CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

INSTITUTE: UIE (AIT-CSE)

Introduction to Data Science

(Types of Big Data)

By : Dr. Jitender Kaushal
Associate Professor

DISCOVER . LEARN . EMPOWER

Introduction to Data Science: Course Objectives

COURSE OBJECTIVES

The Course aims to:

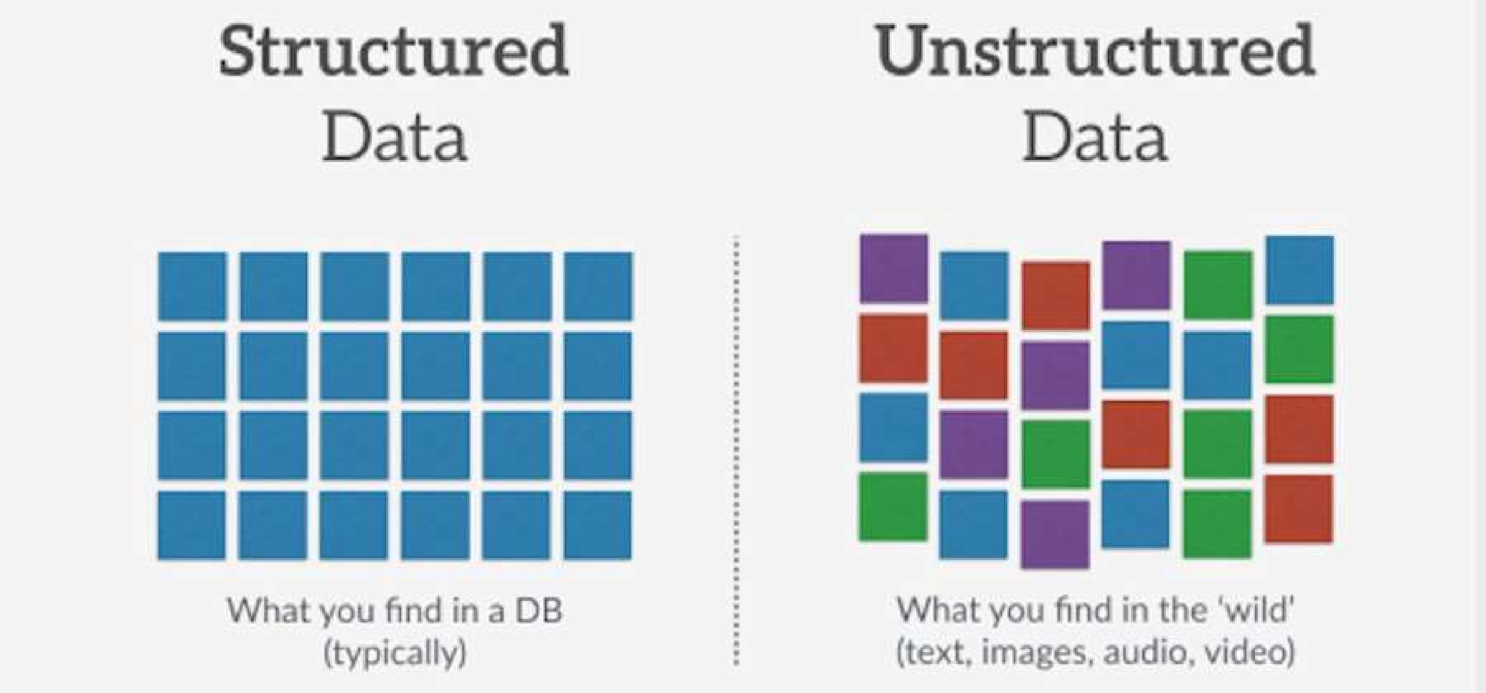
- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation and Loading
- To prepare a sample project in Hadoop Environment

COURSE OUTCOMES

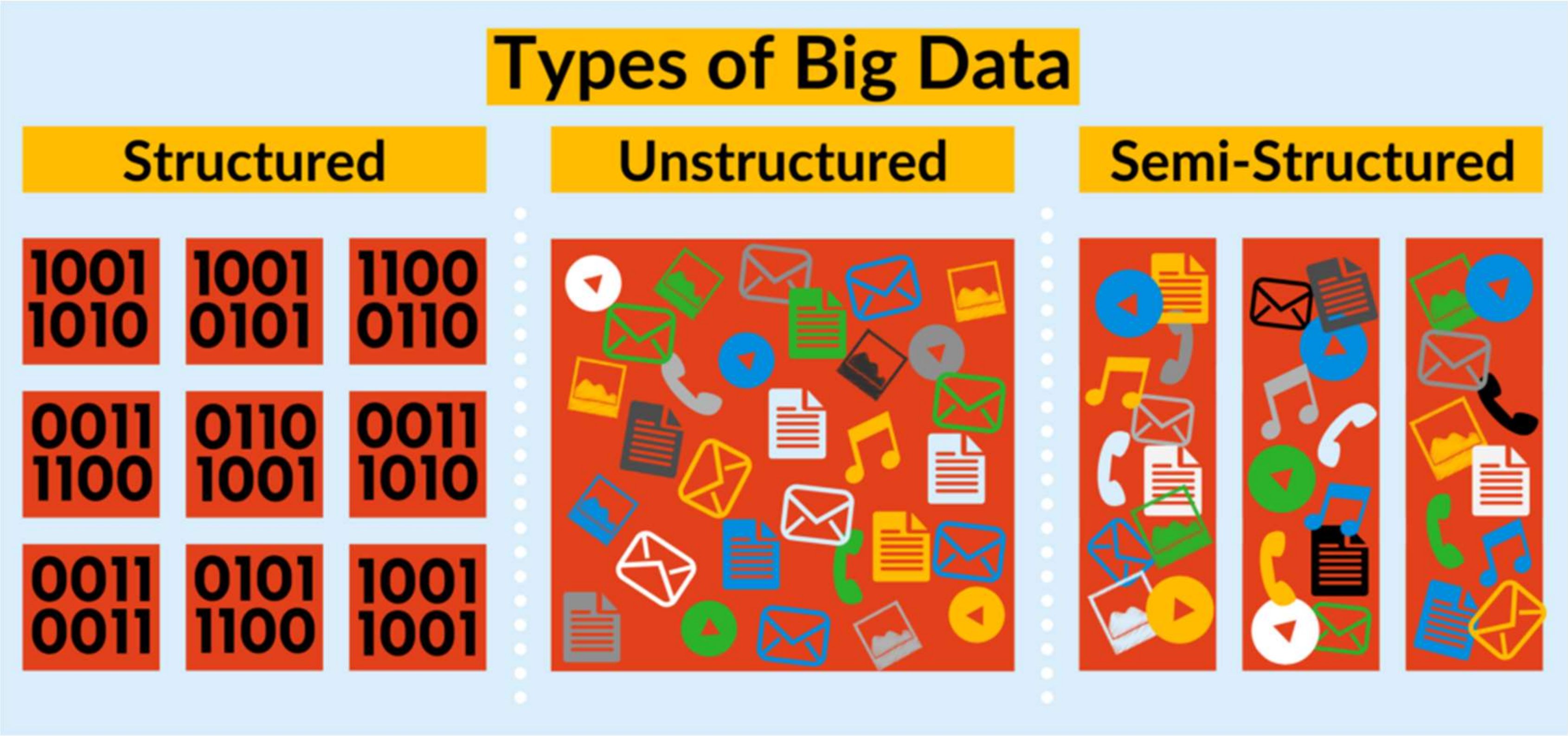
On completion of this course, the students shall be able to:-

