**We do need time-series database**

**Ganglin Liu**

**2019141240107**

In our information-hungry world, access to detailed, feature-rich time series data has become one of the most valuable commodities. Businesses, governments, schools, and communities, large and small, are looking for ways to extract value from analyzing time series data.

The world is changing at a rapid pace, and collecting and analyzing more and more data has become a challenge for industries such as IT. Self-driving cars need to constantly collect data on changes in their surroundings and adjust it to weather conditions, potholes, and countless other variables. Automated trading algorithms need to constantly collect data about market changes in order to optimize short-term and long-term return on investment. Smart homes need to regulate temperature, identify hazards, etc., and handle human-computer interaction by monitoring indoor conditions. The retail industry accurately and efficiently monitors the transportation of every item. Compared to stock market trends, self-driving cars, and accurate predicting when you'll make your next online purchase, there have been many examples in recent years of how time series data collection and analysis can impact an individual's daily lives.

All of the above have a common feature, that is, the collection of a variety of data over time, time series databases are therefore needed and receive a lot of attention. Let's first talk about the related concepts of time series databases: time series data is a series of data that is continuously generated over time, in short, it is data with time stamps. Time series datasets have 3 things in common: the data written is almost new; Data is usually written in chronological order; Time is a spindle, and time intervals can be regular or irregular. A time series database is a database that ingests, processes, and stores timestamp data. Such data may include metrics from servers and applications, readings from IoT sensors, user interactions on websites or applications, or trading activity on financial markets.

Time Series database is different from Mysql and other databases that we commonly use, each data point of time series data contains timestamps for indexing, aggregation and sampling, data can be multidimensional and correlated, time series database is characterized by writing more and reading less, It is necessary to support high-frequency writes in seconds and milliseconds or even nanoseconds, and the queries of data are usually multidimensional aggregate queries, and the latency requirements for queries are relatively high. The problem of time series database is how to quickly and efficiently store a large amount of time data, due to the difference in time, the data measured by the same instrument is also different, the data is stored over time, and the data can also be queried and visualized according to time , which gives us a better understanding of how data changes over time.

Of course, ordinary databases can also store time data, but, you know, if with the increase in time sampling rate, the data to be stored will increase, for so much data, the use of traditional database storage and query methods is not applicable, because it will consume a lot of space and query time 。 For example, traditional database storage uses B tree, for random writing B tree will consume a lot of time on disk seek, resulting in slow speed, for every moment is writing data in the time series database, B tree is obviously inappropriate, Instead, it adopts technologies such as LSM trees, in addition, time series databases need to adopt new technologies in distributed storage and data sharding. Therefore, the design and organization of time series databases and traditional databases are very different, and the demand for time series data collection and query in various industries has increased greatly, so time series databases are urgently needed in current society.

You know, time series data is everywhere, and the primary purpose of developing time series databases is to store and analyze massive amounts of time series data. Time series data, that is, time series data, data that is recorded and indexed in chronological order. Various types of devices such as smart cities, The Internet of Things, the Internet of Vehicles, and the Industrial Internet generate massive amounts of time series data. Massive amounts of data can cause a range of problems, whether it's storage or fast queries, which is why people are more inclined than ever to use time series databases. The world demands faster and better data-driven decisions. Traditional static data cannot solve this problem. To meet industry demands, data needs to be collected with the highest possible fidelity – that's what time series data provides: everything that happens in a system can be stored like a movie, whether it's software, a physical power plant, a game, or a customer in an application.

In the past, our view of time series data was static, with daily temperatures, opening and closing prices of the stock market, and even daily or cumulative hospitalizations due to COVID-19. However, we tend to overlook the subtle differences caused by potential changes in these static data. Let's look at some examples:

Average temperatures in one location for several consecutive days. Over the past few decades, average temperatures have been used as a major reference for energy efficiency in buildings. The average daily temperature in the same location may be only slightly different during any given week, but at the same time, factors affecting the environment can change dramatically. Conversely, understanding the temperature changes at each hour of the day, combined with precipitation, cloud cover, and wind speed during that time, can greatly improve the ability of property modeling and optimizing energy efficiency.

The analysis of time series data can also be used in medical care to track various aspects of patient data over time, such as: patient age, admission or discharge, number of days of recovery, etc., helping us understand how to derive daily statistics, allowing us to better analyze trends, accurately report totals, and take action that can even affect government policies.

The analysis of time series data is crucial for every IT group, and the results of the analysis can become an important indicator of the operation of servers, networks, applications, environments, etc. This time series measurement data is critical to guaranteeing service reliability. By tracking changes in each metric, IT can quickly identify problems, plan for upcoming events, and diagnose whether application updates have resulted in changes in user behavior, for the better or for the worse.

The above example shows that modern time series data is very different from what we used to know. Time series data analysis is much deeper than pie charts or Excel. This data doesn't just use time as a measure, the key is to help us analyze data and obtain valuable information. Therefore, time series databases have great value and can help us solve problems that traditional databases cannot solve in time series data analysis.

Compared to generic databases, there are two reasons why time series databases are the most needed databases today: scale and availability.

For scale, time series data accumulates very quickly, relational databases perform poorly on very large datasets, and while relational databases that fine-tune time series data can perform better, in contrast, the benefits introduced by time series databases are only possible if time is considered the first consideration. These benefits enable them to provide large-scale performance improvements, including higher throughput and faster large-scale queries, as well as better data compression.

For availability, time series databases typically also include built-in functions and operations commonly used in time series data analysis, such as data retention policies, continuous queries, flexible time aggregation, and so on. Even when you're just starting to collect this type of data, and you don't need to think about scale right now, these features can still provide a better user experience, making data analysis tasks easier. Use built-in functions and features to analyze readily available trends in your data layer and often uncover unexpected value, regardless of whether your dataset is large or small.

This is why developers are increasingly adopting timing databases and applying them to a variety of scenarios: public safety: Internet records, call logs, individual tracking, interval screening; Power industry: centralized monitoring of smart meters, power grids, and power generation equipment; Internet: server/application monitoring, user access logs, ad click logs; Internet of Things: elevators, boilers, machinery, water meters and other networked equipment; Transportation industry: real-time road conditions, intersection traffic monitoring, bayonet data; Financial industry: transaction records, access records, ATM, POS machine monitoring.

In short, the world economy is developing at a high speed, the demand for time series data is also greatly improved, and retaining the inherent time series characteristics of data can allow us to retain very valuable information, such as understanding how data changes over time. Of course, storing data in this way also has an obvious problem: you end up with a huge amount of data, and it grows very fast. So here's the problem: being able to analyze increased time series data is more valuable than ever, but it's piling up very fast. The advent of time series databases has given us a solution to this problem, and time series databases promote the development of science and technology, which is a type of database that we need very much.