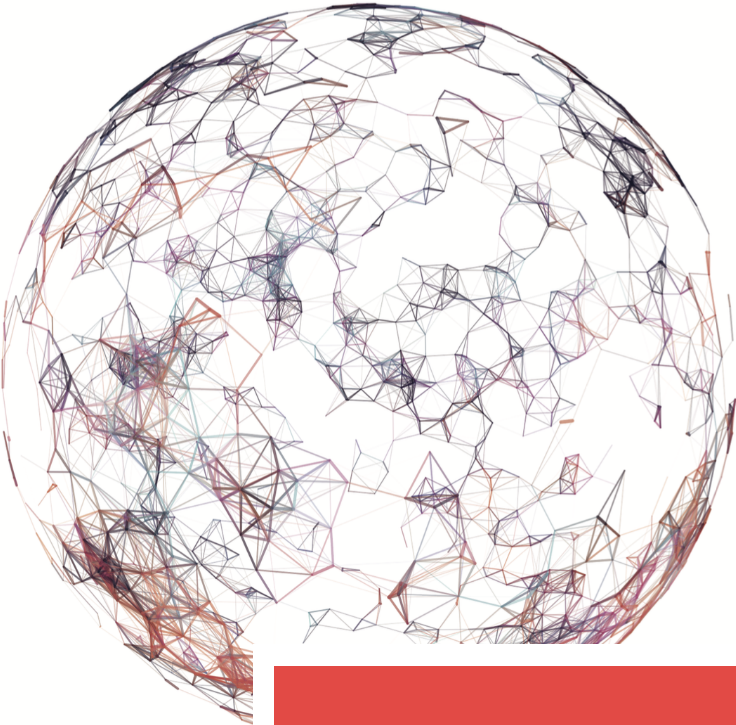
柏睿数据

产品月刊



柏睿2023年度产品盛典

第13期
2023年度

做中国的国际智能数据算力公司

AN INTERNATIONAL DATA + AI COMPUTING POWER COMPANY FROM CHINA

分布式内存数据库 RapidsDB

2月

产品更新

- 发布支持华为Kafka+kerberos认证功能的小版本更新
- 数据湖连接器基础功能完成，等待与4.4大版本一同更新
- 解决表中字段设置自增后，字段类型改变的问题
- 解决部分表结构影响SQL结果的问题

3月

产品进展

- Rapids Pipeline 内置功能升级。该管道可用于提取、转换和加载外部数据。此版本针对 Kafka 和 Hadoop 管道方面的功能进行了优化和提升。
- JDBC 驱动包新增定制功能，用于适配 LOAD DATA HDFS 语法，实现与 HDFS 数据文件的无缝对接。该驱动包已集成于 数据库的安装目录 ./drivers 中。

4月

V4.4更新发布

- 性能大幅度提升，RapidsDB V4.4 性能相比上个版本性能提升数倍。
- RapidsDB 4.4 实现了通过IMPXE连接器并行加载HDFS数据和导出数据到HDFS 中（兼容kerberos认证）。
- 大大提升在大数据体量下的稳定性。

5月

V4.4.0.c更新发布

RapidsDB4.4.0.c 在 RapidsDB4.4 的基础上新增加了 Iceberg 连接器。通过 Iceberg 连接器， RapidsDB 可以在基于 Iceberg 构建的数据湖之上进行数据分析，打通 RapidsDB 和数据湖，形 成湖仓一体。

- Iceberg 连接器主要有以下特性:
- 创建和删除 Iceberg 连接器，支撑存储数据在 HDFS 或者 S3 上;　在 Iceberg 连机器上创建普通表和分区表;
- 插入数据到 Iceberg 表;
- 删除整个 Iceberg 表的数据;
- 删除 Iceberg 表;
- 删除满足指定条件的行数据;
- 更新满足指定条件的行数据;
- 将其它连接器的数据导入到 Iceberg 连接器表中;　跨 Iceberg 连接器和其它数据源进行查询;

