BDA 4-3

1

Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering

School of Computer Science, in Beijing Institute of Technology,

in this session, we discuss about what is NoSQL DB.

2

The big data computing system can be summarized into three categories:

Data storing system, Data processing system, Data application system

The data storage architecture is the foundation of big data computing.

In data storing system, there are 4 parts to accomplish different tasks,

which is Data collection and modeling, **Distributed file system, Distributed database/data warehouse and Unified Data Access Interface.**

**we learned data collection and modeling, distributed file system, now we look at database.**

**Database normally include relational database like Oracle, SQL server, and NoSQL Database, like MongoDB, Hbase etc.**

3

The distributed database and data warehouse built on the file system not only realize the storage management of data, but more importantly, it supports the fast data query function and data analysis service of the upper computing engine and application software.

Currently, there are many types of database products that support big data applications, and their storage structures and technologies used are also different.

Relational database RDBMS has always been the mainstream of data storage management system with its good data consistency, complete transaction mechanism, standardized query language, high query efficiency, and mature technology.

Its representative products include Oracle, MS SQL Server and My SQL. When faced with big data computing problems with large amounts of data and large amounts of unstructured or semi-structured data, relational databases face the following challenges:

1） Large-scale (PB-level) storage management of big data requires that the system has good flexibility and can be easily expanded in a distributed environment.

However, the traditional relational database emphasizes data consistency and integrity and centralized deployment methods. Its scalability is poor, and it is difficult to adapt to the explosive growth of data

2） RDBMS is suitable for storage management of structured data based on strictly defined key indexes and data table storage modes, and can provide an efficient SQL-based query mechanism. But for unstructured or semi-structured data, RDMS is difficult to process, and query efficiency is greatly reduced

3） Big data calculation processing requires the storage structure to well support the upper-level calculation model.

For example, the Mapreduce calculation model adopts the Divide-and-Conquer strategy, which first divides a large data set into multiple subsets, and then runs the map program for each subset for processing. After the processing of the subset is completed, the reduce program is run to complete the summary of the calculation results, which requires the underlying data storage structure to support the division and fusion of this data set (or data table structure).

Due to strict data consistency and completeness requirements of relational databases, it is difficult to perform such segmentation processing on data tables. so, it is difficult to support various calculation models for big data calculations

In recent years, non-relational databases based on distributed file systems that have emerged with big data processing (not only SQL, NoSQL)

4 Now let’s watch a video about “How do NoSQL databases work”.

播放视频的 0:00-6:27！！！

5 from the video we understand How do NoSQL databases work; and advantages, disadvantages relational databases and NoSQL Database.

non-relational databases based on distributed file systems have good scalability, can effectively process unorganized and semi-structured data, and support high concurrency

The characteristics of the calculation model have been widely used in the Internet, medical and health, e-commerce and other fields.

Compared with traditional relational databases, this type of NoSQL database has the following Common structure characters

1） No pre-defined data format is required

We don’t need to define a strict data table structure in advance. Each record of data may have different attributes and formats. When inserting data, there is no need to define their format in advance.

2） Shared nothing architecture

NoSQL databases often divide data sets and store them on various local servers.

The performance of reading data from local disks is often better than that of reading data through network transmission, thereby improving the system's read and write speed.

(A shared-nothing architecture (SN) is a distributed computing architecture in which each update request is satisfied by a single node (processor/memory/storage unit) in a computer cluster.

The intent is to eliminate resource contention among nodes.

3 ）Flexible and scalable

Nodes can be dynamically added or deleted while the system is running without downtime for maintenance, and data can be automatically migrated.

4) Data partition

In contrast to storing data on the same node, NoSQL databases partition data and distribute records across multiple nodes.

In addition, replication is usually performed at the same time as the partition, which not only improves parallel performance, but also ensures that there is no single point of failure.

5 ) Asynchronous replication

NoSQL replication is often a log-based Asynchronous replication, so that data can be written to a node as soon as possible, rather than being delayed by network transmission.

The disadvantage is that consistency is not always guaranteed. NoSQL database provides ultimate consistency.

6

In this session, we learned what is NoSQL DB.

thank you for your attention, if you have any question, feel free to contact me.