DTCC 2010 2022

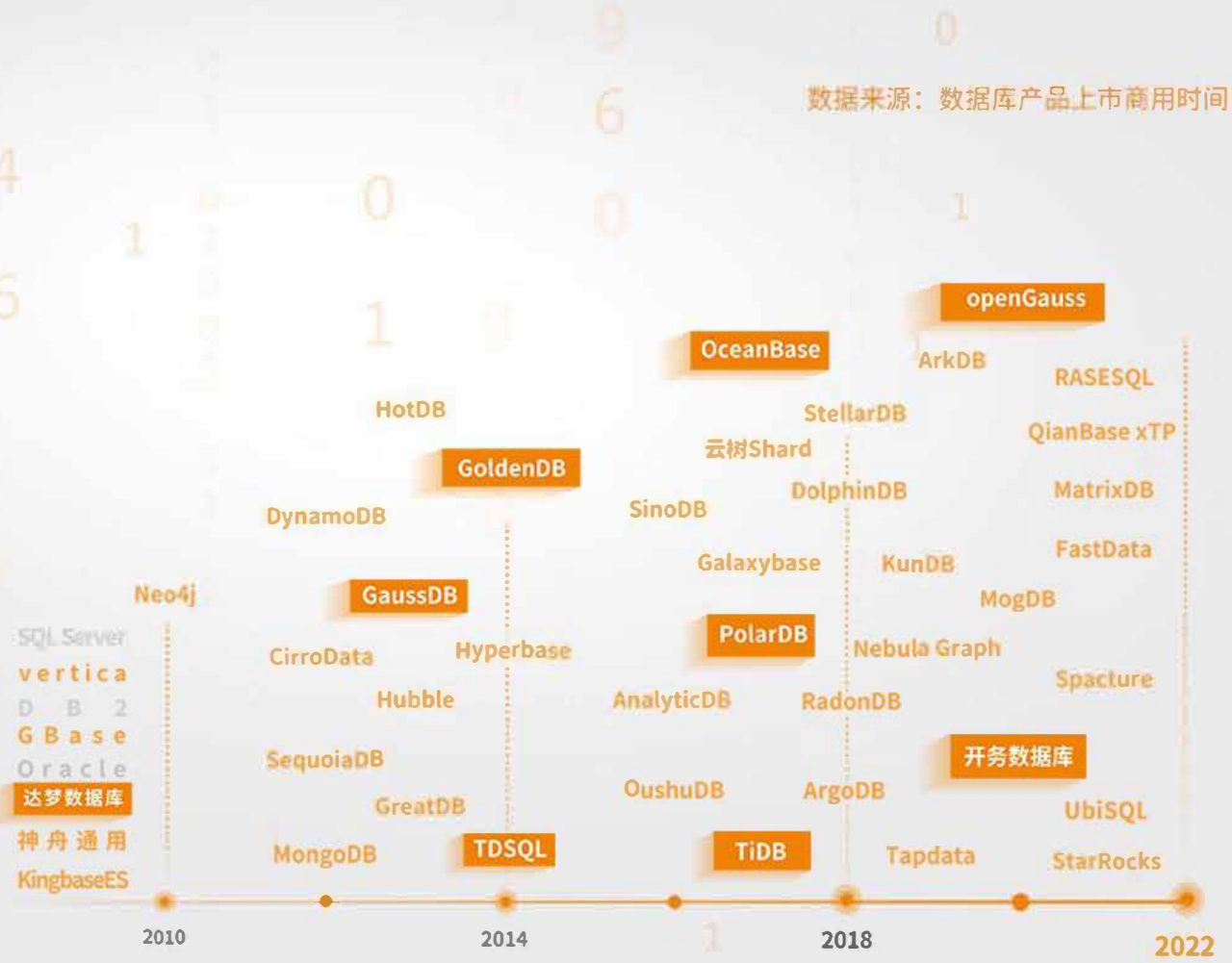
第十三届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix ITPUB

