ATLANTIS PRESS

Advances in Computer Science Research, volume 92

International Conference on Mathematics, Big Data Analysis and Simulation and Modeling (MBDASM 2019)

Performance Analysis of PostgreSQL and MongoDB Databases for Unstructured Data

Yinyi Cheng, Kefa Zhou* and Jinlin Wang
State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography,
Chinese Academy of Sciences, Urumqi 830011, China
Xinjiang Research Center for Mineral Resources
Xinjiang Laboratory of Mineral Resources and Digital Geology,
Chinese Academy of Sciences, Urumqi 830011, China
University of Chinese Academy of Science, Beijing 100049, China
*Corresponding author

***Abstract*—The storage of unstructured data plays an important role in the implementation of big data environment, thus, choosing an efficient database can provide an excellent solution for data mining. In this paper, two database technologies, MongoDB and PostgreSQL, are used as performance tests for storing unstructured data. Remote sensing data in GeoTIFF format is the most representative unstructured data. Large amounts of data are stored in MongoDB and PostgreSQL databases by designing metadata table for GeoTIFF data to test the performance of both. The results show that MongoDB storage is six times faster than PostgreSQL, however PostgreSQL compresses data up to 95%. Therefore, MongoDB is suitable for rapid storage of remote sensing data, while PostgreSQL is more suitable for operations with small data volumes. In a word, this research work has completed the database performance test of unstructured remote sensing data. (Abstract)**

Keywords—unstructured data; MongoDB; PostgreSQL;storage; GeoTIFF

I. INTRODUCTION

In recent years, big data technology has drawn attention from many fields, and even been closely followed by governments around the world. Big data is regarded as the "oil" of modern society, and the value of information it contains has become the focus of scientific research. However, as the global data volume has entered the ZB level, the unique characteristics of big data with large volume, fast speed, multiple modes, difficulty in identification and low value density bring higher requirements for the processing and analysis of big data[1]. Before the era of big data, the data is basically structured data with two-dimensional table schema in standard format, which is relatively easy to manage and operate. However, the data types in the big data environment are very complex, and it is impossible to pre-judge the data storage mode, which is also the reason for the transformation from the previous model-driven research mode to the current data-driven research mode[2]. In the traditional database management mode, unstructured or semi-structured data management is very difficult. As there are many ways to acquire scientific data, multi-source heterogeneity of scientific data appears. Different data organization forms different data structures. The time extension and complexity of these unstructured data are all problems that need to be solved. Remote sensing data is typical unstructured data with large amount of data[3]. This study took remote sensing data as the target data to evaluate the performance differences of unstructured data in PostgreSQL and MongoDB database technology.

II. METHODS AND DATA

A. PostgreSQL

PostgreSQL is a powerful, open source client/server relational database management system. PostgreSQL supports the standards and functions required by the SQL standard, in which consistency does not contradict traditional features[4], PostgreSQL support NoSQL data types (JSON/XML/hstore.) PostgreSQL can deal with unstructured data scenarios very well. First, PostgreSQL supports JSONB data types. Secondly, in terms of customizing data objects, PostgreSQL opens two kinds of interfaces: type extension and index extension. PostgreSQL provides spatial extension and manages unstructured remote sensing data through the same executable files as the raster types compiled in the GDAL dependency library.

B. MongoDB

MongoDB is a high performance, open source NoSQL database oriented to document storage, its characteristic is high performance, easy to deploy and easy to use, very suitable for storing large data, it can support many similar to that of the relational database operations, and its characteristics and relational database SQL statement execution of grammar grammatical syntax is very close[5]. GridFS is a distributed file system built on MongoDB, which realizes file system by storing file data and file metadata in MongoDB, and deals with failover and data integration by Replication. GridFS can also achieve automatic sharding of data through sharding, big data storage and load balancing, and lightweight file system interface and search and analysis through database management and query of documents in the collection[6].

C. *Data*

There are many formats of remote sensing data, such as BMP, HDF, ASC, TIFF, GeoTIFF and so on. Since the data utilization rate of GeoTIFF is the highest, data in GeoTIFF format are selected for experiment in this research[7]. The GeoTIFF file structure inherits the TIFF6.0 standard, so its structure strictly conforms to TIFF requirements. In a TIFF file, all the labels must be in ascending order, and the image data in the file should be processed through the label information. It starts with an 8-byte image header file, the most important member of which is a pointer to a data structure called the image file directory (IFD). All GeoTIFF specific information is encoded in some TIFF reserve tags that do not have its own IFD, binary structure, and other information that is not visible to TIFF.

III. EXPERIMENT

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. *Data Model*

Remote sensing data contains huge information, so this study created physical data tables to be used in PostgreSQL and MongoDB. Metadata table is a standard for remote sensing data input, display, query, editing and other operations. Because the types of image sensors are different, the metadata tables are different. Sensor table stores sensor information including sensor type, task start and end time and other data. Data set tables are used to store data set tile pyramid information.

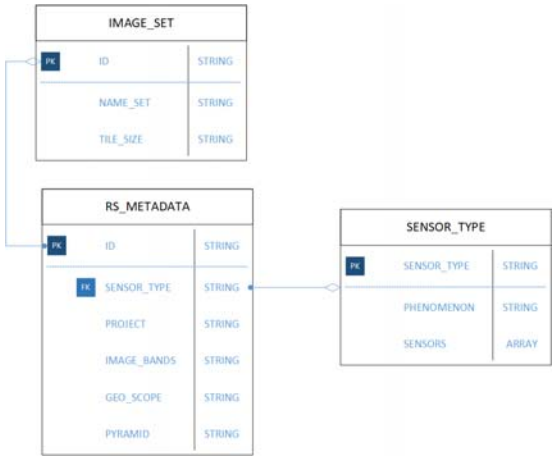

FIGURE I. DATA MODEL OF REMOTE SENSING DATA

B. *Performance Comparison*

To test the performance differences between PostgreSQL and MongoDB for unstructured data, this study evaluated the time consumption of unstructured data and the increment of database data. The computer hardware environment of this research is shown in table 1.In the experiment, about 15000MB of remote sensing data was written to the database in 12 times, and the time consumed by each operation and the space occupied by the data in the database were recorded respectively.

