ALIBABA CLOUD

阿里云

云原生分布式数据库 PolarDB-X

基本原理

文档版本：20201224

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.扩展性原理

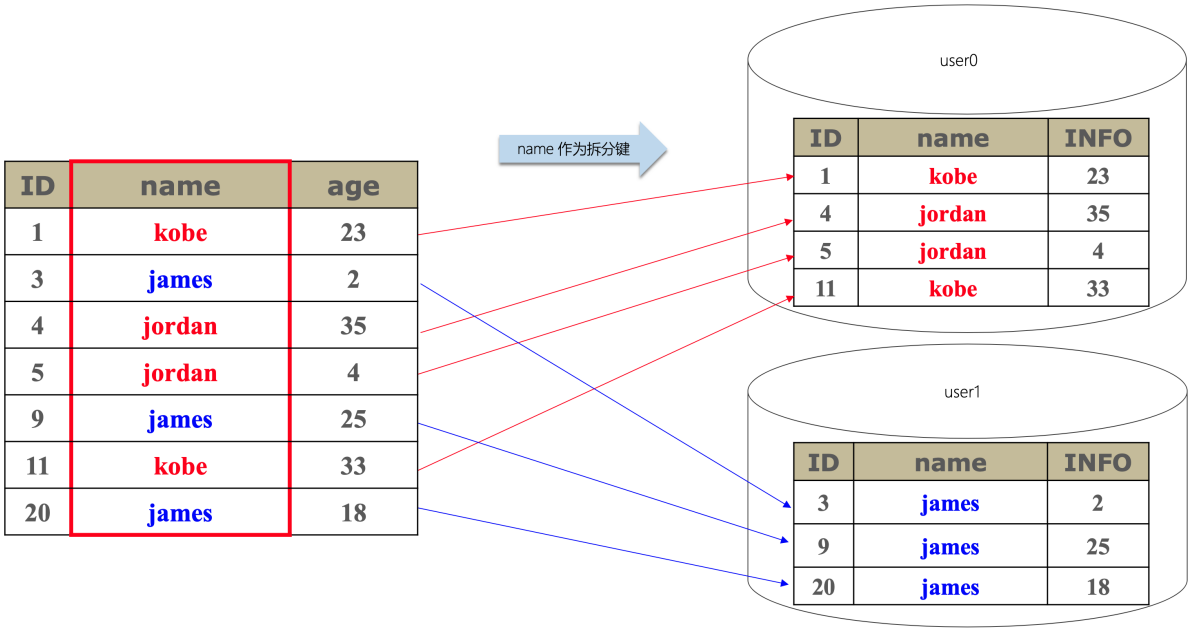
本文将介绍PolarDB-X的扩展性原理。

并发和存储容量扩展性

扩展性本质在于分而治之，PolarDB-X计算资源通过**水平拆分（分库分表）**和**垂直拆分**，将数据分散到多个存储资源MySQL以实现获取数据读写并发和存储容量分散的效果。

水平拆分（分库分表）

您可以通过一定的计算或路由规则放置数据，实现将数据分散到多个存储资源MySQL的目的，实际上PolarDB-X具备相当丰富的算法来应对各种场景。



ID	name	age
1	kobe	23
3	james	2
4	jordan	35
5	jordan	4
9	james	25
11	kobe	33
20	james	18

user0

ID	name	INFO
1	kobe	23
4	jordan	35
5	jordan	4
11	kobe	33

user1

ID	name	INFO
3	james	2
9	james	25
20	james	18

计算扩展性

无论是水平拆分还是垂直拆分，PolarDB-X常常碰到需要对远超单机容量数据进行复杂计算的需求，例如需要执行多表JOIN、多层嵌套子查询、Grouping、Sorting、Aggregation等组合的SQL操作语句。

针对这类在线数据库上复杂SQL的处理， PolarDB-X额外扩展了单机并行处理器（Symmetric Multi-Processingy，简称SMP）和多机并行处理器（DAG）。前者完全集成在PolarDB-X内核中；而对于后者，PolarDB-X构建了一个计算集群，能够在运行时动态获取执行计划并进行分布式计算，通过增加节点提升计算能力。

2.平滑扩容

本文将介绍PolarDB-X平滑扩容的基本原理。

当逻辑库对应的底层存储已经达到物理瓶颈时，则需要对底层存储进行水平扩展。例如当磁盘余量接近30%时，您可以在控制台上通过平滑扩容来改善。

平滑扩容是一种在线水平扩容方式，既把原有的分库平滑迁移到新添加的私有RDS实例上，通过增加私有RDS实例的数量来提升总体数据存储容量，从而降低单个私有RDS实例的处理压力。

