COMP6235 Coursework 1

by Zhennan Huang 28507649 zh4n15@soton.ac.uk

For use in a Data Science application, we can use two different kinds of databases to store data, Relational database and NoSQL database. I will store the data I downloaded in question 1 in MongoDB rather a relational database. Those data can be downloaded in JSON format or XML format. If we want to store those data in a SQL database, we need to design and create many tables to store different kinds of data because of the different data structures. However, if we use NoSQL databases, we can store JSON format directly in document-based databases such as MongoDB.

There are many differences between relational databases and NoSQL databases. Firstly, relational databases store data in tables which have strict data structures, NoSQL databases are able to accept various types of data in different databases, such as key-value, document-based, graph-based and column-based databases. Secondly, a relational database needs to define tables and constraints before storing data, while NoSQL databases are able to insert data whenever we want without define tables. Next, relational databases provide the join operation to search data in different tables. Since NoSQL databases store non-normalized data, it does not have such function. Finally, relational databases are able to avoid data duplication through joining tables.

The costs and benefits of using different kinds of databases are different in various situations. When the dataset is huge and the data structures are various, a relational database has to use different kinds of tables to store the same kind of information, such as the web page access information. However, when we need to pay more attention to the security and stability of data storage, we should use relational databases, such as the bank account information.

I would like to store the data I downloaded in MongoDB which is a document-based database. There are several reasons. Firstly, those data can be downloaded in JSON format, and MongoDB can store JSON documents directly; thus, I neither need to create new tables nor divide each JSON into several records. Secondly, those data have many kinds of data structures. Even if the same kind of establishments may have a different number of attributes. If we store those data in MongoDB, we do not need to pay attention to differences. Since there are a huge number of data, we do not need to focus on every record. When we search for a particular kind of data, if some dictionaries do not have such attribute, it will return a None object which we can use some methods to ignore. Besides, there are various kinds of functions which are supported by MongoDB or Pymongo. Therefore, we can easily to find those data that we want to analyse.

As I mention above, SQL databases are more stable and secure than NoSQL databases, when we need to store our data in a very secure and stable way, we may need to store data in SQL databases. However, while NoSQL databases sometimes may result in data loss, the vast majority data can be stored safely. Since the amount of data is huge, we just need to guarantee the vast majority data safe.

I am aware of the requirements of good academic practice, and the potential penalties for any breaches.