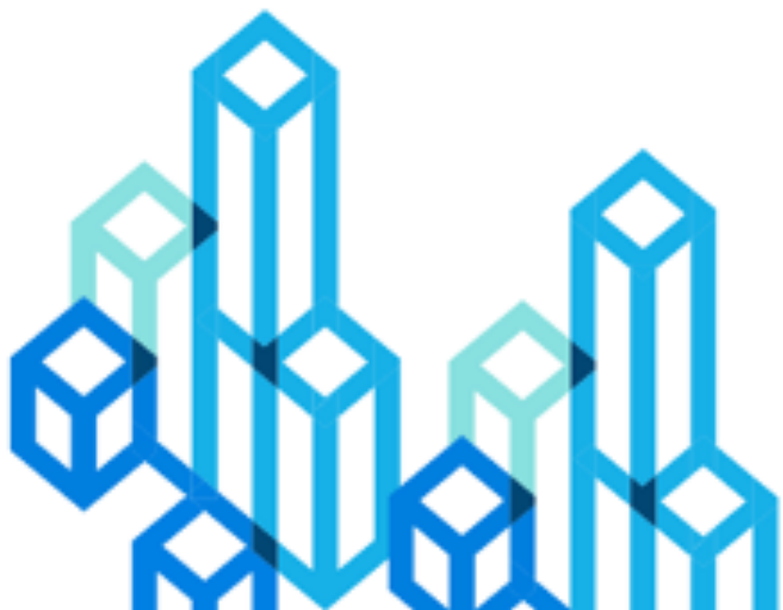
HashData企业级云端数据仓库

简丽荣，HashData



个人简介

简丽荣，HashData, 联合创始人&CEO

- Apache HAWQ，Greenplum Database 的早期核心成员
- IBM中国研究院，雅虎北京研发中心，Pivotal北京研发中心
- 云计算、大数据和分布式数据库
- 10+国际专利，多篇学术会议论文，包括SIGMOD和INFOCOM
- 2008年清华本科，2010年港科大硕士
- 日常爱好：
  - 网球大满贯和F1直播
  - 读书、跑步和冥想

目录：

公司简介

HashData

致力于推动数据分析平民化

公司定位

专注于云端数据仓库的初创公司，总部位于北京，多地设立办事处。

行业经验

为金融、电信、能源、交通等重要行业头部客户，解决最具挑战性的数据仓库难题 。

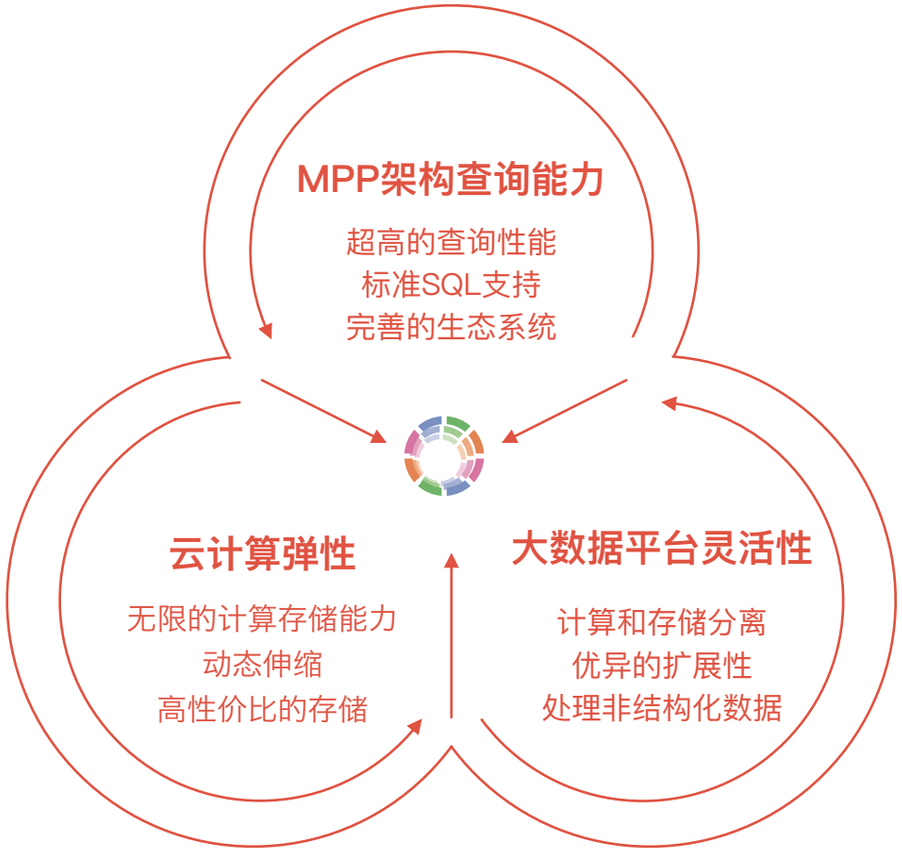
核心团队

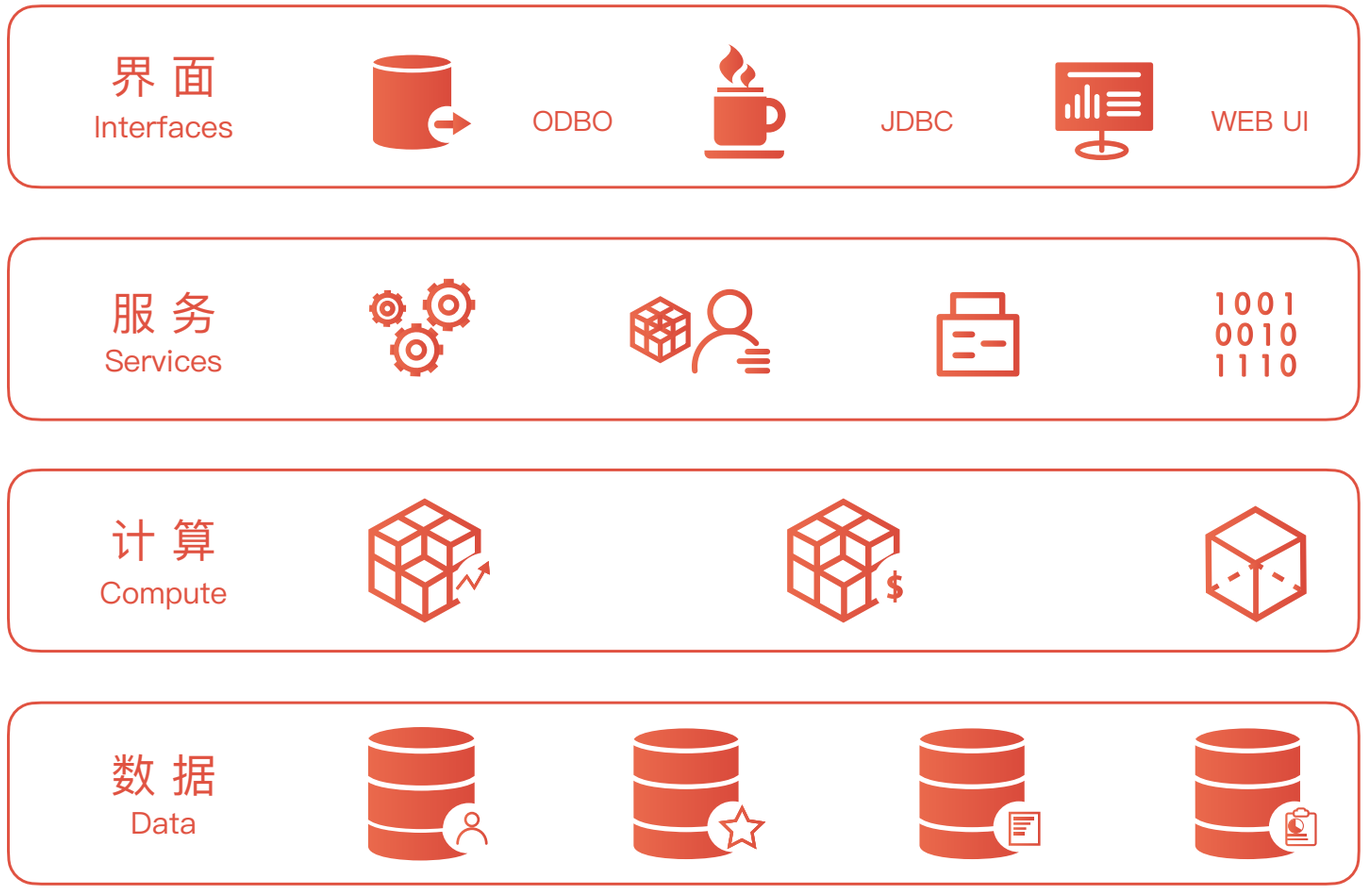
核心团队主要由来自Pivotal、Teradata、IBM、Yahoo!、Oracle和华为等公司资深的云计算、分布式数据库和大数据专家组成

