Advances in Computer Science Research, volume 78

3rd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018)

Research and Application of Large Data Query Technology Based on NoSQL Database

Yin Xiaoqin[1], Luo Qiqiang[2]

[1] Jiangxi Tellhow Animation Vocational College, Nanchang 330020, China

[2] Nanchang Institute of Science and Technology, Nanchang 330108, China

Keywords: NoSQL; big data; query; RDBMS; panorama

Abstract: NoSQL database breaks the traditional relational model, stores data in a free way, provides a new access interface, and overcomes the shortcomings of traditional RDBMS. NoSQL database can be deployed on cheap hardware, supporting distributed storage, and transparently extending nodes. This paper analyzes the characteristics and applications of various types of NoSQL databases, and focuses on the application of the representative MongoDB in the large data query in the NoSQL database. The paper presents research and application of large data query technology based on NoSQL database.

1. Introduction

With the development of big data, non-relational database has become a hot new field. The development of non-relational database products is very rapid. Today's computer architecture has a huge level of scalability in data storage. NoSQL is also dedicated to changing this situation. At present, Google's BigTable and Amazon's Dynamo use NoSQL database. This article introduces 10 excellent NoSQL databases.

Although the NoSQL buzzword has only been around for just a year, it is undeniable that the second-generation movement has now begun, although early stack code is only an experiment. However, today's systems have become more mature and stable. But there is also a stark fact that technology is getting more sophisticated-so much so that good NoSQL data storage has to be rewritten. There are also a few who think this is the so-called version 2.0. Here are some of the more well-known tools for building a quick, extensible repository for big data [1].

NoSQLt real time historical data cloud, is the independent research and development, independent of all intellectual property platform software products, dedicated to processing field time domain measurement data of industrial production, high compression, storage and supply of massive real-time and history data, supports high-speed statistics, high-speed history trend curve, historical data mining application of high-speed batch playback. The data provide parallel expansion, disaster recovery ability, can help enterprise users to easily build the production process data in the cloud.

NoSQLt real-time historical data cloud has been launched. It has been concerned by the military customers and participated in the construction of a big data center of the meteorological department. It has won the praise of users, and is about to enter the core business layer to continue to serve the users.

The west sea data and petroleum system has very deep origin, several core team founder for the oil industry was born, had long been fighting in the oil line. So in the development process, the west is very concerned about the oil data team needs large data, has spent nearly two years, visited a number of design institutes, system integrators, the touch needs to find the problem, and one by one the benchmarking support in NoSQLt to achieve real-time historical data in the cloud.

Google LevelDB announced that it will open, and comply with New BSD license.LevelDB is an embedded key-value database. Its keys and associated values are arbitrary byte arrays, and in accordance with the key sorting, sorting mechanism can be overloaded. Data storage mechanism is very simple, only supports Put, Get and Delete commands, and then the forward and backward

iterative traversal.

The data is automatically compressed using Snappy, this is a compression library for the BigTable, Google, MapReduce and RPC, and in April announced the open source.LevelDB also has some limitations: does not support the SQL query and index, support multi threading single visit process, and can be used for embedded devices, which will make some benefits of the project, but also to the other project to bring trouble.

LevelDB optimized batch write operation [2]. It will be more changes to a sorted in memory, after accumulated to a preset threshold and then the configuration file is written to the disk. For sequential and random writes, and sequential read operation, its performance is very good, according to the Google performance benchmark, it can some tests ahead of SQLite.SQLite of two orders of magnitude in the random read operation slightly better than the LevelDB, and in the larger write data when the speed is two times faster than LevelDB.LevelDB also performed better than Kyoto Cabinet good, Kyoto Cabinet is a key-value database, but Google did not like SQLite in all test items were compared. Also, Riak carried out some tests comparing LevelDB and InnoDB, in some test project, Google LevelDB can achieve the same performance or better than InnoDB.

