

Unit 14

Business Intelligence

DATA AND DATA TYPES

Data everywhere

“The world we live in is complex, random, and uncertain. At the same time, **it’s one big data-generating machine.**”

Scenario 1: you were hired at a startup company and the company just launched its website, your first task is to monitor the website traffic

- Monitoring → A process

Scenario 2: “Imagine spending 24 hours looking out the window, and for every minute, counting and recording the number of cars which pass your house.”

- Monitoring and counting → A process

What is data?

“Distinct pieces of **information**, usually **formatted** in a special way

Data can exist in a variety of forms:

- as numbers
- as text on paper
- as bits and bytes stored in electronic memory
- as facts stored in a person's mind

Since the mid-1900s, people have used the word data to mean computer information that is transmitted or stored

Data is the plural of **datum**, a **single piece of information**.”

Key Terms for Data Types

Numeric

Data that are expressed on a numeric scale.

Continuous

Data that can take on any value in an interval. (*Synonyms*: interval, float, numeric)

Discrete

Data that can take on only integer values, such as counts. (*Synonyms*: integer, count)

Categorical

Data that can take on only a specific set of values representing a set of possible categories. (*Synonyms*: enums, enumerated, factors, nominal)

Binary

A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (*Synonyms*: dichotomous, logical, indicator, boolean)

Ordinal

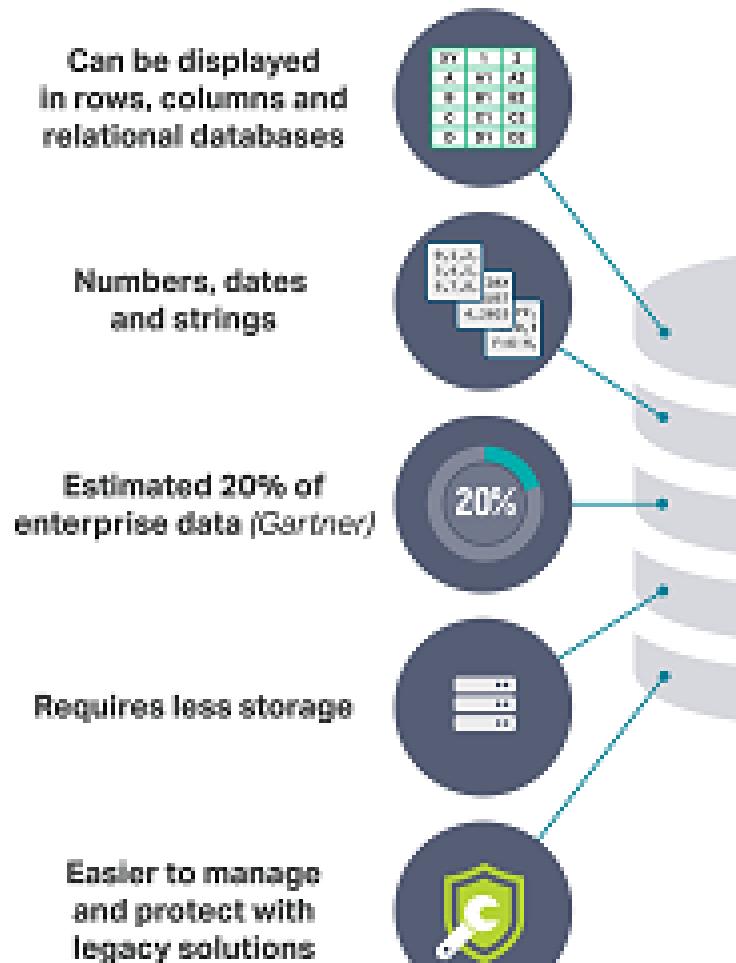
Categorical data that has an explicit ordering. (*Synonym*: ordered factor)

Structured data (Taxonomy of data)

Why do we care about data types?

- ❖ Help determine the type of visual display.
- ❖ Help determine how to manipulate data.
- ❖ Help determine how to store data.
- ❖ Help determine type of data analysis or statistical model.

Structured Data



Structured data

Conforms to a data model

Well defined structure

Can be easily accessed

Can be used by a person or a computer program

Structured data Characteristics

Stored in well-defined schemas such as Databases.