数据量处理规模

千万级的数据库对象
100+PB数据量
数千个并发应用
每天1亿+的复杂SQL查询

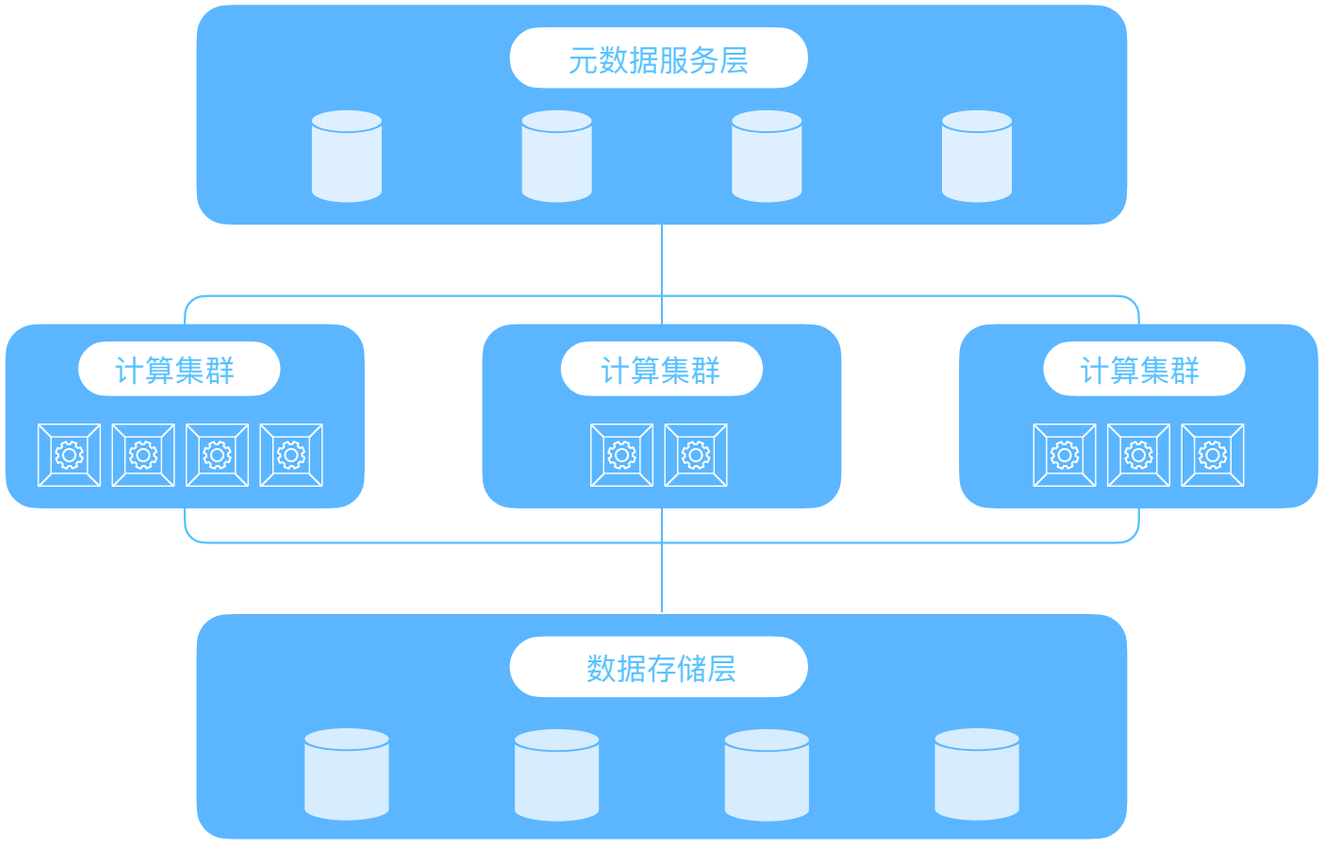
HashData 设计理念





产品介绍

HashData数据仓库



完全托管的PB级数据仓库服务

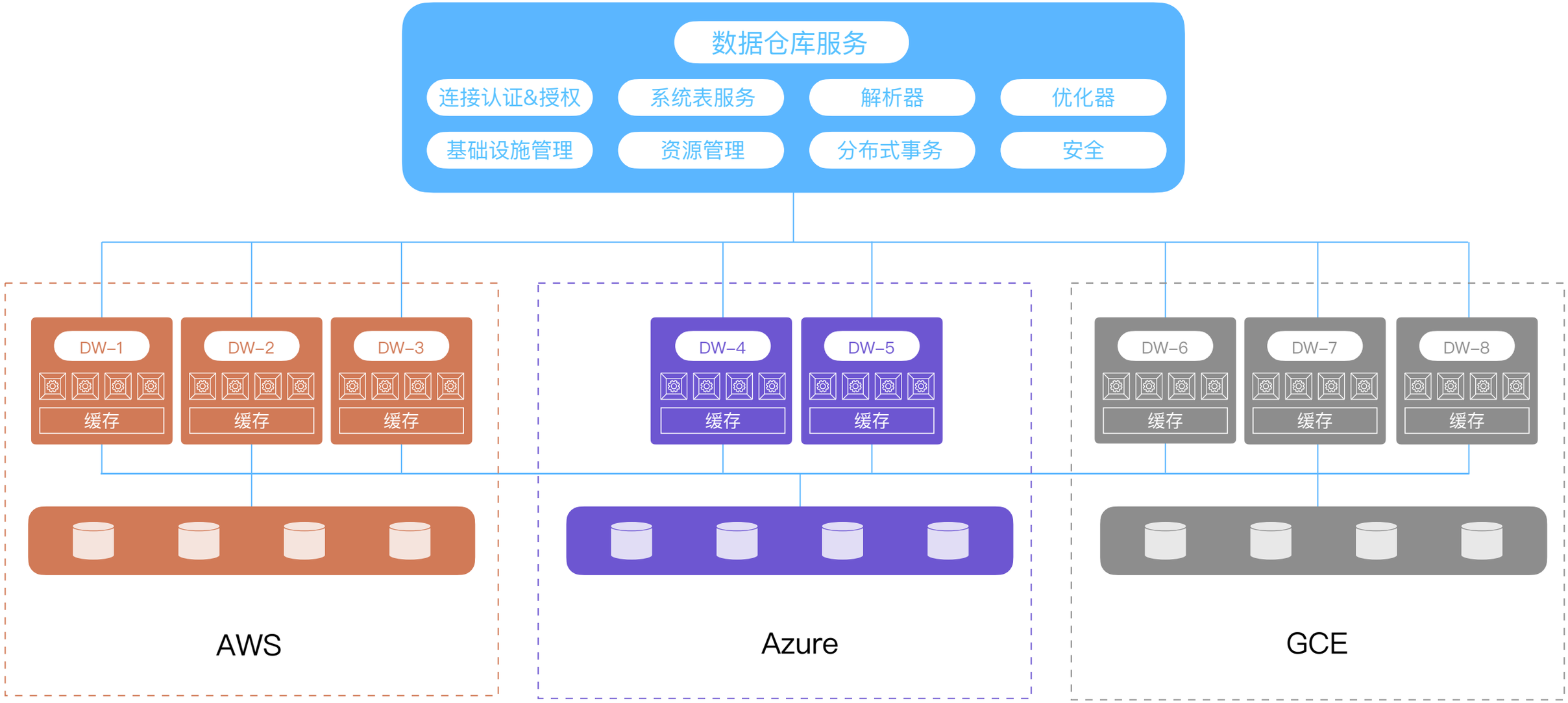
分析接口开放

- 100%兼容开源PostgreSQL和Greenplum Database；

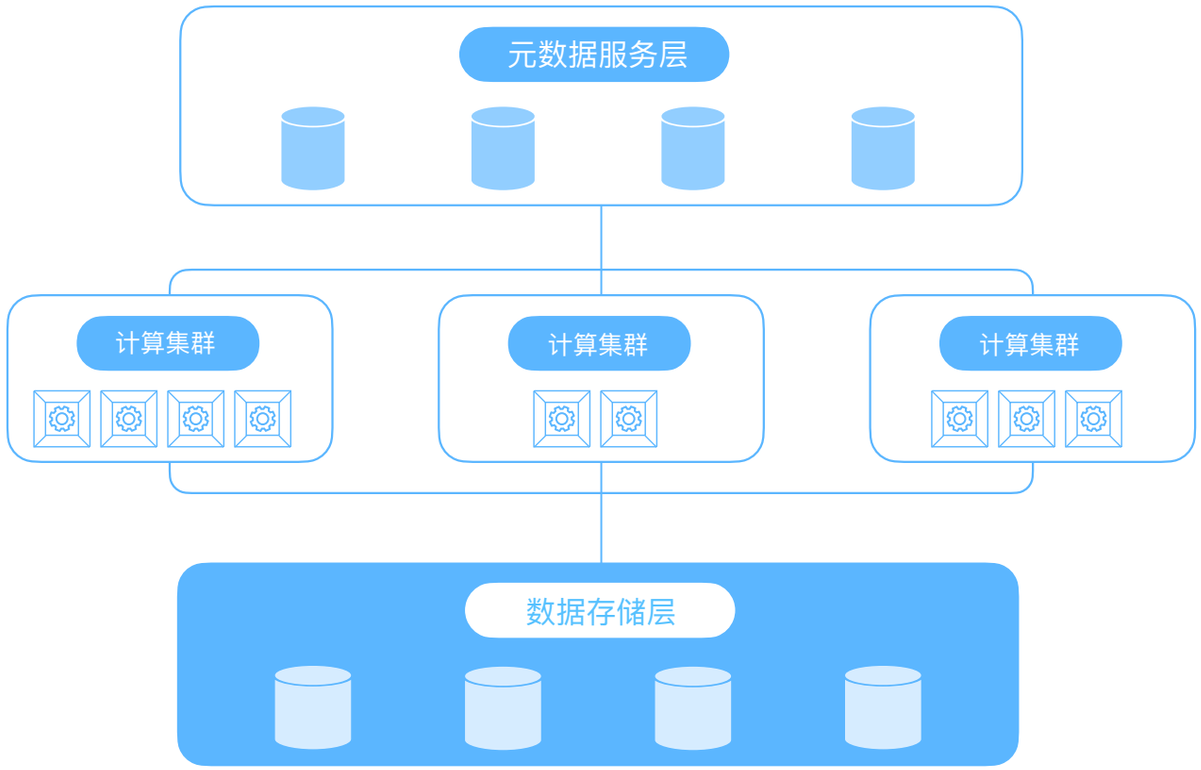
系统架构云原生

- 计算和存储分离；
- 对象存储作为数据持久层；
- 独立元数据服务：在线升级和扩容；
- 一致性哈希的数据分布策略：秒级扩容；

HashData系统架构



数据存储层



目标数据

- 用户表数据；
- 查询结果；
- 运行时临时数据；

对象存储

优　点：
- AWS S3，阿里云OSS，腾讯云COS、华为云OBS，UCloud UFile、金山云KS3；
- RESTFUL API；
- 高可用性和高持久性；
- 按需付费；

