阿里云瑶池

PolarDB分布式版架构介绍

王江颖

PolarDB开源架构师

2024/03/14

PolarDB分布式版是什么

PolarDB分布式版（PolarDB-X）是由阿里巴巴自主研发的**云原生分布式**数据库，是一款基于**Share-Nothing**架构理念，并同时支持在线事务处理与在线分析处理 (**HTAP**）的融合型分布式数据库产品，具备金融级数据高可用、分布式一致性以及极致弹性等能力

10+ 次
经历双11

8.7 千万
TPS峰值

700+ 家
线下用户数

10000+ 台
部署规模

- 2009年集团去IOE
- 2011年7月 TDDL+ALISQL实现商品库去O
- 2012年首次承载双11，迎接零点峰值
- 2013年集团完成去IOE，TDDL成为集团业务接入标准

TDDL+AliSQL

- 产品化输出，产品名:DRDS
- 国内第一家落地分布式技术的云服务
- 2017~19年国家税务、国家路网等基础设施系统上线

DRDS+RDS

- 计算层与存储层深度融合，完整数据库形态输出
- All in PolarDB-X（金融云、公有云、零售云）
- 满足金融行业的一致性、业务连续性要求

PolarDB-X
云原生分布式数据库

Now

PolarDB分布式版典型业务场景

