What is Big Data ?

DataVault LLC

Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time.

In short, such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.


That how Huge Big Data is!

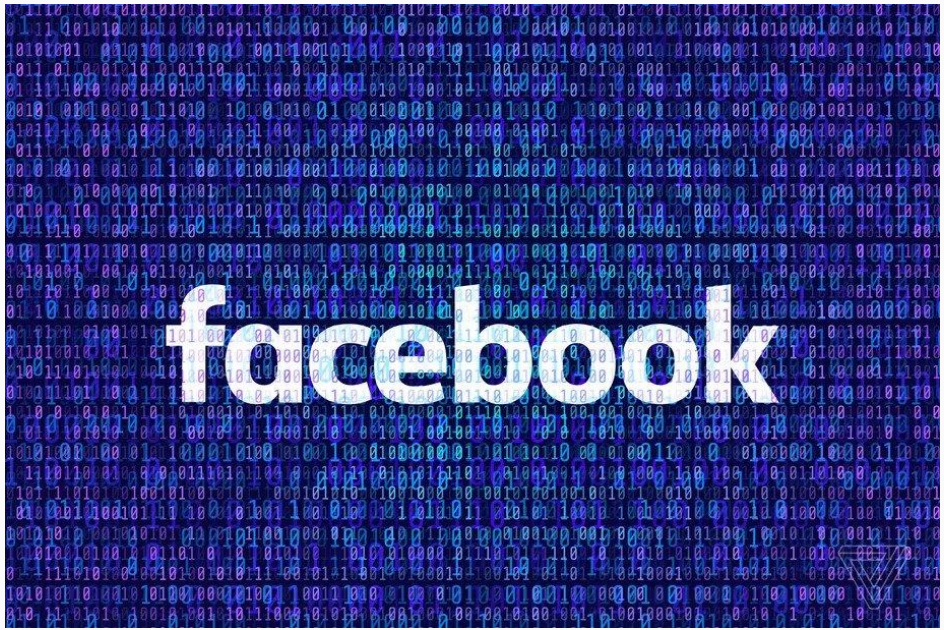
Some Examples...

The **New York Stock Exchange** generates about ***one terabyte*** of new trade data per day.

Some more Examples...

Social Media

The statistic shows that ***500+terabytes*** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments, etc.



Some more Examples...

Aviation

A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time.

With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

Type of BIG DATA

Big Data could be found in three forms:

1- Structured Data.

2- Semi-Structured Data.

3- Unstructured Data.

Structured Data:

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 a size of such data grows to a huge extent, typical sizes are being in the rage of multiple **zettabytes.**

One billion terabytes forms ***a zettabyte***, 1 terabyte = 1000 Gigabytes.

Structured data example:

CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

Semi-structured Data:

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

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

Example of semi-structured data is a data represented in an XML or JSON file.

Semi-Structured Data Example:

Unstructured Data:

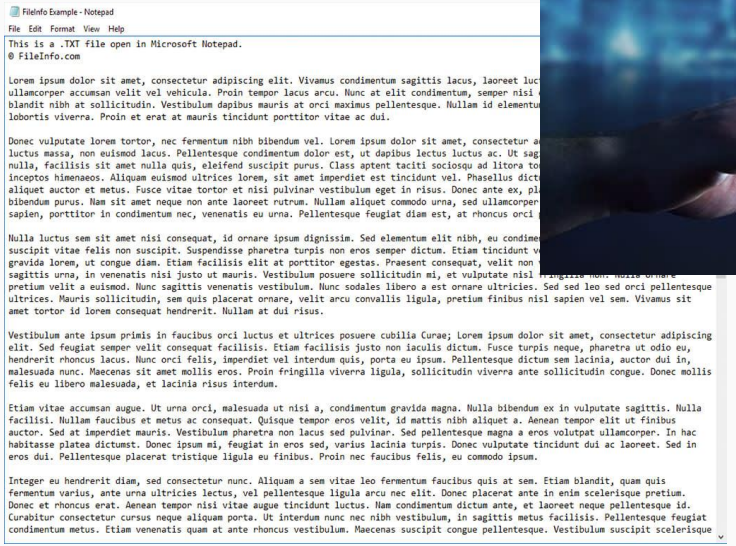
Any data with unknown form or the structure is classified as unstructured data.

In addition to the size being huge, unstructured data poses multiple challenges in terms of its processing for deriving value out of it.

A typical example of unstructured data is a heterogeneous data 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.

Unstructured Data Example:



Characteristics of BIG DATA

1. Volume
2. Velocity
3. Variety
4. Veracity
5. Value

Volume:

It refers to the size of Big Data.

Data can be considered Big Data or not is based on the volume.

The rapidly increasing volume data is due to cloud-computing traffic, IoT, mobile traffic etc.

Velocity:

It refers to the speed at which the data is getting accumulated. This is mainly due to IoTs, mobile data, social media etc.

In the year 2000, Google was receiving 32.8 million searches per day. As for 2018, Google was receiving 5.6 billion searches per day!

Variety:

It refers to **Structured**, **Semi-structured** and **Unstructured** data due to different sources of data generated either by humans or by machines.

Veracity:

It refers to the assurance of **quality/integrity/credibility/accuracy** of the data.

Since the data is collected from multiple sources, we need to check the data for accuracy before using it for business insights.

Value:

Just because we collected lots of Data, it's of no value unless we garner some insights out of it.

Value refers to how useful the data is in decision making. We need to extract the value of the Big Data using proper analytics.

Benefits of BIG DATA Processing

Ability to process

Big Data

brings in multiple benefits,

such as:

- **Businesses can utilize outside intelligence while taking decisions**

Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.

- **Improved customer service**

Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

- **Early identification of risk to the product/services, if any**
- **Better operational efficiency**

Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

In conclusion:

- Big Data is defined as data that is huge in size. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.
- Examples of Big Data generation includes stock exchanges, social media sites, jet engines, etc.
- Big Data could be 1) Structured, 2) Semi-structured, 3) Unstructured.
- 5V’s: Volume, Velocity, Variety, Veracity, Value.
- Improved customer service, better operational efficiency, Better Decision Making are few advantages of Big Data