Institute of Systems Science
National University of Singapore

MASTER OF TECHNOLOGY IN ENTERPRISE BUSINESS ANALYTICS

Graduate Certificate Online Examination
Semester I 2021/2022

Subject: *Big Data Analytics*

Instructions for Paper

Date: Friday 12 Nov 2021
Time: 6.30 p.m.
Duration: Three hours (7.00 p.m. to 10.00 p.m.)
Place: Online Examination

This is an OPEN BOOK examination. This examination paper consists of *three* Sections and *six* Questions. You are to answer *ALL* questions. There are a total of *50 Marks* for this paper.

1. Read **ALL** instructions before answering any of the examination questions.
2. There will be only **ONE question paper**, which may be organized in a series of one or more sections (section A, B, C etc.) with/without appendixes. Each section contains one or more questions. You are required to answer **ALL** questions in the **SEPARATE answer booklet(s)**, according to different sections, downloaded from LumiNUS.
3. The first 30 minutes will be reading time, during which you **must not** start answering your answers.
4. Write your NUS Student ID number on the **front page** of the Answer Booklet(s) in the box provided.
5. This is an ***Open Book*** examination. If you wish, you may use reference materials to answer a question. Reference materials can be books, manuals, handouts or notes, including e-notes on PCs/laptops.
6. All answers provided should be in **digital format**. However, you can **hand draw** diagrams on paper using pens and ensure that it is readable as an image. **Insert this image into your answer booklet**. Do not send images as separate files.
7. Non-programmable calculators may be used if required.
8. **Internet access (except LumiNUS and Zoom) using computers of any form e.g. laptops, tablets, smart watches etc. is not permitted during the examination**. Students who are found with suspected academic dishonesty that give them unfair advantage during assessments will be subjected to disciplinary action by the University, as laid out in NUS Code of Student Conduct.
9. State clearly any assumptions you make in answering any question where you feel the requirement is not sufficiently clear.
10. At the end of the examination:
    a) Convert the answer booklet to PDF format and compress them if necessary.
    b) Please name the PDF file with your **Student ID number** prefaced by the abbreviation of the Course and the Section e.g. **BDA_SecA_A0123456X**.
    c) Upload the answer booklet for **each Section separately**.
11. After submission of answer booklet(s), please wait for Proctor to make announcement on the closure of examination.

SECTION A

Refer to the case study in **Appendix A** and answer the following questions.

Question 1 *(Total: 10 Marks)*

Given the current platform architecture and data sources descriptions, and the aspiration to build a new real-time recommender model for the EatZi case study, answer the following questions and provide justifications for your design choices appropriately:

a. Which kind of ingestion is suitable for the "User Browsing Data" listed in data sources? How will you store the user browsing clickstream data? What are the operational constraints identified? Justify your storage decision against the constraints identified.

(4 Marks)

b. Currently the restaurant data is stored in MySQL RDBMS. As the system scales, will you want to migrate the data store? If yes, to which kind of data store? If no, why is that?

(2 Marks)

c. Identify two critical use cases beyond the current recommendation system and search data processing of current data platform detailed in EatZi case study. Also, briefly explain what tech stack is necessary for your processing proposal.

(4 Marks)

Question 2 *(Total: 10 Marks)*

As the analytics expert, you are proposing a new restaurant sales predictive analytics project for EatZi platform. You plan to use the current historical as well as new real-time data to forecast future strengths, weaknesses, and trends. The project helps to predict the daily, weekly, and long-term sales for EatZi. Design the architecture for this restaurant sales predictive analytics project. Identify the various processing layers. Label the right tools and technologies for each of these layers. The collected Data Set is as described below:

Dataset consists of following fields:

- Id: Restaurant id.
- Open Date: opening date for a restaurant
- City: City that the restaurant is in.
- Type: Type of the restaurant. (FC: Food Court, IL: Inline, DT: Drive Thru and MB: Mobile)
- P1 - P37: These are the p-variables and are a measure of the following:
    - Demographic data: Population, age, gender distribution, development scale
    - Real Estate Data: M2 of the location, front façade, car parking etc.
    - Commercial data: schools, banks etc.
- REVENUE: The revenue column indicates a (transformed) revenue of the restaurant in a given year and is the target of predictive analysis.

SECTION B

Refer to the case study in **Appendix A** and answer the following questions.

