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

**1 What is time series database?**

Time series database refers to data mainly used to process time labels (changes in chronological order, that is, time series), and data with time labels is also called time series data.

Time series data is mainly collected by the power industry, chemical industry, meteorological industry, geographic information and other types of real-time monitoring, inspection and analysis equipment, the typical characteristics of these industrial data are: fast frequency (each monitoring point can produce multiple data within a second), heavy dependence on the collection time (each piece of data requires a unique time), measurement points have a large amount of information (conventional real-time monitoring system has thousands of monitoring points, monitoring points produce data every second, Generate tens of gigabytes of data per day).

So what are the essential differences between time series databases and traditional big data storage solutions? I think the most important difference is structured data.

(1) What is stored is structured data. We all know that the traditional big data solution to store data contains structured, semi-structured, unstructured data, so it determines that we can not decide which fields and define the data types of each field, like hbase is stored uniformly through the byte type, that is to say, the data put into the hbase is a byte array, from the ordinary type to the byte array needs to do it ourselves, we do not know how to convert to byte Its storage efficiency will be higher. However, the data generated by time series data is structured data, and we can define the fields and types of data in advance, so that the database system can choose the optimal compression method according to different field types, which greatly improves the utilization of storage.

(2) Analysis aggregates structured data. Since the analysis aggregation is structured data, we do not need to use a complex computing tool like mapreduce, and generally do not need a data warehouse like hive, but only need to cohesive at the database storage level similar to sum, avg in this kind of computing tool, and even do some simple streaming calculations. Hyperconvergence provides the basis (hyperconvergence is to fuse multiple components similar to the previous big data processing scheme into one component, mainly because the structured data is too simple, the collection and calculation are relatively simple, which is also the development trend of the subsequent time series database, reducing the complexity of the system).

**2 What are the characteristics of a time series database?**

Based on the rapidly growing application needs of time series data and the characteristics that distinguish it from traditional relational data, time series databases generally have the following characteristics :

(1) High-throughput data high-speed writing capability. Because the time series service continues to generate massive amounts of data, and has high requirements for the speed of writes, the concurrency of writes is large, which requires the time series database system to achieve high throughput data high-speed write function.

(2) High compression ratio. Time series databases need to store a large amount of data, and some monitoring data may take a long time to store, 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 business are divided into two categories, one is real-time data query, reflecting the status of the current monitoring object; The second is mainly to query the historical data of a certain time period, the amount of historical data is very large, at this time, it is necessary to optimize a large number of data queries for the time window.

(4) Efficient polymerization ability. Time series business scenarios usually care about the aggregate values of data, such as count, mean, and other aggregate values to reflect the data situation in a certain time period, so the time series database needs to provide efficient aggregate functions.

(5) Batch deletion ability. Time Series business requires batch deletion of expired data.

(6) It is usually not necessary to have the ability to do things.

Time series database is different from traditional relational databases, which focus on adding, deleting, modifying, and transaction functions, while time series databases are written to massive amounts of data, and their read queries are mostly data within a period of time. Traditional relational databases are the use of B tree, is a random read and write mode, will consume more time on the seek, for more than 90% of the scene is written to the time series database is too inefficient, so the mainstream of the time series database is to use LSM Tree (Log-Structured Merge Tree, structured merge tree) to replace the B Tree, such as KairosDB (the underlying use of Cassandra's stand-alone mode), The core idea of openTSDB (the underlying use of Hbase), LevelDB, etc., is to give up part of the ability to read in exchange for maximizing the ability to write. Of course, in addition to LSM Tree as a storage mechanism, time series databases have many other attempts and improvements, and are constantly evolving.

So why did most respondents use time series databases instead of regular databases? Why is TSDB the fastest growing database today? There are two reasons for this:

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.

Availability: TSDB typically also includes some common features and operations for time series data analysis: data retention policies, continuous queries, flexible time aggregation, and so on. 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.

This is why developers are increasingly adopting time series databases and using them for a variety of usage scenarios

**3 The Evolution and Future of Time Series Databases**

Although time series databases have only entered the public eye in recent years, their development can be traced back to the 1990s, when they were born in the field of surveillance

The need for sequential data storage, the resulting first generation of time series databases, represented by RRDtool (Round Robin Database Tool) and Whisper, use fixed-size databases, which can quickly store numerical data over time, but its read performance is still relatively weak, lack of special optimization for time, and the processing data model is single, usually embedded in the monitoring system.

With the development of big data, time series data explosive growth, not only monitoring systems, other systems also have more processing of time series data needs, in 2011 began to appear openTSDB, KairosDB as the representative of distributed storage-based time series database, such time series database in the inheritance of common storage advantages on the basis of optimization for time. For example, the underlying OpenTSDB relies on HBase cluster storage to compress data according to the characteristics of time series to save storage space. Use TSD (Time Series Daemon) to read and write, encapsulate common queries for time series data, and provide data aggregation, filtering, and other operations.

With the development of microservices, time series databases are developing at a high speed, and the shortcomings and deployment complexity of OpenTSDB have promoted the birth of low-cost vertical time series databases. The vertical time series database represented by IfluxDB has become the mainstream of the time series database market, and has more efficient data processing capabilities such as storage and reading and efficient compression algorithms for time series data.

In the past 5 years, the development of time series databases has been very rapid, major Internet companies including Google, Ali, Amazon have launched their own time series databases, DB-Engines since 2014 has also used time series databases as an independent directory for classification statistics.

Time series database is in a stage of rapid development, time series data technology is gradually maturing, but this is by no means the end, time series data technology is still facing a variety of new needs and challenges. While improving the performance of time series databases, major vendors are proposing more solutions for new requirements:

(1) Cloud services. In addition to the stand-alone version, many manufacturers have also released distributed versions, cloud service versions, especially cloud services, which have become an unstoppable development trend.

(2) Visualization services. With the advent of the Internet of Everything, the user's demand for a comprehensive grasp of information is growing, and the visual display of time series data has become a major trend, which puts forward higher requirements for the query capability of time series databases.

(3) Edge computing services. In the era of the Internet of Everything, the huge amount of data brought by more sensors is difficult to load by centralized processing, which makes data computing develop to the marginal, and the device will store the data through the edge device for real-time processing and analysis feedback, which can improve the real-time response ability of the device and enhance the value of time-sensitive data, so the support of the time series database for edge computing will become an important function. In the face of these challenges and opportunities, it is believed that the time series database will have a deeper development, and the future can be expected.