不同数据库的存算分离有何不同

原创 胖头鱼的鱼缸 胖头鱼的鱼缸 2025年04月07日 16:27 四川

数据库管理-第311期 不同数据库的存算分离有何不同（20250407）

作者：胖头鱼的鱼缸（尹海文）
Oracle ACE Pro: Database
PostgreSQL ACE Partner

10年数据库行业经验
拥有OCM 11g/12c/19c、MySQL 8.0 OCP、Exadata、CDP等认证
墨天轮MVP，ITPUB认证专家
圈内拥有“总监”称号，非著名社恐（社交恐怖分子）

公众号：胖头鱼的鱼缸
CSDN：胖头鱼的鱼缸（尹海文）
墨天轮：胖头鱼的鱼缸
ITPUB：yhw1809。
除授权转载并标明出处外，均为“非法”抄袭

数据库的存算分离架构，是一个经久不衰的架构，其中最为出名的当属Oracle RAC集群，但是在不同的数据库中存算分离架构又是有些许不同的，

数据库管理-第311期 不同数据库的存算分离有何不同（20250407）

作者：胖头鱼的鱼缸（尹海文）
Oracle ACE Pro: Database
PostgreSQL ACE Partner

10年数据库行业经验

拥有OCM 11g/12c/19c、MySQL 8.0 OCP、Exadata、CDP等认证
墨天轮MVP，ITPUB认证专家
圈内拥有“总监”称号，非著名社恐（社交恐怖分子）

公众号：胖头鱼的鱼缸
CSDN：胖头鱼的鱼缸（尹海文）
墨天轮：胖头鱼的鱼缸
ITPUB：yhw1809。

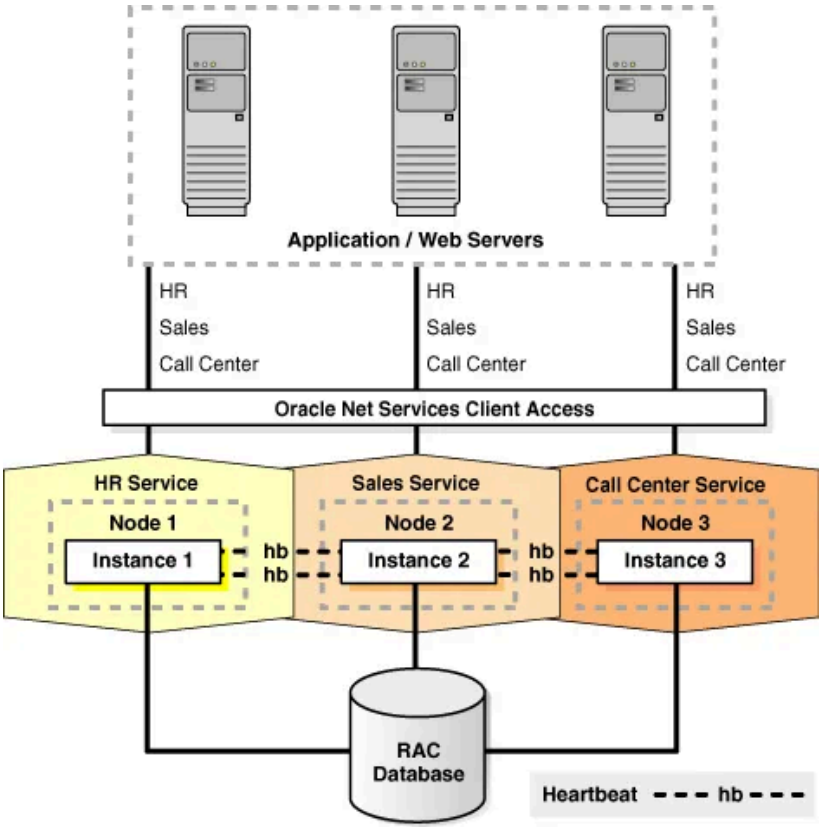
数据库的存算分离架构，是一个经久不衰的架构，其中最为出名的当属Oracle RAC集群，但是在不同的数据库中存算分离架构又是有些许不同的，不同的数据库或者说不同的场景对存算分离架构的概念其实还有有些许不同的：