线上直播 | 2022/12/14-16



云原生数据库 PieCloudDB eMPP架构设计与实现

郭罡 拓数派CTO

关于拓数派（OpenPie）

- 成立于2021年，以“Data Computing for New Discoveries”「数据计算，只为新发现」为使命。
- 现Pre-A轮融资，已完成数亿元融资。
- 核心团队来自于各大厂名校，有丰富的数据库（Greenplum，DB2，ClickHouse等）研发和商务经验。
- 核心产品 PieCloudDB 1.0版本已于 2022.10.24 发布。
- 产品已经在一些金融、医疗等行业开始使用。

关于我

- 毕业于中国科技大学，AI相关专业
- 毕业 1 年后到现在一直从事底层基础软件开发，10多年开发经验
- 领域涉及到：
  - 代码级/算法级/系统级性能优化
  - Linux/Unix内核和系统开发、虚拟化（芯片KVM支持实现）和云计算架构、高速网络开发（内核和应用层如DPDK）
  - 分布式系统（SQL/NoSQL/存储）
- 最近 7+ 年一直从事开源分布式数据库开发

TCloudDB

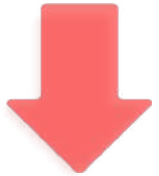
一个eMPP 云原生分布式SQL数据库

一个云原生实时大数据平台基座

愿景：安全可靠 使用简单 功能齐全 性能极致

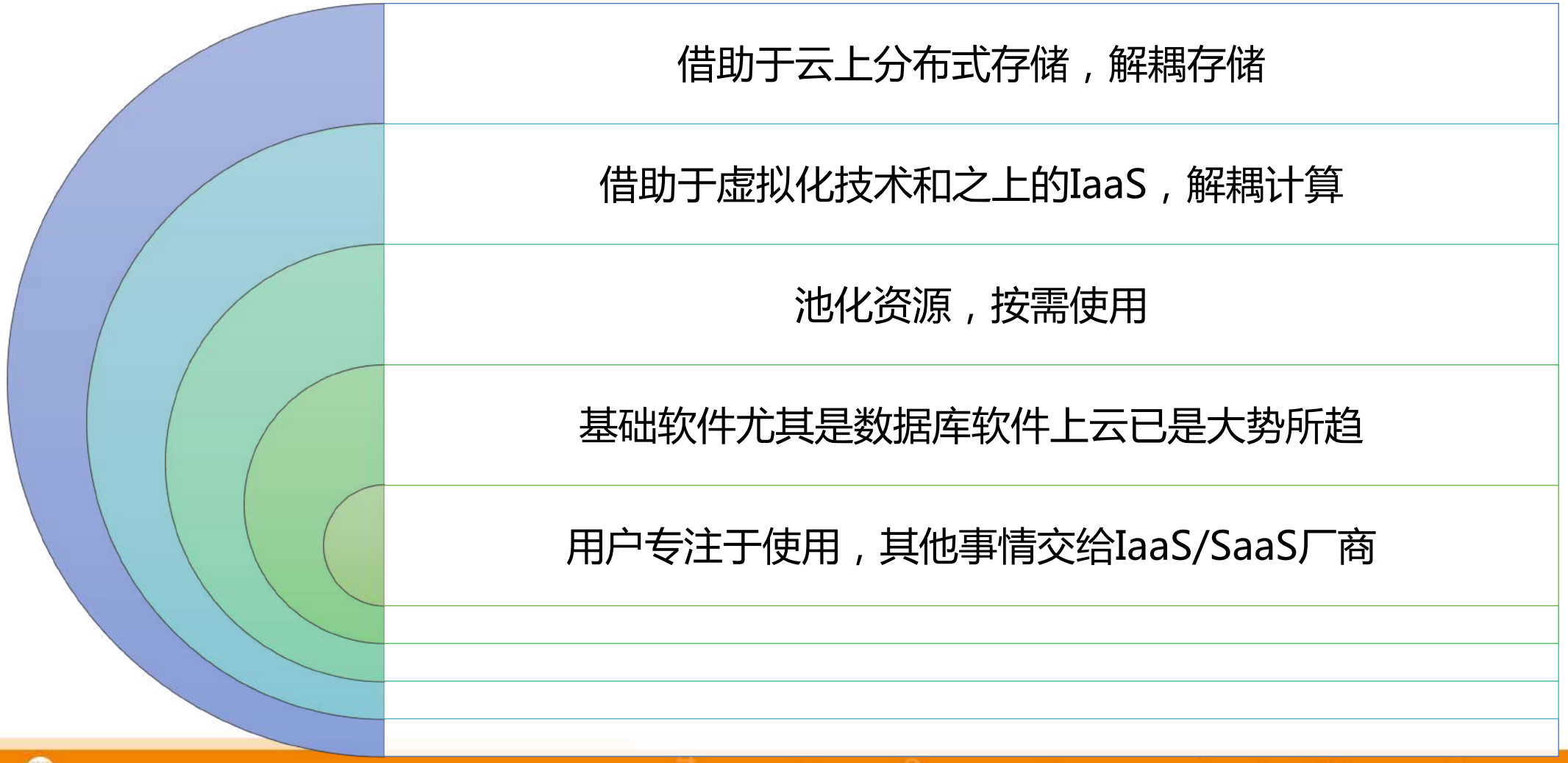
传统分布式MPP架构痛点

缺乏弹性	成本高昂	木桶效应	数据孤岛	运维成本
业务使用不灵活	集群固定，资源利用率低	扩缩容难	元数据和用户数据跨集群访问困难	运维和DBA



我们需要一个云原生数据库

云解决了什么？


借助于云上分布式存储，解耦存储
借助于虚拟化技术和之上的IaaS，解耦计算
池化资源，按需使用
基础软件尤其是数据库软件上云已是大势所趋
用户专注于使用，其他事情交给IaaS/SaaS厂商

上云 ≠ 云原生

- 存储资源和计算资源：分离和隔离
- 资源伸缩快速简单
- 计算、存储：**按需付费**
- 智能管理，复杂交给*aaS厂商

PieCloudDB 重要特点

eMPP

- 弹性计算资源（横向纵向）、极速调整
- 多集群是另外一个弹性的维度
  - 共享用户数据（如按需付费的对象存储）
  - 共享元数据
- MPP架构：分布式，海量数据并行处理
- e代表弹性(elastic)

友好的用户接口（websql, ODBC/JDBC driver等).

云原生 云中立

ACID; 完备的事务支持
(隔离级别：RR, RC)

