

Gdevops

Global DevOps Summit

全球敏捷运维峰会

超大规模数仓集群在大型商业银行的落地实践
——龙跃MPP DB

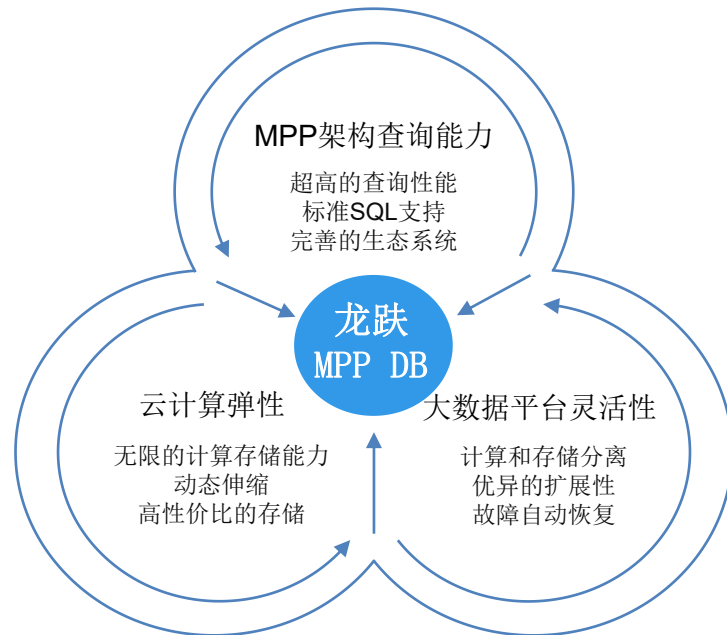
演讲人：建信金科 陈晓新



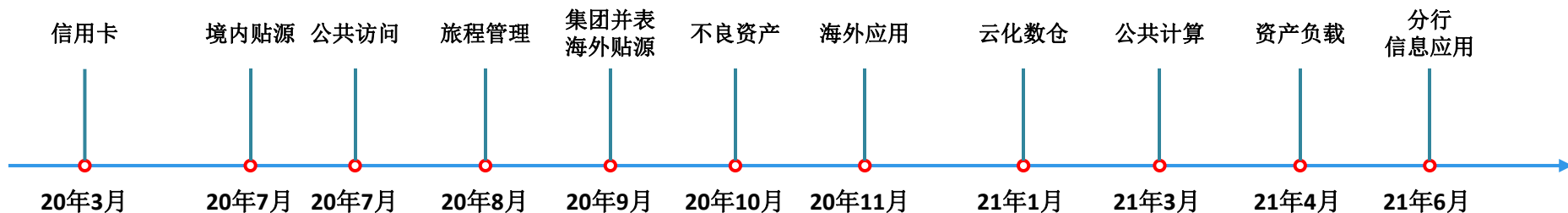
龙跃MPP DB——新一代云原生数据仓库产品



其他数据库、存储等合作公司

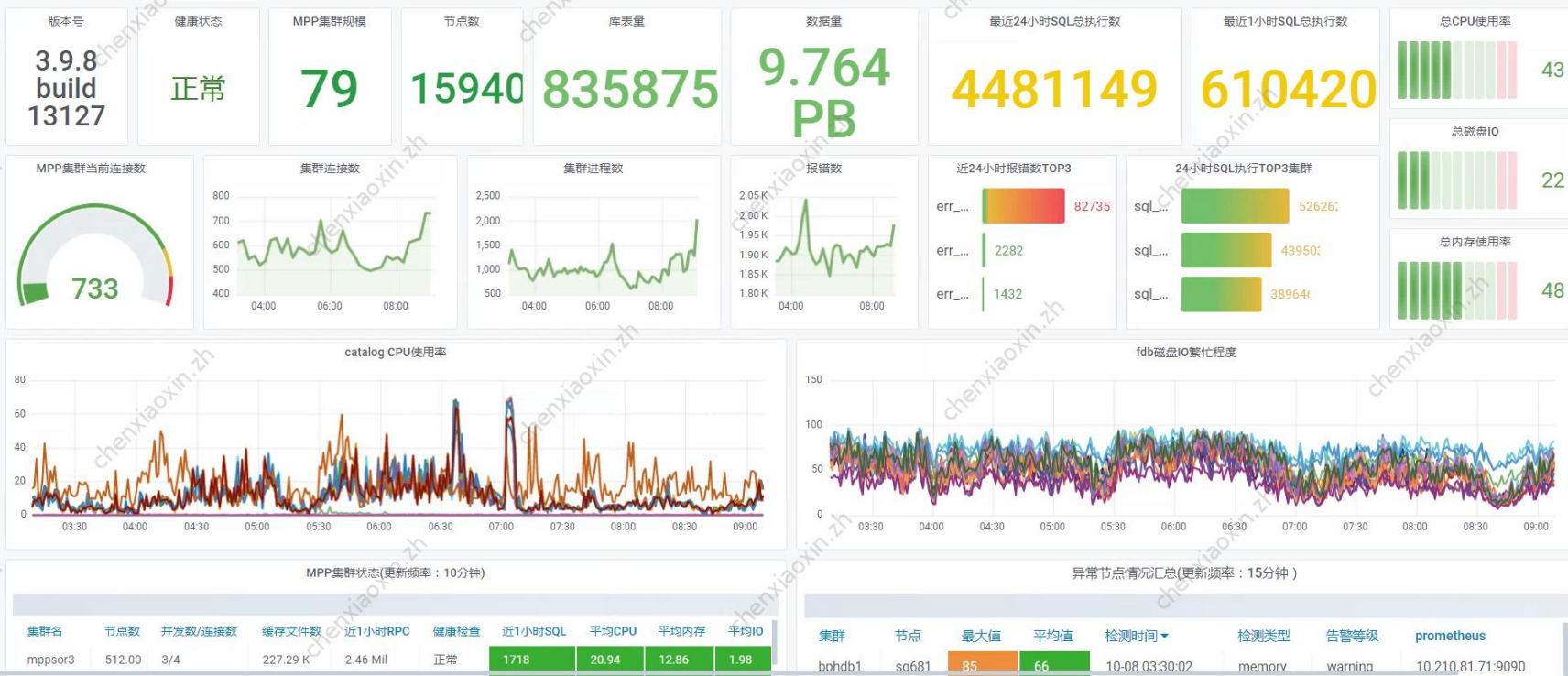


龙跃MPP DB 运行现状	集群规模	数据量	表数量/对象数	负载情况
	15000+服务器	9PB	百万/千万	每天运行作业数达到百万级别，SQL数千万级别

