# 货拉拉大数据Doris稳定性保障实践

杨秋吉-货拉拉-OLAP负责人
梁健聪-货拉拉-大数据工程师

DataFunSummit # 2023

货拉拉 DataFun.

# 目录 CONTENT

# 01

# 背景与挑战

# 货拉拉介绍

360
国内城市

68万
月活司机

950万
月活用户

8+
业务线

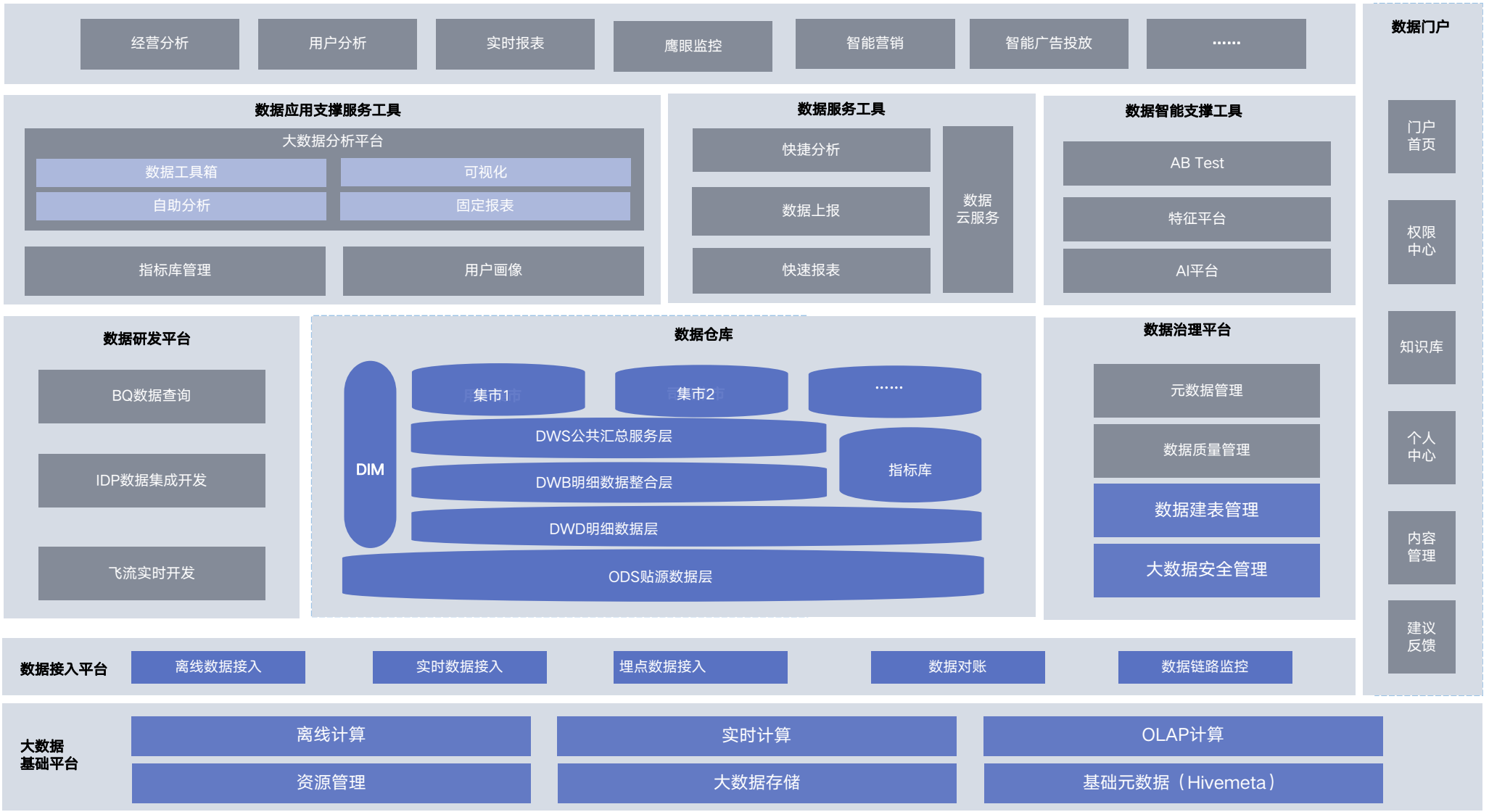
7+
IDC

1000+
机器数

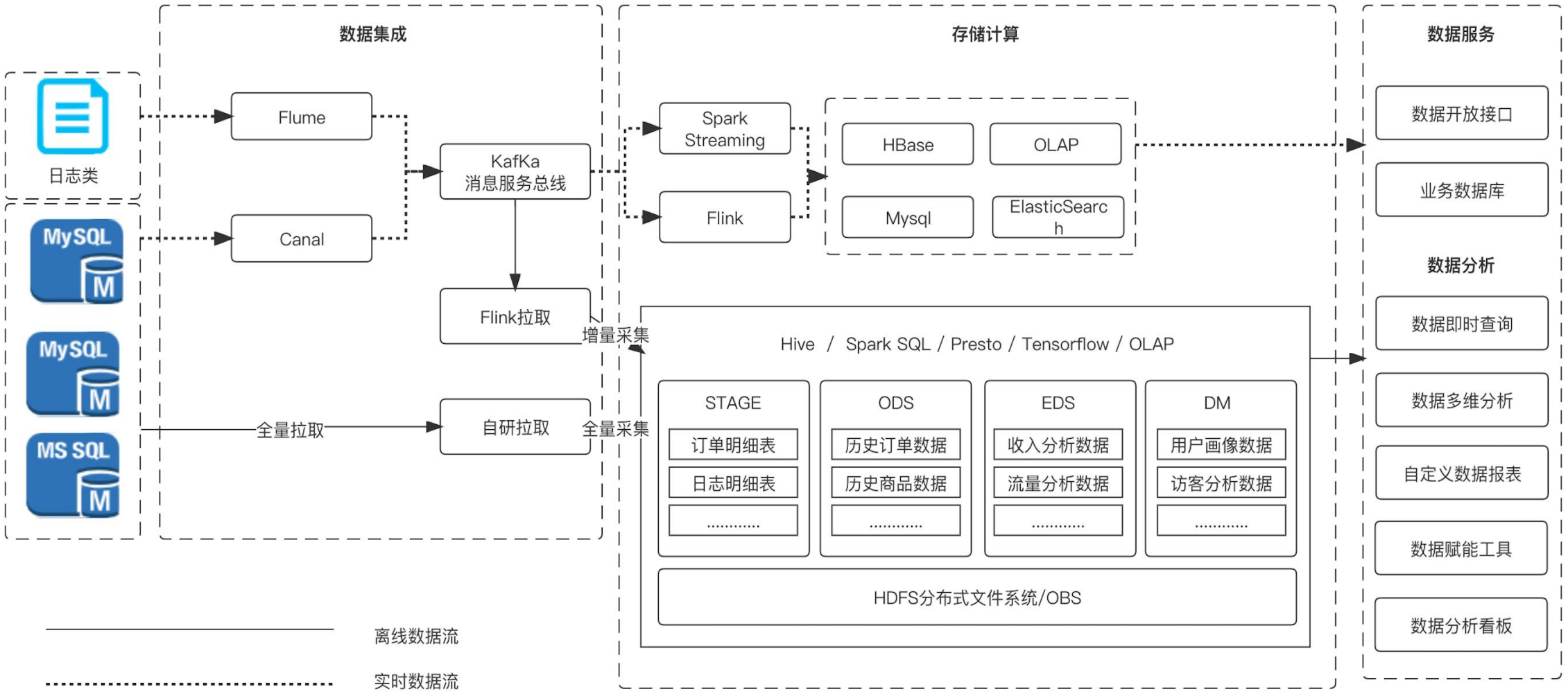
20PB+
存储量

20k+
日均任务数

# 货拉拉-大数据


经营分析
用户分析
实时报表
鹰眼监控
智能营销
智能广告投放
.....
数据门户
数据应用支撑服务工具
大数据分析平台
数据工具箱
可视化
自助分析
固定报表
指标库管理
用户画像
数据服务工具
快捷分析
数据上报
快速报表
数据云服务
数据智能支撑工具
AB Test
特征平台
AI平台
门户首页
权限中心
知识库
个人中心
内容管理
建议反馈
数据研发平台
BQ数据查询
IDP数据集成开发
飞流实时开发
数据仓库
DIM
集市1
集市2
.....
DWS公共汇总服务层
DWB明细数据整合层
DWD明细数据层
ODS贴源数据层
指标库
数据治理平台
元数据管理
数据质量管理
数据建表管理
大数据安全管理
数据接入平台
离线数据接入
实时数据接入
埋点数据接入
数据对账
数据链路监控
大数据基础平台
离线计算
实时计算
OLAP计算
资源管理
大数据存储
基础元数据（Hivemeta）

# 货拉拉-大数据


数据集成
存储计算
数据服务
日志类
MySQL
MySQL
MS SQL
Flume
Canal
KafKa
消息服务总线
Flink拉取
自研拉取
增量采集
全量采集
全量拉取
Spark
Streaming
Flink
HBase
OLAP
Mysql
ElasticSearch
Hive / Spark SQL / Presto / Tensorflow / OLAP
STAGE
订单明细表
日志明细表
.....
ODS
历史订单数据
历史商品数据
.....
EDS
收入分析数据
流量分析数据
.....
DM
用户画像数据
访客分析数据
.....
HDFS分布式文件系统/OBS
数据开放接口
业务数据库
数据分析
数据即时查询
数据多维分析
自定义数据报表
数据赋能工具
数据分析看板
离线数据流
实时数据流

