ELG 5166 – Cloud Analytics

Assignment #1 - (Teamwork)

Due: Please upload to Brightspace before midnight October 8th, 2022.

This is a teamwork assignment (**Only one submission per team**).

Weight: 15% of the final mark**.**

Important Note: Read the following academic integrity statement, type in your full name and student ID, and include a copy in your submission. Submitting this form electronically by one of the team members is considered the same as signing the document by all members of the team.

Personal Ethics & Academic Integrity Statement

Student name: Student ID:

Student Name: Student ID:

Student Name: Student ID:

Student Name: Student ID:

By typing in my name and student ID on this form and submitting it electronically, I am attesting to the fact that I have reviewed not only my work but the work of my team member, in its entirety.

I attest to the fact that my work in this project adheres to the fraud policies as outlined in the Academic Regulations in the University's Graduate Studies Calendar. I further attest that I have knowledge of and have respected the "Beware of Plagiarism" brochure for the university. To the

best of my knowledge, I also believe that each of my group colleagues has also met the aforementioned requirements and regulations. I understand that if my group assignment is submitted without a completed copy of this Personal Work Statement from each group member, it will be interpreted by the school that the missing student(s) name is confirmation of the non-participation of the aforementioned student(s) in the required work.

We, by typing in our names and student IDs on this form and submitting it electronically,

- warrant that the work submitted herein is our own group members' work and not the work of others
- acknowledge that we have read and understood the University Regulations on Academic Misconduct
- acknowledge that it is a breach of University Regulations to give or receive unauthorized and/or unacknowledged assistance on a graded piece of work

Note:

This assignment puts you in the position of a consultant/analyst who is using her/his knowledge of the course to address a real-world problem. There is no "unique" or "best" solution for this assignment question.

Additionally,

- Keep your answers short and succinct.
- Use a diagram if it helps demonstrate or illustrate your answer. Diagrams without appropriate content description and reference will not count as valid responses.
- Please cite all references properly and provide a bibliography or a reference section if needed.

Part 1 (50 points)

1. Describe briefly what a NoSQL database means. Select a NoSQL database (**except MongoDB & Cassandra**) and describe how this database can be used for the storage and management of big data. (10 pts)

2. Investigate and describe one application of Big Data Analytics that was not described in class (10 pts)

3. Briefly describe the transaction management features of Cassandra and MongoDB in the context of ACID vs. BASE properties (15 pts).

4. You are working on a project that requires you to capture data from millions of IoT devices in people's homes. Each IoT device uploads a JSON document with the data elements required for analytics.
    a. Identify potential NoSQL databases that you can to capture data from the IoT devices (5 pt)

b. What are your design and analytics considerations and rationale behind your choice? (10 pts)

Part 2 – NoSQL Labs (50 pts)

1) MongoDB Lab (25 pts)

Setup (Please show evidence of your setup with screenshots)

- Set an account on MongoDB Atlas - https://cloud.mongodb.com
- Load the Sample Netflix Movies Database to your Data Lake
- Set up a connection to this database instance from MongoDB compass or any other MongoDB client.

Provide the following answers:

1. Briefly describe the movies database document model.
2. Filter the documents for **type** "movies" that are **released** before 1970 and **rated** as "PASSED".
3. Build an Aggregation Pipeline that shows all entries of type **movie** that have won at least one award and return the release **year** aggregate counts.

2) Cassandra Lab (25 pts)

Using DataStax Astra Cassandra-as-a-Service (https://astra.datastax.com/)

Setup (Please show evidence of your setup with screenshots)

- Create a Keyspace called **northwind**
- Create the customer tables (the attached SQLite definition will serve as a guide) Review the questions in the queries section below and create one or more tables that partition and cluster data so these queries will execute without using Cassandra "ALLOW FILTERING" that scans all partitions.
- Load the attached data into your table(s) using the insert statements (minor modifications may be needed if your definitions include multiple tables). Please include screenshots of table record counts after loading your data.

Queries

1. Provide the query and the results (screenshots and a copy of your query) that show the customers from **Rio de Janeiro**, **Brazil** ordered by their addresses.

2. Provide a list of customers that are in the **Sales Manager** role without forcing the scan of all partitions across all databases. The result should be ordered by their names.