TU Dublin TU856/TU857/TU858
Advanced Databases
Apache Cassandra CA Task
(using data from a Data Warehouse)

This task will be marked out of 100%.
The lab will contribute 25% to your CA (when weighted to 60%))

IMPORTANT

- You will need to complete Labs from week 8, week 9 and week 10 to be able to complete this lab.

Contents

TASK OVERVIEW

You are going to:

- Setup a Cassandra cluster
- Create a keyspace in that cluster using replication
- Port data from a query on the fact table created in the PostgreSQL data warehouse for the golf exercise for the lab during week 8.
    - You will be writing a Python script to extract the data to a JSON file, create a table in your Cassandra keyspace and then import the data into the table
- You will then execute queries against this database using indexes to improve performance.
- You will also create a second table incorporating a column of type collection and implement indexes on this table.
- You will capture relevant information about the performance of your Cassandra cluster in general and the impact that the indexes have on your query performance.

TASK DETAILS

Task #	Description	Covered in Lab
1.	Setup: a. Create a Cassandra cluster • This should be named with your student number b. Create a keyspace within this cluster • Choose an appropriate partitioning strategy and replication factor.	WK 9
2.	Port data from PostgreSQL to Cassandra: a. Working with a PostgreSQL database, write a query using the fact table in the data warehouse created in the lab class in week 8. This needs to generate some text data in the results. b. Adapt the Python script provided for the lab in week 9 to extract the results of the query to a JSON file, create an appropriate table in Cassandra and populate the table with the contents of the JSON file.	WK 9 (requires WK 8)
3.	Work with tables in Cassandra: a. Write a CQL statement to query the table created in step 2. b. Write a CQL statement to query the resulting table on a non-primary key column – ensure that this can succeed without adding an index. c. Create a secondary index on a non-primary key column. Demonstrate that the secondary index has succeeded. d. Create an SASI index on your table to facilitate pattern matching in a text column. Demonstrate that the SASI index has succeeded.	WK 9 and WK 10
4.	Working with collection data type: a. Create a new table that includes a column of type collection and populate with some data (at your discretion). b. Write a CQL statement to query the resulting table on the collection column – ensure that this succeeds without adding an index. c. Create an appropriate index on your collection column. Demonstrate that the index has succeeded.	Wk 10 and WK 9
5.	Monitor your cluster and query performance a. Capture relevant information about cluster and table performance using nodetool. b. Capture relevant information about query performance using tracing.	WK 9 and WK 10

MARKING

Marking Breakdown		
Setup (cluster and keyspace)		10 marks
PostgreSQL to Cassandra extract and load		15 marks
Working with Cassandra Golf data		40 marks
Basic Query	5 marks	
Adding a secondary index (and verification)	15 marks	
Adding an SASI index to support pattern matching in text (and verification	20 marks	
Working with Second Cassandra table with collection datatype		25 marks
CQL to create and populate data	10 marks	
CQL query using the collection column	5 marks	
Adding an index for your collection column (and verification)	10 marks	
Provide relevant output to demonstrate the existence and performance of your cluster, keyspace and tables for relevant aspects of the above.		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 CQL file* named with your student number, e.g., D123456.cql
   - Containing your create statements and queries
   - Commented appropriately explaining what you are attempting to achieve.
   - NOTE: It should be VERY clear in your CQL where you are addressing each task.
2. A *Python script* which extracts data from PostgreSQL and loads it into Cassandra named with your student number, e.g. D123456.py
   - Commented appropriately.
3. The *JSON file* of data extracted from PostgreSQL, 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-CassandraCA-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 **Cassandra CA**.

What is the deadline?

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