An introduction and analysis of time series database

--Why do we need a time-series database

**1 INTRODUCTION**

**1.1 Time-series data**

Time-series data is a series of data points indexed in time order. With the development of Big Data Era, time-series data become more and more important and are more widely used in society. In our daily life, the location information is time series data, while the health data generated when we use wearable detection equipment is time series data too. Besides, no matter the frontier technologies, such as self-driving technology and smart home, or economic market , such as the rise and fall of stocks and the trends of funds, all produce time-series data. Additionally, The data collected and generated by real time monitoring, inspection and analysis equipment in large industries, such as the power industry and chemical industry, are all time-series data.

Time-Series data has several characteristics. First, time-series data are time stamped data and are structured data. New data are constantly generated as time goes by, and old data do not need to be kept for too long. Secondly, time-series data are basically inserted without updating and delete requirements, and there are more write than read operations. Moreover, Real time data computing is desired and query for the time-series data is always in time and space range. Last but not least, the source of data is like a stream and the data rate is pretty stable. So the amount of time-series data is usually quite large. Take industrial time-series data for example. The characteristics of the data is: fast frequency, each monitoring point can produce multiple data in one second. And it's very sensitive to the time of gathering information. One data corresponds to a unique time. Besides, the real time monitoring system has tens of thousands of monitoring sites that produce data every second, generating tens of gigabytes of data per day.

**1.2 Time-Series database**

A time-series database is a database that processes time series data. To be specific, Time series database needs to realize the management of time series data, including storage, reading and query. The amount of the time-series data is super large, because the data is produced swiftly. So the time series database needs to write and read the data quickly. In addition, the time series databases usually have many query requirements for time series data: First, the user could obtain the latest status, and query the latest data, such as the latest status of sensors. Second, the database should display interval statistics. The user specifies a time range for the database to query statistics, such as average, maximum, minimum, count, and so on. Besides, obtain abnormal data. The database can filter abnormal data according to specified conditions.

Time-Series data is very helpful nowadays. Analyzing these data through the time-series database can help us do forecasting and real time alert. Calculating and processing these data can also help us have a insight of operation and identify the problems. Given these useful functions, the time-series database are becoming more and more important.

**2 Why use time-series database**

The storage and processing of time series data are often processed by using relational database at present. However, based on the characteristics of time series data, the relational database can not meet the effective storage and processing of time series data. Because of the inherent disadvantages of relational database, it is unable to efficiently store and query data. So there is an urgent need for a database system specially optimized for time series data, that is, time series database. Time series database can efficiently store and rapidly process massive sequential big data by using special storage methods, which is an important technology to solve massive data processing. This technology adopts special data storage mode, which greatly improves the processing capacity of time related data. Compared with relational database, its storage space is halved, and the query speed is greatly improved. The superior query performance of the time series function is much better than that of the relational database, which is very suitable for the analysis application of the Internet of things.

Further, we can compare time series databases with relational databases. The first is storage. For relational databases, it is difficult to store time series data. Although for small-scale time series data, it can be processed as ordinary data with time label. However, time series data is usually generated by millions or even tens of millions of terminal devices, and the concurrency of writing is relatively high, which is a massive data scenario.

The following problems also occur when relational databases process time series data. To start with, it will have high storage cost. Relational database have poor compression of sequential data and requires a large amount of machine resources. Secondly, high maintenance cost. For single-machine system, you need to manually divide the database and table in the upper layer, so there is high maintenance cost. Besides, low write throughput. The single write throughput is low, which is difficult to meet the write pressure of tens of millions of sequential data. In addition, poor query performance. Applicable to transaction processing, poor performance of mass data aggregation analysis.

To sum up, for relational databases, it is difficult to solve the storage and reading problems of time series data. The amount of time series data is too large and the corresponding cost is also too large. So there is a need for time series database.

The invention of time-series database products is to solve the shortcomings and defects of the traditional relational database in the storage and analysis of time-series data, and these products are uniformly classified as time-series database. The process of writing, storing and querying is optimized according to the characteristics of time series data. These optimizations are closely related to the characteristics of time series data. First and foremost, the storage cost is reduced. You can reduce the accuracy to aggregate historical data and save storage space. Second, realize high concurrent write. If we batch write data, then we could reduce network overhead. Data is first written to the memory and then periodically dumped as immutable files. Third, reduce the query delay, achieve high query concurrency. The time series database could optimize the common query mode, through the index and other technologies to reduce the query delay. And it can improve query concurrency through caching and routing.

In foreign financial industry, time series database has been used for a long time. Many companies on Wall Street began to use time series database to solve problems in the financial industry before 2000. The domestic financial sector, especially brokerage, private equity and public offering industries, has only started to use time series databases in recent years.

In the Internet of Things industry, time series database has been used for less than 10 years globally, and it has been used for almost 3 to 5 years in China. Most of them are still in the first stage: data collection, data query, monitoring and other simple tasks are completed through extensive use of stand-alone time series database. It is expected that in the next 5 years, sequential database will usher in explosive growth.

In the Internet of Things, as data volumes explode, more and more scenarios exceed 500,000 or even 10 million measurement points per second. For this level of data, the stand-alone sequential database can no longer meet the business requirements. At the same time, in recent years, distributed time series database technology has become more and more mature and perfect, and has entered the stage of large-scale application. Although the application of time series database in Internet of Things is still in a relatively shallow stage, enterprises gradually realize the value of data, and more and more enterprises hope to mine more valuable information by using time series database in the future.

**3 Summary**

In the Big Data Era, the amount of data generated by the Internet of Things is thousands or even tens of thousands of times more than the traditional information technology, and it is real-time collection, high frequency, high density, and dynamic data model can be changed at any time. Time series data is important in many industries. With time series data, we can do real time alert, forecast, have insights of operation, and identify the problems.

Traditional relational database is not suitable for processing time series data due to the large amount and fast generation of time series data and is not enough to store, query and analyze these data, so it urgently needs a database system specially optimized for time series data, that is, time series database. According to the characteristics of time series data, the process of writing, storage and query is optimized in time series database. Time series database reduces storage cost, realizes high concurrent write, reduces query delay, and optimizes common query mode. Therefore, the database can efficiently store and quickly process massive time series big data, and the query speed is greatly improved. The query performance of the time series function is much better than that of the relational database, which is more suitable for the analysis application of the Internet of things.