Assignment One – NoSQL Data Storage

Deadline: **Fri 21st Feb 2020 – 3:30pm (Week 6)**

Submission: TurnItIn Assignment on F21BD Vision Course – one report submission per group

Overview

This is a group coursework assignment (self-selected groups of 4 members; sign up the group to Vision) which takes the form of storing and querying data using NoSQL stores. Please form a group before signing up (i.e. rather than adding your name to a random online group before finding people to work with), also ensure that you sign up to a group in the correct location (Edinburgh / Dubai) and level (UG or PG).

The underlying assumption in the coursework is that you are developing on a sample of a much larger data set. Therefore, you should avoid manual steps of data conversion/tidying/processing.

The submission for this coursework takes the form of a report that explains and justifies the steps that you have taken during the different stages of this project. Your fully commented queries and scripts should be included in the report with sample outputs of the query results, and these must be readable.

There is no need to write any code for this coursework beyond queries to access the data store[1]. However, a scripted approach is an acceptable solution providing it is sufficiently documented.

Groupwork:

Share the work among the group members, ensuring that everyone has an equal share. All members should contribute to all areas of the coursework as experience gained in these topics will help in the examination.

Ensure that all members of the group have access to the files developed by the group. You can use the collaboration tools on Vision to support this, or other cloud-based system. However, you must ensure that it is secure against other groups gaining access.

Please provide a summary in your report stating the contributions of each group member. If necessary, marks will be adjusted if some students have not participated enough. You may also email the lecturer directly if you do not feel happy putting it in the report.

Collaboration and Plagiarism

Coursework reports and code must be the group's own work. If some text or code in the coursework has been taken from other sources, these sources must be properly referenced. Failure to reference work that has been obtained from other sources or to copy the words and/or code of another student is plagiarism[2] and if detected, this will be reported to the School's Discipline Committee. If a group is found guilty of plagiarism, the penalty could involve voiding the course.

Students must never give hard or soft copies of their coursework reports or code to students in another group. Students must always refuse any request from another student not in their group for a copy of their report and/or code. It is expected that all group members will have read and write access to the report and code for their group.

Sharing a coursework report and/or code with another group is collusion, and if detected, this will

[1] Note that some NoSQL stores exploit javascript features, e.g. MongoDB

[2] Heriot-Watt guidelines on plagiarism https://www.hw.ac.uk/students/studies/examinations/plagiarism.htm

be reported to the School's Discipline Committee. If found guilty of collusion, the penalty could involve voiding the course.

Dataset

The Internet Movie Database (IMDB - www.imdb.com) is a website with information about movies including details on the male/female actors, writers, directors, movie ratings from the public, release dates, and other details.

You have been provided with a subset of the IMDB, exported from a relational database, as a zipped collection of text files which can be downloaded from here:
http://www.macs.hw.ac.uk/~pb56/f21bd.html

The text files supplied include details for: Actors, Directors, Movies, Ratings, Film running times, Writers – as well as the files which join these entities (e.g. which actors are in which movies). Take a look at the files to familiarise yourself with their contents.

Scenario

You work for a software development company and have been asked to test the suitability of the graph database **NEO4J** in handling the IMDB dataset. In particular the client would like to know if Neo4J is able to handle types of queries from a list provided.

You should load the data into the system then answer the questions listed below. Your report should include a summary of the data provided, all the commands/code that you used to convert/transform/load the data, and the Cypher commands to answer each question. Ensure you include the question number, query in English, the commands used, and the results with each of the questions listed.

Note:
The neo4J labs (A and B) will provide you with guidance over how to install, load and use CYPHER to query a Neo4J database. You may also need to refer to material on line, such as the Neo4J website.

Required Tasks

Ensure you report is well structured with the tasks in order and clearly labelled, include the group name, course level (i.e. F20BD or F21BD), and team members' names.

TASK 1 **[5 marks]**

- Familiarise yourself with the IMDB data subset and provide an overview of the entities and how they are linked (e.g. Entity Relationship Diagram).

TASK 2: **[10 marks]**

- Load the data provided into **Neo4J** giving the details in your report on any data cleaning/transformations/checks/loading commands used. These should be detailed enough for someone else to implement given they know how to use Neo4J.

[see over page for additional tasks]

TASK 3:

- Answer the following questions using Cypher in Neo4J – your answer must include the Cypher code and the result, as well as the English question and question number. Each question can be answered using a SINGLE Cypher query.

Question		Marks
1	How many female actors are listed in the dataset supplied?	2
2	How many male actors are listed in the dataset supplied?	2
3	Write a CYPHER query that shows the number of female actors and the number of male actors as a single query	2
4	List the movie titles and number of directors involved for movies with more than 6 directors	2
5	Number of movies with a running time of less than 10 minutes	2
6	The movie titles which star both 'Ewan McGregor' and 'Robert Carlyle' (i.e. both actors were in the same film)	2
7	Number of movies directed by 'Spielberg'	2
8	List the male/female actors that have worked together on more 10 films, include their names and number of films they've co-starred in	2
9	List the number of movies released per decade as listed below (1960-69, 1970-79, 1980-89,1990-99,2000-2010)	2
10	How many movies have more female actors than male actors?	2
11	Based ratings with 10,000 or more votes, what are the top 3 movie genres using the average rank per movie genre as the metric? *(Note: where a higher value for **rank** is considered a better movie)*	2
12	Show the shortest path between actors 'Ewan McGregor' and 'Mark Hamill' from the IMDB data subset. Include nodes and edges – answer can be shown as an image or text description in form (a)-[ ]->(b)-[ ]-> (c)…	2
13	List all actors (male/female) that have starred in 10 or more different film genres (show names, and number of genres)	2
14	How many movies have an actor/actress that also directed the movie?	2
15	How many movies have been written and directed by an actor/actress that they didn't star in? (i.e. the person who wrote and directed the movie is a film star but didn't appear in the movie)	2

For 2 marks you need to provide Cypher code and correct answer; correct answer without any cypher code will score 0; Wrong answer but Cypher code demonstrating correct idea may score 1 mark.

- **TASK 4: Additional Task for Level 11 (i.e. MSc) students:**

  a) Provide a summary of using Neo4J in this scenario, highlighting any issues encountered and how these may be overcome. **[5 marks]**

b) Provide 2 further real life Neo4J case studies. You are expected to do background reading and reference the work (i.e. literature review), include tables and figures as necessary. Compare each case study example (i.e. advantages and disadvantages) to it being managed using a relational database management system (e.g. mySQL). **[10 marks]**

Report

Please clearly state your group ID, group participants (name and student ID), degree programme, and for MEng students your year of study on the title page of your report. Also remember to include a summary stating the contributions of each group member.

Reports are to be written in clear concise English and include supporting diagrams (which must be human readable, i.e. if they are screenshots of the output of your queries then I must be able to read the text). All code and queries should be commented. Any code should be included as machine readable text (i.e. not screenshots) in an appendix and referenced and explained in the text of the report.

All material drawn from other works must be appropriately referenced in accordance with the University's policies[3].

Reports should be submitted electronically through Vision F21BD_2019-2020. Please use the appropriate assessment for whether you are a student in Dubai or Edinburgh and taking the course as a 4th year (F20BD) or 5th year/MSc (F21BD).

One report should be submitted per group. Ensure all group members are listed on the title page, and also are listed in the appropriate group on Vision.

The standard policy applies for late submissions, see your programme handbook.

Marking rubrics are available from the TurnItIn Assignment and also as a PDF on Vision.

==========

[3] http://www.hw.ac.uk/students/studies/examinations/plagiarism.htm