CSCI 5408

Data Management, Warehousing, and Analytics

DALHOUSIE UNIVERSITY
FACULTY OF COMPUTER SCIENCE

Test 2

FULL NAME	**Aravind Govindarajan**
Banner Number	B00803336

Instructions

Please read these instructions carefully and thoroughly.

1. This is a take-home exam. You will have 24 hours from the time it is made available until the time it must be submitted on Brightspace.

2. Answer all questions on this document. You may embed diagrams and pictures if desired. You are to submit a PDF version of this document on Brightspace.

3. You are to complete this evaluation individually. Any collaboration or obtaining assistance from others will be considered a violation of academic integrity.

4. You may use online resources (e.g., you can look things up in Google). If you do so, you must provide valid and accurate references and citations for anything you viewed. Using anything (e.g., diagrams, ideas, text) that you did not create yourself is a violation of academic integrity when you do not identify the source.

5. You may use point form if desired and do not have to write full sentences provided that your answers are clear and understandable. As a spell-checker is available, spelling will be considered in your answers.

6. There are 8 questions of equal weight; all are worth 10 marks. The questions vary in terms of difficulty and answer length. Do not assume that every question has 10 "key points" necessary in the answer.

Question 1: Identify 10 different uses for clustering.

Answer: Clustering is the process of grouping data according to similarity and difference. Some of its real-time uses are:

1. **Medicine**: Finding patients with similar symptoms or finding similar diseases upon which a medication is effective.
2. **Biology**: Identifying similar species and environments.
3. **Marketing:** Identifying similar customers for target marketing.
4. **Civil**: Grouping houses or public buildings according to their type, usage, value, and location.
5. **Social Science:** Performing qualitative research. For example, we can identify the group of individuals who think, feel or behave the same way in particular situations.
6. **Insurance**: Finding groups of policyholders with a high average claim cost.
7. **Network:** Identifying the locations for tower installation in a particular region. The location of installing these towers can be found by clustering algorithm so that all its users receive good signal strength.
8. **News**: Grouping news from different sources based on their importance. For instance, Google News clusters content from more than 25,000 publishers.
9. **Social network:** Identifying the group of people with similar interests, likes, and friends. For example, Facebook provides friends suggestions by identifying people of similar interests or similar friends circle.
10. **Recommender systems:** Producing a list of recommendations through content-based filtering. Examples: Netflix, Amazon Music.

Question 2: Explain the differences between OLAP and OLTP. You will not receive any marks for telling me what they are; the marks are awarded for explaining their differences.

Answer:

OLTP	OLAP
1) Online Transaction Processing (OLTP) manages transaction-oriented applications.	1) Online Analytical Processing (OLAP) system reports to multi-dimensional analytical queries like reporting and forecasting.
2) Online database modifying system.	2) Online database query answering system.

3) Uses traditional RDBMS.	3) Uses data warehouse.
4) OLTP uses normalized tables.	4) OLAP uses non-normalized tables.
5) It is used by data critical users like database administrators, clerk etc.,	5) It is used by data knowledge users like managers, CEO.
6) The processing time of a transaction is comparatively less in OLTP.	6) Since complex analytical processes are involved, the processing time is comparatively more in OLAP.
7) OLTP and its transactions are the original source of data.	7) Different OLTPs database becomes the source of data for OLAP.
8) It has short transactions in large numbers.	8) It has long transactions.
9) The space or memory requirements of OLTP can be relatively small since the historical data can be archived.	9) Space requirements are large because of existence of aggregation structures.
10) Periodical backups are required since the operational data is critical for the business.	10) Backups are not very important when compared to OLTP.
11) The main use of OLTP is to control and run fundamental business tasks.	11) The main use of OLAP is to help with planning and decision making.

Question 3: Why is sentiment analysis useful?

Answer: Using sentiment analysis, we can identify the opinions expressed in a piece of text and detect the emotional tone behind it. Sentiment analysis are useful in many ways.

Targeted Marketing : Using sentiment analysis we can find the people who are positive about a product but haven't purchased it. We can also find out what else they like and make specific offers to them. Ideally, we can get insights into our customer behaviour, what our customers want, their likes and dislikes, their decision process, and so on.

Customer Service : Sentiment analysis can be used to identify problems and complaints before they become major issues. We can identify the common problems and prepare solutions for them. Also, we can be proactive in contacting clients before they use contact centres or other assistance. For example, companies like Uber, Dominos etc., use the twitter data to identify the problems faced by the customers and improve the customer service by providing optimal solutions.

Filtering content: Some services try to maintain a style or level of approachability. Sentiment analysis can be used for filtering, sorting and screening content. It allows us to identify and block inappropriate content.

Sentiment analysis helps in market research by understanding what the customer wants and how we can align our products and services with their tastes. This in turn helps in improving the quality of the product/service delivered. Sentiment analysis also helps in crisis management. This is because, the constant monitoring of what is happening in social media helps us in mitigating the damage that might occur in the near future. And finally, sentiment analysis boosts sales revenue. This is the final outcome of successful marketing campaigns, improved customer service and product quality.

Question 4: You need to build a database of tweets about Dalhousie. What issues will you face with respect to obtaining, cleaning, and storing this data?

Answer: There are quite a lot of issues for the given scenario with respect to obtaining, cleaning, and storing the Twitter data.

Obtaining the data:

- For fetching the Twitter data, it is necessary to create a Twitter developer account. Only when the account is successfully verified, we will be able to access the security credentials (API keys) and fetch the tweets. There are more chances for this verification process to be slow. Sometimes it takes 2-3 days and sometimes it takes a week. This impacts the project schedule.
- For the given scenario, it is apparent that live Twitter data should be streamed. Unless the tweets' size (number of tweets) is capped, we will be forced to run the program for an infinitely long time which affects its performance. And also, the streamed data will ideally be stored in a CSV file before feeding it in a database. There are high chances for the size of the CSV file to increase if the tweets' size is not capped.
  Identifying the cap limit is also a challenge.

