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

End-Semester Test
(EC-3 Makeup)

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.

1. Mainwind Traders has several data workloads. Design and explain briefly your data solution to handle this data's ingestion, processing, and analysis based on the following requirements. Tabulate your answer as shown below and draw suitable block diagrams wherever appropriate. Also, comment on whether Lambda architecture will be helpful in this solution. [10]

   - **Manufacturing data.** The company has been storing manufacturing logs collected from the assembly line. You want to analyze these logs to gain insights into material behaviour and quality assurance. To analyse these logs, you need to use reference data, such as material, chemical, and origin information, in an on-premises data store. You want to utilise this data from the on-premises data store, combine it with other log data that it has in a cloud data store, and run stored procedures on the data to gain insights.

   - **Real-time data.** Mainwind Traders need real-time data ingestion and storage for multiple data sources like their websites, point-of-sale systems, and social media sites. It would be best if you had a solution to analyse this data and provide valuable insights to the CEO.

   - **Historical company data.** The company is required to store 5 TB of company data for legal reasons. This data is rarely used or referenced but must not be deleted. You need a cost-effective method for storing this data.

	Manufacturing data	**Real-time data**	**Historical company data**
Sources			

crackbitswilp.in

Ingestion strategy			
Storage strategy			
Tools for Ingestion			
Tools for Ingestion			
Processing strategy			
Processing complexity			
Processing tools			
Storage/usage of analysis results			
Other Architectural components			

Answer:

	Manufacturing data	**Real-time data**	**Historical company data**
Sources	• Log data • Reference data	• Websites • point-of-sale systems • social media sites	• Data Lakes • Databases • Data warehouse
Ingestion strategy	Need basis to pull from sources	Push from sources to intermediator system	Routine data migration from source to target without any transformation
Storage strategy	No additional storage	Intermediate storage required	A cost-effective storage solution is needed
Tools for Ingestion	• Custom code or • ETL / ELT pipeline	• Custom Code • Kafka	• Custom code • Data migration/copying tools
Tools for Storage	• File System for log data / Data Lake • RDBMS for reference data	Intermediate buffers like Kafka or Messaging queues	• HDFS • Data LakeHouse • AWS Glacier
Processing strategy	Batch/offline/ad-hoc processing	Online / Stream processing	No processing, only data migration
Processing	• Joins are involved	• Joins are required	No processing

crackbitswilp.in

complexity	• Hybrid data processing	or not is not clear • The type of data is not clear • What type of analysis needs to be done is not clear	
Processing tools	Spark	• Kafka + Spark • Kafka + Flink • AWS Kinesis	No processing
Storage/usage of analysis results	• Analytical Datastore • Dashboard	• Real-time data store • Alerts • Dashboards	No processing
Analysis objective	to gain insights into material behaviour and quality assurance.	Provide valuable insights to the CEO.	Storage for legal reasons

Block diagram pending

Lambda architecture will only make sense if the sources and objectives of analysis differ.

https://www.openintro.org/data/index.php?data=loans_full_schema

Q2. Consider the following schema definition that is representing the loan status of various applicants and is used to predict whether the customer will be a defaulter or not, what can be loan amount and tenure of the loan.

Attribute	Description
emp_title	Job title.
emp_length	Number of years in the job, rounded down. If longer than 10 years, then this is represented by the value 10.
state	Two-letter state code.
homeownership	The ownership status of the applicant's residence.
annual_income	Annual income.
verified_income	Type of verification of the applicant's income.
debt_to_income	Debt-to-income ratio.

crackbitswilp.in

annual_income_joint	If this is a joint application, then the annual income of the two parties applying.
verification_income_joint	Type of verification of the joint income.
debt_to_income_joint	Debt-to-income ratio for the two parties.
delinq_2y	Delinquencies on lines of credit in the last 2 years.
months_since_last_delinq	Months since the last delinquency.
earliest_credit_line	Year of the applicant's earliest line of credit
inquiries_last_12m	Inquiries into the applicant's credit during the last 12 months.
total_credit_lines	Total number of credit lines in this applicant's credit history.
open_credit_lines	Number of currently open lines of credit.
total_credit_limit	Total available credit, e.g. if only credit cards, then the total of all the credit limits. This excludes a mortgage.
total_credit_utilized	Total credit balance, excluding a mortgage.
num_collections_last_12m	Number of collections in the last 12 months. This excludes medical collections.
num_historical_failed_to_pay	The number of derogatory public records, which roughly means the number of times the applicant failed to pay.
months_since_90d_late	Months since the last time the applicant was 90 days late on a payment.
current_accounts_delinq	Number of accounts where the applicant is currently delinquent.
total_collection_amount_ever	The total amount that the applicant has had against them in collections.
current_installment_accounts	Number of installment accounts, which are (roughly) accounts with a fixed payment amount and period. A typical example might be a 36-month car loan.
accounts_opened_24m	Number of new lines of credit opened in the last 24 months.
months_since_last_credit_inquiry	Number of months since the last credit inquiry on this applicant.
num_satisfactory_accounts	Number of satisfactory accounts.

