

CSL 430: Business Intelligence

Final Examination

20th November 2023

Maximum Marks: 60

Time: 180 Minutes

Instructions:

1. Please be very specific with your answers.
2. Make suitable assumptions wherever necessary and state your assumptions clearly.
3. Please explain the reasoning behind your answers in all the MCQs. You will not get any marks for just writing the options.

Q1. State True or False with Justification.

(8 x 1Mark = 8 Marks)

1. When the size of a database increases beyond a specific volume, it is called a data warehouse.
2. If two datasets have same mean and same standard deviation, they are not necessarily similar.
3. Inventory is a semi-additive measure as it can be aggregated across all dimensions except time.
4. Pie chart is not the right choice of visualization to express proportions as many people find it difficult to compare sectors of circles.
5. A Standard Normal Distribution has a mean of 1 and standard deviation of 0.
6. The choice to use ETL or ELT is based only on the business use case. Either of the two can be used based on choice of the Data Warehouse Architect.
7. Data Visualization best practice suggests Data Ink Ratio should always be minimized.
8. A Modern Data Stack is not as good at managing BI reporting use cases as a Traditional Data Warehouse.

Q 2. After each of the statements below, there are multiple options provided. Please choose all the options that are correct in the context of the statement. For example, in the question below correct choices will be A and C:

(Any 4. 4 x 2 Marks=8 Marks)

VNIT is _____.

- A. At Nagpur B. A medical college C. A Technology institute D. A place of worship

Please explain your choice of options. Partially correct answers may not get any marks.

1. Online Transaction Processing Systems are:

- A. Optimized for retrieval B. Optimized for storage C. Optimized for both retrieval and storage
D. Contain only structured data E. Contain only unstructured data

2. A Modern Data Stack _____.

- A. Supports Real Time Analytics B. Can not be hosted in cloud C. Has multiple layers
D. Supports only Snowflake schema E. Supports Data Visualization

3. In the context of a Bank, following are slowly changing dimensions (SCDs):

- A. Customer ID allocated to the customer B. Marital Status of the customer
C. Address of the customer D. Customer Segment – Gold/Silver/Bronze
E. Account Balance of the Customer

4. All of the following are examples of structured data except:

- A. Server Logs B. Student data in a SQL database C. JSON objects
D. Data from IoT sensors E. Sales data in CSV file

5. Choose all the correct statements about Inventory

- A. It is a physical place where goods are stored B. It is maintained to tackle sudden changes in demand
C. Lead has no impact on Inventory levels D. More inventory has to be maintained if lead time is more
E. Less inventory has to be maintained if lead time is less.

Q3. Look at the images/data below and point out mistakes if any. If you find mistakes provide corrections to correct these mistakes. If there is no mistake please state that clearly. (4 x 2 Marks = 8 Marks)

1. Table below shows the answers to match the pairs question solved by a student of BI.

#	Column A	Column B
1	Relational DBMS	Are always ACID Compliant
2	Data Mart	Data about Data
3	Data Lake	Supports structured as well as Unstructured data
4	Data Warehouse	Best suited for reporting on structured data
5	Data Lakehouse	Combines the best of Data Lake and Data Warehouse
6	Metadata	Departmental Data

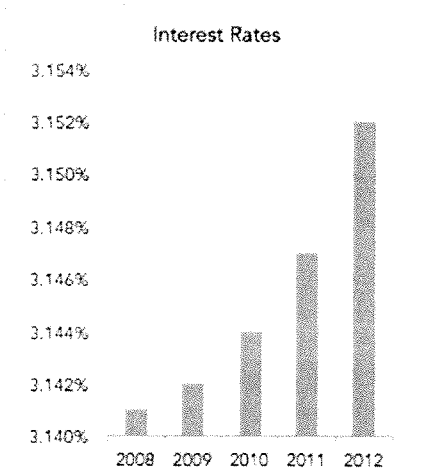
2. An attribute is defined as a unique level within a dimension. For example, Timestamp is an attribute in the Time Dimension.

3. Levels of BI Progression are as follows:

BI Progression



4. Look at the bar chart of Interest rates in the US between the years (2008-2012):



8.4

Q 3. Answer the following questions.

(Any 5. 5 x 4 Marks = 20 Marks)

1. Look at the table below and find out which of the following are facts. For each of the facts write which will be additive, semi-additive and non-additive and why?

Date	Store	Customer	Product	Product Description	Sales in Rs.	Discount	Inventory Balance	Count of Employees
------	-------	----------	---------	---------------------	--------------	----------	-------------------	--------------------

2. What is the explain the concept of granularity of data. How does it determine the ability to roll up or drill down? Explain with an example.

3. Look at the sample data for Final exam conducted by VNIT in the table below. Describe four actionable analyses that you can perform on this data. Assume you have 10,000+ rows of data. Example of analyses – Which branch took the most and least time to complete the exam?

Roll No.	Gender (M/F)	Question (Set Type)	Start Time HH:MM	End Time HH:MM	Section A Score	Section B Score	Total Score (A+B)	Disruption (Y/N)
B1891	F	3-A	09:00	10:21				N
C1901	M	3-B	12:00	13:02				N
C1422	F	4-C	14:30	14:39				N
D1901	F	1-A	15:00	16:30				Y

- The first letter of the roll number indicates branch – there are only 4 branches Architecture, B - Building & Civil Engineering , C – Computer Science, D – Designing and Electronics Engineering.
- The next two digits indicate the year of admission and other two digits indicate the roll number.
- The digit in Question Set Type indicates the year and letter indicates the type of set. There are only 3 set types A,B,C.
- Disruption (Y/N) indicates if there was any power cut or internet disruption while the student was attempting the test.

4. Look at the following table about visitors to the magical island of Good-la-gooba. If this information is stored in a data warehouse, what will be the fact(s), dimension(s) and attribute(s) here? What more attributes are possible within the dimensions?

Number of visitors

Month	Age			Gender M/F
	<30	30-60	60+	
January	781	878	346	1184/821
February	1021	1022	420	1047/1416
March	1211	788	278	1111/1166

5. A product has a lead time of 12 days. Company mandates keeping an safety stock 5 days of demand for each product. What should be order amount if the daily sale of the product is 45 units and current inventory is 920?

- What will happen to the order amount if you choose to place an order after 5 days?
- What will happen to the order amount if daily sale becomes 70 units?

6. How does Predictive Modelling work? Explain with an example. Can Normal Distribution be used as a method for predictive analysis? Give reasons for your answer.

7. Why are companies moving to the Cloud Data Warehouse? Write 5 reasons why Cloud Data Warehouses make sense. Provide an example of a Cloud Data Warehouse.

Q 5. Study the Data Schema for a multinational restaurant . Answer the questions that follow. **(8 Marks)**

POS Bill	ERP System 1	Store DB	Employee DB
Bill No	Item ID (same as Order Code)	Store ID	Employee ID
Date	Item Description (Ordered item)	Store Name	Employee Name
Timestamp	Item Cost	Address Line 1	Address Line 1
Ordered Item	Sub-Category ID	Address Line 2	Address Line 2
Order Code	Sub-Category Desc.	City	City
Quantity	Category ID	State	State
Price	Category Desc.	Country	Country
Amount	Shelf Life	Pin Code	Pin Code
Service Tax	Storage Type ID	Region (as in APAC, EMEA, Americas)	Region (as in APAC, EMEA, Americas)
Vat	Related Items	Total Store Area	Joining Date
% Discount	Storage Type Desc.		Current Designation
Salesperson			Last Designation
Store ID			Salary
Store Type			Employment Type
			Bonus flag

1. What analyses can you perform based on this data to reduce the wastage of food? **(2 Marks)**
2. Which of the following analyses are possible? Explain the steps to perform those analyses. If they are not possible explain why they are not possible and what additional data will be required to make those analyses possible: **(Any 3. 3 x 2 Marks = 6 Marks)**
 - a. The restaurant chain wants to find out the most profitable item on its menu.
 - b. The restaurant decides to award a “Lambi race ka ghoda” award to its top 10 longest serving male employees.
 - c. The restaurant chain wants to order the restaurants by annual sales per unit area.
 - d. The restaurant chain wants to report total number of restaurant (stores) in each region.

Q.6. Answer with the help of a well labelled diagram/visualizations: **(Any 2. 2 x 4 Marks = 8 Marks)**

1.Explain any two of the following visualization charts with the help of a diagram. Mention in which case these visualizations can be used. For example – A multi-line chart can be used to compare two variables over time:

a. Scatter Plot, b. Bubble chart c. Heatmap.

2. Draw a schematic diagram of a Modern Data Stack showing different layers and components.

3. Draw a typical Star and Snowflake Schema. Mention the key difference between the two (2-3 sentences)

X _____ X _____ X

Note 1. All questions are compulsory.

2. It's an open book, open notes examination, already downloaded or physically available.
3. Verbatim answers from textbook/downloaded material will be ignored and will not be credited.
4. Draw the diagrams or sketches wherever necessary.
5. Please write your assumptions.

Q1 Analyse the following statements in Systems Engineering. Give examples in support of your answers. No word limit but precise answers will attract more credit. [CO2, CO4, CO5] [8]

- a) Systems Thinking is not thinking systematically.
- b) Modeling is a conceptual tool to define and grasp complex systems, offering structured representations of their interconnected components and behaviors. (For systems, value is benefit at cost.)
- c) Systems and environments are intricately interdependent, engaging in dynamic interactions that mutually shape and influence each other, while maintaining their distinct identities, structures, and functions.
- d) Optimizing performance requires mastering system complexity for adaptability in dynamic environments.

Q2 a) What emergence properties can be seen in the Amazon on-line shopping system portal? [1 1/2]
[CO4, CO5]
b) Identify the form and functions in the system for packaging of the CPUs using a robot arm. This system should be with conveyor belt and packaging unit based on the size of the CPU packet. It detects the size through optical scanners. [1 1/2]
[CO4, CO5]

Q3 a) Analyse the five issues faced which the System Thinkers face while designing the system in the context of the design of the system for managing startup incubation centre at your institute. This incubation centre will offer space, seed capital, technical know-how. It will also create a Technology & Science Park for the creating institute-industry interface. This will be utilised for the R&D activities jointly with internship programs for the students. Hint: This centre system will be extension to institute ERP system and will generate data like which companies are registered, their products, services etc. and their turnover. [5]
[CO2, CO4]

b) What are the issues involved in designing the Integrated Development Environment (IDE) for any programming language from Systems Engineering perspective? [3]
[CO2, CO4, CO5]
Domain Knowledge: An IDE is a comprehensive software tool that provides a unified interface for coding, debugging, testing, and deploying software applications, enhancing developers' efficiency and workflow.

- Q4 a) **Create** an Object Process Model with textual support for Elevator Control System. You can use the objects like elevators, floors, and passengers. [5][CO1, CO2, CO3]
Identify the processes and their states, and transitions like elevator movement, door operations, and response to calls by person manning the elevator.
 Your model should also have the response of the system to **constraint violations**.
 Also **write what seed** you used in initiating the modeling process and **why**?
- b) A System to do X by (means of) Y in order to achieve Z. [3][CO1, CO3]
 X – What the System does Y – How it does it
 Z – Why is it being done
 Write the X, Y, and Z for the purpose of defining the system's root definition for the sub-system your **Case Study submitted by your group**.
- Q5 Here is description of a system to be designed for the United Nations: [5]
 [CO2, CO3, CO4, CO5]
 In order to stop the wars and skirmishes going all over the world, United Nation decides to collect the on the spot information and get the opinions of the world leaders. This data will be in videos as well as reports and news. It decides to organize weekly assemblies at its headquarters, where world leaders will deliver talks on these incidents, which will be translated into different world languages and gist of the talk will be flashed on the website along with the data received.
 Write the sub-systems and major components required to design this system with diagrammatic representation.
- Q6 Introducing high-tech courses like AI and Bioinformatics in a technical university have brought up about complexity issues. [6]
 [CO1, CO2, CO3, CO4]
 Some of the major are as under:
 1. Intensifying an AI course may require additional computing resources and qualified instructors, which may affect the enrollment.
 2. Modifying the Bioinformatics curriculum may necessitate adjustments in related courses and faculty responsibilities.
 3. The introduction of AI courses might result in increased demand for specialized resources, affecting the university's overall academic landscape.
 4. Offering advanced courses may require prioritizing resources away from foundational courses, impacting overall program diversity.
 Categorise them appropriately and find the solutions for them from systems engineering point of view.
- Q7 What are the building blocks in a Data Center like that of Google or Amazon, which can be categorised into various systems? How are these systems interconnected? [2][CO2, CO4]

Note: 1. All questions are compulsory.

