MODULE 1 :

INTRODUCTION TO BIGDATA PLATFORM

For you to understand Big Data, it is important that you first understand what data is.

“Data can be defined as figures or facts that can be stored in or can be used by a computer.”

BIG DATA is a term that is used for denoting a collection of datasets that is large and complex, making it very difficult to process using legacy data processing applications

“Data which are very large in size is called Big Data. “

Sources of Big Data :

These data come from many sources like

- 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.
- E-commerce site: Sites like Amazon, Flipkart generates huge amount of logs from which users buying trends can be traced.
- Weather Station: All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- Share Market: Stock exchange across the world generates huge amount of data through its daily transaction.

5 ‘V’ S OF BIGDATA:

‘V’ s are also termed as the characteristics of Big Data.

Big data is a collection of data from many different sources and is often describe by five characteristics: **volume, value, variety, velocity, and veracity**.

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

   - Velocity refers to the high speed of accumulation of data.
   - In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
   - There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.

2.Variety:

Nowadays data are not stored in rows and columns. Data is structured as well as unstructured. Data which can be saved in tables are structured data like the transaction data of the bank.

- It refers to VARIETY of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

- **Structured data**: This data is basically an organized data. It generally refers to data that has defined the length and format of data.

- **Semi- Structured data**: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.

- **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can't be stored in the form of rows and columns.

3. Volume:

- The name 'Big Data' itself is related to a size which is enormous.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a '**Big Data'**. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.

4. Veracity:

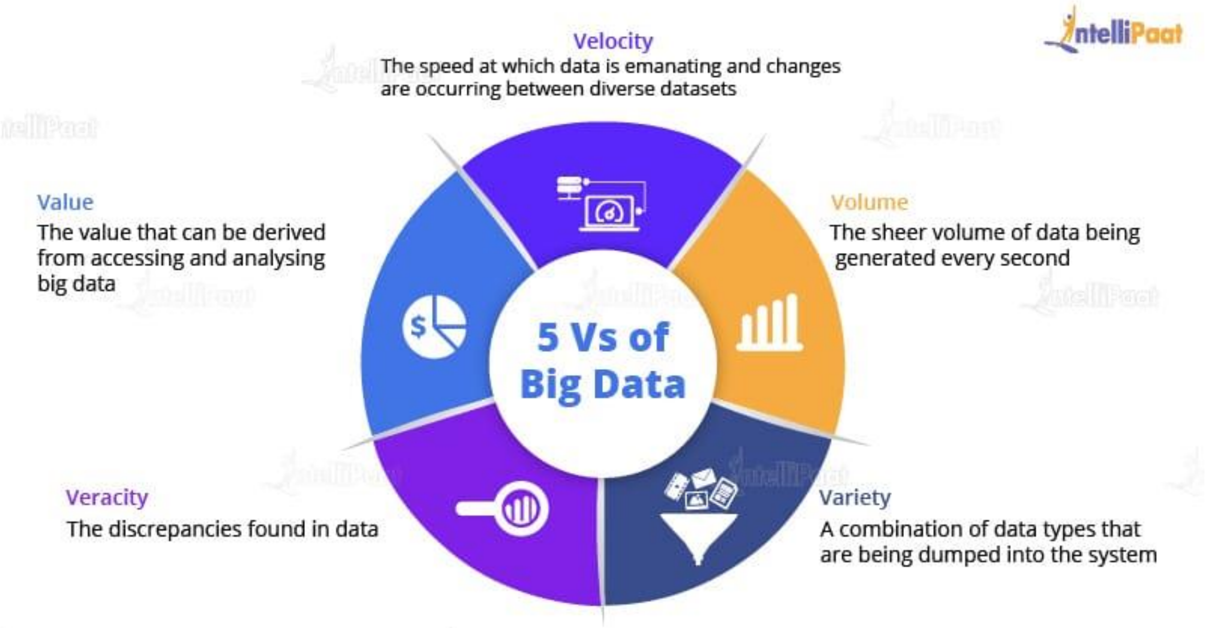
- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

5. Value:

- After having the 4 V's into account there comes one more V which stands for **Value!.** The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V's.

