中国科学院软件研究所学术年会’2022
暨计算机科学国家重点实验室开放周

工具系统

面向CockRoachDB分布式数据库的查询级别参数调优工具

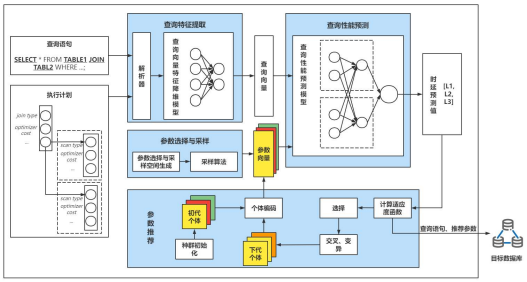
邹逸，方相，许利杰，王伟
软件工程技术研究开发中心
联系方式：xulijie@otcaix.iscas.ac.cn wangwei@otcaix.iscas.ac.cn

工作介绍

场景：数据库系统中包含大量可配置的参数，它们影响着内存分配、数据I/O、任务并行化等模块的运行方式。传统由数据库管理人员手工调优的方法既耗时又容易出错。近年来，使用机器学习算法进行自动参数调优的研究逐渐兴起，其性能和易用性已超过了传统的调优方法。

挑战：现有的数据库参数调优工作普遍专注于单机关系型数据库上的负载级别，**分布式数据库上的研究工作刚刚起步，而分布式数据库上支持查询级别的参数调优工作更少**。

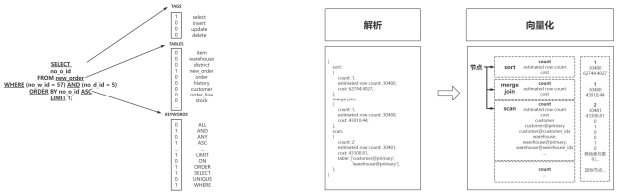
工作内容：本课题以CockRoachDB作为目标数据库，设计并实现了面向CockRoachDB分布式数据库的查询级别参数自动化调优工具。实验表明，**工具能够在较小时间开销下，有效降低查询执行的时延。**



问题分析

本工具采用机器学习方法支持为每条查询语句分别推荐使其在目标数据库上运行性能更优的参数配置。为根据查询特征有针对性的推荐能使其时延降低的最优参数，需要解决以下几个问题：

查询特征提取：输入查询语句以及执行计划，从查询语句中提取静态特征，如查询类别（SELECT、UPDATE、INSERT、DELETE）、数据表、关键字等，从执行计划中提取运行时特征，如每一步操作的类型，以及涉及的结果行数目、数据表、索引等，输出编码后的查询特征向量。

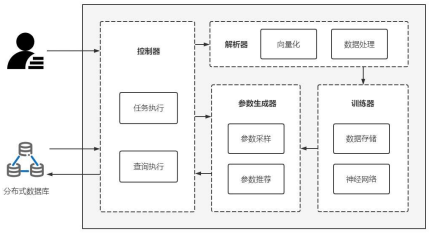


查询性能预测：基于深度神经网络模型，从训练数据中学习，其目标是对于输入的查询向量以及参数向量，输出该查询在该参数配置下运行的时延预测值。

参数推荐：基于遗传算法，在训练阶段得到的查询性能预测模型的指导下，通过搜索参数空间，快速优化推荐的参数配置对查询性能的提升效果。

工具设计

工具的总体架构：工具分为控制器、解析器、参数生成器和训练器四个主要模块。



控制器：用于和目标数据库交互（如执行查询并获取响应、在目标数据库上应用参数、重启数据库、导入数据等）以及处理用户的请求（如生成训练数据请求、推荐参数请求）。

解析器：用于解析查询的响应，响应包括查询向量、执行计划和时延。

参数生成器：用于生成参数，包括采样方法以及推荐方法，由控制器调用（在生成训练数据时调用采样方法，在处理参数推荐请求时调用参数推荐方法）。

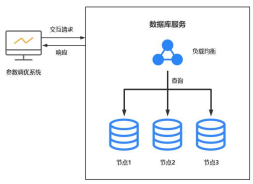
训练器：用于数据持久化以及神经网络的训练，需要持久化的数据包括训练数据以及神经网络模型的权重。

系统评价

实验目标

- 测试工具在典型OLAP负载上对数据库查询的调优效果。
- 测试工具在参数推荐时各方面的时间开销。

实验使用TPC-H测试集。

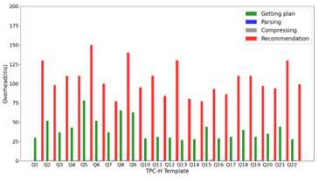


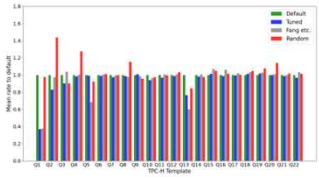
实验流程

- **训练阶段：**从日志中获取TPC-H查询集合，作为初始数据，训练查询特征向量降维模型以及查询性能预测模型；
- **推荐阶段：**基于遗传算法，利用训练阶段得到的查询特征向量降维模型和查询性能预测模型，快速搜索参数空间，选择使得目标查询性能最优的参数配置。检查查询主要性能指标的提升情况以及所需的时间开销。

实验结果

下图表明本工具在负载上运行时，**在仅带来极小的开销前提下，对于所有类别查询，工具平均能够将时延降低9.2%；对于部分类别查询，能将时延降低60%以上。**





实验结论：

- 本工具能够有效的降低分布式数据库场景下查询执行时延。
- 本工具能有效捕捉到各类查询的特点，从实验结果上可以观察到各类查询对参数的敏感性不同。