DTCC

第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

架构革新 高效可控

IT168.com ChinaUnix.net ITPUB

北京国际会议中心 | 2020/12/21-12/23

工行分布式数据库应用实践

林承军 | 中国工商银行软件开发中心 高级经理

工行分布式数据库建设背景

分布式架构下数据库转型核心诉求

快速研发

- 保持对通用数据库最大兼容性
- 产品有完备的方案及配套工具，支持应用快速研发

运维能力

- 具备数据库的运维自动化/智能化能力
- 与行内系统对接和集成，满足定制化需求

支撑能力

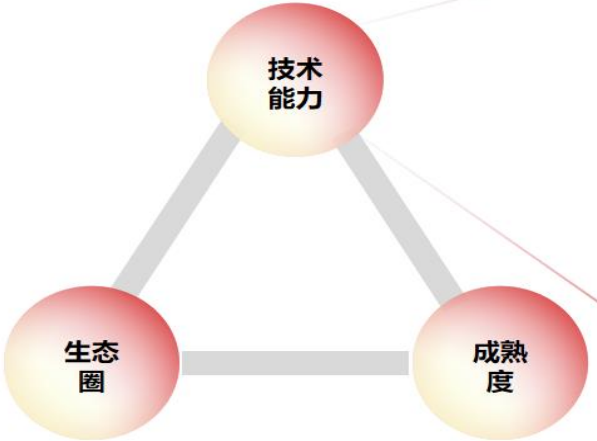
- 高并发、可扩展，海量数据存储的处理
- 满足两地三中心高可用容灾要求

成本控制

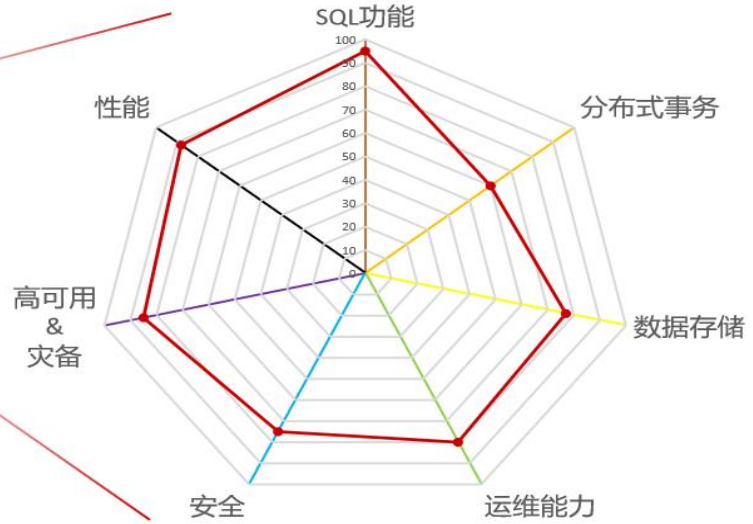
- 通过下移主机业务至平台，使用更廉价的硬件基础设施
- 自主可控,解决对商业产品的过度依赖

分布式数据库技术路线分析和选择

- **工行OLTP分布式数据库建设分两步走的策略：**
  - 近期策略：建设分布式数据库访问层+MySQL的解决方案，支撑行内数据库转型实施。
  - 中远期策略：持续跟踪和研究国内分布式事务数据库产品技术演进和发展，大胆探索谨慎推广。



选择评估



技术能力评估

工行分布式数据库建设历程

- 2016年-至今，逐步建立与云计算融合、国产化适配好的分布式MySQL数据库的应用级、系统级两种解决方案。
- 2019年，结合国内政策导向和行内数据库转型总体规划，加快国产分布式数据库探索实践。

2016
分布式MySQL数据库原型研究

- ✓ 基于个人结算账户开展应用级分布式MySQL数据库原型研究
- ✓ 基于客户信息业务开展系统级分布式MySQL数据库原型研究

2017
分布式MySQL数据库试点

- ✓ 完成个人结算应用在应用级分布式MySQL数据库试点上线
- ✓ 完成客户信息在系统级分布式MySQL数据库试点上线

2018
分布式MvSQL数据库推广

- ✓ 丰富和完善分布式MySQL数据库自动化运维能力，并全面推广实施

2019
- **分布式MySQL数据库云化建设**
- **GaussDB联合创新试点**

- ✓ **分布式MySQL数据库**：开展数据库云化建设，支持云化部署和一键式环境供给，有效提升资源使用率效率
- ✓ **GaussDB**：中间业务 、 生物特征等GaussDB 数据 库应用试点

2020
- **发布分布式MySQL数据库iDB**
- **OceanBase技术探索**
- **GaussDB产品完善及推广**

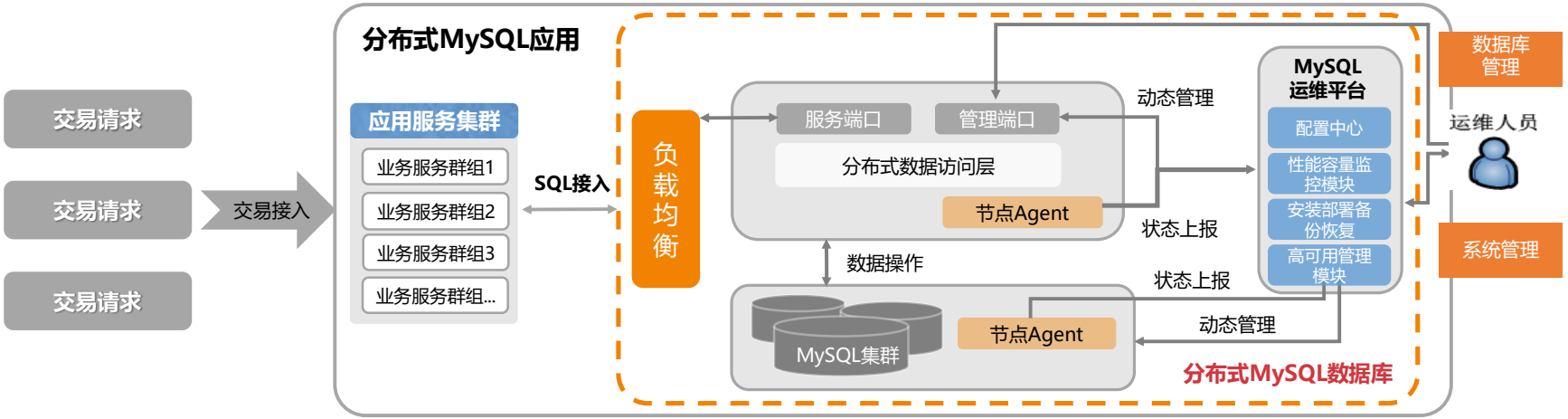
- ✓ **分布式MySQL数据库iDB**：**国产化适配及改造，打造开放、成熟、安全可控数据库软件**
- ✓ **OceanBase**：对公理财应用技术探索
- ✓ **GaussDB**:声誉风险系统 、 商密公文系统 、贵金属交易系统；网讯系统、办公门户系统等5个应用试点