Characteristics of Big Data

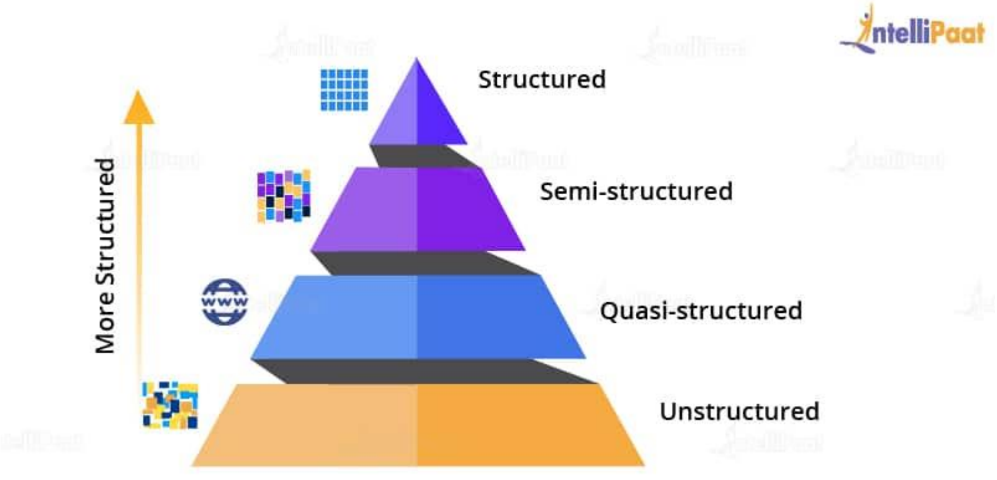
Big Data has the following distinct characteristics:



IntelliPaat
Velocity
The speed at which data is emanating and changes are occurring between diverse datasets
Value
The value that can be derived from accessing and analysing big data
Volume
The sheer volume of data being generated every second
5 Vs of Big Data
Veracity
The discrepancies found in data
Variety
A combination of data types that are being dumped into the system

1. Volume: This refers to tremendously large data. As you can see from the image, the volume of data is rising exponentially. In 2016, the data created was only 8 ZB; it is expected that, by 2020, the data would rise to 40 ZB, which is extremely large.

2. Variety: A reason for this rapid growth of data volume is that data is coming from different sources in various formats.



a) **Structured Data:** Here, data is present in a structured schema along with all the required columns. It is in a structured or tabular format. Data that is stored in a relational database management system is an example of structured data. For example, in the below-given employee table, which is present in a database, the data is in a structured format.

Emp. ID	Emp. Name	Gender	Department	Salary (INR)
2385	ABC	Male	Finance	650,000
4625	XYZ	Male	Admin	5,000,000

b) **Semi-structured Data:** In this form of data, the schema is not properly defined, i.e., both forms of data are present. So, semi-structured data has a structured form but it is not defined; for example, XML, CSV and email. The web application data that is unstructured contains transaction history files, log files, etc.

c) **Unstructured Data:** This data format includes all unstructured files such as video files, log files, audio files, and image files. Any data that has an unfamiliar model or structure is categorized as unstructured data. Since its size is large, unstructured data possesses various challenges in terms of processing for deriving value out of it. An example of this is a complex data source that contains a blend of text files, videos, and images. Several organizations have a lot of data available with them but they don't know how to derive value out of it since the data is in its raw form.

3. Velocity: The speed of data accumulation also plays a role in determining whether the data is big data or normal data.

4. Value it deals with a mechanism to bring out the correct meaning of data. First of all, you need to mine data, i.e., the process to turn raw data into useful data. Then, an analysis is done on the data that you have cleaned or retrieved from the raw data. Then, you need to make sure whatever analysis you have done benefits your business, such as in finding out insights, results, etc., in a way that was not possible earlier.

5. Veracity: Since packages get lost during execution, we need to start again from the stage of mining raw data to convert it into valuable data. And this process goes on. There will also be uncertainties and inconsistencies in the data that can be overcome by veracity. Veracity means the trustworthiness and quality of data.

Challenges of bigdata :

The challenges in BIG DATA are the real implementation . These require immediate attention and need to be handled because if not handled then the failure of the technology may take place which can also lead to some unpleasant result. Big data challenges include **storing, and analyzing** extremely large and fast-growing data.

