

Data Computing for New Discoveries

数 据 计 算 ， 只 为 新 发 现

OpenPie

πDataCS

赋能工业软件创新与实践

吴疆

拓数派产品社区总监

CONTENTS

1 拓数派介绍

2 πDataCS介绍

3 πDataCS应用案例

4 总结

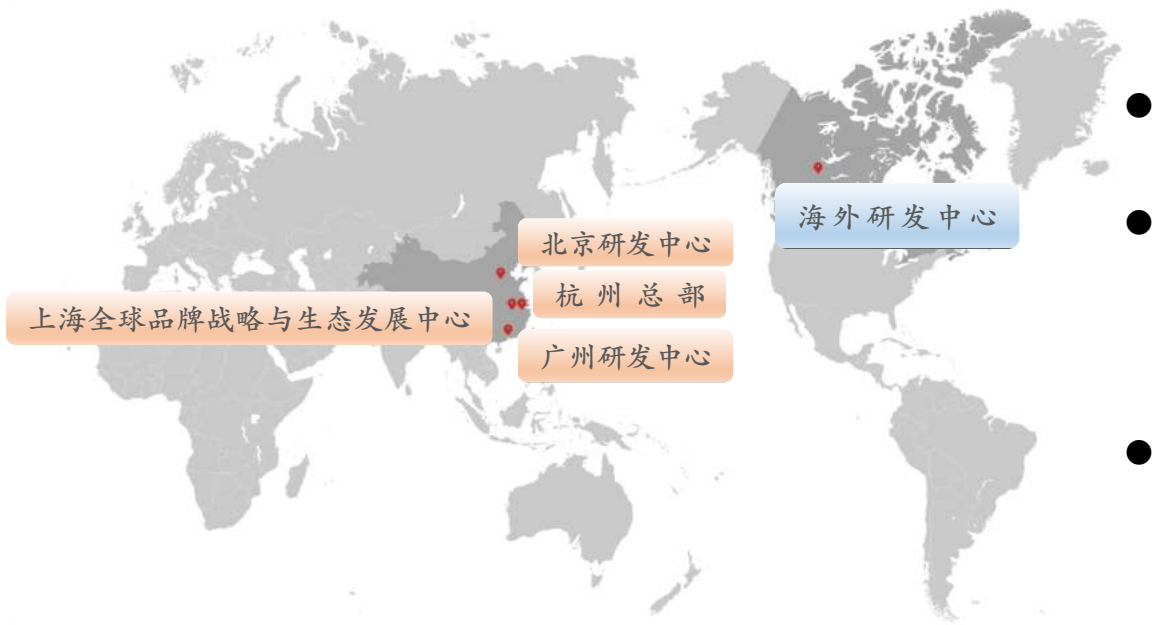


PART 01

拓数派介绍

企业简介

拓数派中国总部与全球分支机构



杭州拓数派科技发展有限公司，简称“OpenPie”

- OpenPie是立足于国内的基础数据计算领域**高科技创新机构**；
- 拥有强大的**数据库内核研发团队**、**数据科学团队**和**数字化转型团队**；
- 国内**虚拟数仓和eMPP技术提出者**，不断在数据计算引擎方向进行创新，**全面拥抱AI技术趋势**。

企业文化

愿景：成为立足中国基础数据计算领域的全球顶级高科技创新机构

价值观：以人为本、开放创新、拥抱变化、诚信正直

使命：数据计算，只为新发现

@2024 OpenPie. All rights reserved.

国际顶级创始团队



冯雷（Ray Von）
创始人&CEO

- 浙江物理奥赛银牌得主
- 北京大学物理经济学双学位
- 卡内基梅隆大学（CMU）硕士学位，校友会主席
- 原Pivotal/Greenplum（中国）创始人



陆公瑜（Brian Lu）
合伙人&COO

- 英国约克大学
- 15年+产品生态建设和运营管理经验
- 原Greenplum中国社区发起人
- Greenplum社区从0到万



郭罡（Paul Guo）
合伙人&CTO

- 中国科技大学少年班
- 18年+底层基础软件领域开发经验
- 原Greenplum首席内核架构师
- Apache HAWQ PMC成员

原Pivotal中国领团队和原Greenplum产品及社区核心成员均已加入拓数派。成员毕业自CMU、北大、清华和科大少年班等顶级学术机构，并有成员获有奥赛奖牌。







@2024 OpenPie. All rights reserved.

蓬勃发展的拓数派 驱动数据计算时代的到来

2021 拓数派 正式成立

- Day-1 **当天即获得头部产业基金天使轮投资 成为Day-1准独角兽**
- 7月 成立**杭州总部、北京研发中心、上海全球品牌战略与生态发展中心**
- 9月 杭州萧山区政府“一事一议”支持政策获批
- 11月 与中国人民大学成立实习基地，**打造中国的云原生数据库世界级智力高地**

2022 极速进入成长期

- 4月 获得**元禾重元**和**东吴证券**Pre-A轮投资 **标志着企业进入快速成长期**
- 6月 上榜 **EqualOcean 2022年源自中国值得关注的新锐全球化科技品牌**
- 7月 PieCloudDB云原生数据**库存算分离等关键技术打造完成引领数据库行业全面走向云时代**
- 8月 成立**广州研发中心**
- 9月 成立**加拿大研发中心**
- 10月 2022拓数派产品发布会 **云原生数仓PieCloudDB 社区版与商业版正式发布**
- 与**东吴证券**在**数仓虚拟化和信创领域**展开试点合作
- 12月 创始人冯雷**再度荣登数字商业周刊“年度智造中国商业领袖”**

2023 引领数据计算时代到来

- 3月 2023拓数派战略暨新产品发布会 重磅发布**πDataCS数据计算系统首个计算引擎PieCloudDB云原生虚拟数仓**
- 拓数派基于**阿里云**构建公共云数据仓库服务 **正式上线**
- 4月 冯雷被评为杭州市所有的独角兽和准独角兽企业中**唯一“年度创业人物”**
- **打造大模型时代 立身中国的世界级团队**
- 7月 拓数派数据计算引擎PieCloudDB虚拟数仓**再获信创认可首家以虚拟数仓通过信通院/可信AP数据库评测**
- 拓数派**再次携WAIC创建智能驾驶科技分会**
- 8月 拓数派入选中国信通院“铸基计划”「高质量数字化转型产品及服务全景图」
- 10月 拓数派上榜**《毕马威中国-未来独角兽 人工智能创新双30企业（长江三角洲地区）》**
- 2023拓数派年度技术论坛 **拓数派大模型数据计算系统正式亮相，让AI模型更大更快**

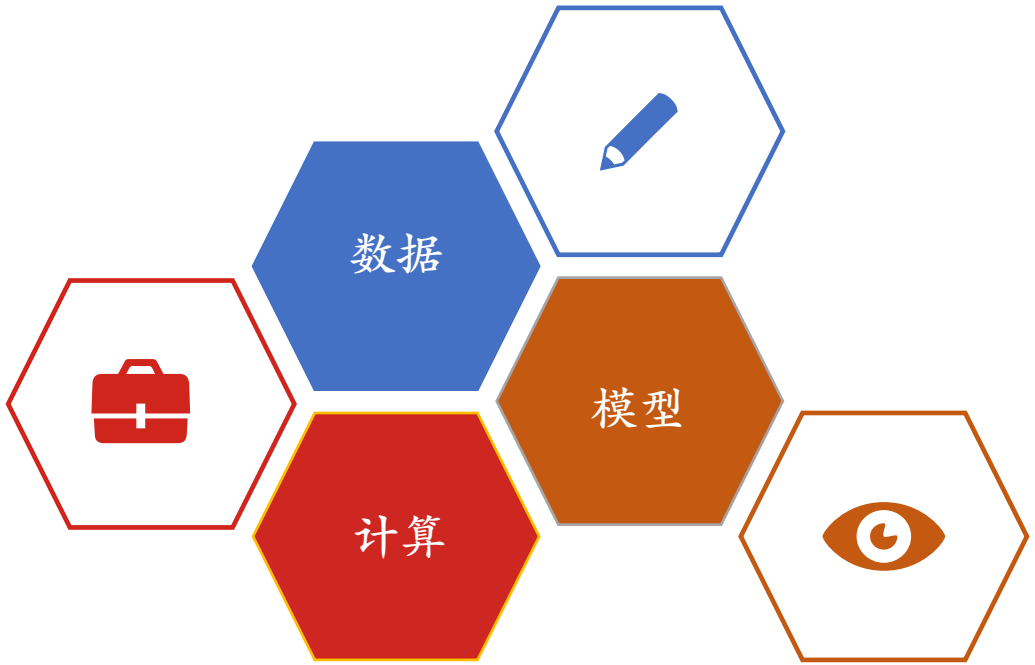
@2024 OpenPie. All rights reserved.

PART 02

πDataCS介绍

πDataCS的产品理念及定位

大模型数据计算系统，以云原生技术重构数据存储和计算，一份数据，多引擎数据计算，AI数学模型、数据和计算三者互为增强，全面升级大数据系统至大模型时代，赋能行业AI场景应用。



1. 灵活扩展的数据引擎，支持关系型数据库SQL、Spark/Flink等流批一体处理、LLM的向量数据库以及GIS地理数据库等。

2. 具备整体数据平台方案，支持多模数据处理（结构化、半结构化以及非结构化数据），实现数据共享和分析。

3. 软件优化＋新硬件（FPGA）加速，实现数据全链路的性能飞跃，让数据存储、SQL查询、向量计算以及机器学习等能力全面升级。

@2024 OpenPie. All rights reserved.

πDataCS 大模型数据计算系统

OpenPie

πCloudDB
虚拟数仓服务HTAP | 点查

πCloudVector
向量数据计算

πCloudML
(大模型)机器学习

…
大模型训练…

πFPGA 智能新硬件技术
数据存储 | 虚拟数仓 | 特定领域（如神经网络）

Mundo元数据管理系统 统一Catalog

自研简墨存储
统一数据格式 | 一份数据多引擎计算 | 兼容主流云存储格式和协议

阿里云　腾讯云　私有云 vmware openstack

- 全新云原生架构「一份数据，多引擎计算」弹性扩缩容，打破数据孤岛
- 全面支持HTAP 湖仓一体和向量计算
- 原生支持数据治理，三权分离
- 支持大语言基础模型和私域数据结合开发垂直应用
- CMU战略合作大模型基础理论，训练合作一事一议

@2024 OpenPie. All rights reserved.

