Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics (CONF-CIAP 2023)
DOI: 10.54254/2753-8818/5/20230258

Spark Computing Framework for Pet Medical Information Management System in Pet Healthcare

Tianyi Fan[1,*]

[1]University of Tasmania Tasmania, Australia 7005

[*]tianyif@utas.edu.au

Abstract. In recent years, the role of big data technology in various industries has become increasingly prominent. With the rise of the pet trend, the pet medical industry has been developing rapidly. However, the current application of big data in the pet medical industry is single and elementary. This study aims to improve the current situation of big data technology in the pet medical industry and build a pet medical information management system supported by Spark computing framework, HDFS, and HBase. This study uses descriptive research method and comparative analysis method to prove that big data analysis based on Spark framework can greatly improve the efficiency of treatment and reduce the time cost. The information management system based on Spark framework can realize the rapid storage and calculation of massive data, reduce the technical threshold of data application area in pet medical industry, and help to promote the accelerated development of big data industry and pet medical industry.

Keywords: Spark, Big Data, Pet Medicine, Database, Information Management System.

1. Introduction

In recent years, with the rapid development of information technology, the field of medical care and medical research is entering the era of big data, and the daily growth of medical data has reached terabytes [1]. It cannot be ignored that a large number of pet diagnosis and treatment activities will also generate a large number of different types of data which have the "4V" characteristics in the general sense of big data.They are having large data volume, wide variety, fast flow speed and low value density. At the same time, pet medical big data also has high dimensional characteristics and uncertainty of use [2].

With the continuous development of medical informatization and the rise of pet ownership, the pet medical industry has developed rapidly, and various pet medical institutions have accumulated a large amount of pet medical and health data. These data themselves contain a lot of medical knowledge and information, which is of great analysis and mining value to pet medical research, pet medical services and other fields. Therefore, the role of big data technology in the pet medical industry is becoming increasingly prominent, and many pet medical industries are beginning to use related technologies to improve operational efficiency. However, the application methods of big data in the current pet medical market are slightly single and elementary, and cannot give full play to the role and value of these technologies. This study makes up for the deficiencies of big data technology in the pet medical market by introducing the big data computing framework and establishing the prototype of the pet medical information management system.

DOI: 10.54254/2753-8818/5/20230258

This paper analyzes the characteristics and storage requirements of pet medical big data, introduces and compares two big data computing frameworks, and explains the current status and role of big data technology in the pet medical industry. The paper also proposes an effective integrated solution to the current technical problems of pet medical big data in terms of data storage and computation speed by using Spark computing framework, distributed file system (Hadoop Dist ributed File System，HDFS) and distributed database (Hadoop Database, HBase) combined technology as the support to build pet medical information management system. At the same time, some problems of the system are proposed in order to make the big data technology better applied in the pet medical industry.

2. Big data architecture

With the development of information technology such as the Internet, more and more TB, PB and even FB level data are generated, and Hadoop and Spark, as the dominant platforms for big data processing today, have been widely used in various fields.

2.1. HADOOP

Hadoop is the earliest big data processing framework, providing a scalable, flexible and reliable distributed computing big data framework for system clusters with storage capacity and local computing power. There are some important components that play an important role in the Hadoop architecture. The first is the Hadoop Distributed File System (HDFS) used as a distributed storage for data, which is a distributed file system that provides maximum throughput access to information. The second is Hadoop YARN, which is responsible for job scheduling and managing cluster resources. In addition, there is Hadoop MapReduce, which are modules that rely on YARN to process data in parallel [3]. Shown in the figure is the master-slave architecture followed by Hadoop, where the Name node contains metadata about all data blocks in HDFS for transformation and analysis of large data sets using the Hadoop MapReduce paradigm [4].