交易订单及相关高并发场景

- 数据量大/并发高；
- 相互联系较弱；

海量数据集中存储、大表拆分+高并发

- 数据归集和查询服务；
- 数据单表过大有并发；

国产化分布式改造

- 核心银行、运营商的部分业务，存在国产化、分布式、去O诉求；

TiDB/MyCat/Sharding-JDBC的用户

- 自建这些产品，运维管理复杂度非常高；
- PolarDB-X在热点扩容、只读实例等有明确优势；

有分布式改造诉求

- 业务未来的数据量非常大；
- 对分布式方向认可。

分布式焦点问题

业务连续性

- 如何提供与单机一致的RPO能力
- 多副本数据如何保证分布式一致
- 是否提供异地容灾多活能力
抵御机房级故障

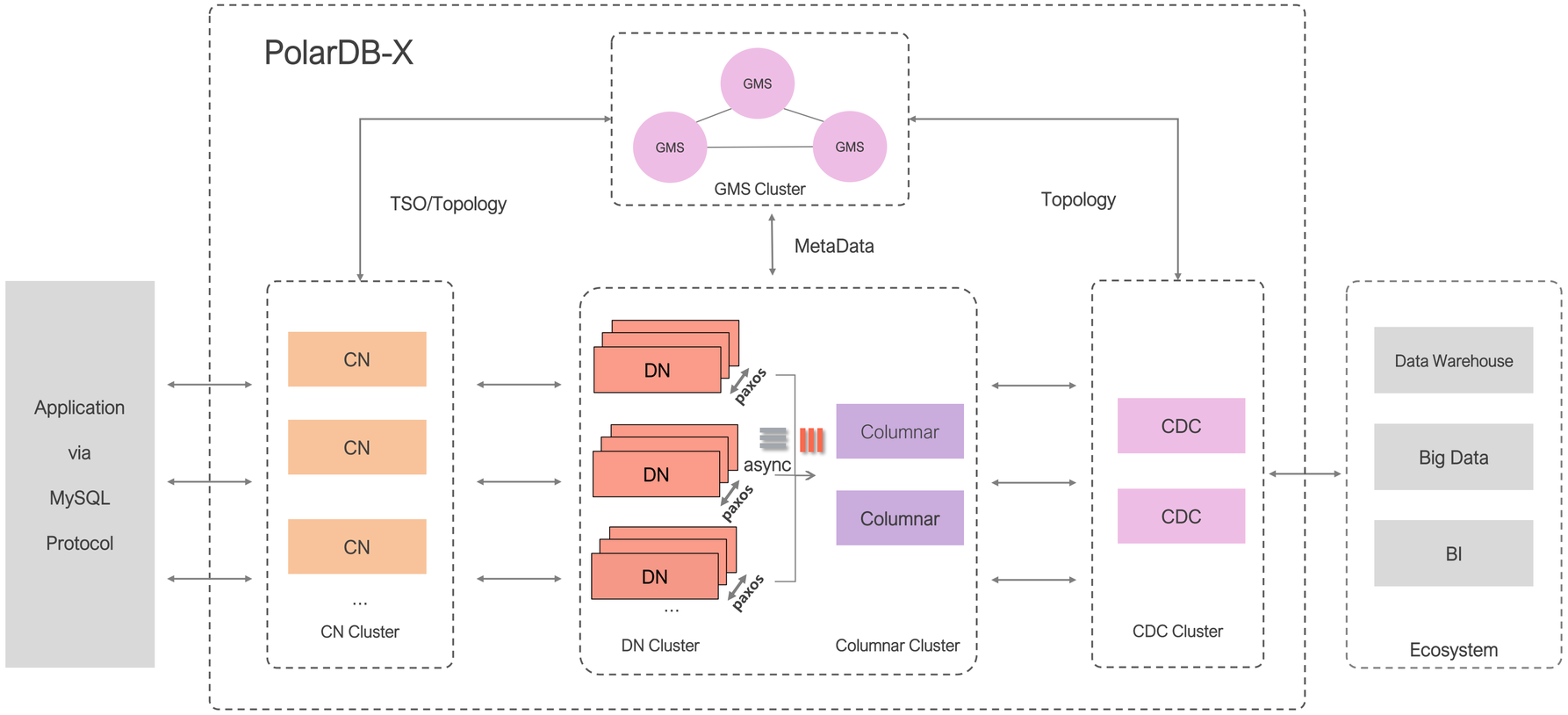
一致性保障

- 跨节点事务如何保证一致
- 如何应对更复杂的批量或查询场景
- 上下游的一致性如何保证
（数据同步、备份恢复）

用户透明

- 如何提供无限接近单机的使用体验
- 如何减少应用的侵入性
- 是否兼容已有生态，降低整体持有成本

PolarDB分布式版技术架构



元数据服务（Global Meta Service，GMS）

- 提供全局授时服务(TSO)
- 维护Table/Schema、Statistic等Meta信息