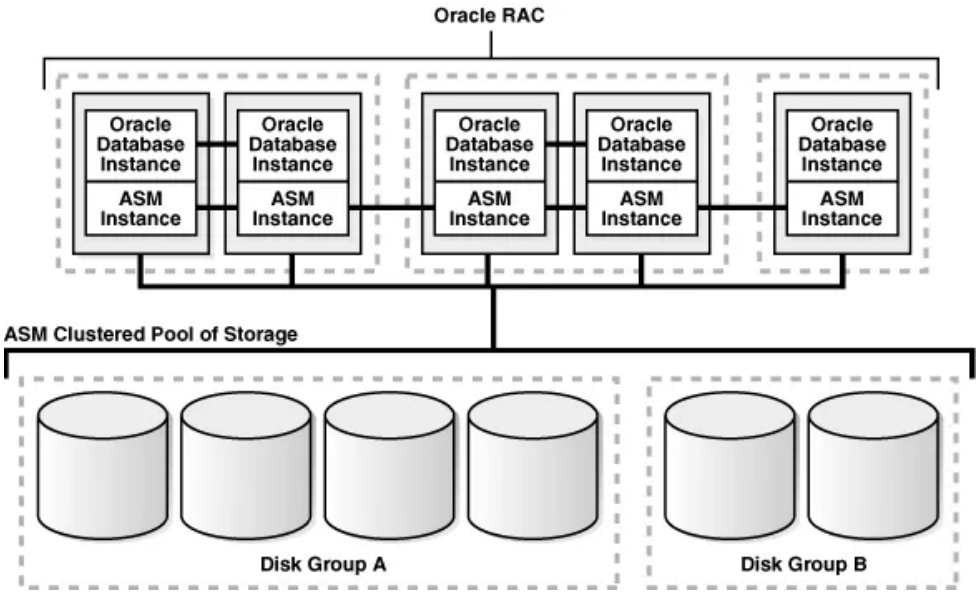
- 对于分布式数据库来说更多的是将计算功能和存储引擎分离，二者都是利用**本机资源**
- 集中式数据库或者从硬件角度来看则是将数据存放在**共享存储**中，计算放在**本机**，数据库可以通过映射为本机资源的方式来使用共享存储

本期根据不同数据库和不同场景的特点，进行简单解析。

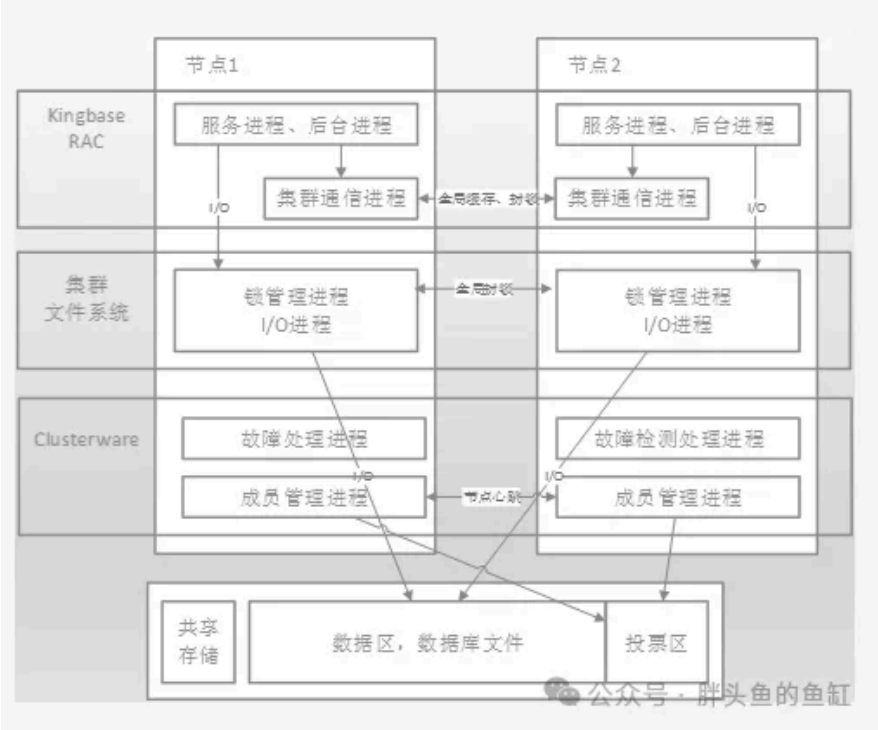
1 基于共享存储

在Oracle RAC集群中，GRID组件的ASM（Automatic Storage Management，自动存储管理）将来自于共享存储映射的磁盘以不同的冗余模式组建为不同的磁盘组，在磁盘组内向数据库各节点提供统一的存储能力：