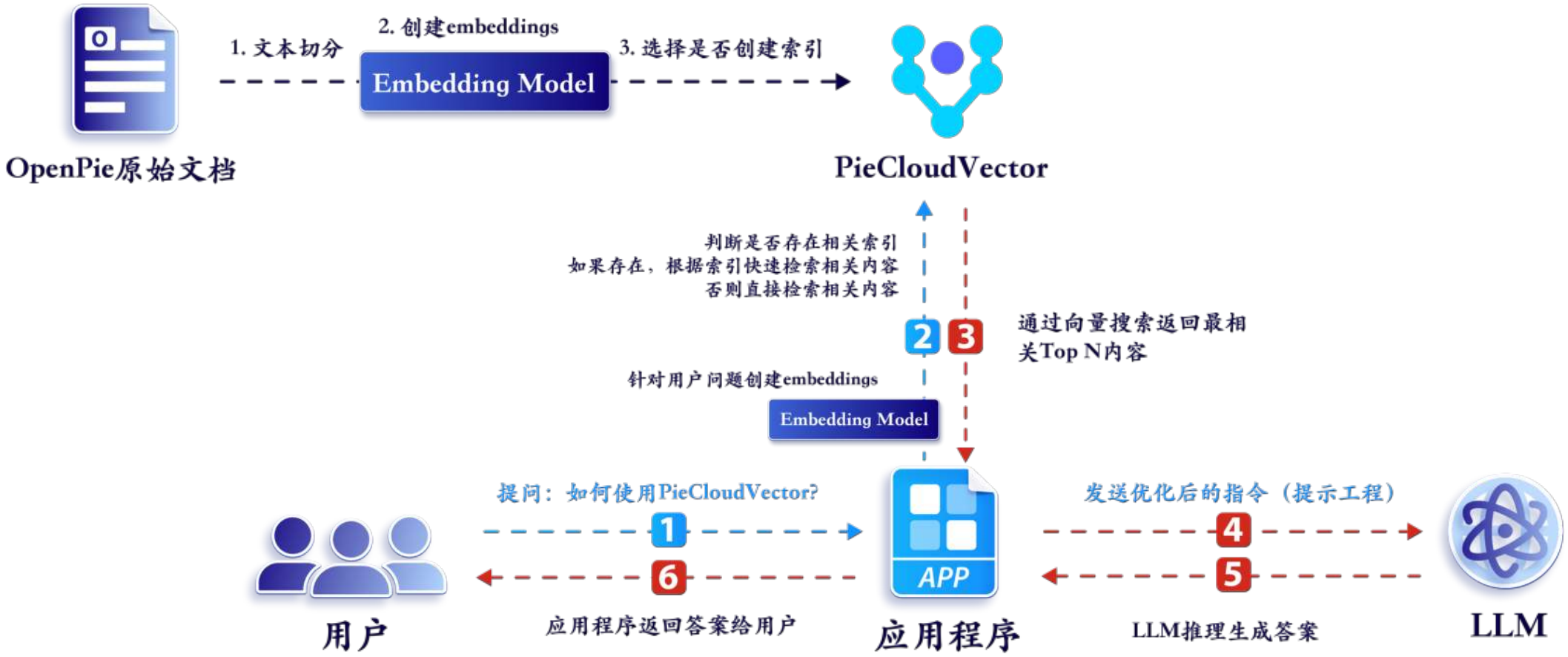
πDataCS 优势1： 全面升级Hadoop大数据和Greenplum数仓至云原生数据平台

πDataCS	Hadoop
大模型数据计算系统，以云原生技术重构数据存储和计算，一份数据，多引擎数据计算。主要解决海量数据的存储和实时计算问题，具备湖仓一体化的能力，用户可根据实际情况去选择合适的数据计算引擎。	是一个存储系统+计算框架的软件框架。主要解决海量数据存储与计算的问题，是大数据技术中的基石。让用户可以在不了解分布式底层细节的情况下，开发分布式程序，以一种可靠、高效、可伸缩的方式进行数据处理。
灵活可扩展的插件式引擎，组件少而精（All in One），提供3种计算引擎、1种Lakehouse引擎，1个硬件加速器，应对各种场景的计算，同时也兼容Spark、Flink等计算任务，保留用户的使用习惯。	组件很多，常见约30个，基础组件包括：Hadoop通用功能库、HDFS、MapReduce以及YARN 。可视化管理是Ambari，其他的计算引擎、列存数据库等都需要额外的组件，应对不同的场景需要安装对应的组件和依赖。
生态完善，支持主流的开发语言和数据科学工具，支持多模数据处理（结构化、半结构化以及非结构化），提供标准的SQL接口和API，完成各种复杂场景的数据处理，业务开发周期短，现存的代码基本可以无缝迁移和复用。	生态丰富，对结构化数据、半结构化数据以及非结构化数据都支持，可以很好的完成各种业务场景的数据处理需求。但是对于开发的要求比较高，需要掌握多种组件的不同使用方法，业务开发周期会比较久。
国内自主研发，具备社区版、商业版以及云SaaS服务，与国产软硬件完美兼容，属于信创产业。支持公有云，可充分利用云资源的弹性能力。组件少，提供可视化管理平台，运维管理简单，通过短暂学习即可掌握，无论是原厂商还是生态合作伙伴都可以直接提供技术服务，7 * 24的安心保障。	国外开源软件，与国产软硬件兼容性差，不符合信创趋势，也无法直接利用云资源的弹性能力。组件太多，导致集群部署和后期运维管理很麻烦，市场上相关人才储备量不多，技术兜底依赖于Cloudera，国内第三方公司主要是基础运维和开发为主。

@2024 OpenPie. All rights reserved.

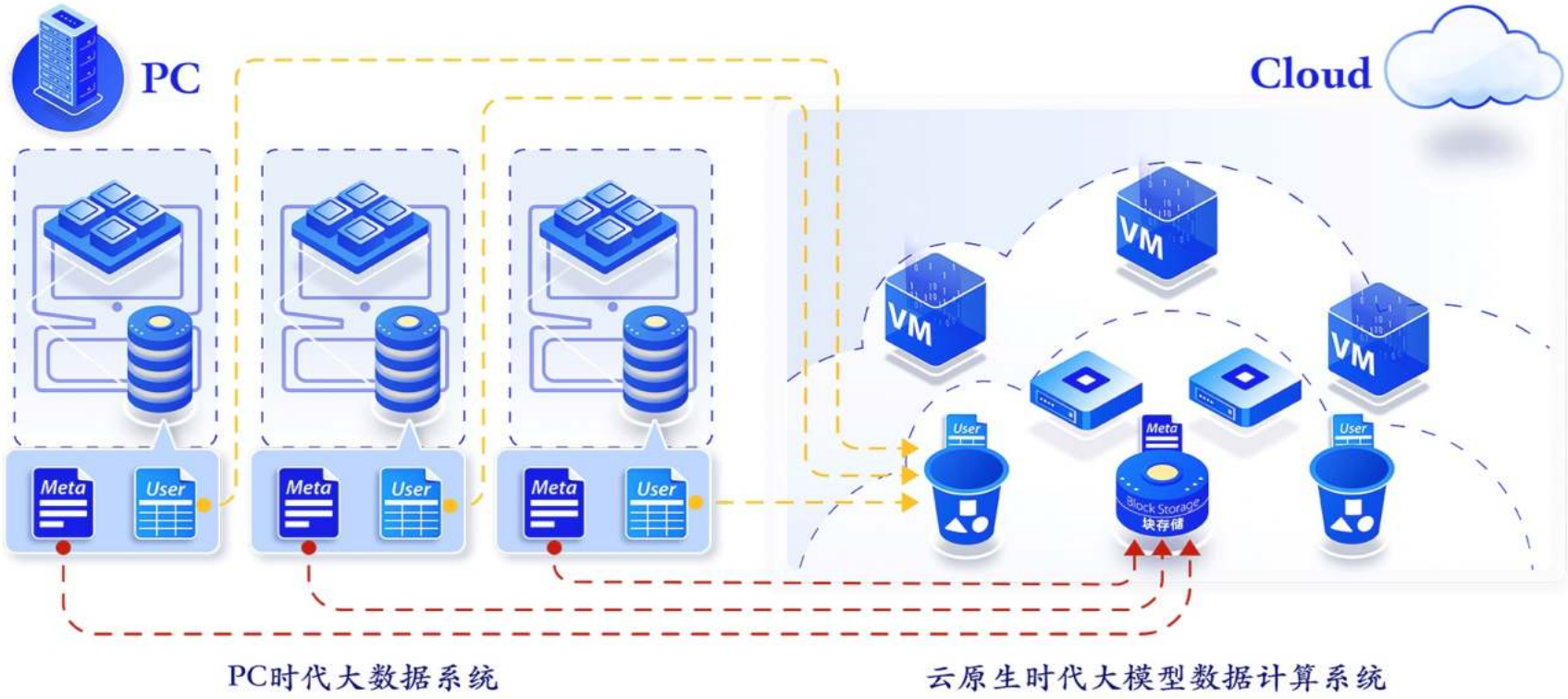
πDataCS优势2：全面支持大语言基础模型和私域数据结合做垂直应用

PieCloudVector与LLM在私域知识库领域的应用实践路径



@2024 OpenPie. All rights reserved.

πDataCS优势3 ： 云原生下eMPP计算引擎全面颠覆MPP技术，大模型数据计算新范式





证书号第5775327号

发明专利证书

发明名称：一种元数据访问方法、装置和存储介质

发明人：王溟舟;杨峻峰;冯雷

专利号：ZL 2022 1 1418015.0

专利申请日：2022年11月14日

专利权人：杭州拓数派科技发展有限公司

地址：311200 浙江省杭州市萧山区经济技术开发区建设三路733号信息港五期一号楼205-18

授权公告日：2023年03月10日授权公告号：CN 115470008 B

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，颁发发明专利证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。专利权期限为二十年，自申请日起算。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

2022114180150

局长
申长雨

2023年03月10日

第1页(共2页)

其他事项参见续页

@2024 OpenPie. All rights reserved.