分布式MySQL数据库建设

国产分布式数据探索实践

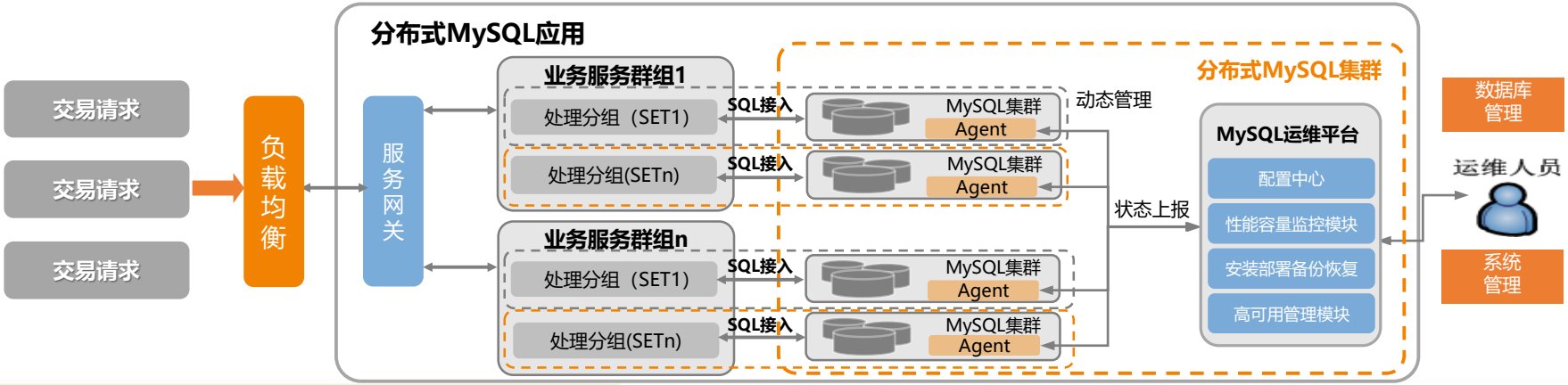
分布式MySQL数据库iDB-系统级

- **系统级方案**：基于分布式数据访问层+开源MySQL+运维平台，构建系统级联机交易分布式数据库。
- **技术特点**：
  - 分布式访问层：通过分布式访问层实现SQL的解析和路由，支持对应用透明的智能路由访问，实现整体集群的横向扩展的能力，降低应用研发复杂度。
  - MySQL数据库集群：采用开源MySQL和原生态数据复制技术，一主多备架构，实现多份数据冗余一致性保障。
  - MySQL运维平台：自主研发，实现对MySQL数据库的安装配置、监控告警、性能容量、健康检查、高可用、节点扩展、数据备份/灾备等全生命周期的管理，提升工行运维自动化水平。



