**1.What is Big data? What are the advantages of big data?**

**Ans**

Big data refers to large and complex sets of data that cannot be easily processed and analyzed using traditional data processing tools.

**1) Enhanced decision-making:**Big Data provides organisations with access to a vast amount of information from various sources, enabling them to make data-driven decisions. By analysing patterns, trends, and correlations within the data, businesses can gain valuable insights that guide their strategies.

**2) Improved operational efficiency:**Big Data analytics helps organisations optimise their operations by identifying inefficiencies, bottlenecks, and areas for improvement. By streamlining processes and workflows, businesses can enhance productivity, reduce costs, and deliver better results.

**3) Personalisation and customer insights:** Big Data enables organisations to understand their customers at a deeper level. By analysing customer data, businesses can identify preferences, behaviours, and needs, allowing them to personalise products, services, and marketing campaigns to enhance the customer experience.

**4) Cost savings:** Big Data analytics can uncover cost-saving opportunities for organisations. By analysing data related to operations, supply chains, and resource allocation, businesses can identify areas where costs can be minimised, such as optimising inventory management or reducing energy consumption.

**5) Competitive advantage:**  By analysing market trends, consumer behaviour, and competitor data, businesses can identify new opportunities, innovate, and stay ahead of the competition.

**6) Improved risk management:**Big Data analytics helps organisations identify and mitigate risks. By analysing historical and real-time data, businesses can anticipate potential risks, such as fraud, cybersecurity threats, or supply chain disruptions.

**7) Efficient resource allocation:**With Big Data analytics, organisations can optimise resource allocation. Businesses can allocate their resources effectively by analysing data related to resource utilisation, demand forecasting, and operational efficiency, reducing waste and maximising productivity.

**8) Enhanced customer service:**Big Data enables organizations to deliver exceptional customer service. By analyzing customer interactions, feedback, and sentiment data, businesses can identify areas for improvement, personalize customer support, and proactively address customer issues.

**2.Explain the sources of the big data?**

**Ans**

**The Primary Sources of Big Data:**

**A significant part of big data is generated from three primary resources:**

* Machine data
* Social data, and
* Transactional data.

### **1. Machine Data**

Machine data is automatically generated, either as a response to a specific event or a fixed schedule. It means all the information is developed from multiple sources such as smart sensors, SIEM logs, medical devices and wearables, road cameras, IoT devices, satellites, desktops, mobile phones, industrial machinery, etc. These sources enable companies to track consumer behaviour. Data extracted from machine sources grow exponentially along with the changing external environment of the market.

### **2. Social Data**

It is derived from social media platforms through tweets, retweets, likes, video uploads, and comments shared on Facebook, Instagram, Twitter, YouTube, Linked In etc. The extensive data generated through [social media platforms](https://www.upgrad.com/blog/how-do-social-media-algorithm-works/) and online channels offer qualitative and quantitative insights on each crucial fact of brand-customer interaction.

### **3. Transactional Data**

As the name suggests, transactional data is information gathered via online and offline transactions during different points of sale. The data includes vital details like transaction time, location, products purchased, product prices, payment methods, discounts/coupons used, and other relevant quantifiable information related to transactions.

**The sources of transactional data include:**

* Payment orders
* Invoices
* Storage records and
* E-receipts

Another critical factor to consider about Big data sources is whether it is structured or unstructured.

* [Structured data](https://www.coursera.org/articles/structured-data) is typically quantitative data that is organized and easily searchable.
* Common types of structured data include names, addresses, credit card numbers, telephone numbers, star ratings from customers, bank information, and other data that can be easily searched using SQL.
* [Unstructured data](https://www.coursera.org/articles/what-is-unstructured-data) is every other type of data that is not structured.
* Unstructured data includes a variety of formats such as emails, images, video files, audio files, social media posts, PDFs, and much more.

**3.Explain 5 V’s of Big data?**

**Ans**

Volume – Volume refers to the amount of data and there is no minimum size level that constitutes big data. But most big data environment contains large amount of data massive ones in many cases.

Variety – Variety refers to the diversity of data types like structured, semi structured or unstructured formats that creates challenges in storing managing data being collected and integrating it for use in analytics applications.

Velocity – Velocity refers to how quickly data is generated and processed. Often a continuous flow of big data is used to make the business decision in real or near real time. Like with medical devices monitoring patients.

Veracity – Veracity refers to the quality and accuracy of the data and how trustworthy it is. A larger amount of data messy or incomplete data could harm more than good to a business, you know, garbage in Garbage out.

Value – Value refers to the what organizations can do with data and how valuable it is. Not all the data collected has real business value, so organization need to confirm that data is relevant and understand its potential business value before using it in analytics projects.

Variability – Variability often applies to sets of big data which can be inconsistent or formatted differently in different data sources and might also change over time.

**4.Explain Hadoop Ecosystem?**

**Ans**

* Hadoop is a framework that enables processing of large data sets which reside in the form of clusters. Being a framework, Hadoop is made up of several modules that are supported by a large ecosystem of technologies.
* *Hadoop Ecosystem*is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are *four major elements of Hadoop* i.e. **HDFS, MapReduce, YARN, and Hadoop Common.**
* Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data

**HDFS:**

* HDFS is the primary or major component of Hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.
* HDFS consists of two core components i.e.

1.Name node

2.Data Node

**Name Node** is the prime node which contains metadata (data about data) requiring comparatively fewer resources than the data nodes that stores the actual data. These data nodes are commodity hardware in the distributed environment. Undoubtedly, making Hadoop cost effective.

* HDFS maintains all the coordination between the clusters and hardware, thus working at the heart of the system.

**YARN:**

* Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.
* Consists of three major components i.e.

Resource Manager

Nodes Manager

Application Manager

* **Resource manager** has the privilege of allocating resources for the applications in a system whereas **Node managers** work on the allocation of resources such as CPU, memory, bandwidth per machine and later on acknowledges the resource manager. **Application manager** works as an interface between the resource manager and node manager and performs negotiations as per the requirement of the two.

**MapReduce:**

* By making the use of distributed and parallel algorithms, MapReduce makes it possible to carry over the processing’s logic and helps to write applications which transform big data sets into a manageable one.
* MapReduce makes the use of two functions i.e. Map() and Reduce() whose task is:
  + Map() performs sorting and filtering of data and thereby organizing them in the form of group. Map generates a key-value pair based result which is later on processed by the Reduce() method.
  + Reduce(), as the name suggests does the summarization by aggregating the mapped data. In simple, Reduce() takes the output generated by Map() as input and combines those tuples into smaller set of tuples.

**Hadloop:**

* Hadoop Common refers to the collection of common utilities and libraries that support other Hadoop modules.
* It is an essential part or module of the Apache Hadoop Framework, along with the Hadoop Distributed File System (HDFS), Hadoop YARN and Hadoop MapReduce.
* Like all other modules, Hadoop Common assumes that hardware failures are common and that these should be automatically handled in software by the Hadoop Framework.