πDataCS 计算引擎之 PieCloudDB虚拟数仓，支撑更大模型所需的数据和计算



πCloudDB（强逻辑计算）

SQL语言实现的结构化数据上的模型计算

打破企业数据孤岛，整合企业所有表格类数据资源

云原生数据库/ 虚拟数仓

收藏 0 0

数仓虚拟化

播报 讨论 上传视频

PieCloudDB产品技术

数仓虚拟化由拓数派（杭州拓数派科技发展有限公司，又称“OpenPie”）于2023年3月[4]正式提出。

数仓虚拟化[1]可将物理数仓整合到云原生数据计算平台[2]，根据数据授权动态创建虚拟数仓，按需灵活计算，打破数据孤岛，支撑更大模型所需的数据和计算[5]。在云上，数据计算资源按需扩缩容，提升数仓的敏捷性和弹性，助力企业降低数仓管理复杂度[3]，实现数量级增加可计算数据空间的同时，数量级降低数仓成本，打开无限数据计算空间，推进AI/BI到下一个精度[2]。

中文名	数仓虚拟化[1]	提出时间	2023年3月14日[4]

技术突破：数仓虚拟化[12]

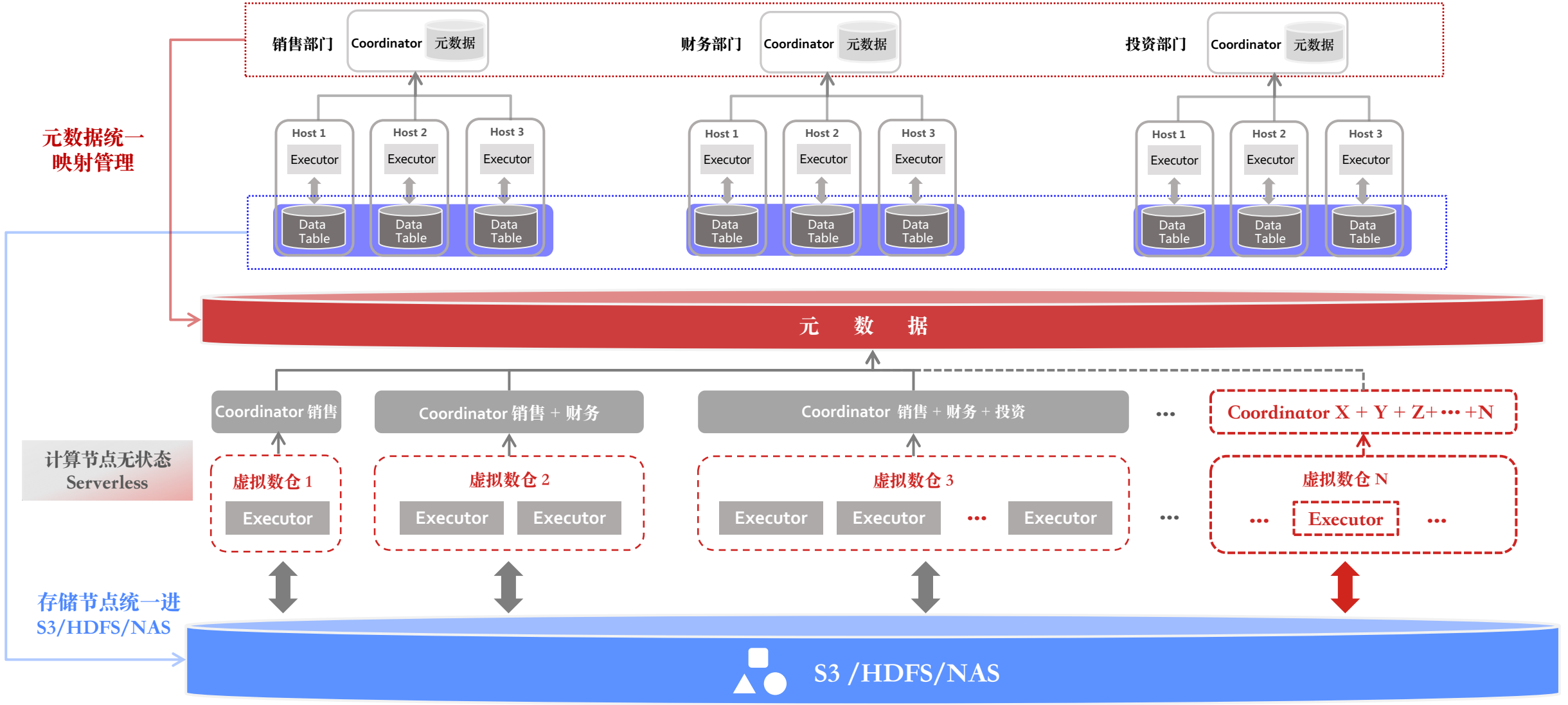
- **云原生存算分离架构**[9]：运用元数据-计算-数据分离的三层架构，实现云上存储资源与计算资源的独立管理。云上计算资源可弹性分配，有查询计算任务的时候按需启动，按照使用时间和规模计算成本。[11]
- **eMPP分布式专利技术**[9]：在云上，PieCloudDB利用eMPP（elastic Massive Parallel Processing）架构，实现多集群并发执行任务。[11]可灵活进行扩缩容，随着负载的变化实现高效的伸缩，轻松应对PB级海量数据。
- **全新的存储「简墨」和缓存架构设计**[9]：在计算层，各个计算节点针对元数据和用户数据都设计了多层缓存结构，避免网络延迟和数据移动，提高计算效率，保证用户的实时性需求。PieCloudDB针对底层对象存储设计了高效的文件格式，可在节省网络请求的同时提高计算效率。[11]
- **全新的优化器「达奇」**[9]：PieCloudDB可以智能高效地生成统计信息，并生成更高效的查询计划，达奇优化器支持聚集下推，预计算，Block Skipping等高级特性，全面满足各种复杂的分析查询需求[11]。

eMPP专利技术发明者/虚拟数仓产品的全球行业领导者

@2024 OpenPie. All rights reserved.

数仓虚拟化：物理数仓整合

- ✓ 按需弹性开启/关闭虚拟数仓，灵活配置存算节点
- ✓ 打破数据孤岛，消除数据多副本
- ✓ 数量级降低成本的时候数量级增加数据计算空间（私有部署更加敏感）



@2024 OpenPie. All rights reserved.

PieCloudDB云原生虚拟数仓 核心价值

突破性创新的核心技术 出色的数仓成本效益

降低数仓硬件和管理成本

物理数仓整合到云原生数据计算平台，根据数据授权动态创建虚拟数仓，打破数据孤岛，解决数据多副本问题，帮助企业降低数仓管理复杂度，以更低的成本实现存算资源在云上更灵活的配置。

提升数据计算资源利用效益

数据计算资源按需扩缩容，实现计算资源配置最优化，提升数仓的敏捷性和弹性，打开无限数据计算空间，支撑更大模型所需的数据和计算。更好地赋能业务发展并走向绿色。

坚如磐石 | 高安全 高在线 高可靠

TDE技术保证了所有数据在落盘前完成加密，服务器无感知技术（Serverless）利用云上无限计算资源和弹性保证了虚拟数仓永远在线可用，S3存储和跨云灾备能力保证了永不丢数。

@2024 OpenPie. All rights reserved.

PieCloudDB技术突破：数仓虚拟化

云原生存算分离架构

运用元数据-计算-数据分离的三层架构，实现云上存储资源与计算资源的独立管理。云上计算资源可弹性分配，有查询计算任务的时候按需启动，按照使用时间和规模计算成本。



全新的优化器「达奇」

PieCloudDB可以更智能高效地生成统计信息，并生成更高效的查询计划，**达奇优化器支持聚集下推，预计算，Block Skipping等高级特性，全面满足各种复杂的分析查询需求。**

全新的存储「简墨」和缓存架构设计

在计算层，各个计算节点针对元数据和用户数据都设计了多层缓存结构，避免网络延迟和数据移动，提高计算效率，保证用户的实时性需求。PieCloudDB针对底层对象存储设计了高效的文件格式，可在节省网络请求的同时提高计算效率。

eMPP分布式专利技术

在云上，PieCloudDB利用eMPP（elastic Massive Parallel Processing）架构，**实现多集群并发执行任务。**企业可灵活进行扩缩容，随着负载的变化实现高效的伸缩，轻松应对PB级海量数据。

@2024 OpenPie. All rights reserved.

eMPP分布式专利技术



证书号第5775327号

发明专利证书

发 明 名 称：一种元数据访问方法、装置和存储介质

发 明 人：王滨舟;杨峻峰;冯雷

专 利 号：ZL 2022 1 1418015.0

专利申请日：2022年11月14日

专 利 权 人：杭州拓数派科技发展有限公司

地 址：311200 浙江省杭州市萧山区经济技术开发区建设三路733号信息港五期一号楼205-18

授权公告日：2023年03月10日　　授权公告号：CN 115470008 B

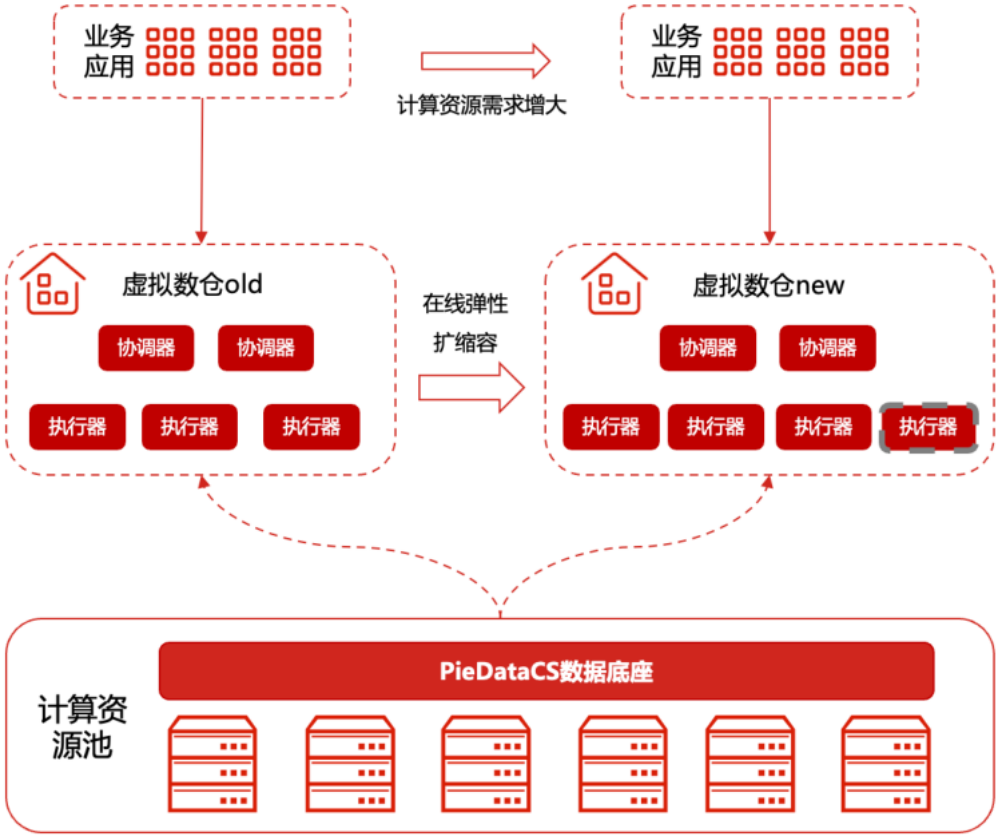
国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，颁发发明专利证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。专利权期限为二十年，自申请日起算。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长
申长雨