完善的SQL标准支持

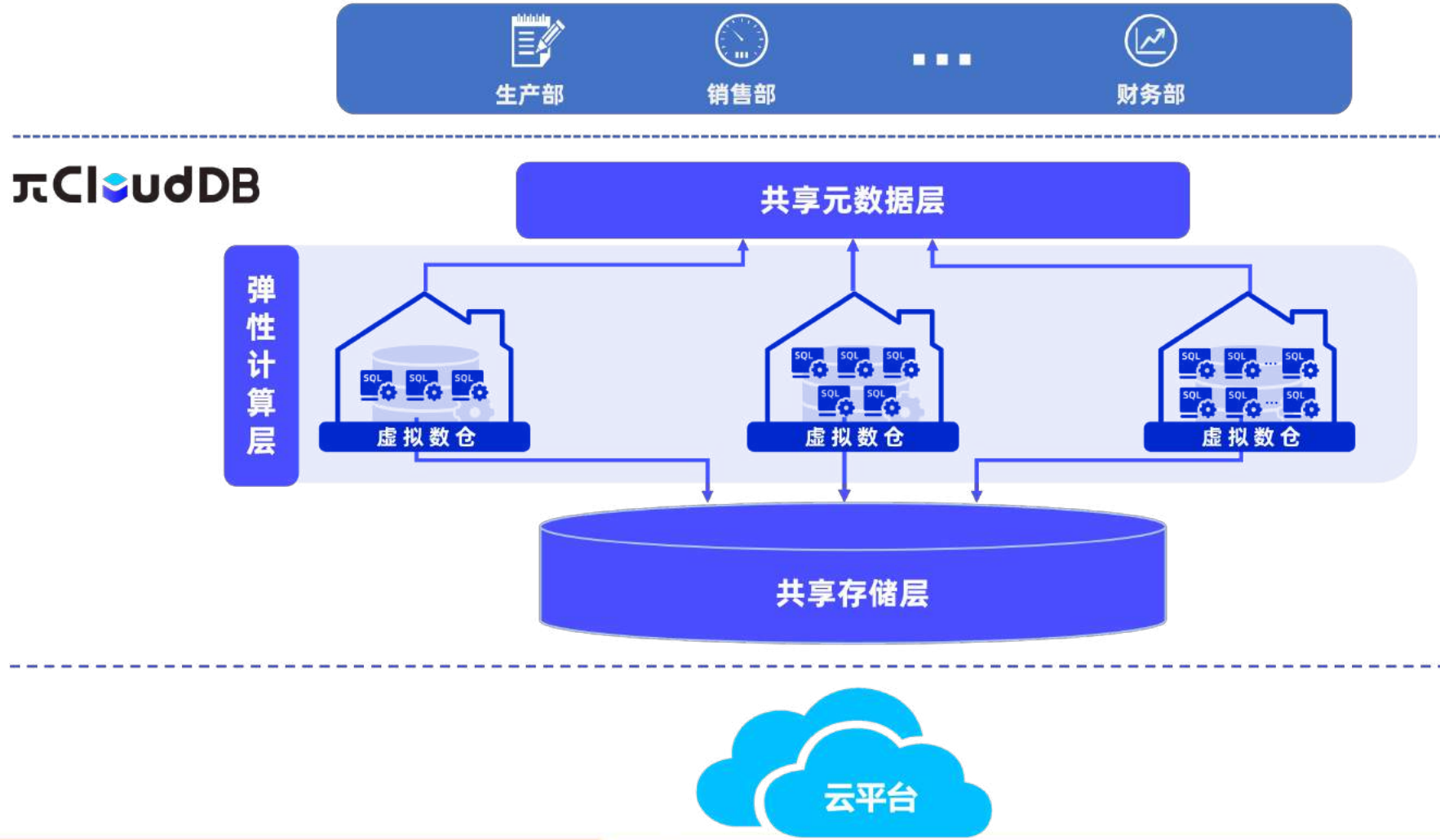
安全可靠

完善的Postgres生态

为什么选择Postgres?

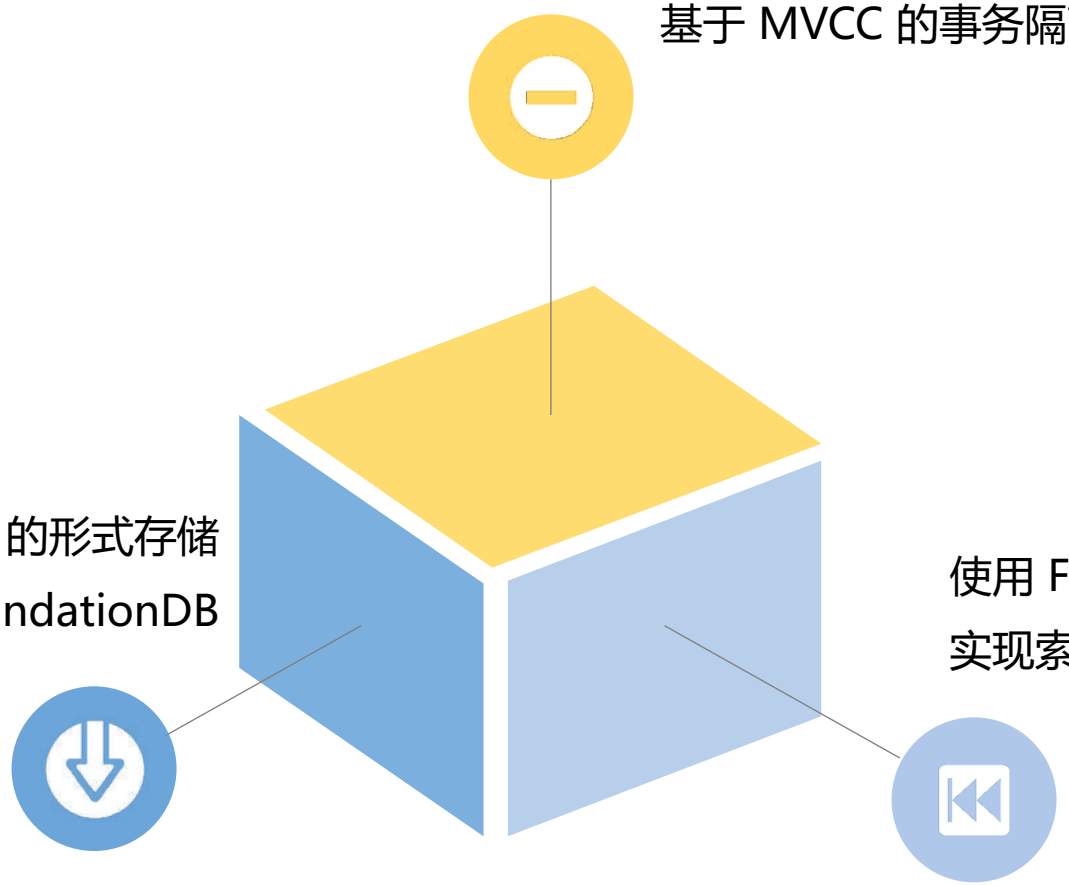
- 关于Postgres
  - 公司中立，开源协议友好，国际一流工程水准的先进开源数据库
  - Postgres对存储扩展，插件扩展支持友好
  - 天然自带一定的多模支持 (原生或者插件)
  - 采用度和流行度持续上升
  - 优秀的生态
- 我们的选择
  - 很多功能不用也没必要重新造轮子
  - 和一流的产品和人才一起成长
  - 团队深度理解Postgres内核代码，在社区参与诸多贡献