CO1	Identify and describe the importance of Big data analysis over Conventional Database management System.

Types of Big Data


Structured Data
Unstructured Data
What you find in a DB (typically)
What you find in the 'wild' (text, images, audio, video)

Types of Big Data



Examples

- **Structured data** include **names, dates, addresses, credit card numbers, stock information, geolocation**, and more. Structured data is highly organized and easily understood by machine language.
- The best example of structured data is the **relational database**: the data has been formatted into precisely defined fields, such as **credit card numbers or address**, in order to be easily queried with SQL.
- **Unstructured data** includes various content such as **documents, videos, audio files, posts on social media, and emails**.

Structured Data

- Data stored in Database Table and spread-sheets
- Can be easily entered, stored, queried and analysed

Unstructured Data

- Information that doesn't reside in a traditional row-column database.
- For example texts, email, Facebook post, audio, video, blog etc.

Types of Big Data

1. **Structured data**
2. **Unstructured data**
3. **Semi-structured data**

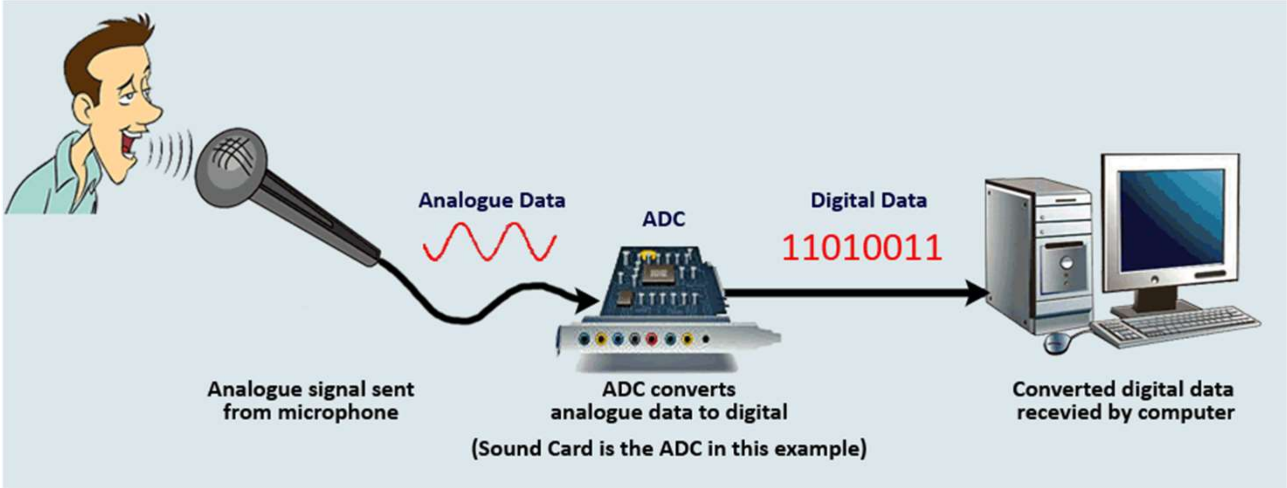
Structured data

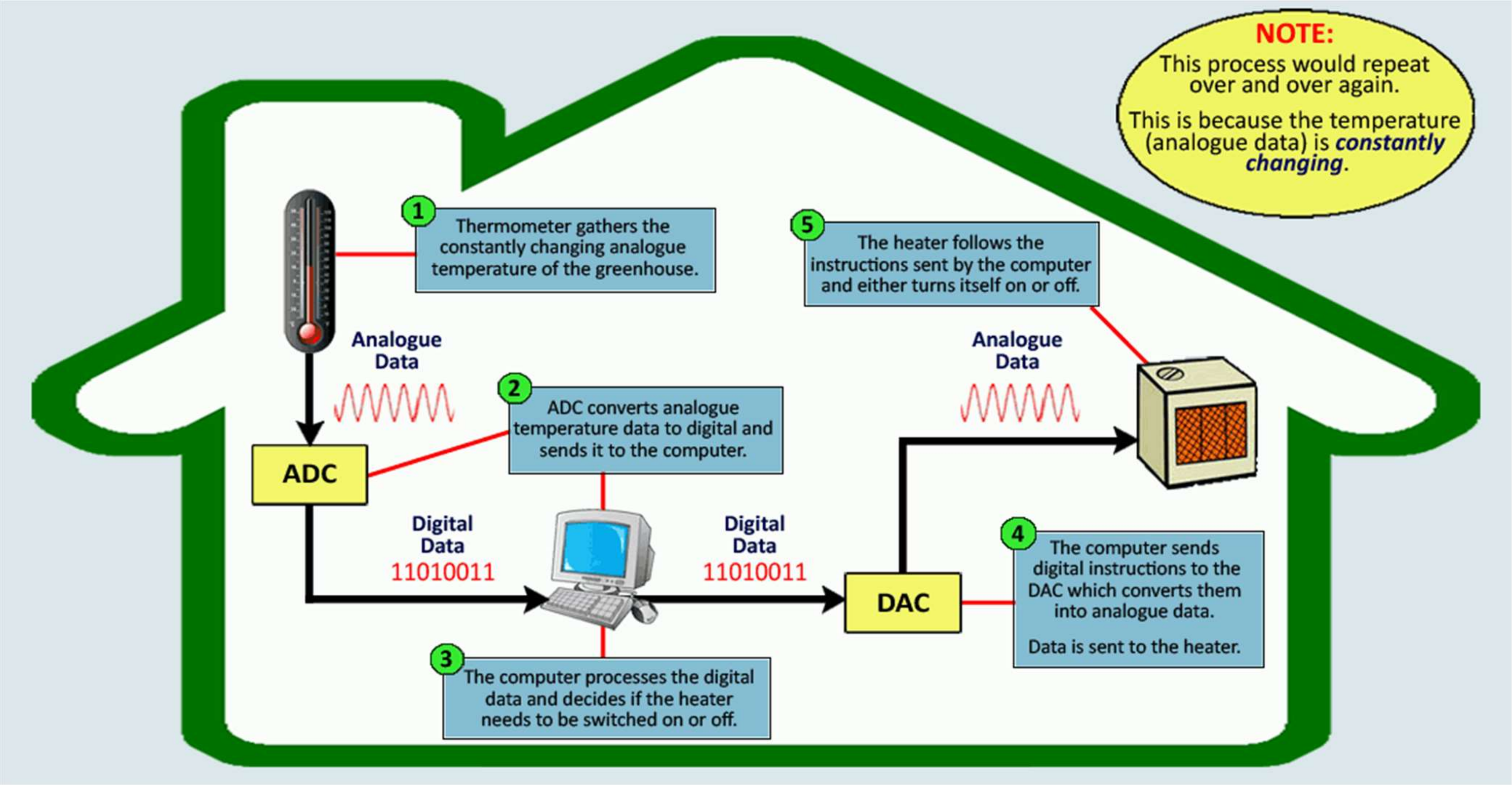
Structured data has certain **predefined organizational properties** and is presented in a **structured or tabular form**, making it easier to **analyze and sort**. In addition, thanks to its predefined nature, each field is discrete and can be **accessed separately or jointly** with data from other fields. This makes structured data **extremely valuable**, making it possible to collect data from various locations in the database quickly.

Database

- **MySQL, SQL Server, MongoDB, Oracle Database, PostgreSQL, Informix, Sybase**, etc. are all examples of different databases. These modern databases are managed by DBMS. Structured Query Language, or SQL as it is more widely known, is used to operate on the data in a database.
- **Real-life examples: grocery store, bank, restaurant, online shopping sites, hospitals, favorite clothing stores and mobile service providers**, for instance, all use databases to keep track of customer, inventory, employee and accounting information.

- Data that resides in fixed fields within a record or file. Example Analog Data, GPS Tracking Information, Audio/ Video Streams.
- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the **format is well known in advance**) and also deriving value out of it.
- However, nowadays, we are foreseeing issues when the size of such data grows to a huge extent, typical sizes are in the range of multiple zettabytes (10^{21}) or (2^{70}).


Analogue Data
ADC
Digital Data
11010011
Analogue signal sent from microphone
ADC converts analogue data to digital
(Sound Card is the ADC in this example)
Converted digital data recevied by computer



NOTE:
This process would repeat over and over again.

This is because the temperature (analogue data) is ***constantly changing***.

