TU Dublin TU856/TU857/TU858
Advanced Databases
MongoDB CA Task
(using data from an Apache Cassandra database)

This task will be marked out of 100%.
The lab will contribute 15% to your CA (when weighted to 60%))

IMPORTANT

- You will need to complete Labs from week 9 and week 11 to be able to complete this lab.

Contents

TASK OVERVIEW

You are going to:

- Setup a MongoDB cluster with replication.
- Create a database.
- Port data from the first table created in your Apache Cassandra CA.
    - You will be writing a Python script to extract the data from Cassandra and create a file with a set of insert statements with JSON versions of this data that you can run in MongoDB.
- You will then execute queries against this database using indexes to improve performance.
- You will then create an aggregation pipeline in this database using indexes to improve performance.
- You will capture relevant information about the performance of your MongoDB replication and the impact that the indexes have on your query performance.

TASK DETAILS

Task #	Description	Covered in Lab
1.	Setup: a. Create a MongoDB cluster • This should be named with your student number. b. Create a replica set • This should include your student number in its name. c. Create a database. • This should include your student number in its name.	WK 11
2.	Port data from Cassandra to a MongoDB collection: a. Adapt the Python script provided for the lab in week 11 to extract the contents of a table in Cassandra and create a set of insert statement using JSON for this data that you can use in MongoDB.	WK 9 and 11
3.	Work with the collection in MongoDB: a. Create a query statement (anything other than finding all documents) which involves a text field. b. Create a secondary index on the text field. Demonstrate that the secondary index has succeeded.	WK 11
4.	Work with aggregation in MongoDB: a. Write a simple aggregation pipeline. b. Create indexes to help improve the aggregation performance. c. Optimize your stage execution.	WK 11
5.	Monitor your query performance a. Capture relevant information about query performance using explain.	WK 11
6.	At some point pause/stop your primary Mongo node. You will see the other nodes elect a new primary node. Capture the relevant information using rs.status()	WK 11

MARKING

Marking Breakdown		
Setup (cluster and replication and database)		10 marks
Cassandra to MongoDB extract and load		15 marks
Working with MongoDB Golf data		20 marks
Basic Queries (and verification)	10 marks	
Adding secondary indexes to support pattern matching in text (and verification)	10 marks	
Working with Aggregation		40 marks
Aggregation Pipeline (and verification)	15 marks	
Adding secondary indexes to support pattern matching in text (and verification)	15 marks	
Optimize your stage execution (with comments)	10 marks	
Provide relevant output to demonstrate the performance of your queries for relevant aspects of the above.		5 marks
Provide relevant output to demonstrate the existence and resilience of your MongoDB replication		10 marks
Total Marks		**100 marks**

SUBMISSION

What needs to be submitted?

You need to **SUBMIT A SINGLE ARCHIVE (.ZIP, .RAR, .7Z)** named with your student number**,** e.g. D123456.zip, containing the following:

1. A *single SQL file* named with your student number, e.g., D123456.sql
   - Containing your create statements and queries
   - Commented appropriately explaining what you are attempting to achieve.
   - NOTE: It should be VERY clear in your SQL where you are addressing each task.
2. A *Python script* which extracts data from Cassandra and creates a file of insert statements to insert JSON data to MongoDB, named with your student number, e.g. D123456.py
   - Commented appropriately.
3. The *file of insert statements*, named with your student number, e.g. D123456.json
4. Either
   - A *companion document named* with your student number (either docx or pdf) e.g. D123456.docx, D123456.pdf
     i. A template outlining the type of content to include is available in the file called ADvDB-MongDBCA-Template.docx attached to the assignment in Brightspace.
     ii. Note: You are free to adapt this template as you see fit.

   OR
   - A link to a *recording of the task/set of recordings of the task* being completed with relevant performance output being created with audio description.
     - Refer to the template for the document to identify what should be addressed.

NOTE: You may be asked to demonstrate your work.

How do I submit?

Submit this via the Assignment section in **Brightspace** into the assignment called **MongoDB CA**.

What is the deadline?

The deadline is **Friday December 16th 2022 @ 23:59.**