PieCloudDB 架构



元数据管理

基于 MVCC 的事务隔离级别



将元组以 key-value 的形式存储到 FoundationDB

使用 FoundationDB Key 的自然排序实现索引

元数据管理

- 临时状态存储（如lock等）也放在FoundationDB
- 依赖于FoundationDB的KV特性、可串行化事务、watcher机制
- 多个集群（虚拟数仓）可以共享一份元数据
- FoundationDB高可用设计、备份恢复保证元数据的可靠性和可用性

元数据管理缓存

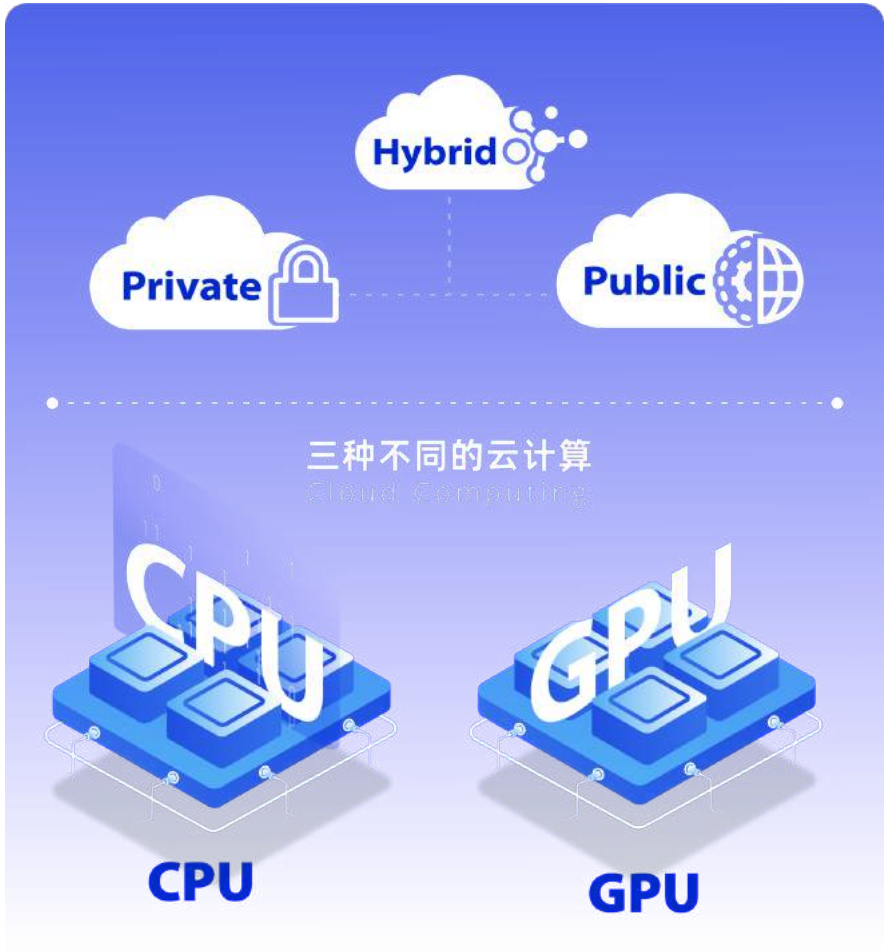
- 目的：
  - 减轻FoundationDB集群负担
  - 加速查询优化（网络延迟远高于内存延迟）
- 以Postgres原生的元数据缓存概念为基础，优化重构实现适用于多集群架构

