

Apache Doris

在网易互娱的管理与应用实践

胡彪

网易游戏 高级大数据开发工程师

个人介绍



胡彪 | 网易互娱 高级大数据开发工程师

- 在网易互娱负责 Trino/Doris/统一查询引擎等组件的开发维护 and 业务支持工作
- 在 OLAP 引擎开发和平台建设上有一定的研究经验
- Trino Contributor
- Apache Doris Contributor

目录

1. 背景介绍
2. 生态建设
3. 场景加速
4. 未来展望

1 背景介绍

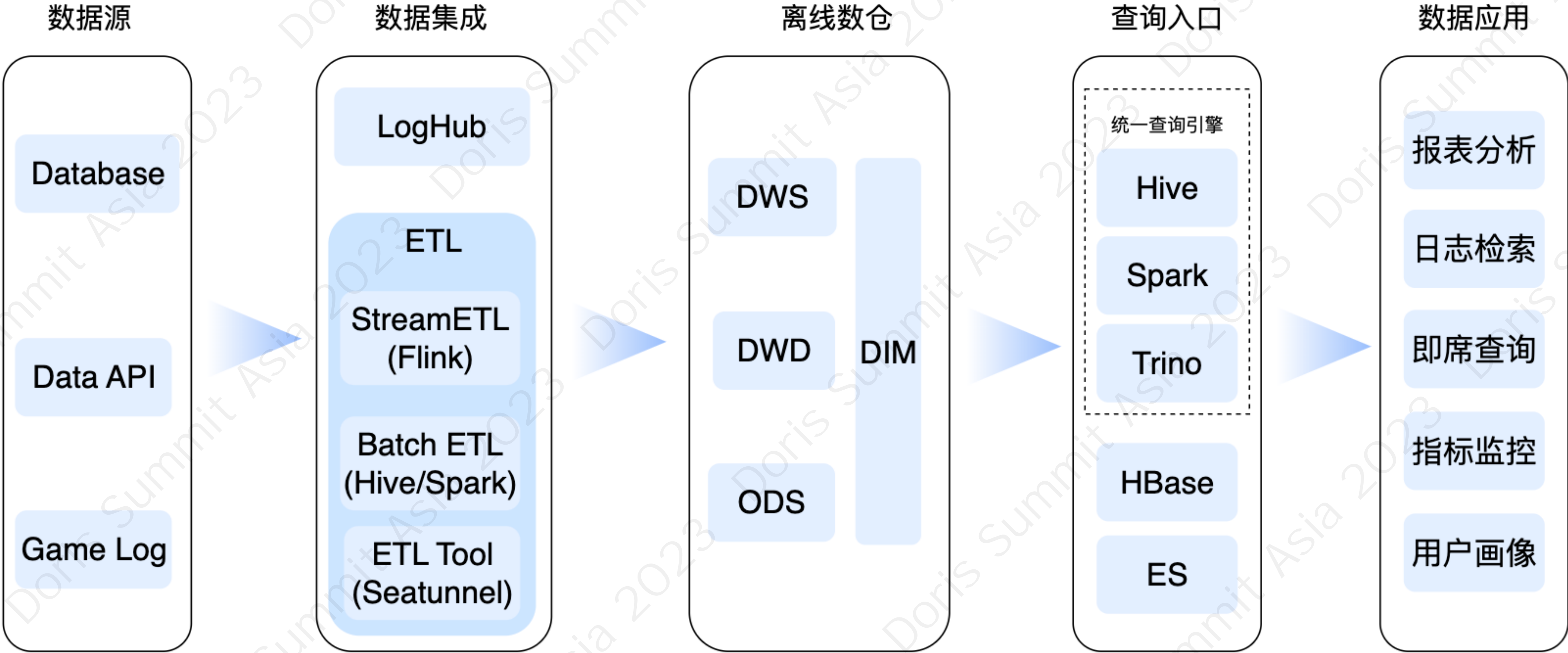
背景介绍



技术中心-数据与平台服务部负责构建和支持迭代网易互娱的数据平台和数据基础架构。

经历了从第一代离线架构到第二代 Kappa 架构演进之后，我们最终**引入了Apache Doris搭建统一的湖上实时数据仓库**。本次分享将详细介绍我们在三代架构演进中的Doris管理平台建设和业务场景落地的实践经验。

