Assignment One – NoSQL Data Storage

Deadline: **Thu 14th Feb 2019 - 11am (GMT) (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 3 members; sign up the group to Vision) which takes the form of storing and querying data using NoSQL stores. The underlying assumption in the coursework is that you are developing a data transformation process using a sample of a much larger data set. Therefore, you should avoid manual steps of data conversion, i.e. you are seeking to develop a data pipeline that can be automated.

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 appendices of your report. Sample outputs of the query results should be included in the report and must be readable.

There is no need to write any code for this coursework beyond SQL and queries to access the NoSQL 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.

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

---

[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

Scenario

You work for a digital music download company, called *Chinook*, that allows people to buy and listen to music tracks. The website currently uses a relational database to store the details of tracks available, music sales, and staff, but the userbase is growing very quickly and some clients have started to complain of slower response times when searching for and purchasing music. The CEO has heard that NoSQL systems provide fast response times and would like you to investigate their use for the company. In particular the CEO has asked you to look at **MongoDB, Neo4J, and Cassandra**. For this task you should **select 2 from this list** and implement your solution for both databases, then compare them in your report with discussion of their suitability for the task.

The Chinook web page allows viewers to preview music tracks, listen to pre-set playlists supported by adverts, and to download digital copies of the tracks. Clients can search on the website for tracks by song title, artist, album, composer and genre. Staff use the intranet version of the webpage to access invoice data, and generate reports on sales per employee. They also generate the playlists, and would like to open this up so clients can generate their own playlists in the future.

The actual files (e.g. MP3, MP4) are stored on a separate cloud based file system (i.e. outside of the DBMS) using the TrackID as the filename. The downloading/streaming of music media is not the cause of any bottleneck in this scenario.

Dataset

For this assignment, we will use the Chinook database[3]. The dataset consists of 347 albums from 275 artists, and 3503 music tracks. There are 59 customers, 8 employees, with 2240 invoices of track sales. In addition, there are 18 playlists consisting of 8715 tracks. Assume that this is a sample of a much larger collection in the full database containing millions of database records.

The dataset is available to download from the Vision module page, as a SQL file to be loaded into mySQL. You may load this into the MACS mySQL server, or your own PC using the VM.

> A VM image for Virtual Box with the necessary software installed is available here: http://www.macs.hw.ac.uk/~pb56/nosql.html.

The schema of the database is shown in Figure 1.

To use the database from a departmental Linux machine enter the command:

```
mysql –u <username> -h mysql-server-1 –p
```

where <username> is replaced with your username and you enter your mysql password when prompted. To have your mysql account reset please contact help@macs.hw.ac.uk.

Tips

Edinburgh students cannot use the `INTO OUTFILE` clause on the departmental MySQL server. Instead they should save their query in a text file and use the following command line argument:

```
mysql -u ab12 -p -h mysql-server-1 db < queryFile.sql > resultsFile.txt
```

You will need to replace the ab12 with your username, db with database name, queryFile.sql is the file containing the query, and resultsFile.txt is the file that the query output will be written to.

Hints for the coursework: You may want to use the `CONCAT` function, also check out the use of header dot notation (e.g. Address.PostCode) when importing CSV/TSV files into MongoDB.

[3] https://raw.githubusercontent.com/lerocha/chinook-database/master/ChinookDatabase/DataSources/Chinook_MySql.sql

