Q 1(a) [9 Marks]

Given the following brief to design a system for a data collection task:

(i) (ii) List three (3) important questions you would ask your client.

Describe the data and/or file formats that you are likely to use in collecting the data.

(iii) Suggest a type of database system to use for this project, giving a reason for your choice.

“ Local councils in Ireland want to understand water usage patterns and customer sentiment in response to proposed water charges. They have historical records of water usage (Litres per day per household), data from water processing plants (e.g., volume processed per day, record of faults, duration of outages, cost per litre processed), maps of the existing water pipelines, access to Central Statistics Office data with population information (e.g., density, age, rate of change) and have commissioned a survey of Irish adults (e.g., opinion on water charges, how they use water). They want to identify where future investment in infrastructure should occur. ”

ANSWER

Question 1

Q1(a) [9 Marks]

1. **Three important questions to ask the client:**
    - What is the desired frequency for data updates and reporting (e.g., real-time or periodic)?
    - What are the key performance indicators (KPIs) for water usage and customer sentiment analysis?

  - Are there any specific regulations or privacy requirements for handling the data (e.g., GDPR compliance)?

2. **Data/File formats likely used:**
   - **CSV/Excel**: For historical records and tabular data (water usage, customer sentiment).
   - **GeoJSON/Shape files**: For maps of water pipelines.
   - **JSON/XML**: For survey results and data from Central Statistics Office APIs.
3. **Type of database system and reason for choice:**
   - **Relational Database Management System (RDBMS)**: For structured data such as historical water usage and population statistics.
   - **Reason**: RDBMS like MySQL or PostgreSQL provides strong support for querying, reporting, and maintaining relationships between data entities.

Q 1(b) [8 Marks]

Categorise the following data attributes as either Qualitative or Quantitative; Discrete

or Continuous (if appropriate) and Nominal or Ordinal or Interval or Ratio.

(i) Type of pet (e.g., cat, dog, bird, fish)

(ii) Number of pets currently owned

(iii) Weight of the pets

(iv) Happiness of pet owners (self-rated from 1 to 5).

ANSWER

Share

MM

You said:

20CA682.pdf

PDF

I have an open book test please find all answers from above pdf

ChatGPT said:

ChatGPT

I can assist you with finding answers to the open-book test. Let me know which specific questions you'd like help with from the exam.

4o

You said:

all

ChatGPT said:

ChatGPT

Providing answers for all questions from the exam might take some time, as I'll need to analyze each question carefully. Let's go step by step. I'll provide the answers for **Question 1** first, and then we can move on to the other questions.

Question 1

Q1(a) [9 Marks]

1. **Three important questions to ask the client:**
    - What is the desired frequency for data updates and reporting (e.g., real-time or periodic)?
    - What are the key performance indicators (KPIs) for water usage and customer sentiment analysis?
    - Are there any specific regulations or privacy requirements for handling the data (e.g., GDPR compliance)?
2. **Data/File formats likely used:**
    - **CSV/Excel**: For historical records and tabular data (water usage, customer sentiment).
    - **GeoJSON/Shape files**: For maps of water pipelines.
    - **JSON/XML**: For survey results and data from Central Statistics Office APIs.
3. **Type of database system and reason for choice:**
    - **Relational Database Management System (RDBMS)**: For structured data such as historical water usage and population statistics.

    - **Reason**: RDBMS like MySQL or PostgreSQL provides strong support for querying, reporting, and maintaining relationships between data entities.

Q1(b) [8 Marks]

Categorization of data attributes:

1. **Type of pet**:
    - **Qualitative**: Describes a category.
    - **Nominal**: No inherent order exists among types of pets.
2. **Number of pets currently owned**:
    - **Quantitative**: Represents a measurable count.
    - **Discrete**: Countable values (0, 1, 2...).
    - **Ratio**: Zero means no pets, and all values can be compared.
3. **Weight of the pets**:
    - **Quantitative**: Represents a measurable amount.
    - **Continuous**: Can take any value within a range.
    - **Ratio**: Zero weight indicates the absence of weight.
4. **Happiness of pet owners (1 to 5)**:
    - **Qualitative**: Represents a subjective assessment.
    - **Ordinal**: Ratings follow a logical order, but intervals may not be equal.

