INTRODUCTION TO BIG DATA

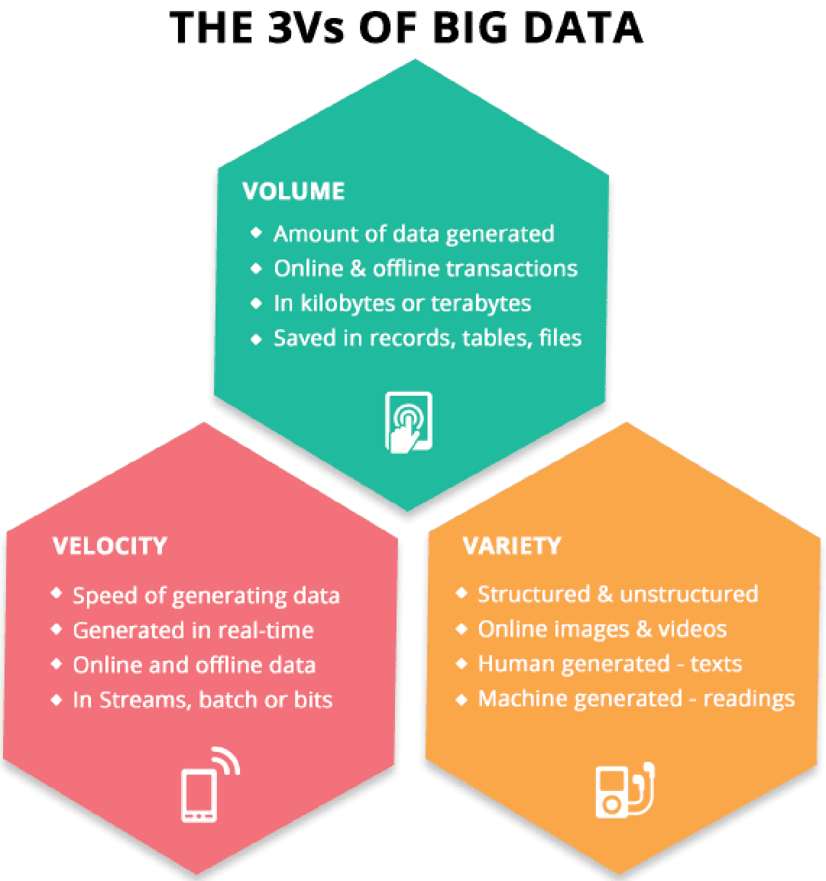
Big Data is becoming one of the most talked about technology trends nowadays. The real challenge with the big organization is to get maximum out of the data already available and predict what kind of data to collect in the future. How to take the existing data and make it meaningful that it provides us accurate insight in the past data is one of the key discussion points in many of the executive meetings in organizations.

