

**Data Analytics (CS3203N)**  
**Practice Set I**  
**(Topic: Introduction)**

**1. Which of the following are not necessarily data analytic techniques?**

Sorting, searching, and data compression are not always data analytical approaches.

Data compression involves removing extraneous or superfluous information to reduce the size of the data. Although it is a widely used data management method, it is not a data analytic method.

Data that needs to be protected from illegal access by encryption is transformed into a code. It is a security measure rather than a data analytics method.

Sorting and searching: Involve arranging and locating data in accordance with predetermined standards. Although these methods are crucial for data management, they are not frequently regarded as data analytics methods.

**2. Give some examples of data analytics applications with reference to the following data?**

**(a) Credit card transactions of a bank:**

Fraud detection: Involves looking through transaction data to find any irregularities or strange trends that can point to fraudulent behavior.

Customer behavior analysis: Recognizing consumer spending trends to enhance product and marketing tactics.

Risk management: Involves determining a customer's creditworthiness and estimating the likelihood that they will default.

**(b) Calls in a call center:**

Call Volume Analysis: Understanding the volume of calls a call center receives will enable you to optimize staffing levels.

Call Routing Optimization: Optimizing call routing involves analyzing call data to better direct calls to the most qualified operators.

Call Sentiment Analysis: Analyzing call content and tone to gauge customer satisfaction and demands is known as customer sentiment analysis.

(c) Geo-satellite data:

Environmental monitoring: Examining satellite photos to keep tabs on shifts in climatic patterns, deforestation, and land use.

Disaster response: Quick assessment of the harm caused by natural catastrophes and prioritization of relief activities using satellite data.

Urban Planning: Utilizing satellite data to comprehend urban growth trends and guide infrastructure and transportation planning decisions.

(d) Facebook messages:

User Behavior Analysis: Understanding how users interact with a platform, such as their posting frequency and timeliness, can help with product development.

Sentiment Analysis: Analyzing the mood of Facebook messages to comprehend the attitudes and feelings of users

Ad Targeting: Targeted advertising is more effective when it is based on user interests and preferences, which can be determined by analyzing Facebook messages.

**3. Discuss whether or not each of the following activities is a data analytic task.**

(a) Gender-based segmentation of a company's clients is a work in data organization, not data analytics.

(b) Segmenting a company's consumers based on their profitability is an example of data analysis, which is the process of organizing and segmenting data.

(c) Calculating a company's total revenue involves data aggregation, a type of data analysis.

(d) Sorting a student database by student ID number is an example of data organization rather than data analysis.

(e) Making predictions based on the results of a (fair) dice roll is a statistical modeling task, which is a type of data analysis.

(f) Using past data, anticipate a company's stock price in the future: this is a predictive modeling problem, which is a type of data analysis.

(g) Monitoring patient heart rates for anomalies is a task that involves monitoring and analyzing data, which is a type of data analysis.

(h) Monitoring seismic waves for earthquake activity is a task that involves data analysis, which includes data monitoring.

(i) A signal processing task, or data analysis, is to extract the frequencies from a sound wave.

4. For each of the following data sets, explain whether or not data privacy is an important issue.

(a) Census data collected from 1950-2010: Because census data includes personal information about people, like their age, gender, occupation, and income, data privacy is a significant concern. To ensure the persons' privacy, this information must be treated with sensitivity.

(b) IP address and visit times of Web users who visit your websites: Given that IP addresses and visit times of Web users can be used to track their online behavior, interests, and personal information, data privacy is a crucial concern. To guarantee the users' privacy, this information must be safeguarded.

(c) Images from earth-orbiting satellites: Since photographs from earth-orbiting satellites do not contain personal information, data privacy is not a significant concern. However, it's crucial to safeguard the privacy of any private information that might be seen in the photographs.

(d) Names and addresses of the people from the telephone book: Given that phone book names and addresses are open to the public, data privacy is not a significant concern. It's crucial to make sure that this data isn't used for nefarious endeavors like spam or unwanted advertising.

(e) Names and addresses of the people from the Web: Names and addresses found online frequently come from social networking sites or other online sources and may contain sensitive personal data, making data privacy a significant concern.

(f) Messages in Social media sites such as Facebook, Twitter, etc.: Because messages on social media platforms frequently contain sensitive information that needs to be protected in order to preserve user privacy, data privacy is a crucial concern.

5. What are the issues to process Big data? (Mention at least FIVE issues/challenges).

- Volume: Processing big data's sheer bulk, which can quickly exceed petabytes or exabytes in size, is one of the largest hurdles. Effective data management and organization techniques are also necessary, as well as a sizable amount of processing power and storage.
- Velocity: Another issue is the rapid generation of large data, which can soon outpace the processing power of conventional data processing systems. High-performance computing and real-time data processing technologies are required for this.

- Variety: Big data can come in many different formats, including structured, semi-structured, and unstructured data, which makes it difficult to process and analyze. This requires the use of flexible data processing tools and techniques that can handle a wide range of data types and structures.
- Veracity: Big data's quality and accuracy can be a problem because it sometimes includes a lot of irrelevant, insufficient, or conflicting information. To ensure that the data being analyzed is of high quality, this calls for efficient data cleansing, validation, and verification methods.
- Privacy and security: Big data frequently includes delicate personal information that must be safeguarded against illegal access and exploitation, such as financial and medical records. To guarantee the privacy and security of the data, this necessitates the adoption of robust security measures, such as encryption and access controls.