crackbitswilp.in

num_accounts_120d_past_due	Number of current accounts that are 120 days past due.
num_accounts_30d_past_due	Number of current accounts that are 30 days past due.
num_active_debit_accounts	Number of currently active bank cards.
total_debit_limit	Total of all bank card limits.
num_total_cc_accounts	Total number of credit card accounts in the applicant's history.
num_open_cc_accounts	Total number of currently open credit card accounts.
num_cc_carrying_balance	Number of credit cards that are carrying a balance.
num_mort_accounts	Number of mortgage accounts.
account_never_delinq_percent	Percent of all lines of credit where the applicant was never delinquent.
public_record_bankrupt	Number of bankruptcies listed in the public record for this applicant.
loan_purpose	The category for the purpose of the loan
application_type	The type of application: either individual or joint.
loan_amount	The amount of the loan the applicant received.
term	The number of months of the loan the applicant received.
interest_rate	Interest rate of the loan the applicant received.
installment	Monthly payment for the loan the applicant received.
grade	Grade associated with the loan.
sub_grade	Detailed grade associated with the loan.
issue_month	Month the loan was issued.
loan_status	Status of the loan
disbursement_method	Dispersement method of the loan.
balance	Current balance on the loan.
paid_total	Total that has been paid on the loan by the applicant.
paid_principal	The difference between the original loan amount and the current

crackbitswilp.in

	balance on the loan.
paid_interest	The amount of interest paid so far by the applicant.
paid_late_fees	Late fees paid by the applicant.

a) Identify instances of data leakages. [2]
b) When developing a model to predict whether the customer will be a defaulter, can k-means algorithms be used as a data partitioning strategy? Why? [1.5]
c) If the model developed in part c is used online to make defaulter predictions, can the feature engineering code used while training the model be used as is for serving the model? [1.5]
d) Explain the approach that can be used for labelling the dataset. [2]
e) Whether data versioning will be required? Justify. [1]
f) Will the tools for model experimentation and model registry be needed for this case?[2]

Answer:

a) There are two cases of data leakage
   - paid_late_fees is feature coming from features like num_accounts_120d_past_due, num_accounts_30d_past_due, months_since_90d_late, num_historical_failed_to_pay, and will not be available in runtime
   - loan amount to be disbursed is direct function of features like debt_to_income, total_credit_utilized, total_debit_limit,balance, paid_total etc

b) No. Model will develop a bias towards the categories of the customers and may attempt to overfit for a particular set of customers, rather than being generic in nature.

c) No. At runtime, while serving the model, many of the feature computed will not be available hence needs to be computed first or needs to be referenced from the feature store to make it available for the prediction code.

d) Yes data labeling will be required in this case as its not explicitly mentioned in the given dataset. As its structured datasets and rules for data labeling (like who is defaulter) are straight-forward, the data scientist (along with domain expert) can complete the labeling, may be by codifying the rule in a program and get it validated from the QA expert.

e) Yes over the period of time (usually monthly) this dataset will be updated by the lenders, hence data versioning will be required to know what dataset is used for developing a model.

f) Yes, both of them will be useful from ML modeling perspective.
   - The model experiment tools will help in tracking various iterations of the experiments carried out before settling on the final version to be released.
   - The model registry tool will help in maintain various versions of the models which might be in different stages like staging, released, retired etc. in its lifetime.

Q3. Use the below product sales data for answering the OLAP sub-questions [2 * 5= 10]

City	Store	Product	Month	Quantity
Pune	NORTH	P1	Jan	17
Pune	NORTH	P1	Feb	19

crackbitswilp.in

Pune	NORTH	P1	March	23
Pune	NORTH	P2	Jan	57
Pune	NORTH	P2	Feb	78
Pune	NORTH	P2	March	45
Pune	SOUTH	P1	Jan	12
Pune	SOUTH	P1	Feb	67
Pune	SOUTH	P1	March	34
Pune	SOUTH	P2	Jan	89
Pune	SOUTH	P2	Feb	54
Pune	SOUTH	P2	March	76
Mumbai	EAST	P3	Jan	76
Mumbai	EAST	P3	Feb	54
Mumbai	EAST	P3	March	89
Mumbai	EAST	P4	Jan	12
Mumbai	EAST	P4	Feb	87
Mumbai	EAST	P4	March	90
Mumbai	WEST	P3	Jan	57
Mumbai	WEST	P3	Feb	32
Mumbai	WEST	P3	March	73
Mumbai	WEST	P4	Jan	97
Mumbai	WEST	P4	Feb	14
Mumbai	WEST	P4	March	65

a) Represent the sales data in a multi-dimensional format. [4]
b) Represent the sales data in multi-dimensional format when the months are rolled up to Quarters.
c) How the multi-dimensional data structure will look when it's sliced for North Store?

crackbitswilp.in

d) How the multi-dimensional data structure will look when it's diced for South Store and Product P2?
e) If "Product" is replaced by "Category" (product P1, P2 belongs to one category, whereas P3, P4 belongs to another category), how the multi-dimensional structure will look like in this case?