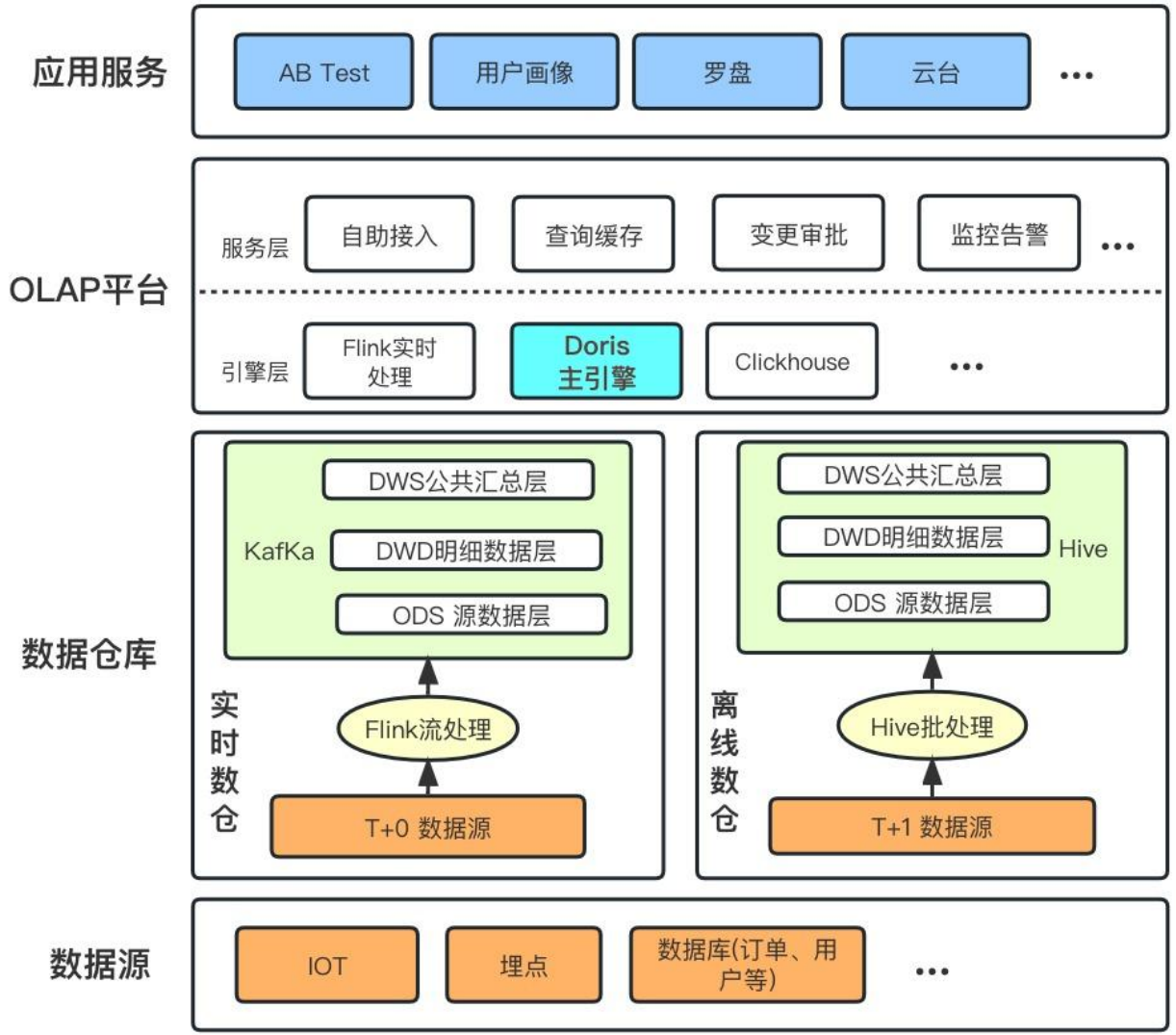
# Doris业务介绍

01 AB平台
关联海量埋点数据
灵活多维分析

02 用户画像
人群圈选

03 罗盘(增长分析决策平台)
漏斗分析、归因诊断

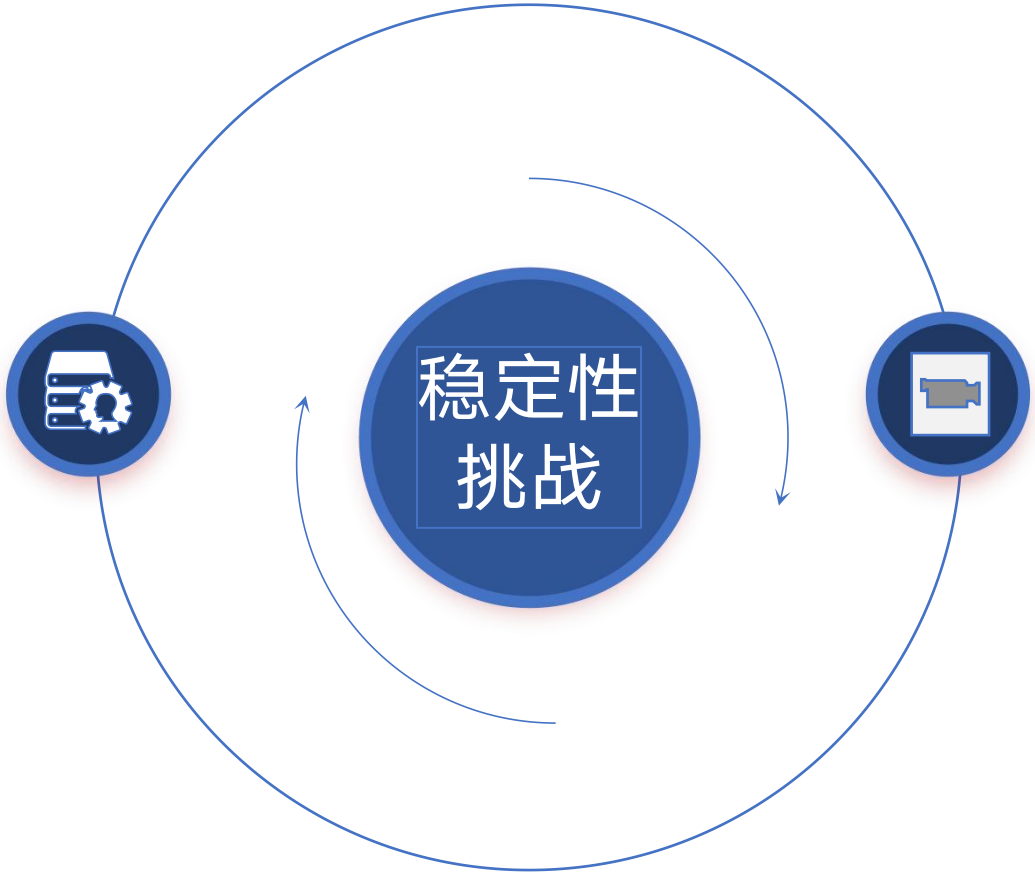
04 云台(数据可视化平台)
自助报表分析



# 稳定性挑战

## 业务对Doris服务稳定性要求高

1. Doris已接入多个核心业务已成为大数据核心基础组件




## 开源软件基本能力和生产需求之间的差距大

1. Doris内核能力完善，但外围平台能力不足，例如监控告警、运维管控
2. Doris内核演进速度快，相应的Issue也较多

| 版本数（2022-2023） | Issue数（2022 ~ 2023） |
| --- | --- |
| 14 | Open：1438<br>Closed：4112 |

# 稳定性保障目标

**少出事**

核心链路数据准确率：全年>=99.45%（2次/年）

