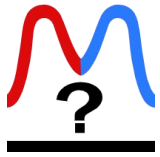
难题1：[易运维] 基于负载驱动的数据库基数估计

出题组织：高斯部　　接口专家：孙佶 sunji11@huawei.com

技术背景

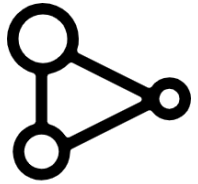
背景：数据库当前使用采样方式得到数据集直方图、max值等统计信息。优化器使用这些统计信息进行基数估计。该种方法在稀疏大表、频繁IUD以及多表连接等场景下会出现基数估计偏差量级以上的问题，直接影响数据库正确执行计划的产生。通过使用基于负载驱动的方式，在SQL执行期间收集查询相关的结果数据作为对统计信息进行基数估计的补充，将由统计信息计算基数估计的方式变为统计信息加负载驱动的方式，从而提升数据库基数估计能力是一个有效的技术方向。

- **稀疏大表基数估计：**业务中经常存在一些数据分布稀疏的大表，通过数据采样技术构建的统计信息往往由于采样数据和完整数据分布不一致导致统计误差很大。这类场景通常会存在一些经常被访问的“热数据”，比如按照时间的消息队列，状态有效的信息会比较集中在最近的时间段。因此基于在线查询负载驱动的基数估计具有较好的潜力。
- **频繁IUD场景基数估计：**数据库频繁IUD场景中，数据统计由于数据采样的时延无法及时更新，进而导致统计信息失效。相比于数据统计，利用查询负载反馈能够更加及时地在这种场景中纠正基数估计结果，提高执行计划效率。
- **多表连接基数估计：**针对多表连接查询，由于缺少相应的多表数据统计信息，基数估计误差比单表上的基数估计会高1-2个数量级。基于数据统计的方法需要对多表进行采样，不同连接条件需要不同的采样方法，效率较低更新不及时，落地困难。而基于负载驱动的方法能够通过分析负载进行基数估计纠正，不受数据库schema的限制。



大表小数据样本
导致分布漂移

数据更新频繁导
致分布漂移



跨表相关性信息
缺失导致误差

技术挑战

- **通过分析历史负载推理查询基数：**利用历史查询推理新查询的方法最简单的是使用查询精确匹配，但是这种应用范围很小，无法提升一批查询的估计准确率。但是利用近似查找技术也无法保证估计结果的鲁棒性，因此如果针对负载查询进行建模是一种很大的挑战。
- **模型漂移判断和更新：**随着负载执行，数据和负载查询都会发生变化，数据变化会导致数据分布和查询基数的变化，负载变化会导致查询分布复杂度增加。这些变化都会导致已经训练的模型针对新查询的基数估计准确率降低。

当前结果

- **人工介入更新统计信息：**当前数据库运行时，通过慢SQL监控等运维手段观察系统，当出现查询性能问题时一般需要人工介入，通过打hint或者更新统计信息来进行修正，存在滞后性和运维效率问题。
- **学术界局限在算法研究：**当前学术界提出了一些负载查询表征和建模的技术，包括混合模型，神经网络等，但是缺少系统性框架支撑商业落地，模型鲁棒性不足。
- **模型缺少可解释性和鲁棒性：**AI神经网络模型由于其强大的拟合能力被广泛使用，但是由于其是一种黑盒模型，不具备解释性；单纯查询表征的技术需要大量样本进行训练，这个过程很难在线进行，针对小样本场景存在鲁棒性和泛化性不足的问题。

