

|  |
| --- |
| **DUBLIN CITY UNIVERSITY** |

**SEMESTER 1 EXAMINATIONS 2015/2016**

**MODULE:** CA682 - Data Management and Visualisation

**PROGRAMME(S):**

MCM M.Sc. in Computing

ECSA Study Abroad (Engineering & Computing)

**YEAR OF STUDY:** 1,X

**EXAMINERS:**

Dr Suzanne Little (Ext:700 6052)

Dr. Brian Lee

**TIME ALLOWED:** 3 Hours

**INSTRUCTIONS:** Answer 4 questions. All questions carry equal marks.

**PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO**

The use of programmable or text storing calculators is expressly forbidden.

Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

***Requirements for this paper (Please mark (X) as appropriate)***

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

*Log Tables*

*Graph Paper*

*Dictionaries*

*Statistical Tables*

*Bible*

|  |
| --- |
|  |
|  |
|  |
|  |

*Thermodynamic Tables Actuarial Tables*

*MCQ Only – Do not publish Attached Answer Sheet*

***QUESTION 1 - DATA MANAGEMENT [TOTAL MARKS: 25]***

**Q 1(a) [10 Marks]**

Describe a data analytics pipeline or lifecycle (e.g., for research or e-commerce) and briefly state the purpose for each stage, identifying a common tool for use in each

stage.

A data analytics pipeline is a series of steps that are followed to collect, process, analyze, and present data. The purpose of a data analytics pipeline is to transform raw data into useful insights.

A generic data analytics pipeline can be described with the following stages: **Gathering, Processing, Analysing, Presenting, and Preserving.**1234

Here is an example data analytics pipeline for research, with a common tool for each stage:

●

**Gathering:** This stage involves collecting data from various sources. **Common tools for this stage include surveys, sensors, and databases.**56

●

**Processing:** In this stage, collected data is cleaned, transformed, and prepared for analysis. **OpenRefine (GoogleRefine) and Python (Juypter notebooks) are common tools used for data processing.**7

●

**Analysing:** This stage involves exploring and analyzing the data to extract meaningful insights and patterns. **Common tools for data analysis include Python libraries such as Matplotlib, Bokeh or Seaborn, and Spreadsheets (Excel, GoogleSheets).**7

●

**Presenting:** This stage focuses on presenting the findings of the analysis in a clear and understandable way. **Common tools for data presentation include Jupyter, Tableau, and spreadsheets for exploratory visualization.**7

●

**Preserving:** This stage involves storing and managing the data for future use. **Relational databases, document databases (e.g., MongoDB), and linked data with metadata are some common tools used for data preservation.**

**Q 1(b) [7 Marks]**

Given the following brief to design a system for a data collection task, list 3 important questions you would ask and suggest a type of database system to use giving a

reason for your choice.

“A major supermarket chain wants to implement a loyalty card system collecting customer purchases, demographics, preferences and calculating a rewards points system. The gathered data will be used to determine stock requirements for specific stores and for creating advertising campaigns.”

1.

**"How much data do I have now? What rate will I get new data?"**8 This question helps to determine the scale of the system and choose appropriate technologies.

2.

**"Is the data structured? What format is the data?"**8 This question will help determine the best way to store and process the data.

3.

**"How many queries will be run? Will they be concurrent? How many users?"**8 This question helps assess the system's performance requirements and design for efficient querying.

A **traditional relational database**, such as Oracle, MySQL, or PostgreSQL, would be suitable for this task.910

●

Relational databases are well-suited for storing structured data, such as customer purchases, demographics, and preferences.9

●

They allow for efficient querying and analysis of the data, which is important for determining stock requirements and creating targeted advertising campaigns.

**Q 1(c) [4 Marks]**

Define “metadata”, what it is used for and provide an example of how it is used.