**快发现**

核心链路问题（主动发现）时间 <= 5min

**快恢复**

P0核心链路恢复时间 <= 5min；P1级（埋点相关指标，容忍度相对高）链路恢复时间 <= 10min

# 02

# 稳定性能力保障

# 稳定性案例


**1、查询问题**

大查询容易打爆BE的内存

查询连接池满，导致查询报错

**2、导数问题**

数据堆积严重，影响到整个集群的读写吞吐

**3、数据质量问题**

Doris-1.1.3 Sparkload的unique模型

数据查询结果不准确

**4、版本升级问题**

1.1 升级至1.2存在OOM问题无法回滚

1.2 版本关闭向量化后数据对不上

**5、业务变更问题**

在问题发生后才发现新增字段变更

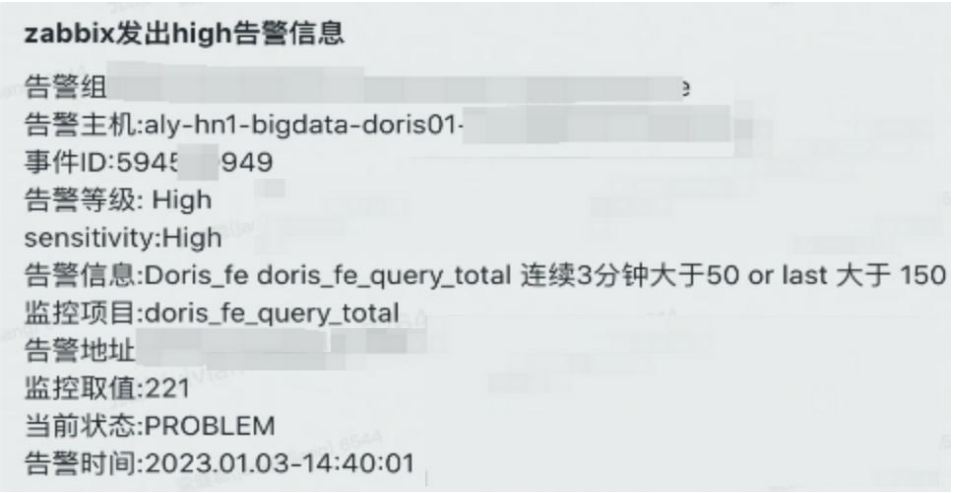
# 稳定性案例

**案例一**
**查询性能问题**

场景:云台查询Doris间歇性报错(Thread pool is at capacity)