3.分布式事务

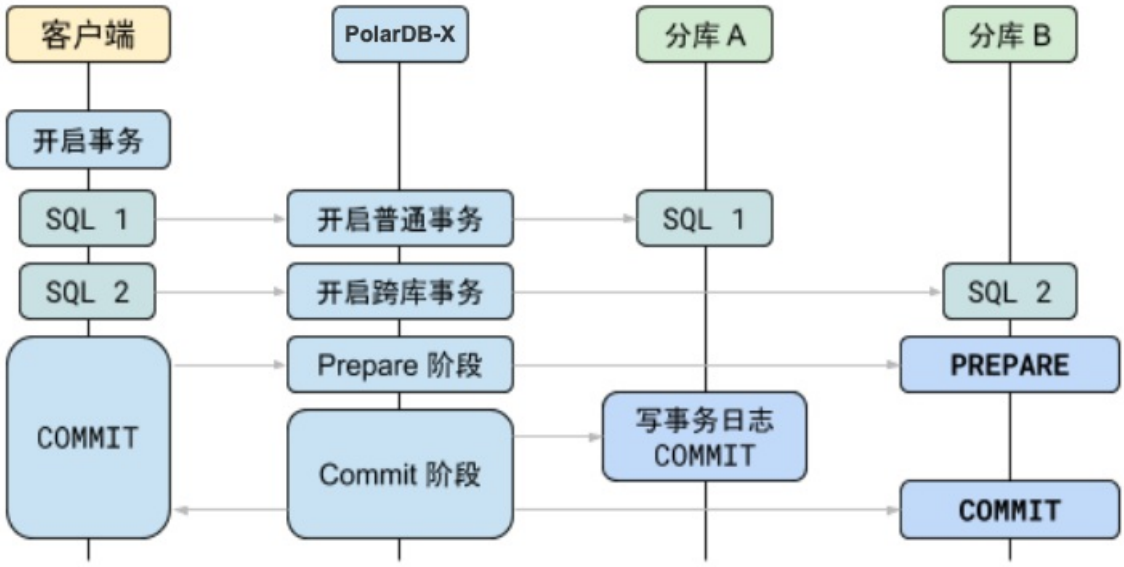
本文主要介绍PolarDB-X分布式事务的基本原理。

分布式事务通常使用二阶段提交来保证事务的原子性（Atomicity）和一致性（Consistency）。

二阶段事务会将事务分为以下两个阶段：

- 准备（PREPARE）阶段：在PREPARE阶段，数据节点会准备好所有事务提交所需的资源（例如加锁、写日志等）。
- 提交（COMMIT） 阶段：在COMMIT阶段，各个数据节点才会真正提交事务。

当提交一个分布式事务时，PolarDB-X服务器会作为事务管理器的角色，等待所有数据节点（MySQL服务器） PREPARE成功，之后再向各个数据节点发送COMMIT请求。



关于如何使用分布式事务，详情请参见基于MySQL 5.7的分布式事务。

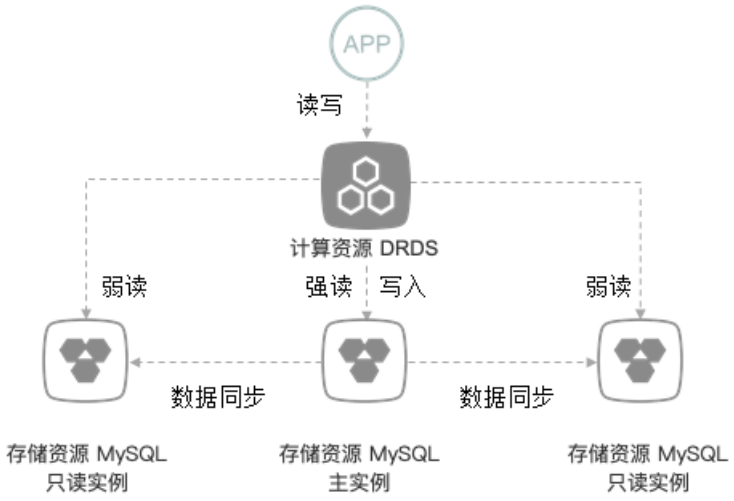
4.读写分离

本文将介绍PolarDB-X读写分离功能的原理。

当PolarDB-X存储资源MySQL主实例的读请求较多、读压力比较大时，您可以通过读写分离功能对读流量进行分流，减轻存储层的读压力。

PolarDB-X读写分离功能采用了对应用透明的设计。在不修改应用程序任何代码的情况下，只需在控制台中调整读权重，即可实现将读流量按自定义的权重比例在存储资源MySQL主实例与多个存储资源只读实例之间进行分流，而写流量则不做分流全部到指向主实例。