缺　点：
- 直接远程访问2～3倍的性能下；
  - 解决方案：本地缓存降低性能惩罚；
- 一旦写入不可更改；
  - 解决方案：精妙实现数据库增删改查操作；

存储访问层优化

- 存储格式：列式存储（每列单独文件）＋多种压缩算法；
- 访问方式：多线程、多个存储桶（Bucket）、动态调整数据包大小；

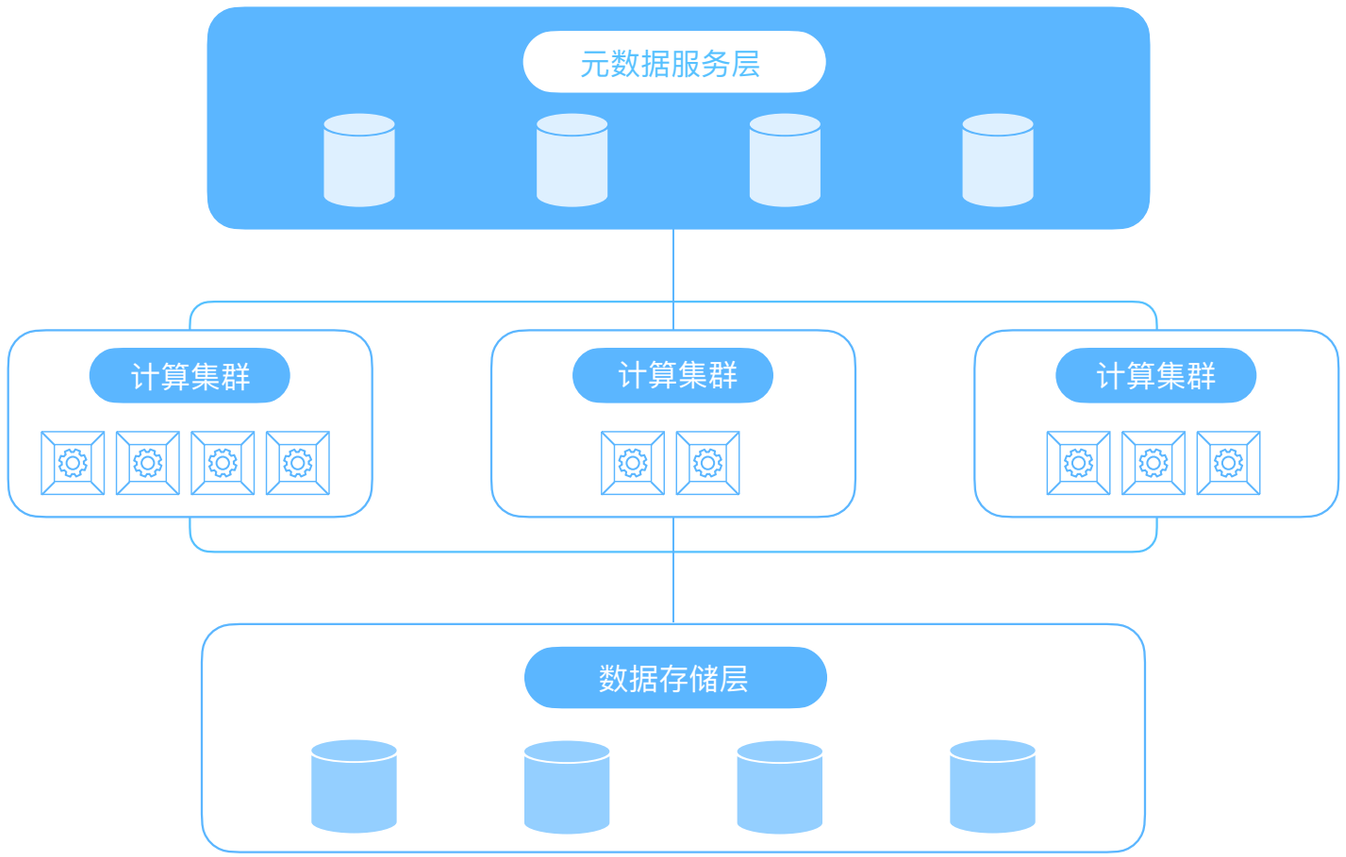
对象存储优势

《Top 5 reasons for Choosing S3 over HDFS》

—— Databricks的官方博客

类型	S3	HDFS	比较
弹性	支持	不支持	S3更加弹性
价格（每TB每月）	US$23	US$206	10x
可用性	99.99%	99.9%（估计）	10x
持久性	99.999999999%	99.9999%（估计）	10x+