原因: 用户提交大量查询以及一些大查询，导致fragment的rpc处理线程池满

解法办法:
1、加大查询缓存容量，增加缓存命中率（query_cache_max_size_mb）
2、查询超时由5min调整至3min
3、增强大查询拦截能力

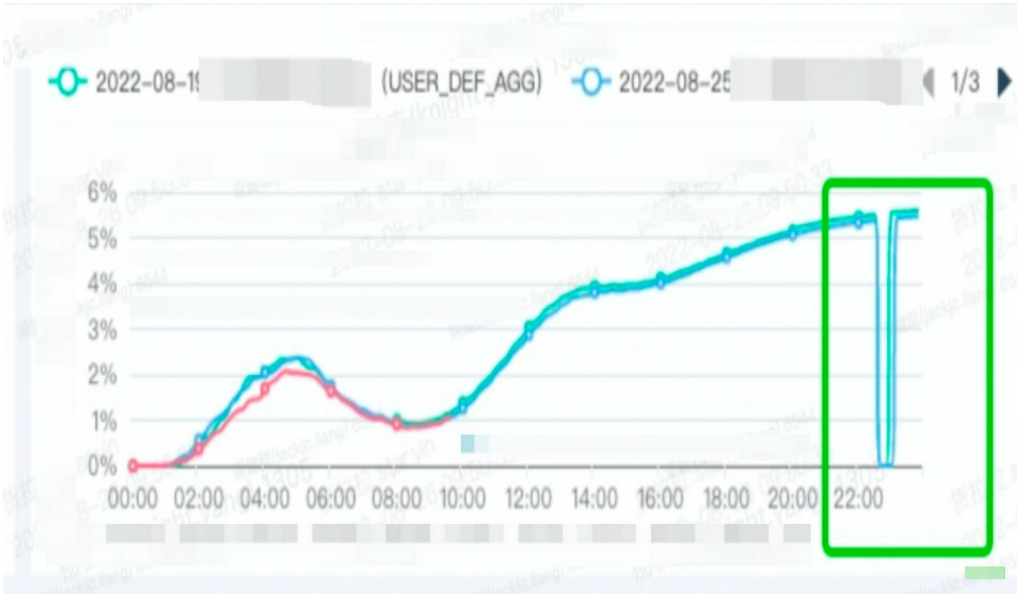
# 稳定性案例

案例二
导数性能问题

场景:准实时场景下5分钟调度任务因多个任务执行超时，导致报表数据更新延迟并跌0

原因: 新增的其他任务存在严重乱序，集群整体写入吞吐变慢，影响了准实时场景

解法办法:
1、Doris任务及导入参数优化 number_tablet_writer_threads (16 -> 32)
2、加强Doris变更规范管控与审批流程
3、业务多租户隔离(进行中)




# 稳定性案例

**案例三**
**数据质量问题**

场景: 业务使用sparkload导入Unique模型表，查询结果不稳定

原因: Unique模型表使用Sparkload导数时存在异常

解法办法:
1、将Unique模型改为Duplicate模型重建表
2、将Unique模型使用注意事项加入准入规范及最佳实践进行宣讲

```
--explain
select
    /*+ SET_VAR(enable_sql_cache = false)*/
     count(*)
   from
     hll_                    .obj_user_1d_in
   where
     test_id = '15676'
     and version_id = '16437'
     and shunt_create_time >= '2023-04-26'
     and shunt_create_time <= '2023-04-26'
--limit 100
```

Execution Time: 200 ms

| count(*) |
| --- |
| 1349418 |

第一次查询

```
--explain
select
    /*+ SET_VAR(enable_sql_cache = false)*/
     count(*)
   from
     hl                      ıser_1d_in
   where
     test_id = '15676'
     and version_id = '16437'
     and shunt_create_time >= '2023-04-26'
     and shunt_create_time <= '2023-04-26'
--limit 100
```

