

A Distributed Data Storage Method based on Hadoop

Yujuan Zou^{1,*}, Yuting Jiang¹, Taizhi Lv²

¹School of Information Technology, Jiangsu Maritime Institute, Jiangsu Nanjing 211170, China

²Nanjing Longyuan Microelectronic Company Limited, Jiangsu Nanjing 211106, China

Abstract

The distribution of data acquisition equipment is scattered, and it is easy to be affected by network environment and other factors. In order to improve the security of data storage, this paper designs and implements a data storage method based on distributed platform. It includes distributed storage server based on Hadoop, data monitoring end, data extraction end and data transmission end. The data storage cluster side records the modification and update of the data, the data backup end collects the change records of the data and copies them. The data repair end collects the change records of the backup data and repairs the damaged data according to the modification records. The data encryption end is mainly used to encrypt and protect the stored data of the data storage end and the data backup end. This method can realize the functions of fast storage, preservation, recovery and encryption of data.

Keywords

Distributed storage; Hadoop; Data Encryption; Data transmission; Cloud computing Technology.

1. Introduction

Data come from all objective existence, including the physical world from macro to micro, the results of various organisms, human social activities, perception, cognition and thinking. The amount of data generated in 2020 is 44 times that in 2009 [1]. With the development of computer technology, the establishment of unstructured data processing and analysis methods is mature, and the emergence of MapReduce, stream computing and other models has greatly reduced the cost and technical threshold of large-scale data processing [2-3]. With the development of Internet of things technologies such as nb-iot and CAT1, data acquisition becomes more convenient, intensive and extensive [4-5].

With the decrease of data storage cost and unit information storage cost, the difficulty of distributed storage technology for massive data is reduced. However, the distribution of data acquisition equipment is relatively scattered, and it is easy to be affected by network environment and other factors. When the data acquisition device cannot send data to the storage server, it can only store the collected data in the data acquisition device, but not in the storage server, which leads to the incomplete data stored in the storage server and unable to ensure the reliability of the data. Moreover, with the continuous development of Internet technology, the security performance of data processing and storage services of existing technologies needs to be improved [6]. This paper provides a distributed data storage service method to solve the problems raised in the background technology, and realize the functions of fast storage, storage, recovery and encryption of data.

2. System Architecture

A data extraction and storage method based on cloud end and server. The cloud data storage server includes data storage end, data backup end, data repair end and data encryption end for

data storage. The data storage side is used to record the modification and update of data. The data backup side collects the change record of data and copies it. The data repair end collects the change record of data backup and changes the damaged data according to the modification record. The data encryption end is mainly used to encrypt and protect the stored data of data storage end and data backup end. The data monitoring terminal includes not only data acquisition, but also data monitoring and evaluation. When the new data collected by the data acquisition equipment of the data monitoring terminal changes, the existing data of the data storage end is retrieved for comparison, and the corresponding evaluation is conducted according to the comparison results. The data extraction end is used to retrieve the latest data stored in the data storage end. The communication connection of each storage service stage is established through the data transmission end. The distributed data storage architecture is shown in Figure 1.

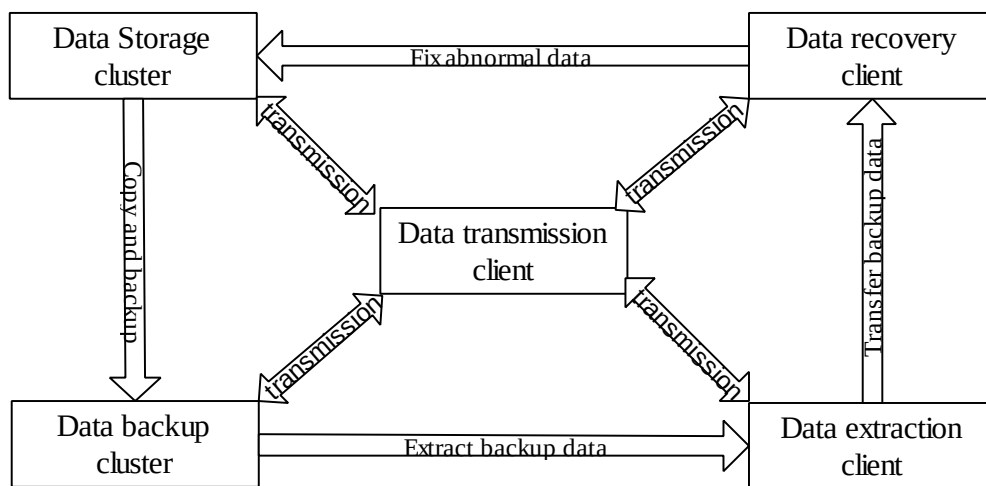


Figure 1. System Architecture

3. Data Storage Flow

As shown in Figure 2, the proposed algorithm is divided into four steps: data annotation, data comparison, data storage and data backup.