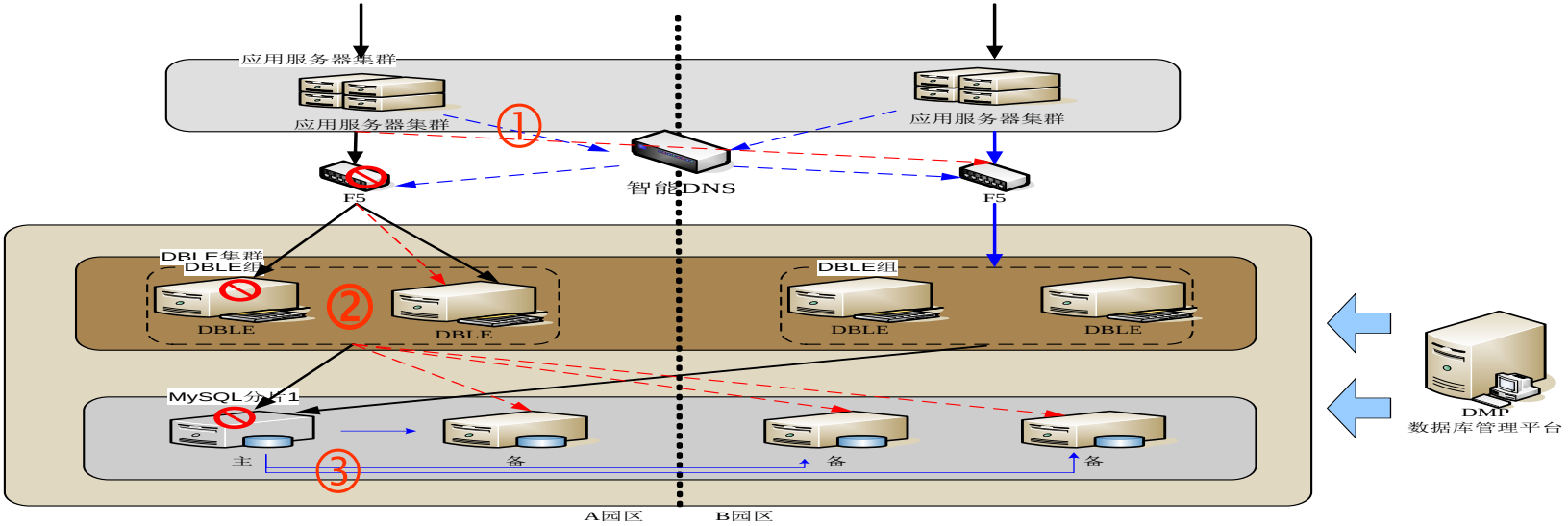
分布式MySQL数据库iDB-应用级

- **应用级方案:** 服务网关（分布式服务）+处理分组（SET）+运维管理平台实现，处理分组（SET）包含完整联机处理功能的服务器集合。
- **技术特点**:
  - 交易路由：服务网关根据接入交易请求的特定字段（如用户或者协议字段）进行路由，分发到不同的处理分组（SET）
  - 数据分片：按明确的业务特征字段进行数据分片，每份一套MySQL数据库集群（一主多备多份数据冗余）
  - 处理分组（SET）：**是针对一个特定业务字段（例如用户或者协议）分片的完整服务**，每套处理分组包含完整联机处理功能的服务器集合，含若干应用服务器、数据库服务器等
  - 数据库相关运维管理和高可用等能力与系统级解决方案一致。



分布式MySQL数据库特性iDB-高可用故障切换

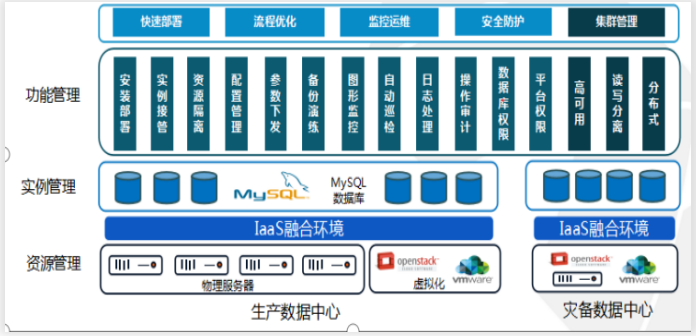
- 支持两地三中心高可用架构部署，实现本地、同城节点故障自动切换，RTO<60s，RPO=0，提供业务连续性、数据一致性保障；
- 与行内多个系统联动，实现应对资源域级故障自动切换、园区级故障一键式并行切换能力；
- 结合各类应用场景，实现了RPO优先、RTO优先、园区优先等多种高可用策略，满足各类应用场景差异化需求。



分布式MySQL数据库特性-专业运维管理

全面的数据库监控

- **对数据库可用性、可靠性、性能等进行监控，涵盖100多项监控项指标**,异常状态可实时通过邮件、微信、短信等形式告警。

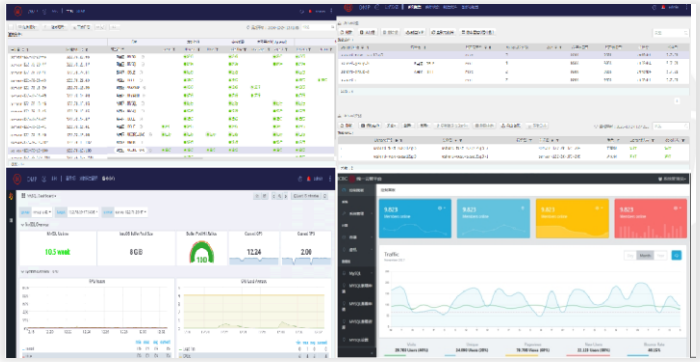


自动化安装部署管理

- 自动化安装部署管理技术，实现对分布式MySQL数据库组件的安装配置、高可用、数据备份/灾备、升级维护等DBA运维管理功能，功能全面、覆盖面广。

智能的高可用保障手段

- 集成了数据库高可用切换的功能实现，支持数据优先、同城优先等多种高可用切换策略；融合智能告警和高可用切换动态管理，**提供切换异常智能化告警和高可用保障建议能力**。

