Triple Storage for Constructing Pathway Network

Min-Ho, Lee*, Yun-Soo, Choi, Dong-In, Park
*Korea Institute of Science and Technology Information, Korea
E-mail : cokeman@kisti.re.kr, armian@kisti.re.kr, dipark@kisti.re.kr

1. Introduction

Pathway which expresses relations between molecules or proteins in the form of network is very important in system biology. Good pathway database is a basic knowledge resource to efficiently support researches such as understanding life activity mechanism, finding causes of decease occurrence or healing, chemical synthesis for the development of new drugs.

For pathway construction, it is general that person who has biology knowledge identify various terminologies and relations between the terminologies in technical literature, and manually draw the pathway. Because research papers have been exponentially increased in recent, researches on automatic pathway construction with text mining technology has getting grown.

Automatic pathway construction consists of following steps. We first identify biological elements such as proteins, genes, cells and relations between the elements. Extracted elements and relations are converted in the triple form of subject – verb – object, and saved into storage. Lastly, we connect common elements and make network-formed pathway.

In case of simple pathway with certain parts of living things or certain signal transduction, there are not many elements. However, in case of complex pathway with whole part or whole signal transduction, there are a very large number of elements from a few million to a few billion.

Large-capacity and high-speed storage are essential for searching common elements and making pathway in large number of data. People have used RDBMS for triple storage. NoSQL storages which store data with only key-value have recently known as suitable for a large data processing and have been used in various fields. In this paper, we explain which one between legacy RDBMS and NoSQL is more suitable for triple storage in terms of saving and searching data to make pathway.

2. Storage for triples and performance experiment

NoSQL storage focuses on availability and scalability rather than consistency which RDBMS focuses on. Depending on how data are saved, NoSQL is classified into paper-based, column-based key-value based.

We chose MongoDB among various NoSQLs for triple storage to make pathway. MongoDB has features such as document-based storage, fast saving and searching, ease of use, supporting many programming languages, SQL-like commands. Researchers who create and use pathways are familiar to SQL of RDBMS and easy programming language such as Python. Features of MongoDB are well suited to the properties of researches.

The number of data for saving and searching is 390billion triples and the triples were extracted from PubMed that is a medical database. We individually measured loading time for saving data and indexing time for searching data in saving performance experiment. We measured searching time to find exact matching results in subject field with 10,000 randomly chosen identified elements for searching performance experiment. We tested RDBMS in only single server environment, but MongoDB in not only single server environment but also 4 machine-cluster environment. This is to find out scalability of MongoDB. Figure 1 is a configuration diagram of the experimental environment for MongoDB. In single-server architecture, the client which asks queries are connected as 1:1 to mongod which keeps triples. However, in cluster architecture, there is mongos which routes queries between the client and 3 mongods, and configd which keeps the locations of triples are connected to mongos. Each server has 4 core Intel i7 3.5GHz CPU and 32GB main memory.

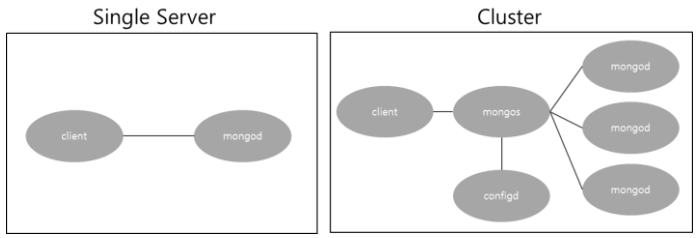

Figure 1. MongoDB experimental environment

3. Experiment result and consideration

Table 1 shows the experimental results.

[Table 1] Experiment result

	RDBMS	MongoDB (Single Server)	MongoDB (Cluster)
Loading (390M triples)	4h 17m	4h 03m	4h 43m
Indexing (390M triples)	1h 31m	1h 29m	28m
Searching (10,000 queries)	460 s	110.9 s	100.9 s

In comparison result of RDBMS and MongoDB in single server environment, MonoDB was slightly faster than RDBMS in both loading time and indexing time. In searching time, MongoDB was much faster than RDBMS. In MongoDB performance comparison of single-server environment and cluster environment, single-server environment was faster than cluster environment for loading, but cluster envirnment was much faster than single-server environment for indexing. It seems that the reason is because data for loading are distributed to mongods via mongos in cluter environment. Faster indexing time in cluster environment can be also described by the same reason. It is because indexing is just performed on each server and does not need data transfer.

MongoDB searching time is much faster than RDBMS. There is no difference between single-server environment and cluster environment in searching time. The reason is because searcing in cluster needs more time for configd to find out which mongod has the wanted data, though the numer of data that shoud be found is small.

4. Conclusion and future works

We compared NoSQL with RDBMS for suitable triple storage that is need in the course of automatic pathway creation from technical listerature. We first found out characteristics of researchers who deal with pathway, and MongoDB is suitable for the researchers. Trough the experiment to save 390 million triples and search results with 10 thousand queries, we knew that MongoDB has better overall performance than RDBMS. We found out that cluster envirnmont is superior to single-server environment in terms of indexing speed. This experiment result could be help to pathway researchers who want to configure their suitable research environment.

We need to find out MongoDB setting to be suitable to cluster environment for improving performance. In addition to that, for researchers who work in not pathway but other fields or can be good at compter programming, NoSQL such as Hbase or Cassandra may be more appropriate to them than MongoDB. Therefore, it could be a research topic that performance comparison among NoSQLs for finding out suitable NoSQL to each field.

5. References

[1] C. Kristina, “MongoDB: the definitive guide”, O'Reilly Media, Inc., 2013.
[2] O.Kanae, J.D. Kim, T. Ohta, D Okanohara, T. Matsuzaki, Y. Tateisi, and J. Tsujii, “New challenges for text mining: mapping between text and manually curated pathways”, BMC bioinformatics, Vol. 9, no. 3, 2008.
[3] S., Michael, “SQL databases v. NoSQL databases”, Communications of the ACM, Vol. 53, no. 4, pp. 10-11, 2010.
[4] S. Michael, and S. Decker, “TRIPLE—A query, inference, and transformation language for the semantic web”, In The Semantic Web—ISWC 2002, pp. 364-378. Springer Berlin Heidelberg, 2002