Answer:



Q4. You are tasked with improving the data quality and observability of a dataset containing customer orders for an e-commerce website. The dataset contains 1,000 records, and you've identified several data quality issues:

- Duplicate Orders: There are 50 duplicate orders in the dataset.
- Missing Values: In the "Shipping Address" field, 100 records have missing or incomplete addresses
- Inconsistent Date Formats: The "Order Date" field contains dates in both "MM/DD/YYYY" and "YYYY-MM-DD" formats. There are 30 records with inconsistent date formats.
- Outliers in Order Amount: There are 10 orders with an unusually high order amount that seems like outliers.

a) Explain why data quality and observability are essential in this context. [2]
b) Design and compute four data quality metrics based on the given data. [4]
c) Propose a step-by-step plan to address these data quality issues and enhance data observability. [4]

Answer:

a) Explanation of Data Quality and Observability Importance:

- Data quality and observability are crucial in the context of this dataset for several reasons:
- Trustworthiness: Data quality ensures that the data is accurate and reliable, which is vital when making business decisions based on customer orders.
- Analysis Accuracy: High-quality data improves the accuracy of any analyses or machine learning models built on it, such as sales forecasting, customer segmentation, or product recommendations.
- Customer Experience: Accurate shipping addresses are essential for delivering products to customers correctly and promptly, which impacts customer satisfaction.
- Operational Efficiency: Reliable data reduces errors and inefficiencies in order processing and inventory management.
- Data Drift Detection: Monitoring data quality helps detect changes over time, such as the emergence of data drift, which can impact model performance and business insights.

b) Design and compute four data quality metrics based on the given data. [4]
Four data quality metrics:

- Percentage of Duplicate Orders.
- Percentage of Records with Missing or Incomplete Shipping Addresses.

crackbitswilp.in

- Percentage of Records with Inconsistent Date Formats.
- Percentage of Orders Considered as Outliers (based on order amount).

Computation of four data quality metrics:

- Percentage of Duplicate Orders:
  Percentage of duplicate orders = (Number of duplicate orders / Total number of records) * 100
  Percentage of duplicate orders = (50 / 1,000) * 100 = 5%

- Percentage of Records with Missing or Incomplete Shipping Addresses:
  Percentage of records with missing or incomplete shipping addresses = (Number of records with missing addresses / Total number of records) * 100
  Percentage of records with missing or incomplete shipping addresses = (100 / 1,000) * 100 = 10%

- Percentage of Records with Inconsistent Date Formats:
  Percentage of records with inconsistent date formats = (Number of records with inconsistent date formats / Total number of records) * 100
  Percentage of records with inconsistent date formats = (30 / 1,000) * 100 = 3%

- Percentage of Orders Considered as Outliers (based on order amount):
  Percentage of orders considered as outliers = (Number of outlier orders / Total number of records) * 100
  Percentage of orders considered as outliers = (10 / 1,000) * 100 = 1%

c) Plan to Address Data Quality Issues and Enhance Observability:

- Duplicate Orders (Data Deduplication):
  - Identify duplicate orders by comparing unique order identifiers or a combination of order attributes (e.g., customer name, order date, and product).
  - Decide whether to keep one instance of each duplicate or remove all duplicates based on business rules.
  - Update the dataset accordingly and document the deduplication process.

- Missing Values (Data Imputation):
  - Examine records with missing or incomplete shipping addresses.
  - Use data imputation techniques to fill in missing values where possible (e.g., predicting addresses based on postal codes).
  - If addresses cannot be reliably imputed, decide whether to exclude those records or follow up with customers for address verification.

- Inconsistent Date Formats (Data Transformation):
  - Standardize the date format in the "Order Date" field to a single format (e.g., "YYYY-MM-DD").

crackbitswilp.in

    - Use data transformation scripts or tools to update the date format for the 30 records with inconsistencies.
    - Ensure future data inputs adhere to the standardized format.

- Outliers in Order Amount (Data Cleaning):
    - Identify the orders with unusually high order amounts using statistical methods (e.g., Z-scores).
    - Evaluate whether these outliers are legitimate, perhaps representing high-value orders or corporate clients.
    - Based on business rules, decide whether to keep, remove, or further investigate the outliers.
    - Apply robust statistical methods or transformations if outliers are legitimate but skewing the data.

- Continuous Monitoring and Documentation:
    - Implement automated data quality checks and monitoring to identify and address data issues in real-time.
    - Create a data quality dashboard to provide visibility into data quality metrics and anomalies.
    - Document all data quality issues, resolutions, and data transformation processes for future reference.

By following this step-by-step plan, you can address the data quality issues in the dataset and enhance data observability, ensuring that the data is reliable, accurate, and suitable for making informed business decisions in the e-commerce context.

****

crackbitswilp.in