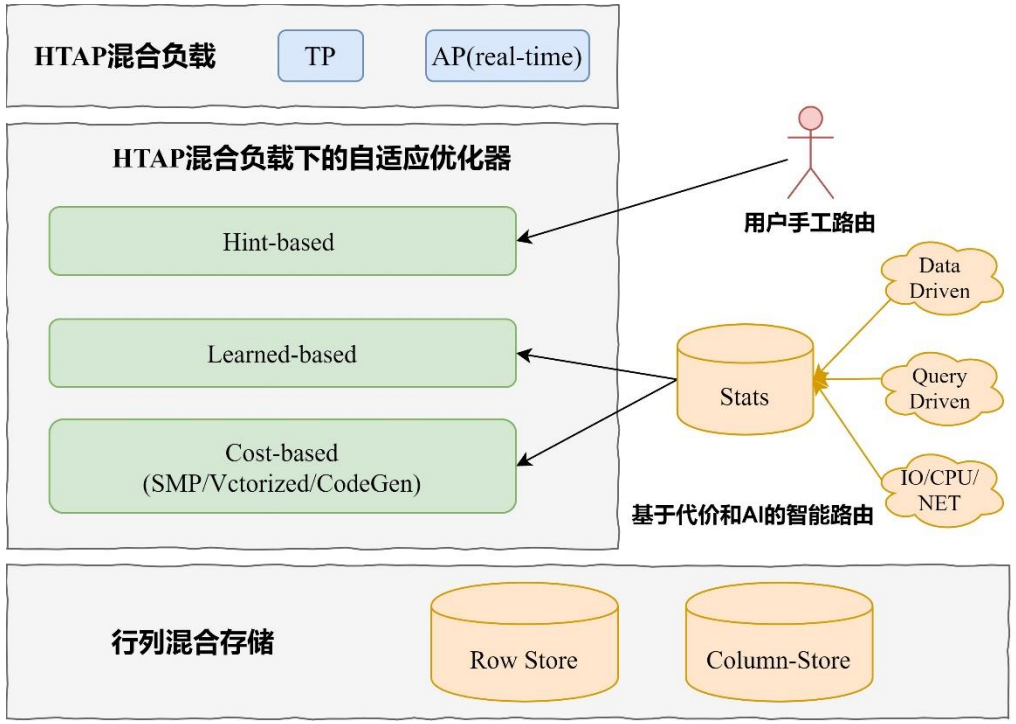
难题1：[易运维] HTAP混合负载下的查询优化技术

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.

