Construction and Research about Big Data Service System of Science and Technology Management Based on HBase

Lin MU\textsuperscript{1,a,*}, Jie GUI\textsuperscript{1,b}, Xi ZHANG\textsuperscript{1,c} and Ai-Dan Li\textsuperscript{1,d,*}

\textsuperscript{1}Institute of Scientific and Technical Information of China, Beijing 100038, China
\textsuperscript{a}mulin@istic.ac.cn, \textsuperscript{b}guij@istic.ac.cn, \textsuperscript{c}zhangxi@istic.ac.cn, \textsuperscript{d}liad@istic.ac.cn

*Corresponding author

Keywords: Big Data, Science and Technology Management, HBase, Hadoop, MapReduce.

Abstract. With the rapid development of science and technology in China, a large amount of scientific and technological management data has been accumulated. In order to avoid the phenomenon of "Information Island", it has become an urgent need to develop a mass information processing system. How to manage it efficiently becomes a problem, while traditional methods are expensive and difficult to extend. This paper builds a massive data processing prototype system with the functions of data preprocessing, data retrieval and data mining for scientific and technological management of big data, based on Hadoop parallel computing platform and HBase No-SQL database.

Introduction

With the rapid development of China's economy, the development of science and technology has made great progress, and the main quantitative indicators of scientific and technological innovation have leaped to the top three in the world. China has become an influential country of innovation. The international ranking of 10 indicators of scientific and technological innovation, such as personnel, funds, papers and invention patents, has all entered the top three in the world. Among them, 6 indicators, such as R&D personnel, scientific and technological journal papers, application amount of invention patents, authorization amount of invention patents and export amount of high-tech products, rank NO.1 in the world. According to the data in 2017, China has 4 million scientific researchers, 1750 billion yuan of research and development costs, ranking 2th in the world, 42% of the United States; about 350 thousand SCI papers, ranking 2th in the world and the cited index of international papers ranks 2th in the world for the first time.

How to manage the massive science and technology data efficiently becomes a problem because traditional ways are expensive and difficult to extend. Taking the science and technology management information system as an example, which manages five kinds of national science and technology plan projects. Facing massive project information, personnel information, organization information and achievement information, science and technology management system has lots of problems in low performance and low scalability in the process of data query, and this data information also include picture, video, text, audio and other information. Alibaba uses big data technology to store, query and analyze massive data with hive data warehouse and HBase database. Facebook has a huge amount of photos and video and audio information. Every week, more than hundreds of millions of photos are uploaded, downloaded and browsed in Facebook, which uses HBase to query and retrieve data in real time.

Preliminary

This section introduces Hadoop, HBase and MapReduce. We use HBase for storage and MapReduce for computation.

Hadoop is one of the most popular cloud computing frameworks that have proven to scale and perform well on clouds [1,2]. Currently, many well-known companies build applications based on
Hadoop, such as Alibaba, Facebook and so on. Hadoop is highly fault-tolerant and is designed to be deployed on low-cost hardware, Hadoop uses hadoop distributed file system (HDFS) to support the low-level distributed storage.

HBase is a distributed database based on Hadoop distributed file system [3, 4]. HBase is different from traditional relational database. It is a high reliability, high performance, column oriented, scalable database. Based on the Hadoop distributed file system, HBase stores data in the form of key \ value, supports transaction processing, but it is more suitable for the massive data processing and query business with high transactional processing requirements, and can only retrieve data through the row key and the range of the primary key, with the ability to process huge data and high concurrency [5].

MapReduce is a well-known parallel programming model presented by Google. The idea wall originally from functional programming language [6, 7, 8]. It’s a programming model and an associated implementation for processing and generating large data sets. Users need to define a map function that processes a key/value pair to produce a group of intermediate key/value pairs, and a reduce function that combines all intermediate values associated with the same intermediate key.

**Methodology**

In the Big Data Service System of Science and Technology Management, the Hadoop + HBase mode can effectively solve the query and application of massive data. Its technical architecture is as follows:

![Technical Architecture Diagram](image)

In the whole technical architecture, the underlying file and data processing can choose Hadoop to solve the storage problem of massive data through Hadoop. In Hadoop framework, the same file can be divided into several copies, and each copy of data is stored on multiple data nodes. This mode can not only effectively ensure the security of data, but also reduce the cost of the whole system by configuring relatively poor performance equipment on the data nodes.

**Storage Architecture**

In the database of HBase, data model is established, data in relational database is transformed, extracted and loaded by using tools such as sqoop according to the technical requirements of HBase, and query database based on HBase + Hadoop is established.
Taking the detailed query of students in the student status system as an example, in the whole architecture, the namenode host stores the metadata and index data of the student status information. After the detailed information of students is indexed, it is divided into several copies. Each datanode stores one of them. At the same time, it can store 2-3 copies of data of other nodes if conditions permit.

MapReduce of Tasks

The feature of data operation based on HBase + Hadoop is to decompose the task first, decompose the task to different data nodes according to different models, summarize and return the results again after the data node operation is completed, and realize the parallel calculation that the relational database cannot solve. By using the map / reduce mode and the computing power of multiple data nodes, the results can be returned quickly.
Experiment

Simulation environment:
CPU: Intel (R) Xeon (R) CPU e5-2660 V2 @ 2.20GHz 32 core; Memory: 128G; Hard disk: 10T;
Concurrent number: 500; Increase mode of concurrent users: initialize 5 concurrent users per second.
Data volume: about 0.5 billion, 1.3TB (including project, personnel, organization, paper, patent, audio, video and other data).

The results show that the average processing power of unified search is 219.837 transactions/second and the average response time is 1.962 seconds.

Conclusion and Future Work

Based on HBase tools, combined with the characteristics of multi-source data of national science and technology management information system, this paper designs and implements a query and retrieval system. Through experiments, in the case of big data, the system responds quickly.
With the development of artificial intelligence, more and more intelligent algorithms will emerge in the field of machine learning. In the storage and query of distributed non relational database, parameter optimization is a very important and necessary work. At the same time, in the interactive system, every layer needs parameter optimization. So in the next step, we will pay more attention to the application of intelligent algorithm and cross layer parameter optimization.

Acknowledgement
This research was financially supported by the Innovation Research Fund of Institute of Scientific and Technical Information of China in 2019 (Project No. ZD2019-14 and ZC2019-07).

References
[1] Tom White. Hadoop: The Definitive Guide[M]. America: O’Reilly Media, 2009:18-25.
[2] The Apache Software Foundation Welcome to Hadoop[EB/OL] http://hadoop.apache.org/
[3] http://en.wikipedia.org/wiki/HBase.
[4] Levin R, Kanza Y. Stratified-Sampling over Social Networks using MapReduce[C]. ACM SIGMOD International Conference on Management of Data. ACM, 2014:863-874.
[5] Dimiduk N, Khurana A, Ryan M H. HBase in action[M]. Shelter Island: Manning, 2013.
[6] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.
[7] Song J, Sun Z Z, Mao K M, et al. Research Advance on MapReduce Based Big Data Processing Platforms and Algorithms[J]. Journal of Software, 2017, 3: 514-543.
[8] Jiang D, Ooi B C, Shi L, et al. The performance of MapReduce: an in-depth study[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 472-483.