Sarna Rani Pal

ID: 193901010

Batch:11th

Semister : 7th

**Ans to the question no-1**

**Big data:** Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate.

**Comes from:** Big data comes from myriad sources -- some examples are transaction processing systems, customer databases, documents, emails, medical records, internet click stream logs, mobile apps and social networks.

**Work**: Big Data comes from text, audio, video, and images. Big Data is analyzed by organizations and businesses for reasons like discovering patterns and trends related to human behavior and our interaction with technology, which can then be used to make decisions that impact how we live, work, and play.

**Ans to the question no-2**

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

* **Volume:** The size and amounts of big data that companies manage and analyze
* **Value:** The most important “V” from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits.
* **Variety:** The diversity and range of different data types, including unstructured data, semi-structured data and raw data.
* **Velocity:** The speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time.
* **Veracity:** The “truth” or accuracy of data and information assets, which often determines executive-level confidence.

The additional characteristic of variability can also be considered:

* **Variability:** The changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases.

**Ans to the questionno-3**

Data is now an important tool for businesses to utilise, whatever sector they operate in and whether they’re a global company or a small firm. Companies are increasingly using big data to gain a competitive advantage over business rivals but how does it benefit them?

Analysing the datasets a business gathers is just one part of the big data process. Big data experts also need to understand what the company wants to gain out of the exercise and how they intend to use the information to their benefit.

**Confident decision making** – Analytics aims to improve decision making and big data continues to support this. With so much information available big data can help business speed up their decision making process while still being confident in the choice they have made. In today’s quick paced world being able to act fast and respond to wider trends and operational changes is a huge business advantage.

**Asset optimisation** – Big data means that companies can monitor assets at an individual level. This means they can better optimise assets based on the data sourced, improving productivity, extending the lifespan of assets and reducing the downtime some assets may need. This provides a competitive advantage by ensuring the business are getting the most out of its assets and links with cutting costs.

**Cost reduction** – Big data can help businesses cut down their outgoings. From analysing energy usage to evaluating the effectiveness of staff working patterns, data collected by companies can help them identify where they can make cost savings without having a negative impact on business operations.

**Improve customer engagement** – When browsing online customers make certain decisions indicating their preferences, habits and tendencies that can then be used to improve and tailor customer dialogue, which could then translate into increased business. Understanding what each customer is looking for through the data stored on them not only means you can target them with certain products but it also gives service a personal touch that many consumers today have come expect.

**Identify new revenue streams** – Big data analytics can also help businesses identify new revenue streams and expand into other areas. Understanding customer trends and decisions allows firms to make decisions about the direction they should go. The data businesses collect can also potentially be sold, adding a further revenue stream and the potential to build partnerships with other businesses.

**Ans to the question no-5**

**Importance of Hadoop**

Hadoop is a valuable technology for big data analytics for the reasons as mentioned below:

* Stores and processes humongous data at a faster rate. The data may be structured, semi-structured, or unstructured
* Protects application and data processing against hardware failures. Whenever a node gets down, the processing gets redirected automatically to other nodes and ensures running of applications
* Organizations can store raw data and processor filter it for specific analytic uses as and when required
* As Hadoop is scalable, organizations can handle more data by adding more nodes into the systems
* Supports real-time analytics, drives better operational decision-making and batch workloads for historical analysis.

**Ans to the question no-6**

**The difference between HDFS & NFS in tabular form :**

|  |  |
| --- | --- |
| **HDFS** | **Tradditional NFS** |
| It is a file system in which data is distributed among many data nodes or networked computers. | It is a file system or protocol which allows its client to access the file over the network. |
| It is mainly used to store and process big data. | It can store and process a small amount of data. |
| Its data blocks are dispersed on the local drives of hardware. | Data is stored on a single dedicated hardware. |
| Its data is stored reliably. Here, data is available even after machine failure. | No reliability, data is not available in case of machine failure. |
| It runs on a cluster of different machines, data redundancy may occur due to replication protocol. | It runs on a single machine, with no chance of data redundancy. |
| It is for multi-domain. | It is for a single domain. |
| Here, client identity is trusted by the OS. | It is for a single domain. |
| It has different calls. It is mainly used for non-interactive programs. | It has the same system calls as O/S. |