设置读写分离后，从存储资源MySQL主实例读取属于强读（即实时强一致读）；而只读实例上的数据是从主实例上异步复制而来存在毫秒级的延迟，因此从只读实例读取属于弱读（即非强一致性读）。您可以通过Hint指定那些需要保证实时性和强一致性的读SQL到主实例上执行，详情请参见读写分离Hint。



读写分离

读写分离对事务的支持

读写分离仅对显式事务（即需要显式提交或回滚的事务）以外的读请求（即查询请求）有效，写请求和显式事务中的读请求（包括只读事务）均在主实例中执行，不会被分流到只读实例。

常见的读、写请求SQL语句包括：

- 读请求：SELECT、SHOW、EXPLAIN、DESCRIBE。
- 写请求：INSERT、REPLACE、UPDATE、DELETE、CALL。

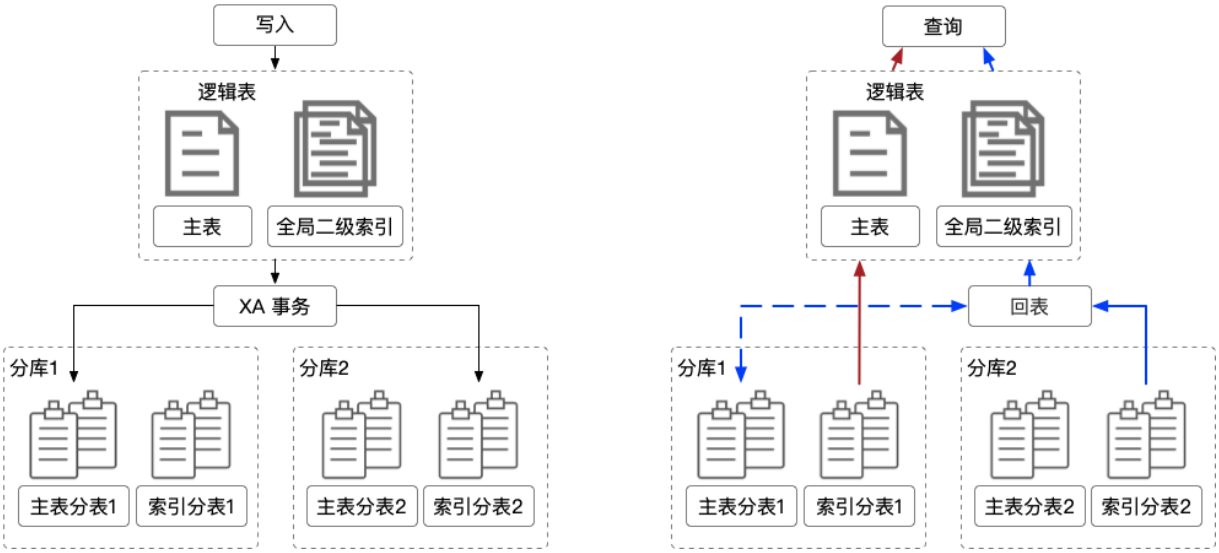
关于如何设置读写分离，详情请参见存储管理。

5.全局二级索引

本文将介绍全局二级索引（Global Secondary Index，GSI）主要功能及常见问题。

功能介绍

全局二级索引（Global Secondary Index，GSI）支持按需增加拆分维度，提供全局唯一约束。每个GSI对应一张索引表，使用XA多写保证主表和索引表之间数据强一致。



全局二级索引支持如下功能：

- 增加拆分维度。
- 支持全局唯一索引。
- XA多写，保证主表与索引表数据强一致。
- 支持覆盖列，减少回表操作，避免额外开销。
- Online Schema Change，添加GSI不锁主表。
- 支持通过HINT指定索引，自动判断是否需要回表。