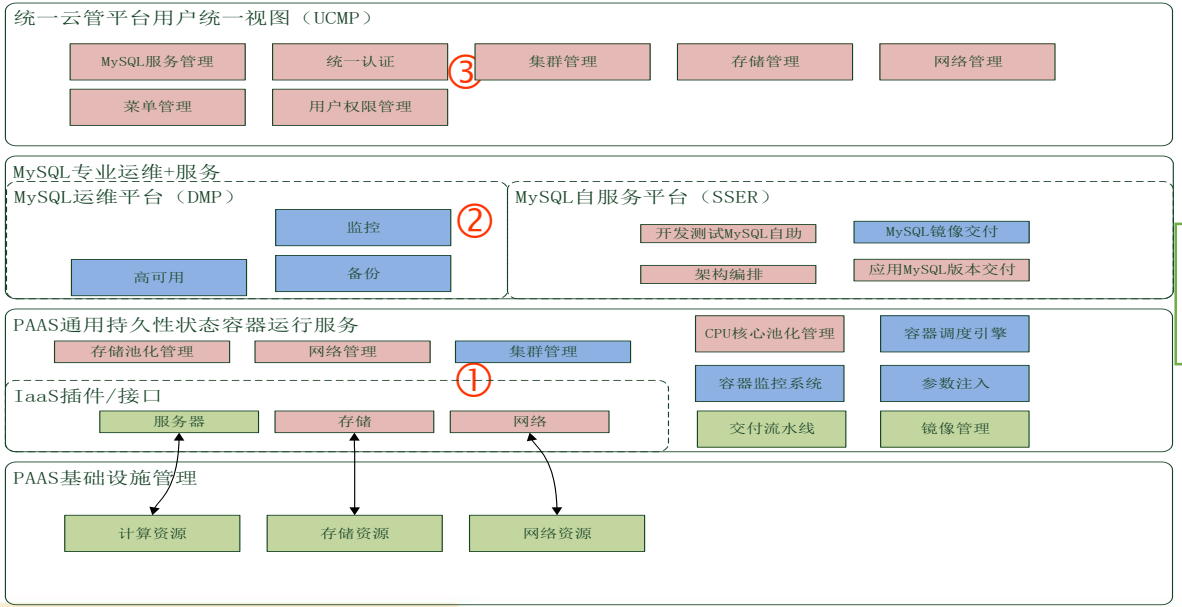


完备的性能容量管理

- 支持数据库性能指标的多维度采集和加工，为性能容量分析和管理、故障定位提供有效的技术手段和支撑。

分布式MySQL数据库特性iDB-数据库云化

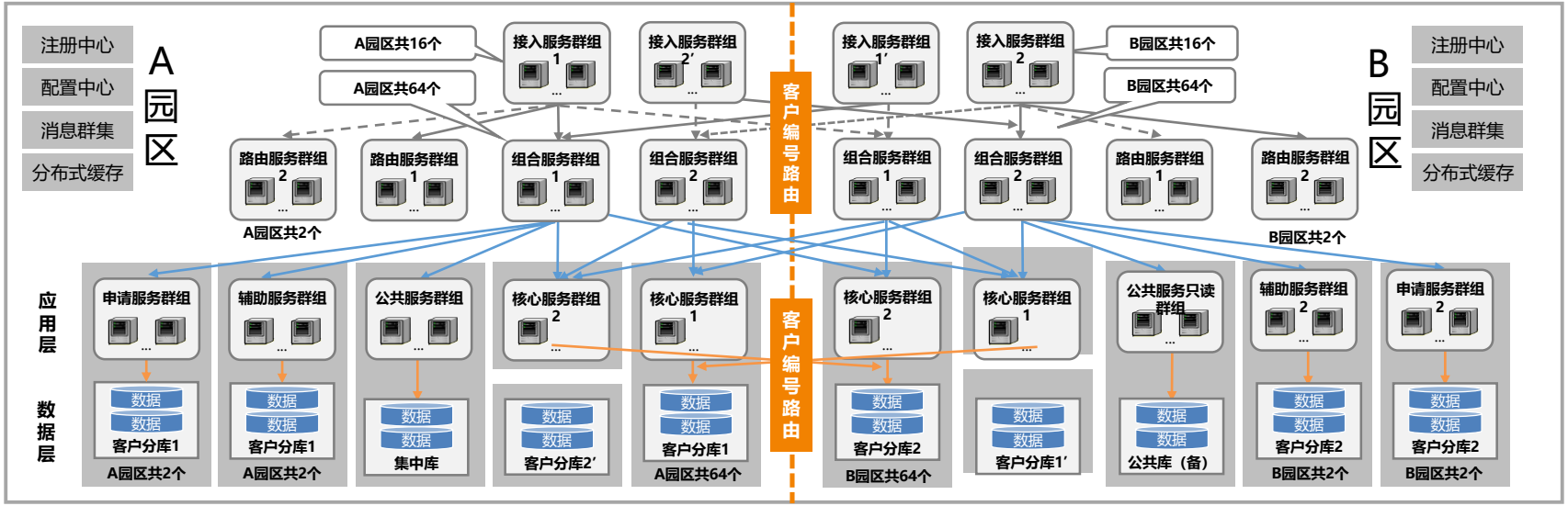
- PAAS层实现对MySQL容器运行的持久化支持，支持MySQL的云化部署，服务器资源使用效率大幅提升；
- 建设了MySQL自助服务平台，支持MySQL云化环境的一键式、批量供给，加快资源部署供给的效率；
- 建立统一MySQL云服务视图，优化管理流程，提升内部管理效率。



- 提升资源使用效率
- 加快资源部署速度
- 提升内部管理效率

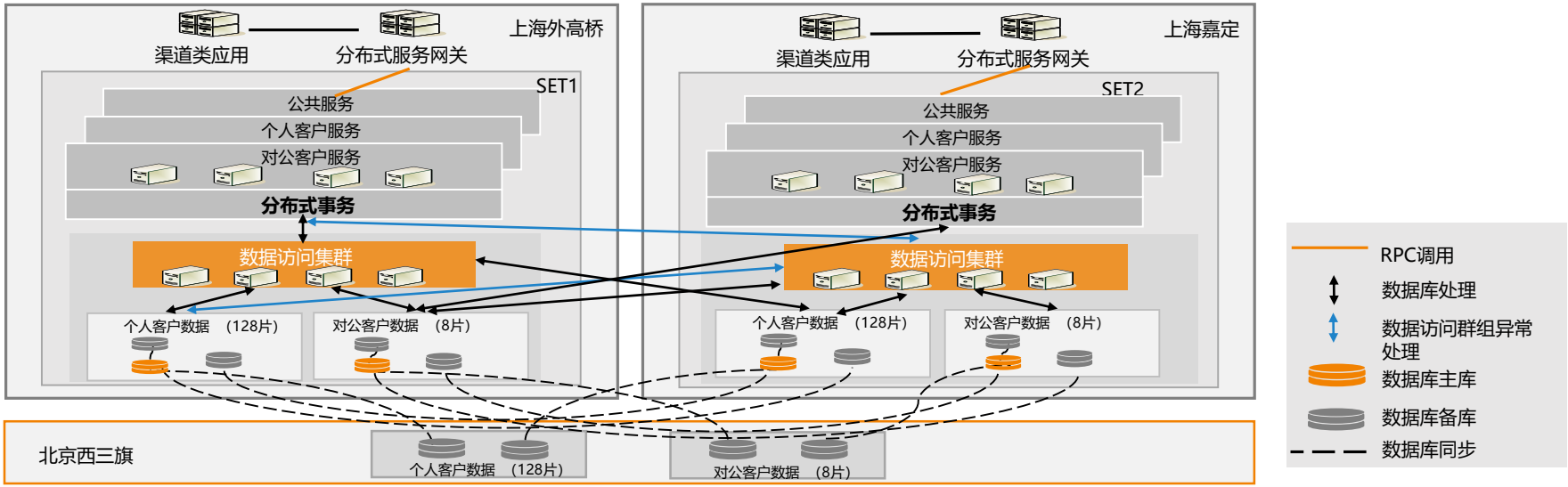
分布式MySQL数据库iDB典型应用场景 - 个人结算

- **场景概述：**为我行个人客户提供资金结算类相关服务，包括个人结算账户开立、变更、撤销及资金结算处理。管理10亿个人结算账户及介质，每日提供2.5亿笔结算服务，1亿笔查询管理类服务。
- **技术方案：**应用级分布式MySQL方案，通过在应用实施服务化改造，结合分布式事务框架、柔性事务机制、分布式服务等分布式技术，以客户、协议等信息进行数据分片的处理分组（SET）设计，降低了应用系统对数据库事务处理机制的依赖。
- **实施效果：**提供日均3.5亿笔服务，最高并发数25000TPS，平均交易耗时小于50毫秒。