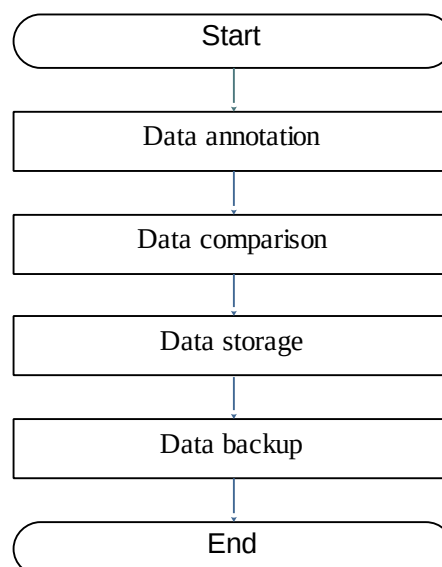


Figure 2. Data storage flowchart

Data annotation step: When the data acquisition equipment receives the new data that needs to be updated, it calls the stored old data through the data extraction client, compares the data of two different periods, marks the same points of the two different periods of data, and evaluates the changes.

The data monitoring client is also used to monitor whether the data update, repair, abnormal data deletion, data backup and storage are normal in each process.

Data comparison step: the data evaluation mainly compares with the data of previous storage and backup, and judges the change trend of new data collected this time according to the fluctuation range of previous data;

The abnormal data is detected at the data monitoring client, and the abnormal data is compared with the old data backed up by the data backup client, and then the two groups of data are compared and monitored. The specific comparison data includes the time when the abnormal data occurs and the specific content of the big data changes caused by the abnormal data. The abnormal data is completely deleted in the data monitoring client, and then the data with the first sequence in the memory of the data backup client is retrieved for replacement.

Data storage step: the data storage client updates and saves the new data according to the collected new data.

All the data in the data storage client and data backup client are sorted, the latest data is ranked first, the rest data are sorted according to the update time, and the data with the earliest update time is sorted last. After receiving the new data that needs to be updated, the old data is saved. After saving, the new data will be sorted to the first place, and the remaining old data will be carried forward to the next.

The encryption program used in data encryption client includes dynamic verification code, system password and iris verification encryption.

Data backup step: at the same time of new data storage, the data backup client will immediately backup and save the old data before storage.

For a certain data a, the updated new data is recorded as A1 in a certain update. Under the condition that the monitoring data at the data monitoring client is normal, the two groups of data are simply evaluated, and then the data storage client stores and saves the A1 data. At the same time, the data backup client copies and backs up the A1 data, and the sequence number of the original data a is postponed by one bit. The A1 data is arranged in the first place of the sequence number, and the remaining historical data are successively postponed to the next.

4. Conclusion

Compared with the traditional data storage technology, the method proposed in this paper compares with the backup data, and judges the trend of the new data collected this time according to the fluctuation range of the data. At the same time of new data storage, the data backup client will immediately backup and save the old data before storage. Through the data encryption client, the whole process is encrypted to ensure the security of data storage service process. The algorithm proposed in this paper monitors whether the data update, repair, abnormal data deletion, data backup and storage are normal in each process. The algorithm proposed in this paper can store the records of abnormal data, and store the comparison records and specific abnormal contents of abnormal data at the data monitoring client.

Acknowledgements

This work was financially supported by the funding of innovation science and technology project of Jiangsu Maritime Institute (The research of data analysis and application for AIS based on Hadoop), the research project of modern educational technology in Jiangsu Province

(2018-R-65176), young academic leaders of Jiangsu colleges and universities QingLan project, and the shipping big data collaborative innovation center of Jiangsu Maritime Institute.

References

- [1] Hu Y. Application of big data in material supply enterprises [J]. Enterprise reform and management, 2020, 000 (008): 31-33.
- [2] Chen C H , Lin J W , Kuo S Y . MapReduce Scheduling for Deadline-Constrained Jobs in Heterogeneous Cloud Computing Systems[J]. IEEE Transactions on Cloud Computing, 2018, 6: 127-140.
- [3] Lv T , Tong L , Zhang J , et al. A real-time physiological signal acquisition and analyzing method based on fractional calculus and stream computing[J]. Soft Computing, 2020(10),1-7.
- [4] Rashmi, Sharan, Sinha, et al. A survey on LPWA technology: LoRa and NB-IoT[J]. ICT Express, 2017,3(1) :14-21.
- [5] Bo C , Zhihui G . Analysis on the Coverage Capability of NB-IoT and CAT 1[J]. Mobile Communications, 2017, 41(22):57-61.
- [6] Li W, Huang Q. Design of multi secret data security storage system based on blockchain technology [J]. Journal of Shenyang University: Natural Science Edition, 2019, 0.