Clear types

Data conforms to a data model Clear types

Data resides in fixed fields

Easy to access and query

Rows & columns

Managed by SQL

Sources of Structured data



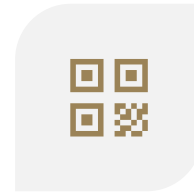
**SQL
DATABASES**



**SPREADSHEETS
SUCH AS EXCEL**



OLTP SYSTEMS



RFID TAGS

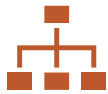


SERVER LOGS



**MEDICAL
DEVICES**

Advantages of Structured data



Well defined structure(easy storage and access)



Data can be indexed(Help in searching)



Mining is easy



Update/delete/insert is easy as well



BI operations is easy as well(ie data warehousing)



Scalable



Security (Column level)



Structured data is foundation to Big data

Semi-structured data Characteristics

Has no data model
but have some
structure

Can't be stored in
rows and Columns

Contains
tags(elements) –
Metadata

Little metadata

Attribute types
may differ even in
the same group

Lack of Structure
makes it hard to be
used in a computer
program

Sources of Semi-Structured data

XML

Email

Binary .exe

Zipped files

Data from multiple sources

Webpages

JSON

```
<studentsList>
  <student id="1">
    <firstName>Greg</firstName>
    <lastName>Dean</lastName>
    <certificate>True</certificate>
    <scores>
      <module1>70</module1>
      <module12>80</module12>
      <module3>90</module3>
    </scores>
  </student>
  <student ind="2">
    <firstName>Wirt</firstName>
    <lastName>Wood</lastName>
    <certificate>True</certificate>
    <scores>
      <module1>80</module1>
      <module12>80.2</module12>
      <module3>80</module3>
    </scores>
  </student>
</studentsList>
```

Source: HubSpot blog.

Advantages of Semi-structured data

**Not
constrained by
a fixed schema**

Flexible

Portable

**Can be viewed
as structured
data**

**Don't need to
know SQL**

**Deals with
heterogeneity
of sources**

Disadvantages of Semi-structured data

Lacks uniform
=> difficult to
store

Hard to
identify
relationships

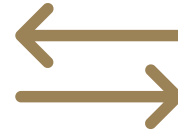
Queries less
efficient



Use special DBMS to
store this type of data



XML – allows for tags
and attributes



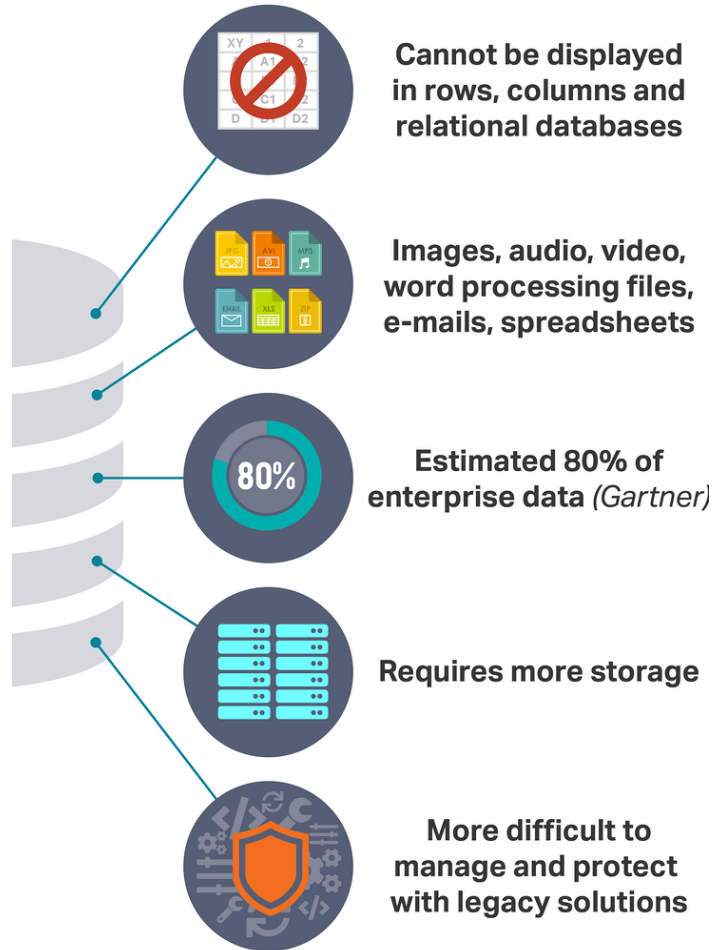
Object Exchange
Model(OEM) – represent
data by graph