Since the last century since 90s, the Internet application has been from the past mainly is the user access to information of the Web1.0 era, the development of more emphasis on user interaction in the Web2.0 era. In the era of Web2.0, the website information provider by the website administrator information into traditional ordinary users, user information is vast. The content from the rigorous application business processes, such as flight booking, stock trading, the development of today's communication, shopping, entertainment, social networking and so on. The amount of data in various fields from the previous TB level to PB level, and continues the explosive growth of Internet applications entering the era of big data. During this period, the computer network and the hardware level has been rapid development.

2. Discussion on NOSQL database

With the rise of the Internet Web2.0 website, non relational database has become an extremely popular in new areas, the development of non relational database products very quickly. But the traditional relational database in Web2.0 website, especially the large scale and dynamic website Web2.0 pure SNS type high concurrency has appeared to be inadequate, exposed a lot to overcome the problem, for example [3].

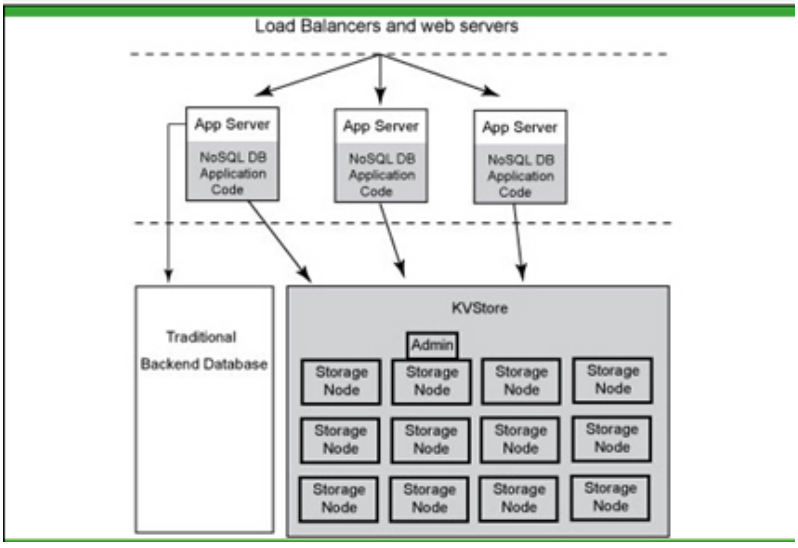

Figure1. NoSQL Database of Oracle

The Web2.0 web site to generate real-time dynamic pages and provides dynamic information according to the user's personalized information, so basically can not use static and dynamic pages, so database concurrent load is very high, often tens of thousands of times per second to read and write

requests. The relational database with tens of thousands of times the SQL query also managed to withstand, but with tens of thousands of times to write SQL the hard disk data request, IO is unbearable. In fact, for ordinary BBS sites, there is also high demand for concurrent write requests, such as JavaEye site real-time statistics for online users, recording popular post clicks, vote counting, so this is a very common demand.

NoSQL(Not Only SQL, which means "not just SQL," refers to non-relational databases. It is a new revolutionary movement for databases, which was proposed early on. By 2009, advocates of .NoSQL were advocating the use of non-relational data storage, as opposed to the overwhelming use of relational databases. This concept is undoubtedly an injection of a new kind of thinking.

The tables in the relational database store some formatted data structures, and each record has the same composition of fields, even if not every record requires all fields. But the database assigns all the fields to each piece of data, and the non-relational database is stored in a key-value pair (key-value), which is structured so that each record can have a different key. Each record can add its own key-value pairs as needed, so that it is not limited to a fixed structure and can reduce some time and space overhead.

What is the NoSQL? The definition of wiki is "NoSQL is a movement promoting a loosely defined class of non-relational data stores that break with a long history of relational databases". In fact, there is not a product called NoSQL, it is the focus of a class of non-relational data stores set.NoSQL is non-relational, while the traditional database is relational [4].

