**HW1-part2 Summary1**

**Application of NoSQL Database in Web Crawling**

This article introduce two different solution, relational database (SQL) and Non-relational database (NoSQL, like MongoDB), for web crawling process in Meteorological BBS information collection system. Comparison are listed in data structure, query performance and scalability.

**Relational Database Solution**

SQL databases store data in the form of a two-dimensional table with strictly row and column format, emphasizing the **consistency** and **integrity** of data.

**MongoDB Solution**

MongoDB use **schema-free** , **document-oriented** storage which is convenient to design. Often support huge data storage, using distributed file system GridFS, split large file into multiple small chunk.

**Comparison & Conclusion**

|  |  |  |
| --- | --- | --- |
|  | MongoDB | SQL |
| Data Structure | Schema-free(Each document fields can be different)  Design conveniently  Modified easily | Data consistency(foreign key join)  Deign complicated |
| Query | Fast query performance  Nested document all in one file, efficient to query | Poor query performance |
| Scalability | Easy horizontal scalability  Low cost of hardware  (Auto-sharding, Automatic failover) | Poor horizontal scalability  High cost (buy server) |

Web crawling collection system need storage huge amounts of data and great query performance. NoSQL is schema-free with easy horizontal scalability, low cost which is more suitable for it.