BIG DATA

Name : Vineeta Verma
Roll No. : 21428BIF031
Branch : Bioinformatics

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

What is Big Data?

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Sources of Big Data

1. **SOCIAL NETWORKING SITES**
   Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.

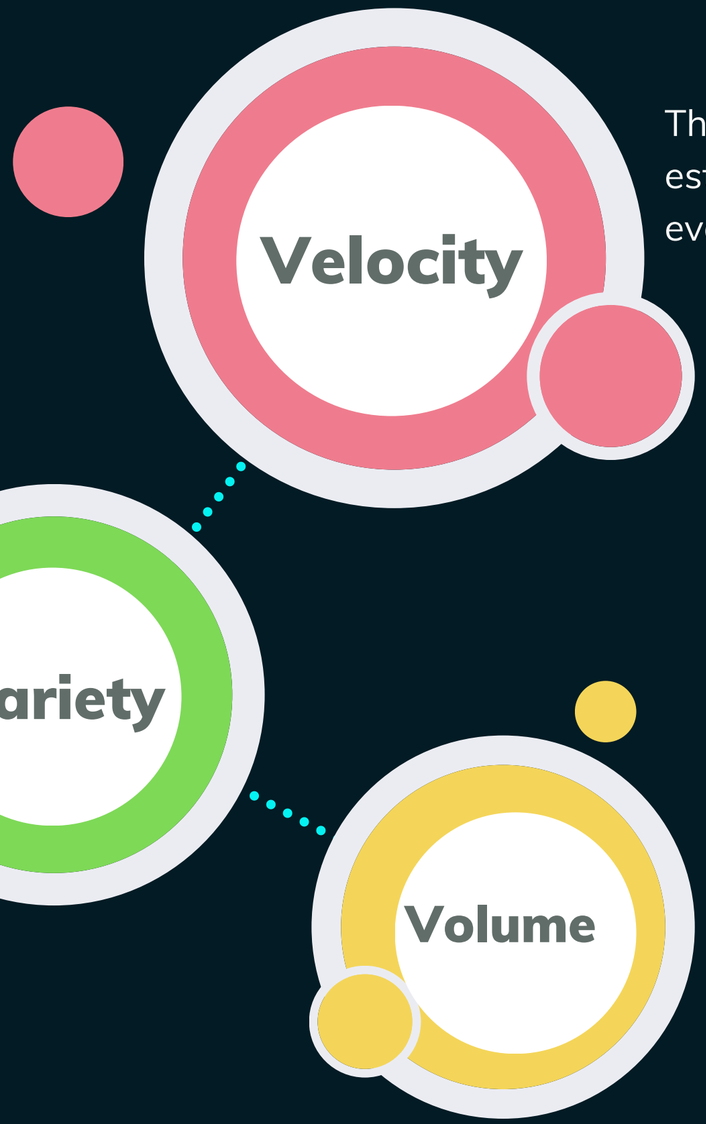
2. **E-COMMERCE SITE**
   Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.

3. **WEATHER STATION**
   All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.

4. **TELECOM COMPANY**
   Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.

5. **SHARE MARKET**
   Stock exchange across the world generates huge amount of data through its daily transaction.

3V's of Big Data



The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.

Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.

The amount of data which we deal with is of very large size of Peta bytes.

Why is big data important?

Data can be a company’s most valuable asset. Using big data to reveal insights can help understand the areas that affect your business from market conditions and customer purchasing behaviors to your business processes.

How big data works

Big data gives you new insights that open up new opportunities and business models. Getting started involves three key actions:

1. Integrate

Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as extract, transform, and load (ETL) generally aren’t up to the task. It requires new strategies and technologies to analyze big data sets at terabyte, or even petabyte, scale.

During integration, you need to bring in the data, process it, and make sure it’s formatted and available in a form that your business analysts can get started with.

2. Manage

Big data requires storage. Your storage solution can be in the cloud, on premises, or both. You can store your data in any form you want and bring your desired processing requirements and necessary process engines to those data sets on an on-demand basis. Many people choose their storage solution according to where their data is currently residing. The cloud is gradually gaining popularity because it supports your current compute requirements and enables you to spin up resources as needed.

3. Analyze

Your investment in big data pays off when you analyze and act on your data. Get new clarity with a visual analysis of your varied data sets. Explore the data further to make new discoveries. Share your findings with others. Build data models with machine learning and artificial intelligence.

Advantages of Big Data :

- Errors inside the business are known immediately.
- Higher conversion rate and additional income.
- The plan of action of your opposition is seen promptly.
- Extortion can be recognized the minute it happens and legitimate measures can be taken to restrict the harm.
- The principal points of interest of Big Data include the increased speed, capacity, and scalability of storage and having the measures and tools to deal with the data all the more proficiently.

Disadvantages of Big Data :

- Data is collected from every source possible over a certain course of time. The data collected is raw, inconsistent and therefore subjected to more noise.
- Security is one of the key issues that Big Data is still struggling with, especially on the social media front.
- Most of the data a user is looking for analysis and interpretation purposes is hidden behind firewalls and private cloud that can only be accessed by having the technical knowledge and expertise to turn the raw data into relevant information.