元数据服务层



目标数据

- 表到对象的映射；
- 数据库数据字典；
- 统计信息；
- WAL日志；
- 索引信息；
- 部分用户表数据；

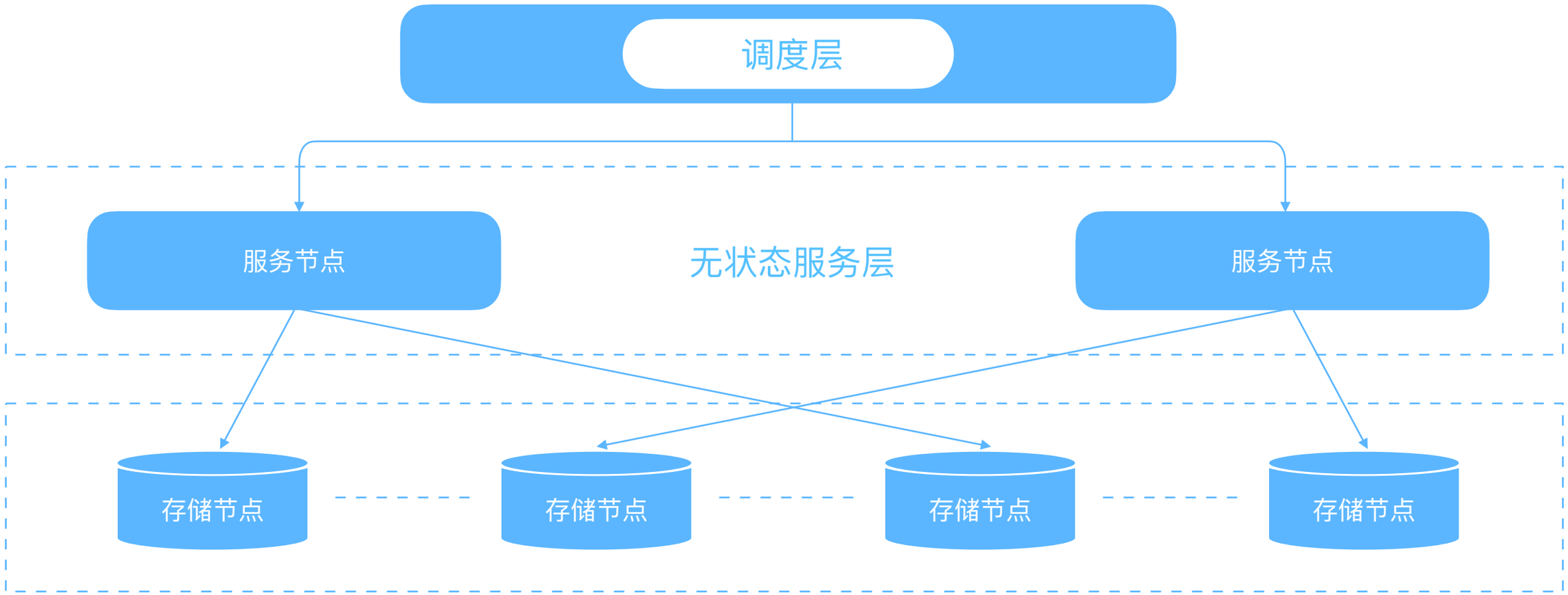
数据持久化

- 分布式K–V数据库；

云服务

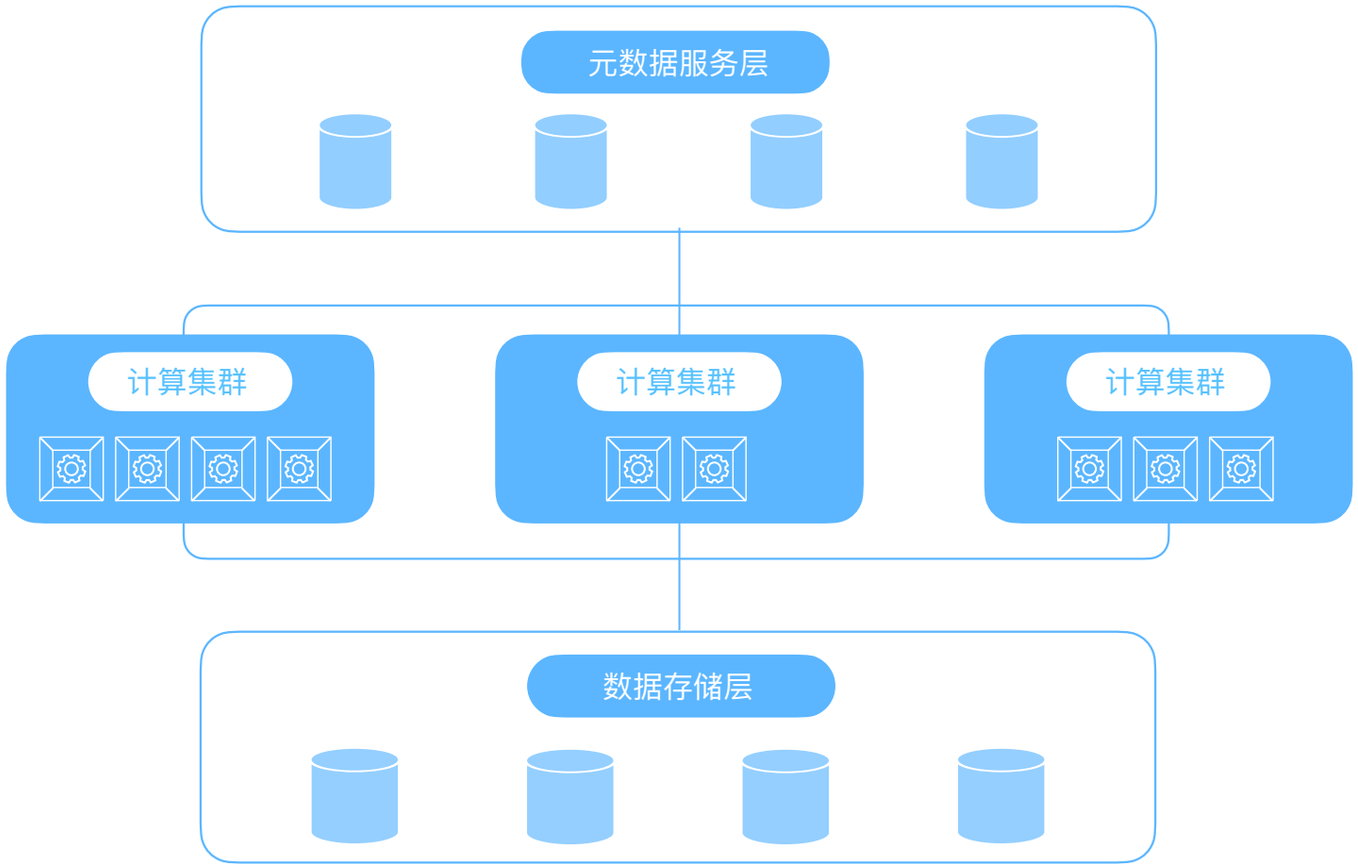
- 访问控制、查询优化、分布式事务、锁管理；
- 集群监控检查、故障恢复、弹性伸缩；

元数据服务层架构图



全球分布式K–V存储

计算集群



资源形态

- 物理服务器；
- 虚拟机；
- 容器；

纯粹的计算资源

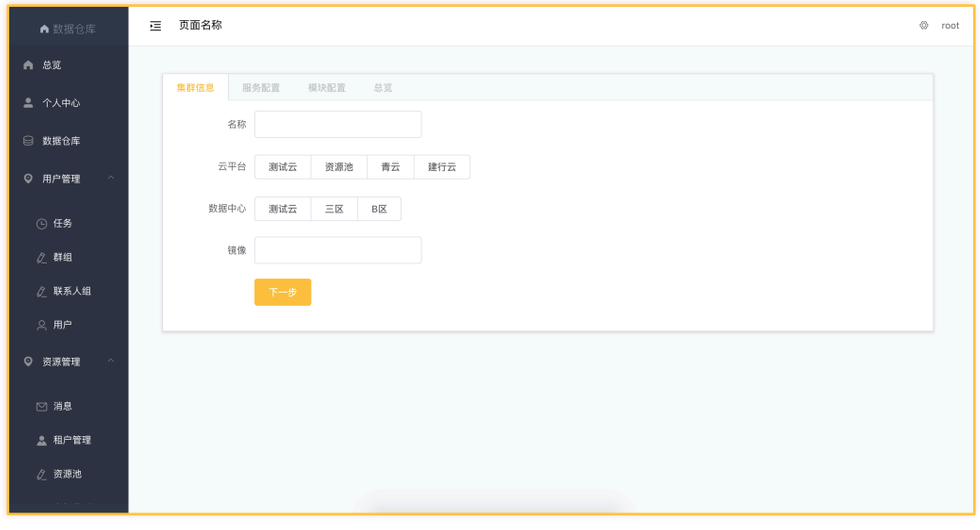
- 按需创建、删除和纵向伸缩；
- 多个虚拟机组成一个数仓集群；
- 集群间性能完全隔离；
- 不需要时释放整个集群资源；

