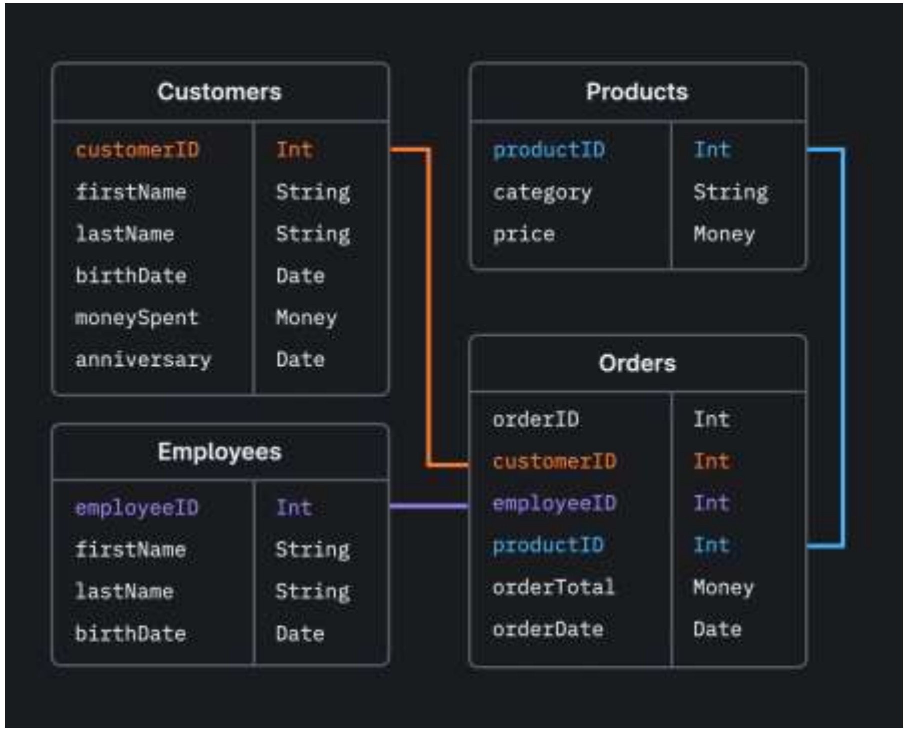
Relational Database Management System (RDBMS)

Organises data into **tables with rows and columns; relationships between the tables**.

- Based on relational model of data.
- Relational database first used by E. F. Codd at IBM in 1970.

Examples

  - MySQL
  - PostgreSQL
  - Microsoft SQL Server
  - Oracle Database



Customers	
customerID	Int
firstName	String
lastName	String
birthDate	Date
moneySpent	Money
anniversary	Date

Products	
productID	Int
category	String
price	Money

Employees	
employeeID	Int
firstName	String
lastName	String
birthDate	Date

Orders	
orderID	Int
customerID	Int
employeeID	Int
productID	Int
orderTotal	Money
orderDate	Date

Relational Database Management System (RDBMS)

Advantages:

- **Data Integrity**: Well defined schema.
- **Robust Transactions**: They follow ACID (Atomicity, Consistency, Isolation, Durability).
- **Standardised Query Language**: SQL used for querying (well establised standard).
- **Maturity**: Around for a long time and have a large ecosystem of tools and best practices.

Disadvantages:

- **Scalability**: Scaled vertically by adding more powerful hardware, not designed to scale horizontally across multiple servers.
- **Flexibility**: Predefined schema, which can limit flexibility when dealing with unstructured or semi-structured data.
- **Complexity**: The strict schema and relationships can make them complex to set up and manage.

