JIA YUAN

Email : yuanjia339@gmail.com — Tel : (+86) 17827063225

EDUCATION

SUN YAT-SEN UNIVERSITY, GuangZhou, China Sep. 2018 - Jun. 2022
Bachelor of Computer Science and Engineering

REASEARCH INTERESTS

AI for DB, autonomous DBMSs, cloud-native databases, data-driven machine learning

WORK/RESEARCH EXPERIENCE

Tencent

ShenZhen, China
Cloud Database Research and Development Department — **AI for Database Project Team**

Position: Intern Oct. 2021 - Apr. 2022

Cross-Machine Histogram

- Contributed to the project ideation by proposing sampling more data from the slave-node of database intance to build high-precision histograms which improve cardinality estimation accuracy.
- Utilized the SpaceSaving algorithm to manage large volumes of data and avoid sorting. Improved its data structure to efficiently track the top K elements and their frequencies within massive dataset. Histograms built using the Top K elements can maintain high accuracy even in highly skewed data scenarios.

Position: Full Time Jul. 2022 - Present

cdbtune Project Phase 2

- cdbtune has introduced reinforcement learning into the field of database knobs tuning for the first time, and it has been successfully commercialized.
- Upgraded the tuning process from being based on sysbench/TPC-C benchmarking to utilizing CDB Workload Generation for more personalized knobs tuning.
- Refactored the cdbtune task scheduling system and achieved stable parallel tuning capabilities.

CDB Workload Generation

- Collaborated with Dr.Baoqing Cai to develop CDB Workload Generation, a new benchmark that reflects the complex and dynamic behavior of real-world workloads. This benchmark generates data with the same distribution as the user's data and preserves the concurrency characteristics of query loads while also retaining the fluctuation characteristics of loads over time.
- CDB Workload Generation generate anonymized queries and also anonymized data, which allowed it be a cloud service. We continuously collected representative user workloads for our serive. Customers can choose the one that best suits their needs by considering the cloud monitoring metrics and metadata from the original workload.

CDB Workload Replay

- Implemented cdb workload replay as an easy to use cloud service that supports stably replay billions of queries. Customers only need to specify the original and target instances, as well as the desired time period, in order to "reproduce the history".
- Now, it is widely used to test changes before they are applied to an instance and assist DBAs in anomaly diagnosis, and it has also successfully assisted commercial POC several times.
- Compared to Oracle Workload Replay. Found that waiting between transactions would cause drift in replay sessions which seriously disrupted the concurrency pattern. So CDB Workload Replay did not implement transaction synchronization.

Anomaly Detection and Root Cause Analysis System—Begining Stage

- Responsible for project design : Due to the instability of OLTP task requests and the complexity of the underlying operating system and database system, cloud monitoring metrics often contain various noises. Decided to detect anomalies by monitoring database load instead of applying temporal anomaly detection algorithms to monitoring metrics.
- Use robust multiple periodicity detection algorithms to extract potential periodic load peaks , thereby reducing false positives in anomaly reporting.
- Use clustering and ranking to locate suspicious SQL queries. Compare the current load with the baseline load from a week ago in real-time to identify hard-to-locate anomalous SQL, such as new queries. Use the wait events from the kernel to analyze anomalies and generate reports.

AWARDS

- Tencent Technology Breakthrough Award for the Second Half of 2022 — Cloud Native Database TDSQL-C And CDB — AI for Database Project was nominated
- Tencent Open Source Collaboration Award — TDSQL-C And CDB Oteam

SKILLS

- **integrate machine learning (ML) with systems** — Two years of experience in AI4DB
- **Languages** — Python, C, C++, SQL, Golang
- **Packages and Frameworks** — Git, Kafka, Flink, TensorFlow, Tensorlayer