Some of the Big Data challenges are:

1. **Sharing and Accessing Data:**
    - most frequent challenge in big data efforts is the inaccessibility of data sets from external sources.
    - Sharing data can cause substantial challenges.
    - Accessing data from public repositories leads to multiple difficulties.
    - It is necessary for the data to be available in an accurate, complete and timely manner because if data in the companies information system is to be used to make accurate decisions in time then it becomes necessary for data to be available in this manner.

2. **Privacy and Security:**
    - It is another most important challenge with Big Data. This challenge includes sensitive significance.
    - Most of the organizations are unable to maintain regular checks due to large amounts of data generation. However, it should be necessary to perform security checks and observation in real time because it is most beneficial.
    - There is some information of a person which when combined with external large data may lead to some facts of a person which may be secretive and he might not want the owner to know this information about that person.
    - Some of the organization collects information of the people in order to add value to their business. This is done by making insights into their lives that they’re unaware of.

3. **Analytical Challenges:**
    - There are some huge analytical challenges in big data which arise some main challenges questions like how to deal with a problem if data volume gets too large?
    - Or how to find out the important data points?
    - Or how to use data to the best advantage?

- These large amount of data on which these type of analysis is to be done can be structured (organized data), semi-structured (Semi-organized data) or unstructured (unorganized data).

4. **Technical challenges:**

- **Quality of data:**
    - When there is a collection of a large amount of data and storage of this data, it comes at a cost. Big companies, business leaders and IT leaders always want large data storage.
    - For better results and conclusions, Big data rather than having irrelevant data, focuses on quality data storage.
    - This further arise a question that how it can be ensured that data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not.
- **Fault tolerance:**
    - Fault tolerance is another technical challenge and fault tolerance computing is extremely hard, involving intricate algorithms.
    - Nowadays some of the new technologies like cloud computing and big data always intended that whenever the failure occurs the damage done should be within the acceptable threshold that is the whole task should not begin from the scratch.
- **Scalability:**
    - Big data projects can grow and evolve rapidly.
    - It leads to various challenges like how to run and execute various jobs so that goal of each workload can be achieved cost-effectively.
    - It must be possible to expand according to the user requirements.

5.Data Growth

One of the most pressing Big Data challenges is storage. Data is growing exponentially with time, and with that, enterprises are struggling to store large amounts of data. Much of this data is extracted from images, audio, documents, text files, etc., that are unstructured and not in databases. It is difficult to extract and analyze all unstructured data. These issues are a part of Big Data infrastructure challenges.

6. Data Security

Security can be one of the most Big Data challenges especially for organizations that have sensitive company data or have access to a lot of personal user information. Vulnerable data is an attractive target for cyberattacks and malicious hackers.

7. Increasing Salaries of Skilled Big Data Professionals

Big Data salaries have increased significantly. According to the 2017 Robert Half Technology Salary Guide, big data engineers have salaries between US$135,000 and US$196,000 on average. while data scientists earn around US$116,000 to US$163, 500.

Nature of data

Big data is a combination of structured, semi-structured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling, decision-making, and other advanced analytics applications.

Big Data can be structured, unstructured, and semi-structured that are being collected from different sources.

Types of Big Data

Big Data is essentially classified into three types:

- Structured Data
- Unstructured Data
- Semi-structured Data

- **Structured Data**

Structured data is highly organized and thus, is the easiest to work with. Its dimensions are defined by set parameters. Every piece of information is grouped into rows and columns like spreadsheets. Structured data has quantitative data such as age, contact, address, billing, expenses, debit or credit card numbers, etc.

Due to structured data’s quantitative nature, it is easy for programs to sort through and collect data. It requires little to no preparation to process structured data.

- **Unstructured Data**

Not all data is structured and well-sorted with instructions on how to use it. All unorganized data is known as unstructured data.

Almost everything generated by a computer is unstructured data. The time and effort required to make unstructured data readable is more..

The challenging part about unstructured data analysis is teaching an application to understand the information it’s extracting. Oftentimes, translation into structured form is required, which is not easy and varies with different formats and end goals

- **Semi-structured Data**

Semi-structured data falls somewhere between structured data and unstructured data. It mostly translates to unstructured data that has metadata attached to it. Semi-structured data can be inherited such as location, time, email address, or device ID stamp. It can even be a semantic tag attached to the data later.