6月

产品更新：

1. RapidsDB 作为轻量化BI加速数据库产品完成在ARM架构麒麟操作系统下兼容性测试。
2. RapidsDB 完成对客户产品推广的验证性测试方案输出，可以为所有希望试用RapidsDB的客户提供一套简单的方案测试RapidsDB的基础功能和性能等特性。
3. RapidsDB研发团队完成对多种常见数据库语法兼容的框架性架构设计。
4. RapidsDB研发团队完成基于手动快照的功能完成自动化快照和宕机重启数据恢复功能。

产品培训：

国内外研发团队举行为期三天的Virtual Offsite Meeting 共同讨论公司核心产品技术的现状和未来。三天的分享的关于产品和技术的主题如下：

【Day 1】： 1. Rapids Lakehouse
2. Vector Database and Generative AI
【Day 2】： 1. SE performance test harness
2. Standard Benchmarking
3. Rapids Pilot
【Day 3】： 1. RapidsDB and LLM
2. Rapids SE Paging
3. Columnar storage: formats, compressions & trade-offs
4. A Survey on Advancing Query Optimizer 5.Rapids SE Roadmap

7月

产品更新：

1. 基于RapidsDB做存储的的 Rapids Vector DB的调研工作完成。
2. RapidsDB研发团队在可配置SQL语法框架设计上完成对Mysql语法的支持。
3. RapidsDB轻量化产品发布。

产品培训：

面向公司内部销售、售前、产品同事做新产品介绍，培训内容如下：

【7月26号】： 基于ARM平台的数据库产品介绍

8月

产品更新

1. RapidsDB完成数据湖连接器增强功能 schema演化、表分区演化、表属性查询、回溯查询、数据回滚能力
2. RapidsDB完成Moxe存储引擎的数据自动备份和恢复功能
3. RapidsDB完成地理空间支持
   a. 支持地理空间数据类型：Geometry、Geography
   b. 支持地理空间分析函数：ST_Area、ST_Length、ST_Distance、ST_IsEmpty、ST_IsValid、ST_IsSimple、ST_Contains、ST_Intersects、ST_Crosses、ST_Disjoint、ST_Equals、ST_Overlaps、ST_Touches、ST_Within

9-10月

产品更新

1. RapidsDB 完成分布式向量引擎的基本功能，能够实现分布式索引创建、相识度查询
2. RapidsDB 完成对JSON数据导入的基础功能实现
3. RapidsDB Geo-spatial 完成SRID初步支持

RPDSQL更新进展

1. RPDSQL V4.4版本新增支持ipv6，并通过本地、远程、jdbc连接测试
2. RPDSQL V4.4版本已成功部署在苏研生产环境中

11月

产品更新

RPDSQL数据库：
1.在SELECT ... INTO HDFS语法中新增将分区数据小文件合并成一个数据文件的功能。

RapidsDB
1.RapidsDB的管理工具Pilot完成，提供方便和强大的集群安装、管理、运维、监控功能。
2.RapidsDB库内推理方面完成库内推理框架设计。
3.自动化测试工具SETH更新，提供并行执行测试功能，加速测试时间

12月

产品更新

RapidsSE完成v1.2.0-dev1的开发版本，主要变更如下：

1. 使用SE Shell替换SE Prompt
2. 新的JDBC驱动程序，通过TCP/IP套接字连接与RapidsSE通信，使用wireline protocol进行通信。SE Shell使用此驱动程序
3. 不再用捆绑的Tiny-C（tcc）编译器来编译为查询生成的动态代码。使用gcc或clang实现
4. 查询动态生成代码的代码合成器更新到v2，以便更好与SQL优化程序和外部编译器配合
5. SQL优化器继续接收修复、改进和性能优化，以处理类似TPC-H的查询
6. 对SQL警告的支持已添加到SE、JDBC驱动程序和SE Shell中

柏睿大数据平台 Rapids Lizard

2月

产品Q1计划

- 1月：完成架构设计、页面设计和Demo版本开发
- 2月：搭建内部生产环境，安装部署测试，完成项目立项评审
- 3月：产品持续开发，准备对内宣传材料，进行内部培训