引入 Apache Doris 前的组件架构



痛点与需求

架构痛点

- 架构复杂，运维困难
- 用户研发成本高
- 数据时效性与查询效率较低

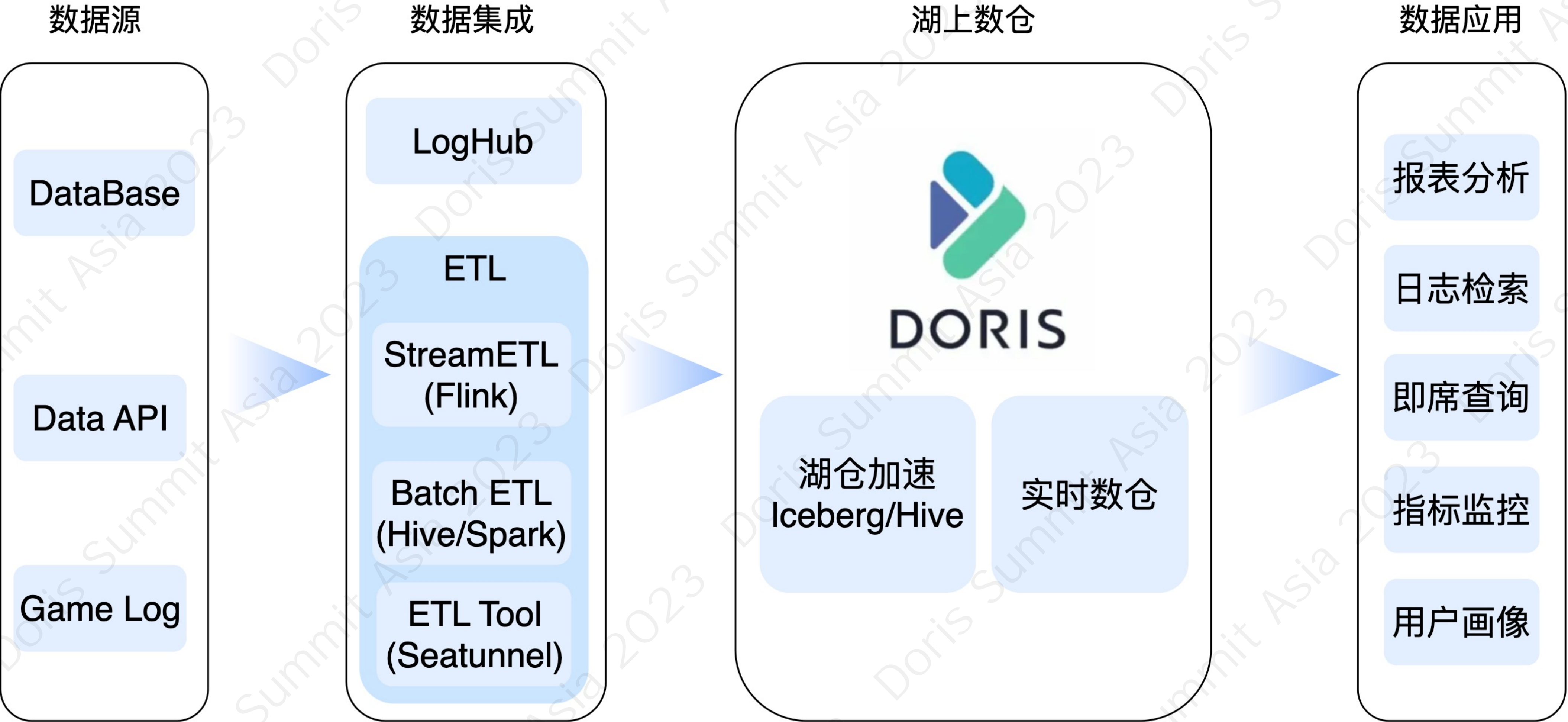
目标需求

- 架构简单，低运维成本
- 统一易用，降低用户学习成本
- 实时高效，支持实时数据导入，湖仓加速实现高效查询

选型因素

关键指标	Apache Doris
查询性能	内部SSB数据集及业务场景测试均满足性能指标要求，1.1 版本以后提供向量化执行、1.2 提供新版查询优化器，性能提升更多。
导入方式	支持离线和实时的导入方式，1.1 版本支持事务导入，能够做到实时数据写入不丢不重。
开源方式	采用 Apache License 协议，安全性高、灵活性强。
社区活跃度	官网文档详细，同时提供了多堂源码解读课程。社区活跃度较高，Issue 问题能够得到及时反馈。
使用方式	兼容 MySQL 协议，接入和迁移成本低
运维部署	不依赖第三方组件，FE/BE 扩缩容简单，同时支持数据自动平衡。

基于Doris构建高效易用的湖上实时数仓



集群规模

10+

当前总集群数
国内/海外均有布局

100+

总节点数接近百台
大部分为FE/BE混部

100+

对接内部项目数

2000000+

平均每日查询总数量

900TB+

最大的一个集群当前
存储数据总量

10TB+

通过实时作业和离线
导入的日增数据量

2 生态建设

大规模集群运维方向

1. 基础建设



- 制定好运维规范，如端口、目录、服务器类型/配置
- 制定好业务接入准则、开发规范，完善基准测试报告
- 建设好元信息，集群、实例、库等维度，集群管理员、业务线等

2. 监控报警



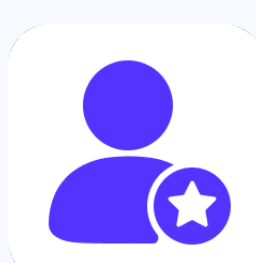
- 梳理集群分级指标、设定报警规则和升级机制
- 用户侧性能指标监控
- 统一入口，查看所有集群重点监控情况

3. 安全性保障



- 制定好备份规范、备份方式，自动化备份与恢复
- 各类故障演练
- 集群巡检、规则治理

4. 平台化



- 利用自动化工具和脚本来快速、可靠地部署和配置集群
- 开发易于使用和管理的用户界面，多租户高可靠的权限模型
- 提供数据报表和分析功能，帮助用户了解业务数据趋势和业务指标

