双剑合璧：openGauss的AI4DB与DB4AI

AI特性是openGauss的关键特性之一，在过去的版本中，openGauss先后开源了AI参数自调优和慢SQL发现功能，获得了广大开发者和用户的关注。为了进一步构筑完善的AI能力，在openGauss的新版本中，我们针对AI4DB和DB4AI两大特性，增加了全新的功能。

AI4DB特性旨在通过AI技术赋能openGauss，使用户操作和管理数据库更加简便，为用户提供端到端的自运维和自优化套件。在新版本中，我们增加了**数据库智能索引推荐**和**数据库监控与异常检测**两项功能。DB4AI特性是在数据库内提供AI计算能力，通过数据库内置AI算法为用户提供普惠AI能力，新版本中增加了**数据库内机器学习算法框架DeepSQL**. 下面将依次详细介绍前述新功能。

一、AI4DB

**数据库智能索引推荐**

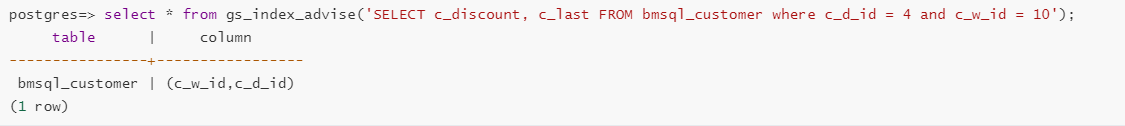
在大型关系型数据库中，索引的设计和优化对SQL语句的执行效率至关重要。一直以来，数据库管理人员往往基于以往的知识和经验，人工设计和调整索引。这样消耗了大量的时间和人力，同时人工设计也往往不能确保索引是最优的。

openGauss提供了智能索引推荐功能，该功能将索引设计的流程自动化、标准化，可针对单条查询语句或工作负载推荐最优的索引，提升作业效率，减少数据库管理人员的运维操作。openGauss的智能索引推荐功能可覆盖多种使用场景，它包含以下三个子特性：

1. **针对单条查询语句的索引推荐**

该特性可基于查询语句的语义信息和数据库的统计信息，对用户输入的单条查询语句生成推荐的索引。

用户可通过openGauss的系统函数“**gs\_index\_advise**”来方便地使用该功能，该功能支持单列索引和联合索引的推荐。具体示例如下，索引的推荐结果包括相应的表名和列名：

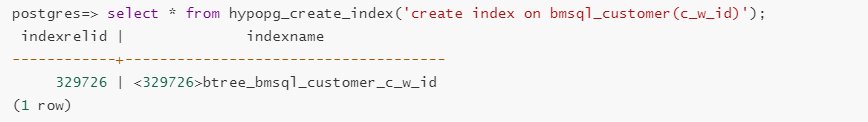


2. **虚拟索引**

该特性可模拟创建真实索引，同时避免创建真实索引所需的时间和空间开销，用户可通过优化器评估虚拟索引对指定查询语句的影响。

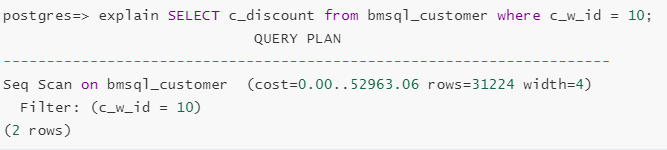
本特性提供了包含虚拟索引的创建、删除、性能和存储空间开销评估在内的一系列操作，用户可通过使用openGauss的系统函数灵活地操作虚拟索引。部分操作的示例如下：