As we all know, the biggest flaw of traditional relational database is extended, although the solutions of each database manufacturers have cluster, but whether it is storage or share nothing share solutions, scalability is very limited. To solve the problem of database scalability has two main ideas: the first is the data sheet (sharding) or zoning, although it could be a good solution to the database scalability problems, but in actual use, once the data partition or partition function, will inevitably lead to the "relational database" at the expense of the biggest advantage of -join is very large, the business limitations, the database can degenerate into a simple storage system. Another idea is through maser-slave replication, through reading and writing separation technology to solve the scalability problem in a certain extent, but this case, because each A database node must save all the data, so that every stored IO subsystem is bound to become the bottleneck of expansion, and masert node is also a bottleneck. In general, the extension ability of traditional relational database is very limited.

NoSQL database breaks the traditional relational model, stores data in a free way, provides a new access interface, and overcomes the shortcomings of traditional RDBMS..NoSQL database can be deployed on cheap hardware, supporting distributed storage, and transparently extending nodes.

Cassandra was originally developed by Facebook, and later became the Apache open source project, it is a social network cloud computing ideal database. It integrates other popular tools such as Solr, has now become a full-fledged large data storage tool.Cassandra is a non relational database of mixed type, similar to Google the main function of the BigTable. than Dynomite (Key-Value distributed storage system) is more abundant, but the main characteristics of support is not to store MongoDB.Cassandra files is that it is not a database, but by a database node constitute a distributed network service, a write operation on the Cassandra, will be copied to the other nodes up to Cassandra and read operation will be routed to a node above to read. In a recent test, Netflix has built a 288 node Cluster.

3. NoSQL-big data, key-value Database with large concurrency

The NoSQLt real time history database is the software product of the national independent intellectual property rights. The west sea data has also been certified by the military. It is the legitimate supplier of military products.

NoSQLt in the process of development, emphasis on product quality, through the harsh pressure test for 18 months, and in the military customers run continuously for more than 24 months. With high stability, each NoSQLt database has its watchdog, support 7*24 hour unattended operation

mode, key event logging operation can meet the strict standards the oil industry field.

The built-in NoSQLt "to rights management strategy group authentication", write data, read, N2N management, synchronization can be accurate to each tag definition, meet user flexible use demand, especially the NoSQLt follow the "standoff security acquisition" principle, the field of unmanned collection site, the data of the least privilege, only can write data acquisition station, hackers can not break through the program, read the other key information into the system.

LevelDB is written using the C++ library dependencies; some external has been successfully transplanted to the Windows Mac OS, X, Android and Unix. In practical application, some experimental version of Chrome LevelDB has been used in the implementation of IndexDB, as API and Riak. It is used in the node level storage. Moreover, the development of a 3D map software company UpNext also use this system [5].

MongoDB is a document oriented database, written by C++ and.2007, in order to solve the problem of a large number of practical application development in the community in October, MongoDB by the 10gen group developed.2009 in February first launched. It is characterized by high performance, ease of deployment, easy to use, storage data is very convenient. The main features a: for collection storage, easy storage object type.

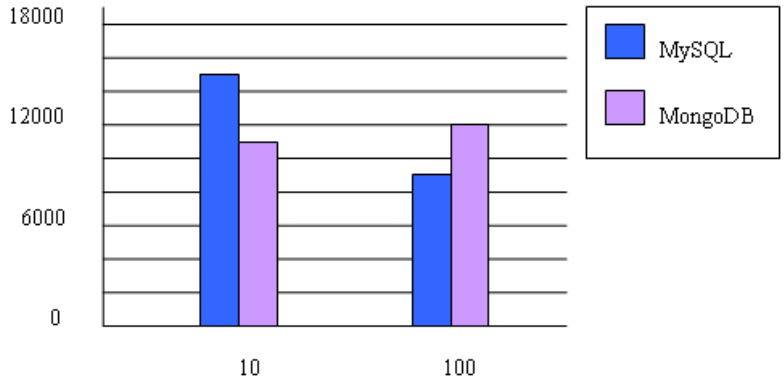

Figure 2. In view of the transformation of Internet applications mysql and mongoDB