3月

产品Q2计划

- 全平台统一登录认证、权限管理
- 平台任务调度系统与数据集成、湖仓一体、数据治理、AI模型预测模块对接
- 内部API对接汇总、DEMO演示环境对内发布
- 全平台UI统一、优化，630达到发布标准
- 销售培训、技术培训、报价体系确认

4月

产品进展更新

- 展开北京、武汉、合肥、广州、上海各研发团队间的深度合作，共同打造平台化产品
- 大数据平台集成Transformer产品能力，进一步提升平台数据同步能力
- 实现全平台的统一登录认证和权限管理
- 实现统一数据源管理，为平台各组件开放中央接口
- 全面升级UI，统一大数据平台整体风格，预计6月低达到发布标准

5月

产品进展更新

- 展开北京、武汉、合肥、广州、上海各研发团队间的深度合作，共同打造平台化产品
- 实现计算模块和Transformer的调度引擎统一
- 实现权限管理模块中数据库层面的权限设计，库表按权限管理
- 实现湖仓一体模块中数据湖的业务存储结构设计，分层、分类、自助
- 监控运维模块实现集成AIworkflow

6月

产品V1.0发布

柏睿大数据平台 Rapids Lizard 是一款主要以柏睿核心自研产品及组件集成的，提供数据采集、数据存储、分析、治理与可视化功能的通用实时大数据基础平台与解决方案。以松耦合多层级架构，提供高性能、低代码、全流程、模块化的体验，赋能行业特色应用，为各行业企业级用户提供一站式数据资产管理和规模化数据智能基础建设服务。

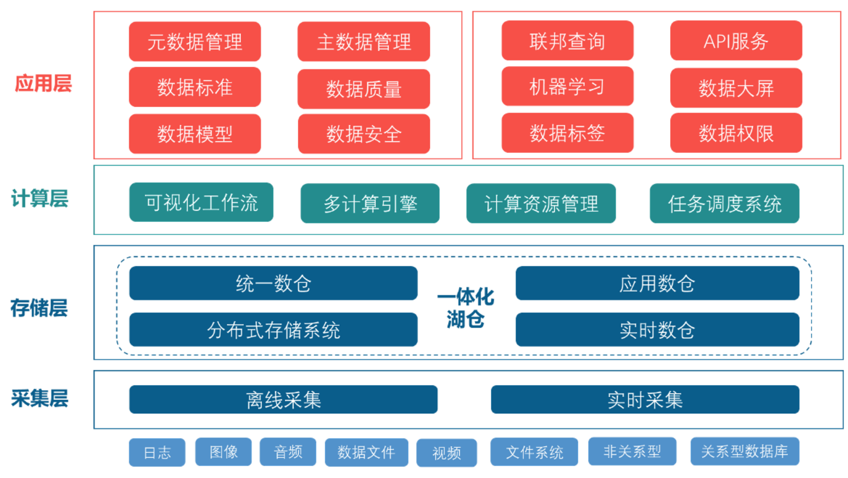
产品架构

核心产品：

- 全内存分布式数据库RapidsDB；
- 低代码机器学习建模平台Rapids AIworkflow；

产品组件部分：

- 一体化流湖仓Rapids Lakehouse；
- 拖拽式数据集成组件Rapids Fusion；
- 全流程任务调度系统Rapids Scheduler；
- 数据治理组件Rapids Datakeeper；
- 可视化数据展示组件Rapids BI；



