Assignment One – NoSQL Data Storage

Deadline: 15:30 (GMT) Thursday 15th February 2018

Submission: TurnItIn Assignment on F21BD Vision Course

Overview

This is individual coursework which takes the form of storing and querying data using a NoSQL store. 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 a NoSQL store[1]. However, a scripted approach is an acceptable solution providing it is sufficiently documented.

Collaboration and Plagiarism

Coursework reports must be written in your own words and all code and queries must also be your own. 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 and/or queries of another student is plagiarism[2] and if detected, this will be reported to the School's Discipline Committee. If a student is found guilty of plagiarism, the penalty could involve voiding the course.

Students must never give hard or soft copies of their coursework reports, queries, or code to another student. Students must always refuse any request from another student for a copy of their report, queries, or code.

Sharing a coursework report, queries, or code with another student 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] https://www.hw.ac.uk/students/studies/examinations/plagiarism.htm

Scenario

You work for a flight comparison website that currently uses a relational database. However, clients have been complaining of slow response times. The CEO has heard that NoSQL systems provide fast response times and would like you to investigate their use for the company.

The airport web page allows viewers to see details of an airport together with the city that it serves including details such as the timezone and day light savings, a list of the airlines, and a list of destinations. Through other web pages it is possible to view details of airlines and the routes.

Dataset

For this assignment, we will use a modified version of the Open Flights dataset[3]. The dataset consists of 6,162 airlines, 7,184 airports, connected by 67,663 routes. There are also 6,649 cities located in 260 countries. It is assumed that this is a sample of a much larger collection in the full database containing millions of records.

The modified dataset is available from the MACS MySQL database server in the movielens database. The schema of the database is shown in Figure 1. To connect to the database from a departmental Linux machine enter the command:

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

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.

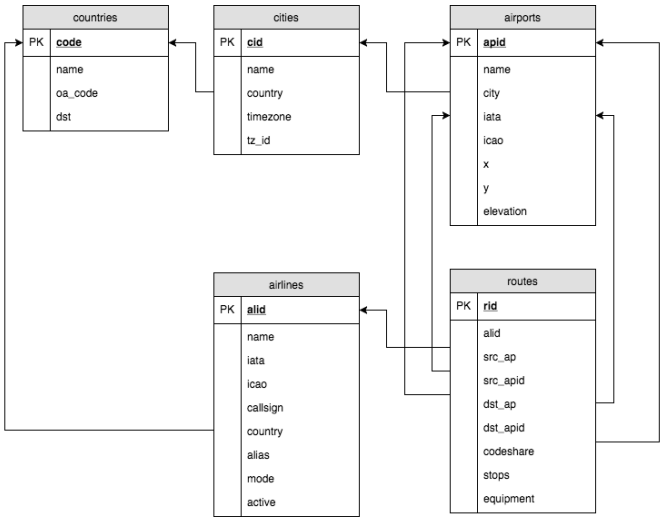

Figure 1: Open Flights database schema

[3] Open flights website: https://openflights.org/
GitHub repository with raw data: https://github.com/jpatokal/openflights

Required Tasks

- Decide upon a NoSQL system to use. Give a brief overview of the data management system (storage paradigm, consistency/replication model) and explain how it meets the needs of the flight comparison website.
- Define a data model for the data in the NoSQL system of choice. 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 flight comparison website.
- *Level 11: Provide a comparison of NoSQL systems for this task. Describe the different types of NoSQL system and their suitability for the given task of the flight comparison website.*
- Extract the data from the relational database, transform to your new data model, and load the data into your NoSQL system of choice. Provide details of your extract-transform-load pipeline, 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.*
- Analyse the dataset using the native query language of your chosen data management system. Aim to go beyond a simple count of the number of records. Five interesting queries[5] that are different from each other should be described and their results (or part thereof) displayed.

Report

Please clearly state your name, user id, degree programme, and for MEng students your year of study on the title page of your report.

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_2017-2018. 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).

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.

---

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

[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