技术诉求

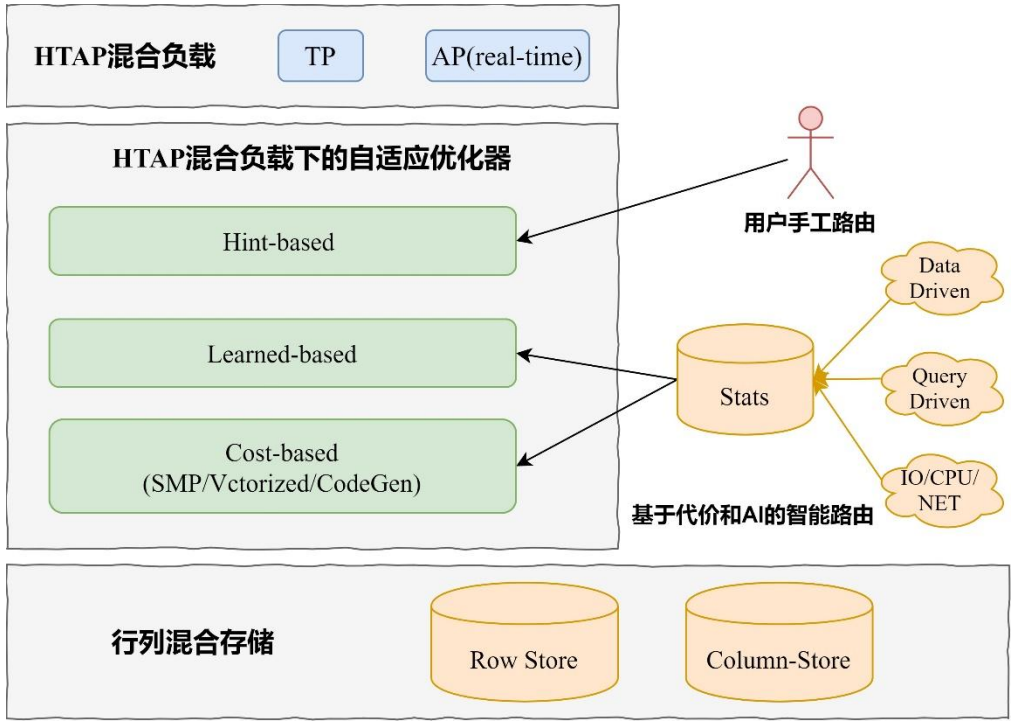
- **具体诉求和目标：**构建基于负载驱动的数据库基数估计引擎，兼顾估计准确率和估计效率，设计算法模型原理和自适应能力，目标是解决大表、IUD和多表连接场景中基数估计误差大，稳定性差的问题，降低实际系统中统计信息的运维工作。
- **精度指标：**与传统基于数据的统计信息相比，在典型的Benchmark场景下, 比如TPC-C(IUD场景),IMDB(多表连接)，以及大表场景等，针对查询的基数估计准确率提升5倍。
- **效率指标：**基数估计效率和传统基数估计在同一数量级，模型构建和更新效率高于传统采样1倍。

参考文献：
[1] Selectivity Estimation for Range Predicates using Lightweight Models. In VLDB 2019
[2] An End-to-End Learning-based Cost Estimator. In VLDB 2019
[3] Self-Tuning, GPU-Accelerated Kernel Density Models for Multidimensional Selectivity Estimation. In SIGMOD 2015
[4] QuickSel: Quick Selectivity Learning with Mixture Models. In SIGMOD 2020

难题2：[易运维] HTAP混合负载下的查询优化技术

出题组织：高斯部　　接口专家：张树杰 zhangshujie1@huawei.com

技术背景



- **混合负载的TP/AP智能识别：**当前的HTAP在存储层面主要是基于行列混存的方式来适应TP/AP的业务请求，因此需要优化器根据业务的特点将SQL查询路由到行存或列存，避免用户手工路由导致业务应用重新配置；
- **TP/AP混合负载下的代价优化：**在HTAP业务负载下，执行器需要混合行存扫描、列存扫描、向量执行、编译执行、并行执行等各种技术，这些需要代价系统即快又好的搜索出最优的执行计划，避免过度使用手工调优干扰，提高系统的易用性；
- **混合负载下的资源管控：**在HTAP混合负载的情况下，在支持既有的TP业务的同时，还需要支持实时的AP能力，对资源的分配是否合理会从 很大程度上影响系统的整体性能；