1 Thermometer gathers the constantly changing analogue temperature of the greenhouse.

2 ADC converts analogue temperature data to digital and sends it to the computer.

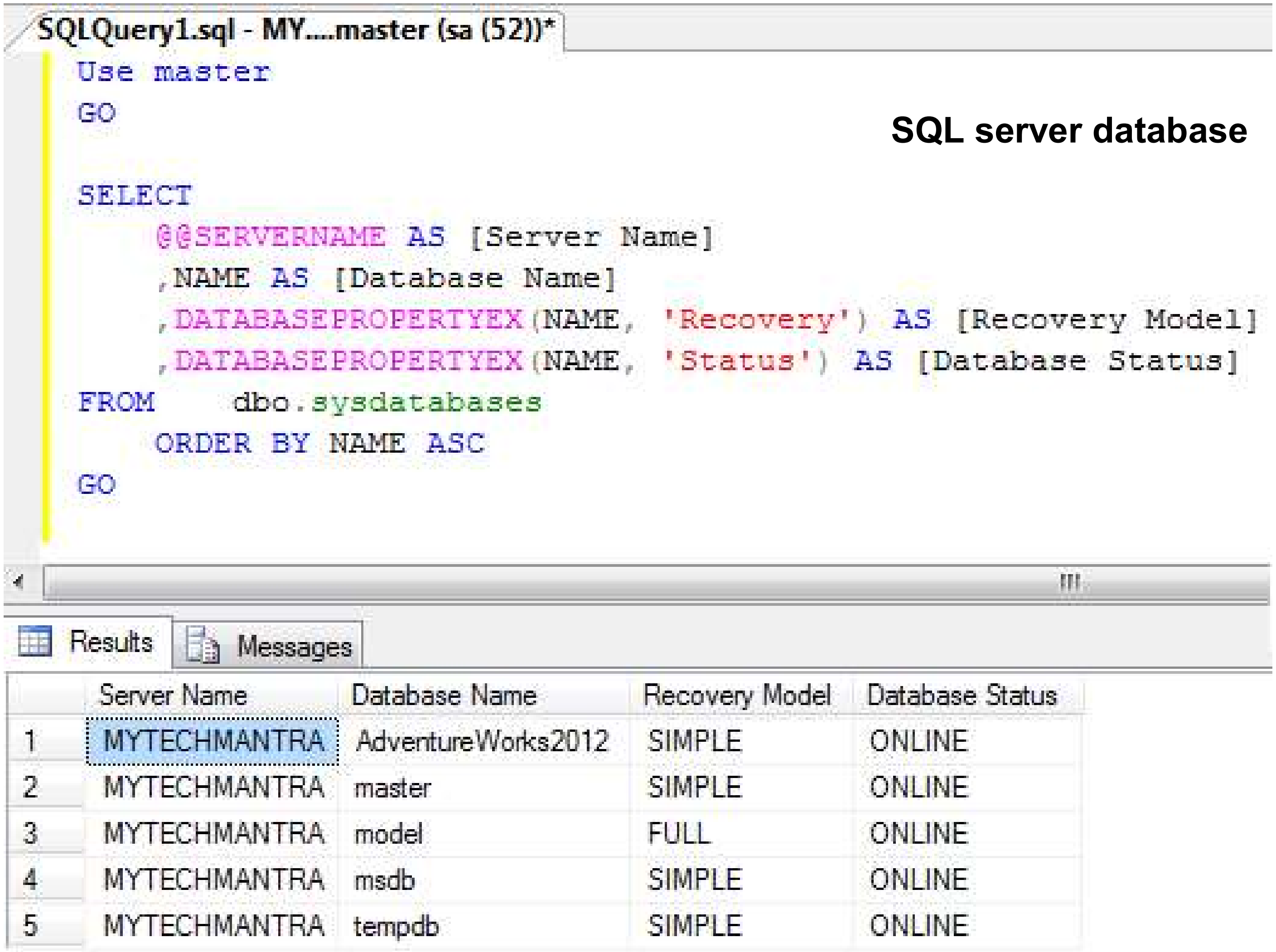
3 The computer processes the digital data and decides if the heater needs to be switched on or off.

4 The computer sends digital instructions to the DAC which converts them into analogue data.

Data is sent to the heater.