7月

产品进展

1.大数据平台演示环境开放：192.168.30.47

- 账号admin密码123456

2.大数据平台UI升级中

3.租户、用户管理功能开发测试，为项目落地做准备

8月

产品进展

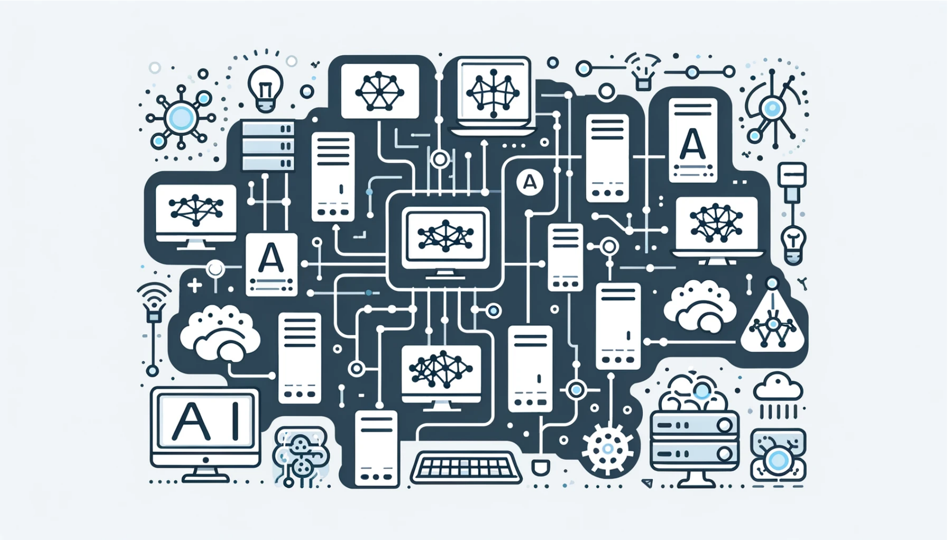
1. 确认前端框架
2. 大数据平台演示环境开放：192.168.30.47，账号admin密码123456
3. 数据计算中心专利申请：交底书内审中

9-10月

产品进展

1. 大数据平台数据目录、数据权限、数据管理等功能初步页面设计完成
2. 大数据平台技术软文输出：
   - 《柏睿大数据平台集成一体化流湖仓助力企业构建面向AI大模型时代的数据能力》
   - 《Gartner发布2024年十大战略技术趋势》
   - 《柏睿大数据平台统一调度解决方案，实现数据资源智能调度》

柏睿大数据平台 Rapids AI



2-5月

产品进展更新

- AIworkflow v4.1完成功能开发，正在进行功能验证

1. 支持 ParallelAI, PyTorch, Tensorflow, XGBoost, sklearn, Prophet的建模框架
2. 支持建模任务的定时调度
3. 支持结构化和非结构化建模，支持构建深度学习模型

- 大语言模型+RapidsDB 演示案例，实现自然语言转SQL和图表

6月

产品进展更新

1. 柏睿数据大模型LLMOps发布，关注大语言模型的构建和运行。
2. 向量数据库Rapids VectorDB 发布，高效适配业务场景的行业大模型。
3. 完成知识问答demo准备,已对多个客户进行展示。
4. 完成自然语言转SQL语句demo准备,已在广州海洋局进行展示。

7月

产品进展更新

1. 发布Rapids AI产品白皮书。
2. 实现大模型本地微调，针对SQL生成进行优化。

8-10月

产品进展更新

- 利用AIworkflow构建了两个应用与机器人的案例
- 完成了YIT的AIworkflow使用培训
- 进行了Data Assistant 架构设计

11月

产品更新

【Data Assistant】

基于大模型和柏睿数据库的智能化数据分析工具，帮助用户通过自然语言进行数据查询和分析

1. 实现基于电力数据的智能问答
2. 对大模型进行微调，提升回答准确率和适用性

【AI4DB】

使用人工智能技术对数据查询优化引擎中的基数进行估计，提升优化器的优化准确率

1. 实现基于查询的基数估计模型
2. 实现数据库内推理的框架搭建

12月

产品更新

【AI4DB】

1. 构建数据库优化器中基数估计的模型

【AIworkflow】

1. 完成AIworkflow 在ARM上的适配
2. 输出AI和大模型在电力行业的产品方案及案例

柏睿流湖仓归档管理产品

2月

产品实施拓扑

5月

产品进展更新

蓝光归档管理平台已适配申威架构

6-7月

流湖仓归档管理系统与申威处理器完成兼容性互认证

北京柏睿数据技术股份有限公司（简称“柏睿数据”）和中电科申泰信息科技有限公司（简称“电科申泰”）经联合性严格测试得出：柏睿流湖仓归档管理系统在银河麒麟高级服务器操作系统V10和电科申泰产品申威3231处理器环境下，产品相互兼容，功能情况良好，满足国产化安全可控要求。

