5.1: Intro to Big Data

1. What's the difference between structured and unstructured data? Can you give examples that you've encountered for both types?

   Structured data examples include CRM databases, product databases, and invoicing systems, whereas unstructured data examples include documents, movies, and audio recordings. Structured data is easily accessible for analysis, but unstructured data is acquired in a disorganised manner, such as emails, social media posts, or unstructured surveys.

2. Given that much of big data is produced by machines and sensors, how trustworthy do you think that big data is? What characteristic of big data relates to the question of trustworthiness?

   It is imperative to recognise that datasets, whether gathered by humans or devices made by humans, are inaccurate. Errors in data recording and processing can result from both human bias and mechanical issues. Big data's veracity, which characterises the data's accuracy and dependability, is an essential quality that pertains to trustworthiness. Big data must be processed and cleansed before analysis, despite being produced by machines or sensors. This is owing to the fact that people are fallible and that the machines they design may have flaws as a result of these constraints. As a result, the authenticity of the data has an impact on the reliability of big data that has been gathered.

3. Assume that you receive a table containing customer data. You notice that some values are missing or incomplete, and the formatting is inconsistent in some columns. Based on what you've learned so far, how would you go about cleaning this table? Think about what you would do first, second, third, etc.

   Analyse the data, find missing values, and decide how to manage them based on the context. This could include assigning values, deleting records, or leaving the data alone. Select and apply a consistent formatting style to the entire data set. Filtering out missing/incomplete values, reformatting columns, and assessing the impact on the database are all steps in the process. Identifying the missing values and taking the context of the study into account will aid in determining the best strategy to manage the missing data. Adjust the formatting of misformatted columns after identifying the solution to ensure consistency in cells.

4. Can you describe tools such as Hadoop and Apache Spark and their role in big data? What do they do and how do they work?

   Hadoop and Apache Spark are analytics engines that distribute data across several computers to alleviate bottlenecks in big data research. Hadoop handles structured and unstructured data efficiently, but Apache Spark optimises query execution and in-memory caching. Spark focuses on streaming and in-memory processing, whereas both engines are utilised for scalable data analysis and querying.

5. How has the application of analytics to big data led to new discoveries and innovations? Can you give some examples?

Analytics to big data has transformed many sectors, including disaster planning, logistics, and financial management. Analytics has shortened emergency response times, saved time and money, and assisted financial institutions in reducing risk and combating fraud. Data analytics is used by sports teams, such as the NBA, to analyse patterns and draught players. Big data has numerous applications, including tracking and predicting patterns, addressing issues, and forecasting severe weather. GPS data analysis can also offer the quickest routes by taking into account aspects such as weather, time of day, and stop lights.