# Unit 3

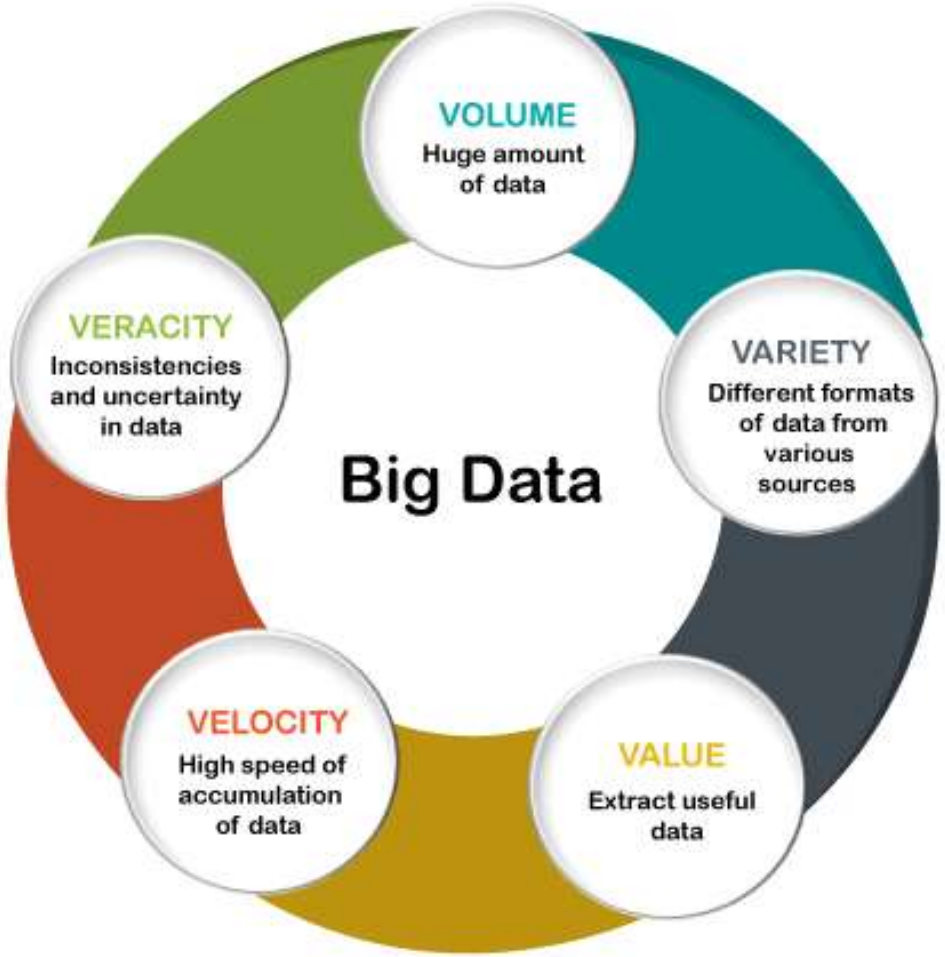
Concept of Data Science

# Traits of Big data

- Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit.
- It is used by many **multinational companies** to **process** the data and business of many **organizations**. The data flow would exceed **150 exabytes** per day before replication**.**
- There are five v's of Big Data that explains the characteristics.

5 V's of Big Data

- **Volume**
- **Veracity**
- **Variety**
- **Value**
- **Velocity**


VOLUME
Huge amount of data
VARIETY
Different formats of data from various sources
VALUE
Extract useful data
VELOCITY
High speed of accumulation of data
VERACITY
Inconsistencies and uncertainty in data
Big Data

# Traits of Big Data

## Volume

- The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more.
- **Facebook** can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.

Variety

- Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos,** etc.

# The data is categorized as below:

- **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.
- **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.
- **Unstructured Data**: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.
- **Quasi-structured Data:**The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.
-

## Veracity

- Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development.
- For example, **Facebook posts** with hashtags.

## Value

- Value is an essential characteristic of big data. It is not the data that we process or store. It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.

# Advantages of Big Data

- Better Decision Making
- Reduce costs of business processes
- Fraud Detection
- Increased productivity
- Improved customer service
- Increased agility

# Web Scrapping

- Suppose you want some information from a website? Let's say a paragraph on Kolhapur! What do you do? Well, you can copy and paste the information from Wikipedia to your own file.
- But what if you want to get large amounts of information from a website as quickly as possible? Such as large amounts of data from a website to train a Machine Learning algorithm?
- In such a situation, copying and pasting will not work! And that's when you'll need to use **Web Scraping**.

# Web scrapping

## What is web scraping

- Web scraping is an automatic method to obtain large amounts of data from websites.
- Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications.
- There are many different ways to perform web scraping to obtain data from websites.
- These include using online services, particular API’s or even creating your code for web scraping from scratch.
- Many large websites, like Google, Twitter, Facebook, StackOverflow, etc. have API’s that allow you to access their data in a structured format.

# Web Scrapping

- Web scraping is used in a variety of digital businesses that rely on data harvesting. Legitimate use cases include:
- Search engine bots crawling a site, analyzing its content and then ranking it.
- Price comparison sites deploying bots to auto-fetch prices and product descriptions for allied seller websites.
- Market research companies using scrapers to pull data from forums and social media (e.g., for sentiment analysis).

# Web Scapping

- **Scraper tools and bots**
- Web scraping tools are software (i.e., bots) programmed to sift through databases and extract information. A variety of bot types are used, many being fully customizable to:
- Recognize unique HTML site structures
- Extract and transform content
- Store scraped data
- Extract data from APIs

# Web Scrap Code

- import requests

- # Making a GET request
- r = requests.get(https://en.wikipedia.org/wiki/Kolhapur/')

- # check status code for response received
- # success code - 200
- print(r)

- # print content of request
- print(r.content)

# Reporting vs. Analysis: What's the Difference?

- **Reporting:** The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

- **Analysis:** The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

| Reporting | Analysis |
| --- | --- |
| ➢ Reporting translates raw data into **information** | ➢ Analysis transforms data and information into **insights**. |
| ➢ Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges. | ➢ The goal of analysis is to **answer questions** by interpreting the data at a deeper level and providing actionable recommendations. |
| ➢ Good reporting should **raise questions** about the business from its end users. | ➢ Through the process of performing analysis you may raise additional questions, but the goal is to identify answers, or at least potential answers that can be tested. |
| ➢ In summary, reporting shows you ***what is happening*** | ➢ n summary, analysis focuses on explaining ***why it is happening*** and ***what you can do about it.*** |

# Tools For Data Sciecne

**Most Used Data Science Tools for Essential Data Science Ingredients are**