Major Sectors Using Big Data Every Day

The applications of big data provided solutions to every sector like Banking, Government, Education, and healthcare, etc.

Banking

Since there is a massive amount of data that is coming from innumerable sources, banks need to find uncommon and unconventional ways to manage big data. It’s also essential to examine customer requirements, render services according to their specifications, and reduce risks while sustaining regulatory compliance. Financial institutions have to deal with Big Data Analytics to solve this problem.

Government

Government agencies utilize Big Data and have devised a lot of running agencies, managing utilities, dealing with traffic jams, or limiting the effects of crime. However, apart from its benefits in Big Data, the government also addresses the concerns of transparency and privacy.

- Aadhar Card: The Indian government has a record of all 1.21 billion citizens. This huge data is stored and analyzed to find out several things, such as the number of youth in the country. According to which several schemes are made to target the maximum population. All this big data can’t be stored in some traditional database, so it is left for storing and analyzing using big data analytic tools.

Education

Education concerning Big Data produces a vital impact on students, school systems, and curriculums. By interpreting big data, people can ensure students' growth, identify at-risk students, and achieve an improvised system for the evaluation and assistance of principals and teachers.

- Example: The education sector holds a lot of information concerning curriculum, students, and faculty. The information is analyzed to get insights that can enhance the operational adequacy of the educational organization. Collecting and analyzing information about a student such as attendance, test scores, grades, and other issues take up a lot of data. So, big data approaches a progressive framework wherein this data can be stored and analyzed making it easier for the institutes to work with.

Big Data in Healthcare

When it comes to what Big Data is in Healthcare, we can see that it is being used enormously. It includes collecting data, analyzing it for customers. Also, patients' clinical data is too complex to be solved or understood by traditional systems. Since big data is processed by Machine Learning algorithms and Data Scientists, tackling such huge data becomes manageable.

- Example: Nowadays, doctors rely mostly on patients' clinical records, which means that a lot of data needs to be gathered, that too for different patients. It is not possible for old or traditional data storage methods to store this data. Since there is a large amount of data coming from different sources, in various formats, the need to handle this large amount of data is increased, and that is why the Big Data approach is needed.

E-commerce

Maintaining customer relationships is the most important in the e-commerce industry. E-commerce websites have different marketing ideas to retail their merchandise to their customers, manage transactions, and implement better tactics of using innovative ideas with Big Data to improve businesses.

- Flipkart: Flipkart is a huge e-commerce website dealing with lots of traffic daily. But, when there is a pre-announced sale on Flipkart, traffic grows exponentially that crashes the website. So, to handle this kind of traffic and data, Flipkart uses Big Data. Big Data can help in organizing and analyzing the data for further use.

Social Media

Social media in the current scenario is considered the largest data generator. The stats have shown that around 500+ terabytes of new data get generated into the databases of social media every day, particularly in the case of Facebook. The data generated mainly consist of videos, photos, message exchanges, etc. A single activity on any social media site generates a lot of data which is again stored and gets processed whenever required. Since the data stored is in terabytes, it would take a lot of time for processing if it is done by our legacy systems. Big Data is a solution to this problem.

Intelligent Data Analysis (IDA)

Intelligent Data Analysis (IDA) is an interdisciplinary study that is concerned with the extraction of useful knowledge from data, drawing techniques from a variety of fields, such as artificial intelligence, high-performance computing, pattern recognition, and statistics.

It is the discovery and communication of meaningful patterns in data. Especially, valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming, and operation research to qualify performance. Analytics often favors data visualization to communicate insight.
Firms may commonly apply analytics to business data, to describe, predict, and improve business performance. Especially, areas within include predictive analytics, enterprise decision management, etc.

Analytics is the scientific process of transforming data into insight for making better decisions. The goal of Data Analytics is to get actionable insights resulting in smarter decisions and better business outcomes.