Q 1(c) [4 Marks]

Which of the following descriptions of data ([A], [B] or [C]) are mostly likely to be

classified as “big data”? Briefly explain your reasoning, including any assumptions,

referring to the standard “V’s” for defining “big data”.

[A] The “Titanic” dataset showing passenger details from the final voyage of the ship.

[B] Records from Spotify of the tracks listened to by each user (est. 232M users).

[C] Sales records from the DCU merchandise store.

ANSWER

1(c) [4 Marks]

Big Data Classification:

- **Spotify records ([B])** are most likely classified as big data.
  - **Reason**: Refers to large-scale data meeting the "3 Vs" of Big Data:
    - **Volume**: Data from 232 million users.
    - **Velocity**: Constantly updated with real-time streaming.
    - **Variety**: Includes audio features, user preferences, and metadata.

Q 1(d) [4 Marks]

Describe the process of scraping data from a website. Give two (2) rules that you

should remember when using this as data source.

ANSWER

Q1(d) [4 Marks]

Process of scraping data from a website:

1. Use a scraping tool (e.g., Beautiful Soup in Python) to extract HTML data.
2. Parse the data and save it into a structured format (e.g., CSV, database).

Rules for web scraping:

- **Respect the website's Terms of Service**: Avoid prohibited actions like overloading servers.
- **Use ethical practices**: Always include proper attribution and avoid scraping sensitive/private data.

Q 2(a) [10 Marks]

“ Local councils in Ireland want to understand water usage patterns and customer sentiment in response to proposed water charges. They have historical records of water usage (Litres per day per household), data from water processing plants (e.g., volume processed per day, record of faults, duration of outages, cost per litre processed), maps of the existing water pipelines (location, capacity), access to Central Statistics Office data with population information (e.g., density, age, rate of change) and have commissioned a survey of Irish adults (e.g., opinion on water charges, how they use water). They want to identify where future investment in infrastructure should occur. ”

Using the UK data archive Data Management Lifecycle (1. Creating Data; 2. Processing Data; 3. Analysing Data; 4. Preserving Data; 5. Giving Access to Data; 6. Re-Using data), explain how you could go about this task and give examples of data analytics tasks, methods and tools that could be used at each stage where relevant.

ANSWER

Question 2

Q2(a) [10 Marks]

Using the UK Data Archive Data Management Lifecycle, here's how to approach the task:

1. **Creating Data**:
    - Collect historical records (e.g., water usage per household, water processing plant records).
    - Conduct surveys to gather customer sentiment data.
    - Use tools like Google Forms for surveys and GPS mapping software for pipeline data.
2. **Processing Data**:
    - Clean the data by removing inconsistencies or duplicates (e.g., resolving missing values in historical records).
    - Use tools like Python (pandas library) or ETL tools like Talend.
3. **Analysing Data**:
    - Perform exploratory data analysis (EDA) to identify trends in water usage and sentiment.
    - Apply machine learning models for predictive analytics (e.g., forecasting future infrastructure needs).
    - Use tools like Tableau, R, or Python (e.g., Scikit-learn).
4. **Preserving Data**:
    - Store data in secure, scalable databases (e.g., PostgreSQL or MongoDB).
    - Ensure regular backups and follow data retention policies.
5. **Giving Access to Data**:
    - Develop dashboards or APIs for stakeholders to access data insights.
    - Use platforms like Power BI or Tableau to create interactive reports.
6. **Re-Using Data**:

    - Share anonymized datasets for future research and policymaking.
    - Use standard formats (e.g., CSV, JSON) to ensure compatibility.

Q 2(b) [8 Marks]

Metadata is used to describe and define the content of data so it can be found and

used more easily. List and explain, with examples, four (4) potential issues with

using metadata created by human users.

