INTERNATIONAL UNIVERSITY OF EAST AFRICA

'Learning to Succeed'

FACULTY OF SCIENCE AND TECHNOLOGY

END OF SEMESTER EXAMINATIONS - APRIL 2025

PROGRAMME: MIT

YEAR/SEM: YEAR 2/SEMESTER 1

COURSE CODE: MIT812

NAME: BIG DATA ANALYTICS

DATE: 2025-04-14

TIME: 2:00-5:00PM

INSTRUCTIONS TO CANDIDATES:

1. Read the instructions very carefully
2. The time allowed for this examination is STRICTLY three hours
3. Read each question carefully before you attempt and allocate your time equally between all the Sections
4. Write clearly and legibly. Illegible handwriting cannot be marked
5. Number the questions you have attempted
6. Use of appropriate workplace examples to illustrate your answers will earn you bonus marks
7. Any examination malpractice detected will lead to automatic disqualification.

DO NOT WRITE ANYTHING ON THE QUESTION PAPER

Section A Section A is Compulsory

Question 1:

QUESTION TWO (40 MARKS)

The Prescriptions-R-X chain of pharmacies has offered to give you a free lifetime supply of medicines if you design its data warehouse. Given the rising cost of health care, you agree. Here's the information that you gather:

- Patients are identified by an SSN, and their names, addresses, and ages must be recorded.
- Doctors are identified by an SSN. For each doctor, the name, specialty, and years of experience must be recorded.
- Each pharmaceutical company is identified by name and has a phone number.
- For each drug, the trade name and formula must be recorded. Each drug is sold by a given pharmaceutical company, and the trade name identifies a drug uniquely from among the products of that company. If a pharmaceutical company is deleted, you need not keep track of its products any longer.
- Each pharmacy has a name, address, and phone number.
- Every patient has a primary physician. Every doctor has at least one patient.
- Each pharmacy sells several drugs and has a price for each. A drug could be sold at several pharmacies, and the price could vary from one pharmacy to another.
- Doctors prescribe drugs for patients. A doctor could prescribe one or more drugs for several patients, and a patient could obtain prescriptions from several doctors.
- Each prescription has a date and a quantity associated with it. You can assume that if a doctor prescribes the same drug for the same patient more than once, only the last such prescription needs to be stored.
- Pharmaceutical companies have long-term contracts with pharmacies. A pharmaceutical company can contract with several pharmacies, and a pharmacy can contract with several pharmaceutical companies. For each contract, you have to store a start date, an end date, and the text of the contract.
- Pharmacies appoint a supervisor for each contract. There must always be a supervisor for each contract, but the contract supervisor can change over the lifetime of the contract.

Task:

a. Based on the information above, identify the Subject dimensions, and facts (metrics **), then d**escribe the hierarchies and categories to be included in the information packages for each dimension. **(05 Marks)**

b. List at least FIVE business metrics or facts for this case. **(05 Marks)**

c. Draw the information Package Diagram for the case above. **(10Marks)**

d. Draw an Entity Relation Diagram for the case above. **(10Marks)**

e. Based on the derived Information package, draw a Dimensional Model for the case above. **(10 Marks)**

Section B Attempt any Three Questions from Section B

Question 1:

QUESTION ELEVEN:

i.A multinational corporation wants to strengthen its cybersecurity measures by detecting potential threats before they cause harm. How can big data analytics be used for real-time threat detection and response?Â Â Â Â Â

Â Â
Â Â Â Â Â Â Â Â Â Â Â Â Â Â Â **(10 Marks)**
ii.A large farm is implementing IoT sensors to monitor soil moisture, weather conditions, and crop health. How can big data analytics help farmers make better decisions to improve crop yield?
Â Â
Â Â
(10 Marks)

Question 2:

QUESTION EIGHT:

i.A meteorological agency wants to improve weather predictions and disaster response. How can big data analytics be used to analyze climate patterns and provide early warnings? **(10 Marks)**

ii. An e-learning platform wants to offer personalized learning paths for students based on their performance and engagement levels. How can big data analytics be used to analyze student behavior and recommend customized courses? **(10 Marks)**

Question 3:

QUESTION FIVE:

i.An online marketplace wants to improve customer engagement through personalized product recommendations. How can big data analytics enhance the customer shopping experience? **(10 Marks)**

ii.A power company wants to predict electricity demand and optimize energy distribution. How can big data analytics help in balancing supply and demand while reducing costs? **(10 Marks)**

Question 4:

QUESTION TWO:

i. You have a **large dataset (10 million rows)** containing sensor readings (sensor_data.csv). Using **data.table**, write an R script to:

  a. Compute the average sensor value per device_id. **(05 Marks)**

  b. Filter rows where sensor_value > threshold. **(05 Marks)**

i. A telecom company wants to analyze call durations. Using the dataset (calls.csv) with columns Call_ID, Customer_ID, Duration, Call_Type, write an R script to:

  a. Create a histogram of call durations. **(05 Marks)**

  b. Generate a bar chart showing the total number of calls by Call_Type. **(05 Marks)**

Question 5:

QUESTION FOUR:

i. A telecom company wants to predict customer churn using a dataset (churn_data.csv) with features Age, Subscription_Length, Monthly_Charges, Churn (Yes/No). Write an R script to:

  a. Train a **logistic regression model** to predict churn. **(05 Marks)**

  b. Evaluate the model's accuracy. **(05 Marks)**

ii. A marketing team wants to segment customers based on spending behavior using a dataset (customer_spending.csv) with Customer_ID, Annual_Income, Spending_Score. Write an R script to:

a. Perform **k-means clustering** to find 3 customer segments. **(05 Marks)**

b. Visualize the clusters using a scatter plot **(05 Marks)**

Question 6:

QUESTION ONE:

i. You are given a large dataset containing customer transactions (transactions.csv), but it has missing values in the Amount and Category columns. Write an **R script** to:

a) Replace missing numeric values with the column mean. **(05 Marks)**

b) Replace missing categorical values with the most frequent category **(05 Marks)**

ii. A retail company has a dataset (sales_data.csv) containing sales transactions with columns: Customer_ID, Product, Category, Price, Quantity, Date. Using **dplyr**, write an R script to:

a. Calculate total revenue per category. **(05 Marks)**

b. Filter transactions where Price > 100. **(05 Marks)**