2023年03月10日

第1页(共2页)



架构阐述：

1. 云原生虚拟数仓，采用eMPP（弹性大规模并行计算）技术，把所有的计算节点组成一个大的计算资源池，然后用户可以在计算资源池上创建虚拟数仓；
2. 虚拟数仓是由执行器和协调器等资源组成，协调器和执行器在操作系统层面来看，他们就是一个个的进程，这些进程分布在不同的计算节点上；
3. 如图中示例，业务应用的计算任务数量增加或需要处理的数据量变大，对应的计算资源（主要是CPU、MEM）也会增加，根据计算资源需求，虚拟数仓可以在线弹性的扩缩容，在此图示例中，执行器可以从原来的3个变成4个，缩容也是同样的原理；
4. 虚拟数仓在执行具体的业务应用请求时，协调器会去统筹整个计算资源，将优化后的SQL下发至执行器，执行器负责具体的SQL执行，多个执行器之间并行的计算，最终结果需要返回给协调器，由协调器最终反馈给业务应用；
5. 多个执行器并行处理数据时，会存在一个现象，其中一个执行器工作中突然故障，导致整个任务失败，此时需要业务应用重新下发SQL指令。

@2024 OpenPie. All rights reserved.

统一存储格式，一份数据，多引擎计算

核心 - JANM Table Format

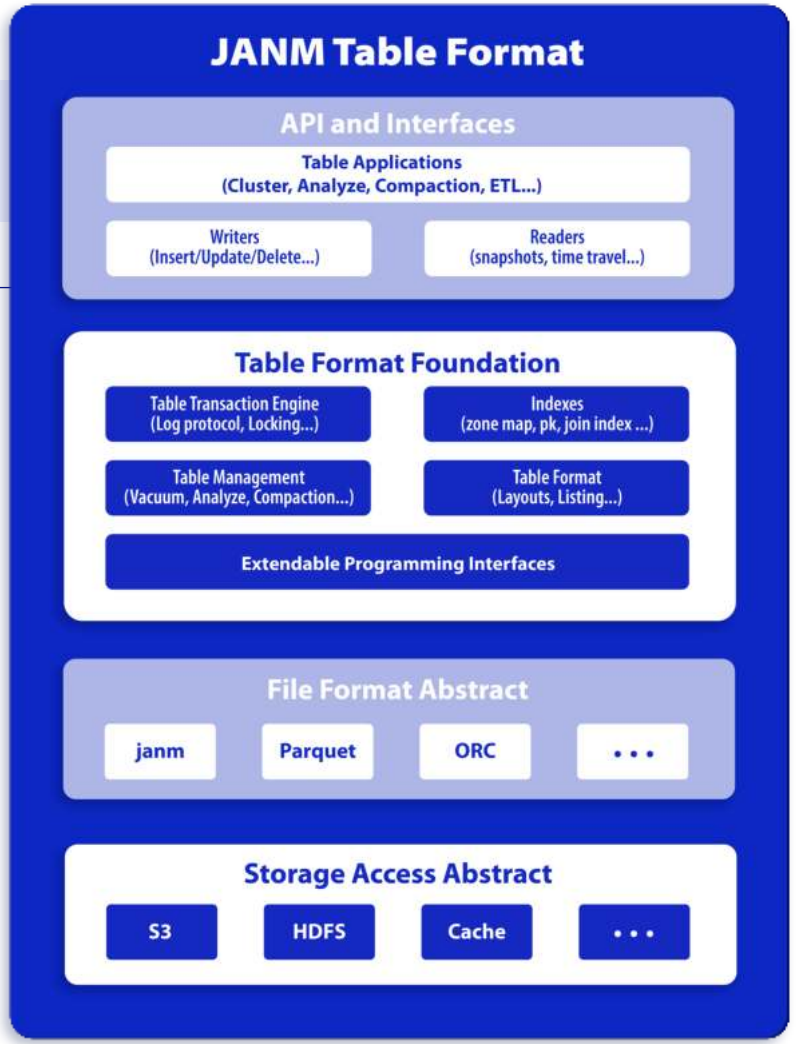
四家最早推出表格式Table Format提出者：

拓数派 自研简墨(JANM)存储

Apache hudi

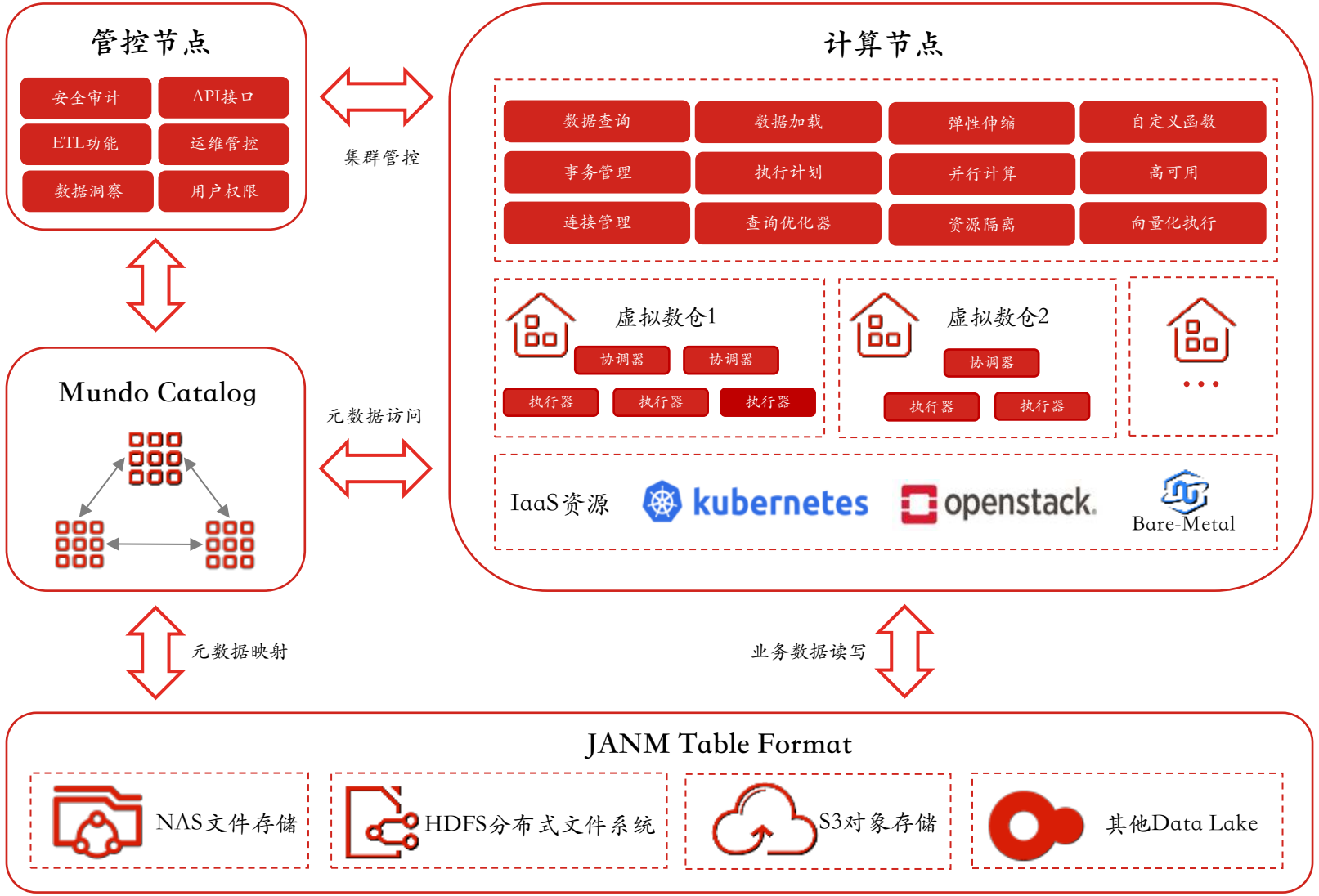
DELTA LAKE ICEBERG

为了实现一份数据，多引擎计算，全面支持大模型的同时也兼容大数据。



@2024 OpenPie. All rights reserved.