用户数据存储引擎

- PAX（行列混存）配以高效压缩
- Block文件为一个存储(MVCC)单位
- 辅助信息存储用于计算优化
- 设计考虑：
  - 高效和精准的统计信息收集
  - 存储和计算成本
  - 各种计算优化
    - SIMD, Cache Line
    - Data Skipping (本地查询和远程读取)
    - 预聚集
    - ……

存储中立

- 公有云，私有云，混合云
- 对象存储 （数据共享，存算分离）按需付费
- 也支持HDFS，NAS



用户数据可靠安全

- 用户数据高可靠实时加解密 (TDE)
- 分布式对象存储多副本多可用区保证数据安全：“一份”数据，避免数据不一致
- 将来Time Travel查询“回收站”数据

用户数据查询效率优化

- 远程访问数据要考虑的点：性能和成本
- 如何解决?
  - 数据和/或辅助信息缓存，同时一致性Hash减少数据移动
  - 读取优化（比如异步并行等）
  - 计算优化（各种功能特性持续优化中）
  - 很多复杂OLAP查询如果不是IO瓶颈，不会受制于它
  - ……

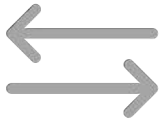
计算引擎之优化器

PieCloudDB Optimizer 是一个基于eMPP架构的云原生分布式优化器，它可以为海量数据集上的复杂OLAP查询提供最优的查询计划。

- 分布式优化器
- 处理复杂OLAP查询
- 云原生优化器

处理复杂OLAP查询

多表连接的最优顺序搜索

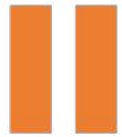


多阶段聚集

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



等等

更多高阶计算功能

- 聚集下推：1.0已经支持，在一些情况下可以十倍百倍更多倍提升

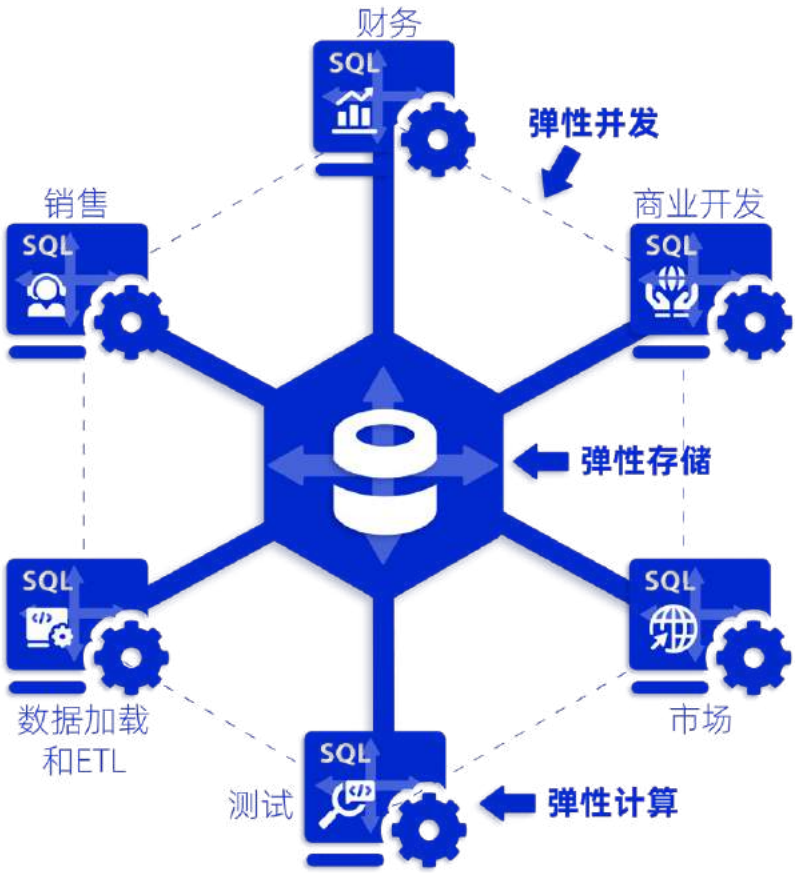
```
SELECT a.i, SUM(a.j) FROM agg_pushdown_t a, agg_pushdown_t b WHERE a.i = b.i GROUP BY 1;
                        QUERY PLAN
------------------------------------------------------------------------------
Gather Motion 3:1  (slice1; segments: 3)
  -> Finalize HashAggregate
       Group Key: a.i
       -> Redistribute Motion 3:3  (slice2; segments: 3)
            Hash Key: a.i
            -> Hash Join
                 Hash Cond: (b.i = a.i)
                 -> Seq Scan on agg_pushdown_t b
                 -> Hash
                      -> Broadcast Motion 3:3  (slice3; segments: 3)
                           -> Partial HashAggregate
                                Group Key: a.i
                                -> Seq Scan on agg_pushdown_t a
```