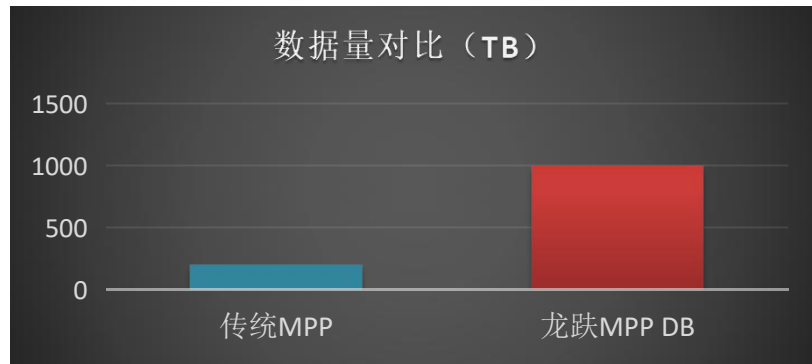
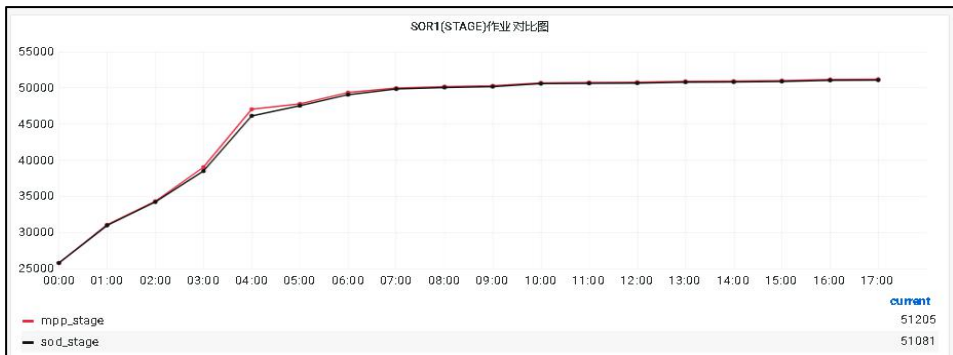
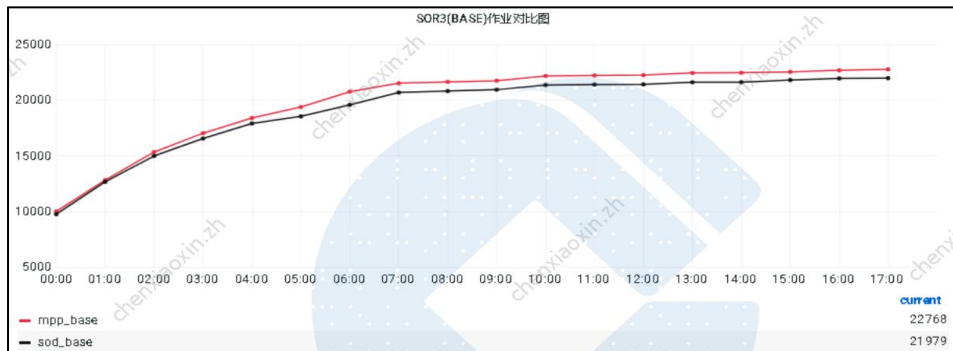


龙跃MPP DB

09:13



贴源集成应用运行效率对比



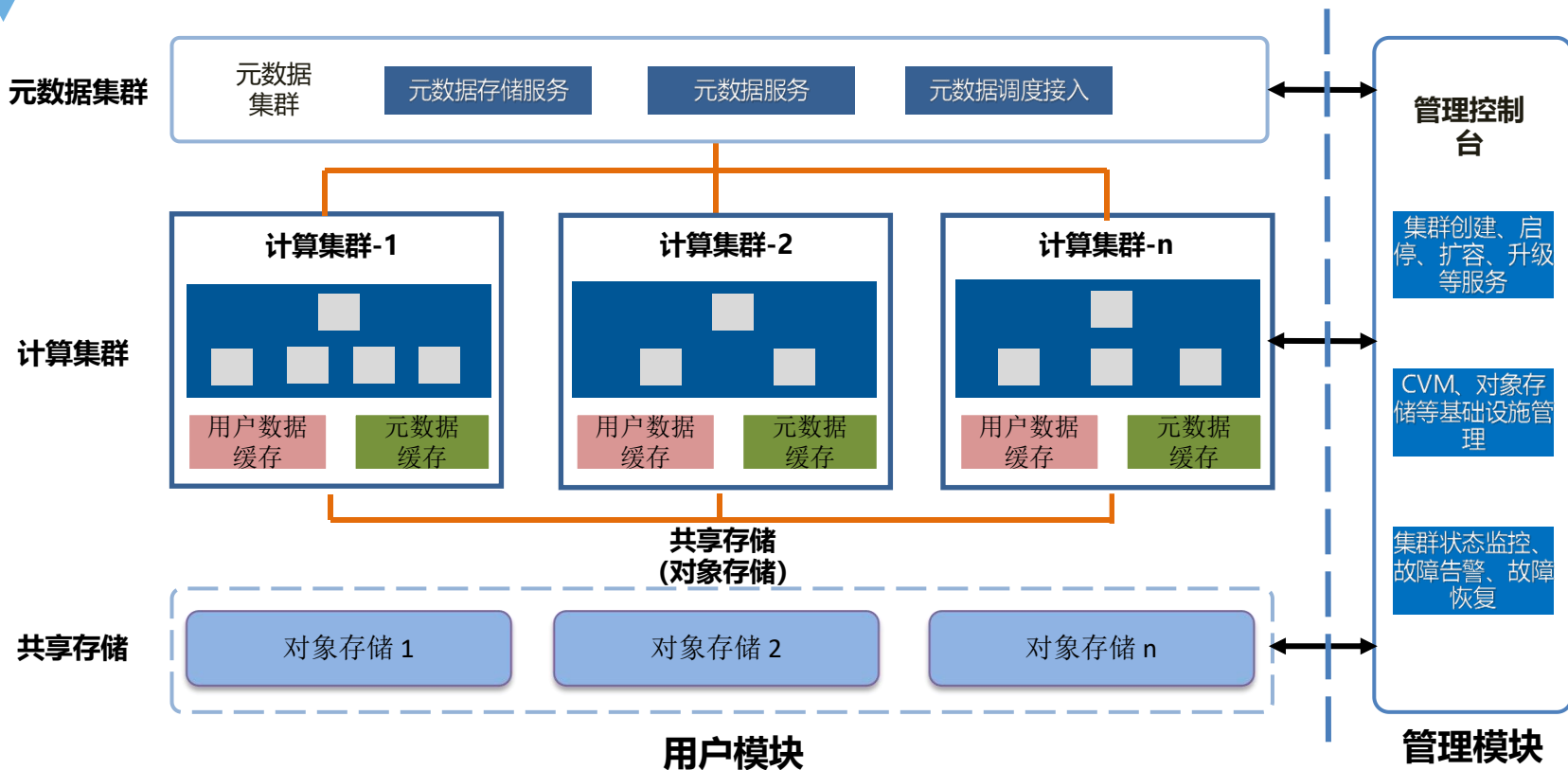
- 龙跃MPP DB的计算资源和传统MPP的计算资源基本相等
- 龙跃MPP DB存储和计算的数据量（1000TB）为传统MPP（200TB）的5倍
- 每天7万个作业、100万个SQL，龙跃MPP DB和传统MPP的运行效率无明显差别