Execution Time: 242 ms

| count(*) |
| --- |
| 675243 |

第二次查询

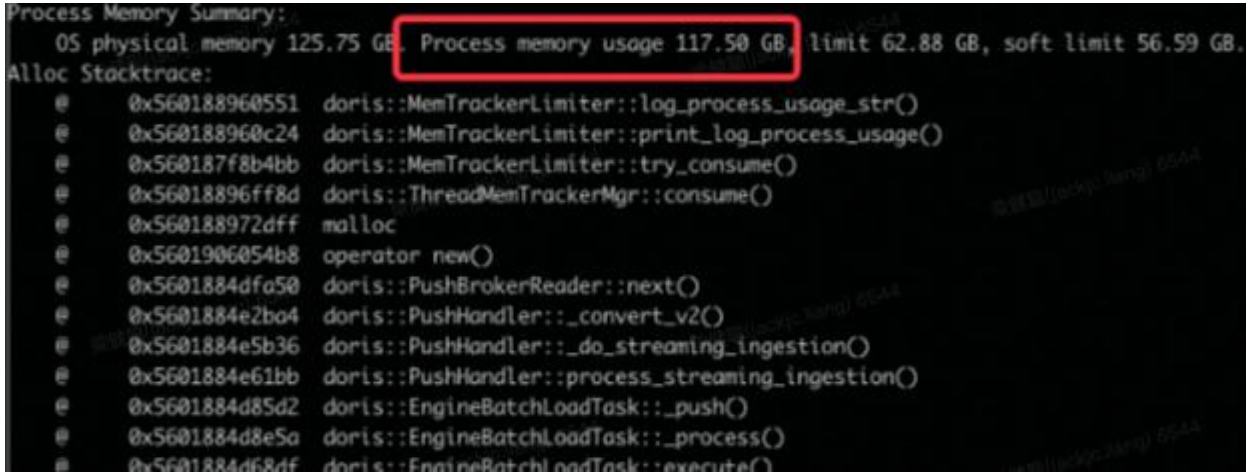
# 稳定性案例

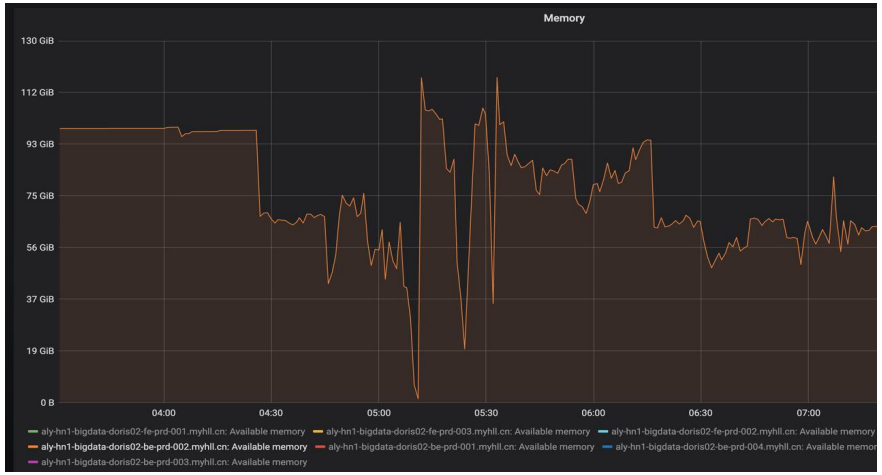
**案例四**
**版本升级问题**

场景:凌晨时间段 broker load任务和insert任务重合时间段，BE内存出现OOM被kill导致任务报错

原因: 升级1.2版本后的bitmap向量化读没有进行谓词下推，导致内存上涨

解法办法:
1、业务对SQL谓词下推的优化，如and和or的条件合并
2、后续集群HA方案（因1.2无法直接回退1.1）



# 稳定性案例

案例五
业务变更问题

场景: 业务侧自行对Doris表进行新增字段,表数据未更新且在无法查询

原因: 触发Doris版本1.0的bug，导致部分segment损坏，无法修复

解决办法:
1、沉淀通过Sparkload快速恢复数据预案
2、宣导用户使用规范、任务上线规范、发布变更规范

Results

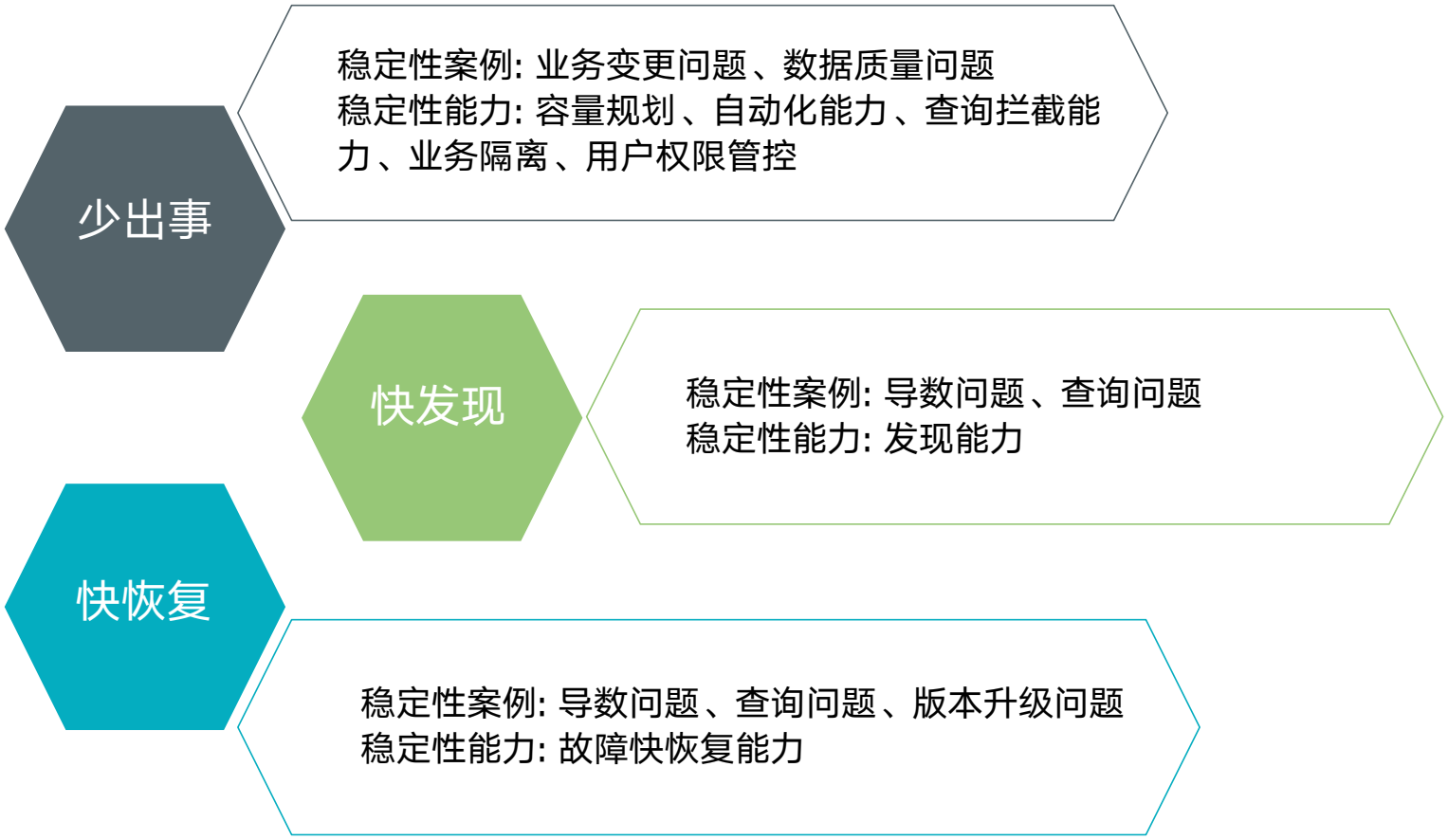
Run successfully

```
SHOW PROC '/dbs/13032/1060207/partitions/9136516/9299229/9299550';
```

Execution Time: 1 ms

| ReplicaId | BackendId | Version | VersionHash | LstSuccessVersion | LstSuccessVersionHash | LstFailedVersion | LstFailedVersio |
|---|---|---|---|---|---|---|---|
| 9299551 | | 350 | 1021618335750951880 | 350 | 1021618335750951880 | 351 | 0 |
| 9299552 | | 350 | 1021618335750951880 | 350 | 1021618335750951880 | 351 | 0 |
| 9299553 | | 350 | 1021618335750951880 | 350 | 1021618335750951880 | 351 | 0 |

# 建设思路


少出事
稳定性案例: 业务变更问题、数据质量问题
稳定性能力: 容量规划、自动化能力、查询拦截能力、业务隔离、用户权限管控
快发现
稳定性案例: 导数问题、查询问题
稳定性能力: 发现能力
快恢复
稳定性案例: 导数问题、查询问题、版本升级问题
稳定性能力: 故障快恢复能力