- 维护账号、权限等安全信息

计算节点（Compute Node，CN）

- 基于无状态的SQL引擎提供分布式路由和计算
- 处理分布式事务的2PC协调、全局索引维护等

存储节点（Date Node，DN）

- 基于多数派Paxos共识协议的高可靠存储
- 处理分布式MVCC事务的可见性判断

列存节点（Columnar Replica，CR） HOT

- 提供表级的列存副本，满足行列混存

日志节点（Change Data Capture，CDC）

- 提供兼容MySQL生态的binlog协议和数据格式
- 提供兼容MySQL Replication主从复制的交互

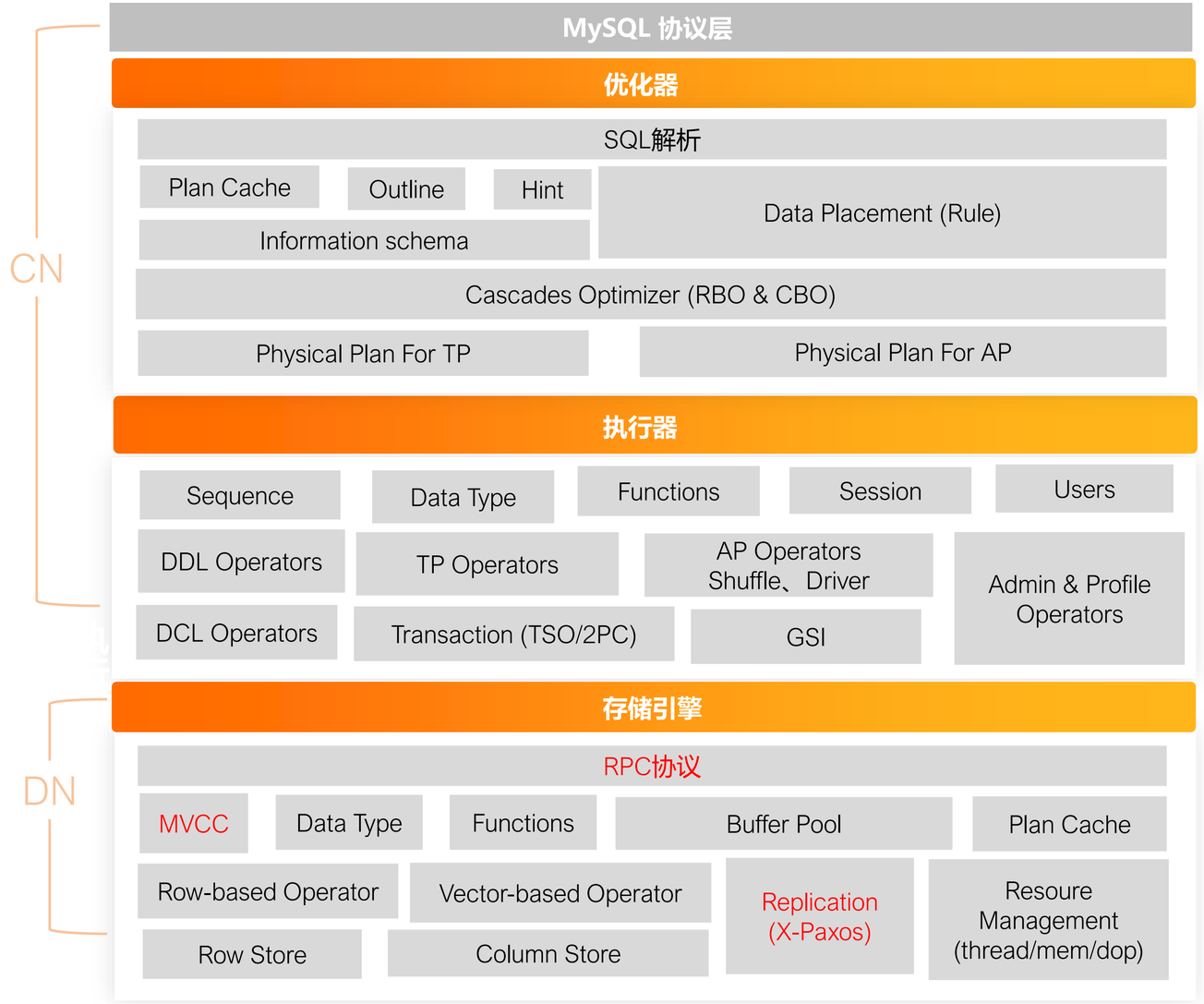
PolarDB分布式版 ~ CN/DN组件

计算节点

- 经历多年实战磨练，MySQL语法高度兼容
- 完整的SQL解析层，实现精准算子下推
- Serverless无状态，弹性能力对业务透明
- 提供HTAP 并行计算能力，应对混合负载场景

数据节点

- 基于AliSQL内核，历经多年考验，稳定可靠
- 基于Paxos强一致协议，高可用能力进一步提升
- 全局MVCC改造，满足持金融级一致性要求
- RPC协议改造，提升节点间通讯性能



