Graduate Certificate Online Examination
Semester I 2020/2021

Subject: ***Big Data Engineering and Web Analytics***

Question	Marks
1	/ 15
2	/ 8
3	/ 12
4	/ 15
TOTAL	/50

Section A

Answer the below questions from both *Analytics* and *Engineering* perspective.

Question 1 *(refer also to Appendix A)* *(Total: 15 Marks)*

The ***GET*** case study is described in Appendix A. You are the data scientist involved in establishing the big data processing capabilities for the new enterprise wide data architecture. Based on the circuit breaker outage situation specified in Appendix A, GET decided to overhaul the existing architecture by migrating to enhanced data lake and real time stream-based event processing architecture. However, to be effective, the data engineers want to categorize the causes for application, order and message outages. *They intend to study the server logs and click stream patterns for understanding user behavior patterns that occur during such outages. Help them implement an efficient solution.*

a. Consider the data enhancement plan described in the case study and analyse the ***three log and clickstream sources*** provided. **Describe** the business (and technical) insights that the company can derive from your proposed log analytics? Also **suggest** relevant ***priority, ingestion tools and data ingestion policy*** for the log and clickstream sources.

(3 Marks)

b. **Identify** a ***key missing capability*** in the current data architecture for dynamic log analytics processing? **List** the different ***processing steps*** you envision for this log analytics data pipeline with appropriate examples drawn from the case study.

(6 Marks)

c. **Suggest** relevant ***processing framework, tools and special libraries*** for achieving the above said log and click stream analytics capabilities. Justify your choice based on the relevance to case study. *(4 Marks)*

d. **Suggest** TWO ***business questions*** that can be answered from your proposed log analytics, that would help GET achieve actionable insight leading to the business outcomes.

(2 Marks)

Question 2 *(refer also to Appendix A)* *(Total: 8 Marks)*

Refer to the GET case study described in Appendix A. You are the data scientist involved in establishing the data lake. Answer the following questions in the context of ***Data Lake***:

a. For the ***two*** new data collection into data lake situations listed below, ***identify appropriate storage model and related tools for processing*** (i.e., A NoSQL data source such as graph, document, in-memory, column-family and key-value stores or *NewSQL* or Raw HDFS or any other useful format).

1. F&B Information, Chef Recommendation, Images of Menu, Review comments from customers, and Location Information. Support for flexible schema is mandatory. Applications built using languages such as Scala and Python needs to communicate frequently to this data. Information scales with time and review usages. Write once read many times access patterns and high availability requirements. Up to 20 million active content browsing customers at any given time.

2. Customer Current Order in Shopping Cart. Accessed always by customer. Persisted when the order is confirmed or the customer logs out. Cart remembers the previous customer conversation state upon fresh login. Cart is actively checked for stock and item availability. Latency expected on cart operation is less than 0.001 milliseconds.

You should advocate for your choice by evaluating against the metrics descriptions mentioned.

(4 Marks)

b. Suggest ***two*** new data store related to *F&B Outlets (example - to increase sales) and delivery services (example - to analyse effect of a promotional event)*. Highlight how you can perform analytics including your suggested data store to add business value to any of the GET business outcomes mentioned in the case study.

(4 Marks)

Section B

Question 3 *(Total: 12 Marks)*

a. GET would like to perform customer segmentation to increase its sales and profits by recommending those food items that customers in the same segmentation often ordered. Given the data in *Sample Click Stream Log* (Appendix A), design a *network based approach* to perform customer segmentation.

(6 marks)

b. Management has decided to enhance delivery providers' operational efficiency when they deliver the food to customers. Design a *network based method* by suggesting the optimal route to their delivery team where they can leverage different transportation modes such as public transport, motorbike, bike, walking etc.

(6 marks)

Section C

Question 4 *(Total: 15 Marks)*

a. Refer to GET objective of usage of IoT data:

GET proposes to have a suitable functional architecture to address the following real-life challenges:

- GET requires optimizing both food delivery and food deliverers. In Singapore, the shops are inside a mall. The deliverers have to queue to get in, find parking, run through multiple levels (elevators, etc.) to get to the shop and return to their vehicle, then get out of the carpark. All these are required to be timed by restaurant locations and built into their delivery.

- Same applies on the other side – food delivered to condos takes time as well. Same with kitchens, everything is required to be timed to the second.

Define and draw the functional architecture in a suitable cloud environment for IoT implementation. Please highlight the key underlying capabilities in the functional IOT Analytics eco-system diagram in order to show how the above challenges can be taken care as “Always-on” IoT eco-system.

(5 marks)

b. Post IoT implementation, GET is required to improve the estimate of the time-of-delivery (ETD), which constantly affects diners’ experience during the entire order lifecycle. For example, whenever diners are wondering how much longer they need to wait for their food, they can check the ETD predictions from GET first.

GET can’t do accurate predictions without traffic data, so GET has already integrated Google Maps API in the architecture.

Please highlight one feature from each of the below events that should improve ETD prediction

i. a diner is browsing the restaurant options
ii. a diner places the order
iii. a delivery partner is matched to pick up the order

(3 Marks)

c. Please highlight the features post IOT implementation (Real time, Near real time, Historical and External API) as you would consider to implement the machine learning algorithms towards the objective of ETD prediction for GET.

(2 Marks)

The management would also like to analyze the energy consumption by different electrical appliances in the smart kitchen to avoid the unnecessary wastage of electrical energy as well as to optimize the workflow. Each kitchen is equipped with smart plugs, which have the ability to measure the energy consumed by appliances and to turn on/off their own outputs. The data collected by smart plugs contain measurement of electricity consumption gathered from different kitchens in different F&B outlets within a continuous period. You are a Data Scientist and is tasked to help conduct the data analytics project for the new GET platform. Answer the following question (d).

d. For the energy usage pattern, one of the applications is to identify anomalous energy usage. If an anomalous energy consumption by the electrical appliance is detected, the system will send warning information through the GET platform. Consider the following two cases pertaining to an electrical appliance:

   i. A normal energy consumption level is not set;

   ii. A normal energy consumption level is set. If the energy consumption perform as normal, it will be predicted as positive; otherwise, it will be predicted as negative.

   For each of the above cases, you are to analyze the situation, identify and recommend an appropriate abnormal event detection approach. As the management will be more concern about undesirable outcomes, based on your recommended approach for each of the cases, evaluate what are the abnormal (not preferred) outcomes in the analysis, in addition to the preferred normal outcomes. Elaborate clearly the evaluation metrics used.

(5 marks)

END