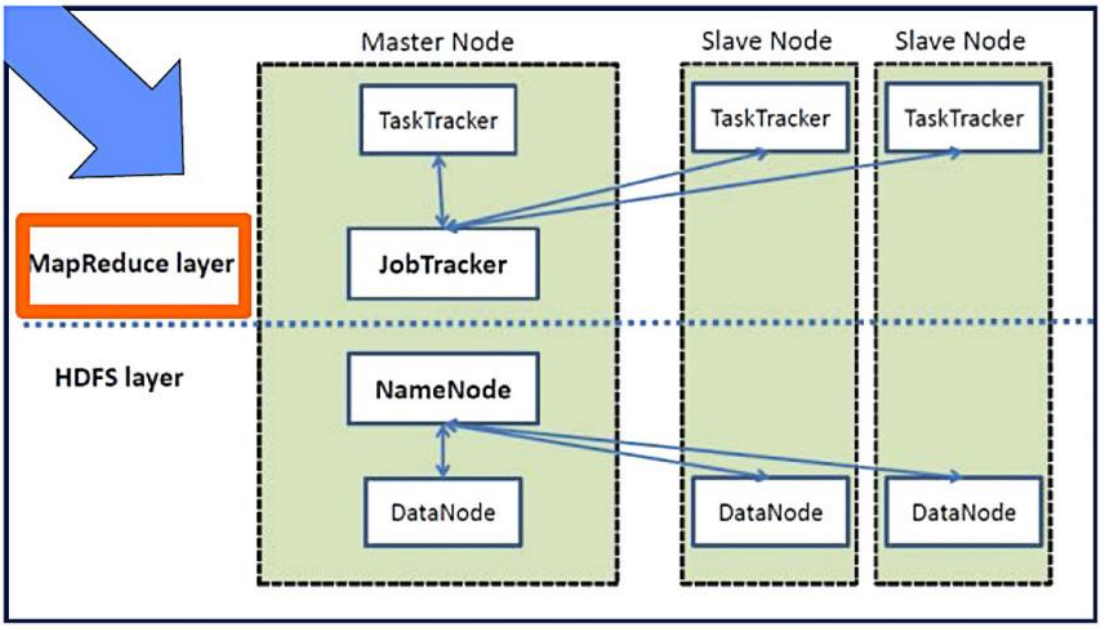

Figure 1. Master–slave architecture of Hadoop.

2.2. Spark

Open-source projects such as MapReduce and HDFS in the Hadoop environment are still indispensable tools for processing big data. However, with the increase of data volume and the emergence of more

DOI: 10.54254/2753-8818/5/20230258

and more applications such as machine learning, big data platforms have higher requirements for the efficiency of data processing, thus Spark was invented. The emergence of Spark is to improve various problems of Hadoop MapReduce in practical applications [5]. Apache Spark is an open-source parallel processing framework that is much faster than Hadoop in large-scale processing because it provides in-memory primitives that minimize data transfers from and to disk. The basic data structure of Spark is RDD (Resilient Distributed Dataset). RDDs help Spark perform MapReduce operations faster, especially for iterative tasks [6]. Spark can be used on Hadoop data sources and fits well into the Hadoop ecosystem [7]. The following figure shows the basic architecture of spark [8].

