CSC4023Z 2020 - Assignment 1

This assignment must either be done in groups of 4 (exactly 4) students, or alone. This assignment is worth 30% of your final mark for the course (assignment 2 is also worth 30%, and the final exam makes up the remaining 40% of your result). Please use the Vula forum for clarification questions, not email/chat room.

(1) **Find or create a suitable data set.** This could be data you already have, are using in your honours project/modules, or are interested in. Otherwise, find one online (google JSON data sets eg), but no 2 assignments can use the same data (check the wiki). If you use data that is not publicly available, please obtain clearance to use your data, and attach this to your assignment submission. *If you go to kaggle.com, please check if the competition rules for that data set prohibit use outside of the competition – if this is not specified, the data may be used for academic purposes (but not commercial use).* Data needn't be large or realistic, as the aim is to show you can design, load and access the database, not build a useful application. You may find you obtain a data set online but it is too large and too simple; if so, extract a small subset to use, and then create additional data for those few instances, so you have a more suitable database with which to show your understanding of NoSQL. Include the data in the zip file submitted for this assignment; also include a file explaining attribute meanings if necessary (eg if this is data for a domain that isn't well-known). [4]

(2) **Design a MongoDB database** for your data, making sure you have at least 2 collections, and specifying as clearly and completely as possible all aspects of the design and why this design was chosen. While you will only be testing with a small data set, assume the system will use big data when deployed, and design accordingly. The less well-known the domain from which your data is drawn, the more detail you must give for the use cases on which your design is based. [6]

(3) **Create and load this MongoDB database**. Show the statements you used to do this, and to test that it is working correctly (along with output of the test/s). [4]

(4) Discuss the relative benefits and disadvantages of MongoDB compared with **each of the other types** of database studied (graph, key-value, column-family, relational) in this context/domain. Also say for each type, what role this could have if polyglot persistence is used, and why (else, if not usable at all, say why). Note you can include data and use cases not mentioned before in this answer. If in a group do 1 db type each; if alone do any 2 types (mark will be doubled). [4x6=24]

(5) Test a wide **variety of MongoDB operations** and show the output produced. Each group member should write and test 4 different operations/statements. If working alone, write and test 8, and that mark will be doubled. [4x4=16]

(6) Write & test **a simple program**, in any programming language MongoDB supports, that runs 1 statement per student (or just 1 if alone) from (5) above. [6]

(7) Groups give the **exact contribution of each** member (-2 if omit, -1 if unclear).