分布式MySQL数据库iDB典型应用场景 - 客户信息

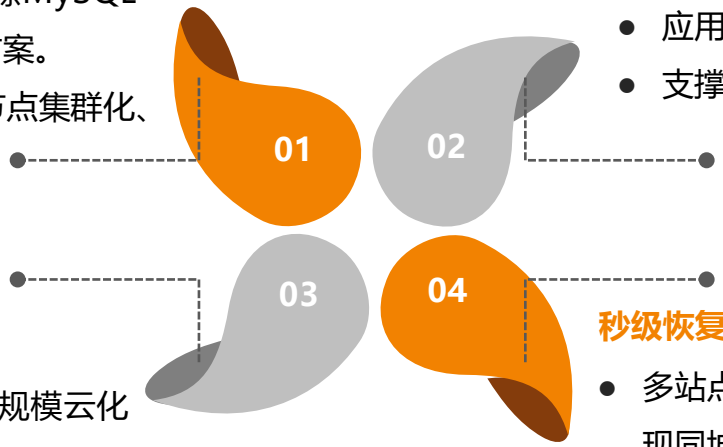
- **场景概述：**为全行业务系统提供客户信息维护与查询服务，管理6亿个人客户和1千多万对公客户，总记录数超过160亿，每日提供2亿次客户信息维护与查询，需要满足超大数据量的高频访问需求。
- **技术方案：**系统级分布式MySQL数据库方案，采用分库分表的设计思路，通过访问层实现SQL的智能路由，降低应用设计复杂度。
- **实施效果：**为全行180多个总分行应用提供日均超2亿次维护与查询服务，最高并发数7600TPS，平均交易耗时小于30ms，支撑应用范围同业最广、日均访问数量同业最多。



MySQL分布式数据库iDB - 核心优势

高度自主可控能力

- 基于开源分布式数据访问层和开源MySQL数据库，构建分布式数据库解决方案。
- 自研运维管理平台，实现大规模节点集群化、自动化、标准化管理。



广泛运用于各业务场景

- 应用于企业客户信息等重点业务场景。
- 支撑双十一、春节业务高峰万级TPS。

领先的数据云化服务

- 在同业率先实现MySQL数据库大规模云化服务，数据库节点数量达到数千个。
- 一键式快速供给，一体化运维管理。

秒级恢复、分钟级切换

- 多站点数据存储，保证数据强一致性，实现同城双活RPO=0。
- 本地故障秒级恢复，无需人工干预，业务系统无感知，园区级故障分钟级同城切换。

国产分布式事务数据库探索实践

分布式MySQL解决iDB问题：

- 分布式MySQL数据库技术路线成熟，能提供对高并发、可扩展应用场景的支撑;
- 通过持续能力建设，能满足分布式系统大规模节点运维自动化要求。

分布式MySQL数据库iDB挑战：

- 该方案与应用耦合相对紧密，研发成本高;
- 无法支持分布式事务，在应用层解决，增加设计复杂度;
- 复杂SQL支持能力弱，无法满足混合负载应用场景，需要定制方案;
- Oracle兼容不够，存量Oracle迁移成本高。

探索通用分布式事务数据库方案

国产分布式事务数据库核心能力诉求：

- 分布式可扩展架构，具备在线扩缩容能力;
- 支持强一致性分布式事务;
- 支持SQL2003，具备较低成本ORACLE系统迁移能力;
- 具体复杂计算能力，满足混合负载场景;
- 国产化适配兼容性好;
- 满足两地三中心高可用容灾，具备同城双活部署架构下RPO=0，RTO分钟级的高可用能力。

国产分布式事务数据库应该挑战：

- 业界分布式数据库产品多，且技术路线差异大，技术路线演进有待明确;
- 分布式数据库金融应用尚处于发展阶段，产品还不成熟，有待完善和实践打磨。

分布式数据库GaussDB探索历程

2018

启动OLTP分布式数据库研究

与华为就OLTP分布式数据库开展技术交流和创新探索

2019

OLTP分布式数据库业务试点

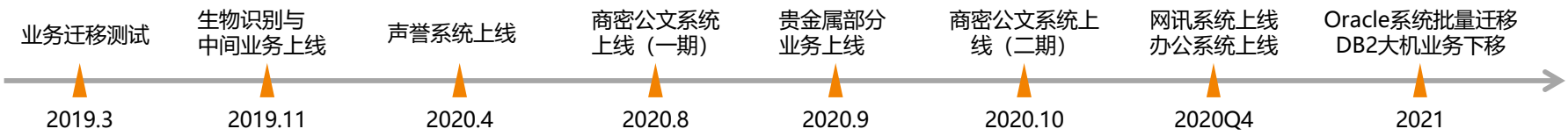
以Oracle存量应用转型和主机下移能力验证为牵引，先后上线生物特征识别和中间业务系统，初步完成产品能力验证