2. It's an open book, open notes, open lecture slides exam

3. As far as possible, please express yourself in your own words. **Verbatim answers from the book or reading material others publish will be ignored and will not carry any credit.**

4. Please write your assumptions wherever required.

Q1. With the advances in medical science, human organ replacement with either donated organs from other human bodies or artificially made organs is the long term objective of the medical fraternity. You are required to write a technical part of the proposal for the the system for making repositories of the various organs at the national level. (Hint: You need to devise a methodology for data management of the donors as well as recipients) [3]
[CO1,CO2,CO3]

Q2. Analyse the following statements on the Systems Engineering (each 200 words max): (any 4) [6]
[CO1,CO4,CO5]

- a) *Systems Engineering handles the complexity in the design and development of the systems by enhancing the Systems Thinking.*
- b) *Biological systems have synthesis as the system building functionality.*
- c) *Boundaries in computer systems are permeable.*
- d) *Feedback loops drive systems behavior.*
- e) *Requirements drive the long-term sustainability in the software systems.*

Q3. Academic Information Systems for technical institutions and universities are becoming complex due to changing demands of the students and parents financing the education. As a systems engineers what interfaces you can define as a set of services and (user) flows in this changed scenario for technical education institutes with data analytics for continuous improvement of the system. [6]
[CO1,CO2,CO3]

Q4. A System to do X by (means of) Y in order to achieve Z.

X – What the System does

Y – How it does it

Z – Why is it being done

Write the X, Y, and Z for the following systems, which state the purpose of defining the system's root definition.

- a) *Multi-speciality Hospital Management System*
- b) *Sell and logistics in a Mall Specialised in only sports goods and equipments*
- c) *Managing the online courses for a technical institution*
- d) *System for managing the maintenance of the residential facilities in a technical institution which include student hostel*

(Domain knowledge for these is already provided in the last semester SE course and this course)

[12]
[CO1,
CO2, CO4]

Q5. Give any example from the software systems you are aware of where you can apply the spectrum of failures from blameworthy to praiseworthy for designing the feedback system of that system.

[3]
[CO2, CO5]

Note: 1. All questions are compulsory.

2. It's an open book, open notes, open lecture slides exam

3. As far as possible, please express yourself in your own words. **Verbatim answers from the book or reading material others publish will be ignored and will not carry any credit.**

4. Please write your assumptions wherever required.

- Q1.** With the advances in medical science, multi speciality hospitals in the tertiary sector(specialised medical care involving advance and complex diagnostics) is growing in India. You are required to write a technical part of the proposal for the the system for making a regulatory framework for these hospitals at the national level. (Hint: You need to devise a methodology organization of the data from these hospitals. Any tertiary hospital will have any long term patients, staff, and other resources) [3]
[CO1,CO2,CO3]
- Q2.** Analyse the following statements on the Systems Engineering (each 200 words max): (any 4) [6]
[CO1,CO4,CO5]
- a) *Solutions to complex issues like Climate change can be found by enhancing the Systems Thinking approach to various solutions for this.*
 - b) *Photosynthesis is a system-building function in natural systems. (Photosynthesis converts sunlight,CO₂, and H₂O into glucose and oxygen.)*
 - c) *Real-time software systems like internet manage and adapt to permeable boundaries for interaction.*
 - d) *Understanding and analysing systems behavior is crucial for effective systems thinking.*
 - e) *Systems thinking can shape the requirements with long-term implication of software systems.*
- Q3.** Academic Information Systems for technical institutions and universities are becoming complex due to changing nature of technological and basic science research. Research funding soon will become very crucial for maintaining the standards of the education and attracting the students. As systems engineers what interfaces you can define as a set of services and flows towards various funding sources and for technical education institutes with data analytics for continuous improvement of the system. [6]
[CO1,CO2,CO3]

Q3. A System to do X by (means of) Y in order to achieve Z.

X – What the System does

Y – How it does it

Z – Why is it being done

Write the X, Y, and Z for the following systems, which state the purpose of defining the system's root definition.

- a) *Multiple Restaurant Management System*
- b) *Sell and logistics in a Mall Specialised in garments.*
- c) *Managing the hostels for a chain of technical institution*
- d) *System for managing the six-month technical institution which inch internship program for students of a technical institute catering to the needs of various industries all over the country.*

(Domain knowledge for these is already discussed in the class through various examples directly or indirectly taken during the regular teaching)

[12]
[CO1,
CO2, CO4]

Q4. Give any example of application of a spectrum of failures, ranging from blameworthy to praiseworthy, to evaluate and understand the performance and outcomes of a software-intensive systems.

[3]
[CO2, CO5]

Sessional – I Examination, September 2023
VII Semester - Computer Science and Engineering

Subject: - Intro. to IoT (CSL429)
Date/Time: - 29/09/2023 (3:00 – 4:30 PM)

Slot: - H
Max. Marks: - 25

Q1.

- a. For what purpose MQTT.fx or MQTT Explorer is used? Write all necessary steps to connect them with AWS IoT Core . (write either MQTT.fx or MQTT Explorer) CO2 (3)
- b. What was the limitation of ZigBee protocol stack? How it is addressed? CO3 (2)

Q2.

- a. Name the communication model used for (WMS) weather monitoring system? Why it is more appropriate than other models? Also draw the IoT deployment level used for WMS. CO1 (4)
- b. In IPv6 fragmentation is done at which layer? Why? CO1 CO2 (2)

Q3.

- a. Smart objects are constrained nodes. Why? CO2 (2)
- b. Give pros and cons of wireless based solutions. CO2 (3)

Q4.

- a. In traditional IT models, data is filtered, aggregate, stored and analyze in cloud/data center. Where all these processes took place in IoT data management and compute stack? Show diagrammatically the hierarchical nature of IoT system. CO1 CO2 (3)
- b. Match the Pairs CO-1 (6)

A	B
Bluetooth	WNAN (wireless neighborhood area network)
ZigBee	LPWA (low power wide area)
802.11 a/b/g/n/ac/ax	Cellular network
6LoWPAN	WLAN (wireless local area network)
SIGFOX	WHAN (wireless home area network)
3GPP	WPAN (wireless personal area network)

End Sem Examination
CSL 436
Information Retrieval
Winter 2023 (VII sem)
Slot F
Maximum Marks 50
Maximum Time: 3 hours

1. Consider the documents:
Doc1: "Environmental changes are impacting global weather patterns."
Doc2: "Weather forecasting has improved with the advent of advanced computational models."
Doc3: "Global warming is a significant factor in drastic environmental shifts."
Doc4: "Computational models help in predicting the effects of global warming."
a) Draw the term document incidence matrix for collection.
b) Draw the inverted index representation for this collection.
show the returned results for **global AND weather** (2+2=4)
2. Given a two-word query, the following is the postings lists for first term with skip pointers.
2, 5,20,35,67,90,112,135,145,260,267; with 2,67,145, as skip pointers and the posting list for second word with skip pointers is given below.
3,5,7,20,35,150,260,261,270,300; with 3,35,270 as skip pointers
a. How many comparisons would be done to intersect the posting list without skip pointers?
b. How many comparisons would be done to intersect both the posting list with skip pointers? (2+2=4)
3. Suppose you wish to find economic reports regarding the impact of oil extraction in the North Sea on the Scottish economy. A commercial document retrieval service offers the following suggested matches: the table shows how often some key phrases appear in each report.
North Sea oil Scotland economy
Report A 12, 0, 3, 24
Report B 10, 5, 20, 10
Report C 0, 12, 9, 8
Query 1, 1, 1, 1
Actually obtaining the reports will cost real money, so you would like to select the one most likely to be relevant. Your task now is to assess this using the cosine similarity measure.

(a) Write out the general formula for calculating the cosine of the angle between two 4-dimensional vectors (x1, x2, x3, x4) and (y1, y2, y3, y4).
(b) Use this formula to rank the three documents in order of relevance to the query according to the cosine similarity measure. What do you think of the results? (2+2=4)
4. Assume that the posting lists are gap encoded using γ codes. Using this encoding, suppose that the posting list for the term 'information' is the bit sequence:
1111 1111 1011 1100 1101 0011 1110 0000 0

And the posting list for the term 'retrieval' is the bit sequence:

1111 1111 1100 0000 0011 1011 1101 111

What doc-ids match the following query:

information AND NOT retrieval

(4)

5. What is term partitioning and document partitioning in distributed indexing? Differentiate between Block Sort based Indexing (BSBI) and Single Pass In Memory Indexing (SPIMI). (2+2=4)
6. A collection of 3000 documents, 3 documents are relevant to a query q : $REL\ q = \{A, B, C\}$ Below are rankings given by three systems for that query q . Only the first 5 documents are shown to the user. (Here we mark nonrelevant documents with "N".)

System 1	System 2	System 3
N	N	A
N	A	N
A	N	N
B	N	N
N	B	N

Which of the three systems would you prefer and why?

(4+1=5)

7. Answer the following questions in short. Some questions carry one mark and some carry 2 marks. Answer accordingly.
- a) What is the inverse document frequency (using log to the base 10) for a term that appears in 100 documents – with 10000 occurrences in these documents – in a text corpus of 1000000 documents? (1)
- b) Consider a collection with one billion tokens (i.e., with 10^9 tokens). Suppose the first 1,000 of these tokens results in a vocabulary size of 1,000 terms, and the first 100,000 tokens results in a vocabulary size of 10,000 terms. Use Heap's law to estimate the vocabulary size of the whole collection. (1)
- c) The following are entries of a positional index for the terms "peanut" and "butter", each showing the entries for one of the four documents 201, 202, 203, and 204. Which document contains the phrase "peanut butter"?
- document 201: <peanut: ... 201: 4, 30; ... > <butter: ... 201: 3, 28; ... >
document 202: <peanut: ... 202: 17; ... > <butter: ... 202: 13, 21; ... >
document 203: <peanut: ... 203: 8, 10; ... > <butter: ... 203: 9; ... >
document 204: <peanut: ... 204: 5, 15; ... > <butter: ... 204: 7, 8; ... > (1)
- d) What are the two assumptions behind modeling of the Binary Independence Model (BIM) and how do these assumptions help to reach to the BIM formula? (Need not do the complete derivation. Be precise.) (2)
- e) When (or in what situation) is the Normalized Discounted Cumulative Gain (NDCG) metric used/applicable? (1)

- f) What is the precision if a search engine retrieves 10 documents, 5 of which are relevant, from a corpus that contains in total 10 million documents, of which 100 are relevant? (1)
- g) A search engine is given the query “restaurants in Amsterdam”, upon which it returns 10'000 documents. In total, there are 100'000 documents that are indexed by the search engine, 4000 of which are relevant for the given query. Out of these relevant documents, 2000 are part of the returned results. What is the recall in this specific case? (1)
- h) Given a search engine that has a precision of 0.4 and a recall of 0.1, what is its F1-Measure? (1)
- i) What is the edit distance between “form” and “from” assuming the three operations insert, delete, and replace (but not transpose)? Use the dynamic programming algorithm and reach to the answer methodically. Do not guess the answer. (1)
- j) How do you estimate the size of the dictionary in an IR system? Give the formula. (1)
- k) Explain with a suitable diagram how cluster pruning is effective in IR? (1)
- l) Illustrate the mechanism of pseudo relevance feedback and implicit relevance feedback with suitable examples. (2)
- m) The Probabilistic Model of Information Retrieval is based on the foundation of the probability ranking principle. Justify in brief. (1)
- n) What do you mean by document likelihood model and query likelihood model in Language modeling? Which one is preferred over the other in the realistic setting and justify the reasoning for the same. (2)
- o) Latent Semantic Indexing (LSI) is observed to be effective as compared to the Vector Space Model (VSM). Justify this statement. (1)
- p) Why is HITS algorithm called as a mutually recursive algorithm? Give the outline of its pseudo code. (2)
- q) How/when does the page rank algorithm terminate? What is the termination mechanism called as? Describe the mechanism with an example. (1)
- r) Draw the BOW-tie structure of the web with all its components. Draw suitable arrows indicating how a surfer can surf from a part of the web to the other. Label the parts correctly. (1)
- s) The average in degree of all nodes in a snapshot of the web graph is 9. What can we say about the average out-degree of all the nodes in this snapshot? (1)
- t) How does index elimination mechanism help to improve the efficiency of the search mechanism? Illustrate with a suitable example. What are the heuristics used to do index elimination? (2)

CSL 374 Business Analytics 1st Sessional Exam.