缓存进程

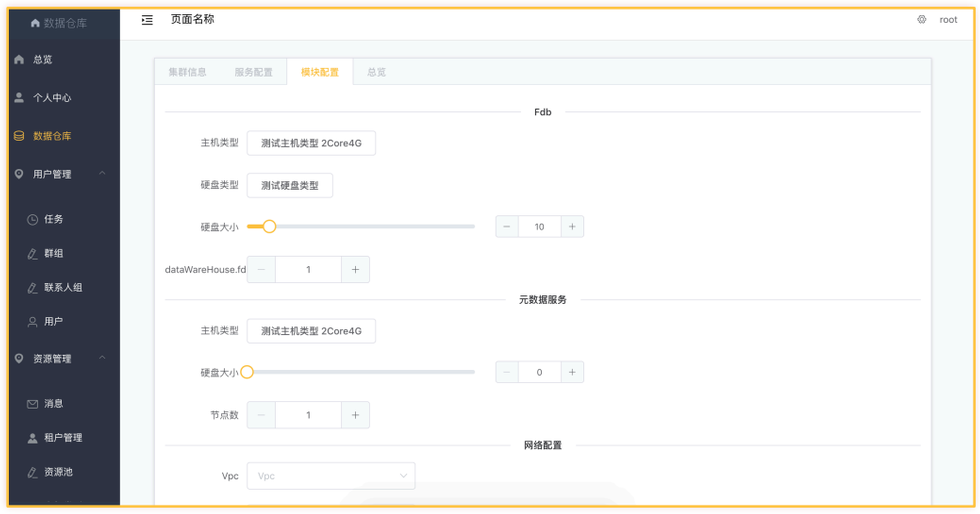
- 本地SSD作为缓冲介质；
- 同一个集群的所有缓存进程组成分布式缓存；

HashData产品形态

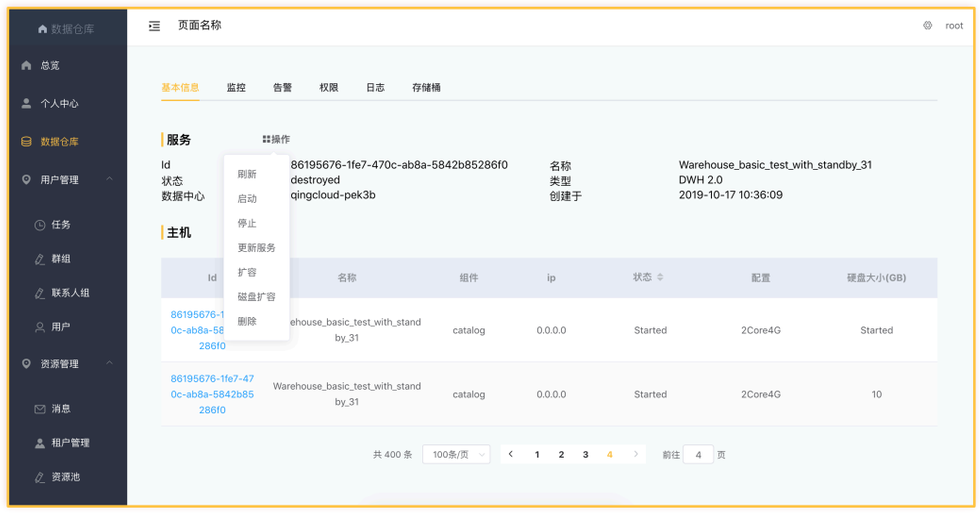
登陆界面



选择云平台和数据中心



集群配置



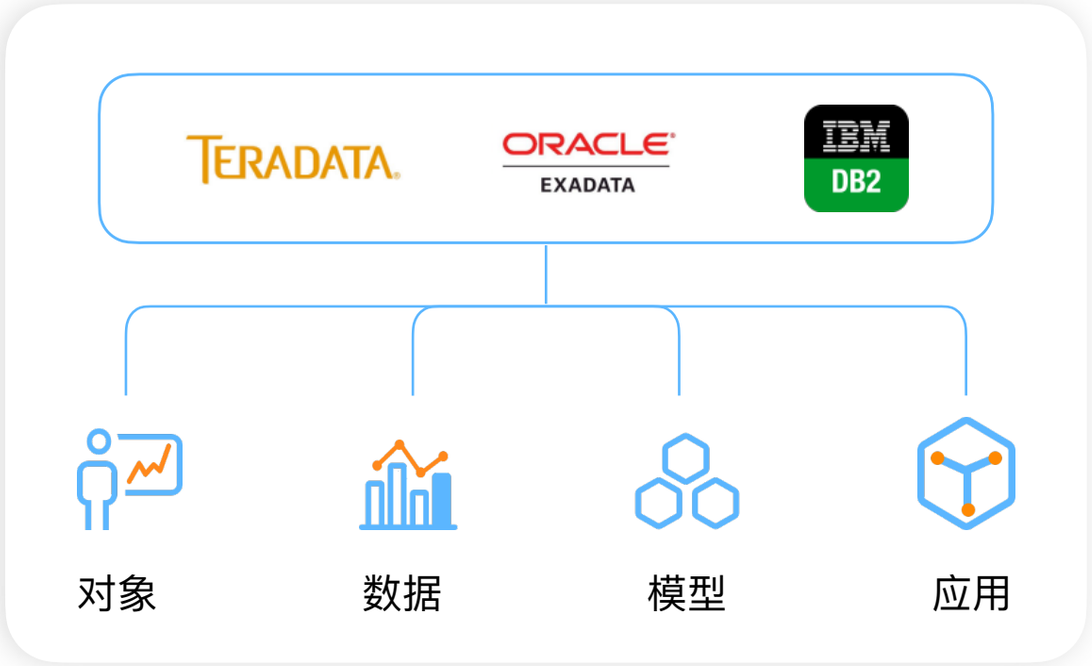
集群操作

HashData

解决方案

HashData

数据仓库



完善迁移工具链

高性能

- MPP架构、列式存储；

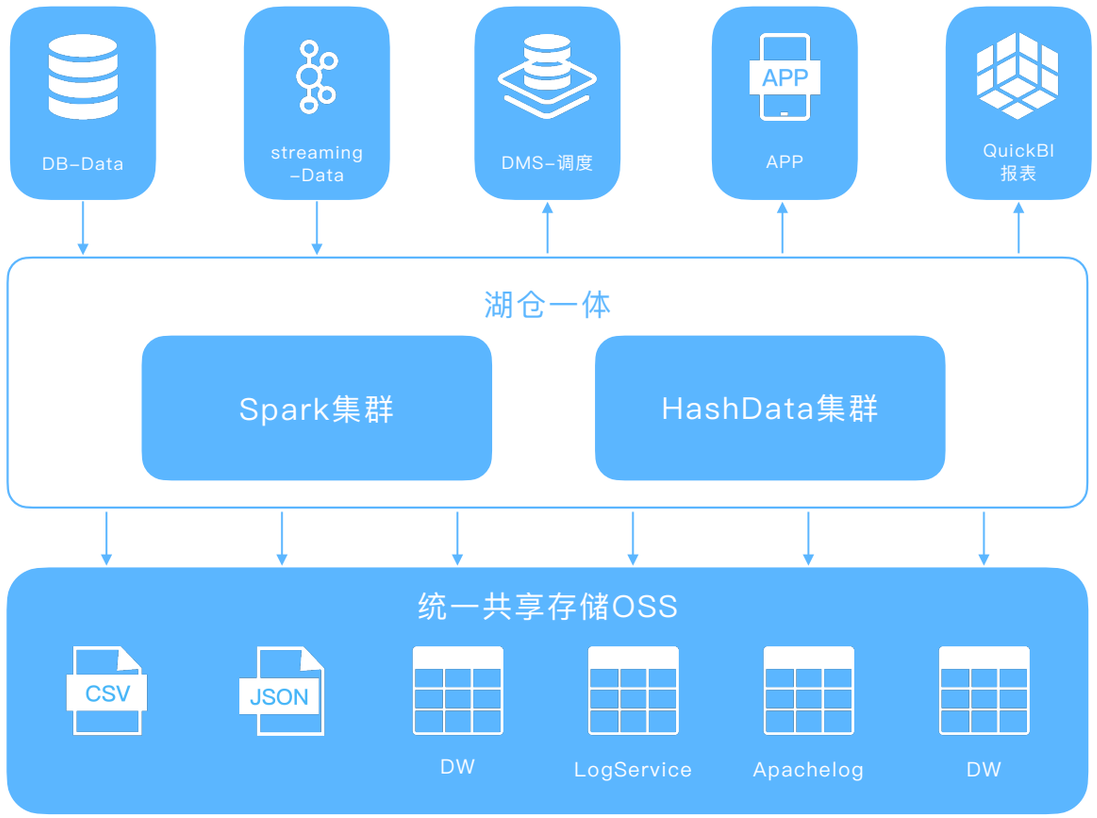
低门槛

- ACID、CRUD、ANSI SQL 2008、OLAP 2003；

数据安全

- 数据库安全管理机制，持久化数据加密；

数据湖



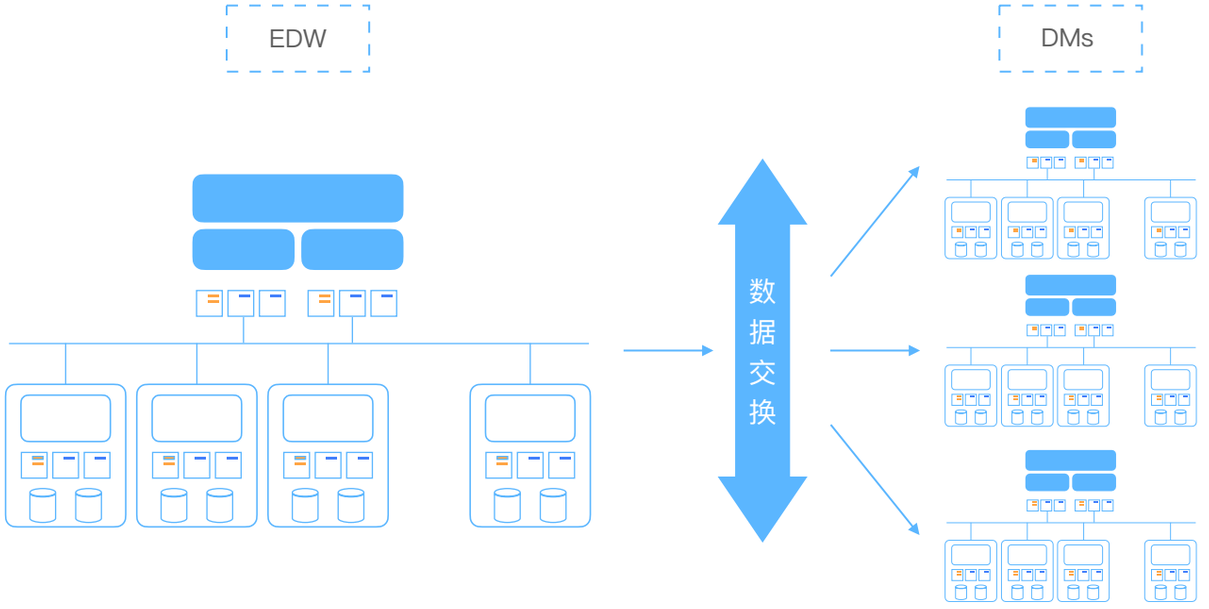
Spark – – > HashData

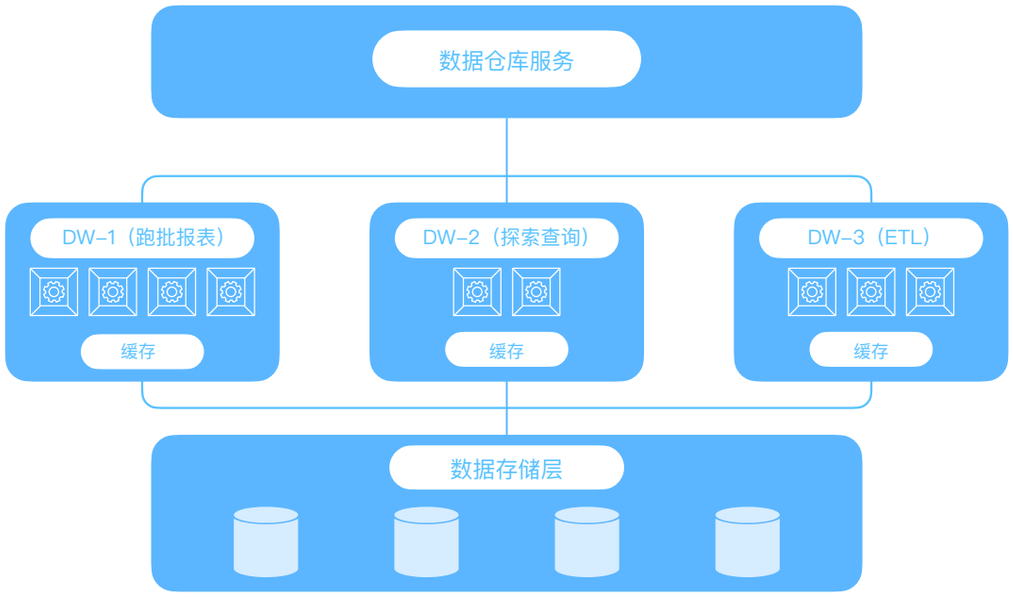
- Spark读对象存储上原始数据，并进行ETL处理；
- Spark将ETL结果写回对象存储（通过湖仓一体SDK）：
  - 数据按照湖仓一体格式保存；
  - 更新HashData对应的元数据；