PieCloudDB--虚拟数仓引擎



元数据：

- 分布式KV，存储系统表
- 基于key的自然排序实现索引
- 高性能
- 基于watcher机制的高效分布式锁管理
- 基于MVCC的事务隔离级别
- 多集群数仓数据共享

计算节点：

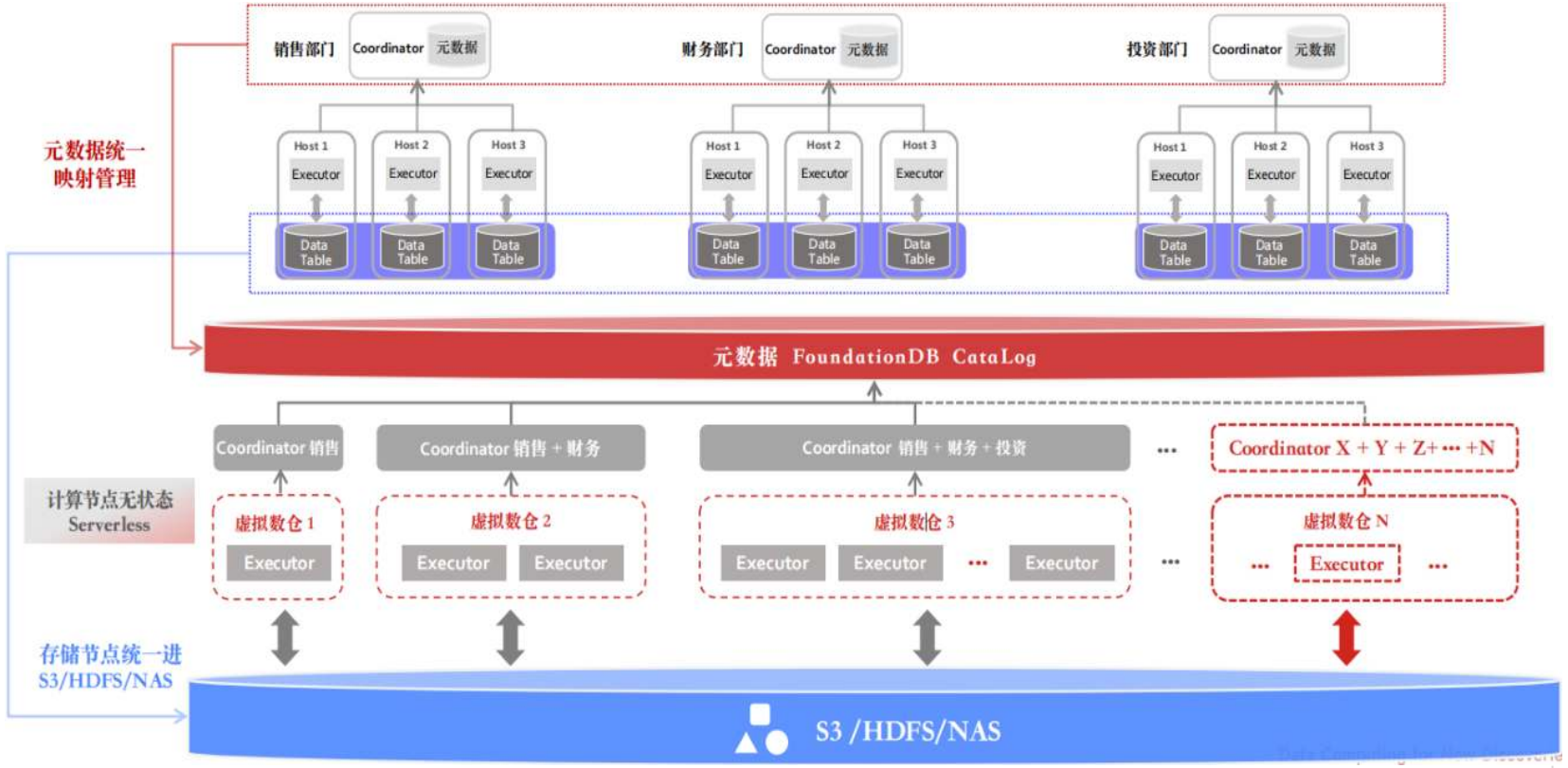
- 自定义虚拟数仓，资源隔离
- 向量化执行器（SIMD）和JIT
- HTAP，弹性扩缩容
- eMPP专利，弹性大规模并行计算
- 分布式优化器
- 存算分离架构

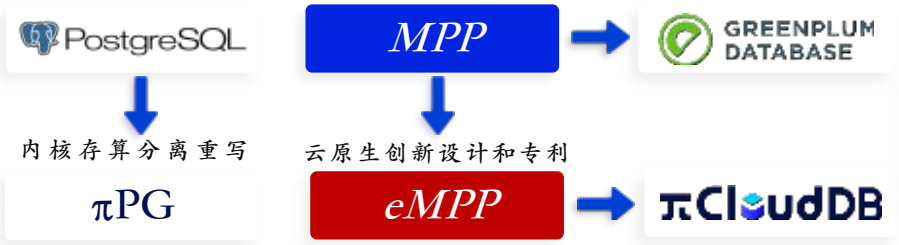
可视化管控：

- 监控告警
- 自定义和动态服务启停
- 备份恢复
- ETL和数据导出
- 数据洞察
- 细粒度权限管控，与LDAP无缝集成
- Rest API

@2024 OpenPie. All rights reserved.

国内首创虚拟数仓技术





行业顶级数据库的
抽象思考和设计原则复用

打破数据孤岛，解决数据多副本

物理数仓整合到云原生数据计算平台，根据数据授权动态创建虚拟数仓，打破数据孤岛，解决数据多副本问题，帮助企业降低数仓管理复杂度，以更低的成本实现存算资源在云上更灵活的配置。

提升数据计算资源利用效益

数据计算资源按需扩缩容，实现计算资源配置最优化，提升数仓的敏捷性和弹性，打开无限数据计算空间，支撑更大模型所需的数据和计算。更好地赋能业务发展并走向绿色。

坚如磐石 | 高安全 高在线 高可靠

TDE技术保证了所有数据在落盘前完成加密，服务器无感知技术（Serverless）利用云上无限计算资源和弹性保证了虚拟数仓永远在线可用，S3存储和跨云灾备能力保证了永不丢数。

@2024 OpenPie. All rights reserved.

云原生分布式优化器--达奇

达奇

PieCloudDB Optimizer

- 基于eMPP架构的云原生分布式优化器；
- 为海量数据集上的复杂OLAP查询提供最优的查询计划。



聚集下推

通过把聚集操作下推到连接操作之前去执行，极大的减少连接操作需要处理的数据量，使得查询性能显著提升。在很多场景下经过严格的测试，聚集下推会取得百倍或千倍的性能提升。

预计算

根据对象存储的特性，为每一个数据块预先计算好块内数据的聚集值。通过把聚集操作下推到对数据表的扫描之上，可以直接利用数据块内预计算的结果，从而避免了大量的计算。





文件裁剪

利用每个数据块内预计算的结果，提前过滤掉不符合条件的数据块，从而减少对数据的扫描和后续的计算。



多表连接的最优顺序搜索



多阶段聚集



分区表的静态和动态裁剪



相关子查询的提升转换



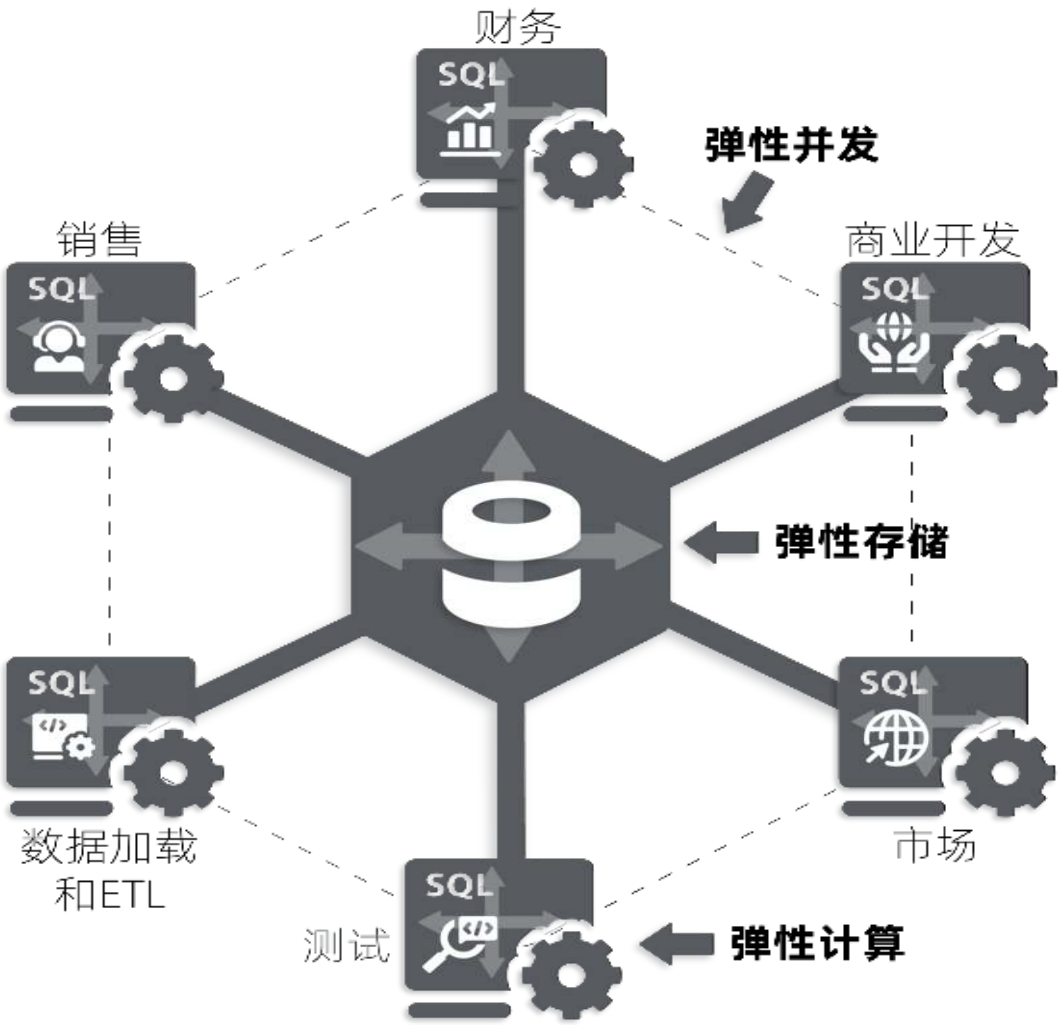
CTE和递归CTE的优化



其他相关优化

@2024 OpenPie. All rights reserved.

分布式设计，保证可靠性



计算层：

- eMPP：根据计算任务智能调度合适的资源，将一个单一计算任务在大量独立的计算机上并行执行。
- 多租户，资源隔离。
- 弹性伸缩，支持多集群协同。

存储层：

- 存算分离，存储独立扩展。
- 高可靠，支持跨多数据中心的灾备复制。
- 节约成本，全局数据唯一存储，通过共享存储来实现数据共享，避免拷贝和维护多份数据副本。

事务管理：

- 存两种隔离级别：Read Committed、Repeatable Read
- 高性能事务管理器，支持自动扩展。
- 高可靠性，支持故障自动容错。
- 租户之间的事务管理器完全隔离，互不干扰。

@2024 OpenPie. All rights reserved.