Despite having rigorous knowledge of the benefits and pitfalls of Big Data, there are various firms and enterprises keen on taking the challenge of creating meaningful data from this nerve wracking amount of data. However, knowledge and expertise on upcoming tools and technologies don't seem enough to cater to the needs of the end user to give data some meaning. Here are a few reasons as to why the Big Data projects fail on such a large scale.

How to big data impacts on ibusiness processes :

Modern enterprises deploy a variety of specialized internal systems to track operational activity and collect useable information. This approach, a result of the big data revolution, complicates already complex business processes, most notably those that fall into the information systems realm.

What is big data analytics?

- Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. These processes use familiar statistical analysis techniques like clustering and regression and apply them to more extensive datasets with the help of newer tools. Big data has been a buzz word since the early 2000s, when software and hardware capabilities made it possible for organizations to handle large amounts of unstructured data.

- Since then, new technologies from Amazon to smartphones have contributed even more to the substantial amounts of data available to organizations. With the explosion of data, early innovation projects like Hadoop, Spark, and NoSQL databases were created for the storage and processing of big data.

- This field continues to evolve as data engineers look for ways to integrate the vast amounts of complex information created by sensors, networks, transactions, smart devices, web usage, and more. Even now, big data analytics methods are being used with emerging technologies, like machine learning, to discover and scale more complex insights.

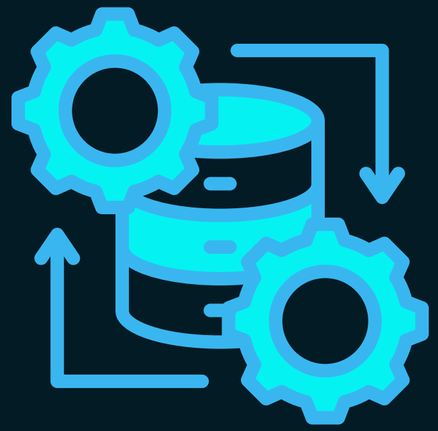
How does big data analytics work?

- **Collect :** The data, which comes in structured, semi-structured, and unstructured forms, is collected from multiple sources across web, mobile, and the cloud. It is then stored in a repository a data lake or data warehouse in preparation to be processed.

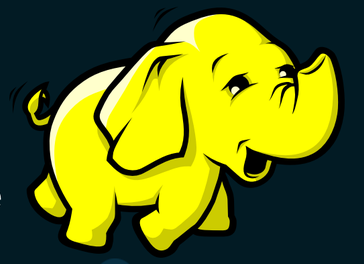
- **Process :** During the processing phase, the stored data is verified, sorted, and filtered, which prepares it for further use and improves the performance of queries.

- **Scrub :** After processing, the data is then scrubbed. Conflicts, redundancies, invalid or incomplete fields, and formatting errors within the data set are corrected and cleaned.

- **Analyze :** The data is now ready to be analyzed. Analyzing big data is accomplished through tools and technologies such as data mining, AI, predictive analytics, machine learning, and statistical analysis, which help define and predict patterns and behaviors in the data.

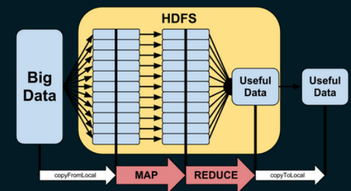


Big data analytics tools and technology



- **Hadoop** is an open-source framework that efficiently stores and processes big datasets on clusters of commodity hardware. This framework is free and can handle large amounts of structured and unstructured data, making it a valuable mainstay for any big data operation.

- **NoSQL databases** are non-relational data management systems that do not require a fixed scheme, making them a great option for big, raw, unstructured data. NoSQL stands for “not only SQL,” and these databases can handle a variety of data models.



- **MapReduce** is an essential component to the Hadoop framework serving two functions. The first is mapping, which filters data to various nodes within the cluster. The second is reducing, which organizes and reduces the results from each node to answer a query.

- **YARN** stands for “Yet Another Resource Negotiator.” It is another component of second-generation Hadoop. The cluster management technology helps with job scheduling and resource management in the cluster.

- **Spark** is an open source cluster computing framework that uses implicit data parallelism and fault tolerance to provide an interface for programming entire clusters. Spark can handle both batch and stream processing for fast computation.

- **Tableau** is an end-to-end data analytics platform that allows you to prep, analyze, collaborate, and share your big data insights. Tableau excels in self-service visual analysis, allowing people to ask new questions of governed big data and easily share those insights across the organization.

References:

https://cloud.google.com/learn/what-is-big-data?hl=en
https://assignmentpoint.com/define-business-advantages-using-big-data-analytics/
https://www.simplilearn.com/tutorials/big-data-tutorial/what-is-big-data
https://www.guru99.com/what-is-big-data.html
https://www.javatpoint.com/what-is-big-data
https://www.oracle.com/big-data/what-is-big-data/
https://www.geeksforgeeks.org/world-big-data/?ref=lbp
https://www.tableau.com/learn/articles/big-data-analytics
https://azure.microsoft.com/en-in/resources/cloud-computing-dictionary/what-is-big-data-analytics/#layout-container-uidb190


Thank You!