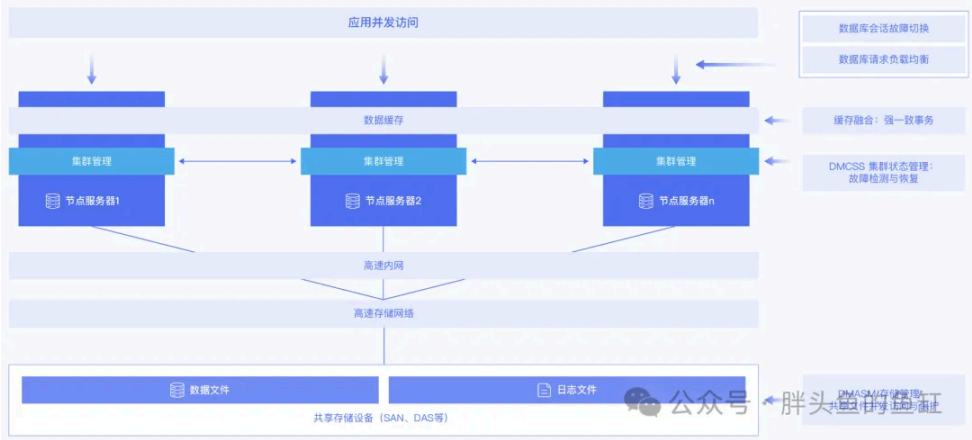




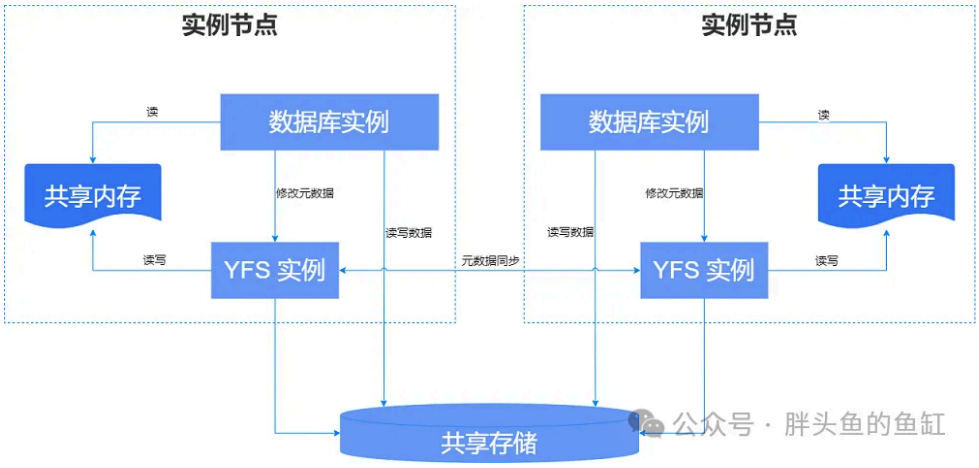
除了冗余模式，也就是简单来说的一份数据有几个副本之外，ASM还会将磁盘组内的磁盘条带化以提升性能，这样可以最大化利用共享存储的性能。这时候Oracle RAC集群的性能上限往往依赖于共享存储的性能，国产数据库中类似的架构还有金仓数据库的KES RAC、达梦数据库的DSC、崖山数据库的YAC、PolarDB的存算分离架构等。