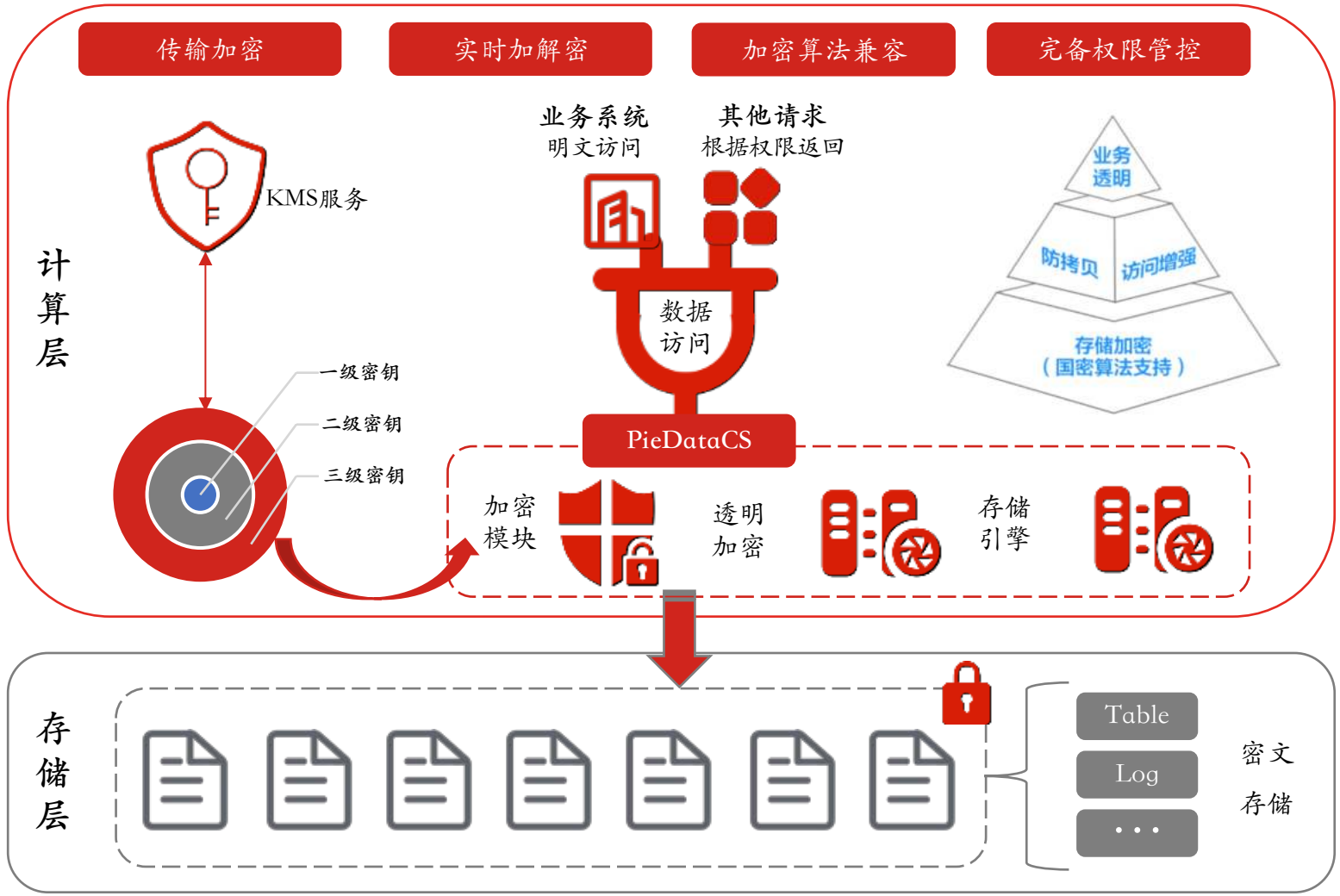
数据安全

国家法律与全国性标准

- 中华人民共和国网络安全法
- 中华人民共和国数据安全法（草案）
- 中华人民共和国个人信息保护法（草案）
- 信息安全技术网络安全等级保护（等级保护2.0）

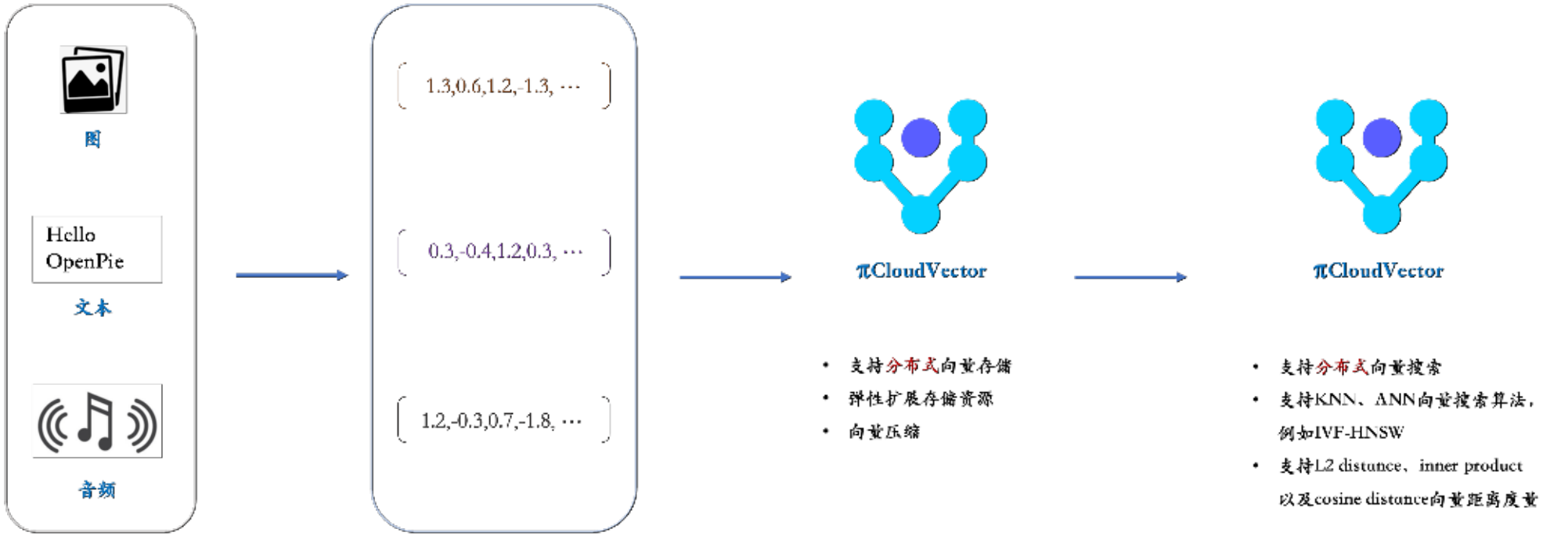
- 2019年：中国网络安全等级保护2.0（等保2.0）起步年。同年《新密码法》和《个人信息保 护法（专家建议稿）》等发布。
- 《网络安全法》第四十二条： 网络运营者不得泄露、篡改、毁损其收集的个人信息；未经被收集者同意，不得向他人提供个人信息。但是，经过处理无法识别特定个人且不能复原的除外。
- 《中国银行业“十二五”信息科技发展规划监管指导意见》中提出，完善敏感信息存储与传输等高风险环节的控制措施，对数据、文档的访问应建立严格的审批机制，对用于测试的生产数据要进行脱敏处理，严格防止敏感信息泄露。

满足等保合规要求



@2024 OpenPie. All rights reserved.

πDataCS的第二个计算引擎πCloudVector



@2024 OpenPie. All rights reserved.

PieCloudVector 竞争优势

专用向量数据库



- 冗余数据、过度的数据搬运、分布式组件之间的数据缺乏一致性
- 专业技能的额外劳动力成本、额外的许可成本
- 有限的查询语言能力、可编程性和可扩展性
- 有限的工具集成
- 较差的数据完整性和可用性

具备向量搜索能力的云原生虚拟数仓



πCloudVector

- 统一的数据平台，在动态扩缩容过程中无需移动数据，充分保障数据的一致性
- 使用简单，学习成本低，无需额外投入
- 既满足了向量存储和向量搜索的需求，又升级实现了云上分布式向量化计算的技术突破
- 支持多种向量搜索算法，为不同的业务场景提供更灵活高效的解决方案

打破专用向量数据库的局限性

具备向量搜索能力的传统数据库



- 无法弹性扩缩向量化存储和计算的资源
- 在向量化计算的场景下，易用性和性能较差
- 对元数据变更的功能有限，导致了数据一致性等问题
- 传统数据库存在的安全性、可靠性、在线性方面的技术瓶颈

突破传统数据库的技术瓶颈

@2024 OpenPie. All rights reserved.

πDataCS 计算引擎之PieCloudML（大模型）机器学习，更大模型多模机器学习数据计算

πCloudML (Beta)

（大模型）机器学习

Machine Learning

Graph Parallel Computing

Large Language Models

…

服务器无感知（Serverless）计算引擎

πFPGA

数据存储 | 虚拟数仓 | 特定领域（如神经网络）

Mundo 元数据管理系统

简 墨 (JANM) 存 储

统一 Catalog | 统一数据格式 | 一份数据多引擎计算

Parquet

Orc

…

阿里云

腾讯云

私有云

vmware

openstack

@2024 OpenPie. All rights reserved.