## Definition and Uses of Metadata

**Metadata is data that provides information about other data.** It describes and defines the content of data, making it easier to find and use. [Q2(b)] Metadata can enhance data quality by promoting consistency, reducing ambiguity, and enabling better data discovery and integration. [Q2(b)]

**Uses of Metadata:**

* **Discovery:** Metadata helps users locate relevant data by providing descriptive information about the data's content, format, and other characteristics. [Q2(b)] For example, metadata for a research dataset might include the title, author, keywords, and abstract, making it easier for researchers to find relevant studies.
* **Management:** Metadata helps organizations manage and organize their data assets by providing information about data lineage, ownership, and access rights. [Q2(b)] For example, metadata can track the origin, modifications, and responsible parties for a data file, ensuring data integrity and accountability.
* **Interoperability:** Metadata enables data exchange and integration between different systems and applications by providing a common understanding of the data's structure and meaning. [Q2(b)] For example, using standardized metadata schemas allows different databases to share information seamlessly.

**Example of Metadata Usage:**

Consider an image file of a family vacation photo. Some metadata examples for this image file include:

* **File Name:** FamilyVacation2023.jpg (structural metadata)
* **Date Taken:** 2023-08-15 (administrative metadata)
* **Location:** Dublin, Ireland (descriptive metadata)
* **Camera Model:** Canon EOS R5 (descriptive metadata)
* **Image Size:** 5472 x 3648 pixels (structural metadata)

This metadata helps users understand the context of the image, search for it based on keywords, and manage it within a photo library.

**Q 1(d) [4 Marks]**

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

explain 2 potential problems that may arise in either making data open or using open data.

Open data refers to datasets that are made freely available to anyone to access and use. While open data has many benefits, there are also potential problems that may arise in either making data open or using open data:

1.

**Privacy concerns:** Making data open can raise privacy concerns if it contains personal or sensitive information. Even if data is anonymized, there is always a risk of re-identification, especially when combined with other datasets.13

2.

**Data quality issues:** Open datasets may contain errors, inconsistencies, or biases that can affect the reliability of analyses and conclusions. Users of open data need to be aware of these potential issues and take steps to validate the data before using it.

***[End of Question1]***

***QUESTION 2 - BIG DATA [TOTAL MARKS: 25]***

**Q 2(a) [8 Marks]**

Big Data is characterised by Volume, Variety, Velocity and, recently, Veracity. Explain what each of these terms mean in relation to Big Data.

**Volume:** This refers to the **sheer amount of data**. Big data involves datasets that are so large and complex that **traditional data processing methods are inadequate**.12

●

**Variety:** Big data comes in **many different forms, including structured, semi-structured, and unstructured data.** Examples of data variety include log files, social media content, photographs, microblogs, surveys, news, CCTV video, movies, television, sales records, clicks, adwords statistics, audio recordings, playlists, search terms, sensors, pedometer/activity monitor, spectrographs, microscopy, and genomes.23

●

**Velocity:** This refers to the **speed at which data is generated and collected.** Big data is often characterized by **real-time or near real-time data streams.**2

●

**Veracity:** This refers to the **quality and trustworthiness of data.** Big data often comes from a **variety of sources, and the quality of the data can vary widely.** Ensuring the veracity of big data is crucial for making sound decisions.

**Q 2(b) [6 Marks]**

Map-Reduce can be used to process large amounts of data (e.g., text). Explain, with the aid of a simple word-count example, how map-reduce works from a programming model point of view.

Here's how MapReduce works for a simple word-count example:

1.

**Map Phase:** The input text is divided into smaller chunks, and each chunk is processed by a **"mapper" function.** The mapper function **reads each word in the chunk and emits a key-value pair, where the key is the word and the value is 1.** For example, if the chunk contains the sentence "the quick brown fox jumps over the lazy dog", the mapper function would emit the following key-value pairs:

○

(the, 1)

○

(quick, 1)

○

(brown, 1)

○

(fox, 1)

○

(jumps, 1)

○

(over, 1)

○

(the, 1)

○

(lazy, 1)

○

(dog, 1)

2.

**Shuffle Phase:** The key-value pairs emitted by the mappers are **grouped by key.** This means that all the key-value pairs with the same word are sent to the same **"reducer" function.**

3.

**Reduce Phase:** The reducer function **receives all the key-value pairs for a given word and sums the values.** This gives the total count of each word in the input text. For example, the reducer function would receive the following key-value pairs for the word "the":

○

(the, 1)

○

(the, 1)

○

It would then sum the values to get a total count of 2 for the word "the".

