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?

Answer: **Structured** data refers to organized and formatted data that can be easily processed, stored, and analyzed. It is typically stored in databases and follows a predefined schema. Examples include sales data with columns for date, product, and revenue, or customer information with fields for name, address, and contact details. **Unstructured** data, on the other hand, lacks a specific structure and is more difficult to organize and analyze. It includes data in formats such as text documents, social media posts, emails, images, videos, and audio recordings. For instance, social media feeds, customer reviews, or recorded customer service calls are unstructured data. Both types have their challenges. Structured data is easier to analyze but may not capture the full context, while unstructured data provides more context but requires advanced techniques for analysis.

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?

The trustworthiness of big data depends on the quality and reliability of the data sources, as well as the accuracy and integrity of the data itself.
The characteristic of big data that relates to trustworthiness is data quality. This includes factors like data accuracy, completeness, consistency, and timeliness.
If the data sources and collection processes are reliable, and the data is properly validated and verified, then the trustworthiness of the big data increases. However, challenges like data errors, biases, and inconsistencies can undermine trust.
Implementing robust data governance practices, ensuring data quality checks, and validating data against known benchmarks can help enhance the trustworthiness of big data.

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.

Answer:When cleaning a table with missing or incomplete values and inconsistent formatting, I will follow these steps:

1. First, identify the missing or incomplete values. Determine the extent of missing data and evaluate its impact on analysis. Decide whether to impute missing values or exclude the corresponding records, depending on the data's importance and quantity of missing values.

2. Next, address inconsistent formatting. Standardise the formatting across columns by applying appropriate transformations such as capitalization, removing leading/trailing spaces, or converting data types.

3. Clean the missing or incomplete values. Depending on the type of data, we can use techniques like mean/median imputation or regression-based imputation to fill in missing values.

4. Check for data consistency and accuracy. Validate data against known benchmarks or external sources. Remove duplicate records if applicable.

5. Finally, document the cleaning steps performed and communicate any assumptions or limitations associated with the cleaned data to maintain transparency.

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?

Answer: Hadoop and Apache Spark are popular tools for processing and analyzing big data:

Hadoop is a framework that enables distributed storage and processing of large datasets across clusters of computers. It uses the Hadoop Distributed File System (HDFS) for storing data and the MapReduce programming model for processing data in parallel. Hadoop divides data into chunks and distributes them across multiple nodes for efficient processing. It is particularly useful for batch processing and handling structured and unstructured data.

Apache Spark, on the other hand, is an open-source data processing engine designed for speed and real-time analytics. It provides an in-memory computing framework that allows data to be processed in-memory, leading to faster data processing and iterative computations. Spark supports various data processing tasks, including batch processing, interactive queries, streaming, and machine learning. It provides a high-level API for data processing and supports multiple programming languages, making it versatile and user-friendly.

Both Hadoop and Apache Spark play crucial roles in big data analytics by enabling the processing and analysis of large datasets in a distributed and scalable manner, facilitating efficient data storage and processing across clusters of machines.

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

Answer: The application of analytics to big data has resulted in groundbreaking discoveries and innovations across various domains. For example:

1. Healthcare: Big data analytics has led to the identification of patterns in patient data, enabling personalized treatments and improving disease prediction models.

2. Marketing: Analyzing large datasets has helped businesses understand customer preferences and behavior, leading to targeted marketing campaigns and improved customer experiences.

3. Finance: Advanced analytics on big data has enabled better fraud detection, risk assessment, and algorithmic trading strategies.

4. Transportation: Big data analytics has optimized traffic flow, improved logistics, and enabled the development of autonomous vehicles.

These examples demonstrate how big data analytics has revolutionized industries, leading to transformative discoveries and innovations.