为什么需要研发龙跃MPP DB

- ◆ 并发能力和可扩展性不足，分库分表造成大量数据冗余
- ◆ 数据的存储和计算不分离，数据库孤岛情况严重
- ◆ 升级、扩容等操作复杂，运维成本高，应用影响大
- ◆ 木桶效应，服务器故障会导致集群性能严重下降
- ◆ 非云原生架构，难以融入建行云建设

传统MPP数据库在建行落地实践中遇到的困难

龙跃MPP DB架构



权限管理

- 多租户/用户管理

集群生命周期管理

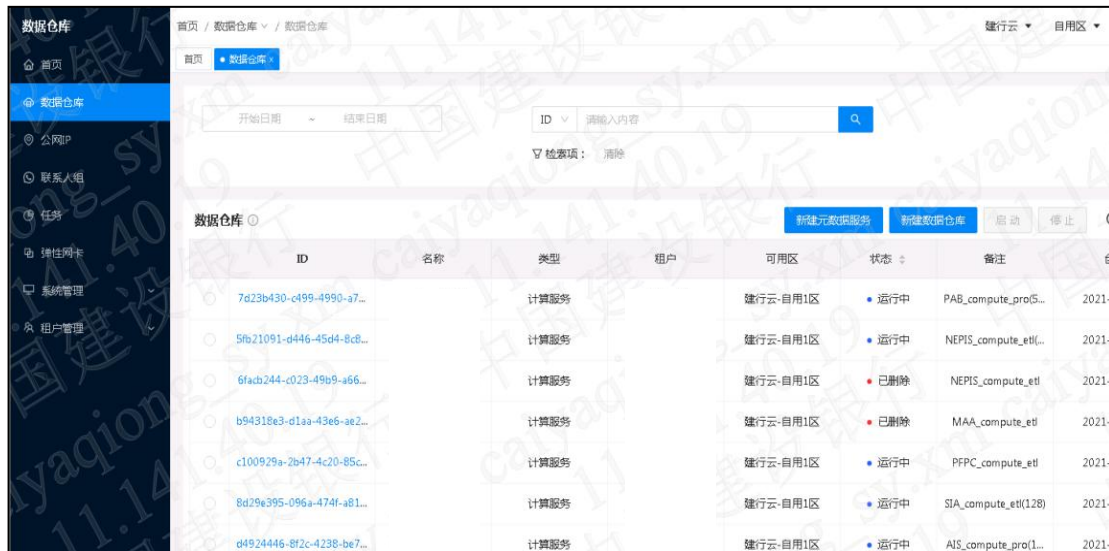
- 创建、删除、扩缩容、升级、启动、停止

IaaS资源交互和调度

- 自动化申请基础设施资源，包括计算、存储和网络资源等

自动化/智能化运维

- 自动化安装部署
- 监控、告警
- 运维
 - 扩容、升级、备份
 - 故障自愈



The screenshot displays the '数据仓库' (Data Warehouse) management interface. On the left is a dark sidebar with navigation links: 首页, 数据仓库, 公网IP, 联系人组, 任务, 弹性网卡, 系统管理, and 租户管理. The main panel shows a table of data warehouses with columns for ID, Name, Type, Tenant, Available Zone, Status, and Remarks. The table lists several instances, some in '运行中' (Running) and others '已删除' (Deleted). At the top right of the main panel, there are buttons for '新建元数据服务' and '新建数据仓库', along with '启动' and '停止' buttons. A search bar and date range filters are also present.

ID	名称	类型	租户	可用区	状态	备注
7d23b430-c499-4990-a7...		计算服务		建行云-自用1区	运行中	PAB_compute_proIS...
5fb21091-d446-45d4-8c8...		计算服务		建行云-自用1区	运行中	NEPIS_compute_eti...
6facb244-c023-49b9-a66...		计算服务		建行云-自用1区	已删除	NEPIS_compute_eti...
b94318e3-d1aa-43e6-a62...		计算服务		建行云-自用1区	已删除	MAA_compute_eti...
c100929a-2b47-4c20-85c...		计算服务		建行云-自用1区	运行中	PFPC_compute_eti...
8d29e395-096a-474f-a81...		计算服务		建行云-自用1区	运行中	SIA_compute_eti(128)
d4924446-8f2c-4238-be7...		计算服务		建行云-自用1区	运行中	AIS_compute_pro1...