Q 2(c) [4 Marks]

Open datasets are made freely available for all people to access. Identify and

explain two (2) potential problems that may arise either in making data open or using

open data.

Q 2(d) [3 Marks]

HDFS enables Map-Reduce processing in Hadoop by replicating large files into

blocks across DataNodes, managed by NameNodes. How does HDFS prevent or

limit data corruption errors?

ANSWERS

Q2(b) [8 Marks]

Four potential issues with metadata created by human users:

1. **Inconsistency**:

    - Example: Different formats for dates (e.g., "DD/MM/YYYY" vs. "MM-DD-YYYY").
    - Solution: Enforce strict metadata standards.
2. **Subjectivity**:
    - Example: Tags or keywords may reflect personal biases (e.g., labeling data as "important").
    - Solution: Use predefined vocabularies or taxonomies.
3. **Errors or Omissions**:
    - Example: Missing metadata fields (e.g., source, creation date).
    - Solution: Automate metadata generation where possible.
4. **Duplication**:
    - Example: Multiple users entering overlapping or redundant metadata.
    - Solution: Implement validation rules to avoid duplicates.

Q2(c) [4 Marks]

Two potential problems with open datasets:

1. **Privacy Concerns**:
    - Example: Open datasets may inadvertently include sensitive or personal information.
    - Solution: Anonymize data before sharing.
2. **Quality Issues**:
    - Example: Data may be incomplete, outdated, or inaccurate.
    - Solution: Validate and clean the data before use.

Q2(d) [3 Marks]

How HDFS prevents/limits data corruption:

- **Replication**: Each file is divided into blocks and replicated across multiple DataNodes.
- **Error Detection**: HDFS uses checksums to verify the integrity of blocks during read/write operations.
- **Automatic Recovery**: If a block becomes corrupted, HDFS fetches a healthy copy from another DataNode.

Q 3(a) [8 Marks]

(i) Give simple example metadata (3-4 attributes) describing your pen (ie,

what you are using to write this exam paper).

(ii) For each example you've identified, is it Descriptive, Administrative or

Structural metadata?

(iii) How could a standard be used for this type of metadata?

(iv) Identify one (1) problem with enforcing a standard.

Q 3(b) [6 Marks]

Having successfully gathered data for local councils to understand water usage (Q1

or Q2), give two (2) examples of possible *data glitches* and explain how they might

result in poor decision making.

Q 3(c) [6 Marks]

Identify three (3) possible data errors in the sample view of a simple expenses table below. What methods for data cleaning would you recommend to clean this dataset?

Total Expenses €27 €29 €21 €32 €21

Jan Feb Apr May Jun

Jane €5 €5 €10 €9 €8

John 6.2 €7 €6 €5 €6

Sally €10 €7 €10 €9 €7

Simon €6 €10 €5 €9

Q 3(d) [5 Marks]

Give an example of data that could be considered sensitive or personal data under

GDPR regulations. You are working for DCU, someone breaks into your office and

steals a laptop with this data stored on it! What actions should you take?

ANSWERS:

Q3(a) [8 Marks]

1. **Example metadata describing a pen**:
    - Color: Blue.
    - Ink type: Gel.
    - Brand: Pilot.
    - Length: 15 cm.
2. **Type of metadata**:
    - **Color**: Descriptive metadata (describes an attribute).
    - **Ink type**: Descriptive metadata (specific characteristic).

    - **Brand**: Administrative metadata (identifies the manufacturer).
    - **Length**: Structural metadata (physical property).
3. **Use of standards for this metadata**:
    - Standards (e.g., Dublin Core) ensure consistency and interoperability across systems.
    - For example, using controlled vocabularies for color ("Blue" vs. "#0000FF") avoids ambiguity.
4. **Problem with enforcing a standard**:
    - Adoption may be inconsistent if users are unfamiliar or unwilling to follow it, leading to non-standardized entries.