5 The heater follows the instructions sent by the computer and either turns itself on or off.

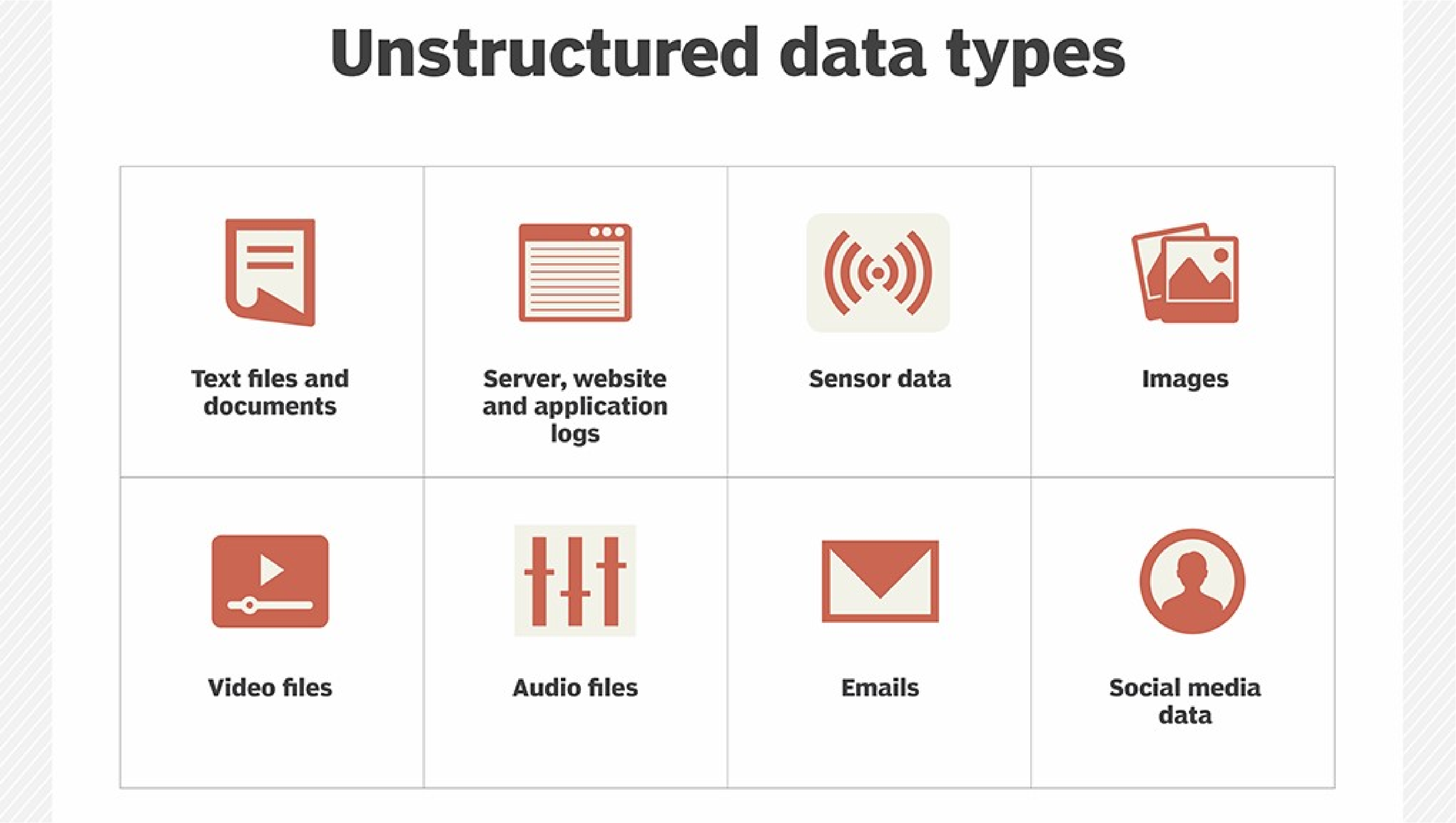
SQL server database



```
Use master
GO

SELECT
    @@SERVERNAME AS [Server Name]
    ,NAME AS [Database Name]
    ,DATABASEPROPERTYEX(NAME, 'Recovery') AS [Recovery Model]
    ,DATABASEPROPERTYEX(NAME, 'Status') AS [Database Status]
FROM    dbo.sysdatabases
    ORDER BY NAME ASC
GO
```