调度层

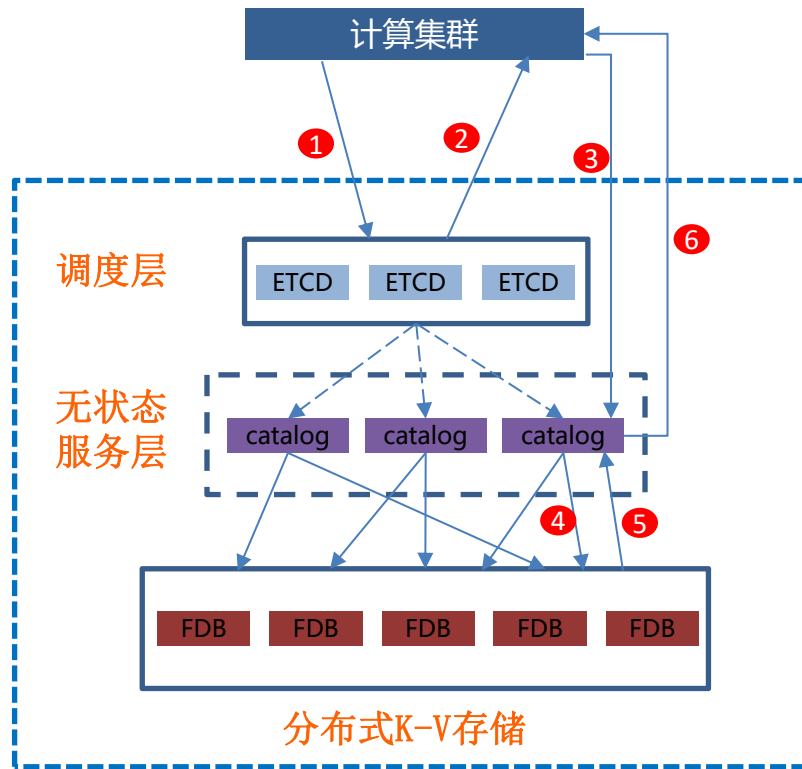
- 服务发现和监控
- 负载均衡

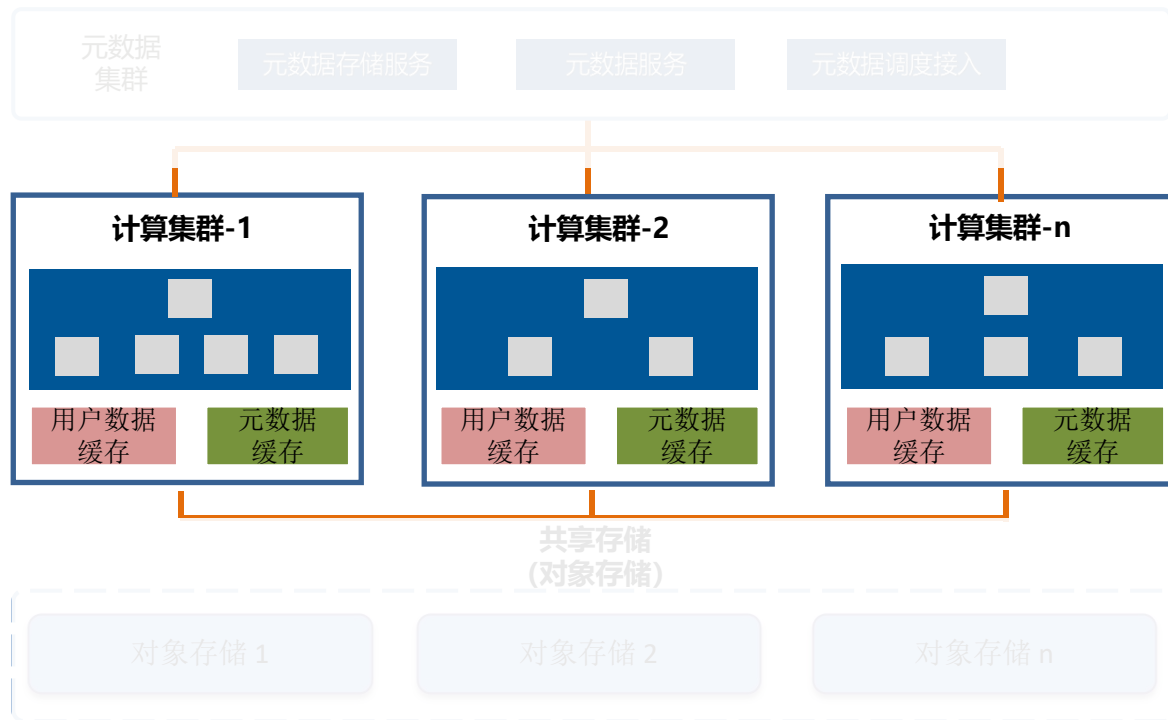
无状态服务层

- 服务层由一组服务节点组成，每个服务节点其实是无状态的服务进程，负责接收和处理计算集群的元数据请求；

元数据持久层

- 元数据持久化存储服务，存储数据字典、统计信息、表到对象映射等



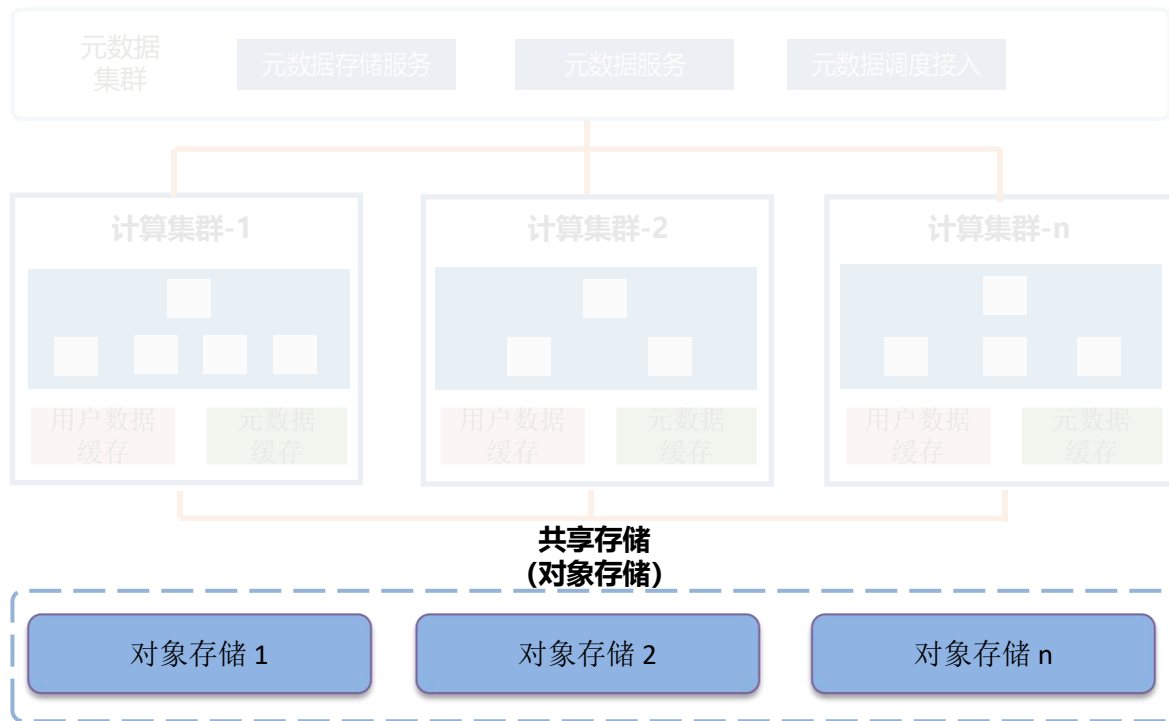


资源灵活分配

- 按需创建、删除、扩缩容
- 集群间资源完全隔离
- 作业可在不同集群建灵活调配
- 并发能力线性扩展

缓存服务

- 本地SSD作为缓存介质
- 小文件合并

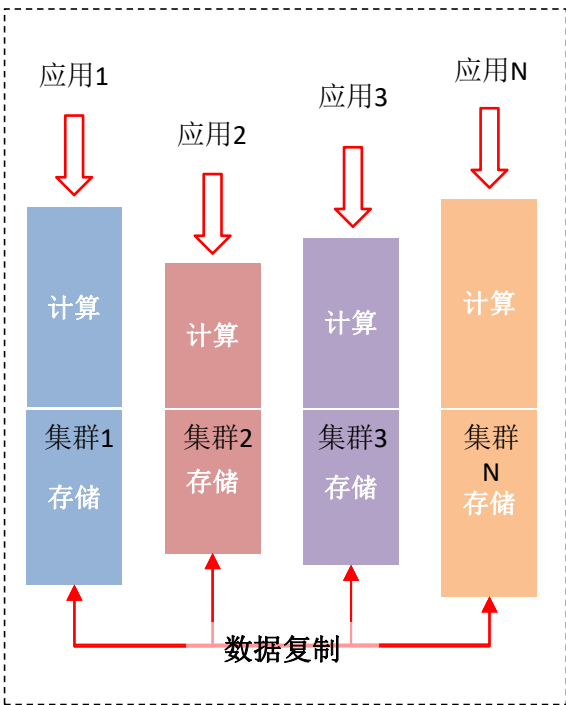


使用对象存储作为数据持久化存储

