National University of Computer and Emerging Sciences

Lahore Campus

Fundamentals of Big Data

Final Exam

Date: May 06th 2024 DS2004

Course Instructor(s)

Asbah Khalid

Mamoona Akbar

Total Time (Hrs):	**3**
Total Marks:	**120**
Total Questions:	**6**

22L-7529 Roll No

BDS-4A Section

Student Signature

Do not write below this line

Attempt all the questions.

CLO #: Design and analysis the concepts of RDD in pySpark

Q1: You are working as a data engineer for an e-commerce company. The company has a large dataset of transaction records stored in a distributed file system. Each transaction record contains the following fields: [15 marks]

- Transaction ID
- User ID
- Product ID
- Product Category
- Transaction Amount
- Payment Method (Credit Card, PayPal, etc.)
- Timestamp

Your task is to analyze this dataset to extract several insights regarding user behaviour, product performance, and transaction trends. Specifically, you need to perform the following tasks using Spark's RDD:

T1,U1,P1,C1,100.0,Credit Card,2023-01-01 10:00:00
T2,U2,P2,C2,50.0,PayPal,2023-01-01 10:05:00
T3,U1,P1,C1,100.0,Credit Card,2023-01-01 10:10:00
T4,U3,P3,C3,200.0,Credit Card,2023-01-01 10:15:00

1. Calculate the total amount spent by each user and identify the top 10 highest-spending users.
2. Calculate the total spending per product category and determine the percentage contribution of each category to the overall revenue.

3. For each user, determine the most frequently used payment method and the total amount spent using that method.
4. For a given user, generate a sequential list of products they purchased, ordered by the timestamp of the transactions.
5. Identify the top 5 products by total sales amount and find the users who purchased these products the most frequently.
6. For each product category, compute the total transaction amount for each month and identify any seasonal trends.

CLO #: Concepts of DataFrames

Q2: You work as a data analyst for a video streaming service. The company has a large dataset of user interactions with the platform, including viewing history, user demographics, and subscription details. The dataset contains the following fields: [20 marks]

Field Name	Description	Data Type
InteractionID	Unique identifier for each interaction	String
UserID	Unique identifier for each user	String
VideoID	Unique identifier for each video	String
VideoCategory	Category of the video (e.g., Drama, Comedy)	String
WatchTime	Total watch time in minutes	Integer
Rating	User rating of the video (1-5)	Integer
SubscriptionType	Type of subscription (Free, Basic, Premium)	String
Timestamp	Timestamp of the interaction	Timestamp
UserAge	Age of the user	Integer
UserGender	Gender of the user (M/F/Other)	String

You need to perform the following tasks using Spark DataFrames.

- Calculate the total watch time for each user.
- Determine the average rating given by each user.
- Identify the top 10 users with the highest total watch time.
- Find the distribution of ratings across different video categories.
- Calculate the total watch time for each video.
- Determine the average rating for each video.
- Identify the top 10 most-watched videos.
- Find the top 5 video categories by total watch time.
- Calculate the total number of users for each subscription type.
- Determine the average watch time per user for each subscription type.

- Identify any patterns or trends in subscription upgrades or downgrades over time.
- Analyze the watch time distribution across different age groups and genders.
- Determine the most popular video categories for each age group and gender.
- Identify any significant differences in viewing behaviour between different demographic groups.
- Compute the total watch time for each month.
- Identify any seasonal trends or patterns in viewing behaviour.
- Perform a year-over-year growth analysis for the total watch time.

CLO #: Concepts of Transformations

Q3: Draw the execution plan for Question#01 and mention the transformation types for every task done in question.01.[20 marks]

CLO #: Understanding of MapReduce

Q4: You are given a large dataset of web server logs where each entry contains a timestamp, a user ID, and a URL visited. You need to use the MapReduce framework to find the total number of unique URLs visited by each user. [10 marks]

Part A: Design the map function. What key-value pairs would it emit?

Part B: Design the reduce function. What will it receive as input, and what should it output

CLO #: Hive and Recommendations systems and YARN working

Q5: Answer the following questions: [20 marks]

- On what architecture Hive works.
- Which is better RBMS or Hive and explain why?
- Which Recommendation system you would prefer to use and in what scenario.
- Explain the implicit and explicit feedback with example.
- What is meant by the resource manager in High Availability Mode.
- What is the role of elector.
- Describe the process of application running in YARN.

CLO #: Working of Recommendation Systems

Q6: Consider a small subset of a movie ratings matrix as follows, where each row represents a user, each column represents a movie, and the cells contain rating scores from 1 to 5. A blank (NaN) indicates that the user has not rated that movie. [10]

User	Movie A	Movie B	Movie C	Movie D	Movie E
User1	3	4	NaN	5	1

User	Movie A	Movie B	Movie C	Movie D	Movie E
User2	2	NaN	3	4	2
User3	1	2	2	3	NaN
User4	NaN	4	4	4	4

Tasks:

- Calculate the average rating for each user based only on the movies they have rated.
- Center the ratings for each user by subtracting their average rating from each rated movie. Fill the resulting centered ratings matrix.
- Calculate the similarity scores.
- Using the similarity scores, predict User1's rating for Movie C
- Would you like to suggest any movies from the dataset to the users and explain why?

CLO #: Types of Recommendation Systems

Q7: You are given a dataset with attributes of various movies and user preferences. The goal is to find out how interested a particular user might be in a newly released movie.[10 marks]

Given Data:

Movie Attributes (Movie Profile):

Movie ID	Genre	Director	Length	Audience Score
M1	Action	Spielberg	120	85
M2	Comedy	Tarantino	90	75
M3	Action	Nolan	150	95
M4	Sci-Fi	Scott	130	90

User Preference (User Profile):

User ID	Preferred Genre	Favorite Director	Ideal Length	Minimum Audience Score
U1	Action	Nolan	120	80

Newly Released Movie:

Movie ID	Genre	Director	Length	Audience Score
M5	Action	Nolan	140	88

Estimate User U1's interest level in the newly released Movie M5 based on their preferences.

National University of Computer and Emerging Sciences

Lahore Campus

CLO #: RDDs Understanding

Q7: You work for a company managing a high-traffic web application, which generates large volumes of access logs. These logs are stored in real-time across a distributed system using Spark RDDs. [20 marks]

Logs are continuously appended to RDDs, and each log entry is structured as follows:

timestamp | IP address | request type | response code | response time

Example log entries:

2024-06-05T12:00:00Z | 192.168.1.1 | GET | 200 | 15
2024-06-05T12:00:01Z | 192.168.1.2 | POST | 500 | 20
2024-06-05T12:00:01Z | 192.168.1.1 | GET | 200 | 10

Do the following tasks:

- calculate the total number of requests per IP address.
- summarize the average response time and the most common response code for each IP within these intervals.
- Identify IPs with the highest and lowest average response times
- Detect IPs with abnormal request rates compared to the global average
- Identify IPs with frequent 404 errors
- Calculate the hourly distribution of requests for each IP
- Identify IPs with a sudden spike in request volume compared to the previous hour
- Identify IPs with a consistent pattern of accessing specific endpoints