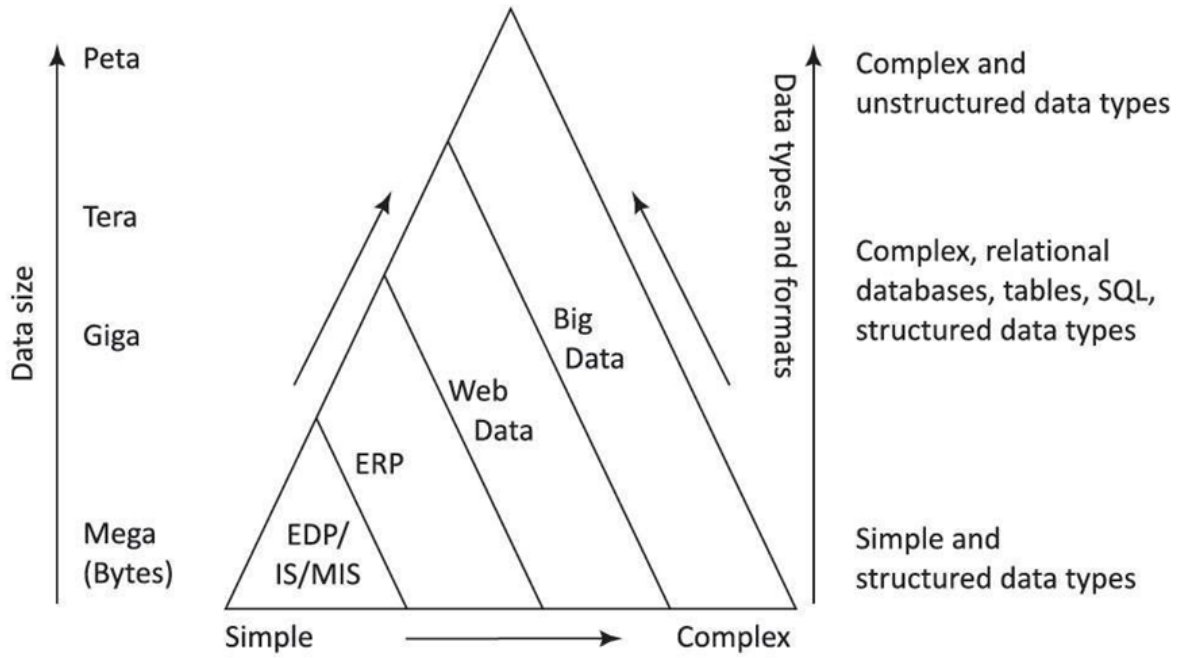
MODULE 1: INTRODUCTION TO BIG DATA ANALYTICS

Need of Big Data

The rise in technology has led to the production and storage of voluminous amounts data. Earlier megabytes were used but nowadays petabytes are used for processing, analysis, discovering new facts and generating new knowledge. Conventional systems for storage, processing and analysis pose challenges in large growth in volume of data, variety of data, various forms and formats, increasing complexity, faster generation of data and need of quickly processing, analyzing and usage.

Figure shows data usage and growth. As size and complexity increase, the proportion of unstructured



Evolution of Big data and their Characteristics

Data

Data has several definitions.Usages can be singular or plural.

- “Data is information, usually in the form of facts or statistics that one can analyze or use for further calculations.”
- "Data is information that can be stored and used by a computer program"
- "Data is information presented in numbers, letters, or other form".

- "Data is information from series of observations, measurements or facts".
- "Data is information from series of behavioural observations, measurements or facts",

Web Data

. Web data is the data present on web servers (or enterprise servers) in the form of text, images, videos,audios and multimedia files for web users. A user (client software) interacts with this data. A client can access (pull) data of responses from a server. The data can also publish (push) or post (after registering subscription) from a server. Internet applications including web sites, web services, web portals, online business applications, emails, chats, tweets and social networks provide and consume the web data.

Classification of Data-Structured, Semi-structured and Unstructured

- Data can be classified as structured, semi-structured, multi-structured and unstructured:
- Structured data conform and associate with data schemas and data models. Structured tables (rows and columns). Nearly 15-20% data are in structured or semi-structured form, Unstructured data
- Unstructured data do not conform and associate with any data models.

Structured Data:

Structured data enables the following:

- Data insert, delete, update and append , Indexing to enable faster data retrieval
- Scalability which enables increasing or decreasing capacities and data processing operations such as, storing, processing and analytics
- encryption and decryption for data security.

Semi-Structured Data

- Examples of semi-structured data are XML and JSON documents.
- Semi-structured data contain tags or other markers, which separate semantic elements and enforce hierarchies of records and fields within the data.
- Semi-structured form of data does not conform and associate with formal data model structures.
- Data do not associate data models, such as the relational database and table models.

Multi-Structured Data

- Multi-structured data refers to data consisting of multiple formats of data, viz. structured, semi-structured and/or unstructured data.
- Multi-structured data sets can have many formats. They are found in non-transactional systems.
- For example, streaming data on customer interactions, data of multiple sensors, data at web or enterprise server or the data warehouse data in multiple formats.
- Multi-or semi-structured data has some semantic meanings and data is in both structured and unstructured formats.

Unstructured Data

- Unstructured data does not possess data feature such as a table or a database.
- Unstructured data are found in file type such as TXT, CSV.
- Data may be have internal structures, such as in e-mails.
- The relationships, schema and features need to be separately established

BIG DATA:

- Big Data is high-volume, high-velocity and/or high-variety information asset that requires new forms of processing for enhanced decision making, insight discovery and process optimization .

- "A collection of data sets so large or complex that traditional data processing applications are inadequate."

- "Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges."

- "Big Data refers to data sets whose size is beyond the ability of typical database software tool to capture, store, manage and analyze.