Q3(b) [6 Marks]

1. **Examples of data glitches and their consequences**:
    - **Example 1**: Duplicate data entries (e.g., duplicate water usage records).
        - Consequence: Inflated usage metrics might lead to overinvestment in infrastructure.
    - **Example 2**: Incorrect geolocation data for pipelines.
        - Consequence: Misaligned planning and allocation of resources.
2. **Explanation**:
    - These glitches can result in poor decision-making by distorting the true state of the system.

Q3(c) [6 Marks]

1. **Three possible data errors in the table**:

  - **Error 1**: "6.2" (John’s value in January) should be formatted consistently (e.g., as "€6.2").
  - **Error 2**: Missing value for Simon in June.
  - **Error 3**: The column total for April seems incorrect (does not sum up to €21).

2. **Recommended data cleaning methods**:
   - **Standardization**: Ensure consistent formatting (e.g., all monetary values prefixed by "€").
   - **Imputation**: Fill missing values using averages or contextual data.
   - **Validation**: Recalculate totals to ensure accuracy.

Q3(d) [5 Marks]

1. **Example of sensitive or personal data under GDPR**:
   - Names and addresses of employees or students.
2. **Actions to take after a data breach**:
   - **Notify authorities**: Report the breach to the Data Protection Commission within 72 hours.
   - **Inform affected individuals**: Advise on steps to protect their data (e.g., changing passwords).
   - **Secure systems**: Investigate and implement measures to prevent future breaches.

Q 4(a) [12 Marks]

In the appendix, Figure 1 shows a graphic.

(i) (ii) What visual communication goals are evident? [2 marks]

Identify two (2) design principles and explain how the graphic applies them

to fulfil the communication goals. [4 marks]

(iii) The figure has been converted to greyscale. What colours would you

recommend to use to highlight important points and why? [4 marks]

(iv) Identify two (2) attributes that the graphic uses to encode data. [2 marks]

Q 4(b) [6 Marks]

Identify the gestalt principles of visualisation present in each of the 3 images ([A], [B]

& [C]) below.

[A] [B] [C]

Q 4(c) [4 Marks]

In some visualisations, size of objects represents a quantity. If using a 2d shape

such as a circle to represent a quantity, what does the designer need to be careful

of? You can draw a sketch to illustrate the problem.

Q 4(d) [3 Marks]

Which is of greater importance in a visualisation - luminance (brightness/contrast) or

colour (hue)? Justify your answer.

ANSWERS:

Q4(a) [12 Marks]

1. **Visual Communication Goals (2 Marks)**

    - Convey data trends clearly and effectively.
    - Highlight important insights, enabling quick interpretation.

2. **Two design principles and their application (4 Marks)**
    - **Clarity**: The graphic uses appropriate labeling and avoids clutter, ensuring the information is easy to read.
    - **Consistency**: Similar data points are encoded using the same symbols or styles to avoid confusion.
3. **Recommended colors and reasons (4 Marks)**
    - Use contrasting colors such as **blue** and **orange** to highlight key points.
    - **Reason**: These colors are distinguishable in color-blind palettes and ensure clear visibility even in print.
4. **Two attributes used to encode data (2 Marks)**
    - **Size**: Represents magnitude or importance.
    - **Position**: Encodes numeric values along a defined axis, ensuring easy comparisons.

Q4(b) [6 Marks]

Gestalt principles in the images:

- **Image A**: **Proximity –** Grouped elements suggest they belong to the same category or share a relationship.
- **Image B**: **Similarity –** Elements with similar shapes, colors, or styles are perceived as related.
- **Image C**: **Continuity –** A visual flow guides the viewer's eye along paths or curves in the design.

Q4(c) [4 Marks]

Using 2D shapes (e.g., circles) to represent quantity:

- **Potential problem**: Misinterpreting the area rather than the diameter. Viewers often perceive differences in size incorrectly, as they compare areas rather than proportions.

Illustration: A small circle with a diameter of 2 and a large circle with a diameter of 4 are interpreted as "twice as large," but the large circle's area is actually four times bigger.

Solution: Use linear scales (bars) or annotate circles to clarify values.

Q4(d) [3 Marks]

Greater importance in visualization: Luminance (brightness/contrast) or Color (hue)?