21st February 2024

Maximum Marks: 30

Time: 90 Minutes

Instructions:

1. Please be very specific with your answers. Marks will be proportional to sense the answers make and not the length of the answers.
2. Make suitable assumptions wherever necessary and state your assumptions clearly.

Q1. State True or False with Justification. Correct answer without justification will not get any marks.
(Any Five) (5 Marks)

1. In statistical analysis of data terms Population and Sample can be used interchangeably as they refer to the same data set.
2. ERP, CRM, POS, HRMS, MIS can all be source systems for finding business insights.
3. Unless you have some historical data you cannot perform Predictive Analytics.
4. Order Processing, Payroll, Point-of-Sales (POS) are all examples of Transaction Processing Systems.
5. Tactical decisions have long term impact while strategic decisions have short term impact.
6. Break even analysis is a type of Predictive Analytics
7. In context of Business Intelligence ETL stands for Extensive Technical Literature.

Q 2. After each of the statements below, there are multiple options provided. Please choose all the options that are correct in the context of the statement. For example, in the question below correct choices will be A and C:
(5 Marks)

VNIT is _____.

- A. At Nagpur B. A medical college C. A Technology institute D. A place of worship

1. Which of the following is **always true** about a Linear Relationship between two variables (say x and y):

- A. If $x = 0$ then $y = 0$
B. y can be expressed as $y = mx + c$ (c constant)
C. If we know the slope we can predict values of x and y
D. x can be expressed as $x = ny + d$ (d constant)
E. If we know the value of one pair of x and y, we can find values of all other values of x and y

2. All the following about outlier is **wrong except**

- A. An outlier can skew the averages B. Outlier belongs to the same dataset as normal values
C. Outlier can never be detected D. Outlier seems to be generated using a different mechanism
E. An Outlier if from a different dataset while normal values belong to same dataset

3. Choose all options that **are NOT true** in case of a Normal Distribution :

- A. All distributions are normal distributions.
- B. A normal distribution is symmetrical about its mean
- C. A normal distribution looks like an inverted bell and hence is called as bell shaped distribution
- D. Every normal distribution has mean 0 and standard deviation 1
- E. Normal Distribution and Standard Normal Distribution are same.

4. Choose all options that **are true** about a statistical model :

- A. It should replicate a real life scenario
- B. can be made up of wax
- C. is useful only if we choose the right sample
- D. can often be imprecise
- E. always makes accurate predictions

5. A predictive model for a scenario has an equation

$$N = 3.4x + 3.2y - 1.45. \text{ where } x, y \text{ are variables}$$

Choose **inaccurate** statements of the following:

- A. R is a dependent variable
- B. R is an independent variable
- C. Relationship between R and x,y, is non-linear
- D. R is constant
- E. Value of R depends only on x and y

Q 3. Archis and Jai run a Samosa business. The fixed cost for running the business are as follows:

Head	Expense (Rs. Per month)
Rent	10,000
Utility Bills	4,000
Salaries	10,000
All other costs	3,000

The cost of ingredients for each piece of Samosa is as follows:

Ingredient	Cost in Rs.
Maida	2
Oil	3
Potatoes	1.5
Condiments	2.5
Fuel Cost	2
All Other Costs	1

- a. Plot a Break Even Analysis curve for their business. Assume that they sell Samosa at 30 per plate (each plate has 2 pieces)
- b. How many Samosas they need to sell per month (assume 30 days every month) to break even?
- c. What will happen to the break-even point if price of Potatoes triples (taking cost from 1.5 to 4.5)?

(2+2+2 = 6 Marks)

Q 4. What is Proximity or Distance Based Outlier Detection Method? On what principals is this method based? Enlist any 3 real life use cases where Outlier Analysis is used. **(4 Marks)**

Q5. Devesh is a business analyst with a large bank. The loan information system (LIS) of the bank has the following set of information fields available. Based on the available information which of the following analyses are possible for to Devesh present to his boss Yash? If an analysis is not possible, what more information should their colleague Genie make available to the team to make that analysis possible?

Assume there are 1001 records in LIS.

Information Available in LIS	Format	Information Available in LIS	Format	Information Available in LIS	Format
Loan Account Number	29448	Date of commencement of Loan	1 st April 2022	Rate of Interest	8.2%
Principal Amount of Loan	100000	Date of Completion of Loan	1 st March 2032	EMI Amount	16,000
Branch where the Loan was issued	Jalandhar	Number of EMI payments missed	0	Principal Outstanding as of Today	840,234
Name of Borrower	Avishkar Karnik	Address of Borrower	Pune	Current age of Borrower	22

Analyses required by Yash:

1. Top 10 loan account numbers by Principal Outstanding as of today.
2. Percentage of loans that have Rate of Interest more than 10%
3. How many loans are given to unmarried males.
4. Number of Loans where at least 1 EMI payments were missed
5. Which state issued most number of large loans? (A large loan is defined as loan with principal amount > 50,00,000)
6. What was the Median value of Principal amount?
7. What is the mean Principal outstanding for all loans issued at Jalandhar branch?
8. How many Loans are of tenure more than 15 years?

(10 Marks)

X _____ X _____ X

Mid-Sem Examination, Feb 2024
2nd Semester M.Tech. Computer Science

Subject : Advances In Data Mining And Data Analytics Course Code: CSL 544

Time: 1.5 HR.

Max Marks: 25

Note: All questions are compulsory. Use suitable example to answer questions.

Assume suitable data wherever necessary. It's an open notes examination.

Assume suitable data wherever necessary. Credits are reserved for analytical answers.

Que. 1 Mention and justify which Data Analytic/ Mining Functionalities is/are useful for the following task.

- (i) Smart City Planning (ii) Launch of new Pre-paid plan (2+2) [CO4,CO5]

Que. 2 Find similarities and dissimilarities of the following with suitable example: (6) [CO2,CO3]

- (i) Quantile and Quantile-Quantile Plot.
(ii) Bar Graph and Histogram.
(iii) Binning by Median and Binning by Boundaries

Que. 3. Test Marks out of 20 and students' Attitudes are given in some datasets. Demonstrate the use of Regression for finding out missing values and smoothing of data: (2+2)

Test Marks are 16,13,12,15,16,14,16,16,18,19 and

Attitude are 94,73,59,80,93,85,66,79,77,91

[CO3,CO4,CO5]

- (i) Find out missing values for Attitude when Test Marks are 15.5 and 11.5
(ii) How same line can be used to smoothing the data? Explain with a suitable example

Que. 4 Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result a

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(2+2+2) [CO3,CO4,CO5]

- (i) Calculate the central tendency of both features. State descriptive analytics for both features.
(ii) Find the outliers at least by two methods, if any.
(iii) Draw a scatter plot and write your observations about the plot.

Que. 5 Suppose we have the following two-dimensional data set:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

Consider the data as two-dimensional data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, and cosine similarity. (2) [CO1, CO2]

Que. 6 Resting heart rate is known to be 71 beats per minute on average, with a standard deviation of 4 beats per minute. A set of researchers believe that heart rate will increase in men when they are waiting to go into a job interview. To test this hypothesis, a group of 9 men attending job interviews are fitted with a wireless heart rate monitor to wear on their chest in the hour preceding their interviews. Their average heart rates over this hour are shown in the table below. (1+1+1) [CO1, CO2]

Participant	Heart rate (bmp)
1	80
2	74
3	73
4	72
5	78
6	75
7	70
8	74
9	69

- Should a z-test or a t-test be used to check if there is significant evidence to suggest heart rate increases in men while they are waiting to attend a job interview?
- Conduct the test at the 5% level and interpret your result.
- Calculate a 90% confidence interval for the population mean.

Mid-Sem Examination, Feb 2024

2nd Semester M.Tech. Computer Science

Subject : Advances In Data Mining And Data Analytics Course Code: CSL 544

Time: 1.5 HR.

Max Marks: 25

Note: All questions are compulsory. Use suitable example to answer questions.

Assume suitable data wherever necessary. It's an open notes examination.

Assume suitable data wherever necessary. Credits are reserved for analytical answers.

Que. 1 Mention and justify which Data Analytic/ Mining Functionalities is/are useful for the following task.

- (i) Smart City Planning (ii) Launch of new Pre-paid plan (2+2) [CO4,CO5]

Que. 2 Find similarities and dissimilarities of the following with suitable example: (6) [CO2,CO3]

- (i) Quantile and Quantile-Quantile Plot.
(ii) Bar Graph and Histogram.
(iii) Binning by Median and Binning by Boundaries

Que. 3. Test Marks out of 20 and students' Attitudes are given in some datasets. Demonstrate the use of Regression for finding out missing values and smoothing of data: (2+2)

Test Marks are 16,13,12,15,16,14,16,16,18,19 and

Attitude are 94,73,59,80,93,85,66,79,77,91

[CO3,CO4,CO5]

- (i) Find out missing values for Attitude when Test Marks are 15.5 and 11.5
(ii) How same line can be used to smoothing the data? Explain with a suitable example

Que. 4 Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result a

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(2+2+2) [CO3,CO4,CO5]

- (i) Calculate the central tendency of both features. State descriptive analytics for both features.
(ii) Find the outliers at least by two methods, if any.
(iii) Draw a scatter plot and write your observations about the plot.

Que. 5 Suppose we have the following two-dimensional data set:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

Consider the data as two-dimensional data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, and cosine similarity. (2) [CO1, CO2]

Que. 6 Resting heart rate is known to be 71 beats per minute on average, with a standard deviation of 4 beats per minute. A set of researchers believe that heart rate will increase in men when they are waiting to go into a job interview. To test this hypothesis, a group of 9 men attending job interviews are fitted with a wireless heart rate monitor to wear on their chest in the hour preceding their interviews. Their average heart rates over this hour are shown in the table below. (1+1+1) [CO1, CO2]

Participant	Heart rate (bmp)
1	80
2	74
3	73
4	72
5	78
6	75
7	70
8	74
9	69

- Should a z-test or a t-test be used to check if there is significant evidence to suggest heart rate increases in men while they are waiting to attend a job interview?
- Conduct the test at the 5% level and interpret your result.
- Calculate a 90% confidence interval for the population mean.

CSL 374 Business Analytics 1st Sessional Exam.

21st February 2024

Maximum Marks: 30

Time: 90 Minutes

Instructions:

1. Please be very specific with your answers. Marks will be proportional to sense the answers make and not the length of the answers.
2. Make suitable assumptions wherever necessary and state your assumptions clearly.

Q1. State True or False with Justification. Correct answer without justification will not get any marks. (Any Five) (5 Marks)

1. In statistical analysis of data terms Population and Sample can be used interchangeably as they refer to the same data set.
2. ERP, CRM, POS, HRMS, MIS can all be source systems for finding business insights.
3. Unless you have some historical data you cannot perform Predictive Analytics.
4. Order Processing, Payroll, Point-of-Sales (POS) are all examples of Transaction Processing Systems.
5. Tactical decisions have long term impact while strategic decisions have short term impact.
6. Break even analysis is a type of Predictive Analytics
7. In context of Business Intelligence ETL stands for Extensive Technical Literature.

Q 2. After each of the statements below, there are multiple options provided. Please choose all the options that are correct in the context of the statement. For example, in the question below correct choices will be A and C: (5 Marks)

VNIT is _____.

- A. At Nagpur B. A medical college C. A Technology institute D. A place of worship

1. Which of the following is *always true* about a Linear Relationship between two variables (say x and y):

- A. If $x = 0$ then $y = 0$
B. y can be expressed as $y = mx + c$ (c constant)
C. If we know the slope we can predict values of x and y
D. x can be expressed as $x = ny + d$ (d constant)
E. If we know the value of one pair of x and y , we can find values of all other values of x and y

2. All the following about outlier is *wrong except*

- A. An outlier can skew the averages B. Outlier belongs to the same dataset as normal values
C. Outlier can never be detected D. Outlier seems to be generated using a different mechanism
E. An Outlier is from a different dataset while normal values belong to same dataset

3. Choose all options that are *NOT true* in case of a Normal Distribution :

- A. All distributions are normal distributions.
- B. A normal distribution is symmetrical about its mean
- C. A normal distribution looks like an inverted bell and hence is called as bell shaped distribution
- D. Every normal distribution has mean 0 and standard deviation 1
- E. Normal Distribution and Standard Normal Distribution are same.