1. 使用系统函数“hypopg\_create\_index”创建虚拟索引，输入参数为创建索引的SQL语句。

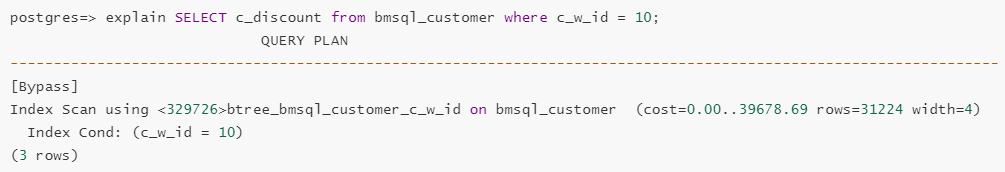


1. 通过对特定的查询语句执行EXPLAIN，用户可根据优化器给出的执行计划评估该索引的性能。

创建虚拟索引前的执行计划是：



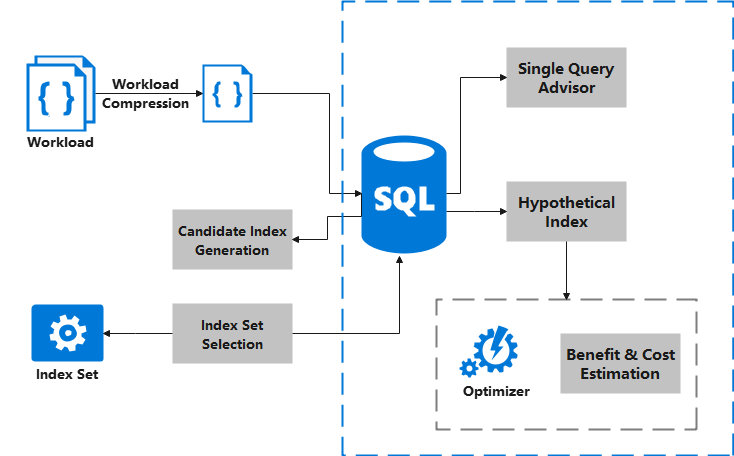
创建虚拟索引后的执行计划发生了改变，如下所示：



通过对比两个执行计划可以观察到，该索引预计会降低指定查询语句的执行时间，基于此结论，用户可考虑创建真实索引。

3. **基于工作负载的索引推荐**

将由多条DML语句组成的工作负载作为输入，算法可以推荐出一批建议创建的索引，利用这些索引可以优化整体的工作负载，该算法的流程如下图所示：



该特性将前述两个特性为基础，首先对工作负载进行压缩，过滤出具有代表性的一批SQL语句，然后通过单条查询语句的索引推荐功能，为每个SQL语句生成候选索引。最后，通过虚拟索引功能再进一步筛选，将对工作负载具有最大正向贡献的索引作为输出。

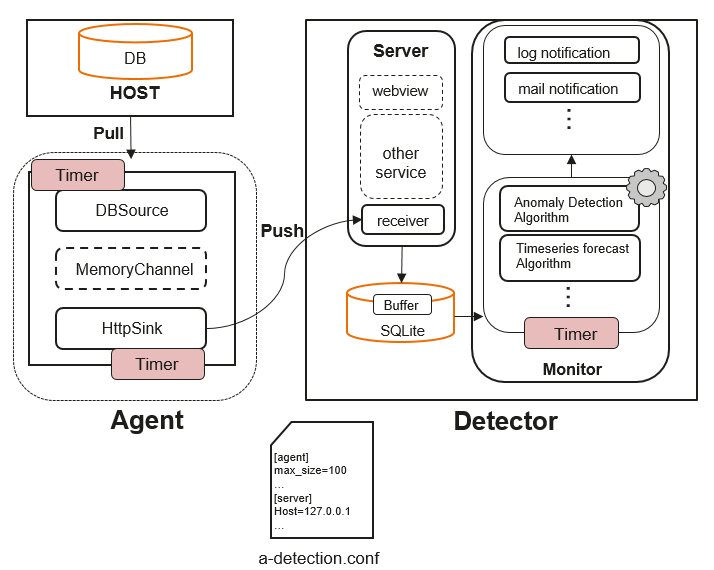
**数据库指标监控与异常检测**

数据库指标能够反映出数据库的健康状况，用户的异常操作或数据库性能的下降都可能导致数据库指标发生变化，因此对数据库的指标进行监控十分有必要。其价值主要体现在以下两个方面：

1. 可以帮助用户从多个视角了解数据库的运行状态，从而更好地进行数据库的规划。
2. 帮助用户提前发现数据库异常以及潜在的性能问题，并及时将情况反映给用户，以避免产生不必要的损失。

anomaly-detection是openGauss集成的AI工具，可用于数据库指标采集、预测以及对异常的监控与诊断。

anomaly-detection 的结构如下图所示：



该工具由Agent和Detector两部分组成，Agent是数据采集模块，负责收集数据库指标数据并将数据推送到Detector. Detector是异常检测模块，该模块有三个作用：① 收集Agent端推送的数据并进行本地化存储；② 对收集的指标数据进行异常检测；③ 将异常信息推送给用户。