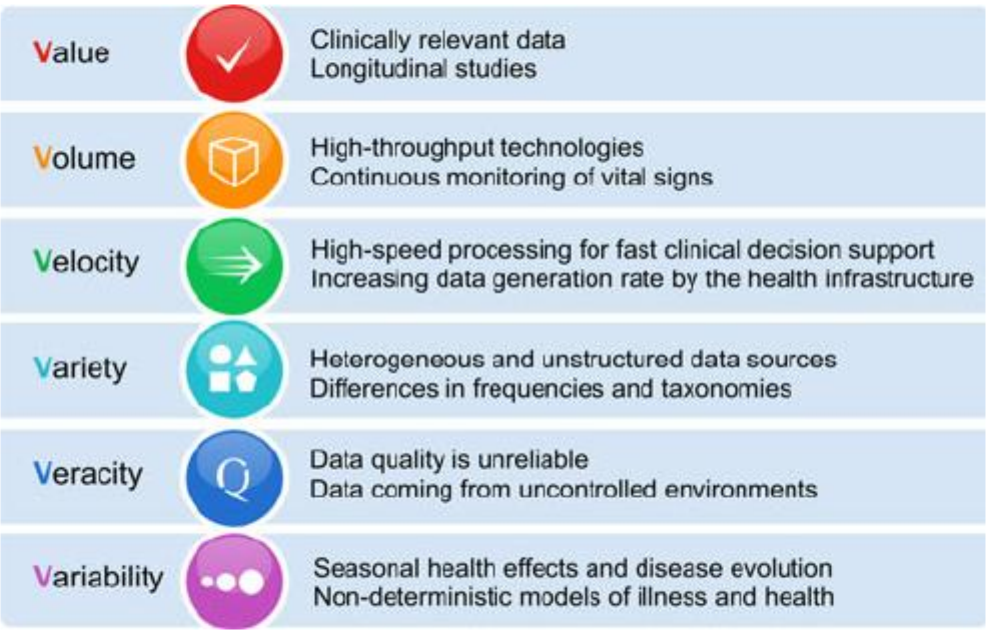
With the explosion of the data the challenge has gone to the next level and now a Big Data is becoming the reality in many organizations. The goal of every organization and expert is same to get maximum out of the data, the route and the starting point are different for each organization and expert. As organizations are evaluating and architecting big data solutions they are also learning the ways and opportunities which are related to Big Data.

There is not a single solution to big data as well there is not a single vendor which can claim to know all about Big Data. Big Data is too big a concept and there are many players – different architectures, different vendors and different technology.

Big Data 3 V’s and 6 V’s

In recent years, Big Data was defined by the “3Vs” but now there is “6Vs” of Big Data which are also termed as the characteristics of Big Data as follows:





Value: Clinically relevant data
Longitudinal studies

Volume: High-throughput technologies
Continuous monitoring of vital signs

Velocity: High-speed processing for fast clinical decision support
Increasing data generation rate by the health infrastructure

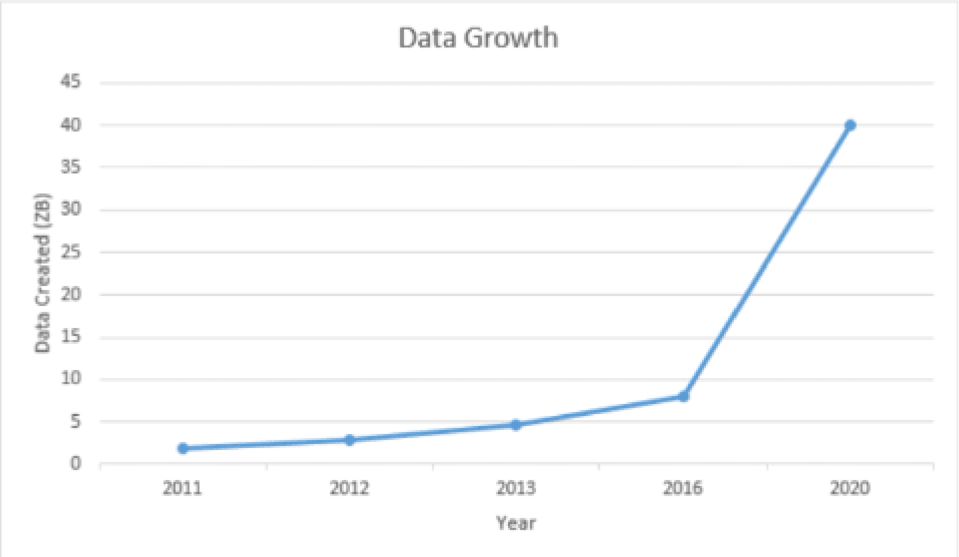
Variety: Heterogeneous and unstructured data sources
Differences in frequencies and taxonomies

Veracity: Data quality is unreliable
Data coming from uncontrolled environments

Variability: Seasonal health effects and disease evolution
Non-deterministic models of illness and health

1. 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’.
- *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabytes (6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 Exabytes of data.