4. Choose all options that **are true** about a statistical model :

- A. It should replicate a real life scenario
- B. can be made up of wax
- C. is useful only if we choose the right sample
- D. can often be imprecise
- E. always makes accurate predictions

5. A predictive model for a scenario has an equation

$$N = 3.4x + 3.2y - 1.45. \text{ where } x, y \text{ are variables}$$

Choose **inaccurate** statements of the following:

- A. R is a dependent variable
- B. R is an independent variable
- C. Relationship between R and x,y, is non-linear
- D. R is constant
- E. Value of R depends only on x and y

Q 3. Archis and Jai run a Samosa business. The fixed cost for running the business are as follows:

Head	Expense (Rs. Per month)
Rent	10,000
Utility Bills	4,000
Salaries	10,000
All other costs	3,000

The cost of ingredients for each piece of Samosa is as follows:

Ingredient	Cost in Rs.
Maida	2
Oil	3
Potatoes	1.5
Condiments	2.5
Fuel Cost	2
All Other Costs	1

- a. Plot a Break Even Analysis curve for their business. Assume that they sell Samosa at 30 per plate (each plate has 2 pieces)
- b. How many Samosas they need to sell per month (assume 30 days every month) to break even?
- c. What will happen to the break-even point if price of Potatoes triples (taking cost from 1.5 to 4.5)?

(2+2+2 = 6 Marks)

Q 4. What is Proximity or Distance Based Outlier Detection Method? On what principals is this method based? Enlist any 3 real life use cases where Outlier Analysis is used. (4 Marks)

Q5. Devesh is a business analyst with a large bank. The loan information system (LIS) of the bank has the following set of information fields available. Based on the available information which of the following analyses are possible for to Devesh present to his boss Yash? If an analysis is not possible, what more information should their colleague Genie make available to the team to make that analysis possible?

Assume there are 1001 records in LIS.

Information Available in LIS	Format	Information Available in LIS	Format	Information Available in LIS	Format
Loan Account Number	29448	Date of commencement of Loan	1 st April 2022	Rate of Interest	8.2%
Principal Amount of Loan	100000	Date of Completion of Loan	1 st March 2032	EMI Amount	16,000
Branch where the Loan was issued	Jalandhar	Number of EMI payments missed	0	Principal Outstanding as of Today	840,234
Name of Borrower	Avishkar Karnik	Address of Borrower	Pune	Current age of Borrower	22

Analyses required by Yash:

1. Top 10 loan account numbers by Principal Outstanding as of today.
2. Percentage of loans that have Rate of Interest more than 10%
3. How many loans are given to unmarried males.
4. Number of Loans where at least 1 EMI payments were missed
5. Which state issued most number of large loans? (A large loan is defined as loan with principal amount > 50,00,000)
6. What was the Median value of Principal amount?
7. What is the mean Principal outstanding for all loans issued at Jalandhar branch?
8. How many Loans are of tenure more than 15 years?

(10 Marks)

X _____ X _____ X

Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology Nagpur
MidSem Examination

Course: **CSL 443 – System and Network Security**

BTech 7th Semester

Sept 2023

Time: 1 ½ Hrs

Max Marks: 20

Precisely answer all the questions.

Wherever applicable use R as (15 + last 3 digits of your Rollno).

For example, if Rollno is BT20CSE020 then $R = 15 + 020 = 035$.

1.	Suppose we are living in the Traditional symmetric cipher key age. In this world, there is a group of 58 friends. The scenario is, each friend wants to communicate with all the friends in the group using a cipher security.	CO2 CO3	
	a) How many keys are needed in order to perform this communication?		1
	b) Also, find how many keys are required by each friend?		1
2.	Let R_3 = last 3rd digit of R, R_2 = last 2nd digit of R, R_1 = last digit of R. Consider each R_i as a Hexadecimal character; thus 4 bits are used to represent each R_i . In Key expansion of AES-128 (10 rounds), if cipher key 1 st word " w_0 " bytes are as follows, $k_0 = E E$, $k_1 = R_3 R_3$, $k_2 = R_2 R_2$, and $k_3 = R_1 R_1$, and 4 th word " w_3 " bytes are same as " w_0 ", then compute the value of " w_4 " (that is the 1st word of round one key). Give answer in Hexadecimal format, that is 8 hexadecimal characters, without spaces. Where $ $ represents concatenation, if R is 035 then $w_0 = k_0 k_1 k_2 k_3 = EE003355$. Consider the following RCon and SubBytes tables for calculations,	CO1 CO2	3

Table 7.4 RCon constants			
Round	Constant (RCon)	Round	Constant (RCon)
1	(01 00 00 00) ₁₆	6	(20 00 00 00) ₁₆
2	(02 00 00 00) ₁₆	7	(40 00 00 00) ₁₆
3	(04 00 00 00) ₁₆	8	(80 00 00 00) ₁₆
4	(08 00 00 00) ₁₆	9	(1B 00 00 00) ₁₆
5	(10 00 00 00) ₁₆	10	(36 00 00 00) ₁₆

Table 7.1 SubBytes transformation table																
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	63	7C	77	7B	F2	6B	6F	C5	30	01	67	2B	FE	D7	A8	76
1	CA	82	C9	7D	FA	59	47	F0	AD	D4	A2	AF	9C	A4	72	C0
2	87	FD	93	26	36	3F	F7	CC	34	A5	E5	F1	71	D8	11	15
3	04	C7	23	C3	18	96	05	9A	07	12	80	E2	EB	27	82	75
4	09	83	2C	1A	1B	6E	5A	A0	32	3B	D6	83	29	E3	2F	84
5	53	D1	00	ED	20	FC	81	5B	6A	CB	BE	39	4A	4C	58	CF
6	D0	EF	AA	FB	43	4D	33	85	45	F9	02	7F	50	3C	9F	A8
7	51	A3	40	8F	92	9D	38	F5	BC	86	DA	21	10	FF	F3	D2
8	CD	0C	13	EC	5F	97	44	17	C4	A7	7E	3D	64	5D	19	73
9	60	81	4F	DC	22	2A	90	88	46	EE	B8	14	DE	5E	0B	DB
A	E0	32	3A	0A	49	06	24	5C	C2	D3	AC	62	91	95	E4	79
B	E7	C8	37	6D	8D	D5	4E	A9	6C	56	F4	EA	65	7A	AE	08
C	8A	78	25	2E	1C	A6	34	C6	E8	DD	74	1F	4B	BD	8B	8A
D	70	3E	85	66	48	03	F6	0E	61	35	57	B9	86	C1	1D	9E
E	E1	F8	98	11	69	D9	8E	94	9B	1E	87	E9	CE	55	28	DF
F	8C	A1	89	0D	BF	E6	42	68	41	99	2D	0F	80	54	BB	16

3.	a) Draw the Linear Feedback Shift Register (LFSR) with 5 cells in which $b_5 = a_4b_4 \oplus a_3b_3 \oplus a_2b_2 \oplus a_1b_1 \oplus a_0b_0$ where, $a_4 = 1$ and $a_3a_2a_1a_0$ are the binary representation of (last digit of your rollno + 1). If rollno = 032 then last digit + 1 = 3 and $a_3a_2a_1a_0 = 0011$.	CO2	3
----	---	-----	---

	b) What is the output bit of the above LFSR after 6 transitions, i.e. k_5 value, if the seed is $(10101)_2$.	CO2	2
4.	If $p ab$, p being a prime, then prove that $p a$ or $p b$.	CO1	3
5.	Determine the value of $2^{100} \bmod N$, where $N = (R \bmod 40) + 10$. e.g. if $R = 135$ then $N = 135 \bmod 40 + 10 = 15 + 10 = 25$.	CO1	2
6.	a) Find whether M is a primitive root under modulo 31, where $M = (R \bmod 25) + 3$, show calculation steps?	CO1	2
	b) Justify your answer.	CO1	1
7.	How many primitive roots exists under modulo R , where R is as mentioned above?	CO1	2

At the end of the course the students will be able to:

CO1	Develop an understanding of Mathematics involved in Cryptography.
CO2	Gain the familiarity of various ciphers and cryptosystems executions.
CO3	Analyze security strength of a cryptosystem.
CO4	Assess various system and network attacks.
CO5	Apply cryptographic techniques into network protocols and system applications, to make them secure

Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology Nagpur
Re-End Sem Examination

Course: CSL 443 – System and Network Security

BTech 7th Semester

Dec 2023

Time: 3 Hrs

Max Marks:50

Precisely answer all the questions.

Wherever find use A_u as (15 + last 2 digits of your Rollno).
For example, if Rollno is BT20CSE120 then $A_u = 15 + 20 = 35$.

1. a) Compute ($A_u^{10000000002} \bmod 22$). 2
b) Why Elgamal Enc. scheme is not secure in IND-CCA2 model, prove its insecurity cause? 8
2. a) Show the process of Man in the middle attack in Diffie Hellman key exchange method. 6
Draw the diagram to show the process with example numeric values of your own in the figure.
b) Explain the following notation of SSL: SSL_DH_RSA_WITH_3DES_EDE_CBC_SHA? 4
3. a) Explain the contents of Security Association Database (SAD) in IPSec protocols. 5
b) What is Security Association (SA) and who creates it for IPSec protocols? 2
c) When do we call Tunnel mode in action in IPSec protocol? 3
4. a) Use a single-letter frequency attack to break the following ciphertext. You know that it has been created with an additive cipher 8
OTWEWNGWCBPQABIZVQAPMLJGZWTTQVOBQUMAPMIDGZCAB
EQVBMZLZIXMLAXZQVOQVLMMXAVWEIVLLIZSNZWAB
JQZLWNLMTQOPBVIUMLGWCBPAEQNBGTMTNBBPMVMAB
ITIAKWCTLVBBQUMQBEPQTMQBIEAQVUGBZCAB
b) What is key-only attack model in symmetric cipher? 2
5. a) Explain Buffer overflow attack with example. 5
b) Explain Phishing attack with example. 5

At the end of the course the students will be able to:

CO1	Develop an understanding of Mathematics involved in Cryptography.
CO2	Gain the familiarity of various ciphers and cryptosystems executions.
CO3	Analyze security strength of a cryptosystem.
CO4	Assess various system and network attacks.
CO5	Apply cryptographic techniques into network protocols and system applications, to make them secure

1.	<p>Query: "to₁ be₂ or₃ not₄ to₅ be₆"</p> <p>To, 993427:</p> <p>{ 1: < 7, 18, 33, 72, 86, 231 >; 2: < 1, 17, 74, 222, 255 >; 4: < 8, 16, 190, 429, 433 >; 5: < 363, 367 >; 7: < 13, 23, 191 >;></p> <p>BE, 178239:</p> <p>{1: < 17, 25 >; 4: < 17, 191, 291, 430, 434 >; 5: < 14, 19, 101 >;...></p> <p>Explain the difference between biword indexing and positional indexing. Solve above question for the phrase query: "to be and not to be" using positional indexing.</p>	5
2.	Find the edit distance between "Percolating" and "Calculating", if the insertion cost is 2, and deletion cost is 3 and the substitution cost is 4. Backtrace your solution too. How is the edit distance algorithm used for tolerant retrieval?	5
3.	<p>Given a two-word query, the first-term postings lists with skip pointers are the following.</p> <p>2, 5, 20, 35, 67, 90, 112, 135, 145, 260, 267; with 2, 67, 145, as skip pointers,</p> <p>and the posting list for the second word with skip pointers is given below.</p> <p>3, 5, 7, 20, 35, 150, 260, 261, 270, 300; with 3, 35, 270 as skip pointers</p> <p>a. How many comparisons would be done to intersect the posting list without skip pointers?</p> <p>b. How many comparisons would be done to intersect both the posting list with skip pointers?</p>	5
4.	<p>Explain</p> <ul style="list-style-type: none"> a) index elimination b) champion lists c) vector space model d) Boolean retrieval model e) Global impact ordering 	5

5.	<p>Consider the following set of documents:</p> <p>D1 : Do you like green eggs and ham</p> <p>D2 : I do not like them Sam I am</p> <p>D3 : I do not like green eggs and ham</p> <p>D4 : I do not like eggs</p> <p>D5 : Why are they green Let the query be green eggs and suppose we eventually want to show the user 4 documents. That is, $K = 4$. 1.</p> <p>Q1. Run the phrase query green eggs. What are the documents that get retrieved in this step ?</p> <p>Q 2. What are the scores of the documents you retrieved in part (1) for the phrase query using cosine similarity on a bigram vector space ?</p> <p>Q3. If you did not have K documents in the first step, continue scoring according to the technique. What are the final 4 documents that you return and their scores) ?</p>	3
6	<p>Consider the following documents:</p> <p>doc1: phone ring person happy person</p> <p>doc2: dog pet happy run jump</p> <p>doc3: cat purr pet person happy</p> <p>doc4: life smile run happy</p> <p>doc5 life laugh walk run run</p> <p>(i) Construct the inverted index required for ranked retrieval for these five documents. Assume that no stemming or stop-word removal is required.</p> <p>(ii) Relating to the sample documents above, outline how the processing of the following Boolean query can be optimized: happy AND run AND pet</p>	2

Department of Computer Science & Engineering

Examination Name: **End Sem.** Subject: **CSL510 Introduction to Systems Engineering**

Class: **M. Tech. – Ist semester**

Slot: **C**

Time: **3 hours.**

Date: **6 Dec 2023, 10:00 AM - 1:00 PM**

Max. Marks: **40**

Note 1. All questions are compulsory.