产品版本和部署方式

云上云版

云上云版
- ✓ 领先的数仓虚拟化技术
- ✓ 云原生Serverless架构
- ✓ 第三方计费模式

公有云：阿里云、腾讯云、aws

企业版/社区版

企业版
- ✓ 云原生eMPP专利技术支持与赋能
- ✓ 7*24小时产品故障服务响应
- ✓ 基于业务需求的专家团队服务

社区版
- ✓ 持续产品升级，体验产品最新特性
- ✓ 丰富的社区活动，赋能社区用户
- ✓ 强大的开发者支持

私有云：vmware vSphere、openstack

一体机

一体机
- ✓ 国产化软硬件，技术自主可控
- ✓ 存算分离支持独立扩容
- ✓ 全面适配信创环境

裸硬件

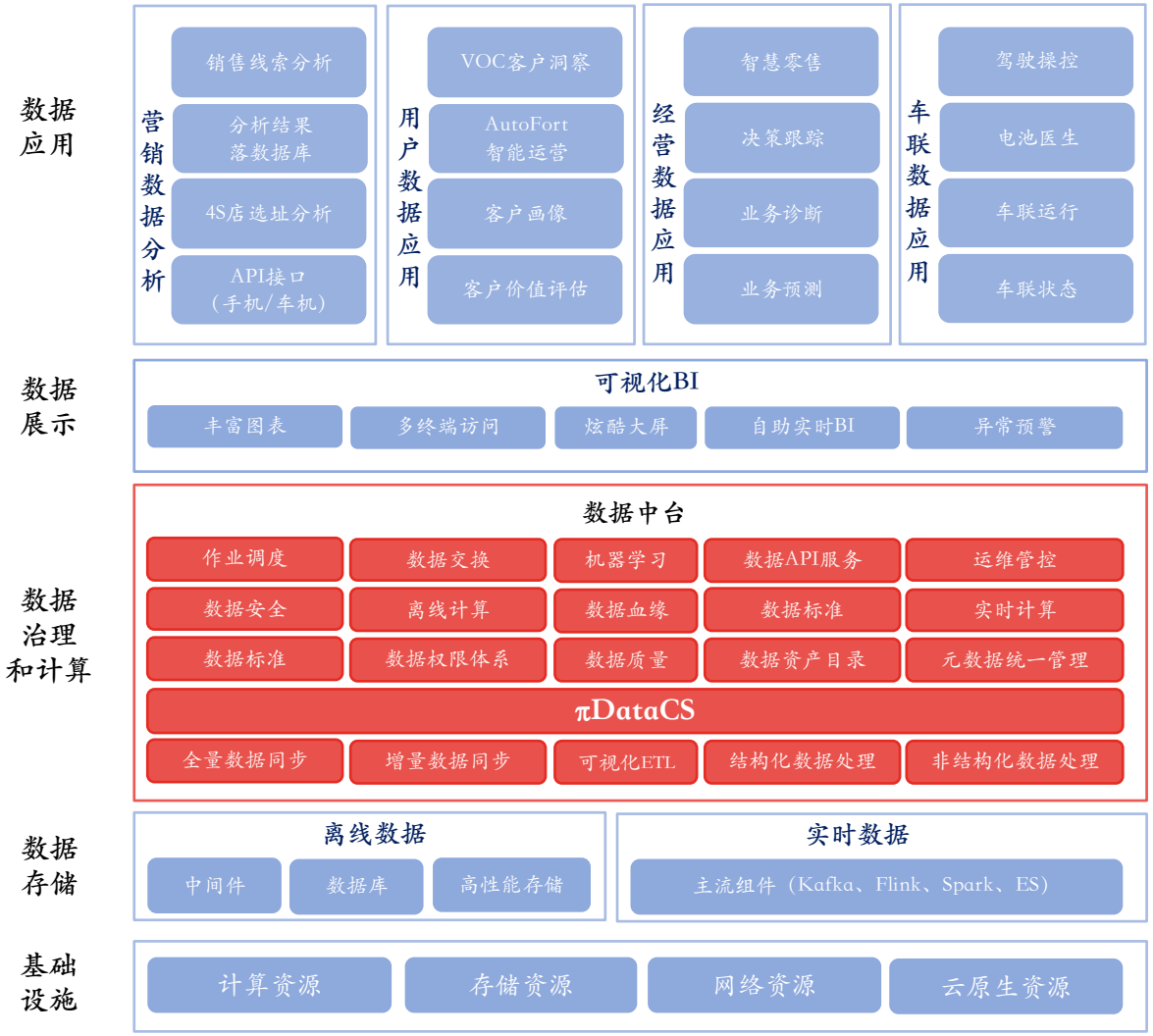


@2024 OpenPie. All rights reserved.

PART 03

πDataCS应用案例

客户案例：某大型汽车集团全面升级大数据平台至云原生πDataCS



项目背景：

该企业依托于集团内部需求，建设了一个云平台，提供基础设施服务。随着子公司接入的数量越来越多，数据量也急剧增长，原本采用的Hadoop平台无法满足业务的分析需求，主要存在以下问题：

- 数据同步和采集周期较长，实时性很差
- 数据量达到PB级别，数据分析性能直线下降
- 采用较多开源组件，技术栈太广，数据应用开发周期长
- Hadoop集群运维成本较高，技术响应慢

解决方案：

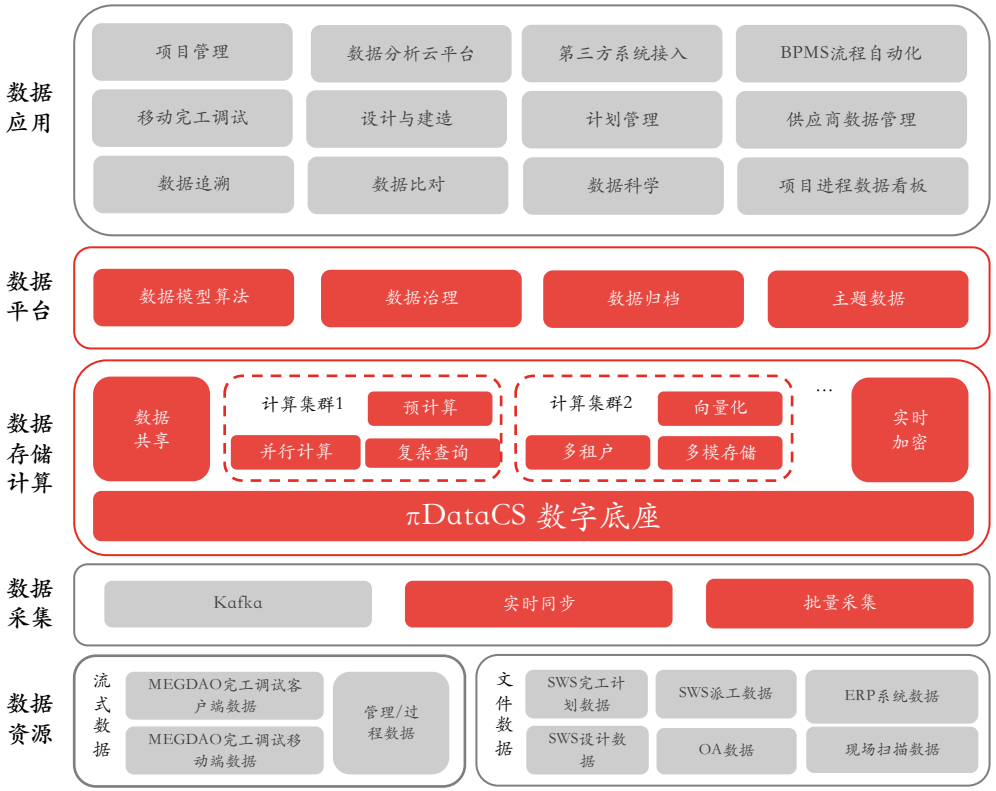
采用PieCloudDB作为核心数字底座，完成原来CDH（Hadoop）平台的替换，构建了制造业行业云的数据中台服务，既面向集团内部服务，同时也面向行业的公司提供服务。在集团内部主要提供营销数据、用户数据、经营数据以及车联网数据等数据的分析和应用。具体如下：

- 替换了Hive数仓，采用PieCloudDB云原生虚拟数仓作为数仓服务，整体数仓的SQL执行性能是原Hive的30倍；
- 支持更多的数据源，具备全量和增量数据同步功能，优化了数据采集和入库的性能，可通过ChunJun进行实时入库，性能较传统GP提升了6倍；
- 升级了数据中台的安全审计应用，采用PieCloudDB替换了ES，支持半结构化数据的查询（json/xml等数据格式），应用开发不用再基于SPL查询语法，改用通用型的SQL语法即可，大大缩短了应用开发和迭代周期；
- 具备流批一体化能力，兼容Spark和Flink等应用程序，替换了部分Spark应用（SparkSQL）；
- 与集团内部的云平台进行了统一集成，通过API接口实现了系统对接，统一进行租户和资源的管理，支持资源管控和隔离、数据透明加密以及租户密钥隔离等；
- 引入PieCloudDB虚拟数仓技术，可以让技术栈更加专精，大大降低了数据中台的开发、运维和使用的门槛，大幅提升了数据应用开发的效率；
- PieCloudDB具备完善的商业化技术服务，提供本地化技术支持，技术响应更加及时，有效提升服务的SLA；
- PieCloudDB采用简洁的技术架构，没有复杂的组件叠加，让集群运维管理门槛更低，帮助集团企业大幅降低后期的运维管理成本；
- PieCloudDB是云原生的存算分离架构，可以无缝对接企业已有的S3协议对象存储和HDFS大数据文件系统，无须再次购买存储资源。同时可以按需进行集群的扩缩容，计算资源和存储资源可单独分别扩容，可大幅减少企业前期的投入成本和后续的扩容成本。
- 运用数仓虚拟化技术，实现全域数据计算

@2024 OpenPie. All rights reserved.

客户案例：大型先进制造企业，通过πDataCS全面构建国产化数据底座





业务需求：

基于核心研发团队在集团首艘豪华邮轮建造项目中的调试完工实施管理及质量数据应用的经验，打造集监测、检测成一体适用于大型复杂装备建造项目的国产数字化调试完工应用平台。旨在为豪华邮轮国产化提供专业的自动化软件应用的同时，加速调试完工数字平台国产化的进程。 本期项目的主要建设目标：

- 在设计数据收集的前端，搭载基于国产设计的智能图像算法芯片和传感器的三维结构光体测量设备，来深度挖掘设计和建造数据，打造孪生设计大数据库，穿透调试计划，串联供应链数据。
- 致力于建设一个安全、可控、软硬件一体的数字建造工业自动化软件应用平台，从设计到交付，发布全流程建造工程调试完工数字化，
- 构建成套云计算架构下的工业软件平台，不仅可服务于豪华邮轮国产化建造，还可应用与数字化升级背景下的各项大型复杂装备建造项目。

解决方案：

基于πDataCS建设一个国产化的调试完工数据平台，平台不仅对标了行业标杆产品，挪威COMPLAN公司WINPCS的核心操作业务模块和数据结构，更将迭代至基于云原生的自主可控eMPP国产数据库，从而搭载一组多专业工种可适用的分布式工业自动化应用系统。 具体如下：