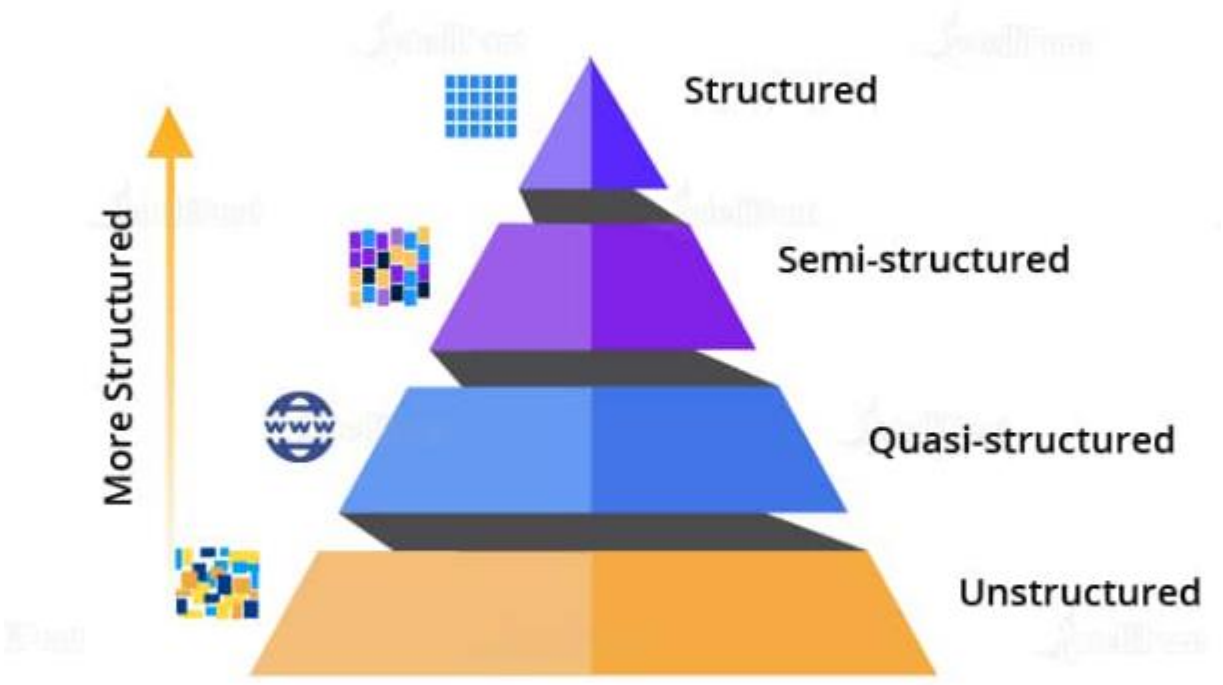
2. Velocity:

- 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.
- Sampling data can help in dealing with the issue like 'velocity'.
- Example: There are more than 3.5 billion searches per day are made on Google. Also, Facebook users are increasing by 22%(Approx.) year by year.
-

3. Variety:

- It refers to nature 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.



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 6V’s.

6. Variability:

- How fast or available data that extent is the structure of your data is changing?
- How often does the meaning or shape of your data change?
- Example: if you are eating same ice-cream daily and the taste just keep changing.