2. It's an open book, open notes examination, already downloaded or physically available.
3. Verbatim answers from textbook/downloaded material will be ignored and will not be credited.
4. Draw the diagrams or sketches wherever necessary.
5. Please write your assumptions.

Q1 Analyse the following statements in Systems Engineering. Give examples in support of your answers. No word limit but precise answers will attract more credit. [CO2, CO4, CO5] [8]

- a) Implementing the latest technology is always the best strategy for system improvement.
- b) A system with a high degree of complexity is inherently more prone to failure.
- c) Once a system is deployed successfully, the initial design assumptions are proven correct.
- d) A system with redundant components is immune to critical failures.
- e) The most efficient way to address a system problem is to fix the immediate symptoms.

Q2 a) What emergence properties can be seen in the IRCTC- on-line ticket booking system? [1 1/2] [CO4,CO5]
b) Identify the form and functions in the system for providing drinking water through kiosks in a municipal area. You can assume that water source is providing treated water. It should start the water can filling only when payment is made. Water cans can be of varied capacity. [1 1/2] [CO4,CO5]

Q3 a) Analyse the five issues faced which the System Thinkers face while designing the system in the context of the design of the system for managing industry-institute research lab at your institute for inter-disciplinary projects. This research lab is given space, seed capital, equipments, and working groups of different domains of the project. The students will be given extra credits in their academic programs. [5] [CO2,CO4]

Hint: This research lab will be extension to institute ERP system as well as data-warehouse and will generate data like which projects of past, present, and future along with other routine data.

b) What are the issues involved in designing the Network Monitoring Software Systems from Systems Engineering perspective? [3] [CO2,CO4, CO5]

Domain Knowledge: A Network Monitoring System is a comprehensive software tool that is used to monitor, manage, and analyze the performance and health of computer networks. It plays a crucial role in ensuring the availability, reliability, and efficiency of network infrastructure.

[P.T.O.]

- Q4 a) **Create** an Object Process Model with textual support for Automatic . You can use the objects like robotic arm, advanced vision cameras, Centralized Control Center, Optical Character Recognition (OCR), conveyors, passengers. [5][CO1, CO2,CO3]
Identify the processes and their states, and transitions.
Your model should also have the response of the system to **constraint violations**.
Also **write what seed** you used in initiating the modeling process and **why?**
- b) A System to do X by (means of) Y in order to achieve Z. [3][CO1, CO3]
X – What the System does Y – How it does it
Z – Why is it being done
Write the X, Y, and Z for the purpose of defining the system's root definition for the sub-system your **Case Study submitted by your group**.
- Q5 Here is description of a system to be designed for a very big NGO [5]
doing environmental watch all over the world: [CO2,CO3, CO4, CO5]
In order to monitor the large scale works like mining, dams, road building etc. this NGO employees the automated systems as far as possible. It collects the on the spot information and also gets the opinions of the various stake holders on the works going on. This data will be in videos as well as reports and news. It decides to organize weekly assemblies at its headquarters, where experts in the concerned works will deliver talks ion theses works, which will be translated into different world languages and gist of the talk will be flashed on the website along with the data received.
Write the sub-systems and major components required to design this system with diagrammatic representation.
- Q6 Introducing two high-tech courses like Data Science and IOT in a [6]
technical university in all the disciplines have brought up about [CO1,CO2, CO3,CO4]
complexity issues for the automated course management system.
Some of the major are as under:
1. Intensifying an AI course may require additional computing resources and qualified instructors, which may affect the enrollment.
2. Modifying the Computer Science curriculum may necessitate adjustments in related courses and faculty responsibilities.
3. The introduction of IOT course might result in increased demand for specialized resources, affecting the university's overall academic landscape.
4. Offering advanced courses may require prioritizing resources away from foundational courses, impacting overall program diversity.
Categorise the appropriately and find the solutions for them from systems engineering point of view.
- Q7 What are the building blocks in a subscription-based streaming [2][CO2, CO4]
service like Netflix or Prime which can be categorised into various systems? How are these systems interconnected?

End Semester Examination, December 2023
VII Semester - Computer Science and Engineering

Subject: - Intro. to IoT (CSL429)

Date/Time: - 30/11/2023 (10:00 – 1:00 PM)

Slot: - II
Max. Marks: - 50

Q1.

- a. Draw a neat and clean labeled diagram of LoRaWAN Architecture and write the function of each component, when data is transferred from LoRaWAN Endpoint to Application server.
CO1, CO2 (4)
- b. For Noise monitoring system, which IoT level is used? Why? Show the working of each component at local level and at cloud level.
CO1, CO2 (4)

Q2.

- a. List out the parameters used as communication criteria while connecting smart objects to networks. Why they are important for constrained nodes?
CO2 (6)
- b. Under MAC & PHY layer for IoT network, which IoT access technology and topology was developed for constrained nodes.
CO2 (4)

Q3.

- a. Draw a neat and clean diagram for RYU Architecture and show what protocols it supports in south bound interface and what is used in North bound interface.
CO3 (4)
- b. What is Backhaul network? How its property/feature is used in LoRaWAN Architecture for connecting heterogeneous smart objects. Explain your answer with the help of LoRaWAN architecture.
CO2 (6)

- a. Using what parameters/features the flow gets removed from the flow table in Openflow Switch. Also write how/when the flow gets removed from the flow table.
CO-4 (4)

- b. In SDN (software define network) when the controller and OpenVSwitch gets configured for the first time, what messages (Openflow messages) gets exchange between them? Write and explain their use in chronological order.

CO3 CO4 (8)

Q5.

a. Match the Pairs

CO-1 (8)

A	B
RYU application	OVSDDB
Mininet topology	GCP, Azure, Rackspace
Table miss entry	ovs-vsctl show
North Bound Plugins	Google Apps (Gmail), Dropbox, Zoom
South Bound Plugins	ofctl_rest
Check number of switches	REST API
Infrastructure as a Service	actions=CONTROLLER:65535
Software as a Service	Single, 4

b. While checking the flow at OpenVSwitch at Bridge S1 using the following command
`ovs-ofctl -O OpenFlow13 dump-flows s1`

If we get following output, then which controller is configured with the OpenVSwitch? How it can be identified?

cookie=0x0, duration=654.092s, table=0, n_packets=72, n_bytes=5040, priority=0 actions=FLOOD
CO-3 (2)

Note: 1. All questions are compulsory.

2. It's an open book, open notes, open lecture slides exam

3. As far as possible, please express yourself in your own words. **Verbatim answers from the book or reading material others publish will be ignored and will not carry any credit.**

4. Please write your assumptions wherever required.

Q1. With the advances in medical science, human organ replacement with either donated organs from other human bodies or artificially made organs is the long term objective of the medical fraternity. You are required to write a technical part of the proposal for the the system for making repositories of the various organs at the national level. (Hint: You need to devise a methodology for data management of the donors as well as recipients) [3]
[CO1,CO2,CO3]

Q2. Analyse the following statements on the Systems Engineering (each 200 words max): (any 4) [6]
[CO1,CO4,CO5]

- a) *Systems Engineering handles the complexity in the design and development of the systems by enhancing the Systems Thinking.*
- b) *Biological systems have synthesis as the system building functionality.*
- c) *Boundaries in computer systems are permeable.*
- d) *Feedback loops drive systems behavior.*
- e) *Requirements drive the long-term sustainability in the software systems.*

Q3. Academic Information Systems for technical institutions and universities are becoming complex due to changing demands of the students and parents financing the education. As a systems engineers what interfaces you can define as a set of services and (user) flows in this changed scenario for technical education institutes with data analytics for continuous improvement of the system. [6]
[CO1,CO2,CO3]

Q3. A System to do X by (means of) Y in order to achieve Z.

[12]
[CO1,
CO2, CO4]

X – What the System does

Y – How it does it

Z – Why is it being done

Write the X, Y, and Z for the following systems, which state the purpose of defining the system's root definition.

- a) *Multi-speciality Hospital Management System*
- b) *Sell and logistics in a Mall Specialised in only sports goods and equipments*
- c) *Managing the online courses for a technical institution*
- d) *System for managing the maintenance of the residential facilities in a technical institution which include student hostel*

(Domain knowledge for these is already provided in the last semester SE course and this course)

Q4. Give any example from the software systems you are aware of where you can apply the spectrum of failures from blameworthy to praiseworthy for designing the feedback system of that system.

[3]
[CO2, CO5]

Business Intelligence and Analytics 1st Sessional Exam.

September 2023

Maximum Marks: 25

Time: 90 Minutes

Instructions:

1. Please be very specific with your answers.
2. Make suitable assumptions wherever necessary and state your assumptions clearly.
3. Please explain the reasoning behind your answers in all the MCQs. You will not get any marks for just writing the options.

Q1. State True or False with Justification.

(2 Marks)

1. A Fact table contains details about the attributes of the data.
2. Data in Data Warehouse is stored in normalized form.

Q 2. After each of the statements below, there are multiple options provided. Please choose all the options that are correct in the context of the statement. For example, in the question below correct choices will be A and C:

(5 Marks)

VNIT is _____.

- A. At Nagpur B. A medical college C. A Technology institute D. A place of worship

Please explain your choice of options. Partially correct answers may not get any marks.

1. In the context of customer support analytics, following are slowly changing dimensions (SCDs):

- A. Gender of the customer B. Marital Status of the customer
C. Address of the customer D. Customer Segment – Gold/Silver/Bronze
E. Name of the customer

2. Choose all the options with correct dimensional hierarchy. Write why the ones you have not chosen are wrong.

- A. Day -> Holiday -> Week -> Month -> Year B. Pincode -> State -> Region -> Country
C. Timestamp -> Day -> Week -> Quarter -> Year D. Continent -> Country -> Region -> State -> City
E. Product -> SKU -> Brand -> Company

3. The primary steps in data model design for a data-warehouse are:

- A. Physical Data Model Design B. Database Normalization C. Logical Data Model Design
D. Hypothetical Data Model Design E. Conceptual Data Model Design

4. All of the following about Data Warehouse are true except:

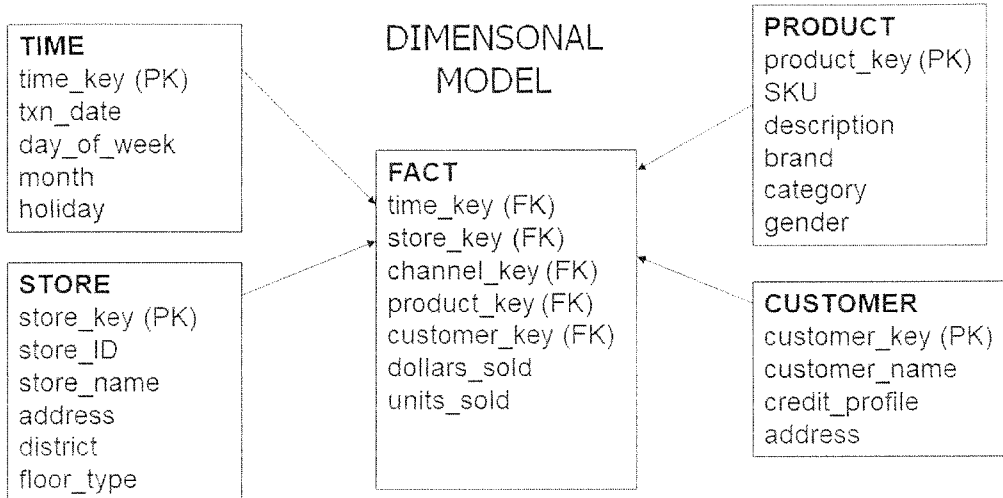
- A. Data Warehouse contains historical data B. Data Warehouse can have multiple Data Marts
C. Data Warehouse can be hosted in cloud D. Data Warehouse is optimized for retrieval
E. Data Warehouse is updated regularly

5. An ETL process involves:

- A. Extraction of Data from source systems
- B. Loading of the data in OLTP systems
- C. Estimation of size of the database
- D. Transformation of data from source systems
- E. Transaction Processing

Q 3. Look at the dimensional model below and answer the questions that follow.

