Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2023-2024

End-Semester Test
(EC-3 Regular)

Course No.	: DSECLZG529/AIMLCZG529
Course Title	: Data Management for Machine Learning
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	: ~~06/03/2021 or 19/03/2021 (FN/AN)~~

No. of Pages = 10
No. of Questions = 7

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

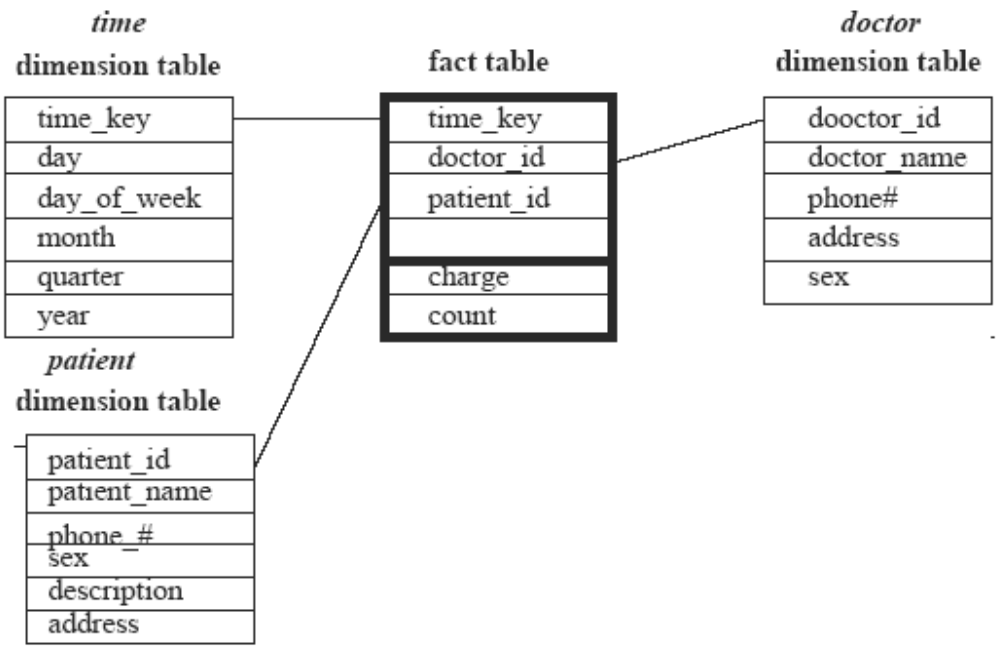
Q1. Suppose that a data warehouse consists of three dimensions: **time, doctor,** and **patient**, and the two measures **count** and **charge**, where charge is the fee that a doctor charges a patient for a visit. **[1 + 2 + 2.5 + 1.5 = 7]**

(a) Enumerate two classes of schemas popularly used for modelling data warehouses.
(b) Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a).
(c) Starting with the base cuboid [**day, doctor, patient**], what specific OLAP operations should be performed to list the total fee collected by each doctor in 2004?
(d) To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (**day, month, year, doctor, hospital, patient, count, charge**).

Answer:

(a) star schema: a fact table in the middle connected to a set of dimension tables
snowflake schema: a refinement of star schema where some dimensional hierarchy is normalised into a set of smaller dimension tables, forming a shape similar to a snowflake.
Fact constellations: multiple fact tables share dimension tables, viewed as a collection of stars, called galaxy schema or fact constellation.
(b) As shown below: Star -schema

crackbitswilp.in



Each table 0.5 marks

(c) Steps

1. roll up from day to month to year
2. slice for year = "2004"
3. roll up on patient from individual patient to all
4. slice for patient = "all"
5. get the list of total fee collected by each doctor in 2004

Each step 0.5 marks

(d)

Select doctor, Sum(charge) – 0.5 marks
From fee
Where year = 2004 – 0.5 marks
Group by doctor – 0.5 marks

Q2. You are working on a machine learning project that predicts monthly sales for an e-commerce website. You have trained a sales prediction model using data from the year's first six months (January to June). The model performs well during initial testing.

