**CAPSTONE PROJECT (55 marks + 15 marks (viva-voce))**

**Business Scenario**

An e-commerce company, ‘***SwiftCommerce***’ is looking to analyze its customer purchasing behavior to improve its sales and marketing strategies. The company plans to collect customer purchase data from various sources such as its website, mobile app, and other transactional data sources. To analyze the data, the company decides to use Spark Framework, which can handle large volumes of data efficiently and provide real-time analysis of customer behavior.

Using Spark Core and Spark SQL, the company can process the customer purchase data to identify popular products, analyze trends in purchasing behavior, and target specific customer segments with personalized marketing messages. The insights gained from the analysis can help the company improve its sales and marketing strategies, optimize pricing strategies, and identify opportunities to upsell or cross-sell products.

Using Spark's powerful data processing capabilities, the company aims to analyze the data in real-time, identifying trends and patterns that could indicate areas for improvement. For example, the system might detect a high volume of purchases for certain products during specific times of the year, indicating that the company could improve its inventory management or promotional strategies.

The system could then generate customized reports, highlighting key purchasing metrics and offering recommendations for improvement. The reports could be accessed by sales and marketing teams, allowing them to make data-driven decisions and track progress over time. With its real-time data analytics capabilities and customizable reports, the system could help the e-commerce company stay ahead of the curve in an increasingly competitive market. The company could market this system to other e-commerce businesses as a valuable tool for improving sales and increasing customer satisfaction.

**Dataset Schema**

1. Customer purchase history dataset:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Customer ID | string | Unique identifier for the customer who made the purchase |
| Purchase Date | date | Date when the customer made the purchase |
| Product ID | string | Unique identifier for the product |
| Discount Applied | float | Discount applied to the purchase |

1. Product catalog Dataset:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Product ID | string | Unique identifier for the product |
| Product Name | string | Name of the product |
| Product Category | string | Category of the product |
| Price | float | Price of the product (without discount) |
| Rating | float | Average rating for each product based on customer reviews. The rating scale is assumed to be out of 5 stars |

1. Website traffic and user behavior dataset:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Date | date | Date of website traffic and user behavior measurement |
| Page Views | int | Total number of page views on the website for the date |
| Unique Visitors | int | Total number of unique visitors on the website for the date |
| Bounce Rate | int | Percentage of visitors who leave the website after viewing only one page for the date |
| Session Duration (Seconds) | int | Average duration of a website visit session (in seconds) for the date |

**Sample Dataset**

1. Customer purchase history dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer ID** | **Purchase Date** | **Product ID** | **Discount Applied** |
| CUST001 | 8/1/2022 | PRD002 | 5 |
| CUST002 | 7/12/2022 | PRD079 | 20 |
| CUST003 | 6/17/2022 | PRD023 | 8.5 |
| CUST004 | 5/20/2022 | PRD055 | 15 |
| CUST005 | 4/21/2022 | PRD013 | 4 |

1. Product Catalog dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Product ID** | **Product Name** | **Price** | **Category** | **Rating** |
| PRD001 | SuperDuper | Electronics | 189.99 | 4.2 |
| PRD002 | MightyGrip | Sports | 49.99 | 4.5 |
| PRD003 | FreshBreeze | Home | 12.99 | 3.8 |
| PRD004 | PowerCrush | Appliances | 199.99 | 4.7 |
| PRD005 | OceanSpray | Food | 3.99 | 4.1 |

1. Website traffic and user behavior dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Page Views** | **Unique Visitors** | **Bounce Rate** | **Session Duration (Seconds)** |
| 2022-03-15 | 500 | 250 | 50 | 60 |
| 2022-03-16 | 750 | 375 | 45 | 80 |
| 2022-03-17 | 1000 | 500 | 40 | 100 |

**Business Requirements**

1. **Using Spark Core (*15 Marks*)**
2. Determine the total revenue earned by the e-commerce company, taking into account the price of products purchased and subtracting any applicable discounts. (*2 marks*)
3. Compute the total discount given by the e-commerce company by summing up the Discount Applied field in the Customer Purchase History dataset. (*2 marks*)
4. Perform an analysis to find the average purchase amount (excluding discounts) for each category of products by joining the customer purchase history and product catalog dataset. (*2 marks*)
5. Identify the days with the highest bounce rate in the Website Traffic and User Behavior dataset. (*2 marks*)
6. Calculate the average rating for each category in the Product Catalog dataset. (*2 marks*)
7. Determine the most popular product category based on the number of purchases in the customer purchase history dataset. (*2 marks*)
8. Identify the top 5 customers with the highest purchase amount by joining the customer purchase history and product catalog dataset. (*3 marks*)
9. **Using Spark SQL (*15 Marks*)**
10. Find the product categories with the highest and lowest average rating in the product catalog dataset. (*3 marks*)
11. Identify the customers who have made a purchase with a discount lesser than or equal to 10 in the customer purchase history dataset. (*2 marks*)
12. Determine the average session duration for each day in the website traffic and user behavior dataset. (*2 marks*)
13. Find the product with the highest rating in each category in the product catalog dataset. (*2 marks*)
14. Determine the average discount amount applied on each product category by joining the customer purchase history and product catalog dataset. (*2 marks*)
15. Identify the top 10 products with the highest total revenue generated by joining the customer purchase history and product catalog dataset. (*2 marks*)
16. Determine the mean bounce rate for the month of March, 2022. (*2 marks*)
17. **Using PySpark Core (*10 Marks*)**
18. Calculate the average session duration in seconds for each month in the year 2021 in the website traffic and user behavior dataset. (*2 marks*)
19. Find the date with the highest number of unique visitors from the website traffic and user behavior dataset. (*2 marks*)
20. Calculate the total number of page views, unique visitors, bounce rate, and session duration for each date from the website traffic and user behavior dataset. (2 marks)
21. Filter the customer purchase history dataset to include only the purchases made on June 13, 2020. (*2 marks*)
22. Filter the product catalog dataset to include only products in the “Electronics” category and sort by price in ascending order. (*2 marks*)
23. **Using PySpark SQL (15 Marks)**
24. Join the customer purchase history and product catalog datasets to create a new dataset with columns: Customer ID, Purchase Date, Product Name, Category, Amount Paid (Price after applying discount). (*3 marks*)
25. Join the customer purchase history and product catalog datasets and filter to only include purchases made on June 13, 2020 to find the total revenue generated for that day. (*3 marks*)
26. Join the customer purchase history and product catalog datasets and filter to include only purchases made by customers with a unique visitor count of 250 or more from the website traffic and user behavior dataset. Then, group the dataset by product name and find the total revenue generated for each product. (*4 marks*)
27. Find the average rating of products in the Clothing category using the product catalog dataset. (*2.5 marks*)
28. Identify the top 10 customers who spent the most money during the month of December 2021 using the customer purchase history dataset. (*2.5 marks*)