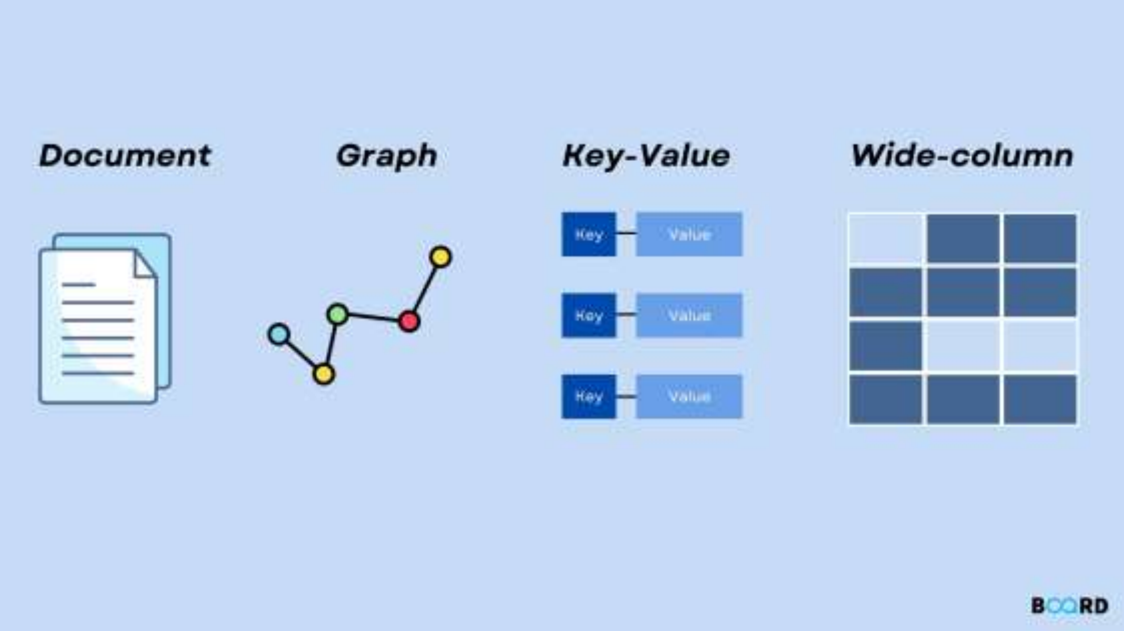
NoSQL Databases

- Around since the late 1960s
- NoSQL coined in the early 2000s
- based on needs of Web 2.0
- NoSQL - "Not Only SQL"

Background Reading

Examples

- MongoDB (Document Store)
- Neo4j (Graph Database)
- BigTable/Apache Cassandra (Wide-column)



Databases - NoSQL

Advantages:

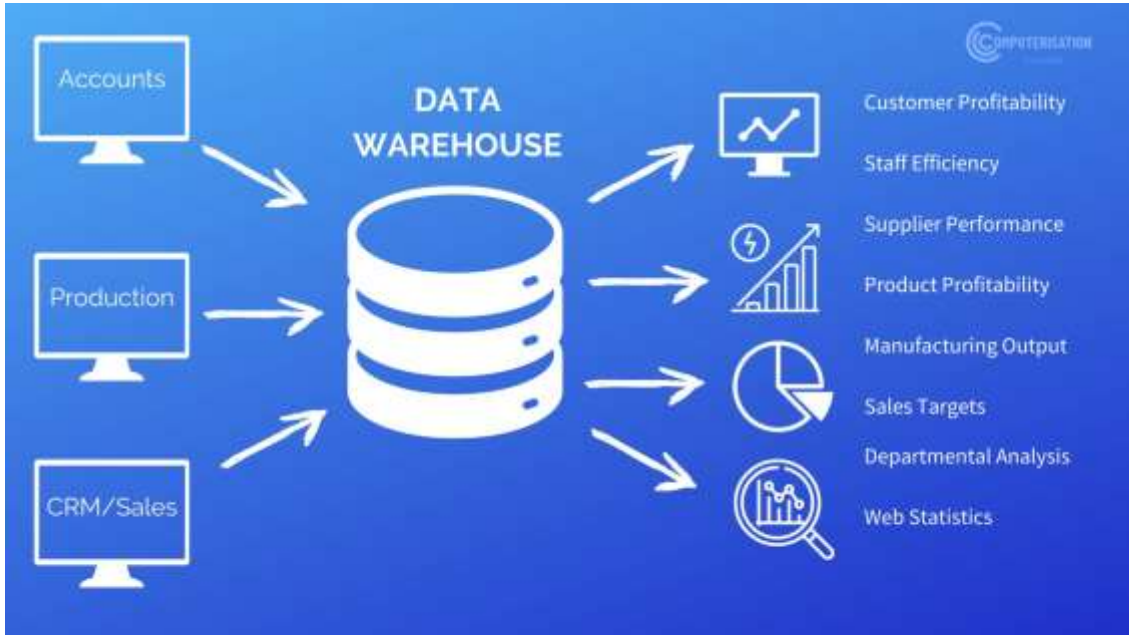
- **Scalability**: Scale horizontally, ideal across multiple servers.
- **Data Modeling Flexibility**: Schema-less, allowing unstructured and semi-structured data
- **High Availability**: Designed for distributed environments.
- **Cost-Effective at Scale**: Especially in cloud environments.

Disadvantages:

- **Consistency**: May not fully support ACID properties, lead to less consistency.
- **Complexity**: The variety of NoSQL databases and their different capabilities can make it more complex to choose the right one.
- **Maturity**: Newer and may not have as many established tools and best practices.

