Graduate Certificate Online Examination

Subject: ***Big Data Analytics***

Sample Examination Questions

Section A

Answer the below questions from both *Analytics* and *Engineering* perspective. All Questions in this section are based on the iTrans case study described in Appendix A. Your answers should pertain to the case study scenario and to earn full credit, you should cite examples and/or justifications based on the case context as appropriate.

Question 1 *(refer also to Appendix A)* *(Total: 12 Marks)*

The existing 'Traffic congestion monitoring' use case (system) described in the case study uses two key data points: speed and travel time. LTA computes the speed and travel time for each vehicle in express highway based on its entry and exit toll readings of passing by timestamp and the type of vehicle. The formulas used for calculation currently is shown below:

$$TTTTTTTTTTTT\ TTTTTTTT \quad tt_{ii,jj} = tt_{mm2,ii,jj} - tt_{mm1,ii,jj}$$

where

$tt_{ii,jj}$ - Travel time of the ith vehicle in jth express highway

$tt_{mm1,ii,jj}$ - passing time of this vehicle at entering toll station $T_{m1,\,i,\,j}$

$tt_{mm2,ii,jj}$ - passing time of this vehicle at exiting toll station $T_{m2,\,i,\,j}$

ii - The ith vehicle

jj - The jth express highway

$$SSSSTTTTSS\ 0000\ TT^{tttt}\ VVTTttTTVVTTTT \quad SS_{ii,jj} = \frac{\sum_{mm1}^{mm2-1} LL_{mm,ii}}{tt_{ii,jj}}$$

where

$SS_{ii,jj}$ - Average travel speed of the ith vehicle in the jth express highway

$LL_{mm,ii}$ - Length of "Link m" which the ith vehicle passing

ii - The ith vehicle

jj - The jth express highway

The common definition of congestion in the state of traffic flow is *"the travel demand exceeds road capacity"*. From the travel time perspective, congestion occurs when the normal flow of traffic is interrupted by a high density of vehicles resulting in excess travel time compared to average. Current model computes congestion as a directly proportional measure of travel time and inversely proportional measure of vehicle speed. It is represented as a binary value (i.e., congested **or** not-congested). A sample dataset table is provided below.

Express Highway	Link Segment	Vehicle	Travel Time	Speed	Congestion
34	1	345	22 mins	60KM/HR	Y
34	2	345	16 mins	80KM/HR	N
34	3	345	14 mins	90KM/HR	N
34	4	345	15 mins	80KM/HR	N

a) Is the case scenario and dataset provided complete in identifying all possible traffic congestion inducing patterns? If not, explain why and mention **TWO** additional data points to consider for initial improvements. How will you improve existing model specifically to reflect non-recurring events such as accidents and repair works? What changes are necessary for the same?

(6 Marks)

b) Based on your proposal for above question, identify **ONE** key shortcoming in the existing architecture. Recommend *suitable 'Big Data Technologies'* with justifications that can help in implementing an efficient real-time traffic congestion prediction system. Draw a reference architecture diagram if necessary. Also, recommend **TWO** additional data points necessary for your proposal.

(6 Marks)

Question 2 *(refer also to Appendix A)* *(Total: 11 Marks)*

For the Accidents identification project discussed in case study, video analytics is proposed on batch processing mode. Due to the increasing number of vehicles in circulation in different express highways, several automatic monitoring CCTV camera sensors have been deployed. In particular, roadside cameras are extensively deployed, as they offer vital technological advantages. iTrans is collecting and analyzing increasing amounts of video data making it difficult for traditional on-premises solutions for data storage, data management, and analytics to keep pace.

a) For the Accidents identification mentioned in case study, you want to implement a self-servicing data storage facility and use query-in-place analytics tools. Propose a new storage facility and necessary implementation steps to satisfy the above requirements. From a platform scaling perspective, suggest public cloud component offerings you will consider and draw a reference architecture to reflect your proposal.

(8 Marks)

b) In data processing of the iTrans platform, there are different types of file formats to store the data. Previously, the traffic speed data is downloaded from LTA DataMall port and ingested into the storage system in CSV format. The attributes are fixed after the data is ingested into the storage system. But as now LTA modifies the datasets by removing and adding attributes to cater to the requirements from clients. The CSV format cannot cater to such change. And more data are collected from different data sources such as cameras, mobile devices etc. Due to the growth of large amount of data and the variety of data sources, the performance of the iTrans platform is not satisfied as it takes more time to read/write data in CSV format. Considering the above scenario, you are asked to recommend TWO appropriate data formats from Optimized Row Columnar (ORC), AVRO and Parquet and justify your choice. When you do the justification, you need to compare these THREE different data formats based on the following aspects: (1) Support both structured and un-structured data (2) Have high compression ratio so as to reduce the size of data being stored (3) Support scalability (4) Support scheme modification/evolution (5) Storage efficiency.