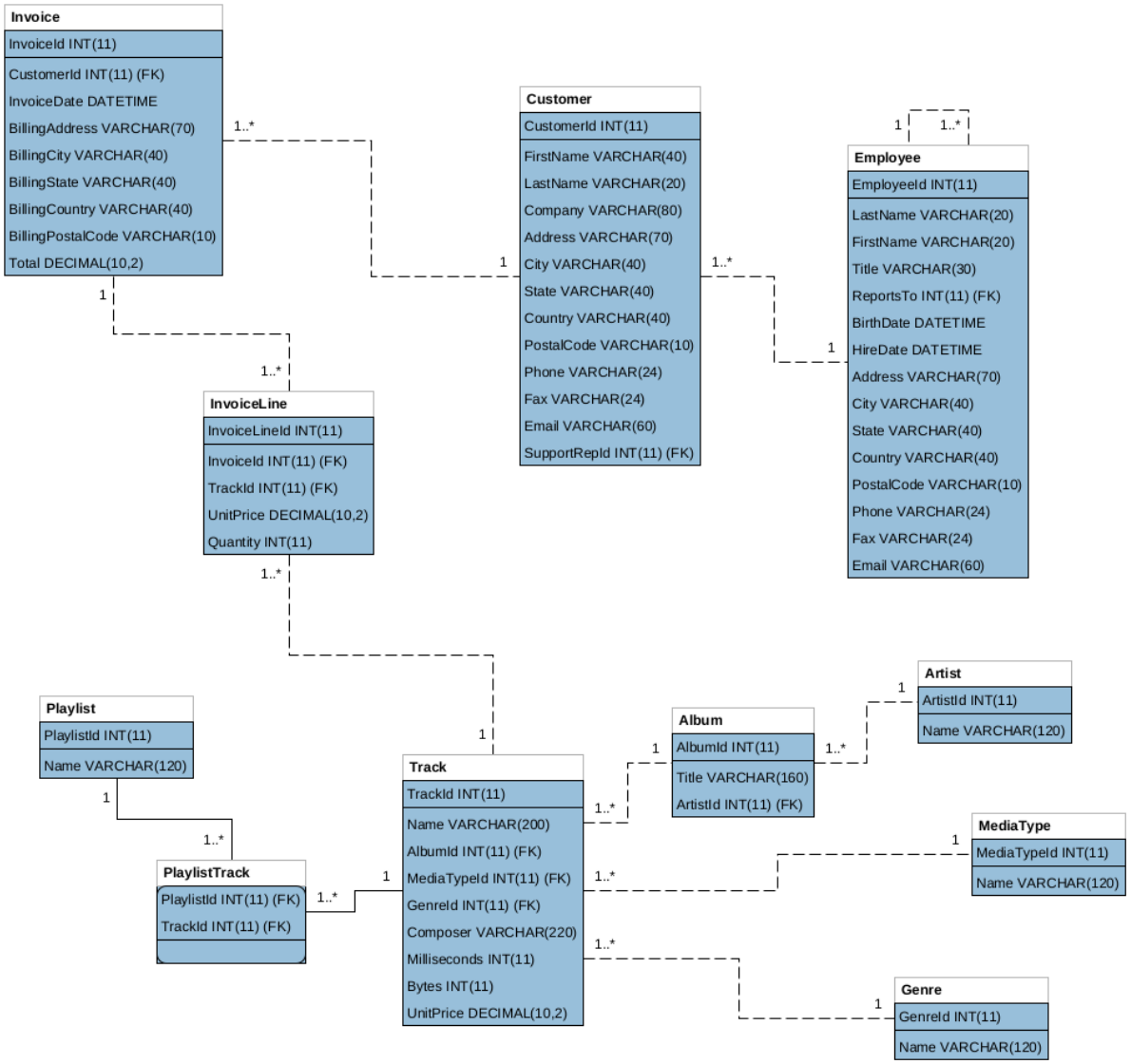

Figure 1: Database schema

Required Tasks

- Decide upon **TWO** NoSQL system to use from the CEO's list (MongoDB, Neo4J, Cassandra). Give a brief overview of each of the data management systems (storage paradigm, consistency/replication model) and explain how it meets the needs of the music download website.
- Define a data model for each of the NoSQL systems you have chosen. This should be represented in JSON Schema[4] and include details of data types and relationships. Provide a discussion of the data model with respect to the relational model provided and linked back to the requirements of the website.

(tasks continued on next page)

[4] http://json-schema.org/

- *Level 11: Provide a comparison of the 4 different types of NoSQL system and their suitability for the given task of a music download website.*
- Extract the data from the relational database, transform to your new data models, and load the data into the 2 NoSQL systems of choice. Provide details of your extract-transform-load pipelines, explain the steps, and provide evidence that the data has been loaded.
- *Level 11: Discuss the limitations of your approach with respect to its scalability for each of the systems implemented. Compare the performance/limitations of 2 systems you implemented for this task.*
- Analyse the dataset using the native query language for each of your chosen data management systems. Aim to go beyond a simple count of the number of records. Five interesting queries[5] per NoSQL system, that are different from each other should be described and their results (or part thereof) displayed. In other words, you need to think of 5 queries and implement them twice (once for each NoSQL solution).

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 extensively 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[6].

Reports should be submitted electronically through Vision F21BD_2018-2019. 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.

[5] An interesting query is one that goes beyond just a simple retrieval of the properties of a single entity.
[6] http://www.hw.ac.uk/students/studies/examinations/plagiarism.htm