Data Warehouses

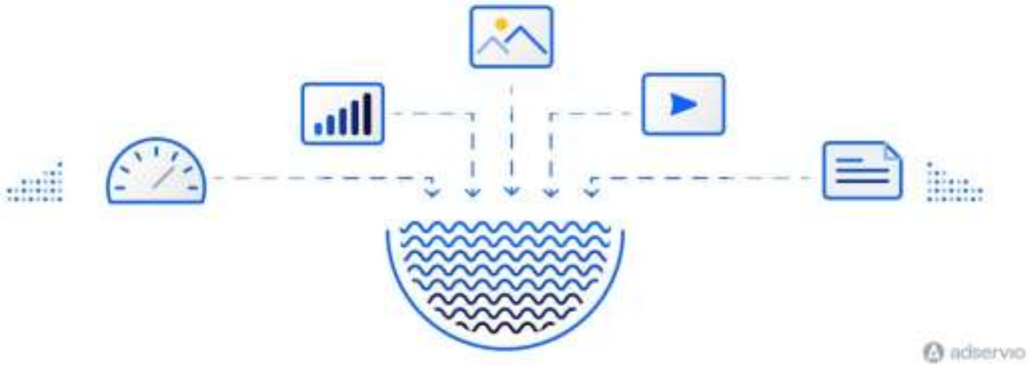
- Data warehouses are optimised for:
    - Analytical processing.
    - Consolidating data from multiple sources .
    - Complex queries and reporting.



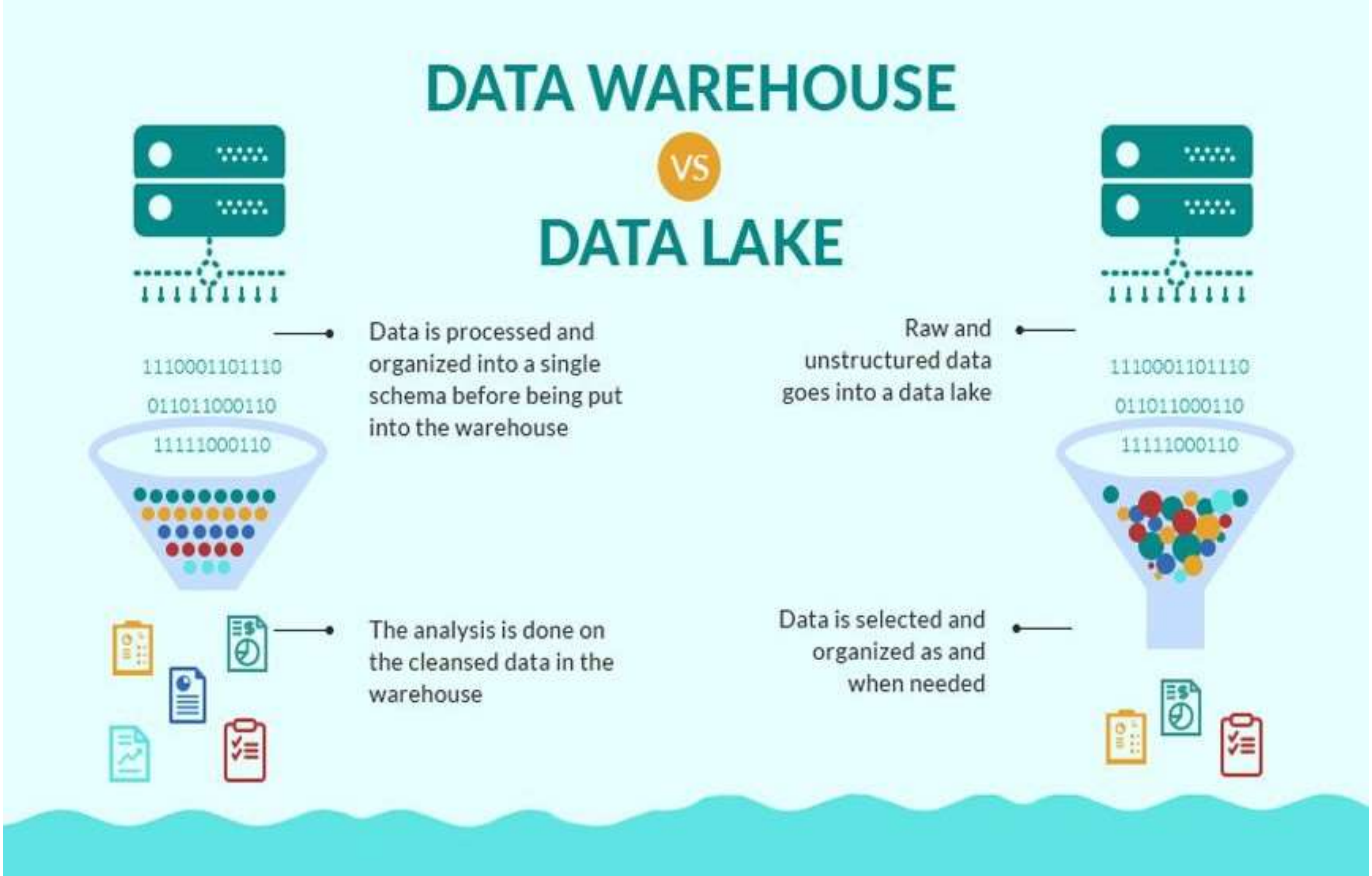
[What is a data warehouse? - Google Cloud]()