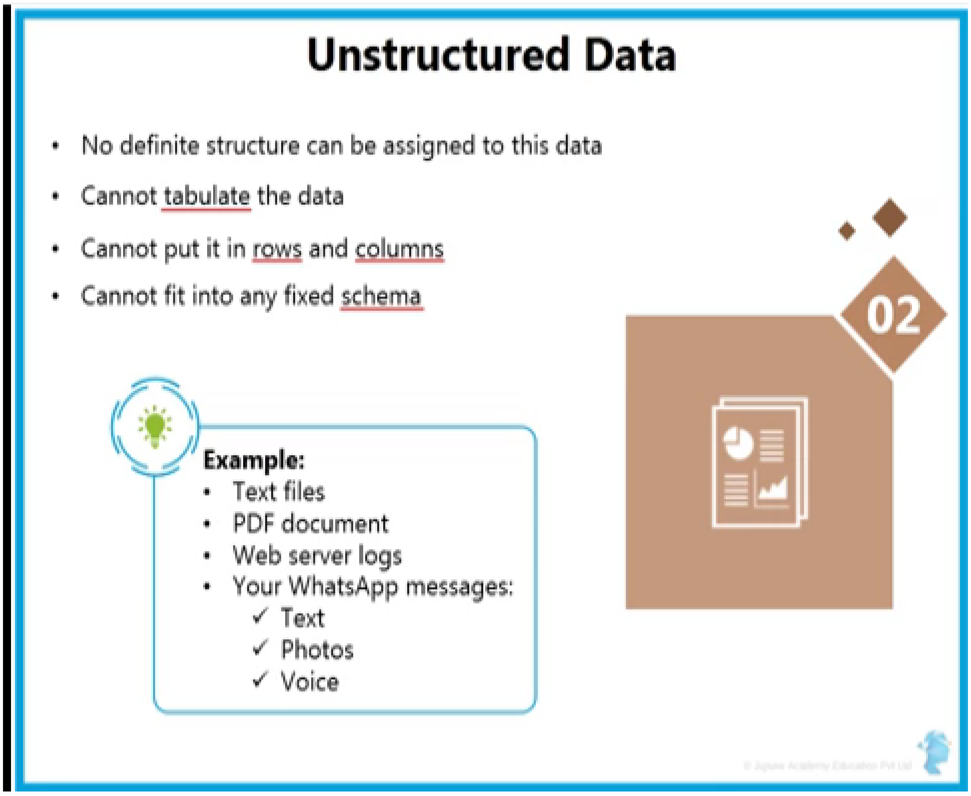
Different Types of Big Data

Big data types in Big Data are used to categorize the numerous kinds of data generated daily. Primarily there are 3 types of data in analytics. The following types of Big Data with examples are explained below:-

1. Structured Data: Any data that can be processed, is easily accessible, and can be stored in a fixed format is called structured data. In Big Data, structured data is the easiest to work with because it has highly coordinated measurements that are defined by setting parameters. Structured types of Big Data are:-

- Address
- Age
- Credit/debit card numbers
- Contact
- Expenses
- Billing

2. Unstructured Data: Unstructured data in Big Data is where the data format constitutes multitudes of unstructured files (images, audio, log, and video). This form of data is classified as intricate data because of its unfamiliar structure and relatively huge size. A stark example of unstructured data is an output returned by ‘Google Search’ or ‘Yahoo Search.’