2020

业务实践进一步夯实和拓展

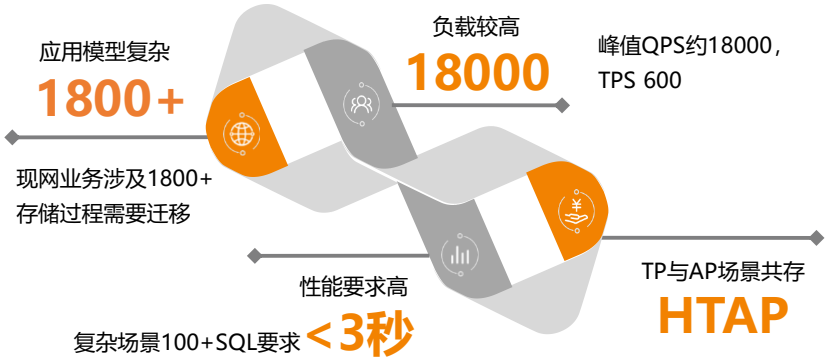
在2019年验证成果的基础上，进一步完成声誉风险系统、商密公文系统及贵金属交易系统的迁移和上线，分布式OLTP数据库能力得到进一步验证

分布式数据库GaussDB的应用情况



业务系统	原数据库	现数据库	上线时间	部署形态
生物特征识别系统	MySQL	GaussDB	2019.11	分布式，同城跨DC双活
中间业务管理平台	Oracle	GaussDB	2019.11	
声誉风险系统	Oracle	GaussDB	2020.04	分布式，同城跨DC双活
商密公文系统	Oracle	GaussDB	2020.08	分布式，同城跨DC双活
贵金属交易系统	DB2大机	GaussDB	2020.09	分布式，同城跨DC双活
网讯系统	SQLServer	GaussDB	2020.Q4	分布式，同城跨DC双活
办公门户系统	Oracle	GaussDB	2020.Q4	分布式，同城跨DC双活

商密公文系统复杂度高，挑战大，双方紧密配合，确保项目如期上线



分布式数据库GaussDB数据库的优化

2020
聚焦去O业务场景，功能、性能具备可规模商用能力

2021
对标DB2高可用方案，具备大机下移能力

	7/30	9/30	12/30	
上线业务	• 生物特征识别 • 中间业务系统 • 声誉风险 • 商密公文系统（一期）	• 商密公文系统（二期） • 贵金属交易系统（大机下移）	• 办公门户 • 网讯系统 • 手机银行系统	• Oracle系统迁移批量商用 • DB2大机下移试点商用
内核优化历程	**分布式存储过程性能优化** • Stream线程池、分布式执行下推、分布式参数化路径、大并发线程池等，商密公文TP响应时延小于3s或不低于Oracle响应时间，满足上线要求 **同城高可用容灾** • 支持同城跨DC双活容灾部署，RPO=0，RTO<60s	**分布式存储过程功能补齐** • 物化视图 • 存储过程Commit/Rollback **运维监控能力增强** • 性能报表能力增强，包括：会话锁定关系历史信息，CN/DN分布式网络通信信息	**两地三中心高可用容灾** • 支持异地跨城市（>1000公里）容灾，满足金融核心业务系统商用上线HA容灾诉求 **支持List/Range分布** • 支持List/Range数据分布策略 **支持小集群部署模式** **运维监控能力增强**	**同城跨DC RPO=0双集群高可用** • 满足大机下移业务商用上线要求 **云化弹性伸缩** • 计算存储分离架构 • 计算弹性伸缩，存储在线扩容 **软硬件结合** • 基于鲲鹏4P高性能服务器的多核设计 • 基于RDMA的高性能网络 **AI4DB，智能运维**
工具优化历程	**语法迁移** • Oracle语法转换率70% **数据迁移与同步** • 数据实时同步，支持LOB等大数量图文信息 • 支持对接第三方平台（工行OSCM系统）	**语法迁移：** • Oracle语法转换率80% • 语法转换异常对象自动识别并定位错误行 **数据迁移与同步** • 全量、增量任务灵活配置，支持多种模式 • 支持暂停、继续、重试等功能	**语法迁移：** • 预迁移评估（源库画像、兼容性分析、迁移工作量分析、TOP风险SQL识别） • 可视化迁移流程，支持迁移对象过滤 **数据迁移与同步：** • 异构数据库数据比对	**语法迁移：** • 应用层SQL语法持续迁移 • PLSQL转Java **数据迁移与同步：** • 在线增量数据实时对比

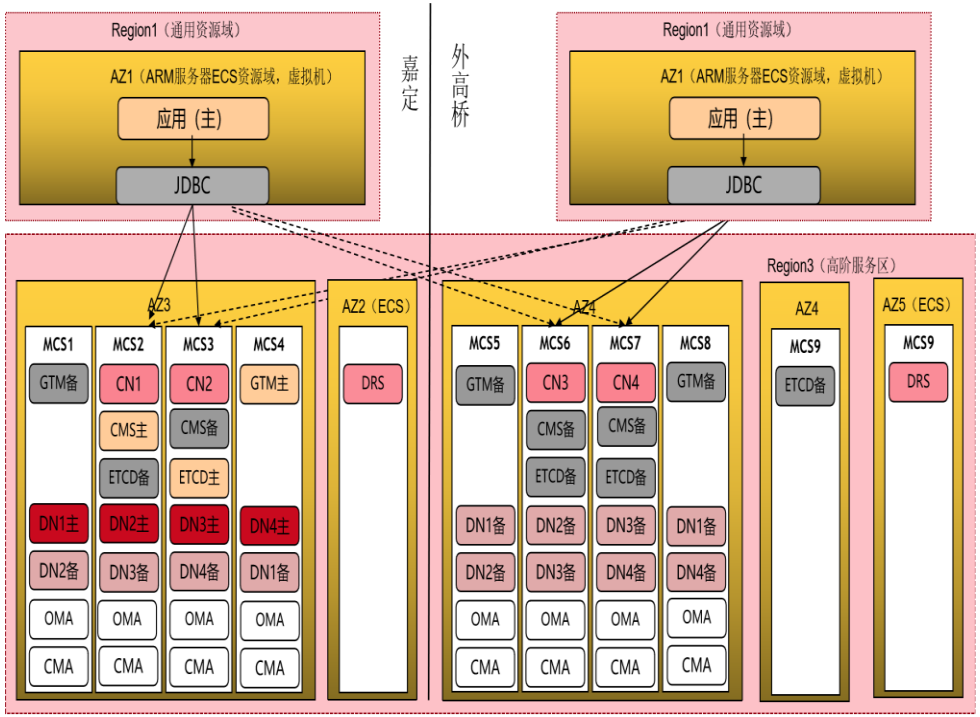
GaussDB分布式数据库探索成果-与云平台融合

高可用架构说明

- 基于MCS容器单元部署，一个容器内部整合部署多个组件，包括协调节点CN，数据节点DN，高可用管节点CMS，事务管理节点GTM，仲裁节点ETCD，OMA和CMA代理服务。
- MCS容器资源标准化，最小标准8核64G内存，最大标准80核640G内存（独占一台服务器设备），支持跨资源域的隔离和部署。