RDBMS utilize mapping
technique

Solution
Semi-
structured
data

Unstructured Data



Unstructured data

- ❖ Has no data model but have some structure
- ❖ Can't be stored in rows and Columns
- ❖ Does not have semantic rules
- ❖ Lacks format or sequence
- ❖ Lack of Structure; Can't be used in a computer program



Images



Videos



Memo



Report



.docx, .ppt



Survey

Sources of
Unstructured
data

Advantages of Unstructured data

Support for lack of
format data

No fixed schema

Flexible

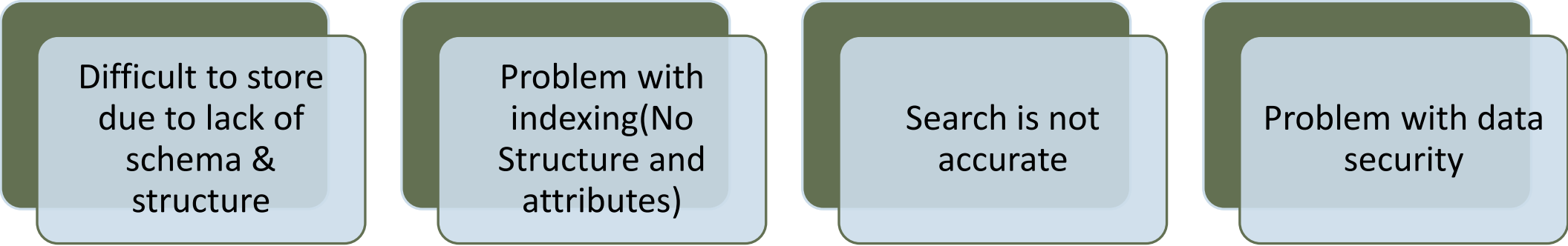
Portable

Scalable

Works fine with
heterogeneity of
sources

Analyzed by
intelligence and
analytics
applications

Disadvantages of Unstructured data



Difficult to store
due to lack of
schema &
structure

Problem with
indexing(No
Structure and
attributes)

Search is not
accurate

Problem with data
security

Problem with Unstructured data



STORAGE



**CHALLENGE TO STORE
VIDEOS, IMAGES AND
AUDIO.**



**HARD TO UPDATE,
DELETE & SEARCH(
NEED STRUCTURED DB)**



STORAGE COST



INDEXING IS DIFFICULT

Solution Unstructured data



**CONVERT TO MANAGEABLE
FORMATS**



**USE CONTENT ADDRESSABLE
SYSTEM(CAS) –METADATA AND
UNIQUE NAME**



STORE VIA XML AND RDBMS (BLOBS)

Data input and data capture



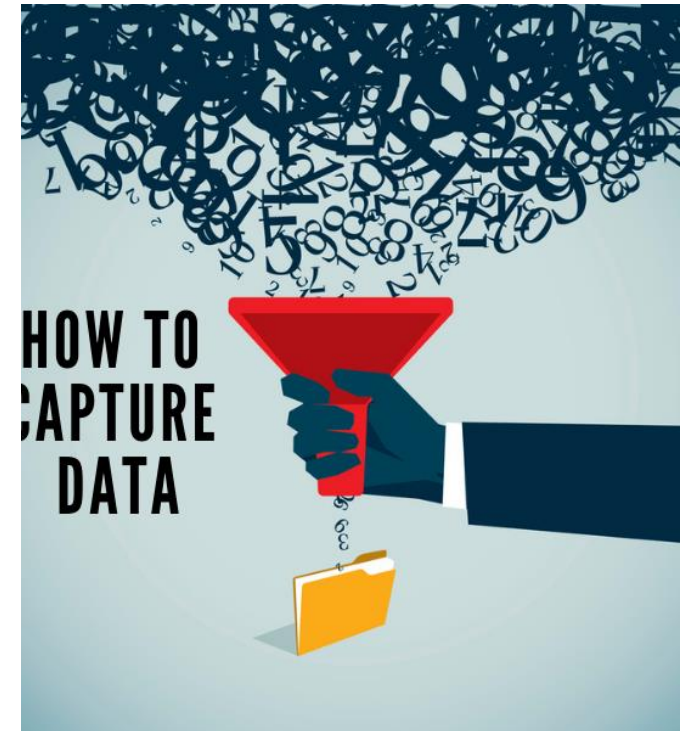
Data Input

- “Any information that is provided to a computer or a software program is known as input.
- Since the information provided is also considered to be data, the process of providing information to the computer is also known as data input.”