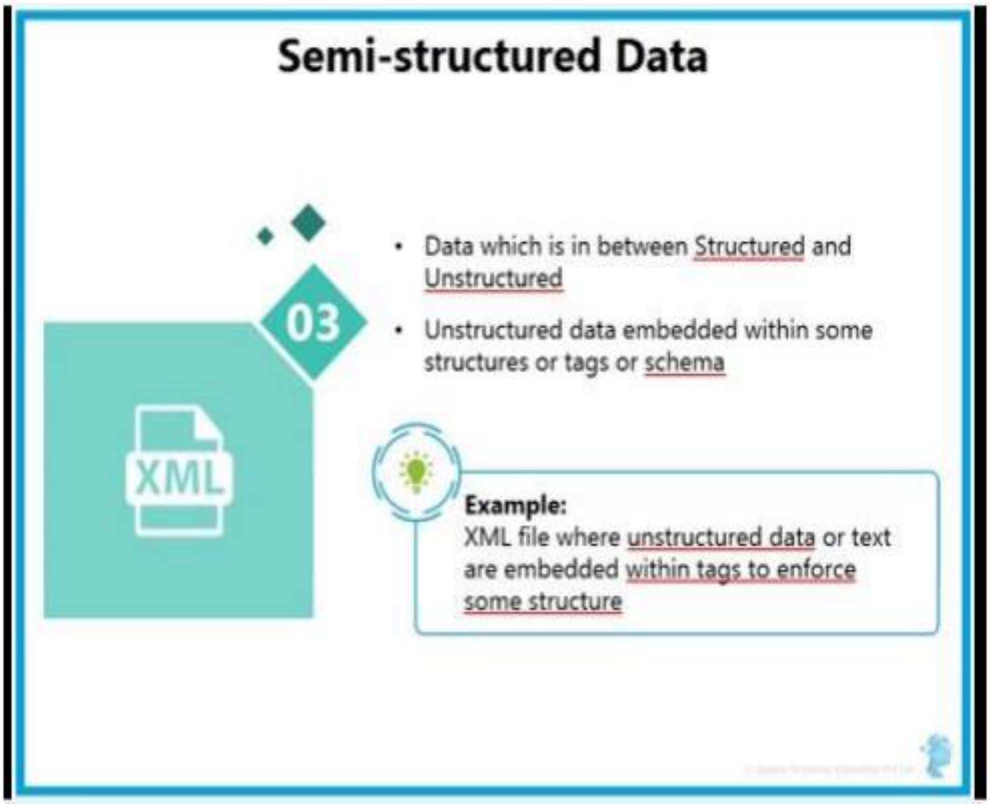
Unstructured Data

- No definite structure can be assigned to this data
- Cannot tabulate the data
- Cannot put it in rows and columns
- Cannot fit into any fixed schema

Example:

- Text files
- PDF document
- Web server logs
- Your WhatsApp messages:
  - ✓ Text
  - ✓ Photos
  - ✓ Voice

3. Semi-structured Data: In Big Data, semi-structured data is a combination of both unstructured and structured types of data. This form of data constitutes the features of structured data but has unstructured information that does not adhere to any formal structure of data models or any relational database. Some semi-structured data examples include XML and JSON.



Semi-structured Data

- Data which is in between Structured and Unstructured
- Unstructured data embedded within some structures or tags or schema

Example:
XML file where unstructured data or text are embedded within tags to enforce some structure

Major Sectors Using Big Data Every Day

Banking

Since there is a massive amount of data that is gushing in 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.

overnment

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 several **Big Data Analytics 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, leveraging 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.

