National Institute of Electronics and Information Technology
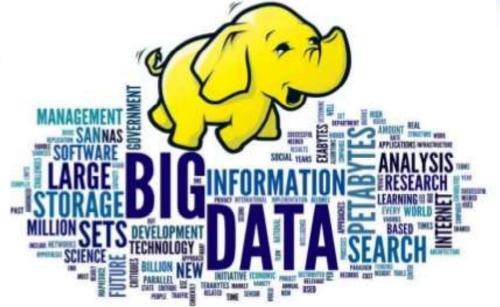
Big Data and Hadoop
Module 1-Introduction to BigData

What is BIG DATA?

"Information is the oil of the 21st century & analytics is the combustion engine"

– *Peter Sondergaard, Senior Vice President, Gartner.*

1. Today's business enterprises are data-driven and without data no enterprise can have a competitive advantage.

2. ***Data is the new currency and oil of our generation.***

3. "Big Data" is the data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

4. Big data are extremely large data sets that may be analyzed computationally to reveal pattern, trends and associations especially related to human behavior and interactions

5. With each passing day, Big data is growing bigger, is more difficult to make sense of, is being generated at a much faster rate and this trend is only going to intensify in our data-driven digital world.

6. Today Big Data is so rampant that one has to look which are the companies that are not deploying Big Data.

1. **Facebook :-** Arguably the world's most popular social media network with more than two billion monthly active users worldwide. Every day, we feed Facebook's data beast with mounds of information.

   ➢ Every 60 seconds, 136,000 photos are uploaded,

   ➢ 510,000 comments are posted

   ➢ 293,000 status updates are posted.

   With data like this, Facebook knows who our friends are, what we look like, where we are, what we are doing, our likes, our dislikes, and so much more. Some researchers even say Facebook has enough data to know us better than our therapists!

2. **Twitter** :- is a gold mine of data. Unlike other social platforms, almost every user's tweets are completely public and pullable.

   ➢ There are 6,000 tweets per second

   ➢ This equates to 500 million tweets every single day

   ➢ Handling the huge quantities of data takes skill

   ➢ Terabytes worth of data uploaded to the servers every day!

3. **Instagram :-** Instagram, the social networking app for sharing photos and videos, launched in 2010. Today, it boasts 800 million monthly active users and is owned by Facebook. People interact by showing their love with a heart, commenting and using hashtags. What all of this activity does is create an enormous amount of data..

   ➢ Every 60 seconds, 136,000 photos are uploaded,

   ➢ 510,000 comments are posted

   ➢ 293,000 status updates are posted.

4. In essence starting from technology companies like Google, Apple, Amazon, Microsoft all the way to mining companies like Rio Tinto, retailers like Walmart and hospitality companies like Airbnb are all using big data.

   ➢ **Amazon** – Getting insights on customer data & providing better user experience

   ➢ **Google** – making sense of what the customer is searching for and providing better search results

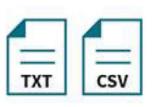   ➢ **Walmart** – providing customers what they are looking for in terms of products, discounts, etc.

1. **Structured Data :-** This is the type of data that is stored in the regular databases in terms of the rows and columns giving it a definite structure. Most of the data that is part of the structured format includes the company employee details, census records, economic data and so on.

2. **Unstructured Data :-** This is the type of data that can be put into a regular row and column based format. Today over 80% of the data is unstructured data and it will only increase due to the huge amounts of sensor and machine-generated data that we will be seeing as part of IoT.

3. **Semi-Structured Data :-** This is the type of data that can be converted into structured data through the addition of certain keys, attributes or other characteristics through which the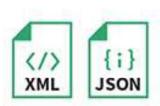y can be arranged or sorted in a database.