ATLANTIS PRESS
Advances in Computer Science Research, volume 50
4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016)

Research of Structured Data Query in the Equipment Support Management System Based on Cloud Platform

Li Dong[1, a], Xu Xinyi[1, b], Wang Huqiang[1, c]

[1]Academy of Armed Force Engineering, Beijing, 100072, China

[a]email:316984759@qq.com, [b]email:316984759@qq.com

Keywords: Structured Data, Hadoop, Cloud Platform, Sqoop, Hive

Abstract. Aiming at deficiencies of current Equipment Support Management System on massive data query and analysis, this article proposed and realized a method of Structured Data Query used by the Cloud Platform based on Hadoop, after analyzing the data source. In the article, we used the tool of Sqoop to import data from a relational database into HDFS, and then use Hive to analyze the data. Finally, the experiment proved that the method makes up the shortage of the data query in the environment of single computer, and has high practical value.

Introduction

Our army developed a series of equipment support management systems in the information construction of equipment support, the study found that most of the systems based on relational database. Because of the various kinds and great number of equipment, if we continue query massive data in the traditional way, it will out of the efficient processing range of relational database. We may use the cloud platform based on Hadoop to solve the problem, but HDFS in Hadoop can usually only support for unstructured data access. Relational database take hold predominance in equipment support management system, if we convert the structured data into unstructured data, it will not only waste a lot of time and finance, but also lose many important information. The Hbase in Hadoop can only suitable for storing unstructured data. So how to use cloud platform based on Hadoop to query massive data in relational database is the focus of this article. On the premise of analyzing data source in equipment support management system, this article design and realize structured data query system architecture based on cloud platform.

Relevant Technologies

Hadoop. Hadoop[1]is the product of the development of parallel technology, distributed technology and network computing technology. It is a model framework for massive data computing and storage [2]. Hadoop has two core technologies, HDFS and MapReduce.

HDFS [3] is a distributed file system running on a large number of low-cost hardware, the underlying file storage system, it is in charge of managing and storing data. AHDFS cluster has one NameNode and more DataNode. NameNode is center server, mainly used to manage metadata and file block and simplify update operation of metadata. One node in cluster usually has one DataNode for storing and querying data block.

MapReduce [4] is a programming framework for processing massive data in cloud computing of Hadoop. It is easy to use, programmers can process data without knowing the underlying implementation details. MapReduce use Map Procedure to divide massive data into many small pieces, assign them to lots of servers to deal, then return the result to Reduce. Finally, Reduce output the summary results to the client.

Hive. Hive [5] is a data warehouse infrastructure based on Hadoop. It can storage, query and analyze massive data in HDFS. Let the people who are familiar with SQL language can easily operate HDFS file system is Hive's original design purpose. Technician can interact with Hive through HiveQL language. HiveQL is a SQL-like language, their syntax are very similar to each other. Its design is influenced much by MySQL. HiveQL language generate several tasks to deal by parsing and converting statement.

Sqoop. Sqoop's full name is "SQL-to-Hadoop", it is used to transfer data between relational database and Hadoop, developed by the Clouder. We can not only import the data in HDFS to relational database (like MySQL, SQLServer, Oracle and so on) but also import the data in relational database to HDFS easily.

Equipment Support Management System Resources Data Analysis

Structure Design Of Relational Database. For the linking and sharing of business data in equipment support management system, after analyzing business data and business composition, we classified the data in every systems. From the perspective of business process, the system was classified to equipment, staff, ammunition, material, facility, establishment, information and expenditure.