Data Capture

Data capture is the process of collecting structured and unstructured information electronically and converting it into data readable by a computer.

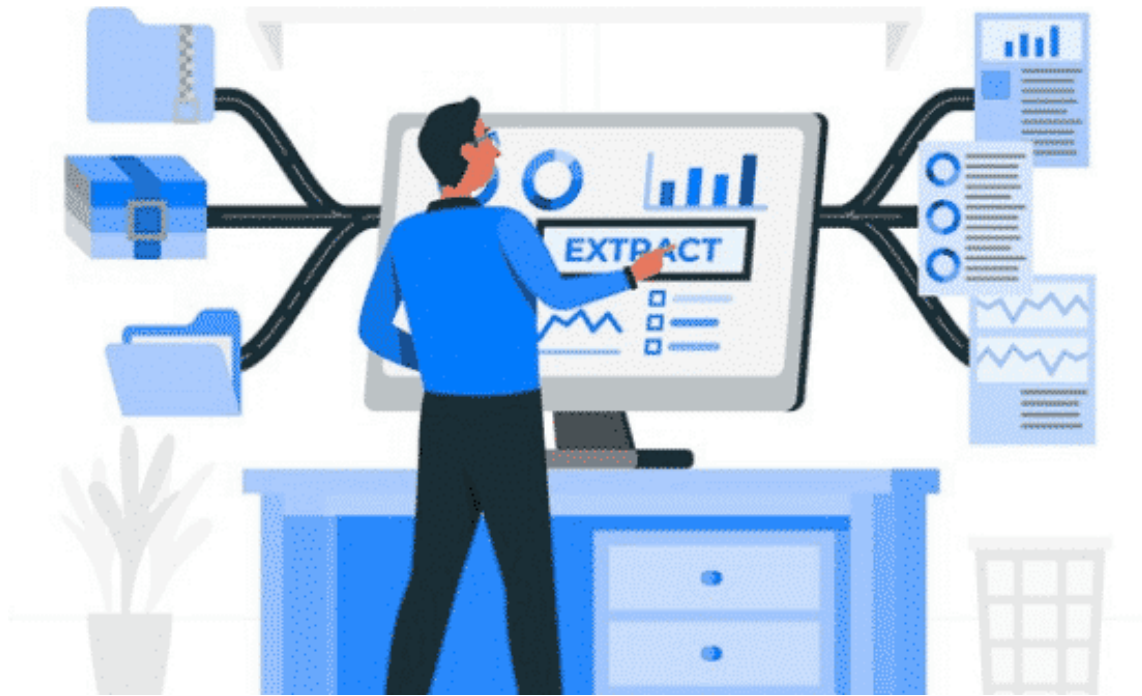
Data capture is very similar to data entry but used mostly on data sources that contain basic response types.



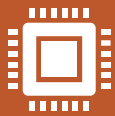
DATA CAPTURE METHODS

Data Capture Methods

- ❖ Manual data capture
- ❖ Automated Data Capture
- ❖ OCR (Optical Character Recognition)
- ❖ ICR (Intelligent Character Recognition)
- ❖ Barcode/ QR Code Recognition
- ❖ Voice Capture
- ❖ Smart Cards