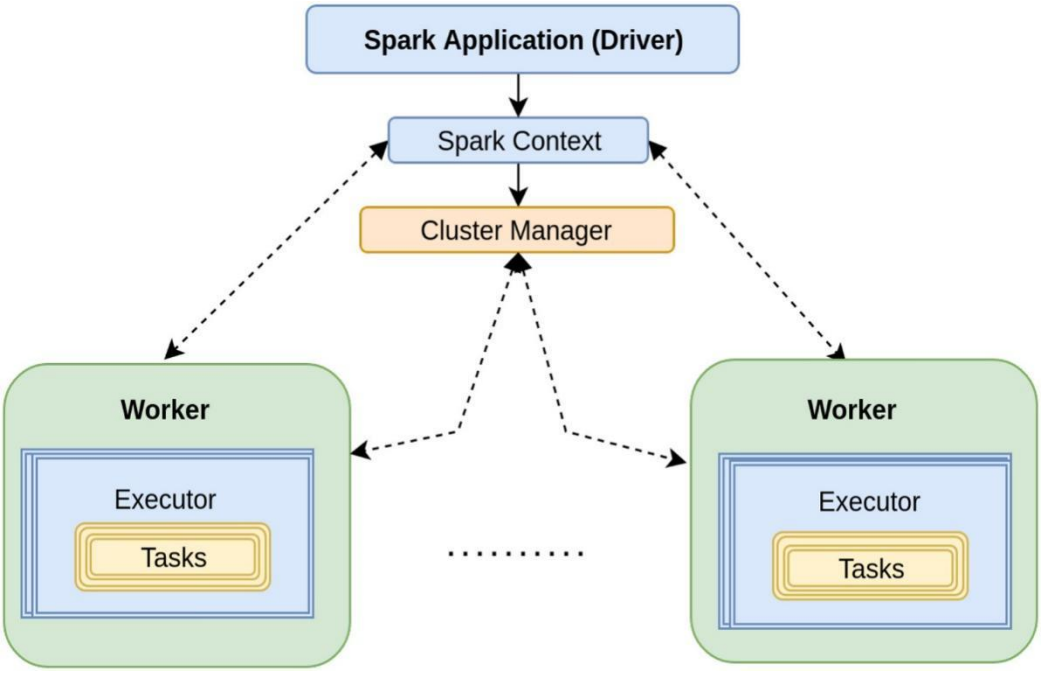

Figure 2. Spark architecture.

2.3. Comparison of HADOOP framework and Spark framework

The following table compares the HADOOP framework with the Spark framework

Table 1. Comparison of the HADOOP framework with the Spark framework.

	Hadoop	Spark
Computational level	Disk-level computing	In-Memory Computing
Performance	Slower, because Hadoop stores data on multiple sources and processes it in batches via MapReduce.	Faster, because it uses RAM.
Data Processing	Batch processing and linear data processing.	Suitable for iterative and real-time streaming data analysis. Run operations with RDDs and DAGs.
Scalability	Easily expand storage by adding nodes and disks.	Scaling is more challenging because it relies on RAM for computation.
Language Support	Hadoop framework is based on Java; MapReduce code is based on Java or Python.	Java, Python, R, and Spark SQL
Cost	Lower	Higher

DOI: 10.54254/2753-8818/5/20230258

Hadoop framework and Spark framework process data in completely different ways. While both Hadoop with MapReduce and Spark with RDDs process data in distributed environments, Spark can process real-time streams of unstructured data more efficiently with the help of in-memory computing and high-level APIs.

After comparing the Hadoop framework and Spark framework, combining the scale of pet medical information data volume and the complex structure of medical data, including structured and unstructured and semi-structured data, with various data types, Spark big data processing framework is used as the core to build the pet medical information management system.

3. The current status of the application of big data in the pet medical industry

Pet medical big data can be applied to a variety of fields such as disease diagnosis and treatment, scientific research, and industry regulation. However, there are several key technical problems that need to be solved urgently when using big data analysis and research in the pet medical industry. First of all, the traditional statistical methods and software are not powerful enough to cope with the actual management needs of massive data. Simple statistical tools such as Excel tables and professional statistical analysis software such as SPSS and SAS are commonly used for data analysis in clinical research. Statistical analysis software only performs statistical analysis on specific data with small sample size and simple structure, and is not suitable for in-depth analysis of large volume and complex structure of big data. Secondly, popular machine learning algorithms in big data analysis in recent years usually require programming, which makes it impossible for many doctors to use machine learning algorithms in scientific research independently, and requires the cooperation of professional data analysts, which increases labor and time costs. Finally, most of the data of medical institutions are stored in relational databases. Traditional relational databases cannot store unstructured data, and there will be problems such as slow data reading and storage, and long computing time, which cannot meet the needs of data storage. In addition, in the current pet medical industry, the real-time and reliability of data acquisition and the accuracy of data analysis are also difficult problems that need to be solved [9].

Distributed technology has been widely used in the storage field due to its advantages of low cost, high reliability and large capacity, which provides a new idea for storing massive pet medical data. This technology stores, manages and processes massive data in a distributed manner by connecting multiple common devices, and supports the storage of unstructured data.

4. Data structuring

The data generated by the pet electronic medical record system is big data, which mainly includes the pet's basic information, medical records, doctor's orders, nursing documents, inspection images, inspection conclusions, etc., which can be summarized into three parts: animal past medical record data, animal medical inspection data and animal medicine image data. These data reflect the basic situation, diagnosis, treatment process and treatment results of animals. Animal medical records are data written in natural language generated by the owner's description of symptoms and the doctor's records, and it is an unstructured data. Animal medical inspection data comes from animal medical inspection equipment, such as blood routine examination, liver function test and electrocardiogram examination, etc. The data generated by these medical instruments is generally mathematical data, and there are standards and norms, so it is a structured data. Animal image data comes from imaging equipment, which is a medical inspection device with image display as the detection structure. The data generated by it is images, which is a kind of unstructured data [10].