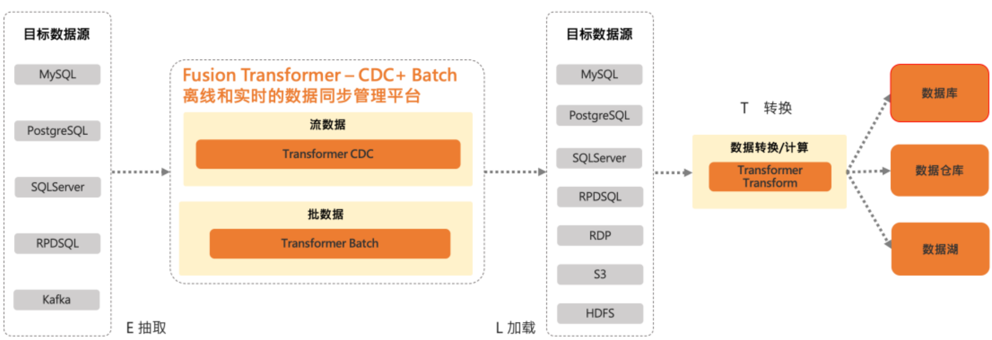
数据同步工具Rapids Transformer

4月

V3.0更新发布

- CDC 同步：捕获源系统中的数据变化，包括新增、修改和删除等操作，将这些变化实时地同步到目标系统中。
- 批处理同步：批处理（Batch）是 Transformer 的 ELT 数据同步，旨在帮助用户轻松地在不同数 据源之间同步结构化和半结构化数据。
- Web 管理界面：Transformer 提供 Web 管理界面，包括用户管理、数据源管理、Pipeline 管理、 Schema 管理功能。
- 高可用：具有完善的 HA(High Availability)机制。

产品架构图



5月

V3.1进展更新

- CDC Datalake Sink (Iceberg)支持，利用Spark + Iceberg技术实现，实现初始化同步和CDC同步
- 新增数据源支持：RapidsDB数据源、本地文件源
- 增加后台关键监控服务

单机版交易型数据库 Rapids RPDSQL

5月

V2.0 产品发布

- Rapids RPDSQL是北京柏睿数据技术股份有限公司自主研发的具有自主知识产权的数据库产品解决方案，此数据库产品**内置SQL、KV双引擎服务**,在此基础之上分别实现**兼容MySQL协议的关系型数据库RPDSQL-TP和Redis协议的非关系型数据库RPDSQL-KV**，使得Rapids RPDSQL为您提供了一种全新的数据管理体验。
- Rapids RPDSQL的独特之处在于其集成了两种不同的数据库技术，将关系型和非关系型数据库的优势融合于一体。通过兼容MySQL协议，Rapids RPDSQL可以与现有的MySQL应用程序完美协同工作，无需任何修改。同时，借助Redis协议的支持，Rapids RPDSQL能够处理结构化和半结构化数据，为您提供更大的灵活性和多样性。此外Rapids RPDSQL已经**和海光、申威，鲲鹏完成国产化适配化，支持国产化替换**。
- 总之，Rapids RPDSQL 是**支持国产化多平台、兼容MYSQL协议、REDIS协议、国密算法及可以主备模式部署安装的交易型数据库**。

培训介绍

- 针对高端客户，数据库产品线未来将附带标准化培训
- 培训讲师由柏睿数据库专家负责，培训结束提供培训证书
- 培训课程安排为3天，具体安排如下

培训内容&安排

- 第1天：产品宣讲（产品架构、技术特性等），安装部署和产品使用（建库建表、数据导入导出、高级函数使用等）
- 第2天：应用开发（Java、Python开发等）及数据源对接
- 第3天：日常运维（集群管理、数据库管理监控等），性能优化，经典案例介绍分析

柏睿数据治理平台 Rapids Datakeeper

6月

Rapids Datakeeper是专为数据治理领域开发的自主数据管理系统，也是柏睿大数据平台的关键组件之一。通过该平台，企业可以便捷地管理和掌握组织的数据资源，实现数据管理规范化和优化，最大程度地实现数据资产的价值。

