F21BD Big Data Management Report

Jamie McCulloch jm7@hw.ac.uk
H00189648

Software Engineering (Year 4)

1. MongoDB Overview

I chose MongoDB as my NoSQL system for this coursework. MongoDB is the “worlds most popular NoSQL database” [1]. Unlike in a relational model where everything is stored in tables and link to each other, MongoDB uses JSON-like document structure to store data. This results in it being a lot easier to combine any type of data. MongoDB, and NoSQL systems in general, can save memory as it does not store NULL values in the table, it simply ignores them. In MongoDB a JSON-like schema can be applied to the JSON data, to ensure that it is in a valid format. For this, the schema must declare what is required within each document. Anything that is not on the required list will be ignored by MongoDB if it is not present (i.e. no warning) and will simply be left out when displaying the document. Comparing this to a relational system like MySQL where everything declared in the table must have a value (even if NULL) which results in complications when querying data.
Especially when considering this open flights dataset, where null values exist frequently (for example some airports do not have a country field or a city field) it would be more appropriate to use a NoSQL system. As the scenario specifies that there would be three webpages:

- Details of airlines and what routes they serve
- Details of airports and what cities/countries they are in
- Details of airports, the timezones, routes they conduct and the airline

MongoDB would be useful as these can easily be split into three different collections, and loaded into MongoDB this way.

2. JSON Data Model

I first created queries in MySQL that would extract the data I wanted to load into the MongoDB collection. Then once I had settled on the queries I then used the fields I was returning as the fields in the JSON schema for mongo. Thus, there were three schemas, one for each collection specified in the scenario. A brief overview of each of the schemas is below and the full JSON documents are in the appendix.

AirlineRoutesSchema.json

Fields:

- AirlineName
- Source
- Destination
- Country

All of the fields are strings, and as source and destination can only be three characters long, I have added this constraint within the schema. I also stated that three fields were required – AirlineName, Source, and Destination as not all

REFERENCES

[1] - https://www.mongodb.com/blog/post/report-mongodb-top-database-next-generation-cloud-workloads