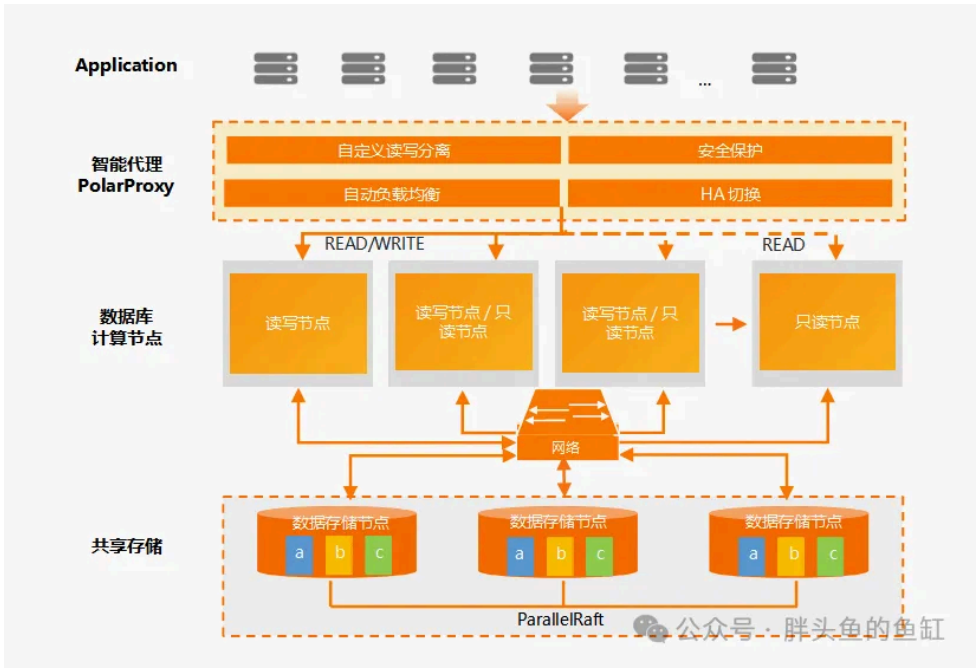
（KES RAC）



（DM DSC）



(YAC)



(PolarDB)

基于共享存储的数据库存算分离架构有诸多好处：

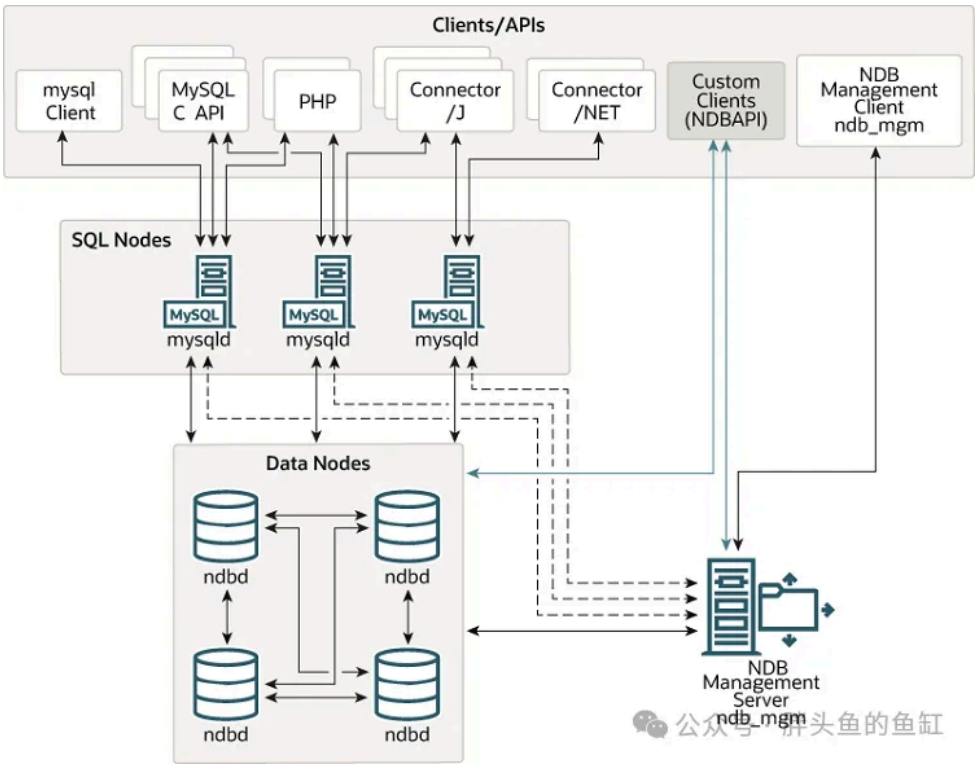
- 除了传统集中存储外，也可以使用分布式存储，可以根据不同的需求和场景选择合适的共享存储
- 节点对数据的获取不一定通过节点间网络进行交互，可以使用更加　定且快速的专用存储网络实现
- 节点异常后整个集群重配置时间较短且故障判断方式除网络外还可以依赖存储链路，节点恢复也不需要占用网络节点同步；同样，集群节点的增减也几乎是无感知的

当然这还是有缺点，比如数据库的安全　定运行依赖共享存储，比如如共享存储出现故障，如存储宕机、链路闪断、静默错误等，将对数据库运行带来灾难性影响，这也要求无论选择何种类型的存储都要慎之又慎，尽可能选择优秀的共享存储，同时在我看来，因为数据库是IT基础架构，该花的钱还是得花。