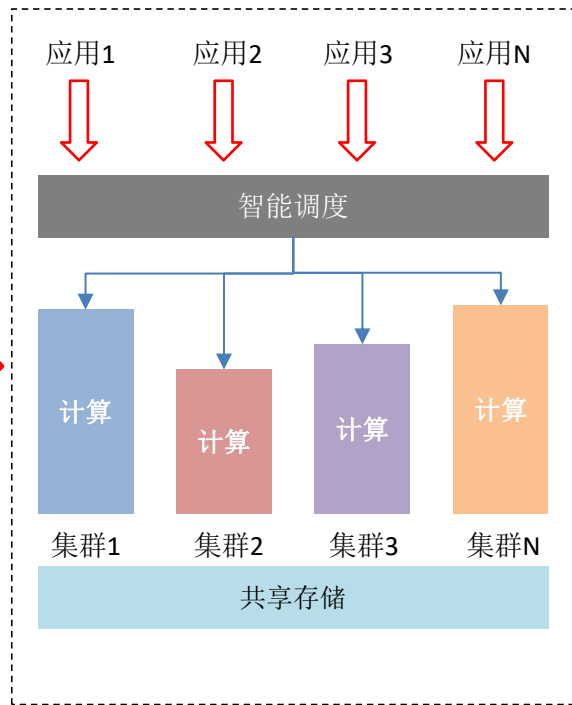
- 支持100亿文件对象，200PB以上的压缩数据
- 使用标准Restful API，支持高并发访问
- 99.99%以上的可用性
- 99.99999999%以上的数据持久性

存储访问优化

- 多桶存储
- 列存+压缩



传统MPP应用解决方案



龙跃MPP DB应用解决方案

	传统MPP	龙跃MPP DB
数据复制	大量集群间数据复制	数据共享，无需
作业动态调度	每个集群运行作业基本固定，无法动态调整	作业可以根据负载需求，在不同集群间动态调整
数据冗余	大量冗余数据	无数据冗余

龙跃MPP DB——运维解决方案

元数据集群

计算集群

Master

Seg1 Seg2 Seg3 Seg4

COS

- ◆ 数据分布：一致性hash的分布方式，避免数据大量重新逻辑分组
- ◆ 独立的元数据共享存储服务，计算节点无状态，随时可增加和减或少，也避免了数据重新物理分布