However, you notice a significant drop in prediction accuracy when you deploy the model in the production environment and start using it to make predictions for the next three months (July to September). You suspect that data drift may be the cause of this decline in performance.

To investigate data drift, you collect the following data for both the training data (January to June) and the production data (July to September):

crackbitswilp.in

Training Data (January to June):

- Mean Monthly Sales: $50,000
- Standard Deviation of Monthly Sales: $7,000

Production Data (July to September):

- Mean Monthly Sales: $45,000
- Standard Deviation of Monthly Sales: $8,500

Calculate the following data drift metrics and provide your analysis: **[2 +2 + 2 = 6]**

a) Percentage Change in Mean Monthly Sales between Training and Production Data.
b) Percentage Change in Standard Deviation of Monthly Sales between Training and Production Data.

Answer:

1. **Percentage Change in Mean Monthly Sales:**

   To calculate the percentage change in mean monthly sales, you can use the following formula:

   $$\text{Percentage Change} = \frac{\text{Production Mean} - \text{Training Mean}}{\text{Training Mean}} \times 100$$

   Plugging in the values:

   $$\text{Percentage Change} = \frac{45{,}000 - 50{,}000}{50{,}000} \times 100 = -10\%$$

   The percentage change in mean monthly sales between the training and production data is -10%. This indicates a 10% decrease in mean sales in the production data compared to the training data.

2. **Percentage Change in Standard Deviation of Monthly Sales:**

   To calculate the percentage change in the standard deviation of monthly sales, you can use the following formula:

   $$\text{Percentage Change} = \frac{\text{Production Std Dev} - \text{Training Std Dev}}{\text{Training Std Dev}} \times 100$$

   Plugging in the values:

   $$\text{Percentage Change} = \frac{8{,}500 - 7{,}000}{7{,}000} \times 100 \approx 21.43\%$$

   The percentage change in the standard deviation of monthly sales between the training and production data is approximately 21.43%. This indicates a 21.43% increase in the variability of sales in the production data compared to the training data.

crackbitswilp.in

Analysis:

The negative percentage change in the mean monthly sales suggests that the average sales amount in the production data (July to September) is 10% lower than in the training data (January to June). This could be due to changes in customer behavior, economic factors, or other external influences.

The positive percentage change in the standard deviation of monthly sales indicates that the variability in sales has increased by approximately 21.43% in the production data compared to the training data. This suggests that the sales data in the production environment may be more volatile or unpredictable, which can significantly impact the performance of the sales prediction model.

To address the data drift issue, you may need to retrain your model using the latest data or implement strategies to adapt the model to the changing patterns in sales data.

Q3. You are tasked with designing a data pipeline to process and analyse log data from a website. The logs contain user interactions, including page views, clicks, and demographics. The pipeline consists of three stages: data ingestion, transformation, and analysis.

- Data Ingestion: The raw log data is ingested into the pipeline at an average of 1,000 log entries per second.
- Data Transformation: During the transformation stage, various data processing tasks are performed, including data cleaning, parsing, and feature extraction. The transformation stage processes data at 800 log entries per second.
- Data Analysis: In the analysis stage, machine learning models are applied to the transformed data to predict user behavior. The analysis stage processes data at an average of 500 log entries per second.

a) Calculate the data throughput for each stage of the data pipeline in log entries per minute. **[3]**
b) Determine the bottleneck stage in the data pipeline based on the calculated throughputs. **[1]**

Answer:

crackbitswilp.in

a) Calculating Data Throughput:

To calculate the data throughput for each stage, we'll first calculate the processing rate per second for each stage and then convert it to log entries per minute.

1. **Data Ingestion:**
   - Processing Rate (per second) = 1,000 log entries/second
   - Processing Rate (per minute) = 1,000 log entries/second * 60 seconds/minute = 60,000 log entries/minute
2. **Data Transformation:**
   - Processing Rate (per second) = 800 log entries/second
   - Processing Rate (per minute) = 800 log entries/second * 60 seconds/minute = 48,000 log entries/minute
3. **Data Analysis:**
   - Processing Rate (per second) = 500 log entries/second
   - Processing Rate (per minute) = 500 log entries/second * 60 seconds/minute = 30,000 log entries/minute

b) Determining the Bottleneck Stage:

The bottleneck stage is the stage with the lowest throughput because it determines the overall capacity of the data pipeline.