**Q 2(c) [4 Marks]**

What is meant by a Data Warehouse? What are some of the advantages that one provides to a business?

A **data warehouse** is a **centralized repository** of data that is designed for **reporting and analysis**. It **stores data from multiple sources** that is transformed into a **common, multidimensional data model**.5

Data warehouses offer several advantages to businesses, including:

●

**Improved decision-making:** Data warehouses provide a single source of truth for data, which can be used to make more informed decisions.

●

**Enhanced business intelligence:** Data warehouses allow businesses to track key performance indicators (KPIs) and identify trends.

●

**Increased efficiency:** Data warehouses can automate reporting and analysis tasks, freeing up analysts to focus on more strategic work.

**Q 2(d) [3 Marks]**

Big data is often varied and unstructured. Explain what this means and why it makes it difficult to store big data in a traditional relational database.

Big data is often **varied** and **unstructured**, meaning that it **comes in many different forms** and **doesn't fit neatly into the rows and columns of a traditional relational database**.67

Examples of unstructured data include text documents, images, video, and audio.6

**Storing big data in a traditional relational database can be difficult for several reasons:**

●

**Schema rigidity:** Relational databases require a predefined schema, which can be challenging to define for unstructured data.

●

**Scalability limitations:** Relational databases can struggle to handle the volume and velocity of big data.

**Q 2(e) [4 Marks]**

Identify the 4 main standards for linked data, giving the full name as well as the acronym.

The sources do not list the four main standards for linked data. The following information is not from the sources and may need to be independently verified:

The four main standards for linked data are:

●

**Resource Description Framework (RDF):** RDF is a standard for representing data as a graph.

●

**RDF Schema (RDFS):** RDFS provides a vocabulary for describing the classes and properties of resources in RDF.

●

**Web Ontology Language (OWL):** OWL is a more expressive language for defining ontologies, which are formal representations of knowledge.

●

**SPARQL Protocol and RDF Query Language (SPARQL):** SPARQL is a query language for retrieving and manipulating data stored in RDF.

***[End of Question2]***

***QUESTION 3 - DATA QUALITY [TOTAL MARKS: 25]***

**Q 3(a) [5 Marks]**

Using an example, explain what data quality is and why it can have important financial or social implications for data management systems.

**Data quality** refers to the **reliability, accuracy, completeness, and consistency of data.** High-quality data is free from errors and artefacts. An **error** is data that is missing or lost due to the capture process and cannot be recovered. An **artefact** is something that has been introduced into the dataset during the gathering, processing, integration or cleaning activities1. **Poor-quality data can have significant financial or social implications for data management systems.**

**Example:** Consider a financial institution that uses a data management system to track customer transactions. If the data quality is poor, the institution may make incorrect decisions about loan approvals, risk assessments, or fraud detection. This could lead to financial losses for the institution and damage its reputation.

**Q 3(b) [6 Marks]**

Identify 3 different sources of data and list a potential point in each collection or import process where data quality can be effected.

Here are three different sources of data and potential points in each collection or import process where data quality can be affected:

●

**Surveys:** During data entry, **typos** or **incorrect data formatting** can occur, leading to inaccurate data. For example, a respondent might accidentally enter their age as "100" instead of "10" or use inconsistent date formats.

●

**Sensors:** **Sensor malfunctions** or **calibration errors** during data capture can introduce erroneous values into the dataset. For example, a temperature sensor that is not properly calibrated might consistently report temperatures that are too high or too low.

●

**Databases:** **Data integration issues** during the import process can lead to inconsistencies or duplicates in the dataset. For example, if two databases have different formats for storing customer addresses, merging them without proper data transformation can result in duplicate or incomplete records.

**Q 3(c) [4 Marks]**

Where errors occur in the data collection phase, what are the two types of solutions that can help improve quality? Give an example of how one of those solutions could be implemented.

Two types of solutions can help improve data quality when errors occur in the data collection phase:

●

**Process mandates:** These are **rules or procedures that are put in place to prevent errors from occurring in the first place**.