# 发现能力

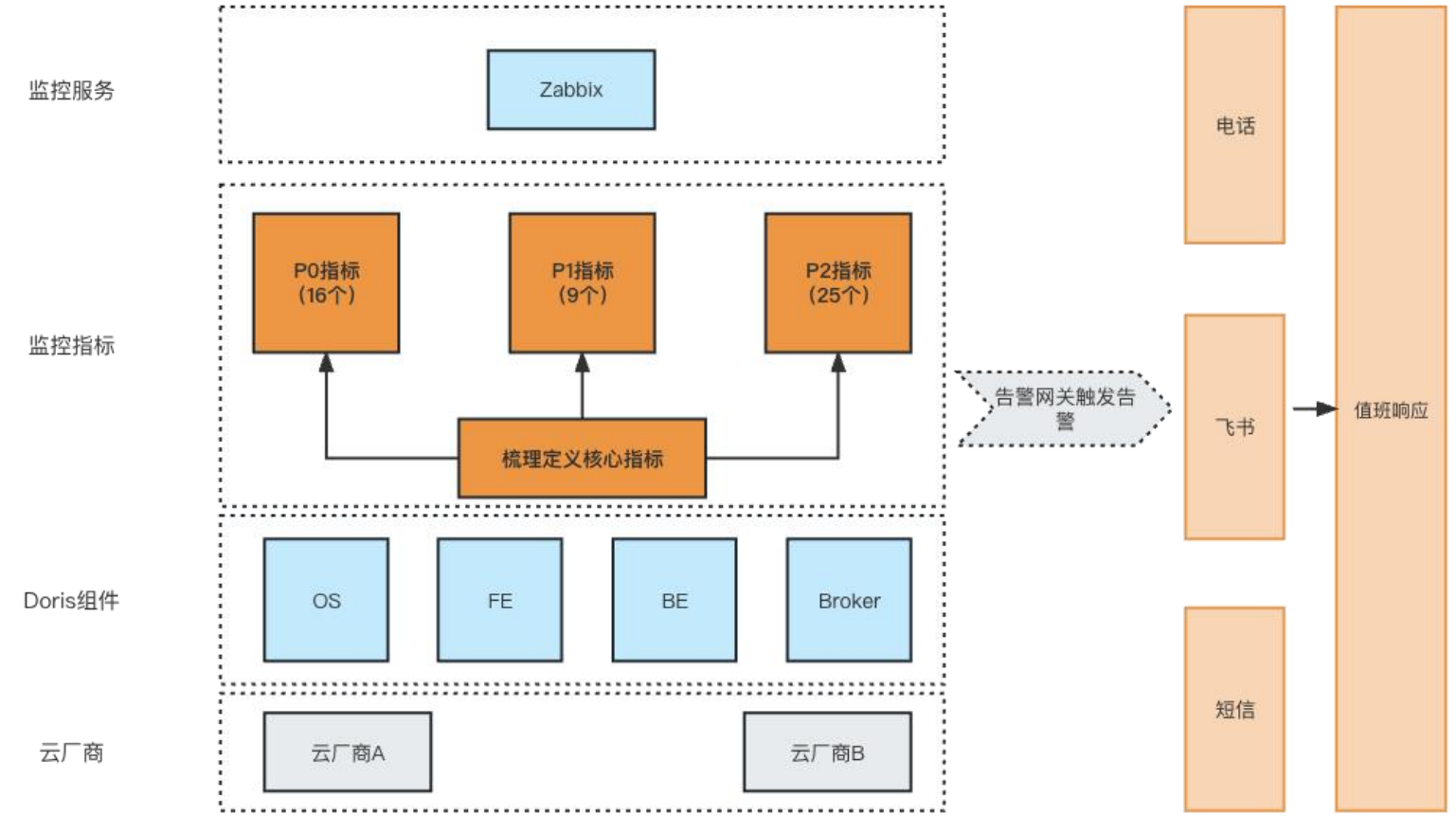
## 目标

快发现：核心链路问题（主动发现）
时间 <= 5min

## Doris监控告警系统

以Zabbix作为大数据基础架构组核心监控系统底座，对Doris服务进行监控和告警。

### 监控告警体系

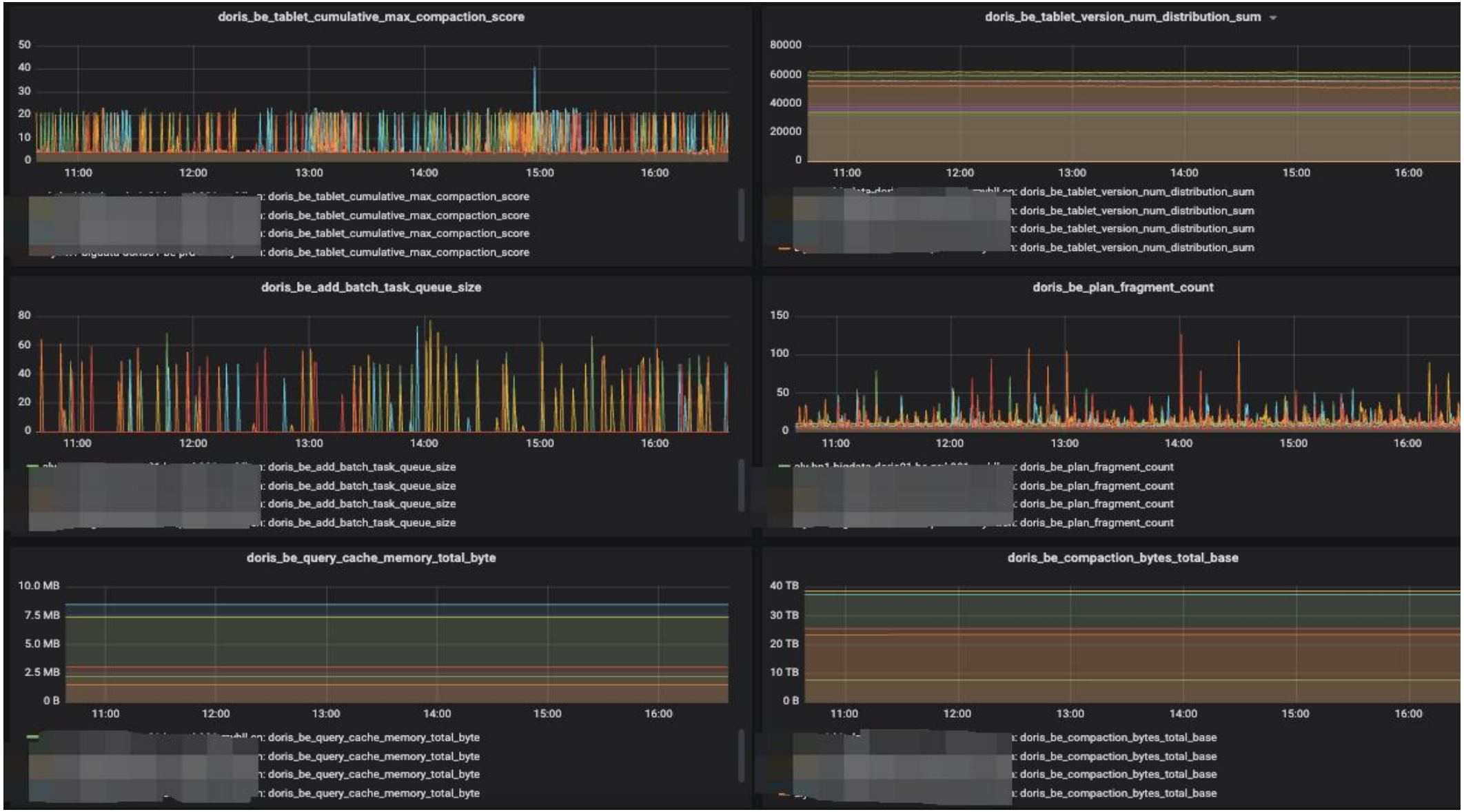


# 发现能力

1、表级监控：监控表容量、状态
2、任务监控：监控导数任务状态
3、组件监控：服务指标(查询、导数)、进程、机器指标

| 指标分级 | 作用 | 告警级别 | 指标项 |
| --- | --- | --- | --- |
| 一级指标<br>(服务不可用) | 发现和定位问题 | Disaster | 服务导数及查询(事务拒绝、事务失败、连接数、队列吞吐、RT、P99、P95)<br>服务内部状态(不健康tablet、进程、探活、JVM)<br>机器(内存、CPU、负载、磁盘、网络不可达)<br>任务状态(导数失败) |
| 二级指标 | 定位问题 | High | 服务查询(QPS、查询请求)<br>服务内部状态(JVM)<br>机器(磁盘IO、网络读写) |
| 三级指标 | 日常巡检、分析问题 | Warnning | 服务查询(表级监控) |