Structured data is the basis for data analysis. How to make computers understand the semantics contained in this medical information and efficiently retrieve, count, analyze and mine this data will be an important issue in the construction of information technology in the medical field.

In the electronic medical record system, the medical images of all animals are stored through distributed files, and real-time parallel computing services and unstructured data processing are provided. The hospital medical data is processed through natural semantic processing technology, combined with the semantic structure of medical terminology, and the medical semantic information is expressed from

DOI: 10.54254/2753-8818/5/20230258

the original natural language, expanded and analyzed into a structured Key-Value model, providing basic data support for subsequent applications. The structuring is mainly carried out from several independent dimensions, and the medical record data is divided according to the subject fields. The main subject fields are: symptoms at the time of onset, animal signs, daily health conditions, clinical pathological diagnosis, etc.

5. Architecture of pet medical information management system

Traditional pet medical institutions usually use manual processing of medical information, there is a risk of information leakage, and there are deficiencies in accuracy and efficiency, resulting in pets not being able to receive timely diagnosis and treatment. Establishing a complete information system for pet medical information, and entering pet information into the database can effectively make up for existing problems and improve pet medical treatment time. Therefore, in order to dig deep into the value of pet medical big data, the most urgent thing is to realize the informatization of pet medical institution management.

The information management system for pet medical information takes electronic medical records as the center, effectively organizes the original case system, imaging system, inspection system and pet owner management system, and connects the electronic medical record systems of different medical institutions through the Internet. Realizing data sharing among many pet medical institutions allows doctors to grasp the pet's past health status and examinations that have been done in a timely manner, reducing repeated examinations and improving the efficiency of medical treatment.

5.1. Distributed computing framework

The system adopts a distributed Spark cluster, and Spark provides a variety of distributed deployment modes, such as Standalone, Yarn and Mesos. The Yarn mode allows Spark to run in the existing general resource management system, which can improve resource utilization, which is the trend of future development [11]. Spark's resource scheduling framework is similar to Yarn. Spark running on Yarn does not require excessive adaptation and modification, and is easier to install and deploy. In the pet medical information management system, yarn acts as a cluster manager, allowing Spark to run on the physical nodes where data is stored to quickly access HDFS. The following figure shows the resource scheduling process of Spark On Yarn.

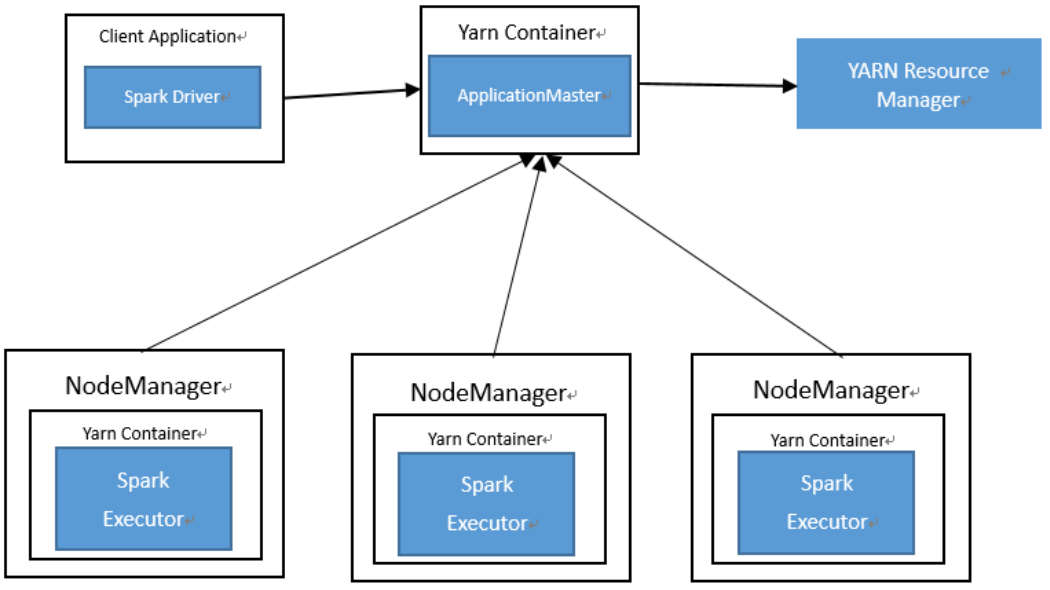

Figure3. Resource scheduling process of Spark on Yarn.