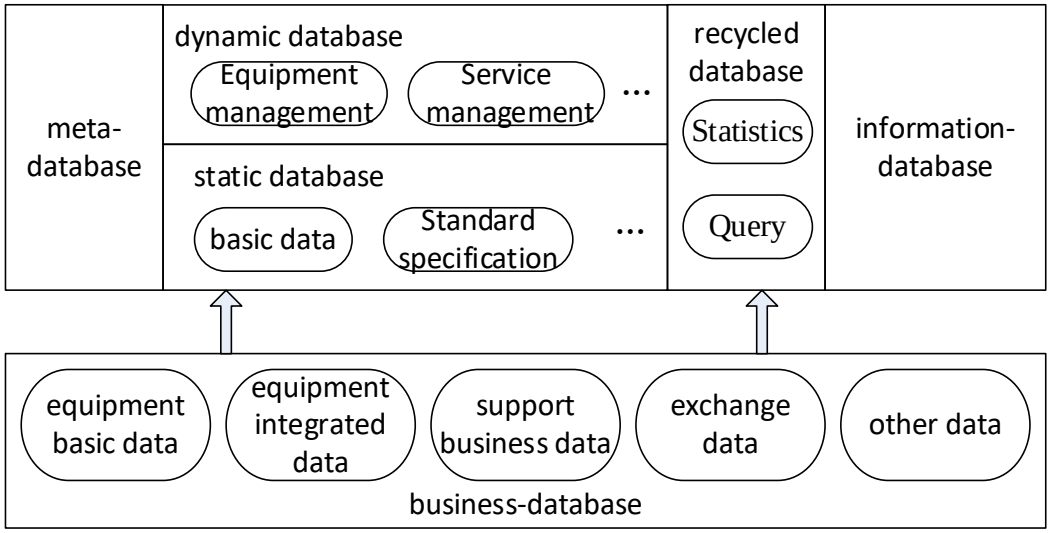

Fig.1Database logical relationship

The system has 3 kinds of database: meta-database, information-database, and business-database. The function of meta-database is managing and organizing information resource, locating resource location quickly, creating a resource directory, exchanging resource, providing information resources transformation and so on.Through analyzing business data in equipment support business information system, we divided the business data into 5 parts: equipment basic data, equipment integrated data, support business data, exchange data and other data.The database logical relationship as shown in the figure 1, business database is divided into static database, dynamic database and recycled database.

Description Method For Metadata. The information resources in equipment support management system are huge, using metadata to extract intrinsic characteristic of data can promote efficiency of data integration. By recording user permissions and data resource location, it can improve accuracy, efficiency, security, reliability of data query.Record transformational rule and vital relations through business metadata can improve system's maintainability and integratability. Metadata can not only make users and developers understand data easily, and also ensure data consistency, accuracy and integrality.Hence, it is necessary to design the metadata standard format. This article design metadata in system considering main from data descriptive elements.

RDF is a metadata system with an XML serialization for the Web, a model for describing collections of formalized statements about a Web resource, it has the characteristics of abstraction, openness, simpleness and so on, therefore, this article choose RDF as metadata description framework.RDF usually consists of three parts：resource, attribute value and attribute type.As shown in figure 2, this system design resource description method process of RDF based on literature, all resources in the system can use this process generate resources description document, interact data and share resource in heterogeneous system.

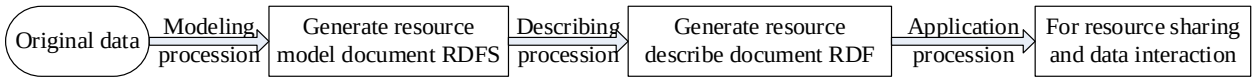

Fig.2 Steps to resource description method

Modeling: First, modeling the original data, make sure the data resources and various properties needed to describe, secondly, generate RDF model according to the relationship between each element; finally, use this model to generate resource model document.

Description: use resource model document, RDF's own vocabulary, and vocabulary defined by ourselves to describe resource, generate resource description document. This document is based on XML.

Application: through exchanging resources description document, we can found and visit resources on other heterogeneous platform easily, realize heterogeneous data sharing and interaction.

Structured Data Query System Design Based On Cloud

Systematic Architecture Design. Query mass data in a traditional database is inefficient. Users always wait for a long time. So, traditional single node calculation ability cannot meet the needs of all sectors. Using cloud computing, distributing the complex computations which consume a large amount of computing resources to multiple nodes through network is a new effective solution.As shown in figure 3, the system in divided into four levels: view-level, business logic-level, Hive-level, and Hadoop-level.