技术挑战

- **TP/AP智能识别的准确性：** 用户的业务系统复杂多变，HTAP要求查询的响应效率比较高，因此如何快捷的识别TP/AP业务、根据业务的特点确定SQL是否路由到只读节点、避免用户手工路由等导致业务应用重新配置成为HTAP数据库是否成功的要点之一；
- **优化器的搜索性能和准确率：**针对多种执行技术，现有的优化器使用基于人工添加规则的改写技术或者由于计划生成需要的统计信息不准确无法做到搜索的准确性。

当前结果

- **主要通过Hint手工指定业务模型：**手工指定是当前业界区分HTAP的最重要的手段，该方法无法根据业务特点的改变进行自适应调整，且需要对业务进行修改；
- **既有智能识别模型准确性低：**目前已有的HTAP混合负载智能路由方法在准确性和资源消耗（实时性）难以做到均衡、自适应的路由识别；
- **针对多种执行技术的自适应搜索准确性不足：**针对多种执行技术，现有的优化器要么手工对执行计划进行转换，要么需要的统计信息不足，无法做到搜索的准确性；

技术诉求

- **具体诉求和目标：**构建基于HTAP混合负载的查询优化技术，实现TP/AP业务的智能路由（本地和远端只读结点），针对多种执行技术，优化器能够即快又好的搜索出最优执行计划，避免过多的手工干预，提高系统的易用性；
- **精度指标：**在HTAP混合负载业务下，应用新的查询优化技术，避免使用过程中依赖DBA的经验，系统资源消耗升高不超过10%，HTAP业务查询性能提升20%；
- **效率指标**：执行效率和传统代价模型在同一数量级；

参考文献：
[1] Li G, Zhang C. HTAP Databases: What is New and What is Next[J]. 2022.
[2] Sirin U, Dwarkadas S, Ailamaki A. Performance characterization of htap workloads. 2021 IEEE
[3] Chen J, Ding Y, Liu Y, et al. ByteHTAP: bytedance's HTAP system with high data freshness and strong data consistency[J]. 2022.