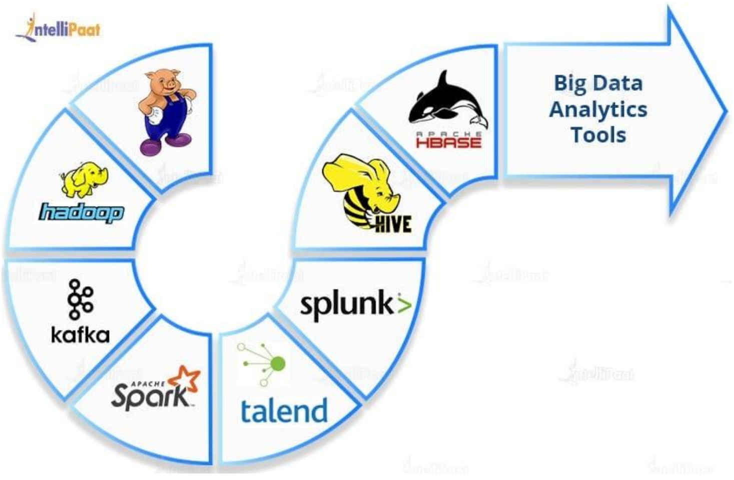
What is Big Data Analytics?

Big data analytics examines large and different types of data to uncover hidden patterns, insights, and correlations. Big Data Analytics is helping large companies facilitate their growth and development. And it majorly includes applying various data mining algorithms on a certain dataset.

Data Analysis :

- FOR DECISION MAKING
- FOR MAKING PREDICTIONS FOR FUTURE

Tools for Big Data Analytics



- **Apache Hadoop**

  Big Data Hadoop is a framework that allows you to store big data in a distributed environment for parallel processing.

- **Apache Pig**

  Apache Pig is a platform that is used for analyzing large datasets by representing them as data flows. Pig is designed to provide an abstraction over MapReduce which reduces the complexities of writing a MapReduce program.

- **Apache HBase**

  Apache HBase is a multidimensional, distributed, open-source, and NoSQL database written in Java. It runs on top of HDFS providing Bigtable-like capabilities for Hadoop.

- **Apache Spark**

  Apache Spark is an open-source general-purpose cluster-computing framework. It provides an interface for programming all clusters with implicit data parallelism and fault tolerance.

- **Talend**

  Talend is an open-source data integration platform. It provides many services for enterprise application integration, data integration, data management, cloud storage, data quality, and Big Data.

- **Splunk**

  Splunk is an American company that produces software for monitoring, searching, and analyzing machine-generated data using a Web-style interface.

- **Apache Hive**

  Apache Hive is a data warehouse system developed on top of Hadoop and is used for interpreting structured and semi-structured data.

- **Kafka**

  Apache Kafka is a distributed messaging system that was initially developed at LinkedIn and later became part of the Apache project. Kafka is agile, fast, scalable, and distributed by design.

DIFFERENCE

On the basis of	Structured data	Unstructured data
Technology	It is based on a relational database.	It is based on character and binary data.
Flexibility	Structured data is less flexible and schema-dependent.	There is an absence of schema, so it is more flexible.
Scalability	It is hard to scale database schema.	It is more scalable.
Robustness	It is very robust.	It is less robust.
Performance	Here, we can perform a structured query that allows complex joining, so the performance is higher.	While in unstructured data, textual queries are possible, the performance is lower than semi-structured and structured data.
Nature	Structured data is quantitative, i.e., it consists of hard numbers or things that can be counted.	It is qualitative, as it cannot be processed and analyzed using conventional tools.
Format	It has a predefined format.	It has a variety of formats, i.e., it comes in a variety of shapes and sizes.
Analysis	It is easy to search.	Searching for unstructured data is more difficult.

Data Analytics

Data analytics converts raw data into actionable insights. It includes a range of tools, technologies, and processes used to find trends and solve problems by using data. Data analytics can ***shape business processes, improve decision-making, and foster business growth.***

There are four types of data analytics:

1. Predictive (forecasting)
2. Descriptive (business intelligence and data mining)
3. Prescriptive (optimization and simulation)
4. Diagnostic analytics

Descriptive Analytics: (WHAT IS HAPPENED & HAPPENING)

Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.

Diagnostic Analytics: (WHY THIS HAPPENED ?)

In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.

Predictive Analytics: (WHAT IS GOING TO HAPPEN)

Predictive analytics turn the data into valuable, actionable information. predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring.
Predictive analytics holds a variety of statistical techniques from modeling, machine, learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event.