Question 3 *(Total: 8 Marks)*

EatZi have asked you to design a new session-based recommender system. The goal is to make individual (not category) restaurant recommendations to users during their browsing session on the EatZi platform. Recommendations are to be made only when the system decides it that it has one or more suitable recommendations to make, hence it is acceptable for some user sessions to not receive any recommendations. How to measure the suitability of a recommendation is left for you to decide, however, you know that a suitable recommendation is one that the user will be happy (or neutral) receiving and will hence not create a negative browsing experience for the user. The recommendations are to appear in a small pop-up window during the user browsing session. You plan to build this recommender system using an association rule approach using the Apriori algorithm to generate the association rules.

Answer the following questions:

a. What data will you use to build the association rules: what are the items/entities that will appear in the association rules and which variables in the data tables will you use to obtain these? How will you group the items/entities into item-sets (e.g. into baskets)? Please describe any choices you made, pre-processing required and assumptions made.

(2 marks)

b. How will the learned association rules be used to make recommendations during a user browsing session? Should all of the learned recommendation rules be used? How will you ensure that only suitable recommendations are made?

(3 marks)

c. Should the association rules be generated during the user-session? Explain your answer.

(1 mark)

d. After your system has been built and deployed for a few months, EatZi begin getting complaints from some of the smaller restaurants that are listed on the EatZi platform that your recommender system never recommends their restaurants. Describe the possible causes for this and suggest how it might be fixed? Your suggested fix/fixes should require only minimal changes and should not involve changing the algorithm/method used. Do you think your fix will succeed, explain your answer?

(2 marks)

Question 4 *(Total: 7 Marks)*

EatZi are still not satisfied with the performance of your recommender system and ask you to build an alternative recommender system using item-based collaborate filtering.

Answer the following questions:

a. You start by building a user-item ratings matrix using only the user rating variable in the user review data table, however after testing you find that this gives very poor results due to the extreme sparsity of data in the ratings matrix (few users supply explicit ratings). Suggest how this might be corrected by utilising more of the available variables. State which additional variables you will use. If you propose combining different variables together then describe any preprocessing required and the method for combining the variables.

(4 marks)

b. You are happy - after implementing the method you suggested in (a) you observe that, on average, all users have now been assigned a rating for at least 50% of the restaurants. Since you now have much more ratings data and the number of users has grown to many millions, you decide to optimize your system by storing the ratings data in memory using a sparse matrix representation. State if this decision is likely to be a good decision or a bad decision, give the reason for your answer and describe the advantages (or disadvantages) of this decision.

(2 marks)

c. You decide to build a hybrid system by combining the item-based collaborative filtering with user-based collaborative filtering. Is this a good decision in this situation? Explain your answer.

(1 mark)

SECTION C

Refer to the case study in **Appendix A** and answer the following questions.

Question 5 *(Total: 8 Marks)*

EatZi is still a small organization and needs confirmation on the idea of near-real-time recommendations before going for a big investment for this project. The management has asked you to design a new session-based recommender system that accounts for user's behavior on the platform in near real-time and provides the best possible recommendations based on other details like location or any other applied filters. A lot of new big data-related technology might get on boarded to test the use-case needs, and if proven, project will be using the technology.

As a big data scientist, first, you need to understand the core deliverables of this project and decide on a type of methodology you will be using to run this project. Second, clearly describe the use-case needs to the key stakeholders into a single framework in an easy-to-understand language.

Answer the following questions:

a. Given all the details, can you select the correct framework and specify the requirements for this use case.

(5 marks)

b. Given the needs of this project, what can be the data pipeline blueprint, you will be suggesting to the stakeholders. State your assumptions and constraints for the design, if any, and new data sources will be integrated with the system over time, and data preparation, processing & validation steps at each stage will be crucial before being served to the recommendation engine.

(3 marks)

Question 6 *(Total: 7 Marks)*

Before the recommender system is deployed to the production environment, you are requested to evaluate the performance of the model. You use the data collected from Jan 2020 to March 2020 as training dataset to build the model, and test the module using data collected from April 2020 (From 7 Apr 2020, Singapore entered the circuit breaker period). However, you noted that there is a degradation of the model's prediction power.

Answer the following questions:

a. You noted that the performance of the model degraded. Assume that there is no issues with the data sources, data pre-processing and recommender algorithm. What is the possible cause for it? Explain your answer and suggest TWO possible methods to address this issue.

(4 marks)

b. Assume that you have solved the issue raised in (a). You decided to collect more data to test your model. Previous users who registered with EatZi platform are domestic users. As the company has expanded its business to other parts of Asia as per the case study, you have opportunities to collect users' data from foreign countries. You noted that the model performance degraded again. What is the possible cause for it? Assume that there are no issues with the data sources, data pre-processing and recommender algorithm. Explain your answer.