DOI: 10.54254/2753-8818/5/20230258

According to the figure, it can be known that the operation process is as follows:

The client submits a request to the ResourceManager (RS) and uploads the jar to HDFS

After the RS receives the request, select a NodeManager (NM) in the cluster to allocate the Container, and start the ApplicationMaster (AM) process in the Container

AM starts, AM sends a request to RS to request a batch of containers to start Executor.

The RS returns a batch of NM nodes to the AM.

AM connects to NM and sends a request to NM to start Executor.

The Executor is reverse registered to the Driver of the node where the AM is located. Driver sends task to Executor.

5.2. Distributed storage

Distributed storage consists of HDFS and HBase. The platform imports the data files uploaded by users into HDFS; the data collected in real time is stored in HBase. Hbase is a distributed and scalable big data storage, which can realize random and real-time reading and writing of big data, which is convenient for real-time online data analysis and rapid visual display [12].

6. The role of pet medical information management system

A lot of information can be parsed and utilized by the different big data generated by the electronic medical record system.

First of all, the pet medical information management system can improve the efficiency of medical treatment. When pets go to the hospital for the first time, all their information will be filled in the electronic medical record, so that no matter which medical institution the pet goes to, each pet hospital can retrieve relevant useful information through the electronic medical record. Doctors can avoid repeated inquiries by looking at the electronic medical records to get a preliminary understanding of the situation. Combined with certain examinations, doctors can diagnose and give treatment plans.

Second, the pet medical information management system can standardize the diagnosis and treatment process. The pet medical information management system assists doctors to complete the consultation process step by step from consultation, testing, diagnosis, prescription to doctor's orders, which can help doctors reduce mistakes and increase the accuracy of diagnosis.

Third, the pet medical information management system is helpful for scientific research. Pet information, diagnosis and treatment information, prescription orders and inspection reports together constitute the big data resources of pet medical care. Through the collection, extraction and transformation of these data, a medical information resource base can be formed to provide data support for veterinarians and scientific researchers.

In short, compared with the traditional paper medical records, this way of seeing a doctor can save the precious time of sick pets and standardize the diagnosis and treatment process, thus playing a huge role in promoting the development of veterinary medicine.

7. Existing problems and analysis

Distributed technology can realize unified storage and query of medical data, but there are still some problems in current research.

First, in terms of data storage, Hbase is less efficient for non-primary key queries. HBase (Hadoop Database) can achieve millisecond-level response in the face of primary key-based data query, and the query performance is very efficient. However, due to the lack of non-primary key indexes, in the face of complex non-primary key queries, the entire table must be scanned, which is time-consuming and cannot meet the data query scenarios that require fast response. In addition, when the Region server in HBase stores data, it first stores the data in memory. After the memory reaches the threshold, the Region server will flush the data to the disk. If the storage file is too large, it will frequently trigger time-consuming operations such as compact and split [13]. This has a large impact on performance.

Second, medical institutions usually use a centralized method to manage data, but this management method is not transparent enough, which easily leads to the problem of data tampering. In addition, in

DOI: 10.54254/2753-8818/5/20230258

the process of collection, transmission, cleaning, storage and analysis of big data, pet owners' information leakage and commercial confidentiality issues are problems that need to be solved urgently. These problems directly threaten the data security in the medical field and the information security of pet owners, making it difficult to share data between institutions at all levels, and unable to fully utilize the value of medical data. To solve the problem of information tampering and leakage, firstly, technical means can be used to automatically filter user private information from the original medical record data through the program. In addition, for the protection of commercial confidential data of pet medical institutions, strengthening legislation is an effective solution. It must be stipulated that only veterinarians and pet hospitals can use pet diagnosis and treatment data; only veterinarians and pet hospitals that provide medical services can access and use the personal information of pet owners being served; must have explicit authorization to access commercial data of pet hospitals [2].

8. Conclusion

The big data analysis platform built in this study uses Spark computing framework as the core architecture, which can process large amount of data concurrently and distributedly to improve the processing efficiency of pet medical data; through in-memory computing technology, it can avoid disk I/O operations and improve the computational efficiency. The platform uses HDFS as the underlying distributed storage environment to realize the rapid storage and processing of massive pet medical data.