更多高阶计算功能 (cont.)

- 预计算：很快面世
- Data skipping：文件裁剪支持很快面世
- 更多计算引擎工作在路上：SIMD, runtime filter，late materization，.....

分布式计算引擎

- MPP弹性计算引擎：按需付费
- 租户隔离（彼此不影响）
- 高可用（自动处理各种错误）
- 高并发



PieCloudDB生态

- 各种外表数据源联邦查询组件天然支持（或者需少量修改）
- 各种Postgres/Greenplum组件或者功能天然支持，如In-database AI组件Madlib, json, text等
- 实时ETL/ELT性能对比PieCloudDB 1.0有巨大提升
- 流处理：原生支持kafka数据导入和查询，在PieCloudDB侧导入实现exactly once语义

智能化云原生平台 （数据服务平台）

设计目的

面向用户，做到开箱即用：离数据分析更近，离繁琐操作更远；

面向运维，降低部署门槛：在不同的基础设施都能发挥实力；

面向管理，让管理更轻松：让数据分析运行更透明；

智能化云原生平台：面向用户、开箱即用

PieCloudDB 是这么来帮助我们的用户的

- 降低上手难度
  — 让用户享受数据分析的乐趣
- 使用门槛低了
  — 扩大平台受众
- 让更多用户离数据更近
  — 离繁琐操作更远

面向运维 部署运维难度小

- 全面拥抱容器化技术，可以适配多种环境
- 已支持私有信创环境和多云环境
- 既实现私有环境离线部署，也可充分利用公有云技术设施
- 数据库维护平台托管

面向管理 多个维度轻松管控

PieCloudDB支持

- 一个数仓多个计算集群同时运行
- 针对不同用户业务负载或者不同场景，可以选择不同集群进行数据计算

云原生平台支持

- 快速启动集群，随时可以关停，随时可以回收
- 结合集群操作记录，用户可以用最低的成本完成数据分析

云原生平台同时提供

- 根据角色访问模型设计的权限系统，所见即可管
- 无论是平台功能还是数据库权限都可以在平台操作

PieCloudDB 的将来

- 理想的PieCloudDB：可靠、高效、简单、完备的SQL数据平台，让用户能专注于应用
- 不论存储、计算、生态还是智能平台都还有不少有挑战性的事情
- 我们需要优秀人才的加入（学习动手能力、创新能力、自驱、团队精神）

关注OpenPie公众号

获得更多资讯

加入PieCloudDB社群

了解更多干货


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
GoldenDB
MatrixDB
DynamoDB
FastData
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
神舟通用
KingbaseES
TDSQL
TiDB
UbiSQL
StarRocks
2010
2014
2018

THANKS