	Server Name	Database Name	Recovery Model	Database Status
1	MYTECHMANTRA	AdventureWorks2012	SIMPLE	ONLINE
2	MYTECHMANTRA	master	SIMPLE	ONLINE
3	MYTECHMANTRA	model	FULL	ONLINE
4	MYTECHMANTRA	msdb	SIMPLE	ONLINE
5	MYTECHMANTRA	tempdb	SIMPLE	ONLINE



Unstructured data types

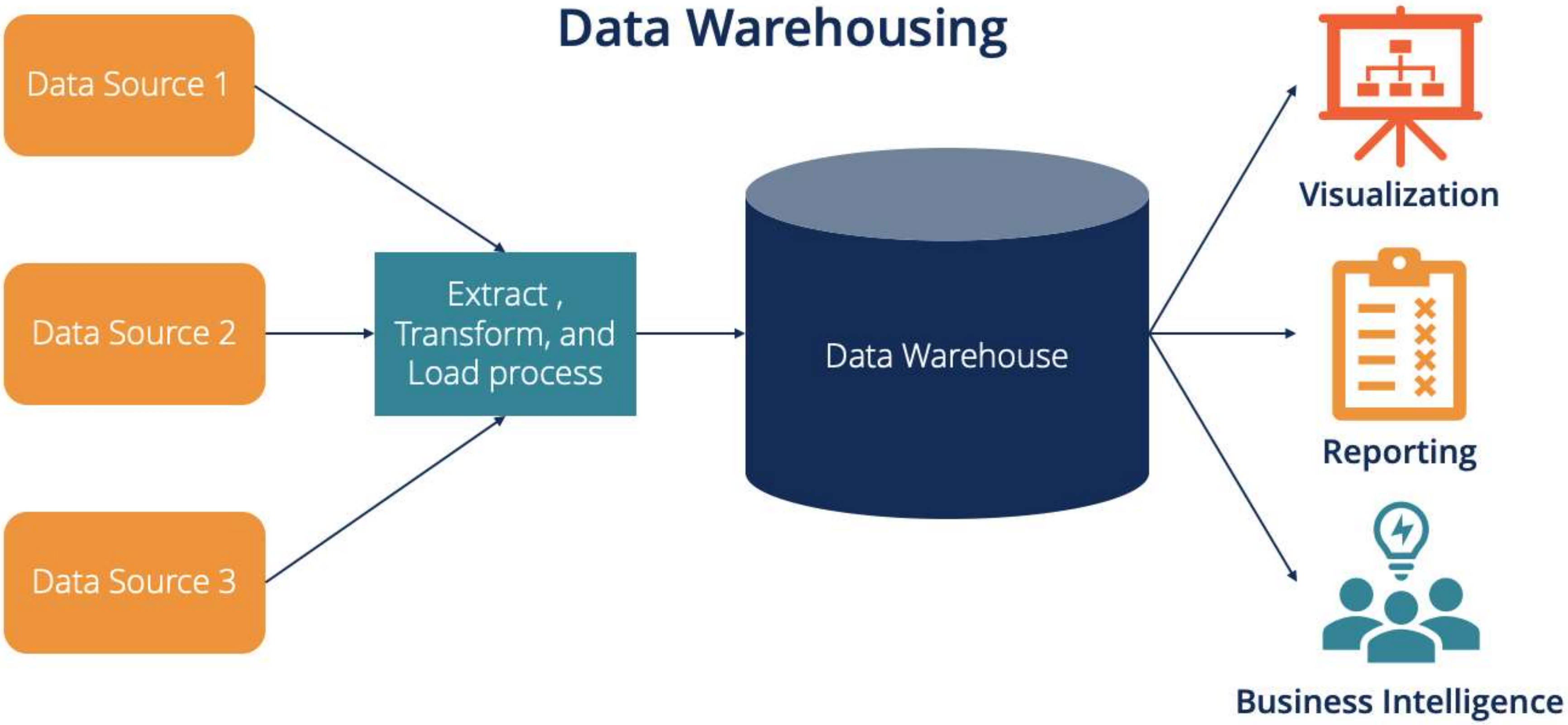
Text files and documents	Server, website and application logs	Sensor data	Images
Video files	Audio files	Emails	Social media data

Unstructured Data

- Unstructured data entails information with **no predefined conceptual definitions** and is **not easily interpreted or analyzed** by standard databases or data models.
- Unstructured data accounts for the majority of big data and comprises information such as **dates, numbers, and facts**.
- Big data examples of this type include **video and audio files, mobile activity, satellite imagery, and No-SQL databases**.
- **Photos** we upload on **Facebook or Instagram** and **videos that we watch on YouTube** or any other platform contribute to the growing pile of **unstructured data**.

