**Why do we need a time-series database?**

**1. Introduction**

With the development of the Internet of Things, the amount of time series data has also shown explosive growth. Time series data (referred to as time series data) refers to a sequence of data points collected at fixed time intervals over a period of time, and each data point is associated with a timestamp to indicate the time when the data was generated. By analyzing time series data within a certain period of time, you can grasp the change trend of the observed things during this period, so as to achieve prediction and early warning of the future.

In the time series data scenario, new data needs to be inserted frequently, the high-precision data retention time is short, and the summary data retention time of medium or lower precision is longer, which means that the data must be continuously deleted from the database, which is different from the normal database design to handle the workload, the database itself must continuously calculate the summary from the high-precision data for long-term storage, which makes the relational database unable to meet the needs of time series data in terms of storage and analysis, etc. Hence the time series database (or time series database for short). The time series database is optimized for the storage and query scenarios of time series data to meet the needs of efficient storage and rapid processing of massive time series data. Compared to relational databases, the advantages of a time series database are not only that the overhead of storage space is greatly reduced, but also that query latency is reduced. At the same time, the time series database also contains some functions and operations for time series data analysis: flexible time aggregation methods, data retention policies, multi-dimensional time range queries, and so on. To elaborate on the necessity of a time series database, let's first introduce time series data.

**2 time series data**

**2.1 Time Series Data Concepts**

Time series data refers to columns of data recorded in chronological order for the same unified metric. Modeling time series data consists of three important parts: subject, time point, and measurement. Time series data is ubiquitous in daily life, for example, in the stock market, the stock price of a stock is a type of time series data, which records the stock price of the stock at each point in time; In the field of operations, monitoring data is a type of time series data, for example, the monitoring data of the machine CPU, that is, the CPU utilization on the machine is recorded at each point in time. In the era of the Internet of Things, time series data will be generated at every moment, and the mining and utilization of time series data is silently changing people's lifestyles. The use of wearable devices for the management of personal health is to continuously collect personal health data through the device, such as heartbeat, body temperature, etc., and evaluate human health according to the collected data calculations. The Internet of Things is also about building a network that allows all objects to produce data and mine its value, and the data collected through this network is typical time series data. In the field of the Internet of Things, the use of time series data can help analysts understand the historical trend of an observation object, and the analysis of time series data is to explore the laws behind its surface phenomena.

**2.2 Characteristics of time series data**

The characteristics of time series data can be divided into three aspects: data writing, querying, and storage, and understanding the characteristics of the data can be used in a targeted manner. These are explained separately below.

(1) Time series data writing characteristics:

1). Stable, continuous, high concurrency. The writing of time series data is relatively smooth, and the amount of data in ordinary applications is generally proportional to the QPS (requests per second) requested. For time series data, QPS is stable, that is, the corresponding time series data is received at each fixed time interval, and the speed of data generation is relatively stable and continuous. In IoT scenarios, the number of subjects being measured is usually very large, and the write concurrency and throughput of time series data will increase accordingly.

2). Write more and read less: Most of the operations on time series data are write operations, in contrast, there are fewer read operations on data, and in actual application scenarios, it is usually necessary to pay attention to some specific indicators, and rarely comprehensively query time series data.

3). Write only more recent time data: time series data is continuously generated with the passage of time, and the time series data received by the common database is new data recently, and at the same time, the old data is rarely updated.

(2) Time series data query characteristics:

1). Query by time range: Time series data in a certain time range is more worthy of attention than a specific data point. At query time, data is specified for a period of time and returned in chronological order.

2). The probability of recent data being read is relatively high: usually time series data is used to record the state of the measurement subject, and the more recent the data, the higher the probability of being concerned, and it is rarely read for historical data.

3). Multi-precision query: The query of time series data can be divided into multiple different precisions, usually the higher the degree of restoration of high-precision time series data to the history, but the number of data points will be very large, and the requirements for storage cost will also increase. However, the query of time series data does not always require high-precision time series data, too high accuracy will occupy more network transmission bandwidth, and does not use observation, by reducing the accuracy for storage, you can store less time series data and save storage space.

4). Multidimensional analysis: Time series data is generated from different measurement subjects, which have different attributes, which may be of the same dimension or different dimensions. For the query of time series data, it can be divided into different dimensions to meet the multi-angle analysis needs of time series data.

(3) Timing data storage characteristics:

1). Large amount of data: In the IoT scenario, the data scale can be terabytes or even petabytes.

2). Hot and cold are clear: the more historical data, the lower the probability of being queried and analyzed.

3). Timeliness: Time series data usually has a validity period, and data that exceeds the age period will be cleaned up and recycled. On the one hand, the more historical time series data, the lower its value, and on the other hand, the recovery of time series data helps to reduce space overhead.

**3. Time Series Database**

**3.1 Time Series Database Features**

Time series database is specially used to store and process time series data database, with the advent of the Internet of Things era, the amount of time series data shows explosive growth, traditional relational database has been difficult to meet the needs of data in storage, analysis, display, etc., and time series database closely fits the characteristics of time series data design, Aggregate queries at the write and interaction levels that can support high concurrency and high throughput are often required.

Time Series Database uses the following key technical points to support Time Series data storage:

(1) High throughput writing ability. To achieve high-throughput writes to the system, two technical requirements are generally met: the system has horizontal scalability and a stand-alone LSM architecture. Horizontal scalability increases write throughput by increasing nodes, while the LSM (Log-Structured Merge) architecture makes data write only need to write memory and append write logs, eliminating the need to randomly write data to disk, ensuring high throughput writes for a single machine.

(2) High compression ratio. Compressed storage of time series data not only saves storage costs but also stores more data in memory. Since the time series data itself is a series of values that change with time, you can take advantage of the characteristics of time increment, dimension repetition, and indicator smoothing to reasonably select the relevant compression algorithm to effectively improve the data compression rate.

(3) Multi-dimensional query. Multidimensional query is a hard demand for time series databases, and most of the time series databases that have been released on the market basically support multidimensional queries, and the means mainly include bitmap indexes and inverted indexes. Multidimensional queries can present data from different perspectives, helping analysts to analyze more flexibly.

(4) Efficient polymerization ability. The query of time series data is usually a query of aggregated data, and it is a very time-consuming operation to query and aggregate the original data that meets the criteria on the basis of large data volume. At present, the industry's more mature solution is to use pre-aggregation, and complete the basic aggregation operation when the data is written in, reducing the amount of computation when the real-time aggregation query is written.

(5) Expiration processing capacity. The value of a lot of historical data is not large, high-precision historical data will be deleted, only low-precision historical data is retained, which requires the system to have a good expiration processing strategy, while deleting a large amount of data to save space while ensuring the normal operation of the system.

3.2 Time Series Database Benefits

So why did most respondents use time series databases instead of regular databases? There are two reasons for this: scale and availability.

(1) Scale: Time series data accumulates very quickly. (For example, a connected car can collect 25GB of data per hour.) Conventional databases are not designed to handle data of this size, and relational databases work very badly with large data sets; NoSQ database L handles scale data well, but it doesn't work as well as a database that fine-tuned time series data. In contrast, time series databases (which can be based on relational or NoSQL databases) treat time as a first-class citizen, processing this large-scale data through efficiency and bringing performance gains, including: higher Ingest Rates, faster large-scale queries (although some support more queries than other databases), and better data compression.

(2) Availability: TSDB usually also includes some common functions and operations for time series data analysis: data retention policies, continuous query, flexible time aggregation, etc. Even if you don't consider scale right now (for example, you're just starting to collect data), these features still provide a better user experience and make your life easier.