Prescriptive Analytics: (WHAT SHOULD BE DONE ?)

Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction.
Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefit from the predictions and showing the decision maker the implication of each decision option.
Prescriptive Analytics not only anticipates what will happen and when to happen but also why it will happen.

Prescriptive Analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option.

.

MODERN DATA ANALYTIC TOOLS :

There are hundreds of *data analytics tools* out there in the market today but the selection of the right tool will depend upon your business NEED, GOALS, and VARIETY to get business in the right direction.

1. APACHE Hadoop

It’s a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel. It can process both structured and unstructured data from one server to multiple computers.
Today, it is the best *big data analytic tool* and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

Features of Apache Hadoop:

- Free to use and offers an efficient storage solution for businesses.
- Offers quick access via HDFS (Hadoop Distributed File System).
- Highly flexible and can be easily implemented with MySQL, and JSON.
- Highly scalable as it can distribute a large amount of data in small segments.

2. Cassandra

It is an open-source NoSQL distributed database that is used to fetch large amounts of data It’s one of the most popular tools for data analytics and has been praised by many tech companies due to its high scalability and availability without compromising speed and performance.
 It is capable of delivering thousands of operations every second and can handle petabytes of resources with almost zero downtime.

Features of APACHE Cassandra:

- Data Storage Flexibility*:* It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
- Data Distribution System*:* Easy to distribute data with the help of replicating data on multiple data centers.
- Fast Processing*:* Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.
- Fault-tolerance*:* The moment, if any node fails, it will be replaced without any delay.

3. Xplenty

It is a data analytic tool for building a data pipeline by using minimal codes in it. It offers a wide range of solutions for sales, marketing, and support. The best part of using Xplenty is its low investment in hardware & software and its offers support via email, chat, telephonic and virtual meetings. Xplenty is a platform to process data for analytics over the cloud and segregates all the data together.

Features of Xplenty:

- Flexibility*:* Data can be sent, and pulled to databases, warehouses, and salesforce.
- Data Security*:* It offers SSL/TSL encryption and the platform is capable of verifying algorithms and certificates regularly.

4. Spark

It is another framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is capable of handling multi-petabytes of data as well. Recently, Spark made a record of processing 100 terabytes of data in just 23 minutes which broke the previous world record of Hadoop (71 minutes). This is the reason why big tech giants are moving towards spark.

Features of APACHE Spark:

- Ease of use*:* It allows users to run in their preferred language. (JAVA, Python, etc.)
- Real-time Processing: Spark can handle real-time streaming via Spark Streaming

5. Mongo DB

It is a free, open-source platform and a document-oriented (NoSQL) database that is used to store a high volume of data. It uses collections and documents for storage and its document consists of key-value pairs which are considered a basic unit of Mongo DB. It is so popular among developers due to its availability for multi-programming languages such as Python, Jscript, and Ruby.

Features of Mongo DB:

- Written in C++: It’s a schema-less DB and can hold varieties of documents inside.
- Simplifies Stack*:* With the help of mongo, a user can easily store files without any disturbance in the stack.
- Master-Slave Replication: It can write/read data from the master and can be called back for backup.

6. Apache Storm

A storm is a robust, user-friendly tool used for data analytics, especially in small companies. The best part about the storm is that it has no language barrier (programming) in it and can support any of them. It was designed to handle a pool of large data in fault-tolerance and horizontally scalable methods. When we talk about real-time data processing, Storm leads the chart because of its distributed real-time big data processing system, due to which today many tech giants are using APACHE Storm in their systems.

Features of Storm:

- Data Processing*:* Storm process the data even if the node gets disconnected
- Highly Scalable: It keeps the momentum of performance even if the load increases
- Fast: The speed of APACHE Storm is impeccable and can process up to 1 million messages of 100 bytes on a single node.
- *Automation:* To cut down the manual chase, datapine offers a wide array of AI assistant and BI tools.
- *Predictive Tool:* datapine provides forecasting/predictive analytics by using historical and current data, it derives the future outcome.
- *Add on:* It also offers intuitive widgets, visual analytics & discovery, ad hoc reporting, etc.