- Data that does not reside in fixed locations generally refers to **free-form text**, which is abundant. Example- **Database, Data warehouse, Enterprise system.**
- Any data with **unknown form or structure** is classified as **unstructured data**.
- In addition to the **size being huge**, unstructured data poses **multiple challenges** in terms of its processing for deriving value out of it.
- A typical example of unstructured data is a **heterogeneous data (possibly ambiguous and low-quality due to missing values, high data redundancy, and untruthfulness)** source containing a combination of simple text files, images, videos etc.
- Now day, organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Data Warehousing



Data Source 1

Data Source 2

Data Source 3

Extract , Transform, and Load process

Data Warehouse

Visualization

Reporting

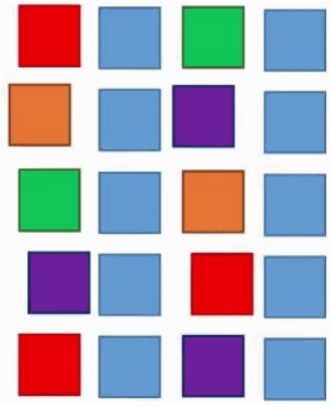
Business Intelligence

Data Warehousing **integrates data and information collected from various sources into one comprehensive database**.

Semi-structured data

Semi-structured data is **a hybrid of structured and unstructured data**. This means that it inherits a few characteristics of structured data but nonetheless contains information that **fails to have a definite structure** and does not comply with relational databases or formal structures of data models. For instance, JSON (JavaScript Object Notation) and XML (Extensible Markup Language) are typical examples of semi-structured data.

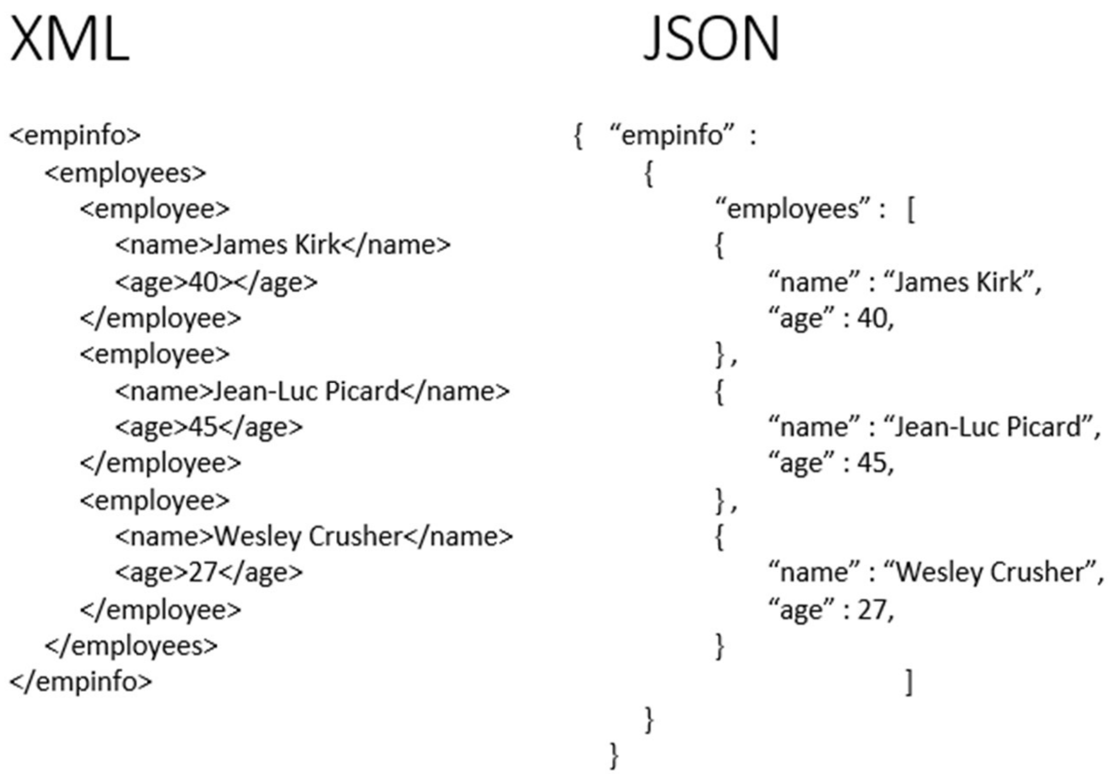
- Between the two forms where “tags” or structure are associated or embedded within unstructured data. XML, E-Mail



- Semi-structured data can contain both forms of data.

- We can see semi-structured data as structured in its form but it is actually not defined within, e.g. definition of a table in **relational DBMS**.

- **JSON** stands for **JavaScript Object Notation**.
- It is a text-based open standard data interchange format.
- JSON is lightweight and easy to read but doesn't provide schema or type information. It's great for sharing data between multiple applications.
- XML stands for an extensible markup language.
- XML is a markup language and file format for storing, transmitting, and reconstructing arbitrary data. It defines a set of rules for encoding documents in a format that is both human-readable and machine-readable