Spark <– – HashData

- Spark读取HashData保存在对象存储的数据（通过湖仓一体SDK）：
  - 访问HashData元数据，获取对象存储上的文件信息；
  - 直接读取解析对象存储上的文件信息；
- 根据应用需求，Spark可以将机器学习的结果模型写回到HashData；

数据共享





原理

- Shared-Nothing：
  每个集群的数据保存在每个计算节点本地的磁盘；
- 集群与集群之间数据无法做任何有效共享；

后果

- 数据孤岛；数据实时性差；
- 大量数据拷贝操作、数据严重冗余；

原理

- Shared-Everything：
  任何一个集群都能够访问任何一份数据；
- 集群之间保证事务的强一致性；

后果

- 统一数据湖，数据完全共享；
- 完全消失数据拷贝、冗余，数据实时性强；

客户案例

HashData

某大型国有银行分布式数据仓库

集群规模庞大

3000+服务器，10+PB数据；

30+生产集群；

单集群100业务场景广泛+服务器；

业务场景广泛

90+%大数据分析业；

20+应用场景；

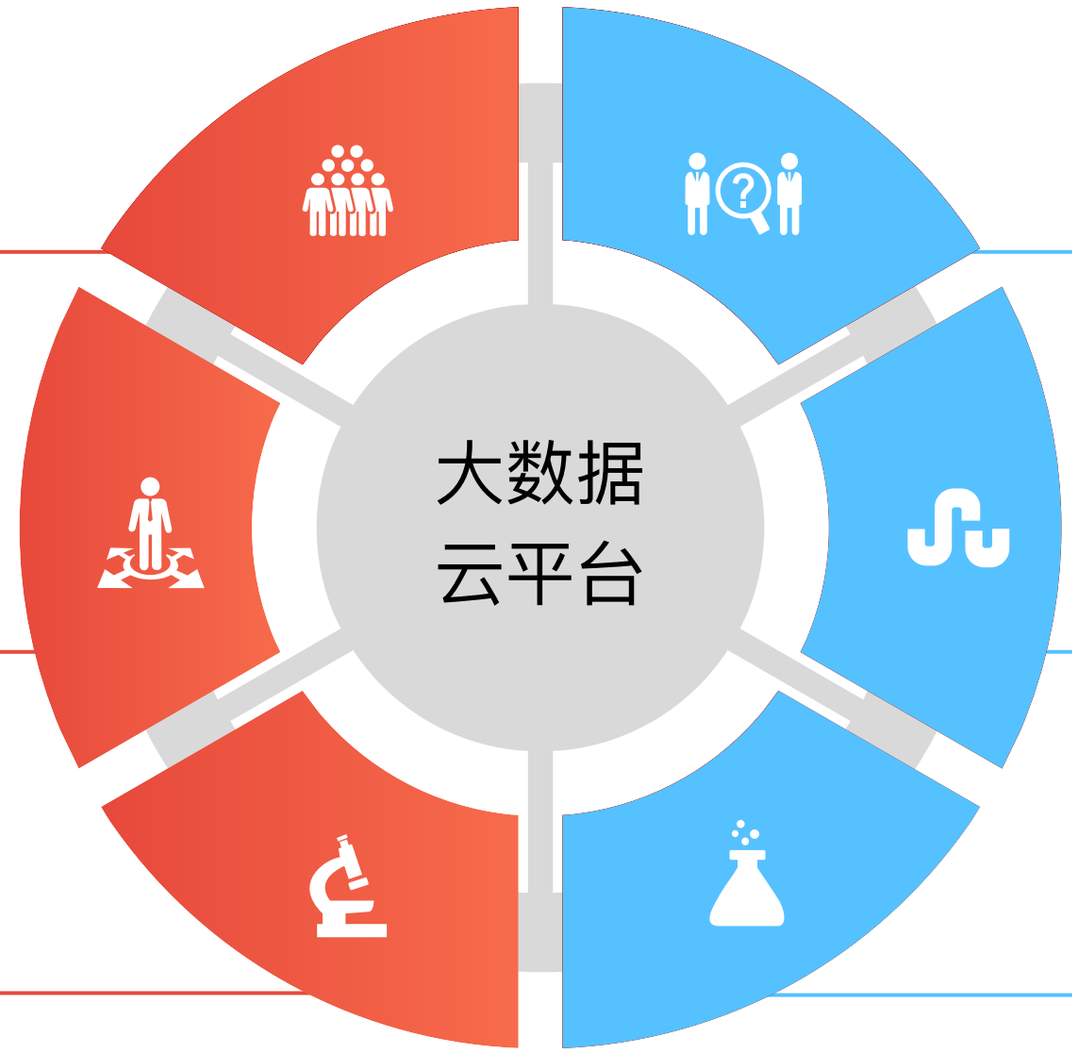
离线跑批到在线实时；

外围工具完善

P9调度系统；

自定义扩展；

应用开发技术积累；



零运维自服务

自动化安装部署；

实时监控和自愈；

多租户；

多维度弹性

管理、计算和存储三者分离；

纵向伸缩和横向伸缩；

秒级扩容；

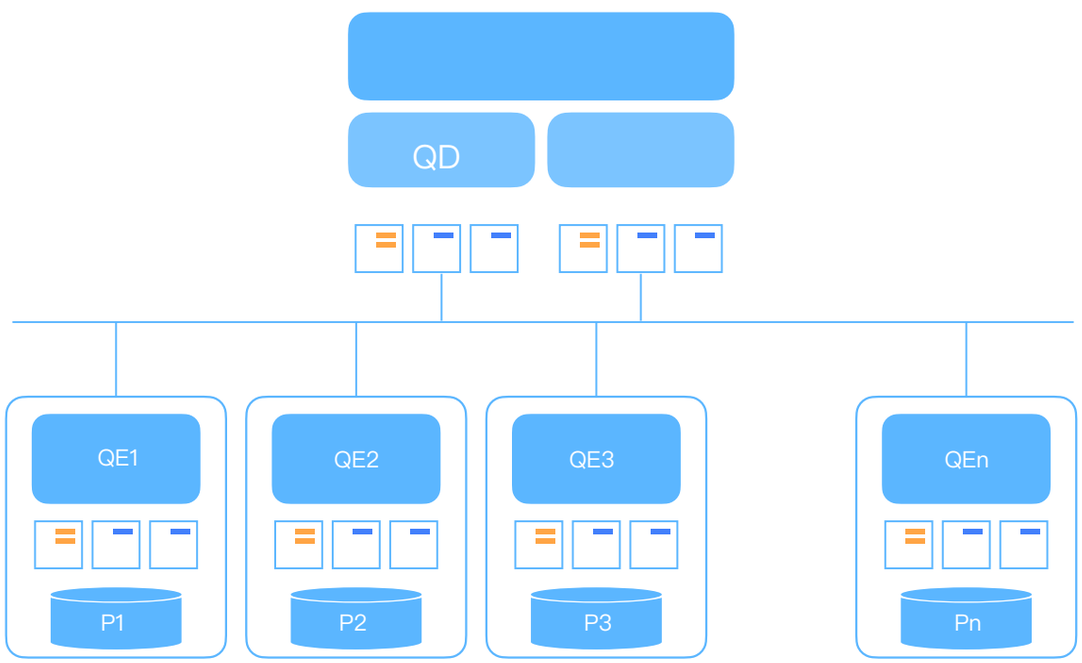
多集群高并发

多集群统一数据存储；

混合负载；

数据共享，性能隔离；

已有MPP系统面临的挑战（1）：高并发

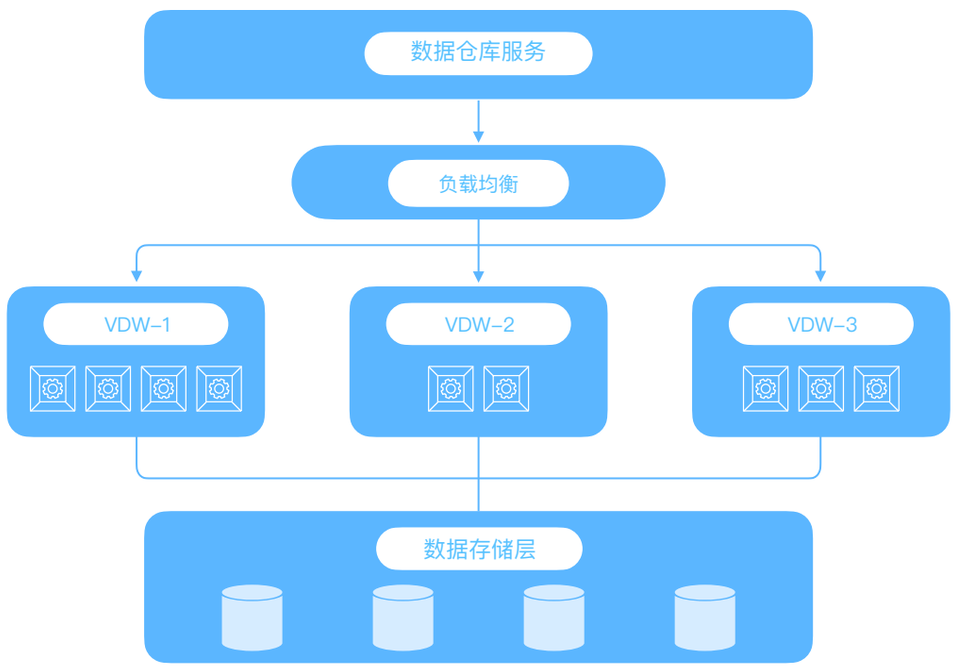


原理

- 每个计算节点参与到每条查询的执行中；
- 系统支持的并发查询数量由单个计算节点的硬件资源决定；
- 扩大集群规模不能提高并发查询数量，虽然能够降低单条查询的延迟（有时候因为调度的开销，甚至可能比原来慢）；