Data Lakes

- Data lakes serve as **repositories for raw data** in its native format.
- Can store **structured**, **unstructured** and **semi-structured** data.
- Store your data as-is, without having to first structure the data.



[What is a data lake? AWS](https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/)



DATA WAREHOUSE

VS

DATA LAKE

Data is processed and organized into a single schema before being put into the warehouse

The analysis is done on the cleansed data in the warehouse

Raw and unstructured data goes into a data lake

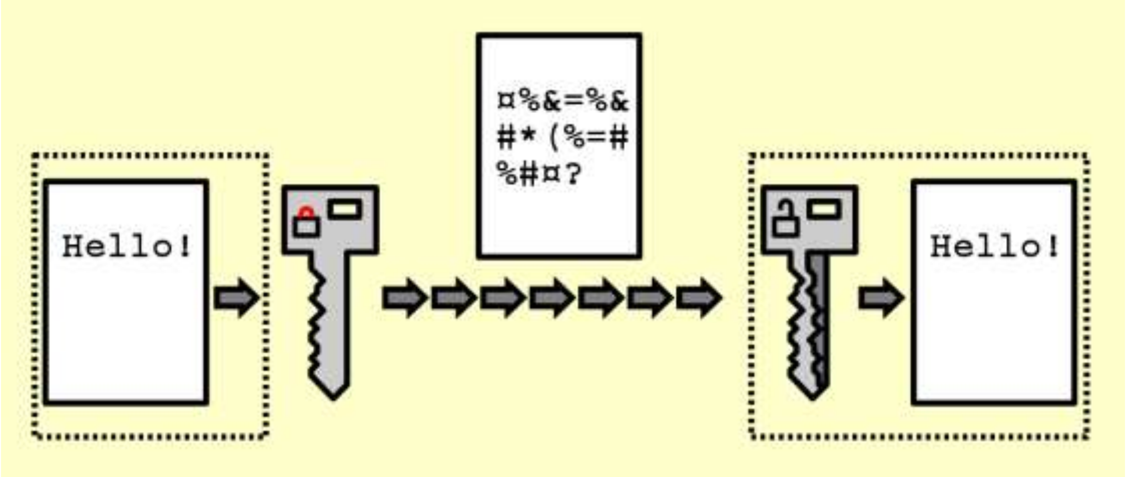
Data is selected and organized as and when needed

Data Security

- Encryption
- Access Control
- Backup and Recovery

Encryption

- Protect data by converting it into a coded format
- Ensures confidentiality and integrity during transmission and storage.



Access Control

- Access control restrict data access based on user roles and permissions
- Safeguard sensitive information from unauthorised users.

Authentication mechanisms (e.g.):

- Passwords
- Biometrics
- Multi-factor authentication

Authorisation mechanisms (e.g.):

- Role-based access control (RBAC)
- Attribute-based access control (ABAC).

Backup and Recovery

Regular data backup procedures and robust recovery plans are essential for:

- Mitigating Data Loss
- Ensuring Business Continuity
- Compliance
- Protecting Against Cyberattacks

Summary

- Covered various aspects of data management:
- Effective data management is crucial for:
  - Leveraging data as a valuable asset
  - Making informed decisions
  - Driving business growth
  - Ensuring compliance with regulatory requirements
- Robust data management practices maintain data integrity, security, and availability throughout its lifecycle.

References

1. Harvard Business School: Steps in the Data Life Cycle
2. Devoteam Belgium: Cloud Data Lifecycle - A Deep Dive
3. Structured Data 'vs' Unstructured Data - AWS
4. What is a Data Warehouse? - Google Cloud
5. What is a Data Lake? - AWS
6. Schema Design 101: Relational Databases - PlanetScale
7. NoSQL Database - Board Infinity
8. Data Lake and Its Benefits in Data Management - Adservio