(2+ 8 Marks)



- The dimensional model was copied from an earlier source and modified by the modeller. However, the dimensional modeller did not do a thorough job. So, there were a couple of mistakes in the dimensional model. Point out two obvious errors in the fact and dimension table above. **(2 Marks)**
- There are 5 analyses mentioned below. For **any four** of these, write if the analysis will be supported by the dimensional model above. Describe how the analysis will be performed.

for example, aggregate dollars sold by store_name and filter on Monday to find the aggregate sales for a store on Mondays

Wherever the analysis is not possible, write what changes need to be made to the dimensional model to make the analysis possible. Clearly state your assumptions (if any) **(8 Marks)**

- Top 5 stores preferred by female shoppers (preferred -> max purchases by female shoppers)
- Analysis of sales on holidays vs non-holidays
- Analyses of sales by region
- Worst performing store by sales of brand "x"
- Most popular SKU for each state

Q 4. Answer the following questions in brief (any 3)

(6 Marks)

- What is a factless fact table, provide an example of a factless fact table.
- What is the grain of a data warehouse? Why is granularity an important parameter in designing data warehouses?
- What are the four levels of analyses? Write a real-life example of each of the levels.
- In the context of Business Intelligence, what are source systems. Give examples of two source systems in any Business Intelligence use case you have studied.

Q 5. Draw a typical diagram of Star Schema highlighting facts and dimension tables. How is Star Schema different from Snowflake Schema?

(2 Marks)

X _____ X _____ X

End Semester Examination, December 2023
VII Semester - Computer Science and Engineering

Subject: - Intro. to IoT (CSL429)

Date/Time: - 30/11/2023 (10:00 – 1:00 PM)

Slot: - H
Max. Marks: - 50

Q1.

- a. Draw a neat and clean labled diagram of LoRaWAN Architecture and write the function of each component, when data is transferred from LoRaWAN Endpoint to Application server.
CO1, CO2 (4)
- b. For Noise monitoring system, which IoT level is use? Why? Show the working of each component at local level and at cloud level.
CO1, CO2 (4)

Q2.

- a. List out the parameters used as communication criteria while connecting smart objects to networks. Why they are important for constrained nodes?
CO2 (6)
- b. Under MAC & PHY layer for IoT network, which IoT access technology and topology was developed for constrained nodes.
CO2 (4)

Q3.

- a. Draw a neat and clean diagram for RYU Architecture and show what protocols it supports in south bound interface and what is used in North bound interface.
CO3 (4)
- b. What is Backhaul network? How its property/feature is used in LoRaWAN Architecture for connecting heterogeneous smart objects. Explain your answer with the help of LoRaWAN architecture.
CO2 (6)

Q4.

- a. Using what parameters/features the flow gets removed from the flow table in Openflow Switch. Also write how/when the flow gets removed from the flow table.
CO-4 (4)
- b. In SDN (software define network) when the controller and OpenVSwitch gets configured for the first time, what messages (Openflow messages) gets exchange between them? Write and explain their use in chronological order.
CO3 CO4 (8)

Q5.

a. Match the Pairs

CO-1 (8)

A	B
RYU application	OVSDDB
Mininet topology	GCP, Azure, Rackspace
Table miss entry	ovs-vsctl show
North Bound Plugins	Google Apps (Gmail), Dropbox, Zoom
South Bound Plugins	ofctl_rest
Check number of switches	REST API
Infrastructure as a Service	actions=CONTROLLER:65535
Software as a Service	Single, 4

b. While checking the flow at OpenVSwitch at Bridge S1 using the following command
`ovs-ofctl -O OpenFlow13 dump-flows s1`

If we get following output, then which controller is configured with the OpenVSwitch? How it can be identified?

cookie=0x0, duration=654.092s, table=0, n_packets=72, n_bytes=5040, priority=0 actions=FLOOD
 CO-3 (2)

Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology Nagpur

End Sem Examination

Course: **CSL 443 – System and Network Security**

BTech 7th Semester

Nov 2023

Time: **3 Hrs**

Max Marks:50

Precisely answer all the questions.

Wherever find use A_u as (15 + last 2 digits of your Rollno).
For example, if Rollno is BT20CSE120 then $A_u = 15 + 20 = 35$.

1. a) Compute ($A_u^{10000000001} \bmod 22$). 2 CO1
b) State and prove the security of Elgamal encryption scheme 8 CO3

Parameters for the Q2:

list = {5,11,17,19,23,25,29,31,37,41,43,47,53,55,59,61,67,71,73,79,83,85,89,95,97,101,103,107,109,113,115,121,125,127,131,137,139,145,149,151,155,157}

$$R_c = \begin{cases} \text{least number greater than } A_u \text{ from the list,} & \text{if } A_u \text{ is even} \\ \text{largest number less than } A_u, & \text{if } A_u \text{ is odd} \end{cases}$$

$$R_s = \begin{cases} \text{second least number greater than } A_u \text{ from list,} & \text{if } A_u \text{ is even} \\ \text{second largest number less than } A_u, & \text{if } A_u \text{ is odd} \end{cases}$$

e.g. if $A_u = 35$ then $R_c = 31$ and $R_s = 29$ and for $A_u = 34$, $R_c = 37$ and $R_s = 41$

2. a) Show the process of computing “pre-master secret” in SSL under Ephemeral Diffie-Hellman key exchange method, using the following values : $g = 2$, $p = 181$, server’s RSA key parameters are ($p = 151$, $q = 155$, $d_s = R_s$ (consider R_s value given as above)), derive other key parameters if required, and client’s RSA key parameters are ($p = 151$, $q = 155$, $d_c = R_c$), choose server’s half key’s exponent as 5 and client’s half key’s exponent as R_c . Put digital signature only on half key and use Hash function as $H(x) = x \bmod 11$. Draw the diagram to show the process with numeric values in figure. 8 CO5

b) Man in the middle attack is not possible in which key exchange method in SSL. Why? 2 CO4
3. a) In ESP protocol, while computing authentication data which field(s) is not included from IP packet. 2 CO5

b) Which service does ESP protocol provide but AH protocol does not provide in IPSec. 2 CO5

c) What does “Next header” field contains in AH and ESP protocol? 1 CO5

4. Suppose that it is known that plaintext letters (p denotes numerical equivalence of plaintext letter) are encrypted using the formula $4c \equiv (a.p + b) \bmod 27$. It is also known that the encryption of letter T is letter K and the encryption of letter I is letter D.
- Use this information to find the encryption keys a and b . 2 CO1
 - Find the decryption keys x and y , where $p \equiv (xc + y) \bmod 27$. 2 CO1
5. Consider the working of Needham-Schroeder Protocol (NSP) for authentication and answer the following questions:
- Why is the nonce R_a and additional nonce R_b necessary? 1 CO3
 - Why is it necessary to encrypt the ticket? 1 CO3
6. a) Suppose Edward wants to communicate with Bella using public key cryptography, how the public keys are advertised, verified by certificate using controlled trusted center and certifying authority? Explain with the help of diagram. 3 CO4
- b) Write the A5/1 algorithm step-wise for keystream generation. 7 CO2
 Suppose that, after a particular instance/step, the current values in the registers are
 $X = (x_0, x_1, \dots, x_{18}) = (10101010101010101)$
 $Y = (y_0, y_1, \dots, y_{21}) = (1100110011001100110011)$
 $Z = (z_0, z_1, \dots, z_{22}) = (11100001111000011110000)$
 Generate the next key bit.
 Print the contents of X , Y and Z after the above key bit has been generated.
7. a) Consider a Feistel cipher where the plaintext and ciphertext are of each 9 bits long and the key are 4 bits long. Assume that the function takes the first, third, and fourth bits of the key, interprets these three bits as a decimal number, cube the number, and interprets the result as a 9-bit binary pattern. Show the results of encryption and decryption if the original plaintext is 101110111 and the key is 1011. 5 CO2
- b) Match the following: 4 CO4
- | | |
|--------------------|---|
| a) Buffer overflow | i) unauthorized access to sensitive data |
| b) SQL injection | ii) injects their malicious code into corrupted memory |
| c) XSS | iii) attempt to steal sensitive information |
| d) Phishing attack | iv) attacker aims to execute malicious scripts in a web browser |

Course Outcome: At the end of the course the students will be able to:

CO1	Develop an understanding of Mathematics involved in Cryptography.
CO2	Gain the familiarity of various ciphers and cryptosystems executions.
CO3	Analyze security strength of a cryptosystem.
CO4	Assess various system and network attacks.
CO5	Apply cryptographic techniques into network protocols and system applications, to make them secure

End Sem Examination
CSL 436
Information Retrieval
Winter 2023 (VII sem)
Slot F
Maximum Marks 50
Maximum Time: 3 hours

1. Consider the documents:

Doc1: "Environmental changes are impacting global weather patterns."

Doc2: "Weather forecasting has improved with the advent of advanced computational models."

Doc3: "Global warming is a significant factor in drastic environmental shifts."

Doc4: "Computational models help in predicting the effects of global warming."

a) Draw the term document incidence matrix for collection.

b) Draw the inverted index representation for this collection.

show the returned results for **global AND weather**

(2+2=4)

2. Given a two-word query, the following is the postings lists for first term with skip pointers.

2, 5,20,35,67,90,112,135,145,260,267; with 2,67,145, as skip pointers and the posting list for second word with skip pointers is given below.

3,5,7,20,35,150,260,261,270,300; with 3,35,270 as skip pointers

a. How many comparisons would be done to intersect the posting list without skip pointers?

b. How many comparisons would be done to intersect both the posting list with skip pointers?

(2+2=4)

3. Suppose you wish to find economic reports regarding the impact of oil extraction in the North Sea on the Scottish economy. A commercial document retrieval service offers the following suggested matches: the table shows how often some key phrases appear in each report.

North Sea oil Scotland economy

Report A 12, 0, 3, 24

Report B 10, 5, 20, 10

Report C 0, 12, 9, 8

Query 1, 1, 1, 1

Actually obtaining the reports will cost real money, so you would like to select the one most likely to be relevant. Your task now is to assess this using the cosine similarity measure.

(a) Write out the general formula for calculating the cosine of the angle between two 4-dimensional vectors (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4) .

(b) Use this formula to rank the three documents in order of relevance to the query according to the cosine similarity measure. What do you think of the results?

(2+2=4)

4. Assume that the posting lists are gap encoded using γ codes. Using this encoding, suppose that the posting list for the term 'information' is the bit sequence:

1111 1111 1011 1100 1101 0011 1110 0000 0

And the posting list for the term 'retrieval' is the bit sequence:

1111 1111 1100 0000 0011 1011 1101 111

What doc-ids match the following query:

information AND NOT retrieval

(4)

5. What is term partitioning and document partitioning in distributed indexing? Differentiate between Block Sort based Indexing (BSBI) and Single Pass In Memory Indexing (SPIMI). (2+2=4)

6. A collection of 3000 documents, 3 documents are relevant to a query q : $REL\ q = \{A, B, C\}$ Below are rankings given by three systems for that query q . Only the first 5 documents are shown to the user. (Here we mark nonrelevant documents with "N".)

System 1	System 2	System 3
N	N	A
N	A	N
A	N	N
B	N	N
N	B	N

Which of the three systems would you prefer and why?

