Overview of Big Data

We are living the world of data; we can easily say that the world is going through datafication where everyday activity and interactions are converted into tabulated information which is being available for the analysis process. Increase in connection of people and smart devices producing huge amount of data every minute every second has brought the requirement of creating large databases and handling them to extract the data for the analysis.

Big Data refers to accumulation of data which is too large and complex for processing using traditional database management tools. Usually over 1 Peta byte is referred as Big Data. It requires specific tools & techniques to process such type of data.

5Vs of Big Data

1. **Volume**
   It refers to size of the datset.
2. **Variety**
   It refers to data accumulation from different data sources having structured, semi-structured & unstructured data.
3. **Velocity**
   It refers to the speed at which the data is being generated.
4. **Veracity**
   It refers to integrity of data. It measures the data quality and trustworthiness of the data.
5. **Value**
   It measures usefulness of the data i.e. after analysis what value it can add.

The longer it takes for the data to generate meaningful insights the lower is the value. At the outset we can say value of data is dependent on veracity. High is the veracity, high the value.

Applications of Big Data

- Healthcare Industry
- E-Commerce and Marketing Industry
- Banking & Finance Industry
- Online Streaming Platforms

Data Types in Big Data

- Structured Data
- Semi-Structured Data
- Unstructured Data

Structured Data

It exhibits a particular order for storing the data working with it. The data attributes are usually related and often the base of analysis. It is usually generated by machines or compiled by humans. It is normally stored in relational databases, CSV files or spreadsheets.

Semi-Structured Data

It has some definitive patterns for storage, but the data attributes may not be inter-related. Here the data could be hierarchical or graph based in nature. It is normally stored in XMLs, JSON. Here data sources can be website feeds, sensors or any other applications.

Unstructured Data

It does not exhibit a fixed pattern. This is most common type of data type in big data. Videos, audios, likes, tweets shares, PDF documents, images, text files are the most common examples of unstructured data. There is need of special tools and techniques to process these types of data.

Comparison factors	**Structured Data**	**Semi-Structured Data**	**Unstructured Data**
Volume of data	Low	Medium	High
Processing Flexibility	Low	Medium	High
Data Generated by	Humans and Machines	Machines	Humans
Data Storage Format	RDBMS	Textual Files	Binary Files
Patterns and Schemas	Fixed	Flexible	Random
Specialized Tools Required	No	No	Yes