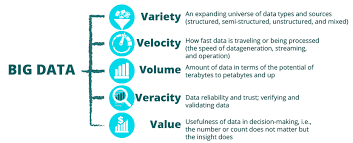
BIG DATA

Big data ***is a blanket term for the non-traditional strategies and technologies needed together, organise, process and gather insides form large data set while the problem of working with data that exceeds the computing power or storage of a single computer is not new and value of this type of computing has greatly expanded in recent years.***

***And exact definition of big data is difficult to name down because projects, vendors and business professionals use it quite differently, so the big data is large data set, the category of computing strategies and technologies that are used to handle large data set.***

***The basic requirements for working with big data are the same as the requirements for working with data sets of any sizes. However, the message scale, the speed of investigating and processing, and the characteristics of the data that must be deal with at each stage of the process present significant new challenges when designing solutions.***

* Characteristics of Big Data



1. ***Volume***
2. ***Velocity***
3. ***Variety***
4. ***Variability***
5. ***Value***
6. **Volume -**

***The scale of the information processed helps define big data system. These data set can be orders of magnitude larger than traditional data set often, because the work requirements exceed the capabilities of a single computer, this becomes a challenge of pooling, allocating and coordinating resources from groups of computers. Cluster Management and Algorithm capable of breaking task into smaller piece become important.***

1. **Velocity -**

***Another way in which big data differs significantly from other data systems is the speed that information moves through the system data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insides and update the current understanding of the system.***

***This focus on instant feedback has driven many big data engineers away from a batch-oriented approach and closer to a real time streaming system. Data is constantly being added, messaged, processed and analysed in order to keep with new information.***

1. **Variety -**

***Big data problems are often unique because of the wide range of both the sources being process and their relative quality. Data can be used from internal systems like applications and server log, from social media and often external API (Application Programming Interface)from physical device sensors, and from other providers. Big data seems to handle potentially useful data regardless of where it is coming from by consolidating int a single system. The format and type of media can vary significantly as well. Rich media like images, videos, files and audio recordings are used alongside text file, structured log.***

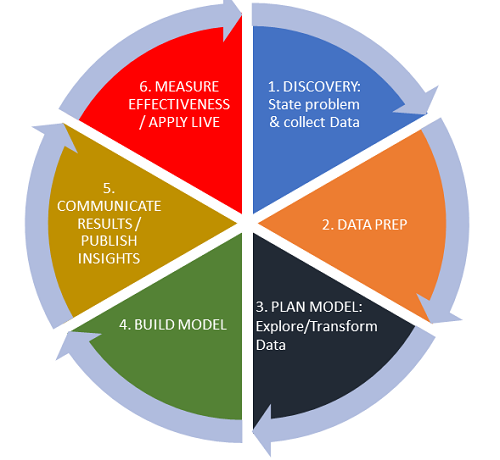
1. **Variability -**

***Variation in the data leads to wide variation in quality additional resources may be needed to identify, process, or filter low quality data to make it more useful.***

1. **Value–**

***The ultimate challenge of big data is delivering value. Sometimes, the systems and process in place are complex enough that using the data and extracting actual value can become difficult.***

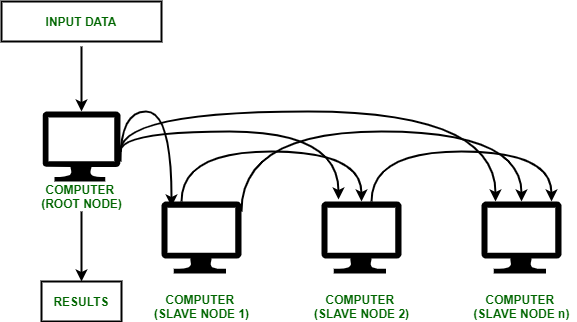
* Big Data Life Cycle



***So how is data actually processed when dealing with a big data system while approaches to implementation differ there are some common properties in the strategies and software that he can talk about generally. While the steps presented below might not be true in all cases they are widely used.***

1. ***Ingesting data into the system***
2. ***Persisting the data in storage***
3. ***Computing and analysing data***
4. ***Visualising the results***

**Clustered Computing–**



***Because of quality of big data, individual computers are often inadequate for handling the data almost stages to better address the high storage and computational needs of big data, computer clusters are a better fit.***

***Big data lusting software combines the resources of many smaller machines seeking to provide a number of benefits:***

1. **Resource Pooling:  *Combining the available storage space to hold data is a clear benefit but CPU and memory pooling is also extremely important. Processing large data sets requires large amounts of all three of these resources: -***
2. High Availability –*Clusters can provide different levels of fault tolerance and availability guarantees to prevent hardware or software failures from affective excess to data and processing.*
3. Easy Scalability - *Clusters make it easy to scale horizontally by adding additional machines to the group this means the system can react to changes in resource requirements without expending the physical resources on a machine.*

* Important terms in Big Data

1. **Big Data: *Big data is an umbrella term for data set that cannot handled by traditional computers or tools due to their volume, velocity and variety. This term is also applied to technologies and strategies to work with this type of data.***
2. **Batch Processing:*It is a computing strategy that involves processing data in large set this is typically ideal for non-time sensitive work that operates an very large sets of data.***
3. **Data Lake:*Data Lake is a term for a large repository of collected data in a relatively raw state. This is frequently used to refer to data collected in a big data system which might be unstructured and frequently changing. This differs in spirit to data warehouses.***
4. **Data Mining:*It is broad term for the practice of trying to find patterns in large sets of data. It is the process of trying to refine a mass of data into a more understandable set of information.***
5. **Data Warehouse:*Data warehouse are large, ordered repository of data that can be used for analysis and importing in contrast to a data lake, a data warehouse is composed of data that has been cleaned, integrated with other sources and it is generally well ordered. Data warehouse are often spoken about in relation to big data but typically are components of more conventional system.***
6. **ETL: *stands for Extract, Transform and Load. It refers to the process of taking raw data and preparing it for the system use. This is traditionally a process associated with data warehouse but characteristics of this process are also found in the ingestion pipe line of big data system.***
7. **Machine Learning: *It is the study and practice of designing systems that can learn, adjust and improve based data fit to them. This typically involves implementation of predictive and stats based algorithm that can continuously zero in on correct behaviour inside as more data flow through the system.***