In view of the transformation of Internet applications, the application architecture has evolved from centralized processing, vertical extension (Scale-Up) interactive system architecture to distributed modern Web applications. Horizontal extension (Scale-Out) system architecture, which can support more users by adding more Web servers to deal with the analysis and processing of mass data in the big data era. The emergence of cloud computing and distributed big data set processing model MapReducebased on large-scale low-cost computing platform is compared with the rapidly converted application architecture and data processing technology. The technology of relational database has been developing slowly in the past several ten years. Relational database is no longer suitable for more and more Internet applications.

NoSQL is a kind of non-relational database technology which is produced under the need of solving this kind of application requirement. Since NoSQL does not define the organization of data as relational as traditional relational databases, so long as the internal data organization is non-relational. It can be called NoSQL database NoSQL database is not intended to replace the widely used relational database. Since its concept was put forward, more than ten kinds of popular database products have emerged, which are widely used in Internet applications.

Similar to Facebook, twitter, Friendfeed such SNS website, users every day dynamic user generated mass, in the case of Friendfeed, a month has reached 250 million users, for a relational database, SQL query on a 250 million record table, efficiency is extremely low and even unbearable. Then the user login system, such as a large web site such as Tencent, Shanda, hundreds of hundreds of millions of accounts, difficult to deal with relational databases.

Based on the web architecture, the database is most difficult for the lateral extension, when the user visits and an application system of the database as you grow with each passing day, web server

and app server as simply by adding more hardware and service nodes to extend the performance and load capacity. But there is no way for a lot of need provide 24 hours of uninterrupted service website, the database system upgrade and expansion is a very painful thing, often need to shut down for maintenance and data migration.

KV Cache type doesn't have persistent storage function. Memcached is widely used to alleviate the pressure of database, and the function of data persistence storage is replaced by database.

KV store has a storage function, the Sina memcachedb is based on memcached, using Berkley DB as the storage layer developed a distributed KV store.Tokyo Tyrand/Cabinet SNS is the largest social networking site mixi.jp the development of Japan's KV store, where TC is a NoSQL database for persistent data storage, TT is the network interface TC (compatible with memcached protocol). The Berkley DB is an embedded database, now lies in the hands of the Oracle.

Eventually consistent KV store is a KV store final design consistency principle, including Amazon Dynamo, Lindedin Voldemort and Facebook Cassandra, the main feature of the Dynamo is distributed (Center), high availability, scalability, and so on can always write the design idea of.Dynamo is one of the most important theories of distributed system the other one is Bigtable.

4. Research and application of large data query technology based on NoSQL database

First, a huge data volume (Volume). The definition of a single data volume that is brought to exponential growth, for example, a single 1080PIPC30 day will produce 2T data; IP network, interconnection of each platform, the number of city network camera as safe level of tens of thousands of huge amount of data, as can be imagined.

Second, various data types (Variety). The video encoding format in the field of video surveillance including: H.264, MPEG-4, MJPEG and other diverse encoding. And at the same time with the various types of networking technology into the video monitoring service, brought together including a variety of sensors, IT, a variety of data produced by the CT system. The system needs to be structured business with the unstructured data interrelated, unified storage.

Lucene is a subproject of the 4 jakarta project team of the Apache Software Foundation, an open source full-text search engine toolkit. This means that it is not a complete full-text search engine, but the architecture of a full-text retrieval engine [6]. However, most people do not agree that Lucene is a database. Because most people just use it to retrieve a lot of text chunks, but it does use a model similar to other NoSQL data stores. Instead, look for words or fields that appear in a block, and Lucene / Solr is no doubt the best way to query.

NoSQLt fully consider the future of the mobile Internet, complex link, massive data terminal application scenarios, unattended, automatic inspection of oil and gas station, mobile patrol, Beidou geographic information tracking needs to do a lot of work, so as to adapt to the needs of the future. Oil users such as NoSQLt each tag allows the definition of latitude and longitude, in support of GIS applications.