TABLE I. CHARACTERISTICS OF WORKING MACHINES USED

CPU	RAM	HDD
INTER Core i7-4790 3.4GHz	16GB,DDR4 3200MHZ	1TB 7200rpm

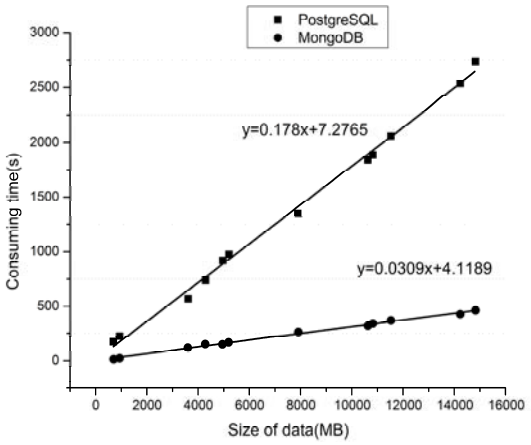

FIGURE II. ERELATIONSHIP BETWEEN THE TIME CONSUMPTION OF POSTGRESQL AND MONGODB

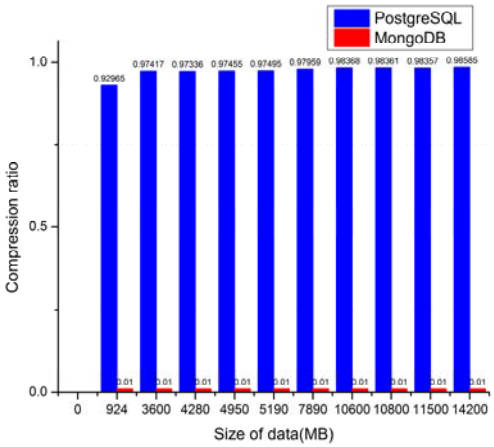

FIGURE III. COMPRESSION RATIO OF POSTGRESQL AND MONGODB

As shown in FIGURE II , for unstructured data with sizes of 3600MB, 10600MB and 14200MB, PostgreSQL time consumption is 567s, 1841s and 2535s respectively, while time consumption for MongoDB is 120s, 331s and 438s respectively. When the data volume reaches 14800MB, the time consumed by them is 2738s and 462s respectively. The time taken by PostgreSQL and MongoDB to store data has a linear relationship with the amount of unstructured data. PostgreSQL's rate is 0.178s/MB, while MongoDB's rate is 0.031s/MB. MongoDB's efficiency is about 6 times that of PostgreSQL. However, as you can see from FIGURE III, each time data is written to PostgreSQL, the compression rate is 95%, while MongoDB is less than 1%.

IV. Conclusion

In this research work, Postgresql and MongoDB are compared by testing the write speed of unstructured data and the incremental performance of write database data. The writing speed of unstructured data in MongoDB is six times that of Postgresql, and the compression performance of Postgresql is much better than that of MongoDB. Postgresql already compresses data when it writes to the database, thus increasing the time consumed to insert into the database. The two databases have different advantages in different scenarios.

Overall, MongoDB can provide a quick solution for storing unstructured data, while PostgreSQL can provide better performance quality for scenarios that require accurate data structure and small data volumes. This comparison work will be used for the design of big data NoSQL database with different requirements. Future work will focus on how to unify the advantages of fast write speed and high compression rate, and provide efficient data management support for big data by combining distributed processing.

Acknowledgment

This study was funded by the National Key R&D Program of China (2018YFC0604001-3) and B&R Team of Chinese Academy of Sciences (2017-XBZG-BR-002). We would like to thank Xinjiang Laboratory of Mineral Resources and Digital Geology of the Chinese Academy of Sciences for guidance and full support.

References

[1] Boyd D, Crawford, Kate. CRITICAL QUESTIONS FOR BIG DATA[J]. Information Communication & Society, 2012, 15(5):662-679.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Mcafee A, Brynjolfsson E. Big Data: The Management Revolution[J]. Harv Bus Rev, 2012, 90(10):60-66..

[3] Zhou, L., Chen, N., Chen, Z., & Xing, C. (2016). ROSCC: An Efficient Remote Sensing Observation-Sharing Method Based on Cloud Computing for Soil Moisture Mapping in Precision Agriculture. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(12), 5588–5598. doi:10.1109/jstars.2016.2574810 .

[4] Vitolo, C. , Elkhatib, Y. , Reusser, D. , Macleod, C. J. A. , & Buytaert, W. Web technologies for environmental Big Data[J]. Environmental Modelling & Software, 2015, 63:185-198..

[5] Liu Y , Wang Y , Jin Y . Research on the improvement of MongoDB Auto-Sharding in cloud environment[C]// Computer Science & Education (ICCSE), 2012 7th International Conference on. IEEE, 2012..

[6] Kang, Y.-S., Park, I.-H., Rhee, J., & Lee, Y.-H. MongoDB-Based Repository Design for IoT-Generated RFID/Sensor Big Data. IEEE Sensors Journal,2016, 16(2), 485–497.

[7] Brodzik, M. J., Billingsley, B., Haran, T., Raup, B., & Savoie, M. H. EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets. ISPRS International Journal of Geo-Information, 2012,1(1), 32–45.

K