(3 Marks)

Section B

Question 3 *(Total: 12 Marks)*

As part of the Big Data Specialist team, you are asked to build an APP to provide route recommendation for drivers to guide them to reach their destination efficiently. Note the road conditions across the city could be changed dynamically due to real-time demand, weather condition and accidents etc. More specifically, you are asked to perform the following two tasks:

i) While existing map systems (APP) can provide drivers with some recommended roads, it did not predict each road's future road conditions (i.e. driving speed) in advance. As a driver will typically take some time to reach his/her final destinations, and the road conditions could be changed over time and lead to delay, originally planned route by map systems may not always be useful and valid during this journey. Your task is to suggest a method which is able to dynamically predict each road's future driving speed in advance based on some important and relevant perspectives.

ii) Construct a road network and design a **network** based route recommendation method by dynamically and proactively suggesting the optimal route for drivers during their journeys based on road conditions, given his/her current source location to the final destination in shortest time.

(12 marks)

Section C

Question 4 *(Total: 15 Marks)*

a. Based on the scenario envisaged under the section Enhancement to the Platform in Appendix A (Case Study), you have been asked to vet through the ***Architectural Blueprint*** of the enhanced platform (Refer to Figure 1). The platform is expected to be able perform traffic congestion monitoring and identification of accidents in real time.

The platform architecture consists of the following 4 layers.

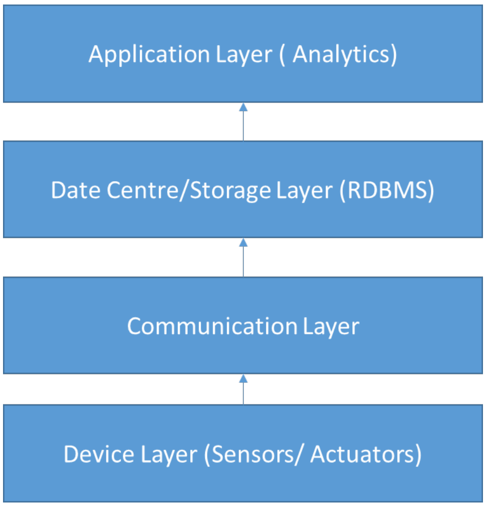
Application Layer (Analytics)
Date Centre/Storage Layer (RDBMS)
Communication Layer
Device Layer (Sensors/ Actuators)

Figure 1 (IoT Architecture)

Evaluate the requirements for the proposed enhancements to the platform. Based on the requirements, recommend changes/improvements to the architecture. This can include add/remove/update of layers or layer related information. Include any details that are important for the efficient functioning of the platform. Provide appropriate justification for the recommendations. Redraw the architectural blueprint with the proposed changes.

(6 marks)

b. Predictive analytics of traffic and transport patterns can reduce congestion and improve the efficiency of public transport services. Predictive analytics can also help to reduce traffic accidents.

i) Recommend TWO (2) ways in which predictive analytics can be used to help reduce traffic accidents.

ii) Identify any THREE (3) data sources that you will need to perform the analytics along with justification for the choices made to ***reduce traffic accidents***.

iii) Identify THREE (3) key features that will help you perform predictive analytics to ***reduce traffic accidents***. Explain the choice of features.

(4 Marks)

c. Refer to the crowd mobility analysis mentioned in the case study, the Call Detail Record (CDR) is used to analysis crowd mobility of people. The CDR Dataset is anonymous cellular phone signalling data. It contains anonymous location estimations from cellular devices which are generated each time the device connects to the cellular network,

The actual dataset contains eight numerical attributes, the description is shown in the following table:

Attribute name	**Description**
square_id	The square grid ID where the activity has happened
country_code	The country code, e.g. 65 represents Singapore
time_interval	Timestamp information when the activity has been started, it is in Unix timestamp format
sms_in	Duration of the message received activity
sms_out	Duration of the message sent activity
call_in	Duration of the call in activity
call_out	Duration of the call out activity
internet_traffic	Duration of the internet activity

A sample dataset table is provided below.

square_id	**time_interval**	**country_ code**	**sms_in**	**sms_out**	**call_in**	**call_out**	**internet_traffic**
1	1488530800	65	2	NA	NA	5	NA
1	1488530800	33	NA	NA	50	NA	NA
1	1488530800	65	1	8	NA	NA	NA
1	1488530800	41	4	57	23	27	11
1	1488600000	0	2	NA	NA	NA	NA

Based on the use case and the sample data, what is the problem with the current raw data? Can you use it directly for crowd mobility analysis? Proposed FOUR pre-processing steps involving data type conversion and feature generation.

(5 Marks)

END