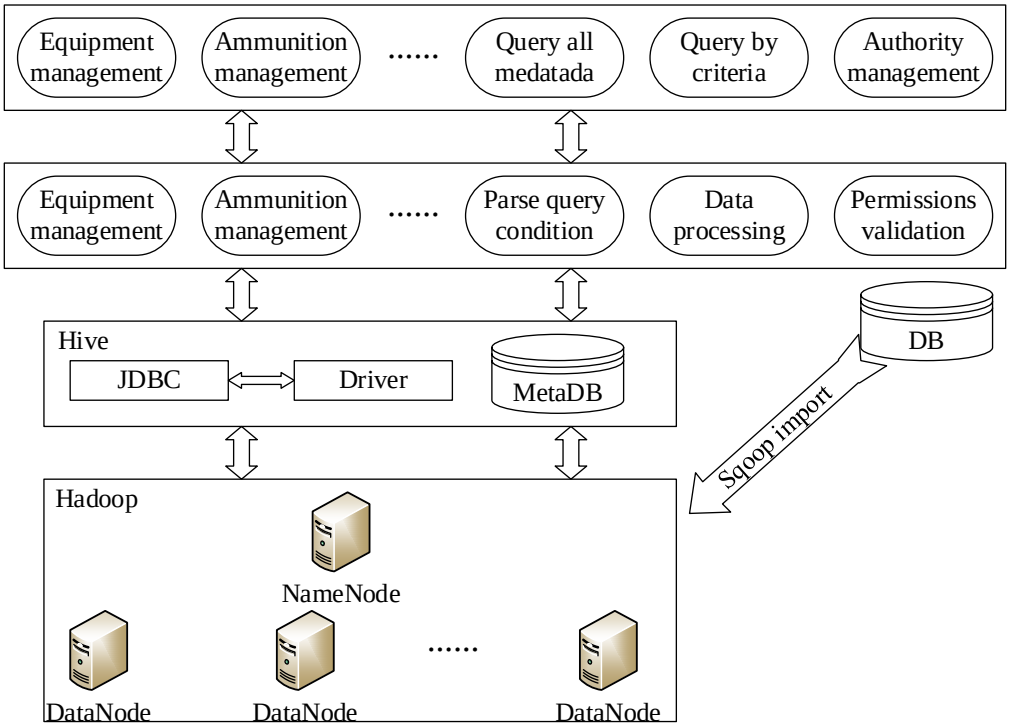

Fig.3 Data query architecture diagram based on Hadoop

View layer: this layer includes equipment management, people management, ammunition management, authority management, log management and so on. After login the system, user can query data by imputing the query conditions.

Business logic layer: this layer includes several modules like equipment management, authority filter, data query, and parse query condition.

Hive layer: this layer accepts query requests from business logic layer, parse and compile user-submitted HiveQL statement through parser and semantic analyzer, then generate MapReduce task based on Hadoop by logic program generator and query program generator. Before system operation, data preparation needs import data from relational database to Hive with the help of Sqoop.

Hadoop layer: Also known as the calculable layer. This layer use MapReduce to carry out Hive layer's computing tasks to read and write different data in HDFS.

Systematic Workflow. First, we use Sqoop to import data into HDFS to establish warehouse model, storage the metadata generated in the process into Hive's Derby metadata database.The completed data warehouse runs as follows: the client requests a data query, query table definition in metadata database according to request contents, if meet the request, the system access to the file directory for the corresponding operation. Finally, system store data query plan in HDFS and return results to client.Work process is shown in figure 4.

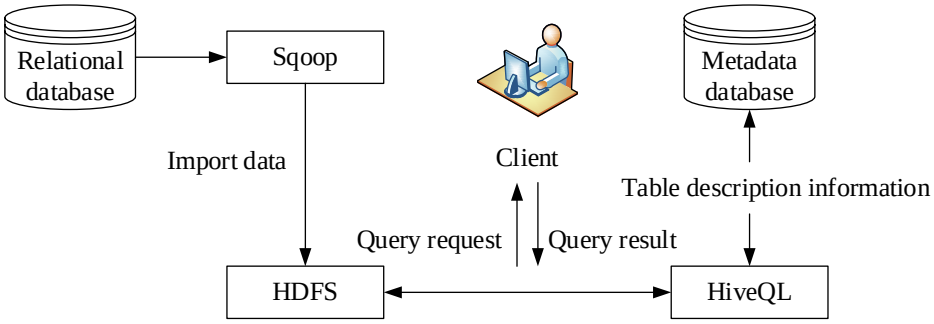

Fig.4 the working process of the query system

Performance Test

The Environment Building And Data Preparation. To test this scheme, this article set up a Hadoop cluster consist of 6 servers, its hardware configuration: processor is Intel Core i5, CPU is 2.4GHz, two gigabytes of memory, 500GB drive; software configuration: OS:Red Hat Linux OS5.0, Hadoop:hadoop-1.0.3, Hive: hive-0.9.0, Sqoop:sqoop-1.4.4.There is one NameNode and five DataNode, Hive and Sqoop is running on NameNode.

If using a cloud platform process the data in relational database, we must import the data into HDFS. The experimental data are shown in table 1.

Table 1 Data size(MB) and number of tuples

data size	61	123	247	614	985	1220
number of tuples	5×10^5	10^6	2×10^6	5×10^6	8×10^6	10^7

Validate Query Performance. This experiment is aimed to give the query time with different amount of data in single environment and cluster environment.

Aggregate query. Query total number of tuples in table project, the SQL is: select count (*) from project. Each group of experiment run for 3 times, record the execution time and calculate the average. Test results as shown in figure 5.