# 发现能力


doris_be_tablet_cumulative_max_compaction_score
50
40
30
20
10
0
11:00
12:00
13:00
14:00
15:00
16:00
doris_be_tablet_cumulative_max_compaction_score
doris_be_tablet_cumulative_max_compaction_score
doris_be_tablet_cumulative_max_compaction_score
doris_be_tablet_cumulative_max_compaction_score
doris_be_tablet_version_num_distribution_sum
80000
60000
40000
20000
0
11:00
12:00
13:00
14:00
15:00
16:00
doris_be_tablet_version_num_distribution_sum
doris_be_tablet_version_num_distribution_sum
doris_be_tablet_version_num_distribution_sum
doris_be_tablet_version_num_distribution_sum
doris_be_add_batch_task_queue_size
80
60
40
20
0
11:00
12:00
13:00
14:00
15:00
16:00
doris_be_add_batch_task_queue_size
doris_be_add_batch_task_queue_size
doris_be_add_batch_task_queue_size
doris_be_add_batch_task_queue_size
doris_be_plan_fragment_count
150
100
50
0
11:00
12:00
13:00
14:00
15:00
16:00
doris_be_plan_fragment_count
doris_be_plan_fragment_count
doris_be_plan_fragment_count
doris_be_plan_fragment_count
doris_be_query_cache_memory_total_byte
10.0 MB
7.5 MB
5.0 MB
2.5 MB
0 B
11:00
12:00
13:00
14:00
15:00
16:00
doris_be_query_cache_memory_total_byte
doris_be_query_cache_memory_total_byte
doris_be_query_cache_memory_total_byte
doris_be_query_cache_memory_total_byte
doris_be_compaction_bytes_total_base
40 TB
30 TB
20 TB
10 TB
0 B
11:00
12:00
13:00
14:00
15:00
16:00
doris_be_compaction_bytes_total_base
doris_be_compaction_bytes_total_base
doris_be_compaction_bytes_total_base
doris_be_compaction_bytes_total_base

# 容量规划



1

## 容量梳理

1. 业务需求
2. 数据量
3. 硬件资源
4. 集群规模

2

## 容量监控

1. 机器指标
2. 服务内部指标
3. 导数及查询指标
4. 表级监控指标

3

## 容量预警

1. 高危/严重告警事件
2. 关注业务需求
3. 高峰期、拉货节保障
4. 监控容量异动

# 容量规划-容量梳理

## 一、业务需求

1、了解业务当前需求，如实时或者离线、数据写入速度、时延、分区情况、查询要求、存储要求、是否有特殊feature支持，且需要业务完成大数据OLAP接入评审

## 二、数据量

1、BE独享节点-本地盘
业务总数据量 = 当前存量+未来增量的数据量
数据量预留比例 = 40%
数据副本 = 3
所需磁盘总大小 = (业务总数据量 / (1 - 数据量预留比例)) * 数据副本
2、FE独享节点-云盘
一般提供300GB

## 三、硬件资源

1、参考官网硬件数据
2、高性能HA模式，3台FE(云盘)+4台BE(本地SSD)
2、中性能HA模式，3台FE(云盘)+ 4台>=BE(云盘)
3、常规高可用配置
FE的配置1:4，如4C16G
BE的配置1:4，如32C128G

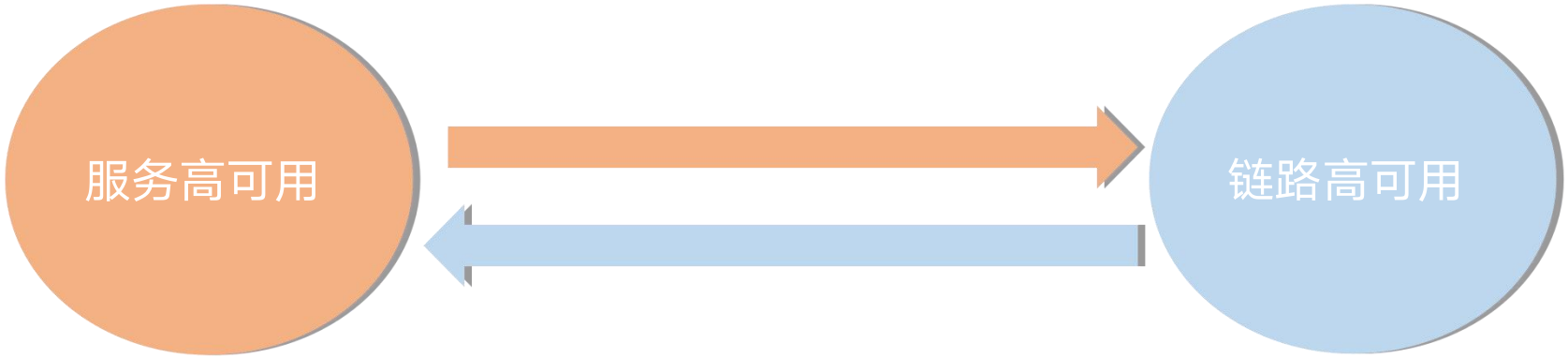
## 四、集群规模

1、参考业务需求、数据量情况来确认最终的集群规模

# 容量规划-容量监控

| 指标分级 | 指标名 | 告警阈值 |
| --- | --- | --- |
| 一级指标 | - 表级存储监控${table}_storage<br>- 服务级别:<br>doris_be_add_batch_task_queue_size<br>doris_fe_query_latency_ms<br>- 机器级别: 磁盘容量、CPU、内存 | cpu >= 80%<br>mem >= 85%<br>disk >= 90%<br>doris_be_add_batch_task_queue_size > 100<br>doris_fe_query_latency_ms >= 5s |
| 二级指标 | - 服务级别:<br>doris_fe_query_latency_ms_95<br>doris_be_plan_fragment_count<br>- 机器级别：磁盘IO | doris_be_plan_fragment_count >= 2200<br>... |

# 高可用能力




1. FE：三台FE高可用部署
2. BE：数据三副本，四台及以上BE，避免一台宕机导致数据不可写
3. LB：使用负载均衡绑定三台FE，实现连接数均衡及读写高可用

1. 离线/准实时导数链路：
   Spark load/Broker load/Select insert into任务，通过离线调度任务平台进行调度，支持异常自动重试或者电话告警
2. 实时导数链路：
   Flink类型任务，通过自研实时任务平台进行调度，支持异常自动重试或者电话告警

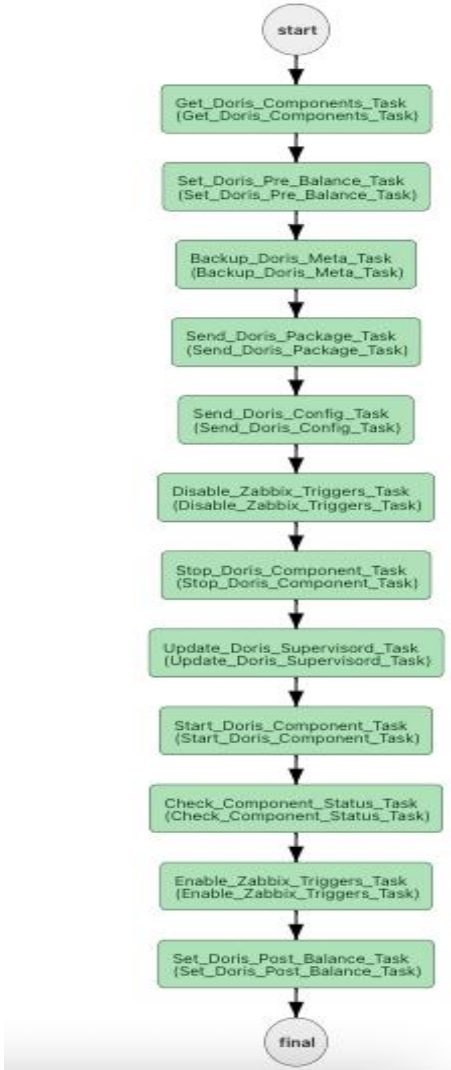
# 自动化能力

背景：初期在构建大数据Doris集群时，我们以标准SOP指引下通过脚本手动操作为主，人为误操作或遗漏的可能，稳定性相对较差。

运维SOP → 脚本化改造 → 运维脚本化 → 自动化封装 → 运维自动化

进展：通过大数据自动化平台构建Doris自动化能力，底座基于Netflix Conductor、Ansible开发，已集成Doris部署、Doris扩容、Doris升级等工作流编排能力。