(3 marks)

APPENDIX A

EatZi– Case Study

A Platform for Restaurant Recommendation
(A Hypothetical Case Study)

Introduction and Background

EatZi is a local food guide platform which is developed by a local company. With EatZi, users can easily find the best food choices and places. It also includes customer reviews and recommendations.

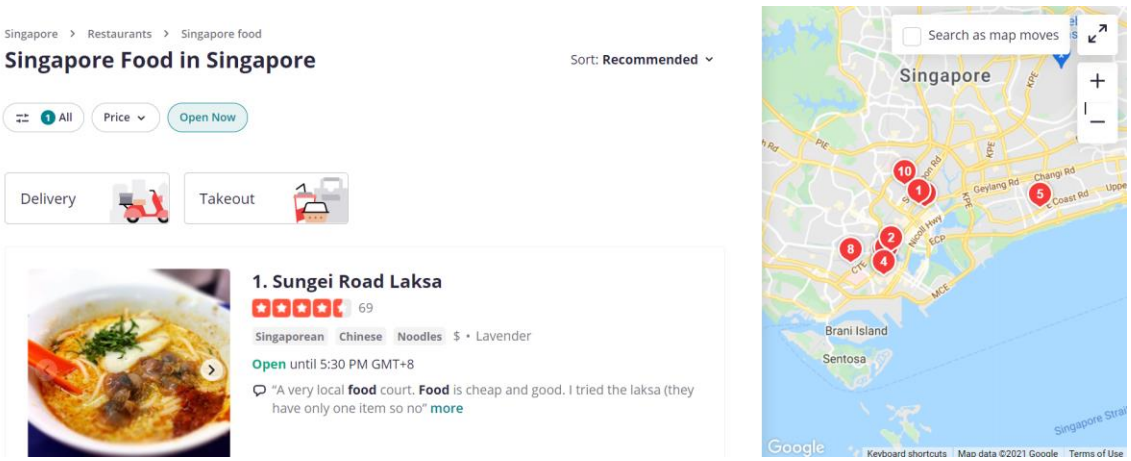
People typically choose restaurants either based on their friends' recommendation, or by finding an interesting outlet. Now, with the EatZi platform, people can check the information online before physically visiting a restaurant. They can check the reviews, price ranges and hours of operations. After evaluating the prices, operating hours and reviews, people can make a better decision whether to visit the restaurant or not.

Current System

The current EatZi platform supports its two key stakeholders: users and restaurant owners.

- Users can browse and search on the website to find the right restaurant, which has the highest potential to satisfy their food preference and appropriate for their budget. They can also have the option to post their reviews and ratings. Only registered users with EatZi may use the platform by signing in.
- Restaurant owners can log in to the platform and upload the restaurant information including location, menus, prices, operating hours, etc.

Figure 1 - Example of an online food guide platform



Source: https://www.yelp.com.sg/search?find_desc=Singapore%20Food&find_loc=Singapore&open_now=5006&sortby=recommended

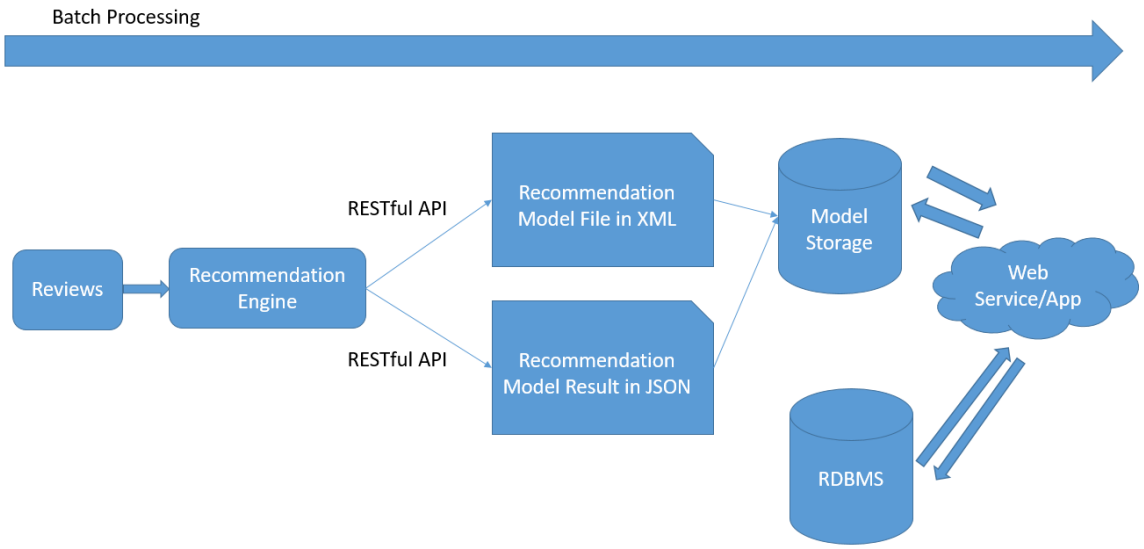
The current platform provides several services for the users:

- Restaurant Search. Users can search on the EatZi website or the app to find a suitable restaurant, e.g.name, location, route etc.
- Ranking. Users can browse the reviews and menu of the restaurants and can rank them based on different sorting strategies, such as Most recommended, Highest rated, Most reviewed, etc..
- Selection. Users can filter the restaurants by prices, category, and opening hours and select the place that best suits their needs.

- Option to Review. Users can post reviews on the website/platform for a specific restaurants that they visited.

The current EatZi platform provides a personalized recommendation engine for food using its own database. The recommendation system makes suggestions on restaurants for each active user by studying their preferences based on their past reviews and feedback.
The general data processing (recommendation engine) of EatZi platform is described in the Figure 2 below:

Figure 2 - EatZi Data Processing



The data processing of the current platform is mainly batch processing. The system processes all the prior data generated batch by batch.
The data process pipeline takes original reviews, extracts important words, builds a topic model and trains a machine learning algorithm by using recommendation engine. The final output of the recommendation engine are saved in XML and JSON format. The saved model is used when the recommendation is made for each of customer.
When the website / app gets a general request for recommending restaurant for a customer, the user ID is passed into the model storage. Based on the user ID, user's model is retrieved from model storage and the selected model is run. The process generates a list of restaurants ranked by user's machine learning model. All the users and restaurants information are stored in MySQL RDBMS.

Data Sources

The main datasets are shown in the tables below.
User data

Columns	Data Type	Description
user_id	INTEGER	User ID
registered_email	STRING	User registered email
first_name	STRING	First name
last_name	STRING	Last name
gender	STRING	Gender
country	STRING	Country

favorite_language	STRING	Favourite languages
signup_date	DATETIME	Date and time that the user signed up (created their EatZi account)
last_active_time	DATETIME	Date and time of the user's last login

Restaurant data

Columns	**Data Type**	**Description**
country	STRING	Restaurant register country
category	STRING	Category of restaurant
restaurant_id	INTEGER	Unique id of restaurant
restaurant_name_en	STRING	Restaurant registration name
latitude	FLOAT	Latitude of restaurant location
longitude	FLOAT	Longitude of restaurant location
restaurant_description	STRING	The description of the restaurant, typically one or two paragraphs
opening time	TIMESTAMP	Opening time of the restaurant
closing time	TIMESTAMP	Closing time of the restaurant
days of week open	LIST	Days of week when the restaurant is open.
menu	STRING	Text description of the menu provided by this restaurant.
item_list	LIST	List of item_ID's for the menu items offered by the restaurant (e.g. chicken rice, cheese pizza etc)

Menu data

Columns	**Data Type**	**Description**
restaurant_id	STRING	Restaurant ID
item_id	STRING	ID of the menu item
text	INT	Text description of the menu item
price	FLOAT	Price of the menu item in dollars

User review data

Columns	**Data Type**	**Description**
restaurant_id	STRING	Restaurant ID
user_id	STRING	User ID
rating	INT	User rating assigned to the restaurant on a scale of 1 to 10 (1 = very bad, 10 = excellent)
rating date	DATETIME	Date and time that the rating was made
review text	STRING	User's text review for this restaurant
review date	DATETIME	Date and time that the review was made

User search data

Columns	Data Type	Description
user_id	STRING	User ID
search_query	STRING	The user's search query
search_date	DATETIME	Date and time the search was made
clicked_ID	STRING	The ID of the content that was clicked by the user in the search results. This field will be blank if no result was clicked