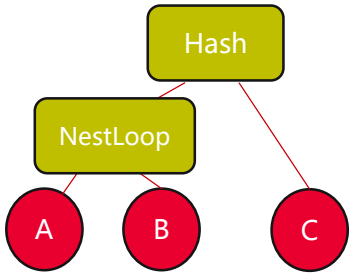
难题2：[高性能] 分布式优化器中模型构建和求解算法

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

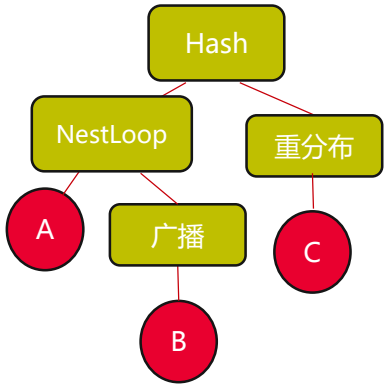
技术背景

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。

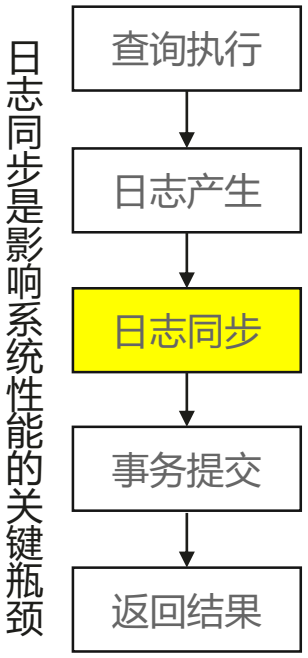
参考文献：

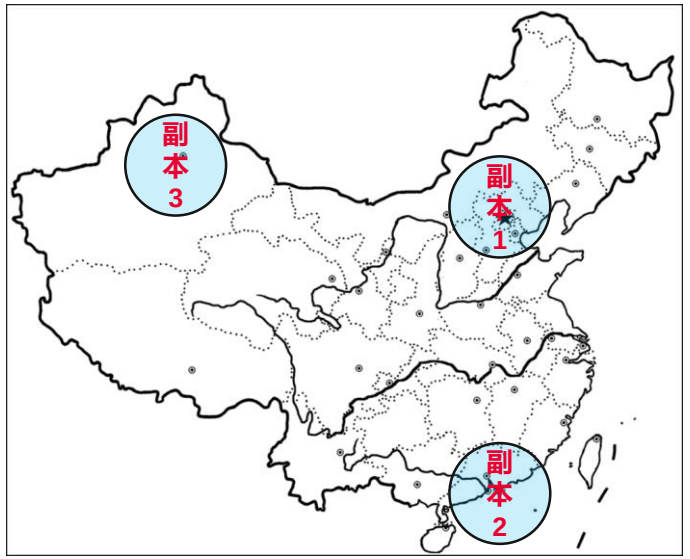
[1] WeTune: Automatic Discovery and Verification of Query Rewrite Rules. In SIGMOD 2022.
[2] SPRINTER: a fast n-ary join query processing method for complex OLAP queries[C]. In SIGMOD 2020
[3] The MemSQL Query Optimizer: A modern optimizer for real-time analytics in a distributed database. In VLDB 2016

难题3：[高可靠] 数据库跨Region一致性日志同步协议 – 打造全球分布数据库竞争力

出题组织：高斯部　接口专家：王磊 wanglei110@huawei.com

价值：副本间日志同步能力是数据库高可用能力的核心竞争力，是金融核心系统、公有云业务等全球多地多中心场景的关键痛点需求





跨Region副本之间的时延高达100ms量级

传统副本同步协议

- **Quorum：**当日志被其他多数派副本接受之后，认为同步完成
- **Paxos/Raft：**由prepare、accept等多个阶段组成的分布式一致性共识协议

传统日志同步协议的技术问题

- **受时延影响严重**：高时延下，同步性能迅速下降[1-4]
- **与事务提交和并发控制紧密耦合**：对于有依赖关系的查询业务，如果前序查询由于日志同步阻塞事务提交，那么后序依赖的查询会阻塞在事务并发控制模块，无法启动查询执行
- **只有单个副本支持写操作**：通常只有主副本才能修改或者插入数据，其它备副本只能支持读业务，影响整体系统的写能力扩展

技术挑战

- **打通同步协议和网络协议：**深入研究网络收发协议和日志同步协议的协作方式，将跨Region网络资源（带宽、时延、CPU等）的利用率发挥到最佳
- **解耦同步协议和事务提交/并发控制协议：**深入研究事务提交/并发控制和日志同步协议的协作方式，在保证并发事务因果依赖序的前提下，将查询执行的效率发挥到最佳
- **探索多副本多写协议：**在日志同步协议层探索支持高性能多副本多写的方案设计，大幅提升整体系统的读、写扩展能力

技术诉求总结

- **诉求和目标：**设计一种适合高时延下的日志同步和事务提交算法，包括日志同步协议、网络收发模型、事务并发控制等，提升系统吞吐量，降低业务时延
- **性能指标：**TPC-C负载模型下，100ms往返时延下的日志传输吞吐量达到2ms往返时延下的80%以上；在100ms往返时延和冲突率30%情况下（包括读写冲突和写写冲突），事务平均执行时延降低10%

参考文献：
[1] Y Amir, C Danilov, M Miskin-Amir, et al. On the performance of consistent wide-area database replication[Technical Report] (2003) cnds.jhu.edu
[2] H. Mahmoud, F. Nawab, A. Pucher, D. Agrawal, and A. El Abbadi. Low-latency multi-datacenter databases using replicated commit. Proc. VLDB Endow., 6(9):661–672, July 2013
[3] T. Kraska, G. Pang, M. Franklin, S. Madden, and A. Fekete. MDCC: Multi-data center consistency. In EuroSys, pages 113–126, 2013.
[4] H. Moniz, J. Leitão, J. D. Ricardo, et al. Blotter: Low Latency Transactions for Geo-Replicated Storage. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, pages 263–272, 2017.

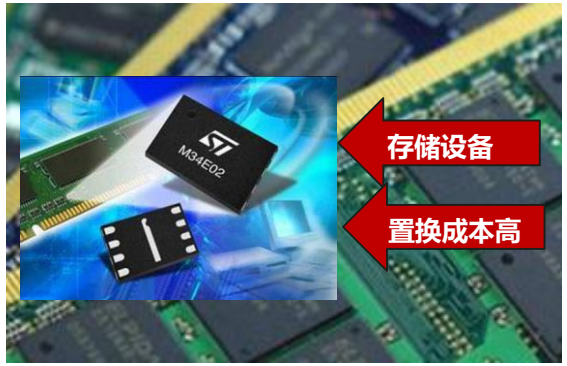
难题4：[泛存储] 数据库感知的高效压缩方案

出题组织：高斯部　接口专家：陈志远 hw.chenzhiyuan@huawei.com

价值：在数据激增的趋势下，存储介质的成本迅速增长，在保障性能的前提下提升存储密度，可构筑数据库的核心竞争力

数据库压缩需求

- 随着大数据、云计算、物联网的急速发展，数据量指数增长，需要更多的存储空间
- 更换已有硬件需大量投资，提升存量硬件的利用率可延长产品生命周期



通用数据压缩 VS 数据库压缩

- **通用数据压缩**
  - ✓ 分为有损压缩和无损压缩
  - ✓ 数据对象特征未知，无法参考数据对象的已有特征
  - ✓ 学术上没有新的理论突破，目前仍基于熵理论或字典理论，工业应用上常见于多种经典算法的组合调优使用
- **数据库压缩**
  - ✓ 只接受无损压缩
  - ✓ 有大量对数据的增删改查操作，因此对于数据压缩的性能有较高的要求
  - ✓ 数据按照固定的格式进行存储，压缩对象自身结构已知