常见问题

- Q：全局二级索引能够解决什么问题？

  A：如果查询的维度与逻辑表的拆分维度不同，会产生跨分片查询。跨分片查询的增加会导致查询卡慢、连接池耗尽等性能问题。GSI能够通过增加拆分维度来减少跨分片查询，消除性能瓶颈。

  > 说明 创建GSI时需要注意选择与主表不同的分库分表键，详情请参见使用全局二级索引。

- Q：全局二级索引和局部索引有什么关系？

  A：全局二级索引和局部索间的关系如下所示：

  - 全局二级索引：不同于局部索引，如果数据行和对应的索引行保存在不同分片上，称这种索引为全局二级索引，主要用于快速确定查询涉及的数据分片。
  - 局部索引：分布式数据库中，如果数据行和对应的索引行保存在相同分片上，称这种索引为局部索引。PolarDB-X中特指物理表上的MySQL二级索引

- 两者的关系：两者需要搭配使用，PolarDB-X通过GSI将查询下发到单个分片后，该分片上的局部索引能够提升分片内的查询性能。

6.HTAP

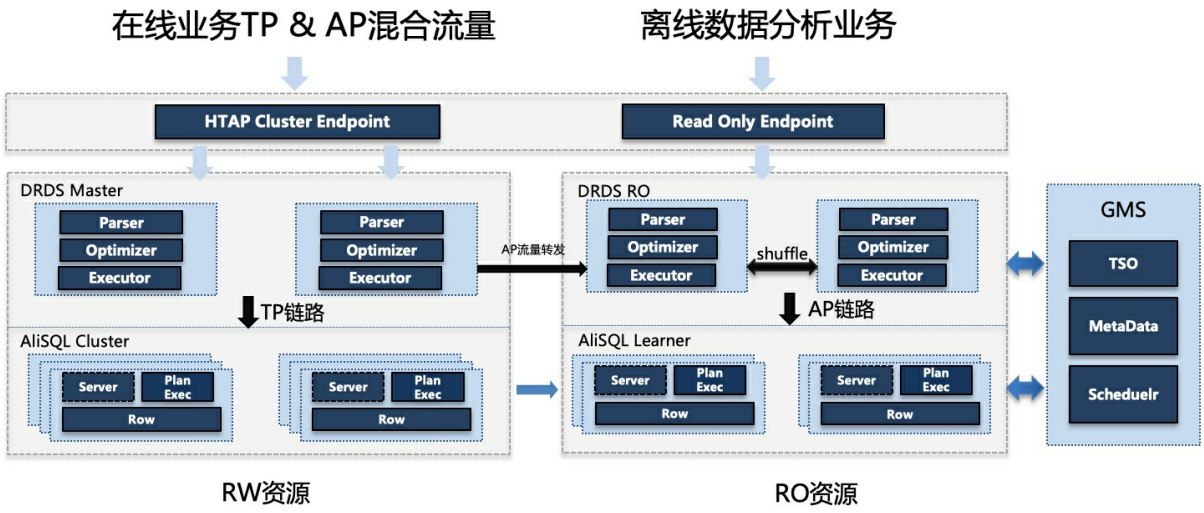
本文介绍PolarDB-X 2.0 混合事务分析处理（Hybrid Transactional/Analytical Processing，简称HTAP）特性。

背景信息

PolarDB-X 2.0解决了OLTP数据库面对海量数据下的存储、并发方面的扩展性问题，但由于缺失多机并行查询加速能力和列存储等能力，无法满足对实时性计算和复杂查询都要求较高的在线业务场景，同时还面临着ETL（Extract-Transform-Loa）数据异步传输链路运维复杂度高、数据一致性和查询实时性无法严格保障等挑战。

PolarDB-X 2.0由多个节点构成计算、存储内核一体化实例，在共用一份数据的基础上避免了ETL（Extract-Transform-Load）操作，实现了在线高并发OLTP联机事务处理以及OLAP海量数据分析，即HTAP。

技术架构



- MPP和只读资源

  PolarDB-X 2.0通过多组DRDS计算节点提供大规模多级并行处理能力（Massively Parallel Processing，简称MPP），针对计算节点进行Scale-out完成MPP处理能力的线性扩展。