动态扩容

计算集群

Master

Seg1 Seg2 Seg3 Seg4 Seg5 Seg6 Seg7 Seg8

动态缩容

计算集群

Master

Seg1 Seg2 Seg3 Seg4

快速升级

计算集群

Master

Seg1 Seg2 Seg3 Seg4

新计算集群

Master

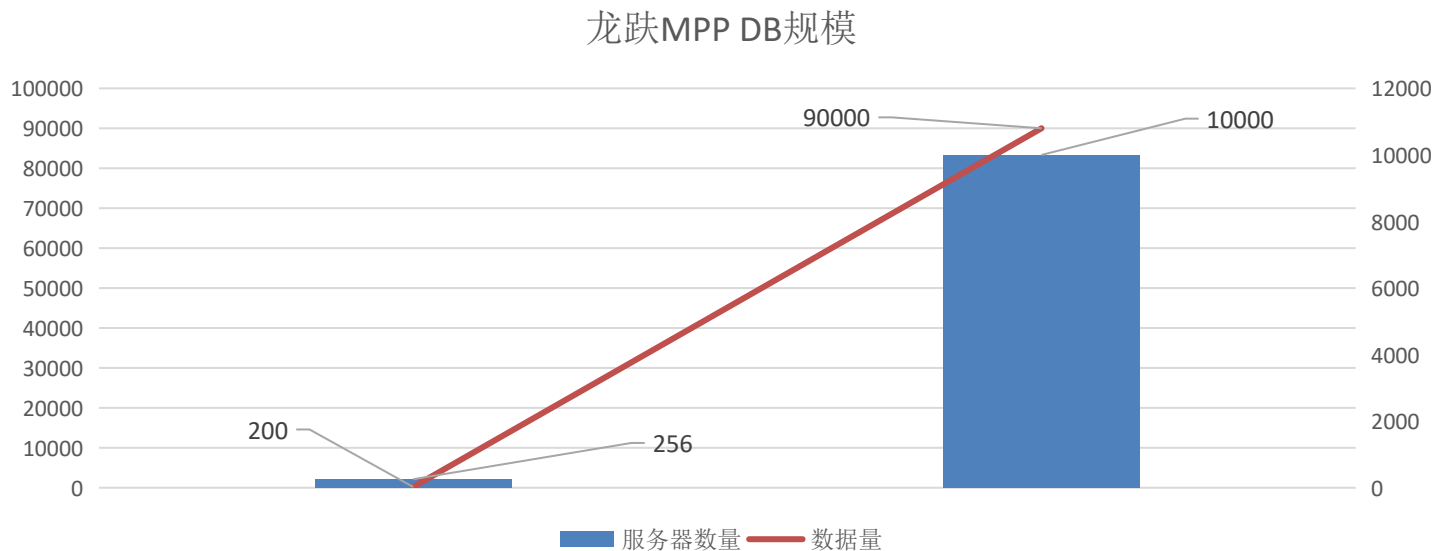
Seg1 Seg2 Seg3 Seg4

故障快速
隔离恢复

计算集群

Master

Seg1 Seg2 Seg3 Seg4 Seg4



过去一年，建行龙跃MPP DB集群的服务器规模增加了50倍，数据量增加了45倍

龙跃MPP DB遇到的问题

◆ 每天百亿级别的元数据RPC请求如何稳定保障

服务拆分、分布式扩展等

◆ 对象存储海量的数据存取需求如何高效满足

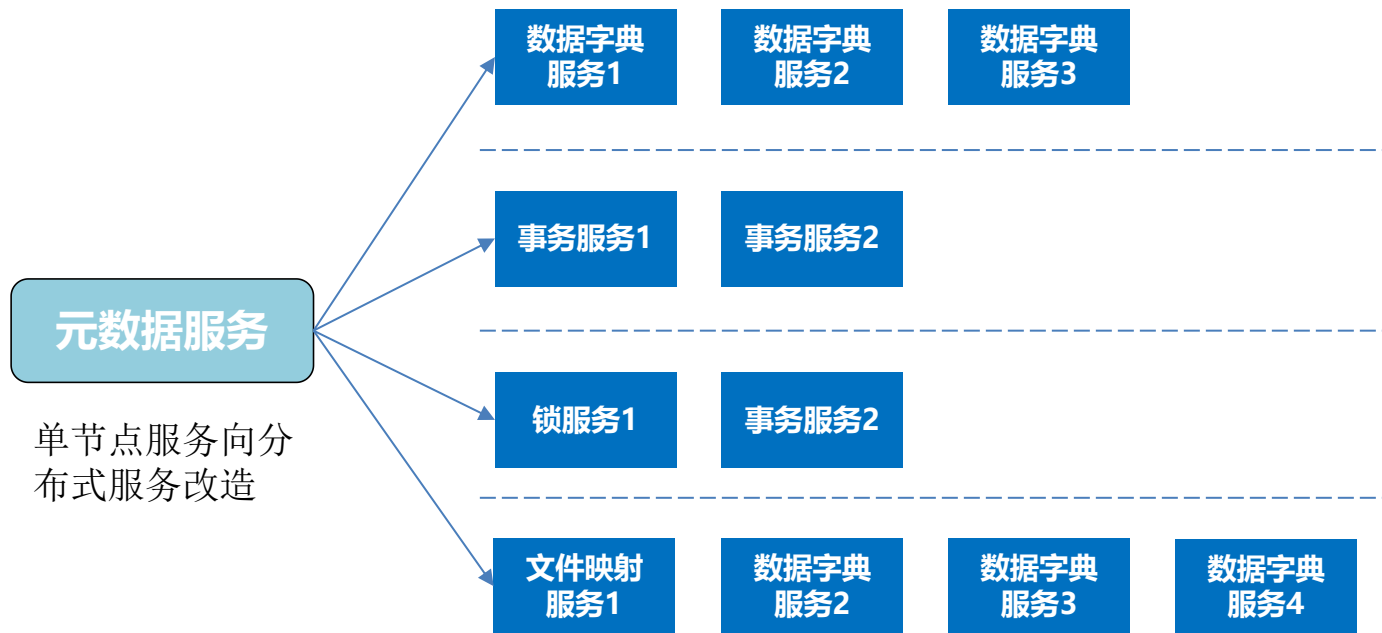
分片、多桶、多线程，共享缓存等

◆ 超大规模的集群如何高效运行维护

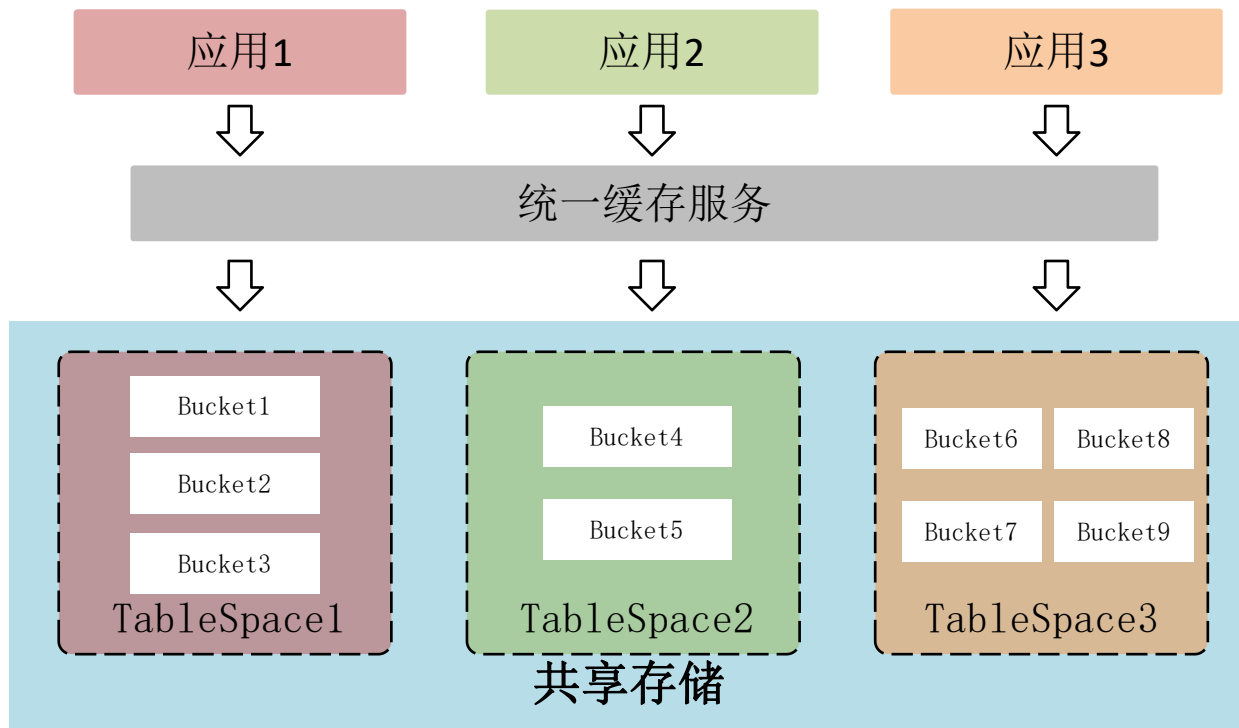
故障自愈、全流程监控、自动化工具

◆ 银行级别的高可用要求如何保障

跨AZ/Region部署、多活、在线备份等



根据服务类型及负载需求，对元数据服务进拆分和分布式改造，提高服务和高可用能力



- 通过统一缓存服务，实现IO加速；减少对象存储压力
- 每个应用创建独立的tablespace，每个tablespace根据需求创建若干个bucket
- 通过tablespace实现共享存储IO能力隔离和流量控制