In this case, the bottleneck stage is the **Data Analysis** stage, as it has the lowest throughput of 30,000 log entries per minute. This means that even if the other stages can process data at a higher rate, the Data Analysis stage cannot process data faster than 30,000 log entries per minute, limiting the overall pipeline throughput.

To optimize the data pipeline's overall throughput, you may need to consider various strategies, such as parallel processing, distributed computing, or hardware upgrades, to improve the performance of the bottleneck stage or balance the processing rates across all stages.

Q4. Imagine you are a data scientist working on a machine learning project for a healthcare organisation. Your task is to build a predictive model to identify patients at high risk of developing a specific medical condition based on their health records.

crackbitswilp.in

In your machine learning project, you trained three models: Model A, Model B, and Model C, each with various hyperparameters and feature engineering techniques. As part of your model metadata, you have recorded the model's architecture, hyperparameters, training data, evaluation metrics, and the date of each model's creation. Additionally, you have documented any noteworthy observations or lessons learned during the modelling process.
You are reviewing the model metadata for Models A, B, and C, and you notice that Model C consistently outperforms the other models in terms of accuracy and recall on the validation dataset. However, Model C is also significantly larger regarding memory usage than Models A and B. Given this information:
a) Why is it important to keep track of model metadata in your machine learning project? **[2]**
b) What are the potential advantages of using Model C despite its higher memory usage? **[2]**
c) How can you ensure that the large memory usage of Model C is smooth in a production environment? **[1]**

Answer:
a) Keeping track of model metadata is crucial for several reasons:

- Reproducibility: Model metadata allows you to recreate and reproduce the same model in the future, which is essential for auditing and validation.
- Performance Evaluation: Metadata helps you compare different model iterations, understand which models perform better, and make informed decisions about model selection.
- Troubleshooting: When issues or unexpected behaviour arise, model metadata provides historical context and insights into what might have caused the problem.
- Documentation: Metadata serves as documentation for your machine learning pipeline, making it easier for collaborators to understand and work with your models.

b) The advantages of using Model C, despite its higher memory usage, may include:

- Higher Accuracy: Model C consistently outperforms the other models regarding accuracy and recall, which may lead to more accurate predictions and better patient risk assessment.
- Improved Patient Care: The improved performance of Model C may help healthcare providers identify patients at higher risk more accurately, allowing for early interventions and better patient care.
- Research Insights: The larger model may capture more complex patterns in the data, potentially leading to valuable insights or discoveries about the medical condition.

c) To ensure that the large memory usage of Model C does not become a bottleneck in a production environment, you can consider the following strategies:

- Model Optimization: Explore techniques for model optimisation, such as model pruning, quantisation, or using more memory-efficient architectures.
- Hardware Scaling: Assess the organisation's hardware capabilities and, if necessary, invest in hardware upgrades or cloud resources that can accommodate larger models.
- Batch Processing: Implement batch processing for predictions if real-time processing is not required, as this can help manage memory usage.
- Monitoring and Scaling: Continuously monitor memory usage in production and set up auto-scaling mechanisms to handle increased loads as needed.
- Model Compression: Explore model compression techniques that reduce memory requirements while retaining most of the model's performance.

crackbitswilp.in

By considering these strategies, you can leverage the advantages of Model C while mitigating potential challenges related to its higher memory usage in a production healthcare environment.

Q5. You are tasked with designing a data architecture for a large e-commerce platform. The platform handles many daily customer transactions, product updates, and user interactions. The architecture must efficiently support real-time analytics, reporting, and data storage. You decide to use a data warehouse for this purpose.
Here are some critical metrics for the platform:

- 10,000,000 customer transactions per day
- 1,000,000 product updates per day
- 5,000,000 user interactions per day

Your data architecture needs to handle and process this data efficiently. Design a data architecture that includes the following components and provide an estimate of the required storage capacity: **[4]**

- Data Ingestion Layer
- Data Storage Layer
- Data Processing Layer
- Data Analytics and Reporting Layer

Answer:
To design a data architecture for the e-commerce platform, we need to consider each layer and estimate the required storage capacity.