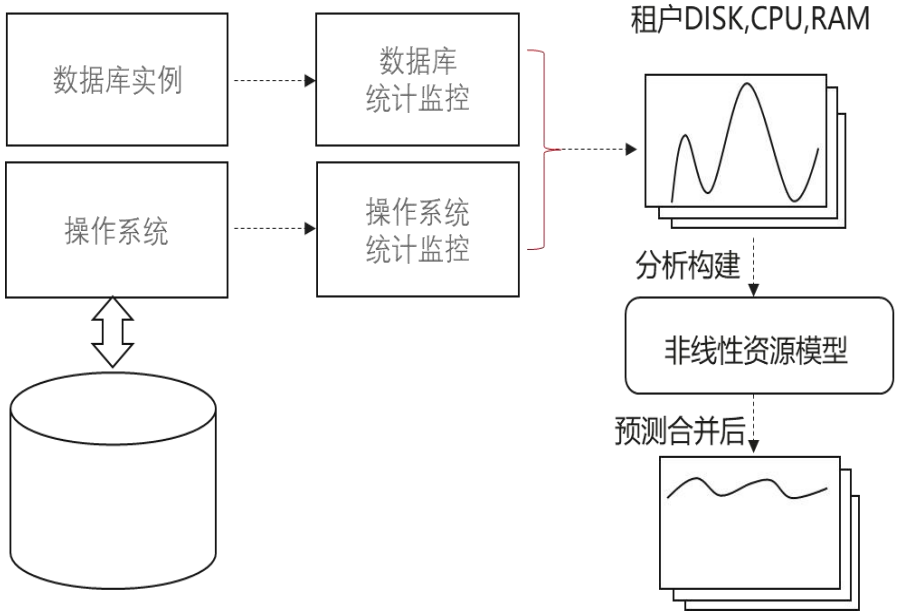
难题3：[高性能] 数据库租户资源动态预测

出题组织：高斯部　　接口专家：田文罡 tianwengang@huawei.com

技术背景

价值：在大多数企业中，数据库部署在专用数据库服务器上。通常这些服务器大部分时间都没有得到充分利用。近期工作（如《Workload-Aware Database Monitoring and Consolidation》）在对不同组织的近200台生产服务器的跟踪调查中发现，服务器CPU平均利用率不超过4%，如果能够很好的进行数据库动态资源预测从而进行调度，将大幅提升服务器CPU利用率，降低云数据库平台的资源消耗。

背景：租户数据库调度的前提是对租户使用的数据库资源进行画像和预估。资源监视器统计每个服务器上运行的租户数据库的CPU、RAM、磁盘I/O、缓冲池利用率和XLOG刷盘信息。通过周期采样获得每个租户数据库资源使用率的时间序列的集合，即工作负载画像，通过刻画较准确的负载画像将有利于租户的调度从而达到削峰填谷且保证租户SLA的目标。



技术挑战

数据库租户调度有两个挑战性问题：

- 需要准确监控每个租户数据库的资源利用率，并估计一组数据库合并后的资源利用率。数据库涉及CPU，内存，IO三种资源，合并时需要同时考虑三种资源的情况。
- 需要算法来选择哪些租户数据库应合并并放置在哪些硬件上，在云上有数百个租户数据库和物理资源可供选择，这个搜索空间很大。租户优化合并的目标是（1）尽量减少支持数据库租户工作负载所需的服务器。（2）最大限度地提高合并后数据库服务器的负载平衡。（3）保证分配到每个数据库服务器的工作负载不过载。

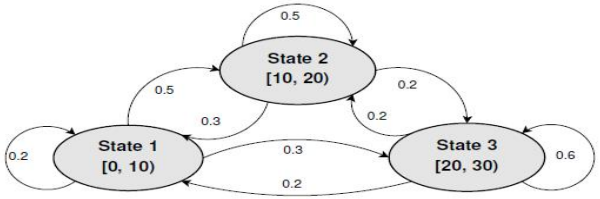
当前结果

当前业界使用的资源算法预测有如下几种，这些算法是通用的预测算法，没有考虑数据库事务，SQL执行特征等。

- **平均值：**预测的资源需求为在窗口采样样本上的平均资源使用量；
- **自动回归：**这个方案尝试基于先前值Xt-1、Xt-2预测系统的当前值Xt，使用公式Xt =a1Xt-1+a2Xt-2...系数（a1、a2等）通过计算自相关来确定系数和求解线性方程组；
- **马尔科夫模型**：根据状态转换的概率实时预测资源需求。

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} \ldots$$

$$\sum_{i=1}^{Z} a_i R(i-j) = -R(j), \text{ for } 1 \le j \le Z$$



技术诉求

- **具体诉求和目标：**根据租户数据库运行N个周期内的CPU、内存、IO等资源消耗数据，充分考虑数据库事务与SQL执行等特征，建立一个数据模型，设计一种数据库负载动态资源时序预测算法来高效准确的预测租户数据库CPU、内存、IO等资源消耗。
- **精度指标**：数据库租户动态资源时序趋势预测准确率达到80%以上，资源预测算法时延在毫秒级。

参考文献：
[1] Multi-Tenant Cloud Data Services: State-of-the-Art, Challenges and Opportunities. In SIGMOD 2022
[2] Scheduling of Time-Varying Workloads Using Reinforcement Learning. In AAAI 2021.