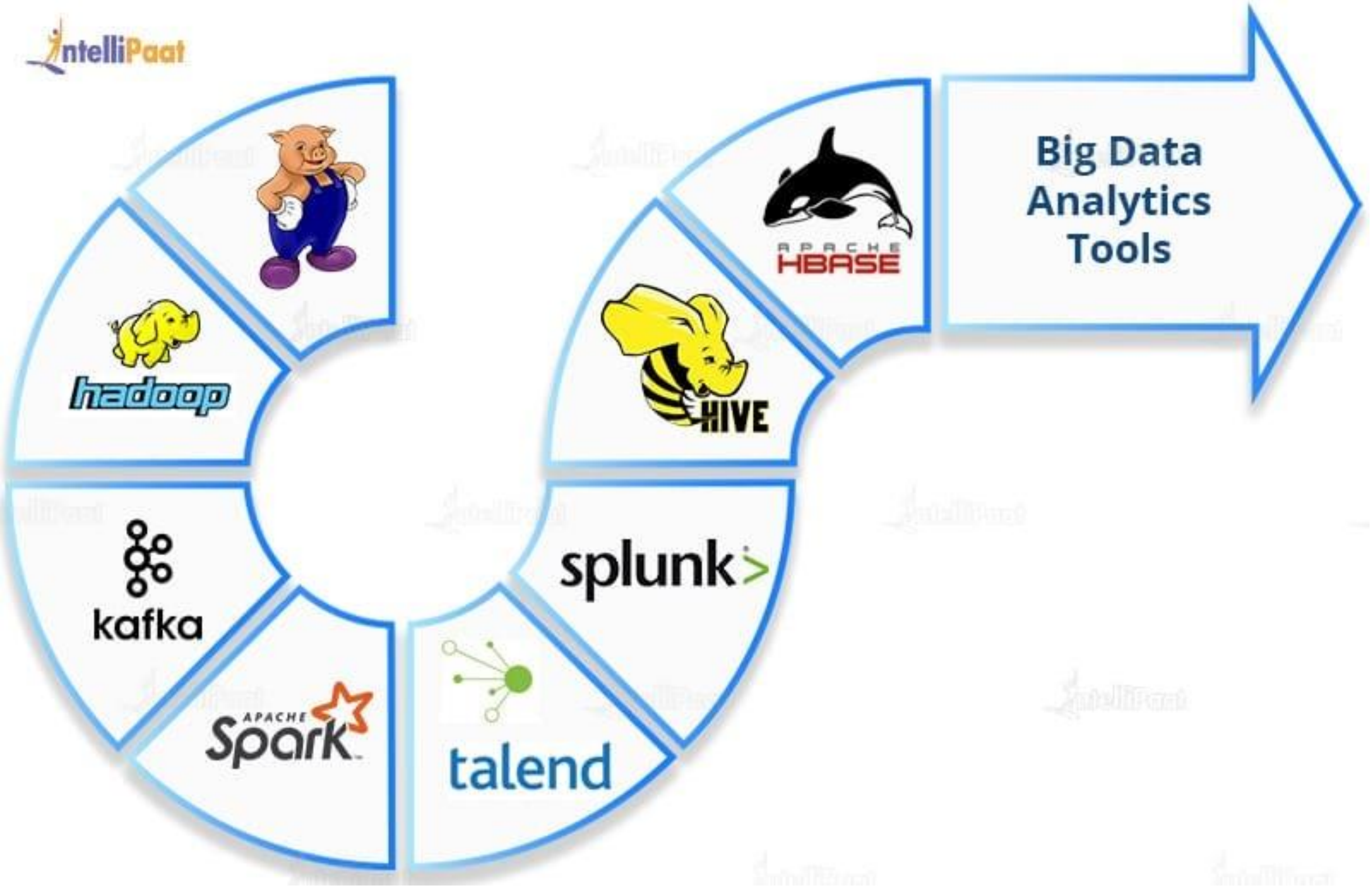
What is Big Data Analytics?

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.

How is Big Data Analytics used today?

Big Data Analytics is used in several industries to allow organizations and companies to make better decisions, as well as verify and disprove existing theories or models. The focus of Data Analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

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 between Traditional data and Big data

Traditional Data	Big Data
Traditional data is generated in enterprise level.	Big data is generated outside the enterprise level.
Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
Traditional database system deals with structured data.	Big data system deals with structured, semi-structured,database, and unstructured data.
Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per seconds.
Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.
Data integration is very easy.	Data integration is very difficult.
Normal system configuration is capable to process traditional data.	High system configuration is required to process big data.

Traditional Data	Big Data
The size of the data is very small.	The size is more than the traditional data size.
Traditional data base tools are required to perform any data base operation.	Special kind of data base tools are required to perform any databaseschema-based operation.
Normal functions can manipulate data.	Special kind of functions can manipulate data.
Its data model is strict schema based and it is static.	Its data model is a flat schema based and it is dynamic.
Traditional data is stable and inter relationship.	Big data is not stable and unknown relationship.
Traditional data is in manageable volume.	Big data is in huge volume which becomes unmanageable.
It is easy to manage and manipulate the data.	It is difficult to manage and manipulate the data.
Its data sources includes ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc.	Its data sources includes social media, device data, sensor data, video, images, audio etc.