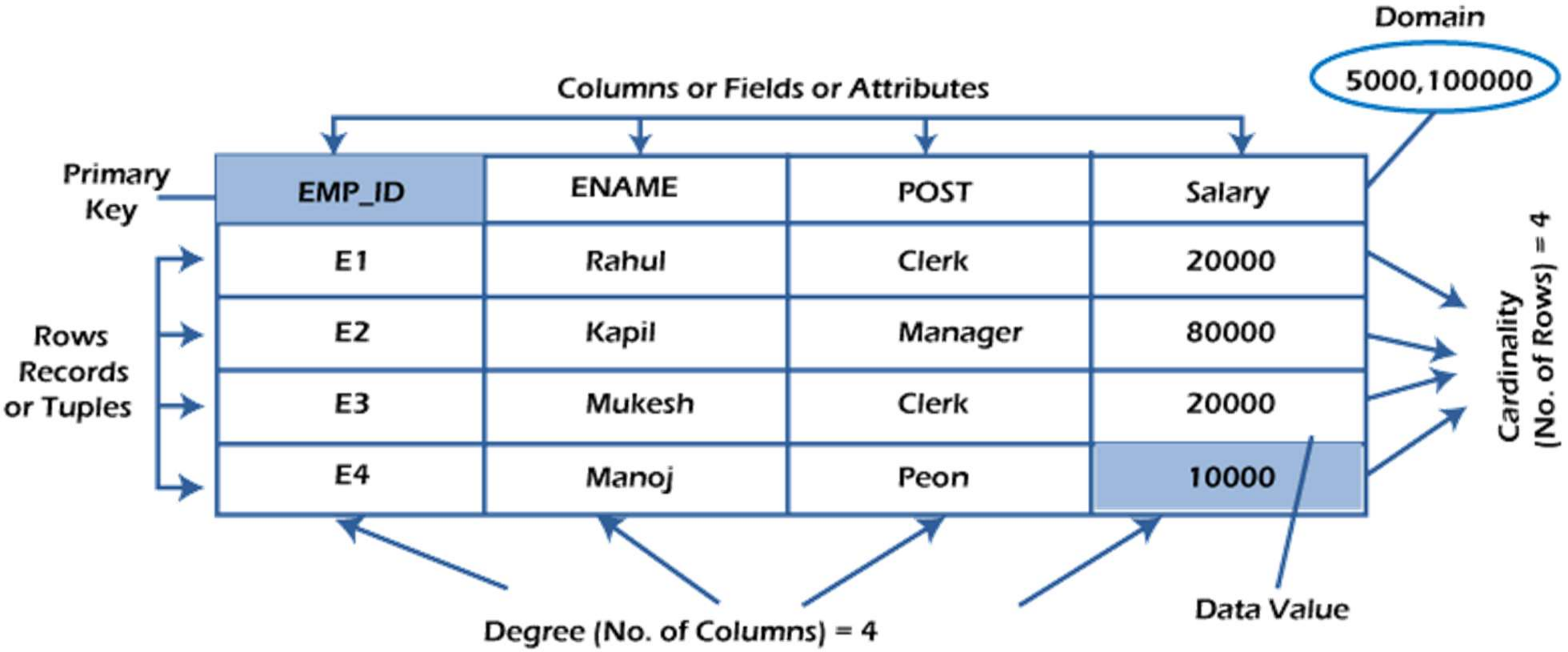


XML

```
<empinfo>
   <employees>
      <employee>
         <name>James Kirk</name>
         <age>40></age>
      </employee>
      <employee>
         <name>Jean-Luc Picard</name>
         <age>45</age>
      </employee>
      <employee>
         <name>Wesley Crusher</name>
         <age>27</age>
      </employee>
   </employees>
</empinfo>
```

JSON

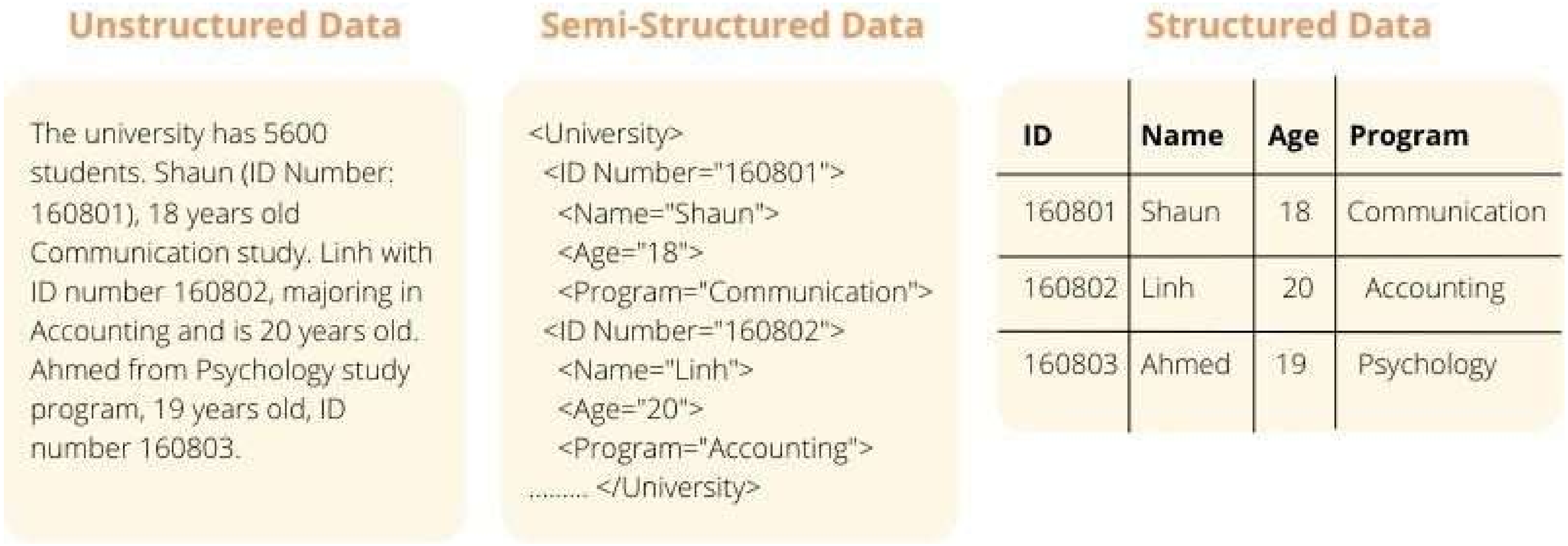
```
{  “empinfo” :
      {
            “employees” :  [
            {
                “name” : “James Kirk”,
                “age” : 40,
            },
            {
                “name” : “Jean-Luc Picard”,
                “age” : 45,
            },
            {
                “name” : “Wesley Crusher”,
                “age” : 27,
            }
                              ]
      }
   }
```



