Goel, Ajay: Nu Id (001897443)

INFO 7250
Engineering Big-Data Systems
Summer Full 2019
Assignment 1

Application of NoSQL - Web Crawling

Web crawling is to filter and collect various information and resource from Internet, according to a certain theme, storage the information into the database, and construct the search engine for users.

- The huge amount of information, the performance of search engine is mostly affected by storage form and the storage database.
- Web crawling requests great capacity of storage space and low hardware costs. Its requirements of consistency and integrity can be reduced, but need high availability, scalability and performance.
- The relational databases store data in the form of a two-dimensional table with strictly row and column format, and emphasize the consistency and integrity of data, the performance in the distributed data management has low efficient, resulting in poor horizontal scalability and high hardware costs when increasing the data storage.
- With the increasing resources of information, the traditional relational database can no longer suffice requirement to query information. Several famous search engines implement their own databases, such as Google's Big Table.

Principles of Web Crawling:

- Web crawling is a process that automatically obtains information from the Web pages through the link relationships between them and expands to the entire Web.
- The spider crawls the pages from the Internet, extracts the URLs to URL database, and saves the pages to the original page library. As crawling a Web page usually takes seconds time to wait network communication, multiple spiders will be started to parallel process so as to improve the crawl rate.
- The controller judges and distributes the URLs from the URL database to control the spider crawling the
- other pages until the URL database is empty.

NoSQL Database:

- DB is a key/value database for the embedded occasion which requires a high speed of inserting, reading and writing with a relatively simple data type.
- NoSQL is formed by reducing the data consistency and integrity constraints, in exchange for high availability and partition tolerance to meet the requirement of Internet application with huge data amount storage, high performance and low-cost scalability.

MongoDB:

- MongoDB uses BSON with loose data structure as the data format and document-oriented storage. It provides auto-sharding to achieve mass data storage and supports full indexes.
- With the powerful query language syntax similar to the object-oriented language, MongoDB can achieve the most function of single-table query in relational database. It also supports atomic in-place update and two replication mechanisms of Master/Slave and replica set.
- MongoDB both has the high performance and scalability of key-value data store and rich data processing functions of traditional relational database

Comparison of Solutions:

Compare the solution between relational database and NoSQL database:

- In MongoDB, post is with all the floors in one document because MongoDB supports schema-free, there is no design that structure beforehand, and it can be modified at run time. The fields in each document do not need to be same, which can be set depending on the actual situation when programming. How many floors there are, and how many fields we can add.

- In relational database, we create two tables, that one stores the post, and the other stores the floors with a foreign key to join the post, which ensures the data consistency.
- Those two data structures can both store the data appropriately but the relational database stores in two tables with a foreign key which is a little complicated.
- MongoDB supports embedded document to implement nested, so we can store a post with all the floors in one document that we can efficiently get the whole post by query the id.
- Relational database stores the post and the floors in two tables, with the post id we query these two tables and get the post and all the floors.
- The amount of the posts and the floors are enormous. Relational database has poor query performance in this large size of table. MongoDB is pretty good at the query with huge amount of data, according to the official documents, when the amount of data is more than 50GB, the access speed of MongoDB is 10 times faster than MySQL.

MySQL	MongoDB
It is a type of relational database which has both DDL and DML.	It is a type of non-relational database which has not having DDL.
Data is stored in tabular form also known as fields and records.	Data is stored as a Document in the pair of key-values.
Development is complex as MySQL needs complex object-relational mapping (ORM) layer that translates objects in code to relational tables	Development is simplified as MongoDB documents map naturally to modern, object-oriented programming languages.

Conclusion:

- Web crawling is one of the most important applications of the Internet, whose select of database storage directly affects the performance of search engines.
- The data storage structure and query are designed and the advantage and the disadvantage are listed in data structure, query and scalability.
- Relational database has multiple tables' storage with foreign key, sharp decline in query performance with huge amount of data, and vertical scalability with high cost.
- Compared to relational database, MongoDB supports schema-free, has great query performance with huge amount of data and provides easy horizontal scalability with low cost of hardware. It is more suitable for data storage in Web crawling.