V1.0 产品发布

数据质量管理系统上线，该系统是柏睿数据治理平台重要的组成组件，通过质检任务的方式进行企业数据质量的跟踪和监控，帮助组织有效管理和改进数据质量问题。主要功能包括：

1. 可创建和管理数据质检任务，并支持按周期自动执行任务。
2. 提供一批常用的质检规则模板，同时支持用户创建符合业务需求的自定义规则。
3. 支持查看任务结果和执行日志，帮助用户高效处理异常问题。
4. 提供统一的数据质量看板，方便管理员集中监控任务执行情况和异常状况。

发布文档

- 产品安装手册
- 产品使用手册
- 产品运维手册

7月

V1.2 产品发布

数据质量管理系统新版本上线，本次产品更新主要按照DCMM标准的四大环节：质量需求、质量检测、质量分析和质量提升，对我们的产品进行重要改进。涉及内容包括：

1. 新增产品首页作为所有模块的统一入口，集中化展示所有业务模块，使其更突显我们系统按照DCMM标准运行的独特优势。

2. 新增了四个功能模块：数据质量分析报告、数据质量提升报告、质量知识库和质量活动建设，这些功能更全面地覆盖了数据质量管理的业务场景。

3. 对质量规则管理、质检任务和质检结果的页面链接逻辑进行了优化，使其更契合当前的框架需求。

4. 为了跃进用户体验，我们对当前的数据质量系统进行了易用性等方面的改进优化。

存算一体机HyperCube

7月

产品进展更新

1. 数据虚拟化DVM已支持arm docker环境中运行。
2. 轻量化一体机产品发布。面向中小企业实时分析或 BI 加速场景，小型交易型数据库使用场景。

产品培训：

面向公司内部销售、售前、产品同事做新产品介绍

8-10月

产品进展更新

基于RPDSQL的分布式向量数据库已可以为客户提供测试

11月

- 由北京市经济和信息化局、朝阳区人民政府指导，华为、鲲鹏联创主办，柏睿数据等协办的“山高海阔·共赢先机——鲲鹏一体机技术应用大会”在北京成功举办。会上，柏睿数据与华鲲振宇、鲲鹏联创正式签署战略合作协议，并联合发布整合各方软硬件优势打造的核心产品——“智算一体机”

柏睿交易型数据库 TXDB

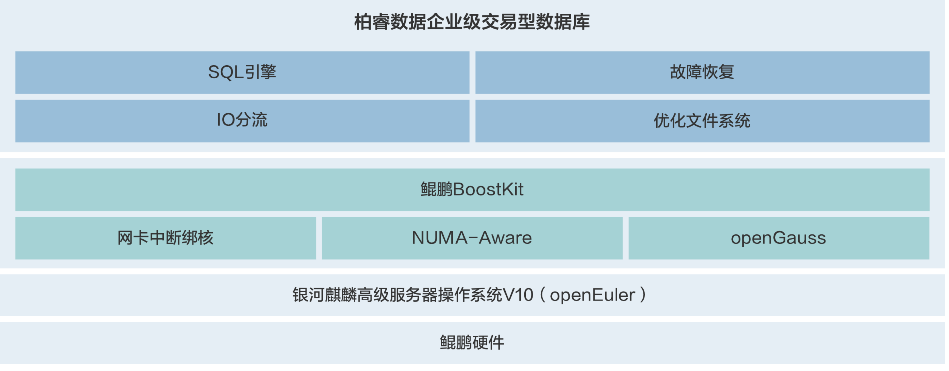
8月

TXDB全面完成国产信创兼容性互认证

柏睿数据基于鲲鹏底座推出企业级交易型数据库，支持多种存储模式，复合业务场景，提供PB级数据处理能力， 同时具备满足7×24小时运行所需的数据保护、节点互备、在线维护等功能属性。既可为超大规模数据管理提供高性价 比的通用计算平台，解决企业级用户的数据处理难题;又可支撑各类数据仓库系统、BI系统和决策系统，统一为上层应用提供数据应用服务。