7. Rapid Miner

It’s a fully automated visual workflow design tool used for data analytics. It’s a no-code platform and users aren’t required to code for segregating data. Today, it is being heavily used in many industries such as training, research, etc. Though it’s an open-source platform but has a limitation of adding 10000 data rows and a single logical processor. With the help of Rapid Miner, one can easily deploy their models to the web or mobile.

Features of Rapid Miner:

- Accessibility*:* It allows users to access 40+ types of files
- Storage*:* Users can access cloud storage facilities such as AWS and dropbox
- Data validation*:* Rapid miner enables the visual display of multiple results in history for better evaluation.

ANALYSIS / VS REPORTING

What is the difference between analysis and reporting?

In reality, reporting is the sorting and organization of data, while analytics derive insights from that data and often influence business decisions. Three key differences to take note of between reporting and analytics are purpose, methods, and value.

Data reporting is the process of collecting and formatting raw data and translating it into a digestible format to assess the ongoing performance of your organization. Your data reports can answer basic questions about the state of your business.

Data reporting is the process of collecting and submitting data which gives rise to accurate analyses of the facts on the ground; inaccurate data reporting can lead to vastly uninformed decision-making based on erroneous evidence.

Data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions.

Prediction error

Prediction error quantifies one of two things:

- “it's a measure of how well the model predicts the response variable”
- it's a measure of how well samples are classified to the correct category.

Prediction Issues:

Preparing the data for prediction is the most pressing challenge. The following activities are involved in data preparation:

- **Data Cleaning:** Cleaning data include reducing noise and treating missing values. Smoothing techniques remove noise, and the problem of missing values is solved by replacing a missing value with the most often occurring value for that characteristic.

- **Relevance Analysis:** The irrelevant attributes may also be present in the database. The correlation analysis method is used to determine whether two attributes are connected.
- **Data Transformation and Reduction:** Any of the methods listed below can be used to transform the data.
  - Normalization: Normalization is used to transform the data. Normalization is the process of scaling all values for a given attribute so that they lie within a narrow range. When neural networks or methods requiring measurements are utilized in the learning process, normalization is performed.
  - Generalization: The data can also be modified by applying a higher idea to it. We can use the concept of hierarchies for this.

Other data reduction techniques histogram analysis, and clustering.

Statistical concepts

• Statistics is a branch of applied or business mathematics where we collect, organize, analyse and interpret numerical facts. Statistical methods are the concepts, models, and formulas of mathematics used in the statistical analysis of data.

• They can be subdivided into two main categories - Descriptive Statistics and Inferential Statistics.

• Descriptive statistics further consists of measure of central tendency and measure of dispersion and inferential statistics consists of estimation and hypothesis testing.

1. Descriptive statistics

Descriptive statistics methods involve summarizing or describing the sample of data in various forms to get an overall gist of the data.

2. Inferential Statistics

In contrast, inferential statistics try to make assumptions about the population of the data, given the sample; or in predicting various outcomes.

RANDOM EXPERIMENTS

• Random experiment is the process to observe the event having an uncertain outcome.

• When we toss a coin the outcome is uncertain and hence, it can be termed as a random experiment.

• The result of a random experiment is known as the outcome and the set of all the possible outcomes of an experiment is known as sample space.

- If we repeat an experiment n number of times, then each time the experiment is done is known as a trial.