  同时通过AiSQL三节点基于Paxos构建Row-based只读Learner配合DRDS只读计算节点，提供TP、AP资源链路隔离机制。

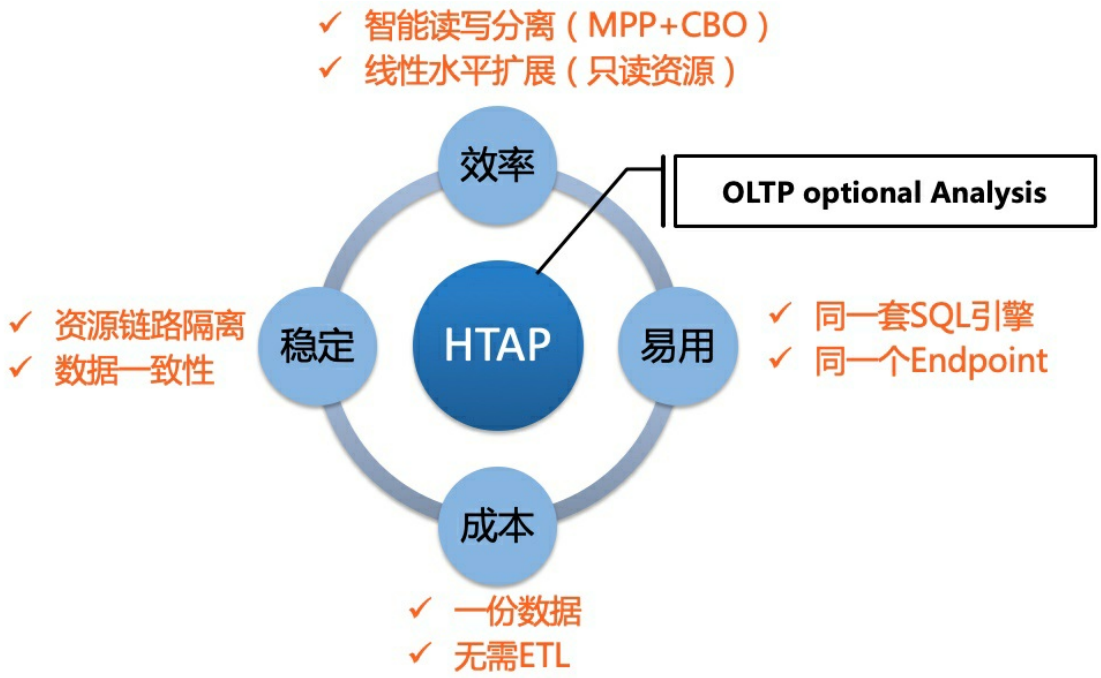
- 连接地址和数据源

  PolarDB-X 2.0TP和AP请求提供了统一连接地址（Endpoint），保持SQL语义以及兼容性完全一致。

  主实例提供HTAP集群地址（Cluster Endpoint）面向在线通用业务场景，提供了智能读写分离和强一致读特性。只读实例提供HTAP只读地址（Private Read Only Endpoint），专注离线拖数、跑批等资源链路隔离场景，确保只读资源可被独享。

  若PolarDB-X 2.0已添加只读实例，默认将AP workload转发至只读实例进行MPP并行加速；若未添加任何只读实例，则转发至主实例内部所有计算节点完成执行。

优势



- 一份数据，一个数据源，一个Endpoint即可覆盖TP和AP业务场景，降低数据库选型成本。
- 支持线性水平扩展提升HTAP复杂查询加速能力，通过横向增加只读实例即可提高复杂查询速率。
- 避免数据异步传输，满足全局数据查询一致性，提升业务实时分析效率。
- 资源链路隔离，确保在线核心业务链路稳定性。

典型业务场景

PolarDB-X 2.0可满足如下典型业务场景需求：

- 在线业务联机查询
  - 少量逻辑表关联、排序、聚合，涉及数据少量。
  - 并发较高，实时性要求高，严格一致性要求。
- 报表BI（Business Intelligence）分析查询
  - 多张大表关联、排序、聚合、子查询以及宽表统计查询，涉及海量数据。
  - 数据一致性、实时性要求不高。
- 离线拖数跑批查询
  - 大批量数据离线抽取、全表扫描、离线归档、T+1离线跑批任务，涉及多张大表，SQL较复杂。
  - 物理资源链路需隔离，不能影响在线业务，少量业务存在INSERT或SELECT需求。
  - 数据一致性、实时性要求不高。
- Adhoc交互式即系查询
  - 后台运营场景交互式标签即系查询，少量并发，少量表关联聚合，WHERE条件不固定。
  - 数据一致性、实时性要求高。