IST 2510: BIG DATA ANALYTICS

HOMEWORK 6

LECTURE REF: Introduction to NoSQL

Fall, 2025

INSTRUCTIONS

PRODUCTIVITY TOOLS ARE NOT ALLOWED. STATE ALL YOUR SOURCES WHEN ALLOWED.

SHOW SCREENSHOTS OF YOUR QUERIES AND RESULT WHEN APPLICABLE. IF IN DOUBT, CREATE A SCREENSHOT.

1. During the class, there was a discussion of modelling a dataset as a Matrix and a record as Vector. Describe your current BA64 SQL database mathematically.

2. MongoDB relaxes the rectangular shape that an SQL table guarantees and allows a database collection to have an irregular shape.
    a. Will your mathematical model work with an irregular data shape like what is possible in MongoDB?

    b. What are the limitations of your current model in Question 1 under the relaxed constraints of MongoDB?

    c. Can you find a way to overcome this? [EXTERNAL SOURCE IS ALLOWED HERE, STATE YOUR SOURCE]

3. Connect to your MongoDB instance using the MongoDB shell using the following connection string: mongodb://127.0.0.1:27017.

    Run the following command to enumerate the databases that exist on your local machine:

    show dbs

    Use the following command to switch to or create a new database named “BA64”:

    use BA64

    Use the following command to create a new collection in your BA64 database:

    db.createCollection(“Students”)

    (a) Insert 10 documents of irregular shape into your “Students” collection using InsertOne function, which is designed to operate on a collection by appending a

document to the collection. “irregular” in this context means, varying number of attributes per record and/or varying attribute definition.

(b) Insert at least 500 records into your Students collection using InsertMany function which accepts an ARRAY([]) of BSON documents. You are allowed to generate up to 10 batches containing at least 50 records. Each batch should define an attribute that is not present in other batches.

(c) Using the lecture, write a query to find the subset of your data with the unique attributes DEFINED over the entire set of students. Do this for all unique attributes introduced during the batching process.

(d) Using the lecture, write a query to find subset of your data with the unique attributes NOT DEFINED over the entire set of students, for all unique attributes introduced during the batching process.

(e) Create a subset of your data using the find function over the data space of Students such that it constrains the resulting data space based on TWO attributes of your choice. For example:

```
db.Students.find({LastName: {$exists: true}, StudentId: {$gt: 1}})
```

(f) Explain why (e) is important in your own words. Write in Vietnamese.

4. From the discussion about the differences between RDBMS and NoSQL, what is your understanding (in your own words) of the main differences between these two database paradigms?