

With big data analytics, you can ultimately fuel better and faster decision-making, modelling and predicting of future outcomes and enhanced business intelligence. Businesses can access a large volume of data and analyse a large variety of data to gain new insights and take action.

Analysing data from sensors, devices, video, logs, transactional applications, web, and social media empowers an organisation to be *data driven*.

Types of Big Data

There are three types of big data: structured, unstructured, and semi structured.

Structured data: Any data that can be stored, accessed, and processed in the form of a fixed format. Nowadays, the size of such data grows to a huge extent with typical sizes being in the range of multiple zettabytes (one billion terabytes forms a zettabyte).

Examples: Data stored in a relational database management system is one example of a ‘structured’ data. Data in tables e.g., timetable, pay tables, list of keywords at the end of the publication, etc.

An ‘employee’ table in a database is an example of structured data.

Unstructured data: Any data with unknown form or the 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.

Example: A heterogeneous data source containing a combination of text files, images, videos etc.

Nowadays organisations have a wealth of unstructured data with interesting information available, but they don’t know how to derive hidden values out of it since this data is in its raw form or unstructured format.

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

Examples: Personal web pages: contain texts, images, videos, links to social networks, links to communicate with friends, etc. Another example is the web pages of online shops with advertising, payment terminals.

The characteristics of Big Data are commonly referred to as the *four volumes of Big Data*.

The *Volume of Data* refers to the size of the data sets that need to be analysed and processed, which are now frequently larger than terabytes¹ and petabytes². The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities.

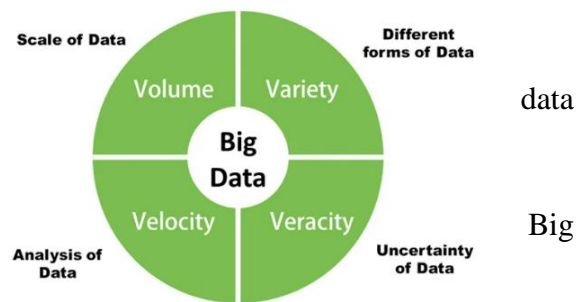
¹ Terabyte: A bit (binary digit) has a single binary value of either 0 or 1, is the smallest unit of data in a computer system. A terabyte is one of the largest units of storage media that products on the market use today. There are units larger than a terabyte: petabyte, exabyte, zettabyte, yottabyte and brontobyte. A geopbyte refers to 10³⁰ bytes. In practical terms, a terabyte of data is equivalent to: 728.177 floppy disks, 40 single-layer Blu-ray discs, 85.899.345 pages of Word documents, 500 hours of movies, 310,000 photos, 17,000 hours of music.

² Petabyte: 1 PB = 1.000 TB, e.g., 1 PB is equal to over 1 quadrillion bytes.

This means that the data sets in Big Data are too large to process with a regular laptop or desktop processor.

The Velocity of Big Data refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. An example of that is generated with high velocity would be Twitter messages or Facebook communication.

The Variety of Big Data makes Big Data really big. Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. An example of high variety data sets would be the audio and video files that are generated at various locations in a city.



The Veracity of Big Data refers to the quality of the data that is being analysed. High veracity data has many records that are valuable to analyse and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.

Data that is high volume, high velocity and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labelled Big Data.

Advantages of Big Data processing

- Businesses can utilise outside intelligence while making decisions.
- Access to social data from search engines and sites like Facebook or Twitter are enabling organisations to fine tune their business strategies.
- Improved customer service.
- Early identification of risk to the product/services, if any.
- Better operational efficiency.
- Big Data technologies can be used for creating data warehouses.

Summary

Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. Big Data can be 1) structured, 2) unstructured, 3) semi-structured. The velocity of Big Data refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct processing techniques. The variety of Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and



ITFARM

unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. The veracity of Big Data refers to the quality of the data that is being analysed: we distinguish between high veracity data and low veracity data. Volume, variety, velocity, and variability are the four Big Data characteristics. Improved customer service, better operational efficiency, better decision making are few advantages of Big Data.

Links to relevant topics

https://en.wikipedia.org/wiki/Big_data

<https://www.guru99.com/what-is-big-data.html>

<https://www.gartner.com/en/information-technology/glossary/big-data>

Key words

quantities

characters

symbols

Big Data

artificial intelligence

mobile devices

social media

Internet of Things

data driven

structured data

unstructured data

semi-structured data

four volumes of Big Data

velocity of Big Data

variety of Big Data

veracity of Big Data



Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Project: Erasmus+ KA220-ADU, Duration: since 01-01-2022 till 30 01-07-2024



**Co-funded by
the European Union**