(4+1=5)

7. Answer the following questions in short. Some questions carry one mark and some carry 2 marks. Answer accordingly.

- a) What is the inverse document frequency (using log to the base 10) for a term that appears in 100 documents – with 10000 occurrences in these documents – in a text corpus of 1000000 documents? (1)

- b) Consider a collection with one billion tokens (i.e., with 10^9 tokens). Suppose the first 1,000 of these tokens results in a vocabulary size of 1,000 terms, and the first 100,000 tokens results in a vocabulary size of 10,000 terms. Use Heap's law to estimate the vocabulary size of the whole collection. (1)

- c) The following are entries of a positional index for the terms "peanut" and "butter", each showing the entries for one of the four documents 201, 202, 203, and 204. Which document contains the phrase "peanut butter"?

document 201: <peanut: ... 201: 4, 30; ... > <butter: ... 201: 3, 28; ... >

document 202: <peanut: ... 202: 17; ... > <butter: ... 202: 13, 21; ... >

document 203: <peanut: ... 203: 8, 10; ... > <butter: ... 203: 9; ... >

document 204: <peanut: ... 204: 5, 15; ... > <butter: ... 204: 7, 8; ... >

(1)

- d) What are the two assumptions behind modeling of the Binary Independence Model (BIM) and how do these assumptions help to reach to the BIM formula? (Need not do the complete derivation. Be precise.) (2)

- e) When (or in what situation) is the Normalized Discounted Cumulative Gain (NDCG) metric used/applicable? (1)

- f) What is the precision if a search engine retrieves 10 documents, 5 of which are relevant, from a corpus that contains in total 10 million documents, of which 100 are relevant? (1)
- g) A search engine is given the query “restaurants in Amsterdam”, upon which it returns 10'000 documents. In total, there are 100'000 documents that are indexed by the search engine, 4000 of which are relevant for the given query. Out of these relevant documents, 2000 are part of the returned results. What is the recall in this specific case? (1)
- h) Given a search engine that has a precision of 0.4 and a recall of 0.1, what is its F1-Measure? (1)
- i) What is the edit distance between “form” and “from” assuming the three operations insert, delete, and replace (but not transpose)? Use the dynamic programming algorithm and reach to the answer methodically. Do not guess the answer. (1)
- j) How do you estimate the size of the dictionary in an IR system? Give the formula. (1)
- k) Explain with a suitable diagram how cluster pruning is effective in IR? (1)
- l) Illustrate the mechanism of pseudo relevance feedback and implicit relevance feedback with suitable examples. (2)
- m) The Probabilistic Model of Information Retrieval is based on the foundation of the probability ranking principle. Justify in brief. (1)
- n) What do you mean by document likelihood model and query likelihood model in Language modeling? Which one is preferred over the other in the realistic setting and justify the reasoning for the same. (2)
- o) Latent Semantic Indexing (LSI) is observed to be effective as compared to the Vector Space Model (VSM). Justify this statement. (1)
- p) Why is HTS algorithm called as a mutually recursive algorithm? Give the outline of its pseudo code. (2)
- q) How/when does the page rank algorithm terminate? What is the termination mechanism called as? Describe the mechanism with an example. (1)
- r) Draw the BOW-tie structure of the web with all its components. Draw suitable arrows indicating how a surfer can surf from a part of the web to the other. Label the parts correctly. (1)
- s) The average in degree of all nodes in a snapshot of the web graph is 9. What can we say about the average out-degree of all the nodes in this snapshot? (1)
- t) How does index elimination mechanism help to improve the efficiency of the search mechanism? Illustrate with a suitable example. What are the heuristics used to do index elimination? (2)

Note 1. All questions are compulsory.

2. It's an open book, open notes examination, already downloaded or physically available.
3. Verbatim answers from textbook/downloaded material will be ignored and will not be credited.
4. Draw the diagrams or sketches wherever necessary.
5. Please write your assumptions.

Q1 Analyse the following statements in Systems Engineering. **Give examples in support of your answers. No word limit but precise answers will attract more credit.** [CO2, CO4, CO5] [8]

- a) Implementing the latest technology is always the best strategy for system improvement.
- b) A system with a high degree of complexity is inherently more prone to failure.
- c) Once a system is deployed successfully, the initial design assumptions are proven correct.
- d) A system with redundant components is immune to critical failures.
- e) The most efficient way to address a system problem is to fix the immediate symptoms.

Q2 a) What emergence properties can be seen in the IRCTC- on-line ticket booking system? [1 1/2] [CO4,CO5]
b) [1 1/2] [CO4,CO5]
b) Identify the form and functions in the system for providing drinking water through kiosks in a municipal area. You can assume that water source is providing treated water. It should start the water can filling only when payment is made. Water cans can be of varied capacity.

Q3 a) Analyse the five issues faced which the System Thinkers face while designing the system in the context of the design of the system for managing industry-institute research lab at your institute for inter-disciplinary projects. This research lab is given space, seed capital, equipments, and working groups of different domains of the project. The students will be given extra credits in their academic programs. [5] [CO2,CO4]

Hint: This research lab will be extension to institute ERP system as well as data-warehouse and will generate data like which projects of past, present, and future along with other routine data.

b) What are the issues involved in designing the Network Monitoring Software Systems from Systems Engineering perspective? [3] [CO2,CO4, CO5]

Domain Knowledge: A Network Monitoring System is a comprehensive software tool that is used to monitor, manage, and analyze the performance and health of computer networks. It plays a crucial role in ensuring the availability, reliability, and efficiency of network infrastructure.

[P.T.O.]

- Q4 a) **Create** an Object Process Model with textual support for Automatic . You can use the objects like robotic arm, advanced vision cameras, Centralized Control Center, Optical Character Recognition (OCR), conveyors, passengers. [5][CO1, CO2,CO3]
Identify the processes and their states, and transitions.
 Your model should also have the response of the system to **constraint violations**.
 Also **write what seed** you used in initiating the modeling process and **why?**
- b) A System to do X by (means of) Y in order to achieve Z. [3][CO1, CO3]
 X – What the System does Y – How it does it
 Z – Why is it being done
 Write the X, Y, and Z for the purpose of defining the system's root definition for the sub-system your **Case Study submitted by your group**.
- Q5 Here is description of a system to be designed for a very big NGO [5]
 doing environmental watch all over the world: [CO2,CO3, CO4, CO5]
 In order to monitor the large scale works like mining, dams, road building etc. this NGO employees the automated systems as far as possible. It collects the on the spot information and also gets the opinions of the various stake holders on the works going on. This data will be in videos as well as reports and news. It decides to organize weekly assemblies at its headquarters, where experts in the concerned works will deliver talks ion theses works, which will be translated into different world languages and gist of the talk will be flashed on the website along with the data received.
 Write the sub-systems and major components required to design this system with diagrammatic representation.
- Q6 Introducing two high-tech courses like Data Science and IOT in a [6]
 technical university in all the disciplines have brought up about [CO1,CO2, CO3,CO4]
 complexity issues for the automated course management system.
 Some of the major are as under:
 1. Intensifying an AI course may require additional computing resources and qualified instructors, which may affect the enrollment.
 2. Modifying the Computer Science curriculum may necessitate adjustments in related courses and faculty responsibilities.
 3. The introduction of IOT course might result in increased demand for specialized resources, affecting the university's overall academic landscape.
 4. Offering advanced courses may require prioritizing resources away from foundational courses, impacting overall program diversity.
 Categorise the appropriately and find the solutions for them from systems engineering point of view.
- Q7 What are the building blocks in a subscription-based streaming [2][CO2, CO4]
 service like Netflix or Prime which can be categorised into various systems? How are these systems interconnected?

CSL 412 - Artificial Intelligence (Slot D)
End Semester Examination (November 2023)

Time: 3 hours

Max. Marks: 50

- Q1 (a) Mention the heuristics used by the DPLL algorithm. Explain how these can be computed efficiently. With an example, explain the concept of unit propagation. (CO3, CO4) (4)

(b) Consider that the knowledge base (KB) has been converted into the conjunctive normal form (CNF) and the seven clauses (C1-C7) with four variables (x1-x4) are as follows. (\neg refers to NOT and \vee refers to OR)

C1: $x1 \vee x2 \vee x3$ C2: $x1 \vee x2 \vee \neg x3$ C3: $x1 \vee \neg x2 \vee x4$
 C4: $x1 \vee \neg x4$ C5: $\neg x1 \vee x2$ C6: $\neg x2 \vee x3$ C7: $\neg x2 \vee \neg x3$

Use the DPLL algorithm, to check if a model exists for the KB. (CO3 CO4) (6)

- Q2 (a) Solve the following cryptarithmic problem using the concepts of solving a constraint satisfaction problem (CSP).

$$\begin{array}{r} T W O \\ + T W O \\ \hline F O U R \\ \hline \end{array}$$

Each letter stands for a distinct digit and no leading zeroes are allowed. (CO1, CO2) (7)

- (b) Design a local search technique to solve a CSP. (CO1, CO2) (3)

- Q3 (a) Show that if there is a function symbol in the Knowledge Base then inference in First Order Logic (FoL) may not be decidable. (CO3, CO4) (3)

(b) Convert the following English Language sentence into equivalent FoL sentence using appropriate predicate names.

A person who plays sports has a few friends. (CO3) (2)

(c) Convert the above sentence to the Conjunctive Normal Form (CNF). (CO3) (3)

(d) Using an example, explain the resolution rule for First order Logic. (CO3) (2)

Q4 (a) With an example, show how alpha beta pruning can lead to better efficiency over the minimax algorithm in the case of a two player zero sum game.

(CO1, CO2) (3)

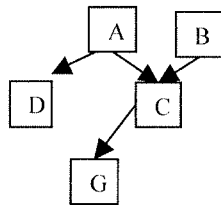
(b) Mention the actions that the memory bounded A* (MA*) algorithm performs when it finds that the memory has been exhausted. Why is MA* called as an optimally efficient algorithm?

(CO1, CO2) (4)

(c) From a given state, how do we proceed to a new state in simulated annealing? What is the role of the temperature variable?

(CO2) (3)

Q5. Consider the Bayesian Network shown below where all the variables are Boolean having values T and F. The conditional probability tables of all the nodes are also given.



$$P(A=T) = 0.3, P(B=T) = 0.1$$

$$P(D=T|A=T) = 0.9, P(D=T|A=F) = 0.2$$

$$P(C=T|A=T, B=T) = 0.8, P(C=T|A=T, B=F) = 0.4$$

$$P(C=T|A=F, B=T) = 0.1, P(C=T|A=F, B=F) = 0.05$$

$$P(G=T|C=T) = 0.8, P(G=T|C=F) = 0.1$$

(a) Using the technique of exact inference, find the answer to the query

$$P(A=T|D=F, G=T)$$

(CO4) (7)

(b) Explain how the same query can be answered using the Likelihood Weighting technique for approximate inference.

(CO4) (3)

Examination Name: **End Sem.** Subject: **CSL510 Introduction to Systems Engineering**
Class: **B. Tech. – 7th semester** Slot: **C**

Time: **3 hours.** Date: **22 Nov 2023, 10:00 AM - 1:00 PM** Max. Marks: **40**

Note 1. All questions are compulsory.

2. It's an open book, open notes examination, already downloaded or physically available.
3. Verbatim answers from textbook/downloaded material will be ignored and will not be credited.
4. Draw the diagrams or sketches wherever necessary.
5. Please write your assumptions.

Q1 Analyse the following statements in Systems Engineering. **Give examples in support of your answers. No word limit but precise answers will attract more credit.** [CO2, CO4, CO5] [8]

- a) Systems Thinking is not thinking systematically.
- b) Modeling is a conceptual tool to define and grasp complex systems, offering structured representations of their interconnected components and behaviors. For systems, value is benefit at cost.
- c) Systems and environments are intricately interdependent, engaging in dynamic interactions that mutually shape and influence each other, while maintaining their distinct identities, structures, and functions.
- d) Optimizing performance requires mastering system complexity for adaptability in dynamic environments.

Q2 a) What emergence properties can be seen in the Amazon on-line shopping system portal? [1 1/2]
[CO4, CO5]
b) Identify the form and functions in the system for packaging of the CPUs using a robot arm. This system should be with conveyor belt and packaging unit based on the size of the CPU packet. It detects the size through optical scanners. [1 1/2]
[CO4, CO5]

Q3 a) Analyse the five issues faced which the System Thinkers face while designing the system in the context of the design of the system for managing startup incubation centre at your institute. This incubation centre will offer space, seed capital, technical know-how. It will also create a Technology & Science Park for the creating institute-industry interface. This will be utilised for the R&D activities jointly with internship programs for the students. [5]
[CO2, CO4]
Hint: This centre system will be extension to institute ERP system and will generate data like which companies are registered, their products, services etc. and their turnover.

b) What are the issues involved in designing the Integrated Development Environment (IDE) for any programming language from Systems Engineering perspective? [3]
[CO2, CO4, CO5]
Domain Knowledge: An IDE is a comprehensive software tool that provides a unified interface for coding, debugging, testing, and deploying software applications, enhancing developers' efficiency and workflow.