Doris Manager 架构

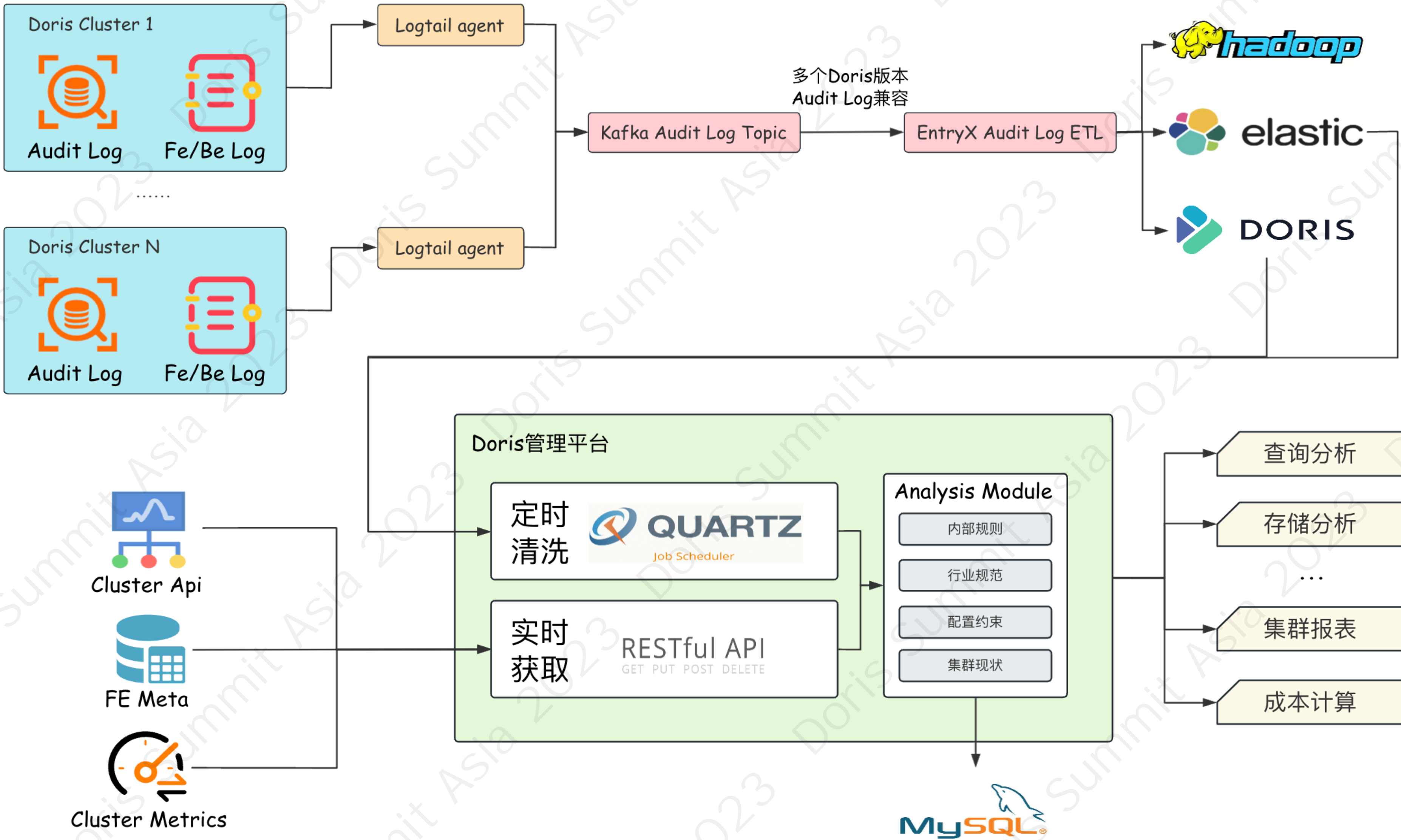
功能特性

- 可视化操作：**提供直观、简便的方式操作 Doris 集群，减小误操作的概率，快速上线业务。
- 多集群管理：**多个集群统一管理，提供一个集中化的控制面板，简化管理流程。
- 丰富的权限管理：**将资源抽象化为部门、项目级别，角色的权限可以更细化的控制，适用于多部门、多项目的统一集中管理。
- 完备的审计：**所有操作都会有完整的审计功能，方便追踪操作历史。
- 集成 Doris 常用功能：**从平台侧帮助用户了解集群使用状态、更好地使用 Doris 服务。
- 多层次监控报警：**提供多级别的监控统计信息，对于异常情况，还能够及时发出报警通知，帮助用户快速发现和解决问题，提高系统的稳定性和可靠性。



Doris 管理平台架构图

Doris Manager数据流转



集群服务设计目标及收益

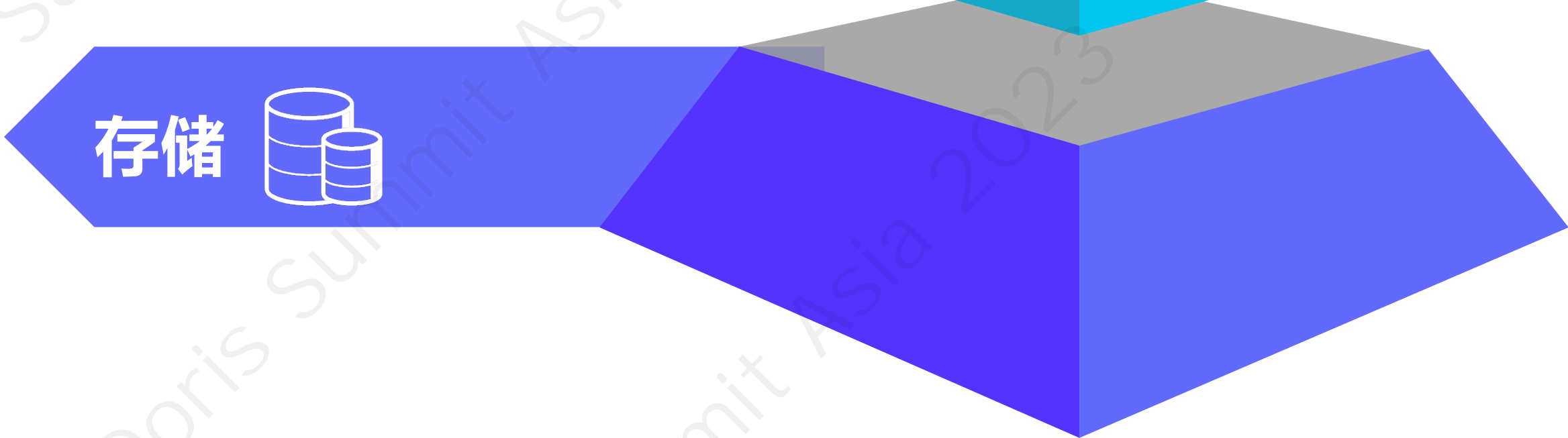
目标：帮助用户全方位了解集群查询情况，提速查询
当前收益：慢查询数减少 **80%**