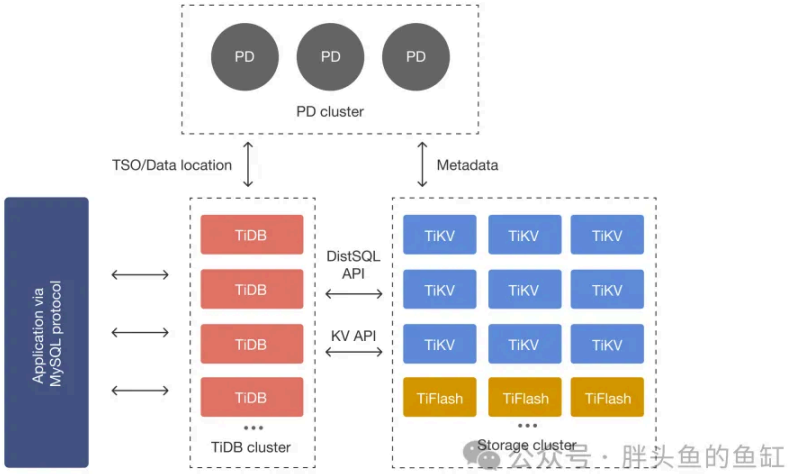
2 基于存储引擎

很多数据库是有存储引擎的概念的，不同的存储引擎是有不同特性的。以MySQL为例，有主流的MyIASM和InnoDB存储引擎，还有一个比较"小众"的存储引擎——NDBCluster。NDB Cluster本身是一个存算分离架构，mysqld提供数据库访问和计算能力，ndb_mgm提供集群管理及元数据

存储，ndbd提供分布式存储能力：



类似的国产数据库架构还有TiDB，TiDB Server提供计算能力，PD Server存储元数据，TiKV/TiFlash（Storage）提供分布式存储能力：



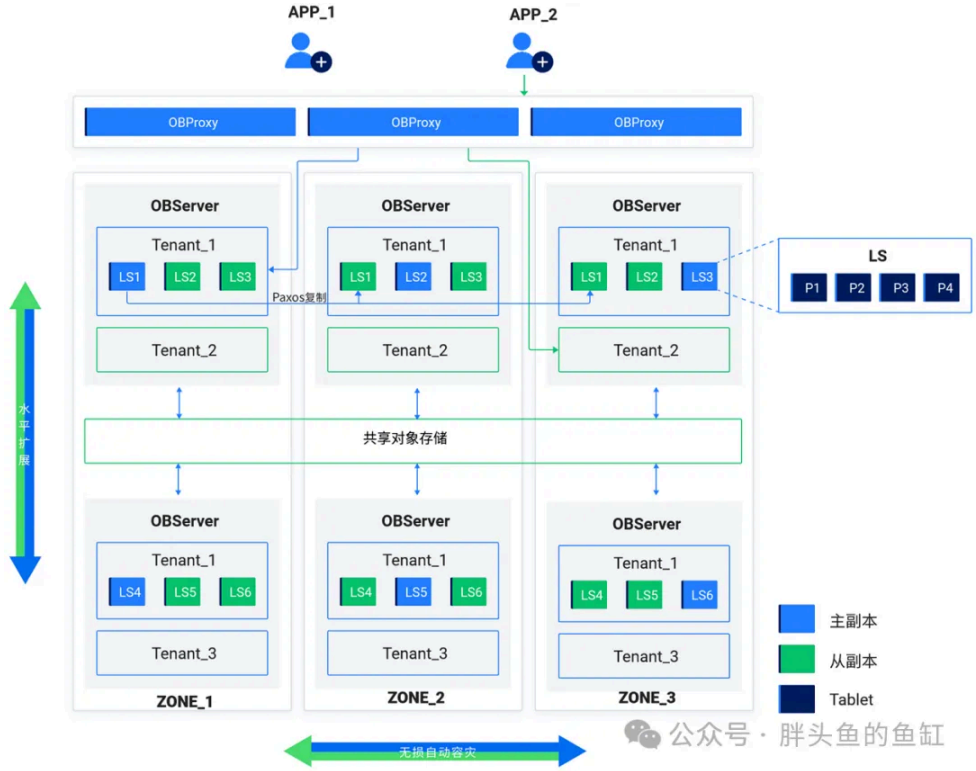
相较于存算一体机的分布式架构，这种架构在设计上是不需要业务方去考　数据在各节点的分布情况。