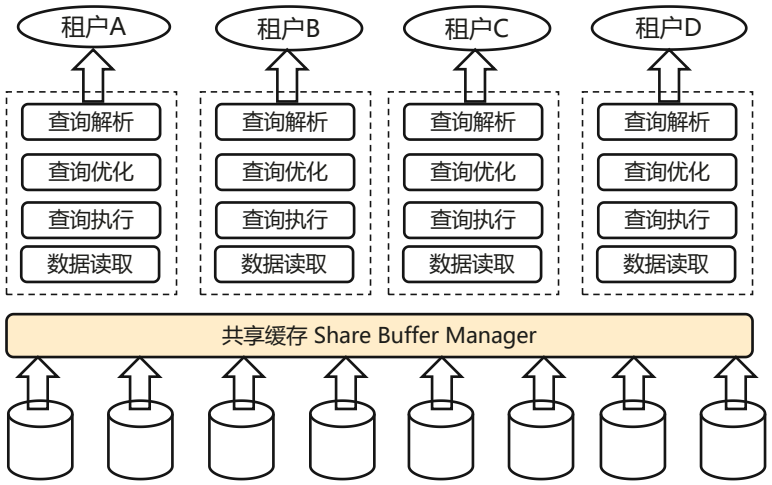
难题4：[易扩展] 多租共享缓冲区管理技术

出题组织：高斯部　　接口专家：钟舟 zhongzhou1@huawei.com

技术背景

价值：多租户(Multi-Tenant)技术可以让多个租户在保证资源、安全隐私方面隔离的前提下共享同一套运算环境或者服务器实例，对云服务商来说可以有效的降低云上环境建置、运维方面的成本，多租环境下数据库共享缓冲区的资源利用率提升对系统成本、服务稳定性等有明显影响。

背景：数据库服务DBaaS多租需要给每个租户提供完备的资源隔离运行环境，在传统的数据库体系结构中可以为租户提供独立CPU计算资源、内存资源、IO资源、网络资源等，但是数据库shared buffer是服务于全局所有租户，在进行dirty-page交换时并没有考虑上层租户之间的影响，因此可能存在租户A的大量读写操作影响到租户B的数据读取，导致共享缓冲区的资源没法隔离、全局资源没未充分利用，造成资源的无效浪费。



内核工作线程（已隔离）
- ✓ CPU计算资源
- ✓ runtime内存资源
- ✓ 连接数资源控制
- …

数据共享缓冲区（未实现）
- ✓ 所有租户共享
- ✓ 刷页策略不考虑租户
- ✓ 租户间相互影响

技术挑战

- **缓冲区页面淘汰算法：**在共享缓冲区进行页面淘汰调度算法从设计上需要考虑租户脏页面之间的相互影响，需要从调度策略上确保租户间隔离不感知，达到多租场景下的资源隔离效果。
- **增强自适应能力：**淘汰算法新增租户维度的逻辑处理需要尽可能无损，即在缓冲区资源相同的前提下，相比原有的页面淘汰算法无新增时延。

当前结果

- **共享缓冲区不感知租户**：缓冲页面调度算法全局调度，并不在调度策略上区分用户/租户，加入某一租户由于进行大量的IO负载的作业时，影响其他租户对共享缓冲区资源的使用，造成租户间缓冲区资源并没有完全做到隔离，无法满足共享实例场景下数据多租的基本要求。
- **共享缓冲区缺少精细化控制：**即便不考虑租户之间的影响，现有的共享缓冲区无法从使用量的角度控制某一租户的共享缓冲区资源使用。

技术诉求

- **具体诉求和目标：**设计和实现租户感知Tenant-Aware的缓冲区管理方案，能够基于租户层级灵活分配buffer缓冲区资源，进一步精细化缓冲区资源的利用，从资源利用率角度做到在满足一定性能的前提下，缓冲区内存资源相比原来减少20%，从性能的角度在总体资源一定的前提下所有租户性能总和提升20% 。
- **精度指标：**在新的缓冲区管理机制下能够对租户级别共享缓冲区使用量在MB级别进行控制，从而达到更加精细、精准资源控制，并提供实时观测手段加以佐证。