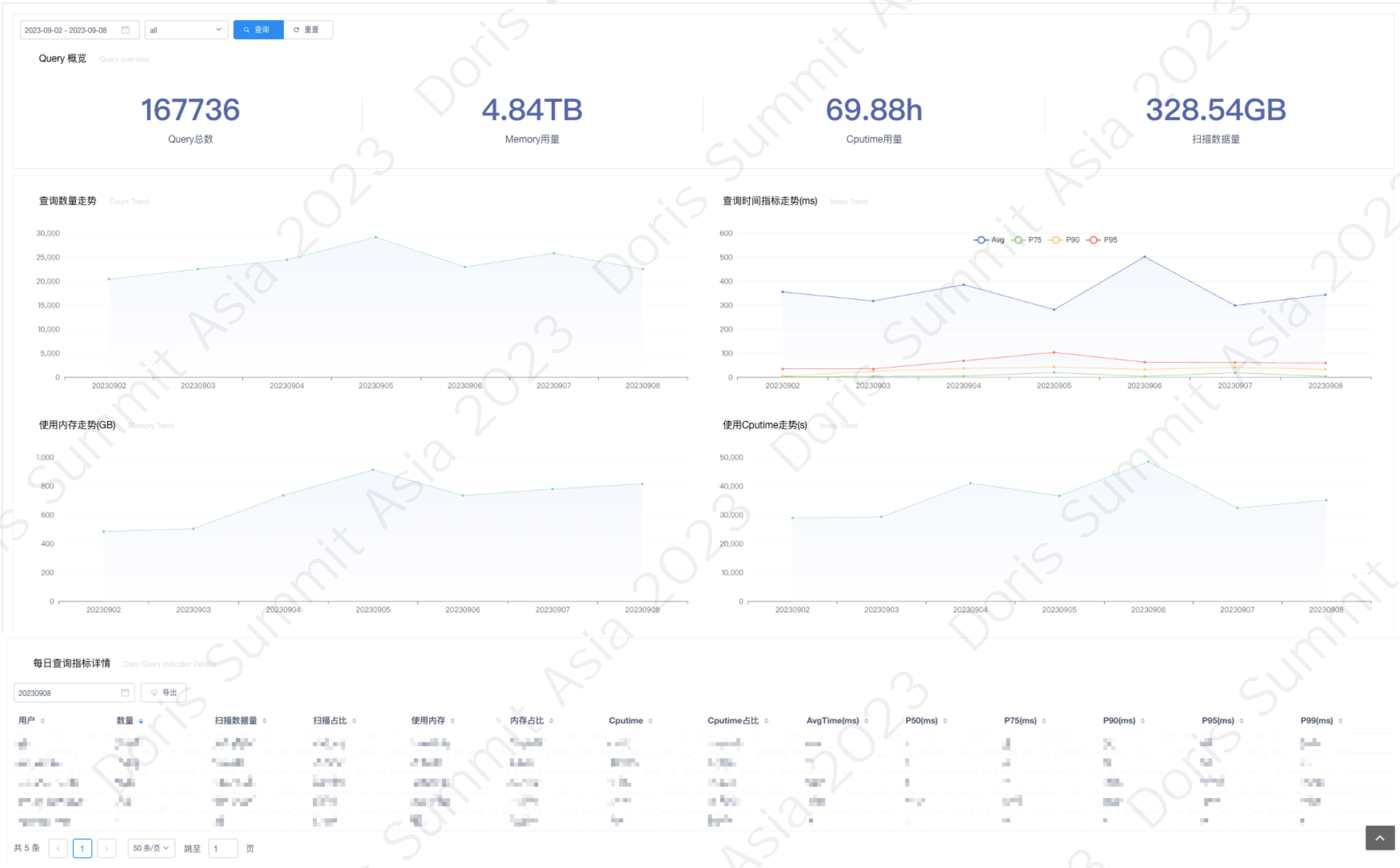
目标：帮助用户掌控实时和离线导入的数据增量情况，导入问题报错辅助排查，减少人力运维
当前收益：月均减少人力沟通次数 **70%**



目标：帮助用户减少冗余存储，节约存储成本，减轻元数据管理压力
当前收益：平均存储下降 **20%**



查询分析 – 查询概览



价值阐述

集群查询消耗的总资源和各用户消耗的资源都需要量化，以便开展用量预警、性能分析等工作。

设计内容

- 支持查看集群近3个月各指标的走势
- 数量、CpuTime、Memory、P系列指标
- 以表格的形式精确展示指标详情
- 统计集群内各用户消耗资源占比
- 所有统计指标支持导出下载

查询分析 – 当前查询

Running Query

刷新

Off

Search Data

Frontend	QueryId	ConnectionId	Database	User	ExecTime	SqlHash	Statement	操作
	8a277aab913b44b7-96c59d810c6aec27	7191474				28b8f544e770df891b4d4946c002dd79		终止 执行计划

价值阐述

Doris 对于 Running Query 的可观测支持度不好，在某些场景下用户需要知道 Running Query 的情况，以及 Kill 误提交或者不合理的大查询。