高可用场景	RPO	RTO
单节点故障场景---数据节点	0	小于30秒
单节点故障场景---管理节点	0	小于50秒
园区级故障场景-嘉定	0	小于60秒
园区级故障场景-外高桥	0	小于10分钟 （手工一键式切换）
园区内网络域故障场景	0	小于60秒

注：对于外高桥园区级高可用故障场景，由于ETCD仲裁节点副本无法满足多数派的要求，因此需要人工介入处理，RTO时间小于10分钟。



GaussDB分布式数据库探索成果-小集群（多租户）

方案目标

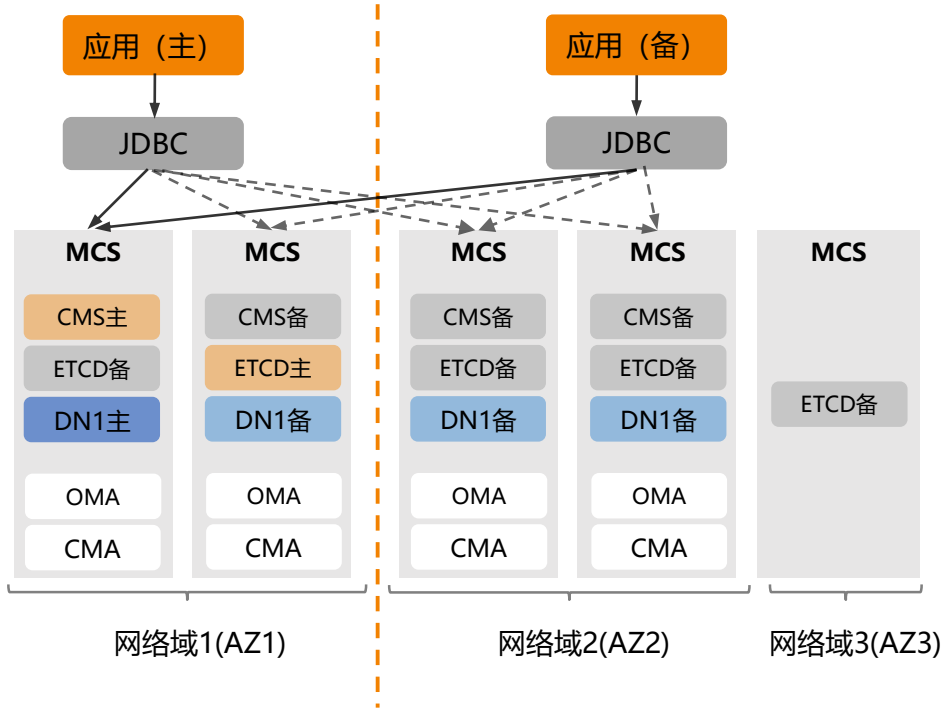
- 新增GaussDB部署功能，满足中小体量应用数据库转型需求，减轻迁移调优的难度和技能要求，降低设备投入成本。

部署特点

- 单租户内数据集中在单个DN节点，1主3备，通过CM实现本地和同城高可用切换。
- MCS容器内轻量化部署，取消CN和GTM节点。
- 小集群部署模式提供纵向扩展，无水平扩展功能。

技术优势

- **优化交易响应时间**：单个租户内数据都集中在一个节点，减少网络交互；
- **减少迁移难度**：无分布式模式数据库的各种限制（如唯一索引需要包含分布列、分布列无法更新），也不需要考虑多表连接的分布列选择；
- **减少应用数据库改造**：不需要应用做分布式改造，将查询局限于单个节点；
- **可以更好的支持存储过程等对象**：因存储过程和SQL执行都在一个节点（分布式模式下存储过程在CN节点，SQL在DN执行），可高效执行存储过程。



