Processing Enron Dataset with NoSQL Database

JUNMING ZHANG, University of Southampton, UK

Abstract

The aim of coursework 2 is using MongoDB to retrieve and process data from a dataset of Enron email. In addition, different types of databases will be compared and the more suitable database will be chosen. This document will be divided into 3 main parts. First part will describe the difference between SQL and NoSQL databases, next part will compare different types NoSQL databases and the final part will discuss which database is more suitable for process Enron dataset.

Declaration

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

1 Database

1.1 Relational database

Relational database is also called as SQL databases, it is architected by structured query language(SQL). This database is defined as a structured query language for selecting information from databases [1]. SQL database is the combination of tables which consist of rows and columns, it follows ACID rules which are atomicity, consistency, isolation and durability.

1.2 NoSQL database

NoSQL database is called as no relational databases and Not Only SQL, it uses the key to store data in a tuple and the structure is unfixed which means it reduces the cost of time and volume. There are 2 general types of NoSQL databases which are close to this coursework [2]:

1. Key-value store: it is the set of key and value are used to store all data and these data are accessed by unique keys.
2. Document store: it is the set of transformed documents of key value and these documents are identified by a unique key. In addition, it has a similar structure with JSON.

1.3 Difference between relational and NoSQL databases

The main difference between them is the method used to store data, relational database uses table to store data but NoSQL database uses key and collection to store data.

For the stored structure, SQL has fixed structure which means it is stable, however, it complicated to modify data. NoSQL has a dynamic structure with a high adaptation of different types data.

For the expansion, SQL uses vertical scaling which means a faster computer with high processing ability is required. No relational data is distributed which means a higher expansion ability because more servers can be used to share the workload.

2 COMPARISON OF DIFFERENT TYPE OF NoSQL DATABASES

2.1 Classification of NoSQL databases

The CAP theorem indicates a distributed system cannot satisfy 3 characteristics which are consistency, availability and partition tolerance at the same time. Therefore, NoSQL databases can be classified by 3 possible configurations: [3]

1. Consistency and availability: OrientDB is one NoSQL database with CA.
2. Consistency and partition tolerance: HBase and Redis are CP databases.
3. Availability and partition tolerance: Cassandra and MongoDB are CP databases.

2.2 Analyze NoSQL databases

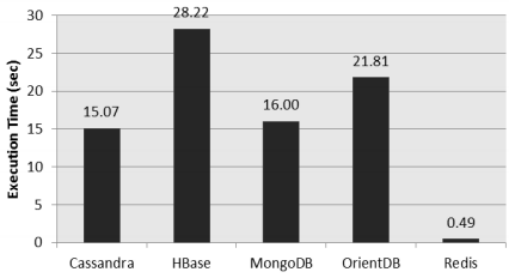

Fig. 1. Execution time compare of different NoSQL databases [4]

Figure 1 is obtained from an open journal which contains the performance test of 5 types NoSQL databases. This report will not research OrientDB because it is a CA type database which not suit for this coursework.

2.2.1 Cassandra. This database consists of codes, clusters, data centres and a partitioner [5]. The benefit of it is high concurrency, flexibility, extensibility and robustness. However, the store method is key-value with simplex value and need an amount of memory to caching all keys for data management.

2.2.2 MongoDB. This is a document-oriented database which suits for XML, JSON and BSON data. In addition, it is an opensource project with low cost. It supports convenient retrieval function MapReduce and aggregation. However, the atomicity of retrieval could not guarantee.

2.2.3 HBase. It is an open source version of BigTable to manage the mess of structured data [6]. The benefit of it is the huge storage space and horizontal sharding extension. However, it is developed by Java which means it is hard to do the configuration.

2.2.4 Redis. It is an open source project by ANSI C language. The value of Redis can be string, hash, list, sets and sorted sets [7]. Therefore, it supports lots of data structures and the execution speed is extremely fast.

3 RESULTS AND DISCUSSION

Four NoSQL databases can be summarized in Table 1.

Databases:	Stored type	CAP	Execution	Characteristic
Cassandra	Document-based	AP	fast	Large cache, simple retrieval
MongoDB	Document-based	AP	fast	retrieval in dynamic and indexes
HBase	Column	CP	lowest	Java-based, huge storage space
Redis	Key-Value	CP	fastest	Rich data structures, power in reading

Table. 1.Conculusion of different NoSQL databases

In this case, Enron email dataset is a JSON file with the structure of 3 layers. Therefore, the databases can be matched by following conditions:

1. Document-based databases suit the processing of JSON file.
2. For part 1 of coursework, availability and partition tolerance of databases are important.
3. For retrieval in the structure of 3 layers, a complex query is required.

According to observe the conclusion of 4 NoSQL databases with previous conditions, MongoDB is a reasonable NoSQL database to process Enron dataset.

REFERENCES

[1] Editors of Encyclopædia Britannica. (2017). SQL computer language. *IEEE Citation Reference [Online]. Available:* https://www.britannica.com/technology/SQL *(accessed 15th December 2017)*

[2] A. Veronika, B. Jorge. 2013. NoSQL databases: MongoDB vs cassandra. *Comm. ACM* New York, NY, USA (2013), DOI: 10.1145/2494444.2494447

[3] B.Laurent, L.Anne, S.Michel. 2011. REDUCE, YOU SAY: What NoSQL can do for Data Aggregation and BI in Large Repositories. *IEEE Citation Reference.* DOI: 10.1109/DEXA.2011.71

[4] A. Veronika, B. Jorge, F.Pedro. 2014. Which NoSQL Databases? A performance Overview. Open Journal of Databases, Vol 1, Issue 2, 2014. ISSN 2199-3459

[5] A. Rodrigo, X. Rene, G. Valeria, H. Fernanda, H. Maristela, W. Maria, L. Sergio. 2015. Evaluating the Cassandra NoSQL Database Approach for Geomic Data Persistency, International Journal of Genomics, Vol 2015, Article ID 502795, 7 pages

[6] C. Dorin, L.Elena, G. Mihai 2010. Hbase-non SQL Database, Performances Evaluation(2010). IEEE Citation Reference. DOI: 10.4156/ijact.vol2. issue5.4

[7] G. Shekhar. 2011. Introduction of Redis – In Memory Key Value Datastore, *IEEE Citation Reference [Online]. Available:* https://dzone.com/articles/introduction-to-redis-in-memory-key-value-datastore *(accessed 15th December 2017)*