- Q4 a) **Create** an Object Process Model with textual support for Elevator Control System. You can use the objects like elevators, floors, and passengers. [5][CO1, CO2, CO3]
Identify the processes and their states, and transitions like elevator movement, door operations, and response to calls by person manning the elevator.
 Your model should also have the response of the system to **constraint violations**.
 Also **write what seed** you used in initiating the modeling process and **why?**
- b) A System to do X by (means of) Y in order to achieve Z. [3][CO1, CO3]
 X – What the System does Y – How it does it
 Z – Why is it being done
 Write the X, Y, and Z for the purpose of defining the system's root definition for the sub-system your **Case Study submitted by your group**.
- Q5 Here is description of a system to be designed for the United Nations: [5][CO2, CO3, CO4, CO5]
 In order to stop the wars and skirmishes going all over the world, United Nation decides to collect the on the spot information and get the opinions of the world leaders. This data will be in videos as well as reports and news. It decides to organize weekly assemblies at its headquarters, where world leaders will deliver talks on these incidents, which will be translated into different world languages and gist of the talk will be flashed on the website along with the data received.
 Write the sub-systems and major components required to design this system with diagrammatic representation.
- Q6 Introducing high-tech courses like AI and Bioinformatics in a technical university have brought up about complexity issues. [6][CO1, CO2, CO3, CO4]
 Some of the major are as under:
 1. Intensifying an AI course may require additional computing resources and qualified instructors, which may affect the enrollment.
 2. Modifying the Bioinformatics curriculum may necessitate adjustments in related courses and faculty responsibilities.
 3. The introduction of AI courses might result in increased demand for specialized resources, affecting the university's overall academic landscape.
 4. Offering advanced courses may require prioritizing resources away from foundational courses, impacting overall program diversity.
 Categorise the appropriately and find the solutions for them from systems engineering point of view.
- Q7 What are the building blocks in a Data Center like that of Google or Amazon, which can be categorised into various systems? How are these systems interconnected? [2][CO2, CO4]

Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology Nagpur
Re-End Sem Examination
Course: **CSL 443 – System and Network Security**

BTech 7th Semester
Time: **3 Hrs**

Dec 2023
Max Marks:50

Precisely answer all the questions.

Wherever find use A_u as (15 + last 2 digits of your Rollno).
For example, if Rollno is BT20CSE120 then $A_u = 15 + 20 = 35$.

1. a) Compute $(A_u^{10000000002} \bmod 22)$. 2
b) Why Elgamal Enc. scheme is not secure in IND-CCA2 model, prove its insecurity cause? 8
2. a) Show the process of Man in the middle attack in Diffie Hellman key exchange method. 6
Draw the diagram to show the process with example numeric values of your own in the figure.
b) Explain the following notation of SSL: SSL_DH_RSA_WITH_3DES_EDE_CBC_SHA? 4
3. a) Explain the contents of Security Association Database (SAD) in IPSec protocols. 5
b) What is Security Association (SA) and who creates it for IPSec protocols? 2
c) When do we call Tunnel mode in action in IPSec protocol? 3
4. a) Use a single-letter frequency attack to break the following ciphertext. You know that it has been created with an additive cipher 8

OTWEWNGWCBPQABIZVQAPMLJGZWTTQVOBQUMAPMIDGZCAB
EQVBMZLZIXMLAXZQVOQVLMMXAVWEIVLLIZSNZWAB
JQZLWNLMTQOPBVIUMLGWCBPAEQNBGTMTNBBPMVMAB
ITIAKWCTLVBBQUMQBEPQTMQBIEAQVUGBZCAB

b) What is key-only attack model in symmetric cipher? 2
5. a) Explain Buffer overflow attack with example. 5
b) Explain Phishing attack with example. 5

At the end of the course the students will be able to:

CO1	Develop an understanding of Mathematics involved in Cryptography.
CO2	Gain the familiarity of various ciphers and cryptosystems executions.
CO3	Analyze security strength of a cryptosystem.
CO4	Assess various system and network attacks.
CO5	Apply cryptographic techniques into network protocols and system applications, to make them secure

CSL 430: Business Intelligence

Final Examination

20th November 2023

Maximum Marks: 60

Time: 180 Minutes

Instructions:

1. Please be very specific with your answers.
2. Make suitable assumptions wherever necessary and state your assumptions clearly.
3. Please explain the reasoning behind your answers in all the MCQs. You will not get any marks for just writing the options.

Q1. State True or False with Justification.

(8 x 1Mark = 8 Marks)

1. When the size of a database increases beyond a specific volume, it is called a data warehouse.
2. If two datasets have same mean and same standard deviation, they are not necessarily similar.
3. Inventory is a semi-additive measure as it can be aggregated across all dimensions except time.
4. Pie chart is not the right choice of visualization to express proportions as many people find it difficult to compare sectors of circles.
5. A Standard Normal Distribution has a mean of 1 and standard deviation of 0.
6. The choice to use ETL or ELT is based only on the business use case. Either of the two can be used based on choice of the Data Warehouse Architect.
7. Data Visualization best practice suggests Data Ink Ratio should always be minimized.
8. A Modern Data Stack is not as good at managing BI reporting use cases as a Traditional Data Warehouse.

Q 2. After each of the statements below, there are multiple options provided. Please choose all the options that are correct in the context of the statement. For example, in the question below correct choices will be A and C:

(Any 4. 4 x 2 Marks=8 Marks)

VNIT is _____.

- A. At Nagpur B. A medical college C. A Technology institute D. A place of worship

Please explain your choice of options. Partially correct answers may not get any marks.

1. Online Transaction Processing Systems are:

- A. Optimized for retrieval B. Optimized for storage C. Optimized for both retrieval and storage
D. Contain only structured data E. Contain only unstructured data

2. A Modern Data Stack _____.

- A. Supports Real Time Analytics B. Can not be hosted in cloud C. Has multiple layers
D. Supports only Snowflake schema E. Supports Data Visualization

3. In the context of a Bank, following are slowly changing dimensions (SCDs):

- A. Customer ID allocated to the customer B. Marital Status of the customer
C. Address of the customer D. Customer Segment – Gold/Silver/Bronze
E. Account Balance of the Customer

4. All of the following are examples of structured data except:

- A. Server Logs B. Student data in a SQL database C. JSON objects
D. Data from IoT sensors E. Sales data in CSV file

5. Choose all the correct statements about Inventory

- A. It is a physical place where goods are stored B. It is maintained to tackle sudden changes in demand
C. Lead has no impact on Inventory levels D. More inventory has to be maintained if lead time is more
E. Less inventory has to be maintained if lead time is less.

Q3. Look at the images/data below and point out mistakes if any. If you find mistakes provide corrections to correct these mistakes. If there is no mistake please state that clearly. (4 x 2 Marks = 8 Marks)

1. Table below shows the answers to match the pairs question solved by a student of BI.

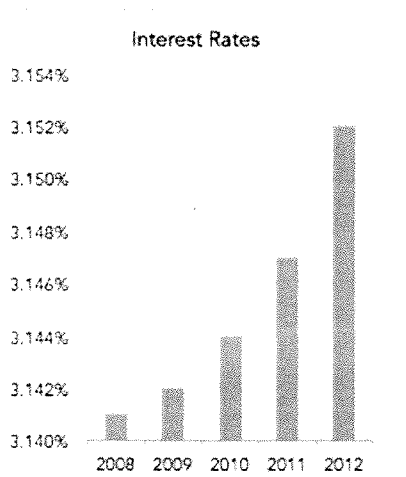
#	Column A	Column B
1	Relational DBMS	Are always ACID Compliant
2	Data Mart	Data about Data
3	Data Lake	Supports structured as well as Unstructured data
4	Data Warehouse	Best suited for reporting on structured data
5	Data Lakehouse	Combines the best of Data Lake and Data Warehouse
6	Metadata	Departmental Data

2. An attribute is defined as a unique level within a dimension. For example, Timestamp is an attribute in the Time Dimension.
3. Levels of BI Progression are as follows:

BI Progression



4. Look at the bar chart of Interest rates in the US between the years (2008-2012):



Q 3. Answer the following questions.

(Any 5. 5 x 4 Marks = 20 Marks)

1. Look at the table below and find out which of the following are facts. For each of the facts write which will be additive, semi-additive and non-additive and why?

Date	Store	Customer	Product	Product Description	Sales in Rs.	Discount	Inventory Balance	Count of Employees
------	-------	----------	---------	---------------------	--------------	----------	-------------------	--------------------

2. What is the explain the concept of granularity of data. How does it determine the ability to roll up or drill down? Explain with an example.

3. Look at the sample data for Final exam conducted by VNIT in the table below. Describe four actionable analyses that you can perform on this data. Assume you have 10,000+ rows of data. Example of analyses – Which branch took the most and least time to complete the exam?

Roll No.	Gender (M/F)	Question (Set Type)	Start Time HH:MM	End Time HH:MM	Section A Score	Section B Score	Total Score (A+B)	Disruption (Y/N)
B1891	F	3-A	09:00	10:21				N
C1901	M	3-B	12:00	13:02				N
C1422	F	4-C	14:30	14:39				N
D1901	F	1-A	15:00	16:30				Y

- The first letter of the roll number indicates branch – there are only 4 branches Architecture, B - Building & Civil Engineering , C – Computer Science, D – Designing and Electronics Engineering.
- The next two digits indicate the year of admission and other two digits indicate the roll number.
- The digit in Question Set Type indicates the year and letter indicates the type of set. There are only 3 set types A,B,C.
- Disruption (Y/N) indicates if there was any power cut or internet disruption while the student was attempting the test.

4. Look at the following table about visitors to the magical island of Good-la-gooba. If this information is stored in a data warehouse, what will be the fact(s), dimension(s) and attribute(s) here? What more attributes are possible within the dimensions?

Number of visitors

Month	Age			Gender M/F
	<30	30-60	60+	
January	781	878	346	1184/821
February	1021	1022	420	1047/1416
March	1211	788	278	1111/1166

5. A product has a lead time of 12 days. Company mandates keeping an safety stock 5 days of demand for each product. What should be order amount if the daily sale of the product is 45 units and current inventory is 920?

- What will happen to the order amount if you choose to place an order after 5 days?
- What will happen to the order amount if daily sale becomes 70 units?

6. How does Predictive Modelling work? Explain with an example. Can Normal Distribution be used as a method for predictive analysis? Give reasons for your answer.

7. Why are companies moving to the Cloud Data Warehouse? Write 5 reasons why Cloud Data Warehouses make sense. Provide an example of a Cloud Data Warehouse.

Q 5. Study the Data Schema for a multinational restaurant . Answer the questions that follow. **(8 Marks)**

POS Bill	ERP System 1	Store DB	Employee DB
Bill No	Item ID (same as Order Code)	Store ID	Employee ID
Date	Item Description (Ordered item)	Store Name	Employee Name
Timestamp	Item Cost	Address Line 1	Address Line 1
Ordered Item	Sub-Category ID	Address Line 2	Address Line 2
Order Code	Sub-Category Desc.	City	City
Quantity	Category ID	State	State
Price	Category Desc.	Country	Country
Amount	Shelf Life	Pin Code	Pin Code
Service Tax	Storage Type ID	Region (as in APAC, EMEA, Americas)	Region (as in APAC, EMEA, Americas)
Vat	Related Items	Total Store Area	Joining Date
% Discount	Storage Type Desc.		Current Designation
Salesperson			Last Designation
Store ID			Salary
Store Type			Employment Type
			Bonus flag

1. What analyses can you perform based on this data to reduce the wastage of food? **(2 Marks)**
2. Which of the following analyses are possible? Explain the steps to perform those analyses. If they are not possible explain why they are not possible and what additional data will be required to make those analyses possible: **(Any 3. 3 x 2 Marks = 6 Marks)**
 - a. The restaurant chain wants to find out the most profitable item on its menu.
 - b. The restaurant decides to award a “Lambi race ka ghoda” award to its top 10 longest serving male employees.
 - c. The restaurant chain wants to order the restaurants by annual sales per unit area.
 - d. The restaurant chain wants to report total number of restaurant (stores) in each region.

Q.6. Answer with the help of a well labelled diagram/visualizations: (Any 2. 2 x 4 Marks = 8 Marks)

1.Explain any two of the following visualization charts with the help of a diagram. Mention in which case these visualizations can be used. For example – A multi-line chart can be used to compare two variables over time:

a. Scatter Plot, b. Bubble chart c. Heatmap.

2. Draw a schematic diagram of a Modern Data Stack showing different layers and components.

3. Draw a typical Star and Snowflake Schema. Mention the key difference between the two (2-3 sentences)

X _____ X _____ X