1. **Data Ingestion Layer:**
   - In this layer, data from various sources (customer transactions, product updates, user interactions) is ingested into the system.
   - Given the metrics:
     - 10,000,000 customer transactions per day
     - 1,000,000 product updates per day
     - 5,000,000 user interactions per day
   - The daily data ingestion rate is 10,000,000 + 1,000,000 + 5,000,000 = 16,000,000 records per day.
2. **Data Storage Layer:**
   - In this layer, data is stored efficiently. We need to estimate the required storage capacity based on daily data ingestion and retention policies.
   - Assume each record on average is 1 KB in size.
   - The daily data ingestion is 16,000,000 records per day.
   - If we want to retain data for a year (365 days), the annual storage requirement would be:
     - 16,000,000 records/day * 1 KB/record * 365 days = 5,840,000 GB or approximately 5.84 petabytes (PB) of storage per year.

crackbitswilp.in

3. **Data Processing Layer:**
   - In this layer, data is processed, transformed, and prepared for analytics and reporting.
   - The required processing capacity depends on the complexity of the transformations and the desired processing speed. It may involve distributed processing using technologies like Apache Spark or Hadoop.
4. **Data Analytics and Reporting Layer:**
   - In this layer, data is used for real-time analytics and reporting.
   - The performance and responsiveness of this layer depend on factors like query complexity, indexing, and hardware infrastructure.

In summary, for this data architecture, you would need approximately 5.84 petabytes of storage capacity per year to handle the daily data ingestion and retain data for a year. The other layers would require scalable and efficient technologies to handle data processing, analytics, and reporting volume and complexity.

Q6. Consider a dataset containing information about customer orders for an e-commerce website. Here are 20 records with various data quality issues. Identify at least 5 data quality issues and compute data quality metrics to help illustrate the importance of data quality assessment. **[10]**

```
Order_ID,Order_Date,Product,Quantity,Total_Price,Shipping_Address
101,2022-01-05,Widget,3,$45.00,123 Main St, Cityville
102,2022-01-10,Widget,5,$75.00,,Cityville
103,2022-02-15,Widget,2,$30.00,456 Elm St, Townsville
104,2022-02-20,Widget,1,$15.00,789 Oak St, Villageville
105,,Widget,4,$60.00,234 Birch St, Townsville
106,2022-03-25,Widget,2,$30.00,567 Pine St,
107,2022-03-30,Widget,3,$45.00,890 Cedar St, Cityville
108,2022-04-05,Widget,,,$70.00,123 Main St, Cityville
109,2022-04-10,Widget,5,$, Townsville
110,2022-05-15,Widget,3,$45.00,123 Main St, Cityville
111,2022-05-20,Widget,-2,$30.00,456 Elm St, Townsville
112,2022-06-25,Widget,2,$35.00,789 Oak St, Villageville
113,2022-06-30,Widget,3,$,, Cityville
114,2022-07-05,Widget,4,$65.00,234 Birch St, Townsville
115,2022-07-10,Widget,2,$30.00,567 Pine St, Villageville
116,2022-08-15,Widget,1,$15.00,, Villageville
117,2022-08-20,Widget,4,,$60.00,789 Oak St, Villageville
118,2022-09-25,Widget,3,$45.00,890 Cedar St, Cityville
119,2022-09-30,Widget,,,$70.00,123 Main St, Cityville
120,2022-10-05,Widget,5,$75.00,123 Main St, Cityville
```

crackbitswilp.in

Answer:

Data Quality Issues:

- Missing values in "Shipping_Address" field (rows 2, 5, 6, 9, 13, 16, 18).
- Missing values in "Total_Price" field (rows 4, 9, 13, 17).
- Negative quantity in "Quantity" field (row 11).
- Invalid date format in "Order_Date" field (row 5, 13).
- Duplicate Order_IDs (rows 1 and 10 have the same Order_ID).