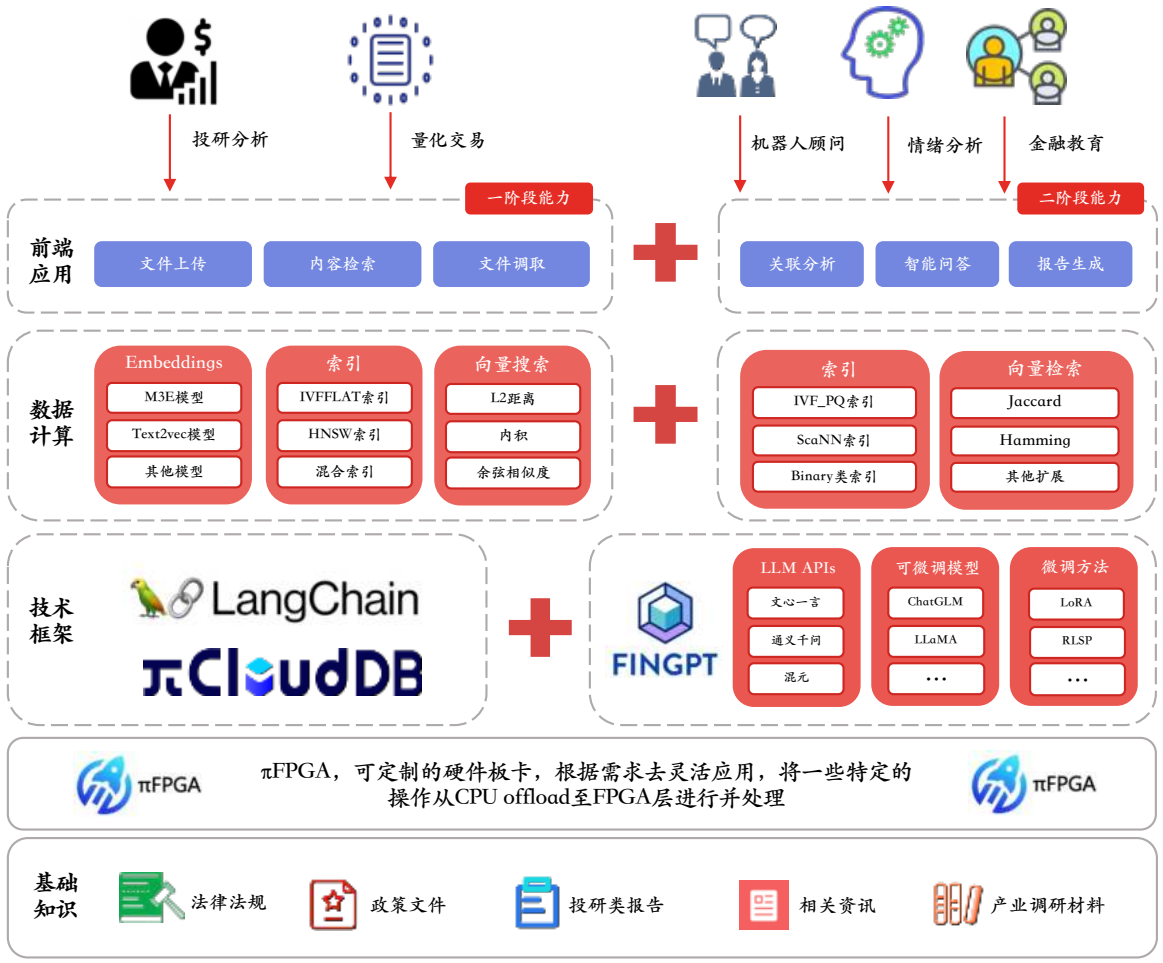
- MEGDAO团队与πDataCS团队双方进行了战略化合作，平台所有产品全面迁移至国产的πDataCS 数据计算系统，优化了原型产品的诸多基础数据驱动程序。
- 基于πDataCS 数据计算系统内置的机器学习算法库，开发自主研发的核心模型算法，结合企业流程管理的本地化业务需求，加速移动端和数据云化的应用升级。
- 采用存算分离架构，构建统一的集群元数据服务，让整个平台扩容更加方便快捷，解决了指数级递增的海量数据存储和数据共享等问题。
- 提供多租户、数据加密以及严格的权限管控等能力，帮助海工能源领域企业，特别是在海外项目中遇到的数据安全问题，实现了跨区域、跨平台、多热点、多并发等数据服务的核心技术能力。

项目收益：

- ✓ 解决关键交付数据的可信度问题
- ✓ 可为建造方提供高效的数据整理手段
- ✓ 自动化创建原本重复而机械化的检验文本
- ✓ 大数据管理分析，提供智能监管及自动控制操作
- ✓ 减少时间成本，降低人为出错概率
- ✓ 保证建造项目准确、完整、安全完成交付
- ✓ 国产工业软件+国产数据库产业应用落地
- ✓ 自主可控的全流程复杂工程调试完工数字化应用
- ✓ 在设计阶段接入项目的调试计划
- ✓ 用国产设计的感应设备、扫描设备及智能芯片记录现场数据
- ✓ 跨台的应用嵌入，移动端工业软件的用户体验
- ✓ 支持数据科学工程和BI工具，提供可视化业务看板

@2024 OpenPie. All rights reserved.

客户案例：某证券客户以πDataCS构建LLM垂直应用



投研分析：

针对日常工作中的法律法规、政策文件以及投研报告进行快速检索，形成对应的分析报告，为客户提供投资相关的数据支撑。

量化交易：

基于 GPT 的大量金融数据训练，可以发现事件的情绪对资产的影响模式，可将这种模式用到量化交易策略中，由 GPT 实时产生交易信号以自动执行交易。这种数据驱动的量化交易策略可以更快地对市场变化进行响应。

机器人顾问：

根据个人需求和偏好实时提供金融建议，提高了获取建议的便利程度，降低了获得服务的成本。GPT 可以学习大量历史案例和研究报告，在此基础上，它可以总结出投资策略与建议。用户可以直接使用这些建议，或根据实际情况进行适当调整，这可以极大提高工作效率。

金融市场情绪分析：

对投资相关的言论和情绪表达等数据进行深度分析，获取市场情绪的指标，帮助投资者更好的把握市场走势，制定合理的交易策略，避免情绪犯错。

金融教育：

形成基础金融知识和产品知识库，针对用户和内部员工传授相关的投资策略和产品功能，用户可以与GPT交互，询问投资工具、风险和回报等具体问题，让用户对投资原理有更深的了解。同时也可以让员工更快速的掌握更复杂的金融产品和工作技能。

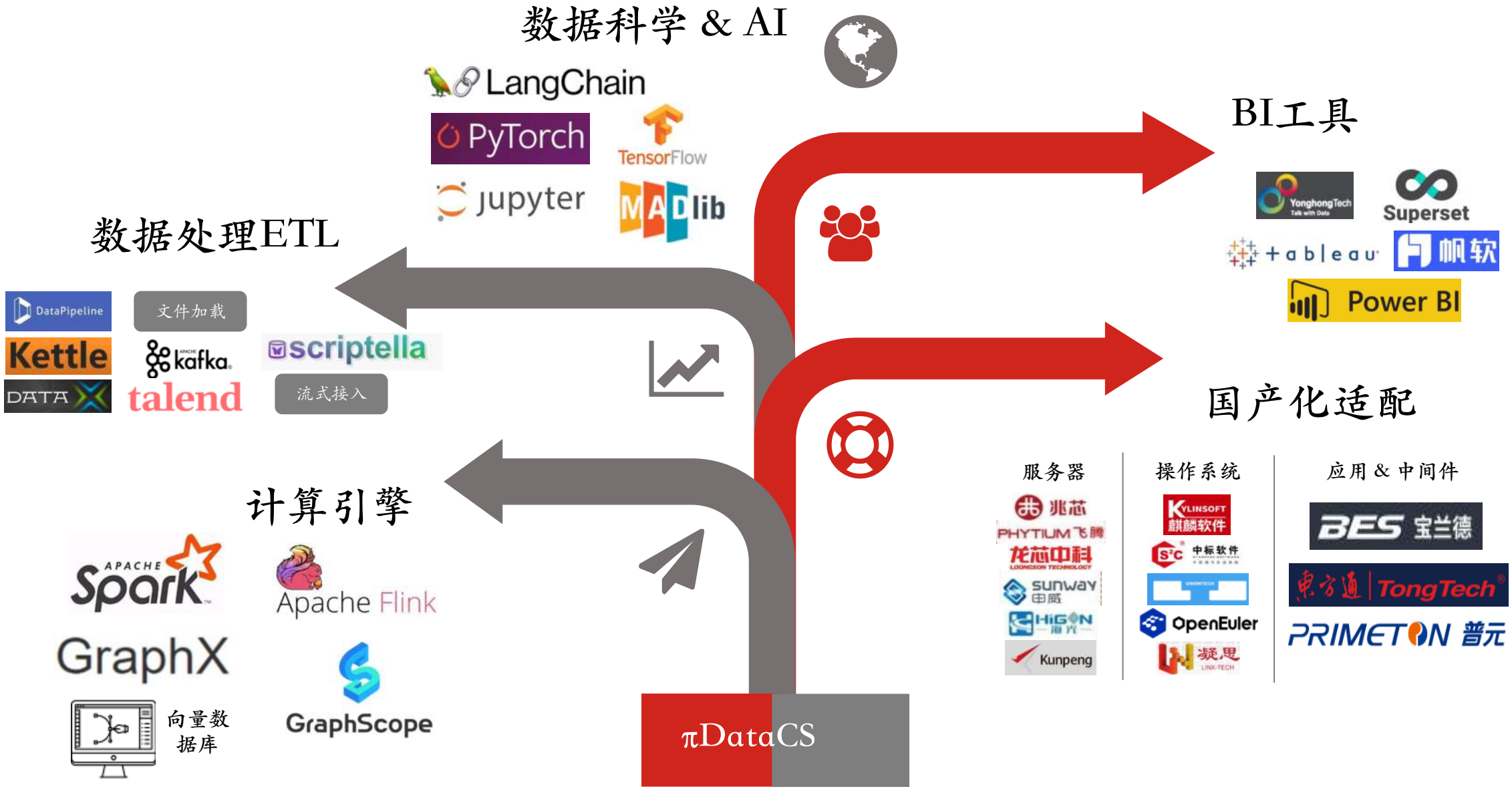
业务运营报告：

结合内容运营的需求，把行情类的数据进行智能化和个性化创作，定时为客户推送实时行情、财经类的资讯以及精准的市场数据和分析，提升内容创作效率。

@2024 OpenPie. All rights reserved.

PART 04

总结



@2024 OpenPie. All rights reserved.

OpenPie



关注PieCloudDB公众号

随时获得产品动态



加入PieCloudDB技术群

获得更多技术干货

@2024 OpenPie. All rights reserved.

Thanks

Data Computing for New Discoveries

数 据 计 算 ， 只 为 新 发 现