后果

- 反洗钱业务需要六个一摸一样的集群，ETL执行六次，数据重复六份；



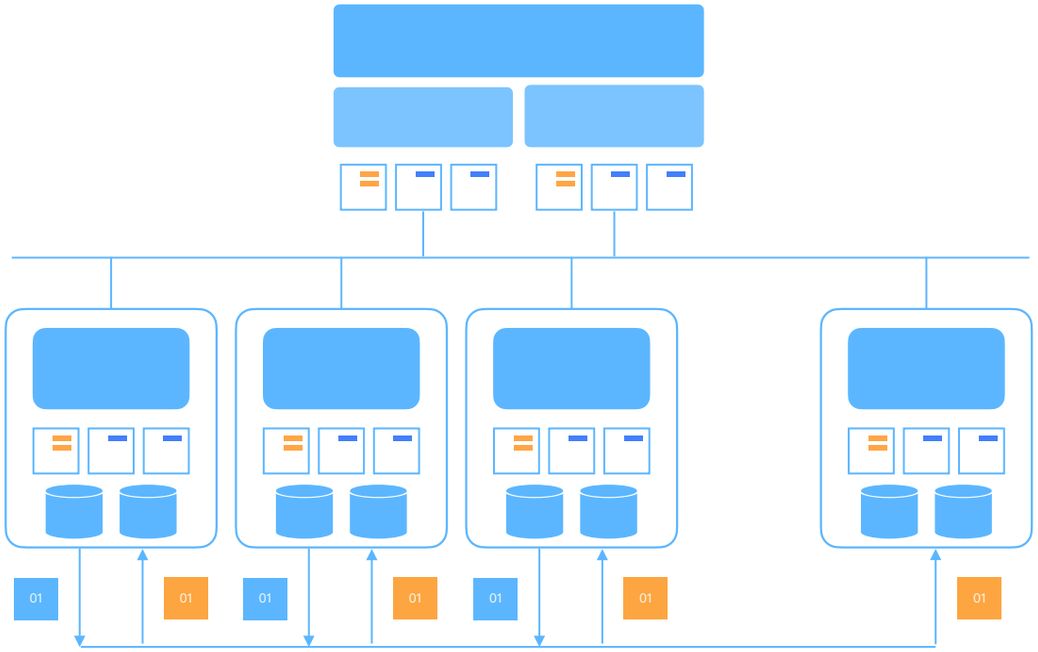
原理

- 多集群共享统一元数据、统一数据存储；
- 集群间不竞争CPU、内存和IO资源；
- 多个物理集群组成一个逻辑集群；

后果

- 6个计算集群，ETL执行一次，数据只有一份；

已有MPP系统面临的挑战（2）：弹性

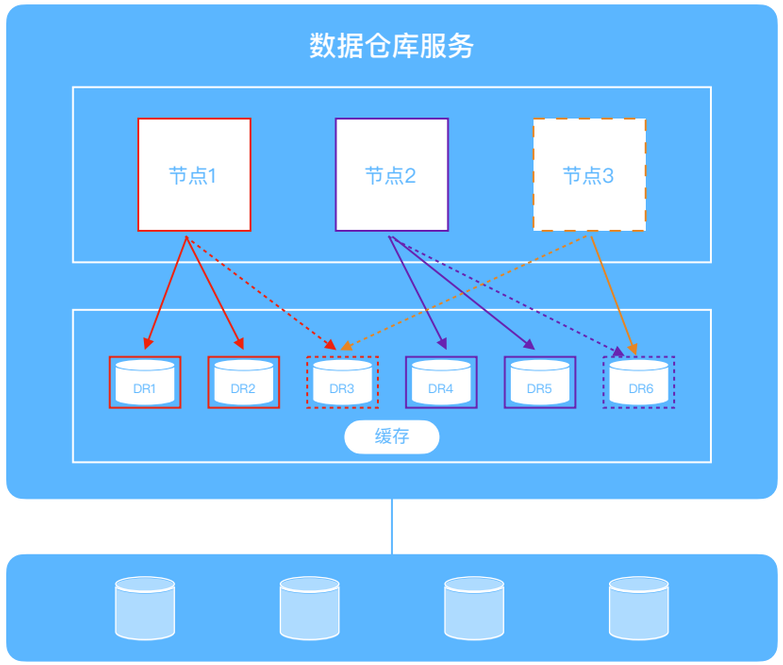


原理

- 数据按照哈希取模的方式均匀分布在各个节点；

后果

- 增加一个节点，所有原有数据都要读出来，重新哈希分布，再次写回磁盘，引入大量磁盘IO和网络IO；



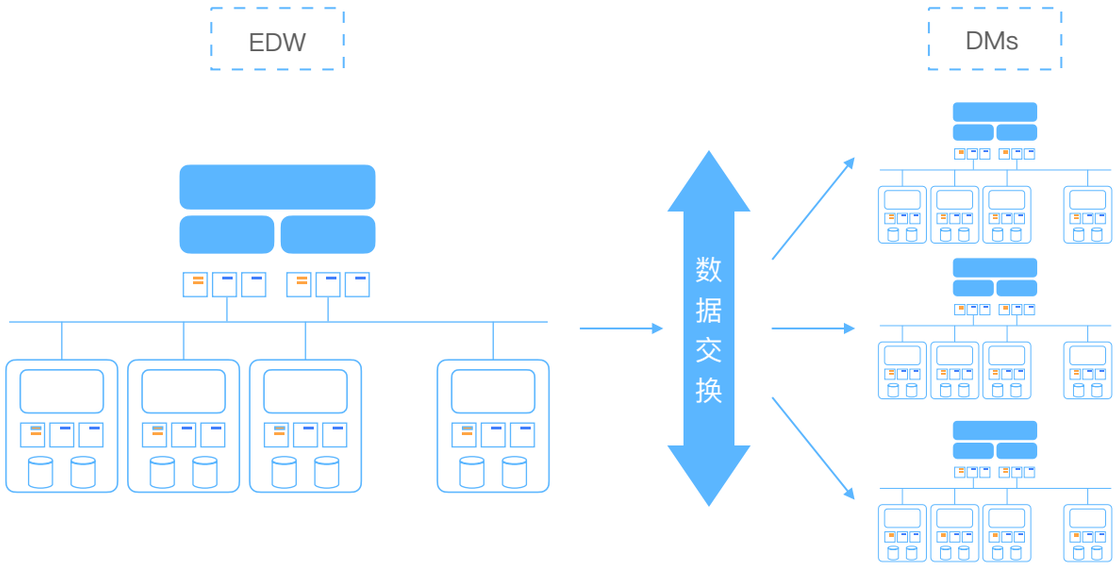
原理

- 一致性哈希避免数据重新逻辑分组；
- 共享存储避免数据重新物理分布；

后果

- 秒级扩容；

已有MPP系统面临的挑战（3）：数据共享

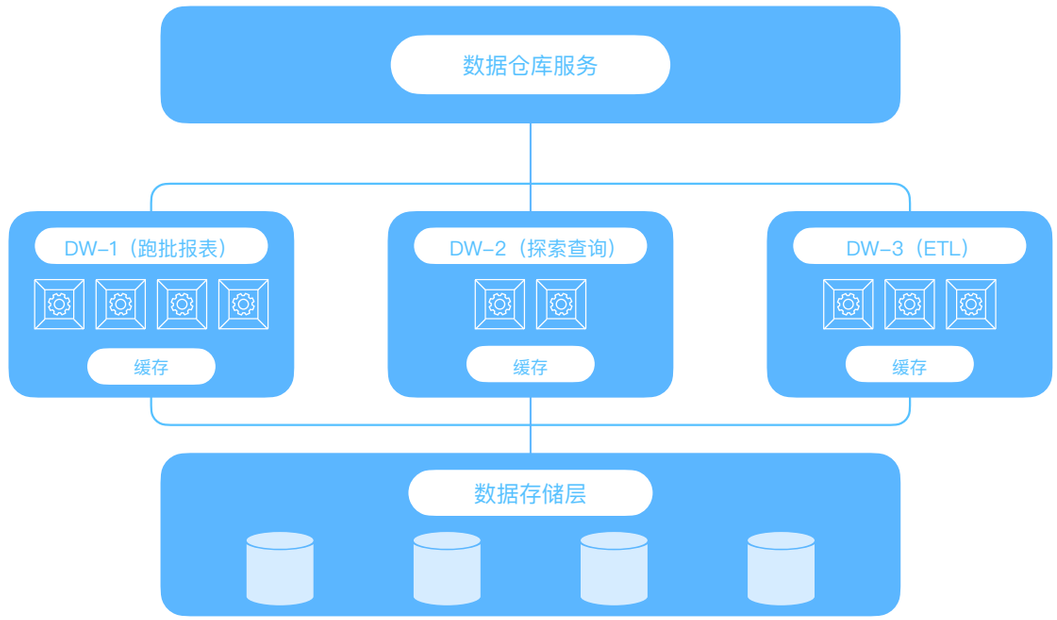


原理

- Shared–Nothing：
  每个集群的数据保存在每个计算节点本地的磁盘；
- 集群与集群之间数据无法做任何有效共享；

后果

- 数据孤岛；数据实时性差；
- 大量数据拷贝操作、数据严重冗余；



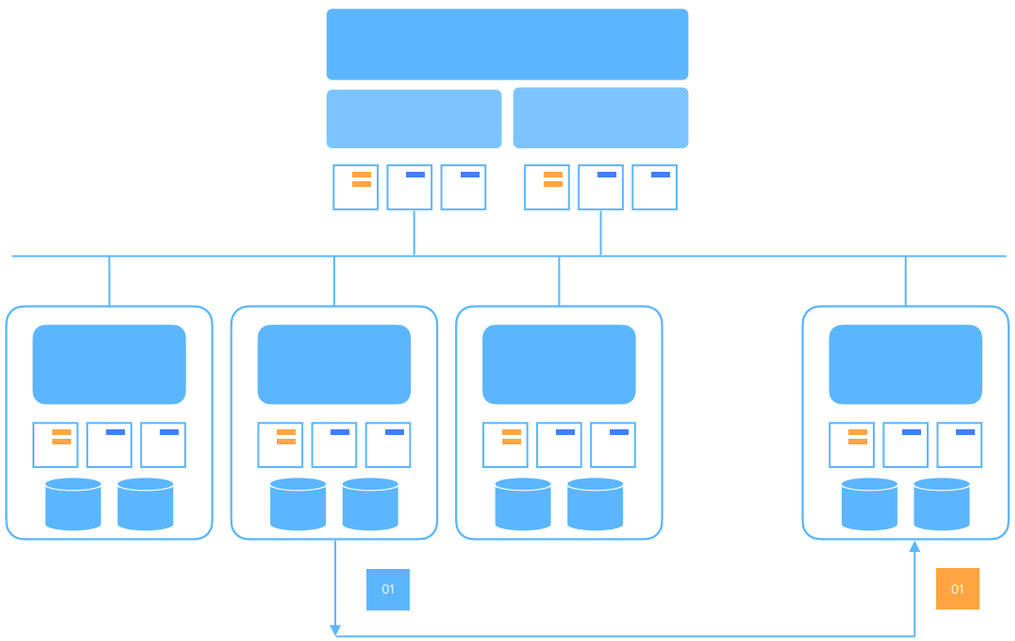
原理

- Shared–Everything：
  任何一个集群都能够访问任何一份数据；