PolarDB分布式版 ~ CDC组件

CDC节点

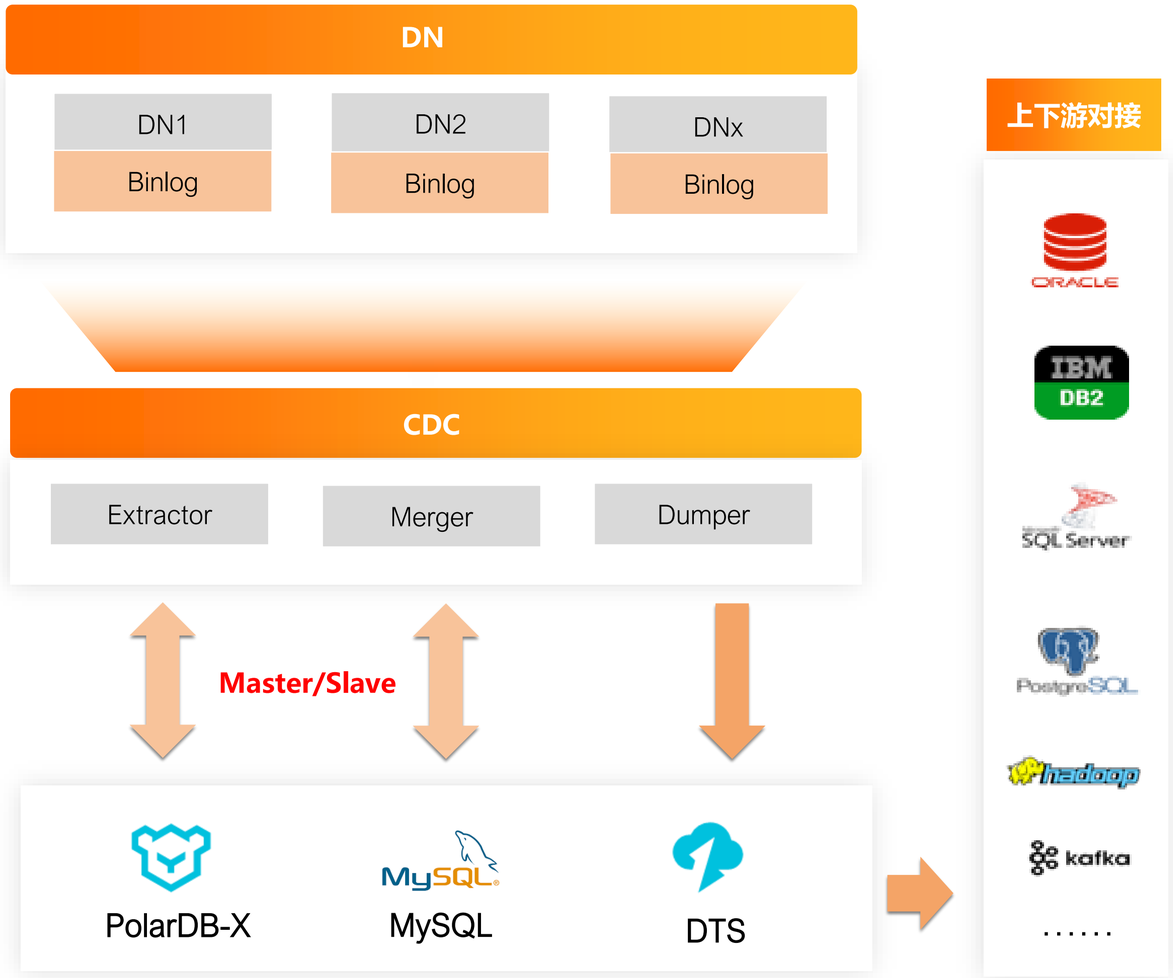
- EX：并行采集所有DN的变更日志
- MR：分布式事务日志/DDL排序重组
- DP：全局日志落盘并提供标准Binlog服务

全局Binlog

- 兼容事务 (分布式事务全局排序)
  例：基于TraceId、TSO信息对Binlog全局排序
- 兼容分布式DDL
  例：可支持DDL同步到下游，比如ADB
- 兼容分布式扩缩容
  例：屏蔽内部分片迁移、广播表、索引等数据干扰

主备Replication

- 兼容MySQL生态的主备复制
- 兼容DTS的上下游生态



PolarDB分布式版 ~ Columnar组件

列存节点

- 提供表级别的列存副本，满足行列混存
- 行存纯异步复制到列存副本，不影响TP行存
- 基于行存事务TSO版本，行和列的副本均满足数据一致性
- 存储采用分布式shard + 共享存储，满足低成本+线性扩展
- 列存对接CN节点的MPP并行计算，一个入口 + 一套SQL引擎
- 优化器智能选择列存索引，提供Select/ETL下的行和列混合执行