Data Quality Metrics:

1. **Percentage of Missing Values in Shipping_Address:**
   - Number of missing values in "Shipping_Address" = 7
   - Total number of records = 20
   - Percentage of missing values = (7 / 20) * 100 = 35%
2. **Percentage of Missing Values in Total_Price:**
   - Number of missing values in "Total_Price" = 4
   - Total number of records = 20
   - Percentage of missing values = (4 / 20) * 100 = 20%
3. **Percentage of Negative Quantity Values:**
   - Number of records with negative "Quantity" values = 1
   - Total number of records = 20
   - Percentage of negative values = (1 / 20) * 100 = 5%
4. **Percentage of Records with Invalid Date Format:**
   - Number of records with invalid date format = 2
   - Total number of records = 20
   - Percentage of records with invalid date format = (2 / 20) * 100 = 10%
5. **Percentage of Duplicate Order_IDs:**
   - Number of duplicate Order_IDs = 1 (Order_ID 101 is duplicated)
   - Total number of records = 20
   - Percentage of duplicate Order_IDs = (1 / 20) * 100 = 5%

Q7. A healthcare organisation is sharing medical research data with a research partner while ensuring privacy protection using a privacy-preserving technique called "k-anonymity." The dataset contains information about patients' medical conditions and their ages. The organisation wants to disclose aggregate statistics about patient ages while protecting individual privacy.
The organisation chooses to achieve 3-anonymity, meaning that there are at least three patients with the same age and medical condition for any combination of age and medical condition. The dataset contains the following information:

- Patient A: Age 45, Medical Condition X
- Patient B: Age 30, Medical Condition Y

crackbitswilp.in

- Patient C: Age 45, Medical Condition Z
- Patient D: Age 35, Medical Condition X
- Patient E: Age 50, Medical Condition Y
- Patient F: Age 45, Medical Condition X
- Patient G: Age 30, Medical Condition Z
- Patient H: Age 30, Medical Condition X

a) Calculate the transformed dataset that satisfies the 3-anonymity requirement. **[3]**
b) Explain how the transformed dataset ensures privacy protection for individual patients. **[1]**

Answer:
a) To achieve 3-anonymity, we need to transform the dataset so that for any combination of age and medical condition, there are at least three patients with the same age and medical condition. We can do this by generalising and grouping patients with similar attributes.
The transformed dataset for 3-anonymity:

- Group 1: Ages 30-35, Medical Condition X
  Patients: B, D, H
- Group 2: Age 45, Medical Condition X
  Patients: A, F, (Any other patient with age 45 and Medical Condition X)
- Group 3: Age 30, Medical Condition Y
  Patients: E (Any other patient with age 30 and Medical Condition Y)
- Group 4: Age 45, Medical Condition Z
  Patients: C (Any other patient with age 45 and Medical Condition Z)
- Group 5: Age 30, Medical Condition Z
  Patients: G (Any other patient with age 30 and Medical Condition Z)

We've grouped patients with similar ages and medical conditions in the transformed dataset. Each group satisfies the 3-anonymity requirement because at least three patients are in each group.
b) The transformed dataset ensures privacy protection for individual patients by obscuring their specific attributes. Here's how it provides privacy:

- Individual Patient Identification: Identifying a specific patient within a group is difficult because multiple patients share the same age and medical condition. For example, in Group 1, patients B, D, and H share similar attributes, making it challenging to distinguish one from another.
- Preventing Attribute Disclosure: The transformed dataset prevents attribute disclosure by only revealing generalised information, such as age ranges and medical conditions, for groups of patients. This reduces the risk of exposing sensitive patient information.
- Meeting the Anonymity Requirement: By ensuring 3-anonymity, the organisation guarantees at least three patients with the same age and medical condition in each group. This makes it statistically challenging to re-identify individual patients from the disclosed data.

Overall, k-anonymity techniques like this one balance the need for sharing aggregated data for research purposes while protecting the privacy of individual patients.

****

crackbitswilp.in