- **Answer**: Luminance is more important because:
    - Contrast ensures readability, particularly in grayscale or low-light conditions.
    - Human perception prioritizes brightness over hue, making data distinctions clearer.

Justification: While hue enhances aesthetics, luminance directly affects the visibility and accessibility of the visualization.

QUESTION 5 [TOTAL MARKS: 25]

Q 5(a) [12 Marks]

(i) Identify two (2) specific problems with the design of the graph below. [2 marks]

(ii) Using the data from the graph, **sketch** an alternative graph for displaying the information to highlight why CA682 (Data Management and

Visualisation) is the most important module in the Data Analytics course. If

you want to use colour then you can write in the colour name. [6 marks]

(iii) Label (ie, identify) on your sketch the following graph components: x-axis,

y-axis, title and the marks used. [4 marks]

Q 5(b) [8 Marks]

Given the following visualisation tasks, suggest an appropriate graph type (specific

chart type not just a category) for each to display the information and give a brief

justification.

[A] Understand the relationship between maximum daily temperature (^{0}C) and

average daily personal water consumption (Litres) in Ireland.

[B] Show the improvement in sales (total profit in €) over the past 5 years for

your product compared to your competitors.

[C] The most popular method of travel to DCU during 2019.

[D] Distribution of grades in CA682 over the past 5 years.

Q 5(c) [5 Marks]

Describe the four stages of understanding that happen when you look at a graphic or

chart. Why does this mean that 3D effects in 2D graphs make understanding

difficult?

ANSWERS:

Question 5

Q5(a) [12 Marks]

1. **Two problems with the graph design (2 Marks)**:
    - **Cluttered layout**: Too many elements or unclear labeling can overwhelm viewers.
    - **Inappropriate scale**: The axis scale might distort the data, making it harder to interpret accurately.
2. **Sketch of an alternative graph (6 Marks)**:
    - Use a **bar chart** to display the importance of CA682 (Data Management and Visualization) compared to other modules.
    - Label axes clearly:
        - **x-axis**: Module names.
        - **y-axis**: Importance score (e.g., based on surveys or course objectives).
    - Use contrasting colors (e.g., highlight CA682 in **blue** and other modules in **gray**) to emphasize its importance.
    - Add a meaningful title like: "Relative Importance of Modules in Data Analytics."
3. **Graph components to label (4 Marks)**:
    - **x-axis**: Module names.
    - **y-axis**: Importance score.
    - **Title**: Descriptive title at the top.
    - **Marks**: Bars representing data points.

Q5(b) [8 Marks]

Appropriate graph types for each task:

1. **[A] Understand the relationship between temperature and water consumption**:

    - **Scatter plot**: Ideal for visualizing correlations between two variables.
    - **Justification**: It shows trends and patterns in continuous data.
2. **[B] Improvement in sales over 5 years (compared to competitors)**:
    - **Line chart with multiple series**: Tracks trends over time for multiple entities.
    - **Justification**: Makes it easy to compare progress across competitors.
3. **[C] Most popular method of travel to DCU in 2019**:
    - **Pie chart or bar chart**: Shows the proportions of different methods clearly.
    - **Justification**: Effective for comparing categorical data.
4. **[D] Distribution of grades in CA682 over 5 years**:
    - **Histogram**: Displays the frequency of grade ranges (e.g., A, B, C).
    - **Justification**: Best for showing distributions in numerical or ordinal data.

Q5(c) [5 Marks]

Four stages of understanding a graphic or chart:

1. **Perception**: The viewer identifies shapes, colors, and other visual elements.
2. **Interpretation**: Patterns, relationships, and trends in the data are discerned.
3. **Evaluation**: The viewer assesses whether the information aligns with their goals or questions.
4. **Decision-making**: Insights from the graphic are used to inform choices or actions.

Why 3D effects make understanding difficult:

- **Distortion**: 3D graphs often skew proportions, making it harder to compare values accurately.
- **Complexity**: They add unnecessary visual elements, increasing cognitive load.