Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2021-2022

Assignment II

Course No.	:	DSECL ZC556
Course Title	:	Stream Processing and Analytics
Nature of Exam	:	Take Home
Weightage	:	15%
Duration	:	15 days

No. of Pages = 3
No. of Questions = 4

Streaming Data Analytics

Assume that you are working as analyst for “Pizzario”, a pizza delivery chain. The group has collected some interesting characteristics of customers who had purchased their pizza earlier. (Refer the attached pizza_customers.csv file for the same). The marketing team is planning a campaign to increase the sales of a newly launched pizza. Before that they want to analyze the segmentation of existing customers so that they can have the clearer picture about the customer categories.

pizza_customers.csv

Exercise 1: [5 marks]

Write a Python program that will take the pizza customers dataset as input and produce the clusters which represents the customer segments present in the dataset.

- You may like to do some preprocessing on the given dataset.
- You have to write your own code matching to the problem statement.
- Add comments at appropriate place so that it’s easy to understand your thought process.
- You are supposed to use k-means clustering algorithm (custom implementation not from any library) for customer segmentations.
- The program should clearly output the cluster number, centroid used and number of records belonging to that cluster.
- The final clusters should be preserved in such a way that it can be used in following exercises.

Your marketing team is now aware about the spending behavior of your customers. Wear hat of marketing professional and think of marketing strategy to attract these customers to your newly launched pizza.

Exercise 2: [2 marks]

As an outcome of Exercise 1, you will obtain clusters in the given dataset.

- Apply appropriate labels to those clusters after careful look at the points belonging to those clusters.
    - Explain your logic behind nomenclature of the clusters. Show the sample dataset.
- Based on this segmentation, think of some marketing offers that can be given to the customers.
    - Briefly explain the logic behind the offers to be made to the various customer categories.

By this time you are well aware about your existing customer base. You know the customers spending habit, you have decided upon offers to be made, now it's time to test it out for the existing customers. Let's assume that they frequently visits the mall where your pizza shop is located. For this purpose you have to have the mechanisms through which their visits to the mall are captured.

Exercise 3: [2 marks]

Write a Python program that will simulate the movement of existing customers around a mall in near real time fashion.

- You can assume that customer's cell phones are enabling the transfer of location data to your centralized server where it's stored for further analysis.

With access to the real time movements of customers around mall, you are feeling more powerful and ready to target these customers. Your accumulated knowledge of streaming data processing and analytics can be leveraged for the same purpose. Now think for an architecture to bring it to the reality through a streaming data pipeline.

Exercise 4: [6 marks]

Construct a streaming data pipeline integrating the various technologies, tools and Programmes covered in the course that will harvest this real time data of customer's movements and produces the offers that can be sent on the customers mobile devices. You can think of various aspects related to streaming data processing such as:

- Real time streaming data ingestion
- Data's intermittent storage
- Data preprocessing – cleaning, transformations etc.
- Data processing – filters, joins, windows etc.
- Business logic for placing the offers
- Final representation of the outcome

Submission requirements:

1) Python program for customer segmentation with sample input and outputs
2) Short note on criteria used for cluster nomenclature
3) Python program for customer movement simulation
4) Short note on describing
    - the streaming data pipeline architecture
    - components used and purposes of the same
    - data flows
    - business logic used
5) Programs / Queries used in exercise 4
6) A short demo describing the overall thought process for approaching this problem and data flow through the pipeline. Share the Google drive link for the same.

****