●

**Data cleaning tools:** These are **software programs that can be used to identify and correct errors in data.**

**Example of implementing a process mandate:** A company could implement a process mandate that requires all data entry staff to undergo training on data quality standards and procedures2. This training would cover topics such as data validation, data formatting, and error detection and correction.

**Q 3(d) [5 Marks]**

How can data quality be measured? Give an example of implementing or using a data quality measurement method.

Data quality can be measured using a variety of methods, including:

●

**Conventional measures:** These include **accuracy, completeness, uniqueness, timeliness, and consistency**3. However, there are **problems with applying these measures to real-world data**, which will be further explored in the next query response.

●

**Test cases:** Test cases involve **creating a set of data with known results and then running the data through the data management system.** This can help identify any glitches or errors in the system4.

**Example:** A company could implement a data quality measurement method that involves **tracking customer complaints related to data errors**. This could be done by setting up a system for customers to report data errors and then analyzing the complaints to identify patterns and trends. This information could then be used to improve data quality processes and reduce the number of errors in the future4.

**Q 3(e) [5 Marks]**

Explain how a tool such as OpenRefine (GoogleRefine) can be used to help data quality.

**OpenRefine (formerly Google Refine)** is a powerful **data cleaning tool** that can be used to improve data quality. It is an offline (desktop) tool that runs in the browser5. OpenRefine offers several features for improving data quality:

●

**Data exploration:** Users can **explore data using graphs to check distributions and outliers, which can help identify potential data quality issues**5.

●

**Cleaning and transformation:** OpenRefine allows users to **fix errors in fields and use heuristics to group data and spot errors**5.

●

**Data enhancement:** OpenRefine can **call services to enhance data, such as geocoding** (converting addresses into geographic coordinates)5. This can be helpful for adding valuable information to a dataset and improving its accuracy.

●

**Handling large datasets:** OpenRefine can **handle moderately large datasets**, making it a useful tool for cleaning and preparing data for analysis6.

***[End of Question3]***

***QUESTION 4 - VISUALISATION DESIGN [TOTAL MARKS: 25]***

**Q 4(a) [8 Marks]**

In the appendix, Figure 1 represents the sales of a product for the years 2006-2012. Identify and explain at least 4 problems with this chart.

**Q 4(b) [8 Marks]**

Given the following visualisation tasks, suggest an appropriate graph type for each to display the information and give a brief justification.

1. Understand the popularity of Apple products (e.g., based on sales of iphone, ipad, ipod, smart watch, etc.)?

2. Show the grades of students in CA682?

3. Summarise all the grades of all students across all modules starting with CA?

4. Show employment rate by industry in Ireland?

**Q 4(c) [6 Marks]**

In the appendix, Figure 2, shows some data. Sketch an appropriate visualisation design and list at least 3 key visual design decisions (e.g., choice of colour, scale, layout, interactivity, animation, etc.).

**Q 4(d) [3 Marks]**

In the west, red and green have cultural significance. Consequently these colours

tend to be used significantly on visualisations. Why is this a problem? Are there any tools to help solve this problem?

***[End of Question4]***

***QUESTION 5 - HUMAN ATTENTION AND INTERACTION [TOTAL MARKS: 25]***

**Q 5(a) [5 Marks]**

In visualisation it is important to understand pre-attentive processing. Explain what a pre-attentive feature is and describe an experiment to determine if a feature is pre- attentive or not.

**Pre-attentive processing** refers to the rapid and subconscious processing of visual information that occurs **before conscious attention is directed to a specific object or area.**1 **Pre-attentive features** are visual properties that are detected very rapidly (less than 200-250ms) and "pop out" from their surroundings.1 These features are immediately perceived and can mislead the viewer.1

**Experiment to Determine Pre-attentiveness:**

A common experiment to determine if a feature is pre-attentive involves presenting participants with a visual display containing a target object among a set of distractor objects.2 Participants are asked to identify the target as quickly as possible.

●

**Pre-attentive Feature Condition:** If the target differs from the distractors by a pre-attentive feature (e.g., color, shape, orientation), participants should be able to identify it quickly and effortlessly, regardless of the number of distractors.