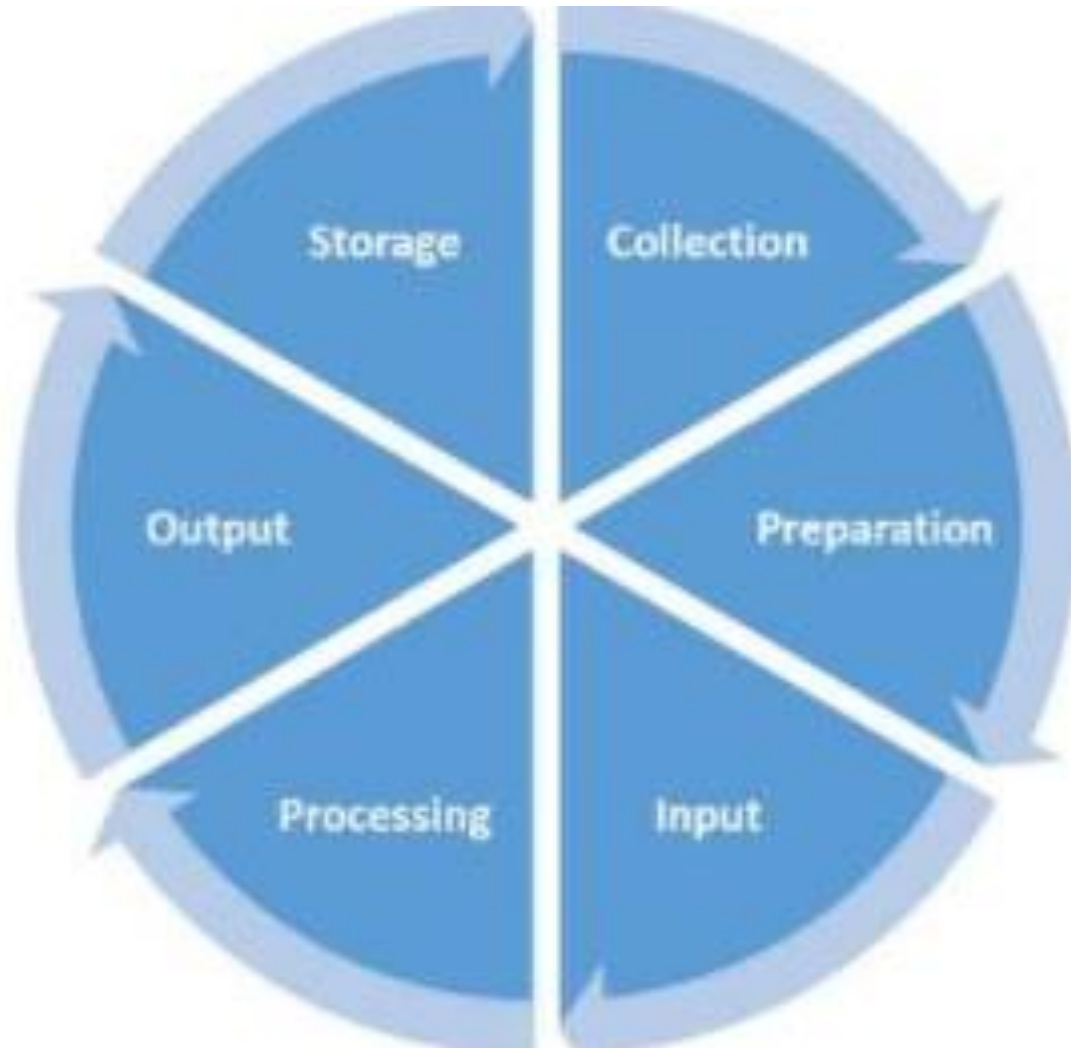
Data Capture Method



Manual Data Capture: This method uses manual keying of required data from written forms into a computer for digitized access. It is suitable for businesses where the volume of data is low and variable. Manual data capture depends on human labor making it susceptible to errors or data omissions, the very reason why automated data capture technology is becoming an ideal solution.



Automated Data Capture: Automated data capture ensures that businesses can function smoothly not only by managing data but also by reducing cost and labor inefficiency. Varied forms of data capture are available to suit the requirements of different businesses.



Data Processing

Data processing is the method of collecting raw data and translating it into usable information.

It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization.

The raw data is collected, filtered, sorted, processed, analyzed, stored, and then presented in a readable format.

Types of Data Processing

- ❖ **Commercial Data Processing**
- ❖ **Scientific Data Processing**
- ❖ **Batch Processing**
- ❖ **Online Processing**
- ❖ **Real-Time Processing**



❖ Security Considerations

- ❖ **Unauthorized access**
- ❖ **Corruption**
- ❖ **Theft**



Databases

“A database is an organized collection of structured information, or data, typically stored electronically in a computer system.

A database is usually controlled by a database management system (DBMS)”



Database Examples



- ❖ Oracle
- ❖ SQL Server
- ❖ MySQL
- ❖ NoSQL

Exercises

How does BI help these companies?

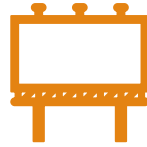
What type of data is produced and used?



Boeing



Grill Mark



Teeba Juice



Shahid



Spotify