For example, in physical exploration, we can deploy points in grid in a single area for seismic detection. Every measuring point has Beidou coordinate property. Data can be described as 2D or 3D data and image, so as to achieve precise interpretation and analysis.

NoSQLt provides a variety of API, and any node in the data cloud is exactly the same API. The user program is developed and used everywhere. It can improve the development of energy efficiency and reduce the cost.

NoSQLt real time historical data cloud, is by the NoSQLt database as the node, with N2N as the huge data link building cluster, system logic itself without restrictions, industrial storage mode based on tag tags, no coupling between data, the user can arbitrarily parallel expansion, users build cloud size, limited only by the user's needs and budget.

NoSQL non-relational database technology uses loosely coupled data patterns, supports horizontal scaling, and has the ability to persist data on disk and / or memory. Data models that support multiple "Non-SQL" interfaces for data access. NoSQL include Key-Value key-value pairs, document-oriented storage. Column storage and graph structure storage. NoSQL supports complex

query, weak transaction mechanism, redundant backup to ensure the reliability of a single machine, and multiple data synchronization methods to achieve multi-machine reliability. Support for hash partitions and scope partitions for distributed extensions, with emphasis on eventual consistency .

The web system of any large amounts of data association are very taboo multiple large tables, and complex data analysis of complex SQL statements types of queries, especially the SNS type of the site, from the perspective of demand and product design, to avoid the generation of this case. More often only a single table query and simple conditions for single table query page, the function of SQL was greatly weakened.

Therefore, the application of relational database in the scene more and more is not so appropriate, in order to non relational database to solve this kind of problems emerged, now this year, a variety of non relational database, especially the key database (Key-Value Store DB) blustery, more confusing. Not long ago, has just held a foreign NoSQL Conference, all NoSQL databases have appeared, but the reputation of the appearance and not, at least have more than 10 open source NoSQLDB.

5. Summary

NoSQL refers to those non relational database, the definition of the concept is not very clear, the data warehouse. NoSQL database no longer use the relational model, to give up the SQL database operation statement. NoSQL database overcomes the shortcomings of RDBMS, can be deployed on inexpensive hardware support, distributed storage, can transparently extend the typical NoSQL database node. In the form of key-values data storage, has the characteristics of mode of freedom.

Key-values refers to a key corresponding to a key, can be accessed through the key name. For example staff record information such as Figure 1 and Figure 2, Name, Age, Profession keys, each key corresponds to a key.

Free mode refers to the use of database no predefined data model. In traditional RDBMS, if you want to store certain employee information, you must define a staff table, there are in the field and the related staff. If there is a change on demand, data model to increase staff have to revise the original information definition. A model free database has no predefined data model to store data.

References

[1] Dileepa Jayathilake, Charith Sooriaarachchi, Thilok Gunawardena, Buddhika Kulasuriya, Thusitha Dayaratne. A Study into the Capabilities of NoSQL Databases in Handling a Highly Heterogeneous Tree [A]. 2012 IEEE 6th International Conference on Information and Automation for Sustainability, 2012: 106-111.

[2] Lu Ming-yu, Li Xiao-yong. Comparative Analysis of NoSQL Database and Relational Database. [J] .Microcomputer applications, 2011, 2710: 55-58.

[3] Zhu Wei-ping, Chen Huan, Li Ming-xin. Using MongoDB to Implement Textbook Management System instead of MySQL. Proceedings of 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN 2011) VOL02, 2011.

[4] Li Lisha. On NOSQL's thinking on [J]. Technology frontiers, 2010 (4): 40-41.

[5] Wang Shan, sa Shixuan. Introduction to database system. [M]. Beijing: higher Education Press, 1983.15-18.

[6] Lu Yi NoSQL data management system overview [J]. Enterprise technology and development, 2011 (17): 31-33.