设计内容

- 支持指定时间间隔刷新查看当前集群运行的查询
- 可以 Kill 当前运行的查询、查看当前运行查询的执行计划

查询分析 – 慢查询统计

价值阐述

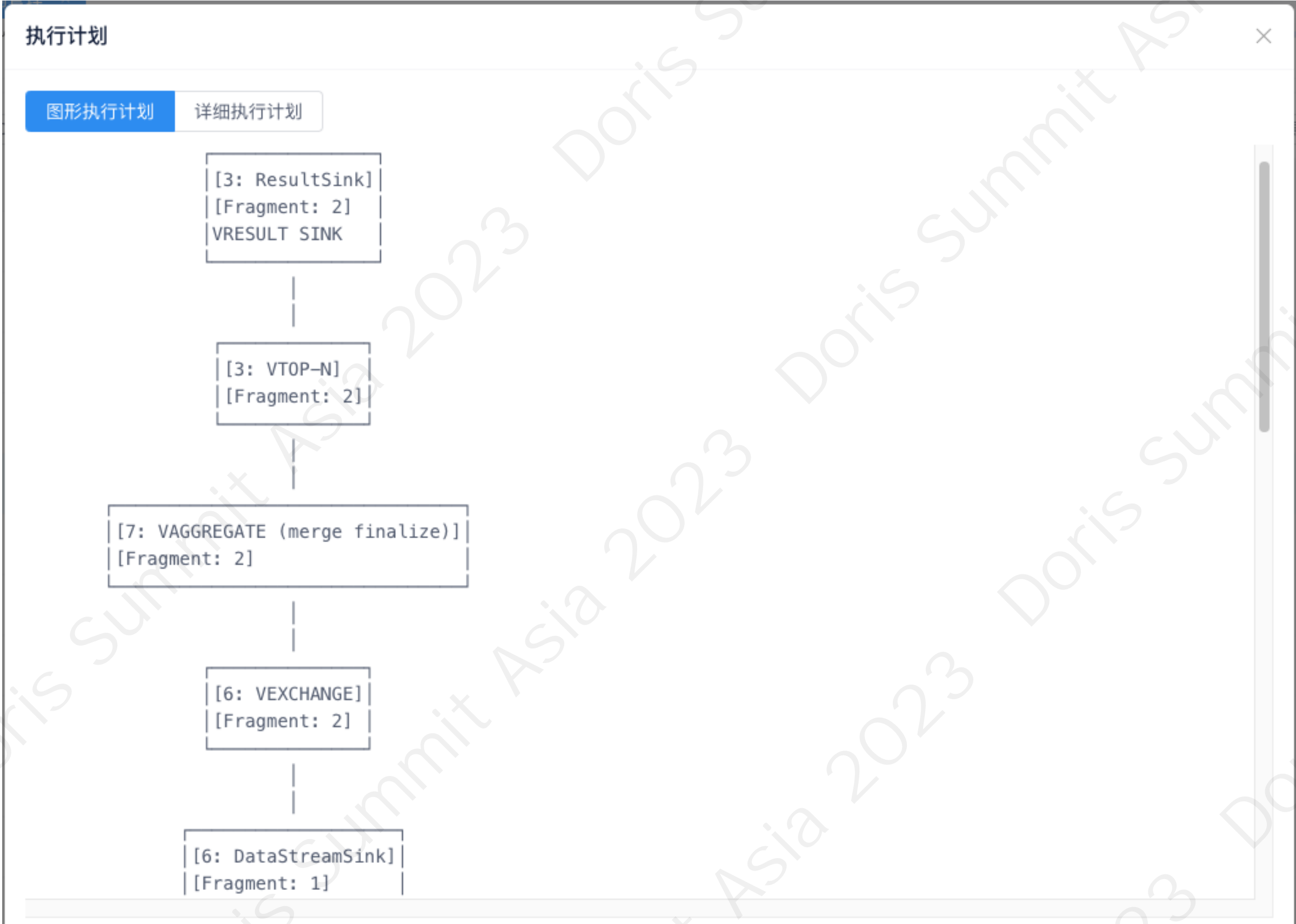
- **用户体验：**通过治理慢查询，提升用户体验
- **集群压力：**通过慢查询治理，及时释放数据库资源，降低安全风险。
- **SQL优化：**通过慢查询治理，提高SQL质量和可维护性。
- **成本节约：**减少慢查询会减少系统的资源消耗，进而降低使用成本。

设计功能

- 支持查看按库和按时间占比分布
- 支持展示慢查询在时间轴上的分布
- 可按照任意时间维度、用户、库、阈值统计各用户的慢查询明细
- 支持查看慢查询执行计划



查询分析 – 执行计划



痛点

用户通过慢查询模块只看到了SQL，无法初步判断为何慢，需要切换终端查看执行计划。



解决方案

可视化操作，提供文字和图形两种执行计划，快速定位慢查询问题。

查询 – Profile统计

[illegible]

痛点

- 对于一些无法通过执行计划准确判断缓慢原因的SQL查询，需要获取更详细的执行信息。官方提供了Profile功能来解决这个问题，但不是所有用户都了解或熟悉这个功能的存在和用法。

[illegible]

解决方案

- 平台集成该功能，指导用户开启Profile信息，查看执行过程各阶段具体的资源消耗情况；
- 在平台侧结合Profile信息给予诊断和优化指导，帮助用户优化SQL。

查询分析 – 慢查询报表



痛点

- 对于用户而言，登录 Doris Manager 平台不是必选项，日常稳定使用集群服务时，用户可能不会上平台关注查询情况。

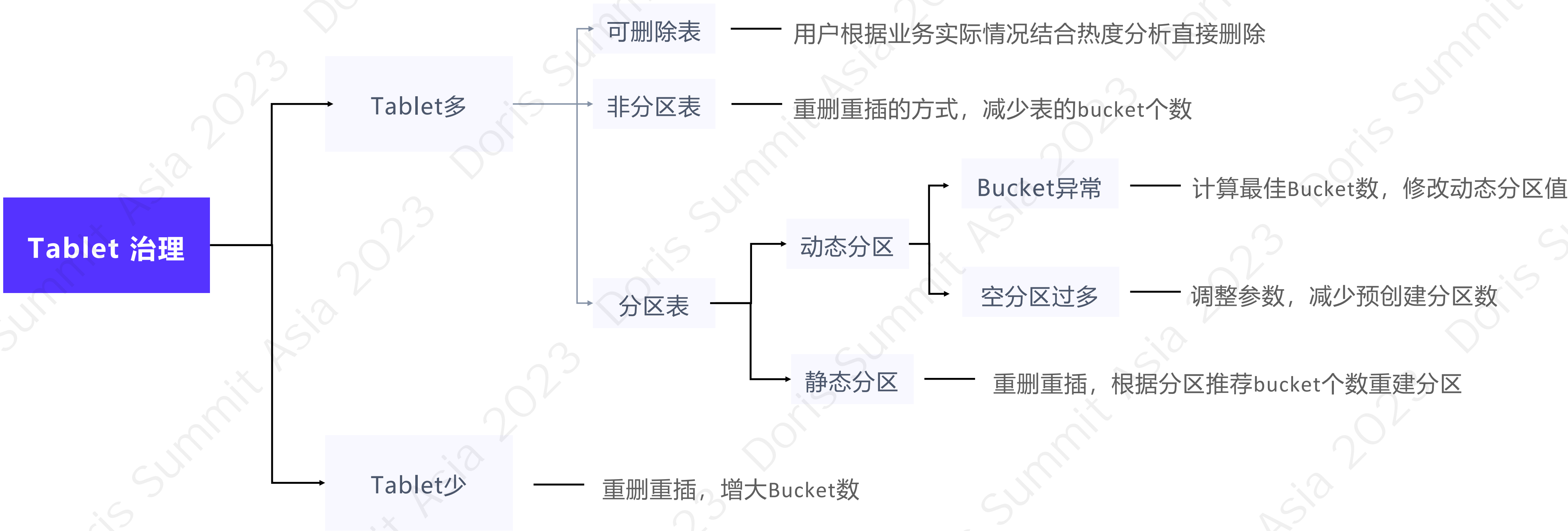
解决方案

- 每日上班前推送查询报表，报表中包含上一日慢查询分布、明细数据以及各项查询指标值。通过发送 T+1 报表，让业务洞察上一日的查询情况，及时做出相应的优化调整。

查询 - 热度分析



存储 – Tablet治理



存储 – Tablet分析

[illegible]