柏睿数据在和多家国产信创公司联合严格测试后，得出结论：

- TXDB在兆芯处理器和银河麒麟操作系统V10平台上完成兼容认证，功能良好
- TXDB与统信服务器操作系统V20完成兼容认证，成为统信产品生态伙伴
- TXDB与飞腾FT2000+/64 +银河麒麟服务器操作系统V10完成兼容认证
- TXDB与华为技术有限公司Kunpeng 920 Taishan 200完成兼容认证
- TXDB与银河麒麟高级服务器操作系统V10完成兼容认证
- TXDB与海光CPU7000 、5000、3000 系列产品完成兼容性认证
- TXDB为强制性国家标准GB 18030-2022《信息技术 中文编码字符集》首批测试通过产品



柏睿向量数据库Rapids VectorDB

11月

显著降低AIGC“入门”成本，柏睿向量数据库助力打造高效易用的向量数据引擎

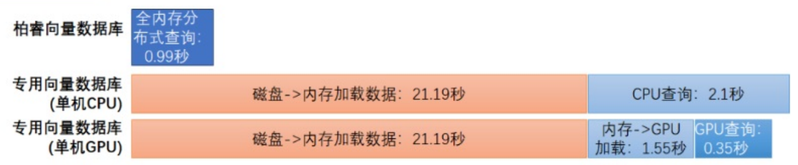
- Rapids VectorDB 在检索过程中无需使用 CPU / GPU，即可实现比在 CPU / GPU 上运行的Faiss高 20+ 倍的整体性能

测试拓扑

配置环境

	主机类型	主机配置	GPU
柏睿向量数据库	虚拟机	计算节点：8vCPU,16G内存 *1 存储节点：8vCPU,32G内存 *3	无
专用向量数据库	物理机	CPU：32线程(IntelE5-2620 v4 2.10GHz*2)，128G内存 *1	NVIDIA RTX 3060(12G显存)

测试结果



根据性能测试结果可知，**Rapids VectorDB 仅需0.99秒**即可返回查询结果，而**运行在 CPU 上的专用向量数据库则用23.29秒**返回结果，**运行在 GPU 上的专用向量数据库则需23.09秒**。可见，Rapids VectorDB 在向量数据检索查询方面不仅具备**显著的性能优势**，还具有**极高性价比**。

12月

基于柏睿向量数据库，一分钟构建“以图搜图”功能

如今我们日常使用的百度、搜狗等通用搜索引擎均提供了相似图像检索功能。淘宝、京东等电商平台也利用自己庞大的商品图像库开发了垂直领域内的图像检索功能来满足消费者们不易用文字描述的商品搜索需求。虽然图像检索技术已广泛应用于生活，但目前这一技术还未完全成熟。比如传统使用机器学习方法，实现以图搜图的能力，流程与实现比较复杂。

使用柏睿向量数据库的向量相似搜索功能，在不需要提前AI训练模型，仅使用通用图像识别算法，就能实现简单的“以图搜图”功能。除去将图片数据Embedding入库的过程，基本1分钟就能构建“以图搜图”。

搜索图

VectorDB搜索

搜出与搜索图类似的图片

示例代码

```
from rapidsvdb_client import RapidsClient
import rapidsvdb_client.embedding_functions as em
import numpy as np
from PIL import Image

client = RapidsClient(host='192.168.31.62', port=3456)
#连接到柏睿向量数据库

embedding_function = em.OpenCLIP_ViTB32_EmbeddingFunction()
collection = client.get_or_create_collection(
    name="rapidsvdb_pic_db",
    embedding_function=embedding_function)
#使用测试图片向量数据集

query_image = np.array(Image.open("search1.jpeg"))
#定义要查询的图片search1.jpeg
retrieved = collection.query(query_images=[query_image], include=[uris], n_results=4)
#从柏睿向量数据库中查询与search1.jpeg相似的4张图的存储位置，其中uris是指图片存储位置
print (retrieved)
#显示查询结果
```

柏睿数据

产品系列月刊

柏睿数据

北京柏睿数据技术股份有限公司

Borrui Data Technology (Beijing) Co. Ltd.

北京市朝阳区利泽西街6号院东湖国际中心A座7层