监控信息获取

RPC分类统计

长SQL

RPC分集群统计

SQL分类统计

CPU

连接数、运行数

IO、IOPS

作业运行数、连接数

内存（虚拟、物理等）

表访问统计

磁盘空间使用

字段访问统计

SQL运行数、报错数等

倾斜统计

进程数、线程数

膨胀统计

作业、SQL、存储全流程数据收集

统计信息和运行状态集成

关键作业完成

异常低/高负载

高连接低负载

异常报错

历史值对比偏离

趋势轨迹

数据聚合分析

服务是否存活

性能是否正常

负载是否倾斜

资源是否充足

智能资源调度

故障辅助定位

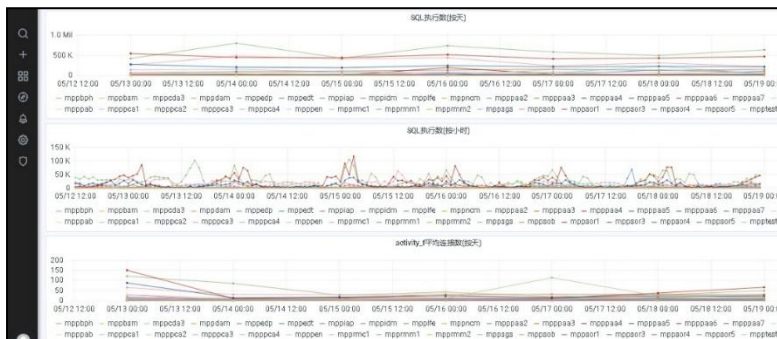
智能运维

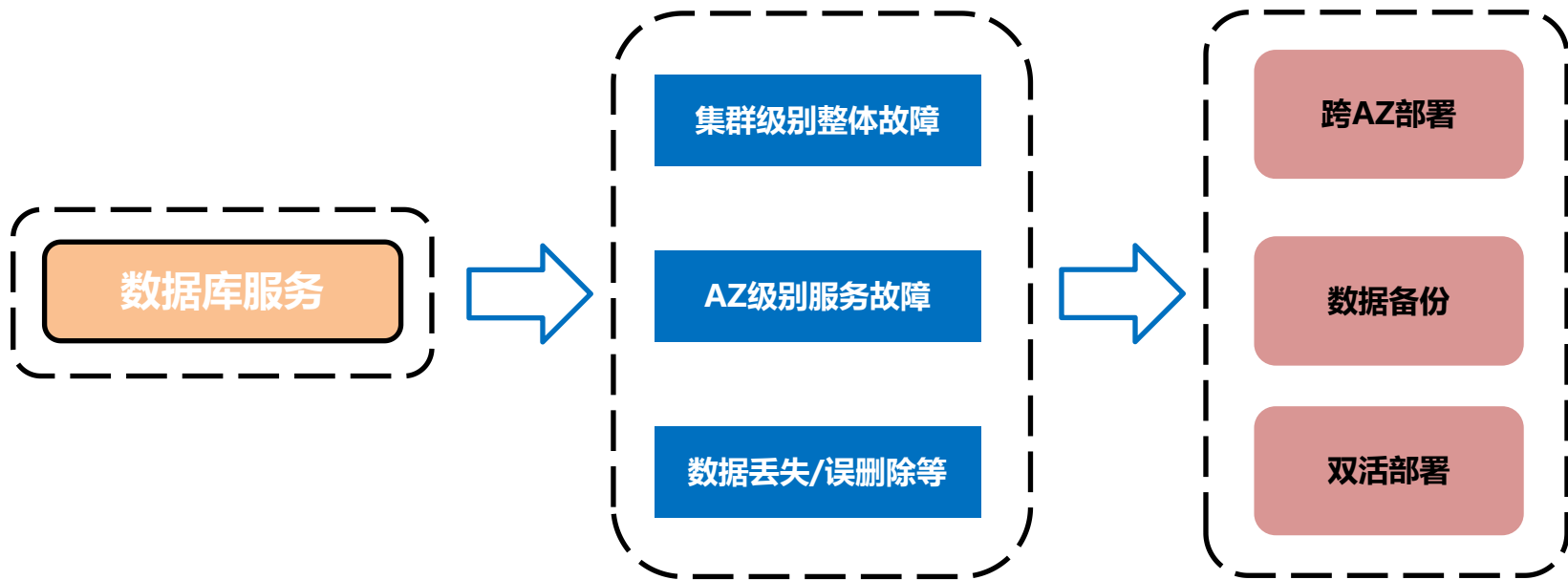
MPP集群状态(更新于: 10分钟前)									
集群ID	节点数	节点数/容量	节点类型	节点状态	节点IP	节点CPU	节点内存	节点磁盘	节点网络
mp001	312/30	4/5	227.20 K	5.55 MB	正常	4447	8.28	13.07	3.97
mp001	312/30	2/3	169.41 K	1.51 MB	正常	4416	2.15	16.60	2.88
mp001	256/30	23/27	85.12 K	5.15 MB	正常	33381	7.24	11.28	2.95
mp001	256/30	13/48	183.91 K	68.63 MB	正常	19336	30.27	22.18	6.10
mp001	128/30	1/2	262.31 K	32.80 K	正常	161	2.19	17.08	3.78
mp001	256/30	17/60	237.18 K	18.31 MB	正常	30609	16.49	16.49	12.32
mp001	312/30	4/7	63.10 K	828.97 K	正常	234	6.36	6.32	6.87
mp001	64/30	1/2	83.15 K	25.23 K	正常	177	0.54	6.33	1.15