技术挑战

- **数据库压缩：结合数据库页面存储特点，针对数据库场景的高效数据压缩**
  - ✓ 压缩数据块改动需要整体解压/压缩，需要定位数据待变更点，进行定点处理，减少不必要的对象压缩/解压
  - ✓ 针对数据频繁变化的特点，快速动态修正数据库统计模型，及时调整最优编码方案，保证压缩率和压缩性能
- **数据库+通用数据压缩：针对数据库的特点，设计数据库操作感知的压缩算法**
  - ✓ 数据库操作频繁且性能敏感，因此压缩/解压吞吐量一般要求大于300MBps。在典型3:1压缩比下，通用压缩算法性能无法满足需求，需要利用结构化存储特征优化压缩/解压速度
  - ✓ 数据压缩存储需要考虑数据库的ACID等约束，同时存储方式有利于数据库读写

技术诉求

- **具体诉求和目标**：探索针对数据库的高效无损压缩方案，针对数据进行自适应选择性压缩，实现整体存储空间占用比当前业界最优方案降低30%以上。优化目标需包含openGauss关系型数据库
- **性能目标**：相对非压缩方案，数据库在通用场景下读写性能下降不超过5%，读多写少场景下性能挑战不下降

参考文献：
[1] A novel encoding algorithm for textual data compression. In bioRxiv, 2020
[2] NNCP v2: Lossless data compression with transformer. https://bellard.org/nncp/nncp_v2.1.pdf, 2021
[3] Making compression algorithms for unicode text. In arXiv:1701.04047, 2017
[4] A two-stage data compression method for real-time database. In ICSEM 2012

难题5：[高性能] 数据库智能基数估计算法

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

价值：金融、电信等关键行业对数据库性能有很高要求，基数估计精度直接影响查询性能，可显著提升GaussDB等的竞争力

传统基数估计方法

传统方法基于采样/统计信息和多列独立性假设进行基数估计，不具有普适性，对于不满足以下(1)或(2)的场景，误差可达10^3甚至更多：

（1）要求业务场景满足以下**基本假设**：

- 独立性假设（Independence）：$P(A+B) = P(A) + P(B) - P(AB)$
  $P(AB) = P(A)P(B)$
- 均匀性假设（Uniformity）：数据服从均匀分布；
- 包容性假设（Inclusion）：估计时假设满足查询谓词的数据存在；

（2）需要**样本**具有代表性，基于采样的**统计信息**模型、维度准确：

- 采样：通常针对属性的域进行伯努利采样或蓄水池采样；
- 统计信息：统计信息包括直方图、NDV、元组数量等；

技术现状

- **查询驱动（Query-Driven）**：从历史查询结果中总结规律，改善估计精度；
  例如：MSCN[1]，基于卷积神经网络，输入为表和关联查询，输出为基数估计结果；
  局限：训练耗时较长，对训练数据的要求较高，难泛化，难以融入数据库内核；
- **数据驱动（Data-Driven）**：从数据中总结概率分布规律，替代传统直方图模型；
  例如：基于深度自回归模型的NARU[2]和基于贝叶斯网络的概率图模型[3]；
  局限：数据分布发生变化时**必须重新训练**，支持的**谓词较少**，处理复杂查询时资源消耗较高；

技术挑战

- **轻量化问题：**现有AI基数估计模型对算力和资源要求高，与数据库内核融合困难，难以商用；
- **自适应问题：**对业务场景的普适性较差，无法自动识别优势场景精准发力；

技术诉求

- **具体诉求和目标**：构建算力要求低，资源消耗少，估计结果精准的**轻量型**智能基数估计算法；能与数据库内核融合，在复杂查询场景中给出比传统基数估计更准确的结果
- **精度指标**：与传统基数估计相比，估计精度(Q-error指标)提高2倍以上，使TPC-H、TPC-DS、Join-Order-Benchmark等常见测试集上的查询用时端到端降低20%以上
- **效率指标**: 智能基数估计算法的时间和资源开销与传统基数估计方法相差不超过1个数量级

参考文献：
[1] Learned Cardinalities: Estimating Correlated Joins with Deep Learning. In CIDR 2019.
[2] Deep Unsupervised Cardinality Estimation. In VLDB 2019.
[3] Efficiently adapting graphical models for selectivity estimation. In VLDB Journal 2013.