这类的存算分离架构直接使用服务器本机资源，在全局层面节点本身硬件资源没有共享，在网络足够健壮的情况下理论上可以做到很大规模的横向扩展。但是这也带来了一些挑战：

- 数据库集群的健壮性与数据安全依赖于数据库自身的高可用特性
- 数据库节点间交互严重依赖网络，而网络是不　定的，基于CAP原理需要牺牲一些东西
- 因为不共享的特性，任何节点异常都需要通过网络进行故障判断和集群相关管理操作，恢复后也需要占用网络进行同步

3 混合架构

从4.3.4版本开始，OceanBase数据库在支持无共享（Shared-Nothing，SN）模式的同时，也开始支持共享存储（Shared-Storage，SS）模式部署，这是一个基于通用的对象存储实现了共享存储的存算分离架构，利用云原生基础架构，可以在一定程度上降低数据库使用成本。



每个租户在共享对象存储上存储一份数据和日志，每个租户在节点的本地存储上缓存热点数据和日志。每个主副本负责将一份全量基线数据上传到对象存储，副本间共享对象存储上的基线数据，所有副本自动识别数据热度，仅在本地缓存热点数据。租户的每个副本独立转储，转储数据不在副本间共享。

相较于前面介绍的基于共享存储的存算分离架构，OceanBase的SS模式限定了使用对象存储，本机也会存储日志和部分数据，数据库架构相对复　。个人认为大多数对象存储性能一般不强，数据库的性能发挥还是依赖于本机内存和热数据缓存。当然这也给不同的需求场景提供了更多的选择。

4 优点

相较于一般单纯依赖本机硬件资源的存算一体的数据库架构（集中/分布式），存算分离架构的出现解决了以下一些问题：

- 单机CPU和内存性能提升明显，但是可承载存储磁盘（尤其是高性能磁盘）容量提升有限，通过各种类型的存算分离架构可以更加便捷扩展存储容量与性能
- 在分布式架构下，存算分离可以降低甚至是消除在数据逻辑层面带来的数据拆分要求（即不需要考　数据分片）
- 计算与存储解耦，计算节点故障不影响数据库访问使用，数据存储依托自身高可用可以确保其安全性

而在使用专用存储设备的情况下还能带来以下一些优势：

- 更加　定高效的存储链路
- 磁盘高可用可完全托管
- 更加　定高效的IO性能表现
- 更加全面的磁盘性能及隐患监控，实时解决性能衰退，提前发现并处理故障
- 对于高性能磁盘，尤其是NVMe SSD，现在几乎没有在单机RAID的大规模应用案例，而专用存储早已解决这一问题

- 专用存储可以用于多套数据库集群，在大量节点的数据库集群中，从全局来看对磁盘其实是节省的，并且可以简化本机的磁盘管理

当然专用存储设备的成本也不便宜，但也更适合企业级应用场景。

5 展望

在之前的文章中也聊过那种为了可用性和　定性的异形的分布式数据库架构，即每个节点利用本机的CPU和内存，不使用本机磁盘而是使用共享存储的映射磁盘。这是一种逻辑架构存算一体但物理架构存算分离的架构，在这种情况下如果存储可以和数据库联动，通过存储直接对分布式数据库进行全局一致性快照或者备份，可以极大的提升分布式数据库场景下的备份恢复效能。
无论是专用存储（集中/分布式），还是各类数据库自带的基于存储引擎的分布式存储集群，往往都带有足够多的CPU和内存，在传统使用方式中，这一部分计算资源是大部分被浪费的，如果能充分利用这部分计算资源，对存储的数据持续的优化存储结构、优化存储分布、自动冷热分层、加入计算提前筛选数据等等，会对数据库的整体性能带来更大的助益，同时简化大量的优化管理维护操作。依托于RDMA等高性能低延　网络的加入，可以进一步提升存储链路带宽并降低延　，同时扩展存储在计算中能做的事情。

总结

本期对各主流数据库或不同场景的存算分离架构进行了总结，总结了存算分离架构的特点与优点，对其进行了展望。
老规矩，知道写了些啥。

胖头鱼的鱼缸

钟意作者

数据库 · 目录

上一篇
坏块修复的小坑

下一篇
KES RAC集群部署手册