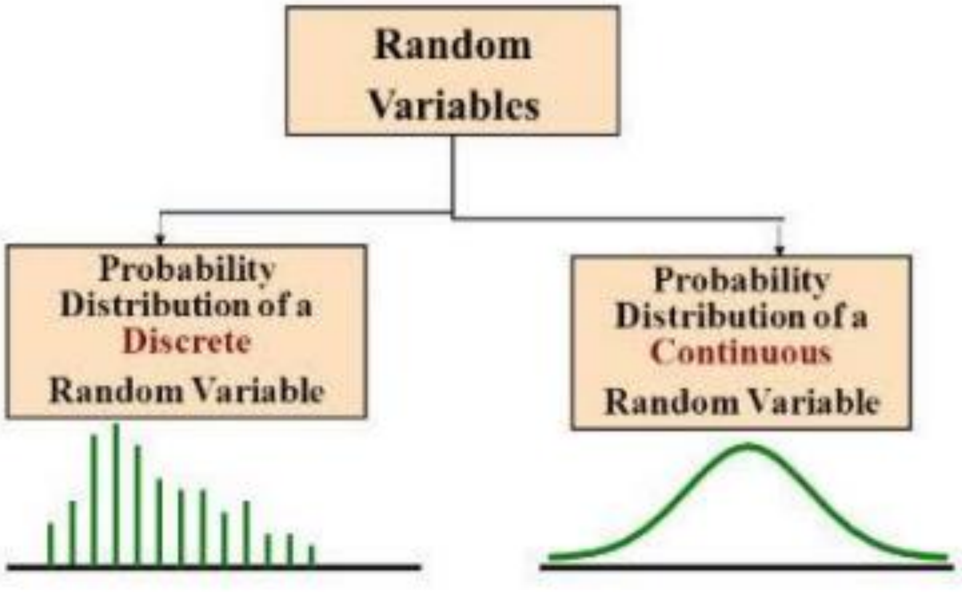
RANDOM VARIABLES

- A random variable is a variable where value is unknown or a function that assigns values to every of an experiment's outcomes.
- They are often designated by letters.
- Random variables can be classified as discrete which are variables that have specific values and continuous which are variables that can have any values within a continuous range.

. • Whereas a random variable has a set of values, and any of those values can be the resulting outcome.

- Example: tossing a coin or dice.

Types of random variables


Random Variables
Probability Distribution of a Discrete Random Variable
Probability Distribution of a Continuous Random Variable

1.Discrete Random Variable

- As the name Suggest, Discrete random variables consist of distinct or discrete unique values. It takes a countable number of distinct values. Now, Consider an experiment where a coin is tossed five times.

- Discrete Random Variable Example: ♦ Tossing a Coin

  Here the number of outcomes that can occur is either a Head or a Tail. Hence we can denote Head, Tail as Random variables as they are distinct in nature.

2. Continuous Random Variable

An example of a continuous random variable can be an experiment that involves measuring the amount of rainfall in a city over a year or the average height of a random group of 25 people.

- Continuous Random Variable Example :►Heights of people playing Basketball.

Here Height can be any value between 4 feet’s to 7 feet’s respectively.

POPULATION AND SAMPLING

Population

- A population is an entire collection of objects or observations from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.
- For each population, there are many possible samples. It is important that the investigator carefully and completely defines the population before collecting the sample, including a description of the members to be included.

Sample

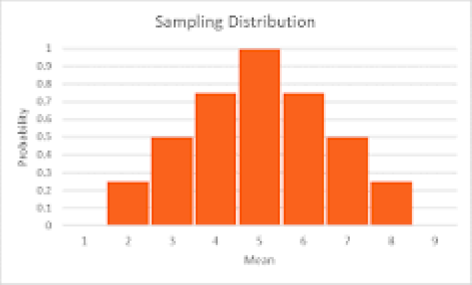
- A sample is a group of units selected from a larger group (the population).
- By studying the sample, it is hoped to draw valid conclusions about the larger group.
- A sample is generally selected for study because the population is too large to study in its entirety. The sample should be representative of the general population. This is often best achieved by random sampling

SAMPLING DISTRIBUTION

- A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population.

- The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

- A lot of data drawn and used by academicians, statisticians, researchers, marketers, and analysts are actually samples, not populations



A sampling distribution refers to a probability distribution of a statistic that comes from choosing random samples of a given population. Also known as a finite-sample distribution, it represents the distribution of frequencies on how spread apart various outcomes will be for a specific population

As a group, sampling methods fall into one of two categories.

- ▪ Probability samples. With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample.

- ▪ Non-probability samples. With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen.

. RE-SAMPLING

- Resampling is the method that consists of drawing repeated samples from the original data samples. The method of Resampling is a nonparametric method of statistical inference. In other words, the method of resampling does not involve the utilization of the generic distribution tables (for example, normal distribution tables) in order to compute approximate p probability values.

- Resampling involves the selection of randomized cases with replacement from the original data sample in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample. Due to replacement, the drawn number of samples that are used by the method of resampling consists of repetitive cases.
- Resampling is also known as Bootstrapping or Monte Carlo Estimation.

STATISTICAL INFERENCE

Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to assess the relationship between the dependent and independent variables. The purpose of statistical inference to estimate the uncertainty or sample to sample variation.