集群ID	节点数	节点数/容量	节点类型	节点状态	节点IP	节点CPU	节点内存	节点磁盘	节点网络
mp001	312/30	4/5	227.20 K	5.55 MB	正常	4447	8.28	13.07	3.97
mp001	312/30	2/3	169.41 K	1.51 MB	正常	4416	2.15	16.60	2.88
mp001	256/30	23/27	85.12 K	5.15 MB	正常	33381	7.24	11.28	2.95
mp001	256/30	13/48	183.91 K	68.63 MB	正常	19336	30.27	22.18	6.10
mp001	128/30	1/2	262.31 K	32.80 K	正常	161	2.19	17.08	3.78
mp001	256/30	17/60	237.18 K	18.31 MB	正常	30609	16.49	16.49	12.32
mp001	312/30	4/7	63.10 K	828.97 K	正常	234	6.36	6.32	6.87
mp001	64/30	1/2	83.15 K	25.23 K	正常	177	0.54	6.33	1.15

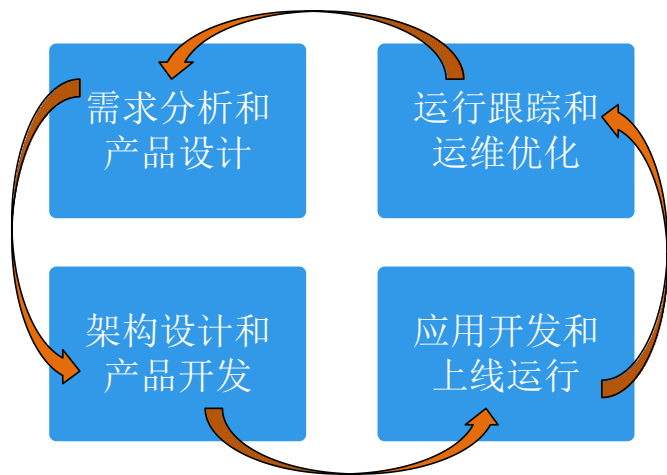
集群ID	节点数	节点数/容量	节点类型	节点状态	节点IP	节点CPU	节点内存	节点磁盘	节点网络
mp001	312/30	4/5	227.20 K	5.55 MB	正常	4447	8.28	13.07	3.97
mp001	312/30	2/3	169.41 K	1.51 MB	正常	4416	2.15	16.60	2.88
mp001	256/30	23/27	85.12 K	5.15 MB	正常	33381	7.24	11.28	2.95
mp001	256/30	13/48	183.91 K	68.63 MB	正常	19336	30.27	22.18	6.10
mp001	128/30	1/2	262.31 K	32.80 K	正常	161	2.19	17.08	3.78
mp001	256/30	17/60	237.18 K	18.31 MB	正常	30609	16.49	16.49	12.32
mp001	312/30	4/7	63.10 K	828.97 K	正常	234	6.36	6.32	6.87
mp001	64/30	1/2	83.15 K	25.23 K	正常	177	0.54	6.33	1.15

集群ID	节点数	节点数/容量	节点类型	节点状态	节点IP	节点CPU	节点内存	节点磁盘	节点网络
mp001	312/30	4/5	227.20 K	5.55 MB	正常	4447	8.28	13.07	3.97
mp001	312/30	2/3	169.41 K	1.51 MB	正常	4416	2.15	16.60	2.88
mp001	256/30	23/27	85.12 K	5.15 MB	正常	33381	7.24	11.28	2.95
mp001	256/30	13/48	183.91 K	68.63 MB	正常	19336	30.27	22.18	6.10
mp001	128/30	1/2	262.31 K	32.80 K	正常	161	2.19	17.08	3.78
mp001	256/30	17/60	237.18 K	18.31 MB	正常	30609	16.49	16.49	12.32
mp001	312/30	4/7	63.10 K	828.97 K	正常	234	6.36	6.32	6.87
mp001	64/30	1/2	83.15 K	25.23 K	正常	177	0.54	6.33	1.15



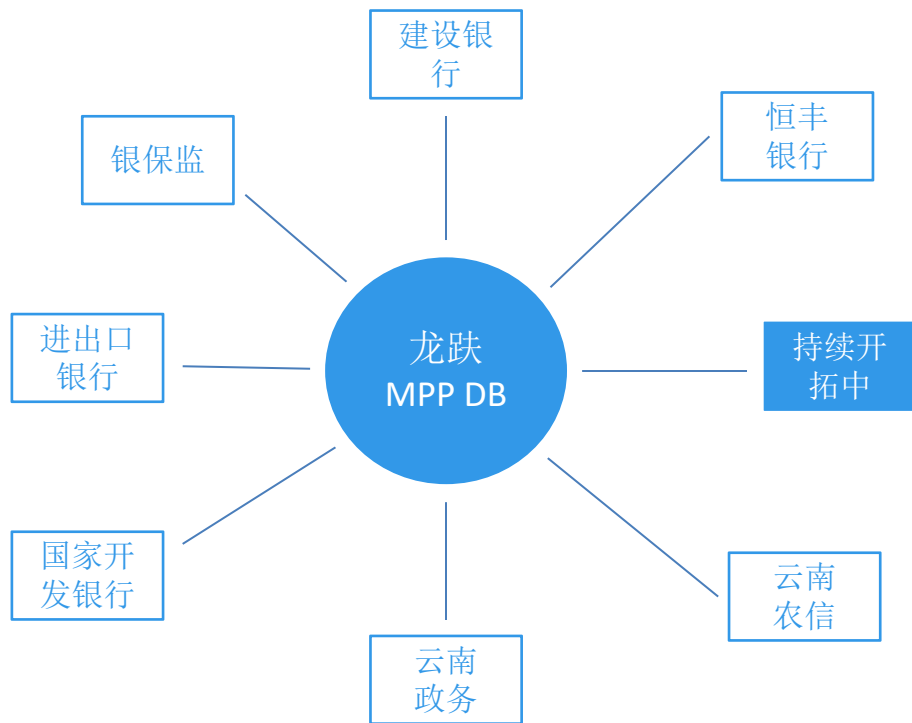


通过跨AZ部署、备份、双活等方式，进一步解决集群故障、AZ故障、数据丢失等问题



过去几年，我们完成了无数次的版本迭代和上线优化。一款数据库产品的成熟发展，需要产品、架构、研发、运维、应用等许许多多人的长期合作和投入。在龙跃MPP DB上，我们：

- 集合了大批建信金科和业界优秀的研发人员；
- 提供了业界最复杂、最丰富、负载最高的应用场景；
- 拥有建行二十几年的数据仓库使用和运维经验，能够最快的发现产品痛点，提出最贴合用户需求的产品设计。



坚持产品研发投入、持续拓展用户、丰富产品生态，打造更为先进、安全的数据仓库产品！！



全球敏捷运维峰会

THANK YOU !