User browsing data

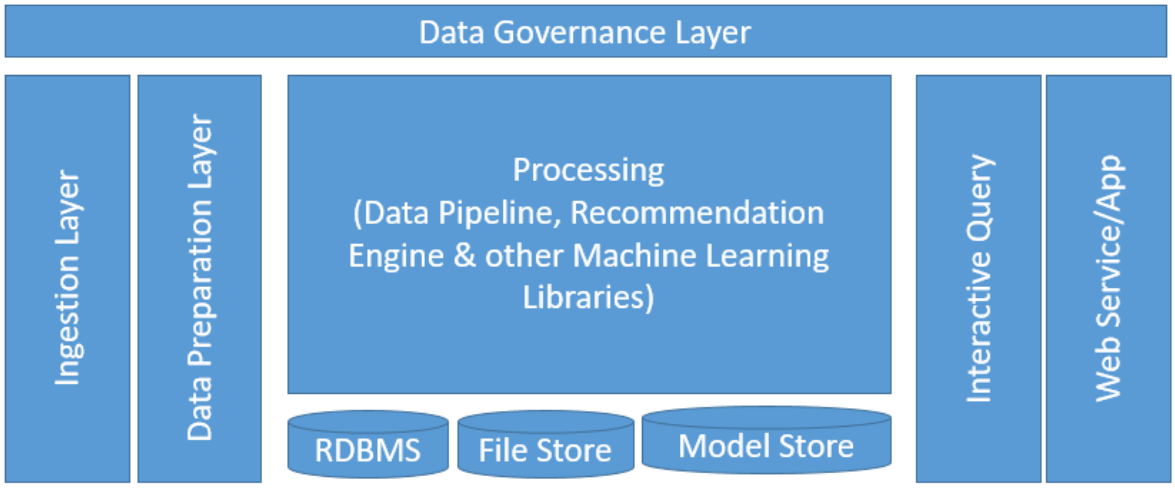
Columns	Data Type	Description
user_id	STRING	User ID
view_date	DATETIME	Date and time of the browsing event
Content_id	STRING	The ID for the content viewed*
Content_url	STRING	URL for the content being viewed
session_id	STRING	The ID assigned to the user browsing session
browser_type	STRING	Type of user's browser, e.g. Chrome, Edge, Firefox

**Note: Content_id will normally be either a restaurant ID or a menu item ID depending on the content being viewed. A small number of content IDs refer to account detail pages and navigation pages*.

Current Architecture

The current architecture of EatZi platform deals with structured and semi-structured data. The users' data is handled via MySQL RDBMS. The main layers involved in the current architecture are ingestion, data preparation, data processing, recommendation engine, query interaction. The current data architecture is show in Figure 3 below:

Figure 3 – Current Architecture



Challenges Encountered Due to Growth

Recently, the company decided to expand its current business in other Asian Markets. This will include food and restaurant products launched on its website and mobile application. This will enable enriched user experiences on EatZi platform and meet users' rising demands. Due to the expansion, the number of users has increased to few millions within few months. This

caused tremendous stress and load on the current EatZi platform. The current platform doesn't support any real-time analytics. The company wants to build a recommendation system which is able to respond based on users' recent browsing activity. So, the company decided to add new capabilities to the platform to improve the situation.

Management Directions for Analytics and Applications

The company wants to improve the Enterprise-level data architecture to ensure that the disintegration of the integrated data stores is minimized. This will enable the EatZi platform to produce more meaningful insights and decisions that are scientifically enabled by effective use of analytics. To enable this automated statistics and machine learning algorithms will be embedded in their data architecture.

- The recommendation system should be able to study users' preferences based on users' behaviors on the platform, perform personalized calculations, and make restaurant recommendations for each active user.
- Another important requirement is that the recommendations made by the platform, need to be real-time. It should be able to provide real-time response to users. Through a real-time personalized recommendation system, suggestions for restaurants will be made to suit people's own preference and current location.
- However, the browsing time of users is around 10 minutes. Users will probably come back to the platform after a few months. Their preferences of restaurants could be totally different from last time when they visited EatZi. As a result, the recommendation engine would focus on single session recommendation.
- Update frequency of the model should also be considered. Features need to be updated in almost real time. For example, a user just browsed a new restaurant, and this user event is immediately updated to the user's behavioral data and being consumed in the next round of model calculation.
- New techniques need to be considered to enrich the current EatZi platform. E.g., big data technologies can be brought in, to process this massive volume of streaming data in a short time.

You are a big data scientist and you are hired by the company to help design the architecture to enrich the current EatZi platform to resolve the aforementioned issues for enhancement. You are required to deal with various topics such as redesign the current system architecture, use new big data technologies to handle massive amount of data and new data resource, enhance current recommendation system, develop new data analytics module, complete data pipeline etc.

Closing Remarks

In addition to the current capabilities of the existing platform, the proposed new platform will enable and integrate more enhanced capabilities. It will have the ability to process and handle increased amount of structured and unstructured data from a larger variety of sources (apps on phones, website etc.). The platform will clean and transform the data and use it to make predictions and analysis to achieve specific outcomes. The enhanced platform will incorporate, predictive and reasoning ability as it will store and make use of the data modelling via statistical and machine learning algorithms based on individual behaviors, preferences, and habits.