- 集群之间保证事务的强一致性；

后果

- 统一数据湖，数据完全共享；
- 完全消除数据拷贝、冗余，数据实时性强；

已有MPP系统面临的问题（4）：高可用

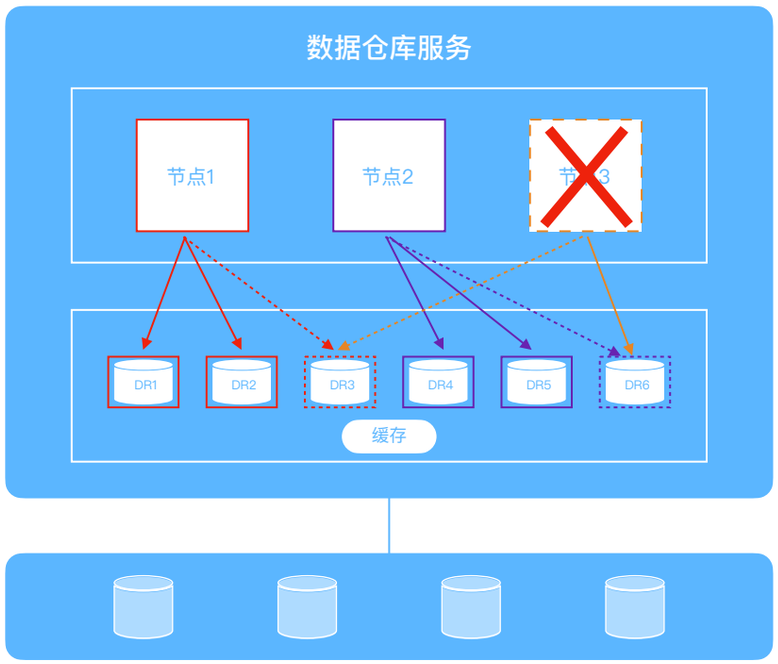


原理

- 计算节点失败，任务调度到Mirror节点；
- 新节点替代失败节点，数据需要从Mirror节点同步到新节点；

后果

- Mirror节点负载加倍，成为系统瓶颈；
- 新节点的数据恢复窗口很长；



原理

- 数据存储在共享存储上面；
- 计算节点与数据块的对应关系动态调整；

后果

- 没有所谓的Mirror节点；
- 分钟级新节点恢复；

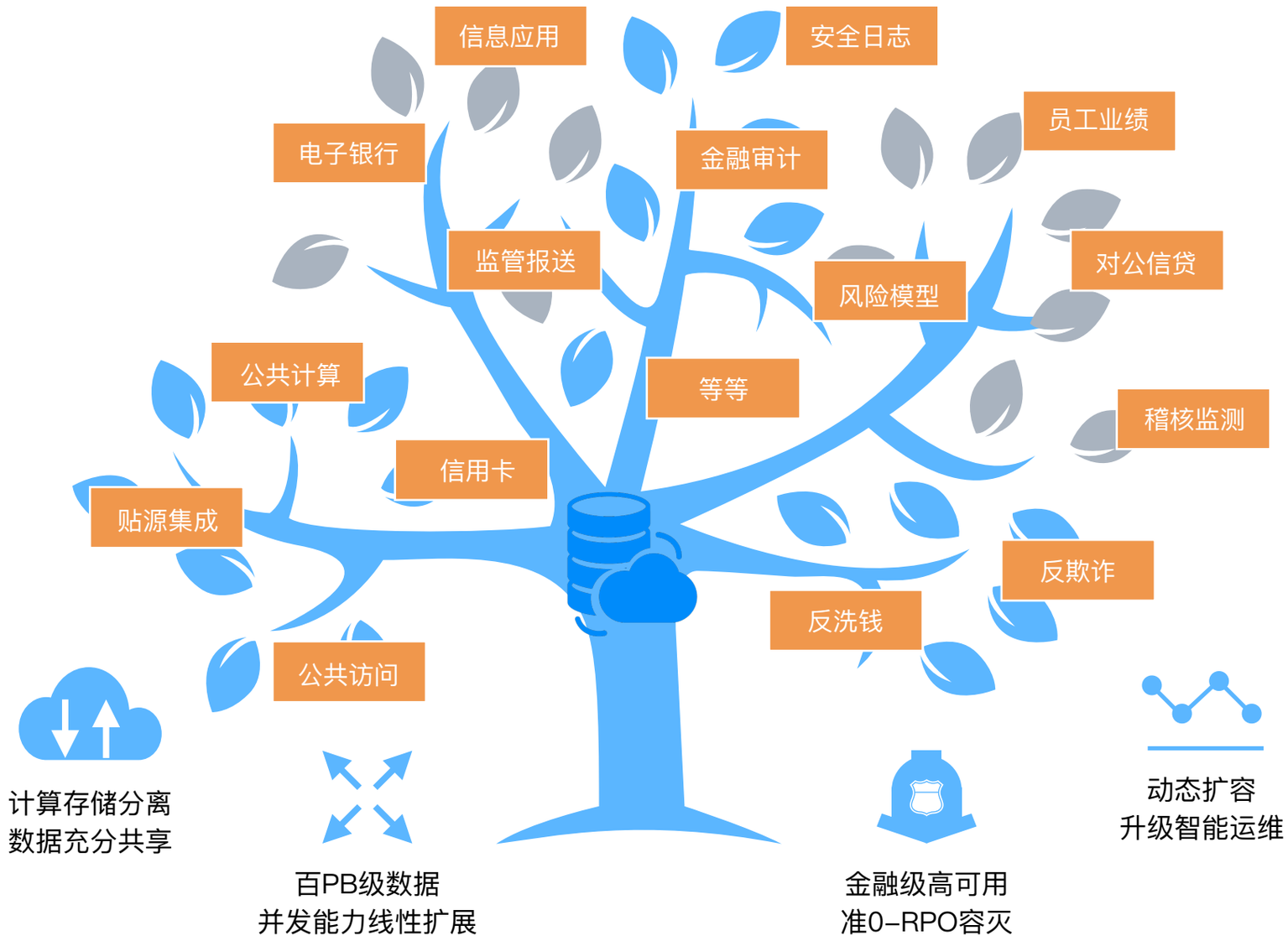
项目成果

内容涵盖：

- 20PB数据
- 数百万个作业
- 几百个应用
- 数千并发

成果效果：

- 并发任意扩展
- 数据充分共享
- 计算存储分离
- 高效数据处理



麦思博(msup)有限公司是一家面向技术型企业的培训咨询机构，携手2000余位中外客座导师，服务于技术团队的能力提升、软件工程效能和产品创新迭代，超过3000余家企业续约学习，是科技领域占有率第1的客座导师品牌，msup以整合全球领先经验实践为己任，为中国产业快速发展提供智库。

高可用架构主要关注互联网架构及高可用、可扩展及高性能领域的知识传播。订阅用户覆盖主流互联网及软件领域系统架构技术从业人员。 高可用架构系列社群是一个社区组织，其精神是“分享+交流”，提倡社区的人人参与，同时从社区获得高质量的内容。