优化建议

基础信息

库名:

表名:

分区数:

总数据量:

现有tablet数:

可优化tablet数:

优化评估等级 ①: 提醒

热度走势

最近半个月热度走势

日期	热度
20230820	1.0
20230821	1.0
20230822	1.0
20230823	1.0
20230824	1.0
20230825	1.0
20230826	1.0
20230827	1.0
20230828	1.0
20230829	1.0
20230830	1.0
20230831	1.0
20230901	1.0

优化措施

方案一：临时表重建

方案二：重导入数据

优化步骤

具体操作步骤如下:

步骤一：新建临时表

步骤二：重新新增所有分区，并指定为符合要求的Bucket数

痛点

用户无法感知当前集群tablet的分布情况是否合理，同时在管理员给出修改要求后，不知道该采用哪种策略去处理。

解决方案

提供tablet概览界面，并根据上面提到的tablet治理类型，结合表类型及查询热度给出相应的优化措施和步骤指导。

导入分析 – 导入概览

价值阐述

- 管理员和业务用户均需要了解集群导入数据的详细情况
- 根据导入的作业数和数据量走势辅助用户评估规模

设计功能

- 离线和实时分开统计
- 支持查看近3个月各指标走势
- 包括作业数量、CpuTime、Memory



导入分析 – 报错指南

http://[redacted]api/_load_error_log?file=__shard_0/error_log_insert_stmt_f64a8f3aca96fd1b-b9bdbb

查询 清空

诊断建议:

报错释义: 数据不在已有分区范围内。

1. 如果数据是脏数据, 可以忽略, 请调大max_filter_ratio, 只要过滤的数量比例不超过这个值, 作业就不会报错退出;

2. 如果数据是正常数据, 静态表请提前创建该数据落到的分区, 动态表调整动态分区数, 直到能够自动创建出该数据落到的分区为止, 数据就能够正常导入。

Reason: no partition for this tuple. tuple=+-----+-----+
|dt(Nullable(Date))|role(Nullable(String))|
+-----+-----+
| NULL | bad |
+-----+-----+
2 rows in block, only show first 1 rows.

底层原理

Flink Doris Connector、Seatunnel 等写入方式底层都是通过Stream Load传输数据。

网络隔离

公司内部网络隔离
相关端口未对用户开放

智能诊断

根据源码分析及日常收集
报错整理到知识库, 智能
诊断用户报错。

3 场景应用

质量数据中台介绍

提升游戏性能，优化设备表现

了解更多

详情联系: [模糊处理]



简介

QData是网易互娱质量保障中心下属的大数据团队。

从质量角度出发，针对游戏产品生命周期中的支付、奖励、性能、登录等主题业务为游戏提供实时监控、离线分析、报表等服务。

关注资源循环，监控奖励发放

- 及时发现折扣异常、奖励超发、购买限制配置错误等问题
- 准确定位购买超额、vip特权异常、道具价格波动等问题

代币充值监控
道具购买监控

提升游戏性能，优化设备表现

- 实时监控外服性能状态，查看性能不达标机型、场景、用户在线性能报警
- 查看配置-性能关系，针对机型调整推荐配置

在线性能报警
机型配置推荐

分析支付过程，定位流失原因

- 梳理支付流程，找出流失订单，整理流失原因并给出召回建议
- 定位购买冲动，分析流失原因，找到有召回潜力的订单

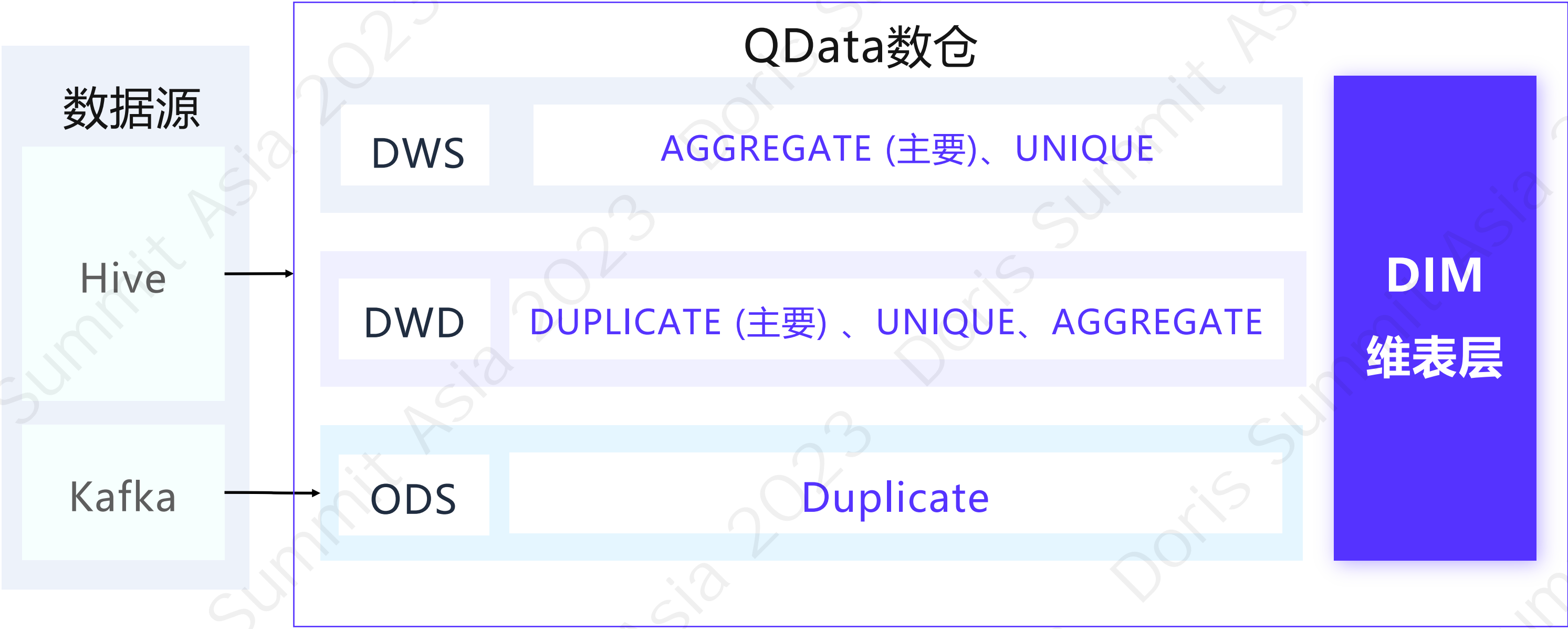
支付转化漏斗
支付窗口分析

聚焦玩家舆情，捕捉外挂黑产

- 监控游戏内外玩家讨论，及时获知玩家反馈
- 实时检测开挂玩家，线上配置处罚方案

玩家讨论分析
反外挂SDK

引擎需求



需求点:

- 日实时流数据量近百亿，并发写入数超过200，需要能够支持高并发写入；
- 支持从Hive中快速同步大量历史数据；
- 需要完整支持行为分析类型的函数，且P95指标不高于3s；
- 日常有变更字段和更新数据的需求，需要引擎支持且不影响正常写入和查询；

bitmap 查询提速

场景描述

游戏产品会在版本日当天放出游戏内容更新或优化。数据团队需要量化玩家打开游戏时，从 Patch 更新到最后登录的过程的转化情况。Patch 转化的数据场景指标需要针对玩家设备ID进行精确去重，数据量往往在 **十亿** 级别以上。

问题现状

直接 COUNT(DISTINCT) 往往会占用大量内存和IO，并且查询时间 **>20s**，特别是当表中有大量不同的值时，查询性能受到的影响更大，无法满足性能要求。



bitmap查询提速

解决方案

将玩家设备id构建全局字典表导入到 bitmap 列，或者针对明细表添加物化视图。

Aggregate 模型：bitmap

- **第一步：构建全局字典表**

由于玩家的设备id是字符串，所以需要先转化为整形，使用Hive全局字典表

- **第二步：Agg模型新增表字段：**

``udid` VARCHAR(256) => `udid_ranks` bitmap
BITMAP_UNION NULL`

- **第三步：改写查询**

`COUNT(DISTINCT udid) => COUNT(DISTINCT
udid_ranks)`

Duplicate 模型：使用物化视图

- **第一步：构建全局字典表**

由于玩家的设备id是字符串，所以需要先转化为整形，两种方式：

1. Hive全局字典表
2. 使用Doris函数bitmap_hash64

- **第二步：Duplicate模型创建物化视图：**

`bitmap_union(TO_BITMAP(udid_ranks))`

- **第三步：改写查询**

`COUNT(DISTINCT udid) => COUNT(DISTINCT
udid_ranks)`



Base表+轮转表
构建全局字典

优化收益

14亿数据	峰值内存	查询时间
优化前	54.0GB	>20s
优化后	4.2GB	<2s

物化视图提速查询

问题描述

游戏性能是玩家游戏时最直观的体验,合适的性能可以确保游戏流畅度、响应速度和画面质量。性能问题可能导致卡顿、延迟或崩溃,严重影响玩家满意度和游戏口碑和留存。因此数据团队需要对玩家游戏时的性能数据进行监控和分析。

衡量游戏性能相关的数据指标有很多,例如:FPS、卡顿次数、内存峰值等8种,单独一个指标相关的维度更多达10个。游戏策划希望在网页端可以针对多种指标和多个维度进行自定义聚合查询,查询响应时延需要控制在2s内。