目前该工具默认采集的数据库指标包括IO\_Read、 IO\_Write、CPU\_Usage、Memory\_Usage以及数据库所占磁盘空间disk\_space等。基于采集的数据，工具会预测指标未来变化趋势并进行异常检测，以实现磁盘空间不足预警、内存泄露预警、CPU资源消耗预警等功能，避免数据库出现异常而导致不必要的损失。

anomaly-detection具有一键部署、一键启动或关闭、指标预测等功能，使用简单方便。此外，用户可以根据业务场景的需要，快捷地添加新的监控参数或时序预测算法。

二、DB4AI

数据库管理系统通过构建有序的文件组织结构，可以方便地对数据记录进行增删改查。在AI领域，人们利用计算机提供的算力对数据进行分析和挖掘。数据的存储和计算是数据治理的关键。

在传统场景中，当数据使用者希望对数据库中存储的数据进行分析、训练时，通常需要将数据从存储系统中提取出来，缓存到内存中后利用Tensorflow、scikit-learn等Python的第三方包进行数据分析或模型训练。从开发角度看，这一流程并不高效。首先，该流程涉及Python语言、第三方机器学习包、数据库等领域，开发的技术栈会比较碎片化。其次，该流程的性能并不优秀，训练模型的数据往往是需要通过网络进行传输的，数据量大时会有较多的网络传输开销，没有实现数据的本地化计算。而且，开发人员的技术水准往往参差不齐，这将不能充分挖掘CPU或GPU的计算能力。在一些数据敏感领域，提取数据需要获取权限、脱敏等操作，这进一步带来了成本。所以将AI计算能力集成进数据库，利用数据库的算力进行计算就非常具有优势了。一方面可以实现数据的本地化计算，另一方面可以利用数据库的优化能力选择最佳的执行计划。最后，用户只需要一条SQL语句就可以实现比自行实现算法更快速的模型训练和预测能力。

DeepSQL兼容Apache MADlib生态，目前支持60多个常用算法。其中包括回归算法（例如线性回归，逻辑回归，随机森林等）、分类算法（比如KNN等）、聚类算法（比如K-means）等。除了基础的机器学习算法之外，还包括图相关算法，比如最短路径，图形直径等等方法；并且也支持数据处理方面方法（比如PCA），稀疏向量数据格式，统计学常用算法（比如协方差，Pearson系数计算等），训练集测试集分割方法，交叉验证方法等。

除了以上通过兼容MADLib获取的算法，DeepSQL还独家支持3个常用算法，分别是prophet、GBDT和XGBoost.

prophet基于时序数据分解、局部贝叶斯等实现的时序预测算法，由facebook开源，是工程场景中比较实用的时序预测算法，相比其他时序预测算法具有更快速、更准确、更鲁棒等优点，计算量远远比RNN神经网络少。

GBDT和XGBoost，同属于利用回归树去拟合残差的Boosting算法。

GBDT算法利用MADlib框架中的树模块设计完成算法。延续MADlib函数接口的风格，利用入参对模型的超参数进行设定。算法支持归回和分类任务，在默认情况下模型为回归模型。算法中选用均方误差(MSE)来计算每次迭代上一个基学习器的残差。对于回归树计算时每一个节点的划分是以最小化平方误差来选定分支策略。

对XGBoost算法集成后实现了gs\_boost模块，提供SQL-like接口，兼容MADlib风格，支持分类和回归任务。gs\_xgboost模块支持通过网格搜索来进行模型的超参数选择和模型评估。

以上便是openGauss最新开源的AI特性，欢迎大家使用和体验相关功能。如果大家有什么意见或建议，欢迎在开源社区进行反馈，我们将诚挚地聆听和参考每一位使用者的宝贵意见，这将是我们不断探索和努力的动力。相信在开发者和用户的共同努力下，openGauss与AI的融合将不断加强，进而为用户带来更智能、更卓越的服务。