参考文献：
[1] Tenant Placement in Over-subscribed Database-as-a-Service Clusters. In VLDB 2022
[2] Multi-Tenant Cloud Data Services: State-of-the-Art, Challenges and Opportunities. In SIGMOD 2022
[3] Scheduling of Time-Varying Workloads Using Reinforcement Learning. In AAAI 2021.
[4] Tempo: Robust and Self-Tuning Resource Management in Multi-tenant Parallel Databases. In VLDB 2016.

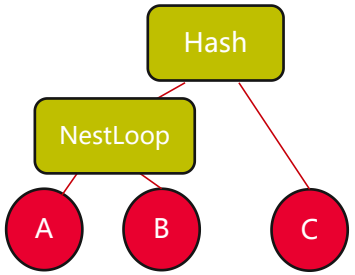
难题5：[高性能] 分布式优化器中模型构建和求解算法

出题组织：高斯部　　接口专家：刘梦醒 liumengxing@huawei.com

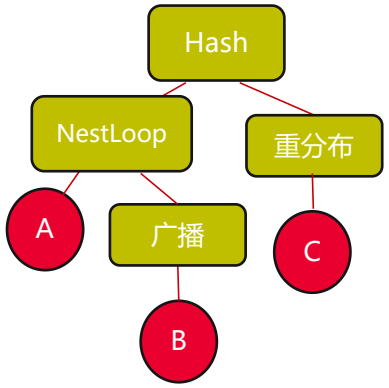
技术背景

价值：分布式优化器是分布式数据库的核心能力，希望可以在约束时间内能找到尽量优的执行计划。

背景：数据库执行查询语句时，会构造一棵查询树来执行，而通常一个查询可以对应很多查询树，需要优化器选出一棵最佳的查询树。集中式数据库中，通常通过动态规划算法来求解；分布式数据库中，子问题的求解方式，会影响子树的数据分布情况，进而影响父节点的算法选择方式。



单机优化器中 (A join B) join C 可以先求解成 A join B 的子问题，再把AB作为整体求解（AB）join C



分布式优化器中需要考虑数据的分布问题。子问题的求解方式，会影响父问题的数据传播方式，因此子问题的最优解未必是全局最优解

技术挑战

- **缺乏有效的求解模型。**在分布式数据库执行计划产生过程中需要考虑数据分布情况，导致局部最优需要考虑网络数据传输的问题，从而导致同全局最优之间的矛盾。
- **同时兼顾**算法准确性和求解速度

当前结果

- **单阶段优化：**枚举所有可能的分布情况
  局限：搜索空间极度膨胀，耗时增加。工程实践中普遍基于规则剪枝，但是也会剪掉可能的最优解
- **两阶段优化：**（1）基于代价模型生成单节点最优计划 （2）基于规则转为分布式计划
  局限：单节点计划最优不一定是全局最优，分布信息在第一阶段被丢掉

技术诉求

- **具体诉求和目标**：建立新的分布式优化器的模型和求解算法，从理论上保证可以找到全局最优解，并且算法复杂度跟当前单机优化器的复杂度接近。
- **准确度指标**：在TPC-H、TPC-DS、TPC-E 等 benchmark 中可以找到最优计划。
- **效率指标**：算法复杂度不超过 O(3^n)，n为表的数量。工程落地时，优化器产生计划的时间耗时不超过当前的2倍。

参考文献：

[1] WeTune: Automatic Discovery and Verification of Query Rewrite Rules. In SIGMOD 2022.
[2] SPRINTER: a fast n-ary join query processing method for complex OLAP queries[C]. In SIGMOD 2020
[3] The MemSQL Query Optimizer: A modern optimizer for real-time analytics in a distributed database. In VLDB 2016