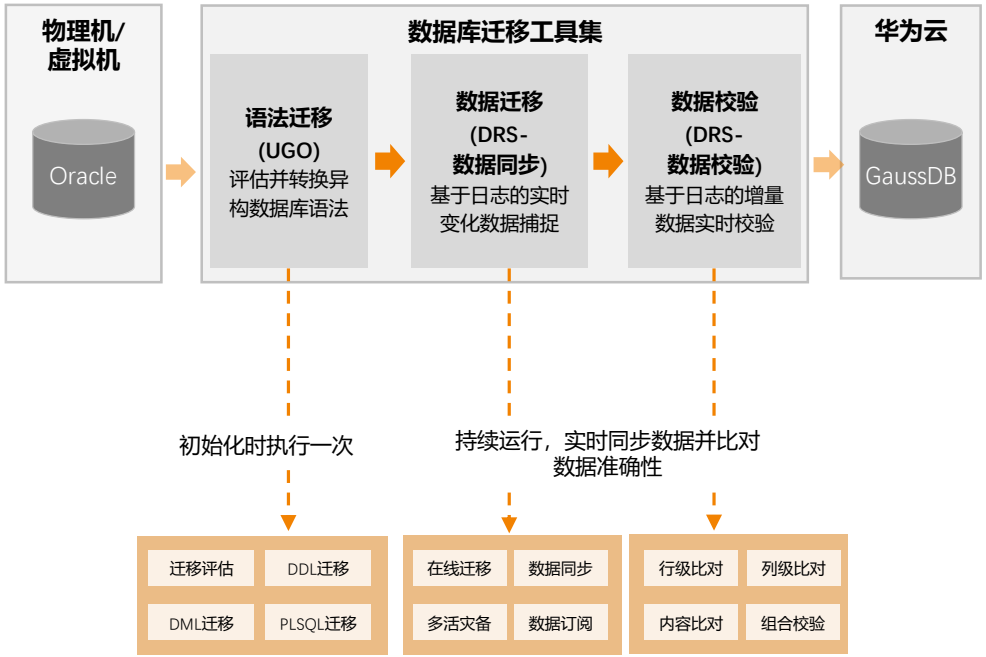
GaussDB分布式数据库探索成果-数据库对象迁移工具

工具成效

- 数据库迁移工具UGO主要承担Oracle数据库对象分析、同义改写、验证对象的工作。对于存储过程的语法迁移自动化率达80%左右，可减少70%以上存储过程改造成本。
- 数据复制工具DRS，基于日志的实时变化数据捕捉支持Oracle数据全量和增量复制迁移，同时也具备数据校验功能。

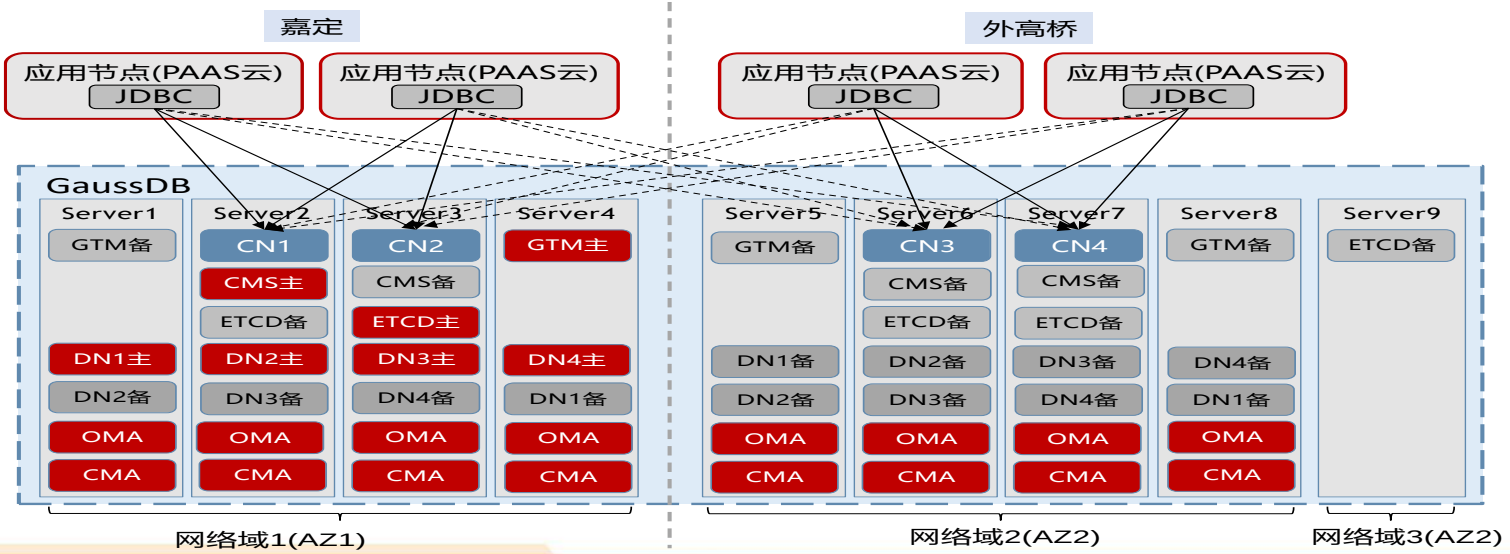
优化目标

- 易用性优化：包括迁移报表展示、代码文本显示、改写后的代码导出功能等后续持续支持和优化；
- 优化迁移效率：目前大部分Oracle数据库对象可以通过自动化迁移，但是存储过程仍然是手工修改的难点，仍需要通过提高迁移成功率来减少人力的投入。同时优化代码迁移点识别、迁移知识库建设等。



GaussDB分布式数据库技术探索-实物贵金属

- **场景概述：**支持融e购、微信小程序、柜面、网银等多渠道接入开展实物贵金属的报价、销售及回购等业务，系统评估用户数50万，日均业务量10万，另还需支撑限量稀缺商品抢购的高并发场景。
- **技术方案：**主机DB2迁移应用，通过数据库并行、渠道开关、应急回切、数据核对和监控等策略，采用灰度发布机制，按流量比例逐步切换，在保障生产稳定运行前提下验证整体迁移方案。
- **实施效果：**从技术层面验证主机DB2迁移到GaussDB的可行性，在同城双活架构模式下，平均交易响应时间<60 ms，RPO=0、RTO<60s，满足业务相关技术指标要求。



GaussDB分布式数据库技术探索小结

明确GaussDB转型场景和策略

- 技术上验证Oracle、主机DB等应用数据库转型。
- 明确了后续转型的技术路线和策略。

推动GaussDB产品的持续优化

- 提交了100多项需求及问题（完成40项需求），促进产品完善。
- GaussDB小集群（云上多租户）的架构部署规划；GaussDB数据库与云底座深度融合。
- 完善和丰富配套迁移技术文档和工具。

初步建立了与行内系统对接

- 与生产运维系统对接，实现监控告警的集成和展示。
- 与应用版本集成，实现应用元数据管理和版本投产上线自动化

形成了完备的迁移方案

- 形成GaussDB开发设计指引等技术指引。
- 应用并行运行的双库并行架构、灰度发布、回切等关键技术方案

分布式数据库后续应用策略

坚持并行并重策略

开源+商业、自主研发+外部技术引入两条腿走路，大胆创新、稳妥推进

聚焦承接主机核心能力建设

与主机能力对标，加强分布式数据库企业级能力建设。

紧跟业界产品技术发展

持续保持对业界主流产品技术路线发展演进跟踪。

加快Oracle替代转型

规划存量Oracle应用迁移，促进国产分布式数据库产品的成熟和完善。

THANKS