EMP_ID	ENAME	POST	Salary
E1	Rahul	Clerk	20000
E2	Kapil	Manager	80000
E3	Mukesh	Clerk	20000
E4	Manoj	Peon	10000

Domain
5000,100000
Columns or Fields or Attributes
Primary Key
Rows Records or Tuples
Cardinality (No. of Rows) = 4
Data Value
Degree (No. of Columns) = 4

Table. Relational DBMS



Unstructured Data

The university has 5600 students. Shaun (ID Number: 160801), 18 years old Communication study. Linh with ID number 160802, majoring in Accounting and is 20 years old. Ahmed from Psychology study program, 19 years old, ID number 160803.

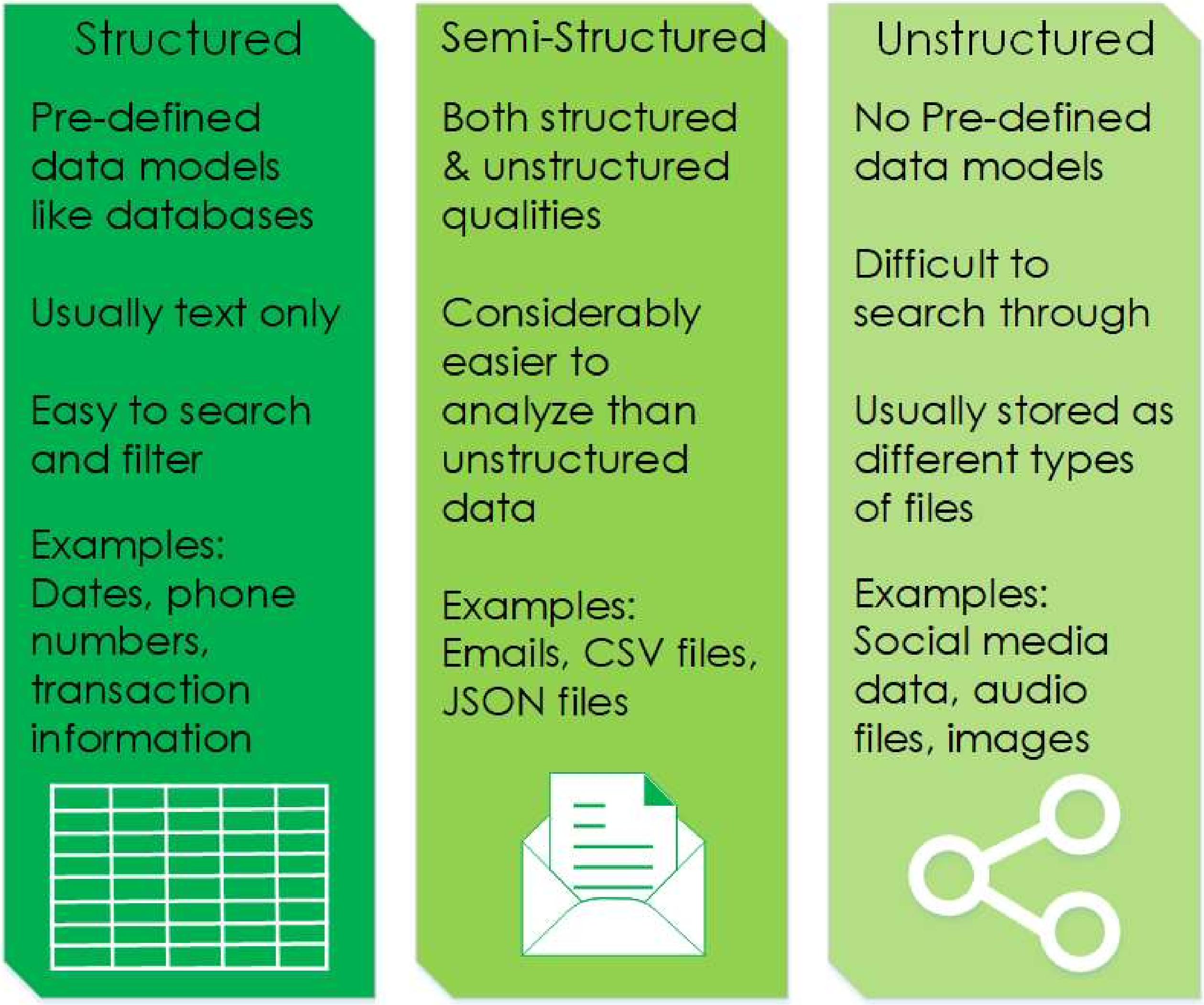
Semi-Structured Data

```
<University>
 <ID Number="160801">
   <Name="Shaun">
   <Age="18">
   <Program="Communication">
 <ID Number="160802">
   <Name="Linh">
   <Age="20">
   <Program="Accounting">
......... </University>
```

Structured Data

ID	Name	Age	Program
160801	Shaun	18	Communication
160802	Linh	20	Accounting
160803	Ahmed	19	Psychology

Summary



Structured

Pre-defined data models like databases

Usually text only

Easy to search and filter

Examples: Dates, phone numbers, transaction information

Semi-Structured

Both structured & unstructured qualities

Considerably easier to analyze than unstructured data

Examples: Emails, CSV files, JSON files

Unstructured

No Pre-defined data models

Difficult to search through

Usually stored as different types of files

Examples: Social media data, audio files, images

CSV (**comma-separated values**) file is a text file that has a specific format which allows data to be saved in a table structured format.

THANK YOU