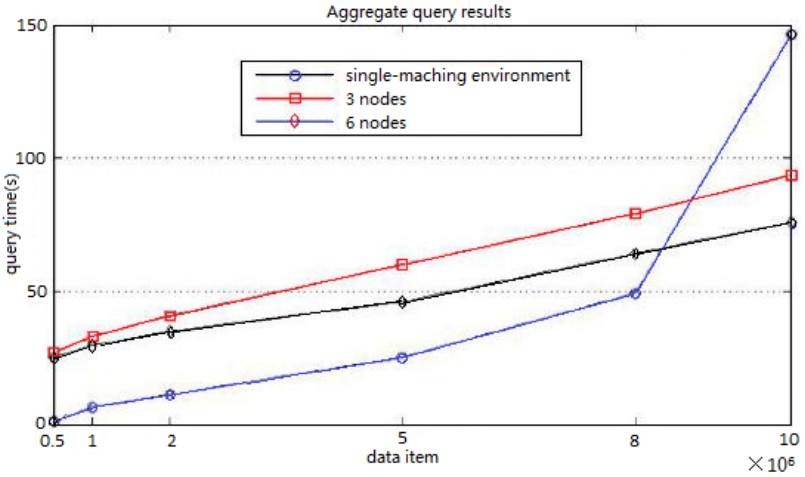

Fig.5 Test results of Aggregate query

Join query. Query two tables' same data in certain fields and sort in descending order by id, the

SQL is: select p.id, p.create_at, t.update_at, t.status from project p join test t on (p.name=t.name and p.description = t.description and p.status=t.status) order by p.id desc. Each group of experiment run for 3 times, record the execution time and calculate the average. Test results as shown in figure 6.

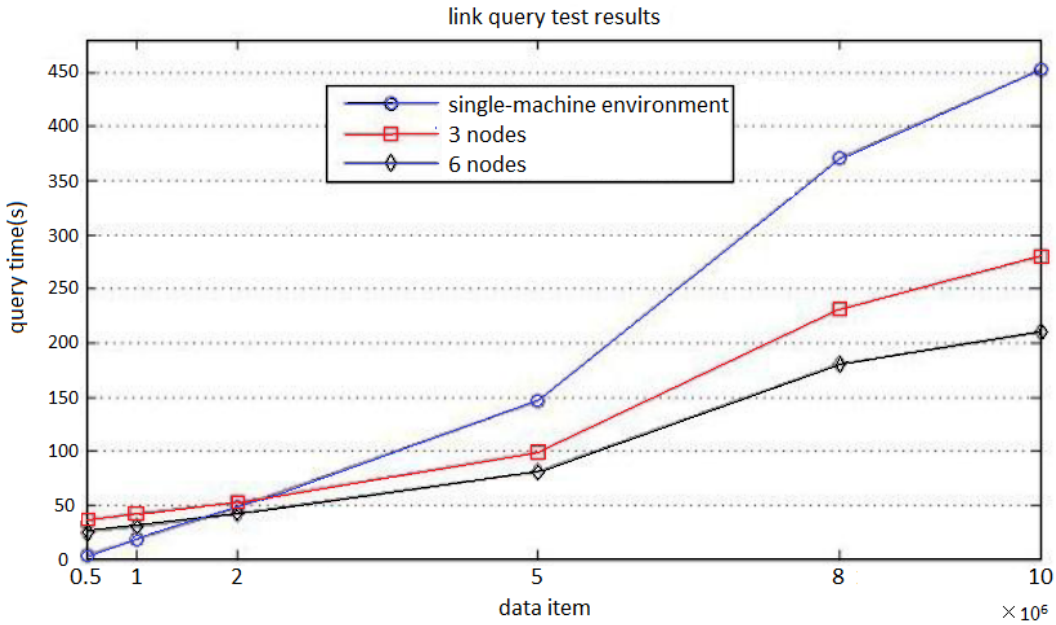

Fig.6 Test results of Joinquery

It can be seen from the diagram when the data volume is small, the relational database query efficiency is higher than Hadoop cluster. But when the data volume reaches a certain degree, query cluster environment advantage begin to emerge.

Conclusion

At present, the relational database to the operation of the small amount of data has been very mature, it usually improve query efficiency through the establishment of the index.But huge amounts of data query is a bottleneck of relational database, it cannot be unlimited expansion and make each property index. So when the query of mass data, the advantage of the cloud storage is obvious. Thus it can be seen that using the idea of cloud computing to solve the problem of insufficient mass data query analysis ability of our equipment management system relational database has extensive application foreground.

References

[1] Konstantin Shvachko, HairongKuang, Sanjay Radia, Robert Chansler. The Hadoop Distributed File System. Sunnyvale, California USA, IEEE 2010:1-10

[2] GaoJichao.Hadoop’s platform storage policy research and optimization [Master’s thesis].Bei Jing: Beijing Jiaotong University.2012.6

[3] Huang Xiaoyun.Cloud storage service system research based on HDFS[Master’s thesis].Da Lian: Dalian Maritime University.2010

[4] Yang Zhiwen.Cloudcomouting technology guide[M].Bei Jing:Chemical Industry Press.2010.10

[5] Ashish T, Joydeep S, Namit J, et al. Hive-A Petabyte Scale Data Warehouse Using Hadoop [C] //Data Engineering(ICDE), 2010 IEEE 26th International:996-1005