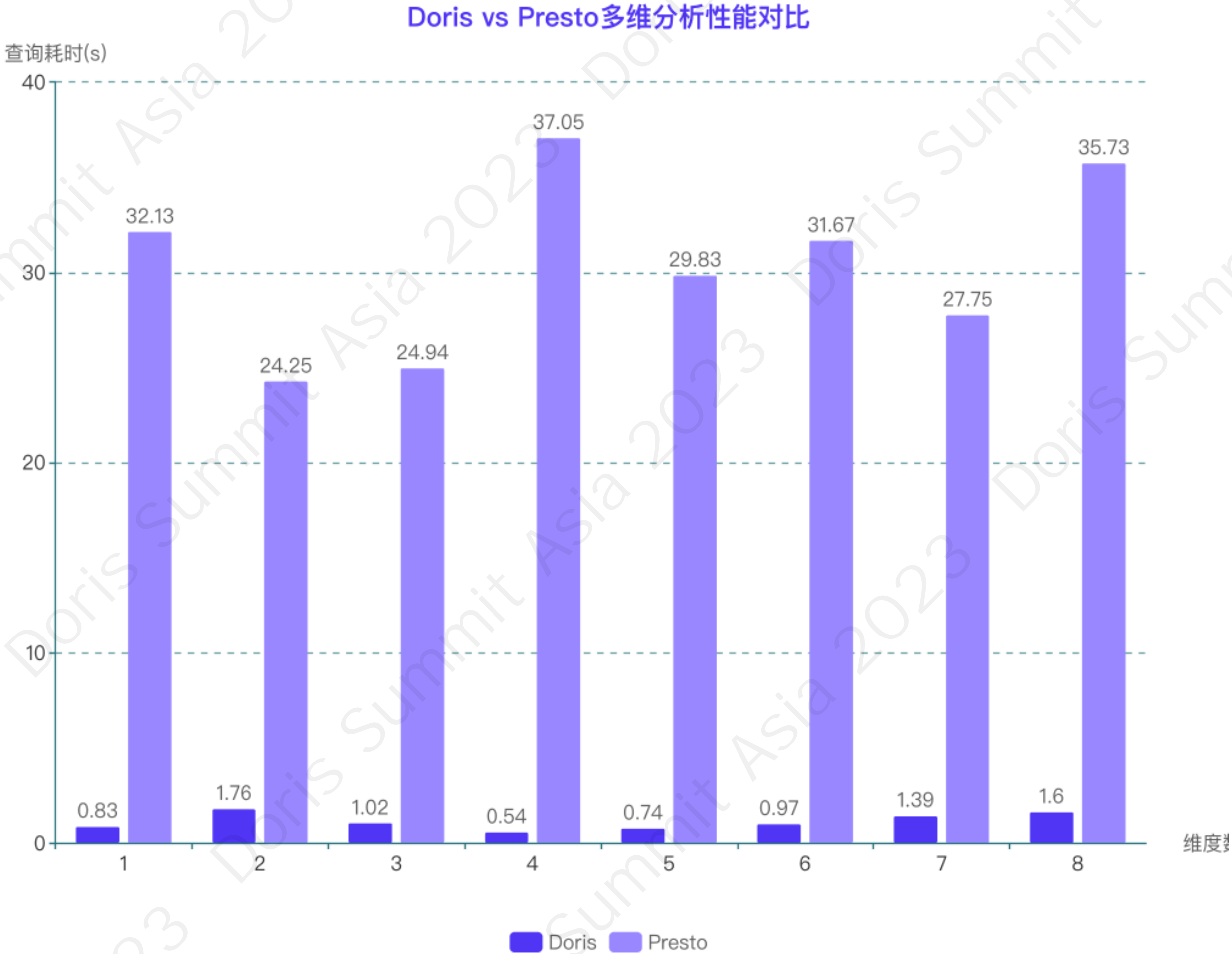


物化视图提速及收益

解决方法

针对常用的维度设计物化视图，可以满足用户绝大部分自定义聚合查询的需求。

列名	类型	列类型
时间	VARCHAR(255)	KEY列
...
数据统计量	LARGEINT	SUM聚合列
FPS 统计值	BIGINT	SUM聚合列
FPS 最大值	DECIMAL(9,0)	MAX聚合列
低FPS统计值	BIGINT	SUM聚合列
内存占用最大值	LARGEINT	MAX聚合列



基于自研大模型问答和拖拽式生成查询

数据探索 / 智能问答

默认工作表 +

Doris - qa_sa

qa_sa

搜索集群名/数据库名/表名

AI分析数据表: 4

▼ Doris_prd

regression_doris_output_cosmos

timestamp DATETIME

index VARCHAR(655...

hostname VARCHAR...

path VARCHAR(655...

game VARCHAR(65...

device VARCHAR(65...

date DATE

excute VARCHAR(65...

uid BIGINT

reply INT

问题描述: 查找设备为 'xiaomi11' 的日期和数量

```
1 SELECT date, count(*) as count
2 FROM regression_doris_output_cosmos
3 WHERE device = 'xiaomi11'
4 GROUP BY date
5
```

查询结果

表格

图表

	date	count
23	2023-05-17	12
24	2023-05-02	19
25	2023-07-29	3

问题描述: 上面查询最晚的七天，按日期倒序

```
1 SELECT date, count(*) as count
2 FROM regression_doris_output_cosmos
3 WHERE device = 'xiaomi11'
4 GROUP BY date
5 ORDER BY date DESC
6 LIMIT 7
7
```

查询结果

表格

图表

保存到图表 (敬请期待...)

Date	Count
2023-09-01	~5
2023-08-30	~5
2023-08-28	~10
2023-08-26	~10

请输入问题开启问答查询，可参考右侧推荐问题。
cmd/ctrl + enter 发送。

查询 清空所有查询 自研模型

推荐问题

换一换

- 在 weather_seatunnel 表中查找城市为 '哈尔滨' 的日期和空调型号
- 在 regression_doris_output_cosmos 表中查找设备为 'oppo' 的日期和数

[illegible]

基于 Seatunnel 的数据集成

2 运行配置

</

[点击前往 Seatunnel 数据同步参数 设置](#)

3 调度设置

* 调度类型 ☐ 周期调度 ☒ 手动触发

* 作业超时设置 ☐ 是 ☒ 否

失败自动重试 ☐ 关

字段映射

字段匹配方式: ☒ 按名称匹配 ☐ 按顺序匹配

④ 字段映射结果如下。如果修改了数据源配置, 请点击 [重新获取映射](#)。

来源表字段	目标表字段
name	name
age	age
first_name	first_name
second_name	second_name
city	city
last_name	last_name
last_name	last_name
street_address	street_address
street_address	street_address
city	city
state_province	state_province
zip_postal_code	zip_postal_code
phone_number	phone_number
email	email
customer_id	customer_id
password	password
username	username

使用中遇到的问题

1.1.3 版本修改 VARCHAR 长度导致 tablet 出现 missing_rowsets, 表不可查

- 关联 Issue-13070, TransactionState 缺少了预提交状态

1.1.3 版本 Flink 写入会丢失最后一行数据

- 关联 Issue-13064, 分隔符指定为多个时每批次会丢失最后一行数据

1.1.5 版本不支持导入Hadoop EC 后的数据

- 升级 Broker 依赖的 Hadoop 版本至 3.x

1.1.5 版本简单查 TB 级大表分区, 做了分区过滤的情况下仍然出现 IO 打满的情况

- **原因:** 用户使用 Unique 模型进行查询的时候, 进行了两次聚合操作, 因此实际上没有用到分区过滤的特性。
- **解决方法:** 升级集群, 借助 1.2 版本的 Merge On Write 特性, 使得查询能够使用索引。

1.2.4 版本 JAVA UDF 不支持 Hive UDF 的重载

- 改造相关源码, 支持 Hive UDF 的重载

4 未来展望

未来规划

存算分离架构

- 引入更为廉价的存储介质以降低成本，数据湖场景下更灵活地弹性部署

倒排索引应用

- 目前有使用 ES 的用户对该功能比较感兴趣，官方介绍在存储及查询速度上均有大幅提升

数据湖分析

- 利用 Doris 的外表物化视图加速数据湖上的查询

Doris Manager建设

- 提供 Doris on K8S 小实例部署模式，降低用户在专属集群需求上的接入门槛
- Doris Manager 对于 2.0 版本新特性的管理支持，如跨集群数据同步等功能



获取更多社区动态与最佳实践

Apache Doris 官方平台:

- Apache Doris 官网: doris.apache.org
- Apache Doris GitHub: github.com/apache/doris/

获取更多峰会资料:

- Doris Summit 峰会官网: doris-summit.org.cn
- Doris Summit 峰会回放: <https://space.bilibili.com/1196172099/channel/collectiondetail?sid=1824324>