Cleaning the data:

Tweets contain a lot of unwanted terms and characters that have very less importance. Hence, it is necessary to clean the tweets before feeding into a database system. Some of the issues in the Twitter data are:

- Presence of smileys and emoticons.

- Presence of stop words (the, is, a, etc.,).
- Presence of URL links.

One problem with removing the above-mentioned items is that, it makes the tweet incomplete. Now, twitter data cleaning is necessary only if we are going to perform sentiment analysis. In the given scenario, it is not mentioned what the database is built for. So, deciding whether to clean the twitter data or not and deciding what to clean is a challenge.

Storing the data:

For the given scenario, the volume of tweets might be huge. Deciding whether to choose RDBMS or NoSQL database can be a big challenge. The Twitter data can be stored in a simple RDBMS system (provided the tweets are cleaned and structured), because just like NoSQL databases, RDBMS can also handle large amount of data. We can also go for NoSQL databases like MongoDB when we want the database to act as both the data store and analytics engine. Both RDBMS and NoSQL have their own advantages and disadvantages, so choosing the right database for the given scenario can be a challenge (should make trade-offs).

Question 5: A friend of yours, who sells cell phone accessories online, wants you to build them a system to perform business analysis on their sales data. What are the first 5 questions that you will ask them? That is, what are the 5 most important things you need to know to begin building them a system.

Answer: According to me, the important things that I need to know to begin building the requested system are:

1) What is the problem that need to be addressed?
2) Where is the data coming from? (What's the source of data?)
3) What kind of business analysis application is required? Reporting application or data mining application?
4) What actions are you trying to derive from the system?
5) What are the questions that can't be answered easily today but should be answered by the business analysis system?

Question 6: How is ElasticSearch different from other kinds of NoSQL databases? What unique features does it contain that ensures fast searching and data retrieval?

Answer: NoSQL databases are primarily used for storing a huge amount of data. Elasticsearch, on the other hand, provides capability to store, index, search, and analyze data in real-time which allows us to extract value from the data.

Elasticsearch is able to achieve fast search responses because, instead of searching the text directly, it searches an index instead. It is like retrieving pages from a book related to a keyword by scanning the index at the back of a book, as opposed to searching every word of every page in the book. This type of index is called an inverted index, because it inverts a page-centric data structure (page to words) to a keyword-centric data structure (words to pages). Elasticsearch uses Apache Lucene to create and manage this inverted index.

Elasticsearch also uses caching techniques for faster searching and data retrieval. It supports three kinds of caches namely, the node query cache, the shard request cache, and the field data cache.

- The node cache is shared by all the shards on a node. It caches the result of queries being used in a filter context.
- The shard request cache caches query results independently for each shard.
- The field data cache stores the field values while computing aggregations.

Question 7: How is a NoSQL database different from a relational database? For mission critical applications (i.e., those upon which the survival or a business or project depend), which database type would you use and why?

Answer: NoSQL and relational database are different in many ways. Relational databases are table-based databases whereas NoSQL databases cane be document based, key-value pairs, graph databases or wide-column stores. Relational databases are best suited for structured data, while NoSQL databases are best suited for semi-structured and unstructured data. Relational databases use Structure Query Language (SQL) for manipulating the data and it is very powerful. The syntax for SQL is almost the same across different vendors. On the other hand, queries in NoSQL are focussed on collection of documents. The syntax for unstructured query language varies from one vendor to another. Relational databases are vertically scalable, while NoSQL databases are horizontally scalable. Relational databases have predefined schema whereas NoSQL databases have dynamic schema.

For mission critical applications, I would choose relational databases because of some straight-forward reasons. Relational databases enable increased interaction with the data. Since it uses a declarative query language, we can state what we want. Interaction is the key for mission critical applications because data that is not interactive is useless. Relational databases also allow us to link information from different tables. Since mission critical applications will have a lot of transactions, it is important that the transactions are processed reliably. The ACID properties of relational database guarantee this reliability. To conclude, mission critical applications demands a lot of complicated querying, database transactions and routine analysis of data and I believe that relational database meets these demands.

Question 8: You are creating a system to categorise MRI images of brains according to various properties of the image. You have been asked to use a machine learning approach. Why will it be difficult to use supervised learning?

Answer: For creating a system to categorise MRI images of brains according to various properties of the image using machine learning approach, unsupervised learning is the best choice.

Supervised learning is a good option if we are trying to predict whether someone has brain cancer and we have a bunch of brain scan images that are labeled as either being of a brain with cancer or a brain without cancer. But that is not the case with the given scenario. We are not told what the answer should be and how the categorization should be made. We are trying to learn the structure of our data without using explicitly provided labels. Hence, unsupervised learning is the optimal solution.

REFERENCES:

[1] System, O. (2018). *Difference Between OLTP and OLAP (with Comparison Chart) - Tech Differences*. [online] Tech Differences. Available at: https://techdifferences.com/difference-between-oltp-and-olap.html

[2] OpenSource Connections. (2018). *Caching In Elasticsearch*. [online] Available at: https://opensourceconnections.com/blog/2017/07/10/caching_in_elasticsearch/

[3] Stat.columbia.edu. (2018). [online] Available at: http://www.stat.columbia.edu/~madigan/W2025/notes/clustering.pdf

HAVE A GREAT DAY!

(end of the exam)