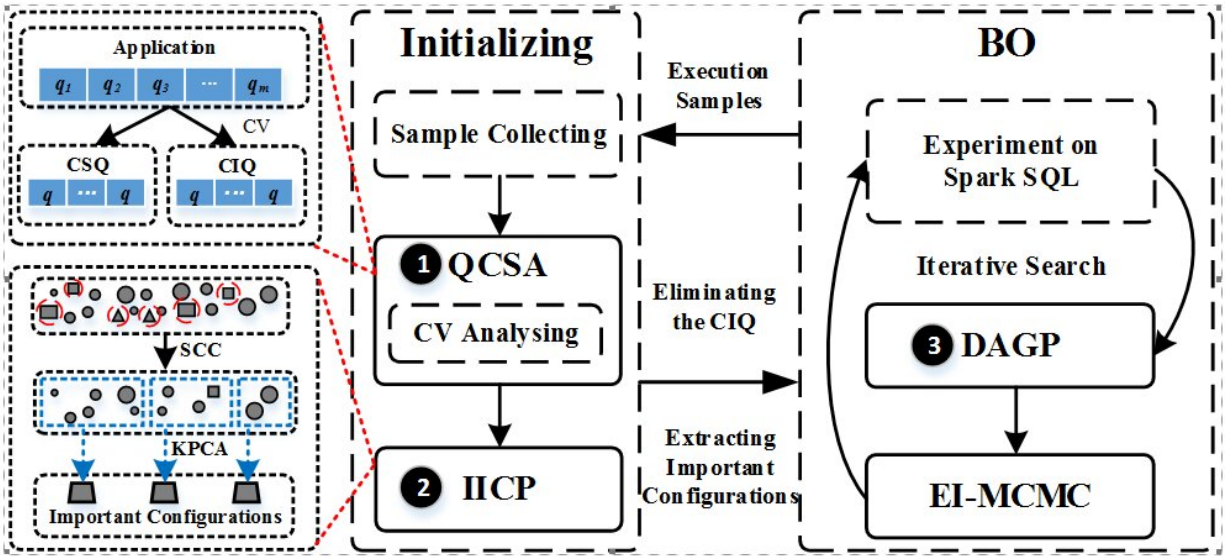
Novel tuning method for Spark SQL applications

June 16 2022, by Li Yuan



An overview of LOCAT. Credit: YU Zhibin

Spark SQL is a Spark module for structured data processing. It has been widely deployed in industry but it is challenging to tune its performance.

Existing machine learning tuning methods are difficult to apply in practice due to the high time cost and failure to adapt to the changes in the amount of data to be processed.

To address these problems, a research team led by Prof. Yu Zhibin from

the Shenzhen Institute of Advanced Technology (SIAT) of the Chinese Academy of Sciences proposed a low-time-cost automatic configuration optimization method named Low-Overhead Online Configuration Auto-Tuning (LOCAT), which could reduce the optimization time and improve performance of Spark SQL.

The results were published at SIGMOD 2022, an international forum for database researchers, practitioners, developers, and users. The associated paper can be found in *Proceedings of the 2022 International Conference on Management of Data*.

The researchers first designed query and configuration parameter sensitivity analysis techniques for LOCAT. Queries that were insensitive to configuration parameters were identified and removed from a given workload when training samples were collected.

"For the remaining queries, LOCAT calculated correlation coefficients to identify important configuration parameters," said Prof. Yu. "Then, it applies kernel principal component analysis to reduce the dimension of configuration parameter search."

Finally, the researchers designed Bayesian optimization for LOCAT, which is aware of the dataset size to search for the optimal configuration so that its performance can be automatically optimized based on the size of the dataset.

The experimental results on the ARM cluster (a cluster of servers for big data computing, in which each server uses CPU based on the ARM instruction) showed that the LOCAT accelerated the optimization procedures of the state-of-the-art approaches by at least 4.1x and up to 9.7x. Moreover, the LOCAT improved the application performance by at least 1.9x and up to 2.4x. On the x86 cluster, LOCAT showed similar results to those on the ARM cluster.

More information: Jinhan Xin et al, LOCAT: Low-Overhead Online Configuration Auto-Tuning of Spark SQL Applications, *Proceedings of the 2022 International Conference on Management of Data* (2022). DOI: 10.1145/3514221.3526157

Provided by Chinese Academy of Sciences