To sum up, although the current big data technology is not fully used in the medical industry, and the current big data technology cannot completely replace the traditional data statistics and analysis mode, the big data technology has shown its unique advantages and bright prospects. Through big data analysis, the correlation in the pet medical treatment process can be found more quickly and easily, providing directional and strategic reference for pet disease diagnosis and treatment, scientific research, etc., and helping to improve the ability of animal disease prevention and control and the efficiency of industry supervision. The combination of the Internet and the Internet of Things, cloud computing and big data will enable more pet owners, pet hospitals and pets to be connected in the future to achieve greater information sharing and promote the further development of the entire industry. At the same time, improving big data analysis technologies and methods, improving information management mechanisms, and strengthening privacy and rights protection are also urgent needs that accompany big data.

This paper constructs a prototype of a pet medical information management system, but it lacks algorithm-level research and is still in the preliminary research stage, and the experimental scheme is not yet complete. In addition, the pet medical information management system designed in the study has an electronic medical record system as its core, which has functional limitations. In the future, a big data-based pet intelligence management platform can be used as the research direction to build a functionally integrated system.

References

[1] LI Huibo, LIU Haitao & WU Yiping.(2022). Practice of Smart Management Platform Based on Big Data Construction at Hospital. Chinese Journal of Health Informatics and Management (01),110-115.

[2] Wu Liping, Chen Penggang & Zeng Jihui.(2016). Exploring the application of big data in pet healthcare. Animal Health(2),42-43.

[3] Rajendran, S., Khalaf, O. I., Alotaibi, Y., & Alghamdi, S. (2021). MapReduce-based big data classification model using feature subset selection and hyperparameter tuned deep belief network. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-03019-y.

[4] Basha, S. A. K., Basha, S. M., Vincent, D. R., & Rajput, D. S. (2019). Master–slave architecture of Hadoop [Figure]. Challenges in Storing and Processing Big Data Using Hadoop and Spark. https://doi.org/10.1016/b978-0-12-816718-2.00018-x.

[5] Quan Zhao Heng, Li Jia Di. (2019). From Hadoop to Spark Technology Innovation. Computer Knowledge and Technology15(8),265-268.

DOI: 10.54254/2753-8818/5/20230258

[6] Kadkhodaei, H., Eftekhari Moghadam, A. M., & Dehghan, M. (2021). Big data classification using heterogeneous ensemble classifiers in Apache Spark based on MapReduce paradigm. Expert Systems with Applications, 183, 115369. https://doi.org/10.1016/j.eswa.2021.115369.

[7] Li, J., Zhang, C., Zhang, J., Qin, X., & Hu, L. (2022). MiCS-P:Parallel mutual-information computation of big categorical data on spark. Journal of Parallel and Distributed Computing, 161, 118–129. https://doi.org/10.1016/j.jpdc.2021.12.002.

[8] Bedi, J., & Toshniwal, D. (2022). Spark architecture [Figure]. Spark Map Reduce Based Framework for Seismic Facies Classification. https://doi.org/10.1016/j.jappgeo.2022.104762.

[9] ZHU Chengzhang, LIU Zixi, LI Wenjing, XIAO Yalong & WANG Han.(2022). Research Review of Distributed Medical Big Data Storage Scheme. Software Guide (04),7-12.

[10] Xie Fanghua, Meng Ge, Bao Xijun. (2016). A discussion of issues related to big data in pet healthcare. ZHONGGUO GONGZUO QUANYE (12),47-50.

[11] Wen Yanqi. (2017).Research and Implementation of Performance Modeling and Optimization Technology of Spark Computing Framework (Master's thesis, XIDIAN UNIVERSITY). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017301815.nh.

[12] Apache HBase – Apache HBaseTM Home (2022, August 6). The Apache Software Foundation. Retrieved August 7, 2022, from https://hbase.apache.org/.

[13] Ji Yimu, Zhang Ning, Yao Haichang, Li Kui, Li Hang, Liu Shangdong & Wang Ruchuan.(2019). HOS:design and implementation of distributed storage system based on HBase. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition) (05),63-71. doi:10.14132/j.cnki.1673-5439.2019.05.009.