●

**Non-Pre-attentive Feature Condition:** If the target differs from the distractors by a non-pre-attentive feature (e.g., a conjunction of color and shape), participants' reaction time should increase as the number of distractors increases. This is because they need to consciously search for the target among the distractors.

**Q 5(b) [6 Marks]**

Define 2 visual communication goals and give a key design principle for each.

Here are two visual communication goals and a key design principle for each:

●

**Information:** The goal is to present structured data clearly and credibly.3 A key design principle is to prioritize **structure**.3 This involves using clear layouts, appropriate levels of detail, and colors that enhance credibility.

●

**Persuasion:** The goal is to use information to elicit a particular response, such as in advertising.3 A key design principle is to **research the target audience**.4 Understanding the audience's values, opinions, and attitudes allows designers to use illustrations, themes, colors, and groupings that effectively attract the viewer's eye and persuade them.

**Q 5(c) [8 Marks]**

Below are some Gestalt Principles. Choose 4 and explain how they can be used in your visualisation designs.

Proximity Similarity Continuity Symmetry Closure

Figure and Ground

The following Gestalt Principles can be used in visualisation designs:

●

**Proximity:** Elements that are close together are perceived as a group.5 This can be used to **group related data points together** or to **highlight an element by separating it from other elements**.

●

**Similarity:** Elements that share similar visual characteristics, such as color, shape, or size, are perceived as a group.5 This can be used to **represent different categories of data** or to **create a visual hierarchy**.

●

**Closure:** The human brain tends to perceive incomplete shapes as complete.5 This can be used to **create a sense of unity in a visualisation** by implying connections between elements, even if they are not explicitly connected.

●

**Figure and Ground:** This principle refers to the relationship between an object (the figure) and its background (the ground).5 A strong figure-ground relationship makes the visualisation easier to understand. This can be used to **highlight important data points by making them stand out from the background**.

**Q 5(d) [6 Marks]**

Explain what D3.js is and give 2 examples of things it doesn't do. Explain the difference between rules and selectors (CSS) and how they are used in D3.js.

**D3.js (Data-Driven Documents)** is a **JavaScript library for creating interactive and dynamic data visualisations in web browsers**.6 It uses web standards such as HTML, SVG, and CSS to manipulate the Document Object Model (DOM) and create visual representations of data.

**Limitations of D3.js:**

The sources provide two examples of things D3.js doesn't do:

●

**Data processing and analytics:** D3.js is primarily a visualisation tool and is not designed for complex data processing or analysis tasks. You would typically perform these tasks in a separate environment (e.g., Python with Pandas or R) before using D3.js for visualisation.6

●

**Handling bitmaps:** D3.js primarily works with vector graphics (SVG) and does not natively handle bitmap images (e.g., PNG, JPEG). However, there are workarounds, such as using other JavaScript libraries like Leaflet.js for integrating maps or images into D3.js visualisations.6

**Rules and Selectors in D3.js (CSS)**

**CSS (Cascading Style Sheets)** is used in D3.js to **style and visually represent data elements**.7

●

**Selectors** are used to **target specific HTML or SVG elements within the DOM**.7 For example, to select all circle elements, you would use the selector "circle".

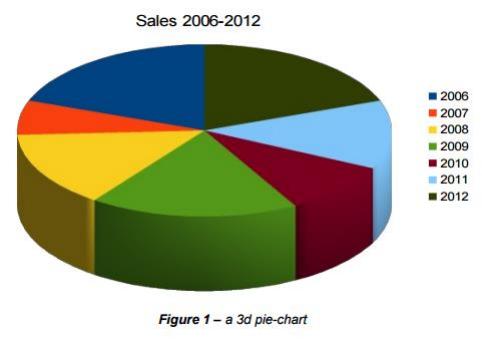
●

**Rules** define the **styles that will be applied to the selected elements**.7 They consist of a property (e.g., "fill", "stroke", "font-size") and a value (e.g., "red", "2px", "12pt").

In D3.js, **selectors are used to bind data to specific elements, and rules are then applied to style those elements based on the bound data**. This allows for dynamic and data-driven styling of visualisations.

***[End of Question5]***

***APPENDICES***





*Figure 2 - Data sample*

***[END OF APPENDICES] [END OF EXAM]***