Why do we need a time series database?

Before we start discussing this issue, we might as well throw out a few questions. What type of data is a time series database used to store? Did time series databases appear in the last few years? Why was the time series database born? What is a time series database? After clarifying these questions, I think: "Why do we need a time series database", the answer is self-evident.

First of all, what type of data is a time series database used to store? What are the characteristics of this type of data? Let's see this example first: imagine the location of a car, other attributes of a particular car over a period of time, including model, color, license plate number, owner, etc., are unchanged, but its location data is As time changes, a series of data composed of position values and other attributes determined according to time is a set of data that needs to be stored in the time series database. When we drive the car to start navigation, we need to The data determines the route to the destination next and stores the driving record. It can be speculated that this is even more essential in the upcoming driverless driving. If we think about it carefully, we will find that the position coordinates here change with time, so the time here is not just a metric, but a main axis of coordinates. That's time series data, and it's gradually playing a bigger role in our world. Think again, what are the characteristics of this kind of data when operating the database? Yes, time series datasets track changes throughout the system and keep inserting new data instead of updating old data. As can be seen from the aforementioned example of car positioning, there is a big difference between time series data and relational data: First, the most obvious feature is that time series data all have unique timestamps, and they are sorted by the size of the timestamp, and the timestamp is used as the Unique identifiers are used to distinguish, while relational data usually has other fields as identifiers. For example, student data is usually distinguished by student ID as a unique identifier. The second is that time series data does not care about relationships. In car positioning, we do not need to know other attributes of the car's owner, such as age, occupation, etc., and there is no association to the car owner's table. Third, the amount of time-series data continues to grow linearly, and new data will be generated at regular intervals, and massive amounts of data will continue to be generated, so the amount of data is huge. The growth of relational data usually does not grow continuously over time. For example, the amount of student data in a school is relatively stable over a period of time. Fourth, the time series data rarely has an update operation, and the measurement value generation at a certain moment will not change, so it is almost unnecessary to update the time series data. For relational data, existing data is frequently updated, such as personal information of students, including attributes such as age and height.

So, did time series databases appear in recent years? Why was the time series database born? Although the time series database has only entered the public field in recent years, its development can be traced back to the 1990s, which created the demand for time series data storage in the monitoring field.) and Whisper, represented by a fixed-size database, can quickly store numerical data over time, but its read performance is still relatively weak, lacks special optimization for time, and processes a single data model, usually embedded in monitoring in the system. It can be seen that time series data also existed in the past, but there are three reasons for the generation of time series database: one is that the current data scale is huge, the amount of equipment is large, and the amount of data generated is so large that the original relational database cannot meet such a large amount of data. The number of concurrency; the second is that traditional database row storage cannot do lossless compression for a large amount of time series data, and the problem of data storage cost is prominent. Third, time series databases are highly available. Time series databases usually include some common functions and operations for time series data analysis: data retention strategies, continuous query, flexible time aggregation, etc. Even if scale isn't a consideration right now (for example, you're just starting to collect data), these features can still provide a better user experience and make your life easier.

What exactly is a time series database? In my opinion, time series database is a database that emerged as the times require to cope with the rapidly growing application requirements of time series data and the characteristics that are different from traditional relational data. It has the following characteristics: (1) High-speed data writing capability with high throughput. Since time-series services will continue to generate massive amounts of data, and have high requirements for the speed of writing, the concurrent amount of writing is large, which requires the time-series database system to implement high-throughput data writing at high speed. (2) High compression ratio. The time series database needs to store a large amount of data, and some monitoring data may need to be stored for a long time, from 5 to 10 years, so the data needs to be compressed according to the characteristics of the time series data. (3) Efficient time window query capability. The query requirements of time series services are divided into two categories: one is real-time data query, which reflects the status of the current monitoring object; the other is mainly to query historical data of a certain period of time. The amount of historical data is very large, so it is necessary to target the time window Optimized for large data queries. (4) Efficient aggregation ability. Time series business scenarios usually care about the aggregated values of data, such as count, mean and other aggregated values to reflect the data situation in a certain time period, so time series databases need to provide efficient aggregation functions. (5) Batch deletion capability. Time series services need to perform batch deletion operations for expired data. (6) Usually do not need to have the ability of transactions. Time-series databases are different from traditional relational databases. Traditional relational databases focus on adding, deleting, modifying, querying and transaction functions, while time-series databases are written for massive data, and their read queries are mostly data within a period of time. Traditional relational databases all use the B tree, which is a random read and write mode, which consumes a lot of time in seeking. For more than 90% of the scenarios written in time series databases, the efficiency is too low. Therefore, the mainstream of time series databases is Use LSM Tree (Log-Structured Merge Tree) to replace B Tree, such as KairosDB (the bottom layer uses Cassandra's stand-alone mode), openTSDB (the bottom layer uses Hbase), LevelDB, etc. The core idea is to give up the ability to partially read in exchange for maximizing the ability to write .

Finally, everyone must understand why we need a time series database. I will make a summary here: one is due to the characteristics of time series data, and the other is because of the rapidly growing application requirements of time series data, we urgently need such a database that can efficiently compress time series data and have high availability at the same time , which is the time series database. There is no doubt that the time series database is in a stage of rapid development, and the time series data technology is gradually maturing. It is here, and it is time to use it.