收益：
1、提升Doris组件服务稳定性
2、提升运维人效




Doris升级工作流

# 其他保障能力


## 一、查询拦截能力

1、设置用户级别拦截规则，根据实际数据量级、查询规模制定拦截规则
2、可快速对异常query进行手动kill，防止对集群整体产生更大影响

## 二、故障快恢复能力

1、分区数据的快速恢复能力
2、tablet状态恢复能力

## 三、业务隔离

1、根据业务重要程度、数据类型属于实时或离线进行集群隔离、多租户。

## 四、用户权限管控

1、通过使用Doris自带的RBAC(Role-Based Access Control)能力，对用户/角色赋予相关权限

# 03

# 稳定性流程规范

# 稳定性流程规范

一、Doris业务准入规范

二、Doris使用规范

三、Doris业务变更规范

# Doris业务准入规范

## 需求评估

1、快速理解业务需求，判断Doris是否最适合业务场景

## 准入评审

1、参加大数据部门的需求准入评审

2、评估业务价值、投入产出比ROI

| | |
|---|---|
| CheckList<br>（重要，填写后请打勾☑） | • **稳定性要求**<br>☐ 可用性 >= xx.xx%（无特殊需求可不填写）<br>• **导数**<br>◦ 实时（如无需实时导数，则无需填写）<br>☐ 数据写入速度<= x 行/分钟，约 y Byte/行<br>☐ 延时：毫秒/秒/分钟级?<br>◦ 离线（如无需离线导数，则无需填写）<br>☐ 导数方式：sparkload/brokerload/streamload<br>☐ 每天导入分区数：一天一个分区/一天n个分区<br>☐ 每天是否需覆盖旧分区数据：是/否<br>• **存储**<br>☐ 日存储数据量（如为hive表，可根据元初表详情存储量估算）：>= x G/天<br>☐ 数据保留天数：1个月/2个月/3个月/半年（若保留3个月及以上，请说明理由）<br>• **查询**<br>☐ 查询QPS <= xx<br>☐ 查询性能要求 P95 <= xx Sec<br>☐ 单次查询结果集最大行数 <= xx 行 （若大于10w行，请说明理由）<br>☐ 单次查询**最大**扫描分区数：<30个 / < 60个/ <90个/ >90个 （若大于60个，请说明理由）<br>• **特殊feature支持**<br>☐ 是否需要精准去重，非精准去重误差1%左右是否接受：是/否，接受/不接受<br>☐ 数据导入（实时导入）是否需要精准一次，是否接受偶发的数据重复：是/否，接受/不接受<br>☐ 查询是否涉及join：是/否<br>☐ 表字段是否涉及复杂结构？（Map/JSON/Struct等）：是/否<br>☐ 是否需要unique模型：是/否 |
| 查询Pattern<br>（重要） | 请尽量全面的列举常用的查询语法（若暂时不确定查询pattern，可用文字描述查询场景）：<br>查询SQL Pattern 1：select xxx from<br>查询SQL Pattern 2：select xxx from |

# Doris使用规范

| 类型 | 关注1 | 关注2 | 关注3 |
| --- | --- | --- | --- |
| 建表 | 1.分桶数建议值16或32，单个tablet 约1G<br>反例:分桶设置太小，导致单个tablet达30G，执行compact很慢，集群吞吐变差 | 1.表模型(优先使用Aggregate/Duplicate)<br>2.前缀索引(根据查询条件设置)<br>3.分区字段 (设置合理的生命周期) | 1. 高频写入表建议放在单独的数据库，避免事务数过多影响到同库的其他表 |
| flink写入 | 1.切勿使用自己的jar，要使用flinksql<br>反例：使用自己的jar,一条数据一个写入事务，导致集群吞吐变慢 | 1.数据无乱序(保证只有单个分区的数据)<br>反例: 数据存在大量乱序，任务每次写入都涉及几十个分区，导致集群吞吐变慢 | 1. 单个batch的发送数据量建议 Batch size >= 100MB或者timeout = 2~ 5min，单表批次导入实例并发 <= 2 |
| Insert写入 | 1. insert into执行成功后，还需等待版本发布后，数据才可见<br>**反例:执行成功后立即查询，查询结果为空** | N/A | N/A |
| 删除 | 1.避免频发删除，导致Compaction压力大，影响写入和查询性能<br>反例:业务高峰期频繁执行delete语句，集群base compact频繁，吞吐变慢 | 1.Delete执行，数据可见性是异步 | 1.Drop语句后不用加force，这样出现误删除可在一段时间内找回 |
| 修改 | 1.一张表同时间只能执行一个alter<br>反例: 执行修改列类型,会对数据做遍历，耗时久，这段时间再次执行则会报错 | 1.支持增删列、修改列，不支持修改列名 | 1.新增/删除数据前停写入任务，且等待5分钟后再执行，防止数据不一致 |
| 查询 | 1.严禁不带过滤条件查询全量数据，且建议业务加上limit兜底<br>反例: 查询全表数据，打爆集群CPU和内存 | 1.Doris不适用高并发QPS场景 | 1.Doris兼容Mysql协议，业务查询时可带traceId，方便排查问题 |

# Doris变更规范

## 一、发布窗口

1、业务低峰期，非节假日前一天
2、离线12-16点，实时20-24点
3、非变更窗口需走紧急变更流程

## 二、发布内容、发布通知

1、发布背景、执行操作需描述清楚
2、通知业务方、执行方、次日Oncall

## 三、审核

1、方向负责人、组负责人审核
2、遵循Doris使用规范
3、不变更就必然产生稳定性风险或无法故障恢复情况下可提前变更，事后补充

## 四、验收

1、服务稳定性验收
2、服务功能性验收
3、异常快速回滚

# 04

# 总结与规划

# 总结

**一、保障目标**

1、数据准确性/可靠性

2、业务链路稳定性

**二、案例分析**

1、数据查询

2、导数性能

3、数据质量

4、版本升级

5、业务变更


**三、保障能力**

1、发现能力

2、容量规划

3、高可用能力

4、自动化能力

5、拦截、隔离、恢复、权限管控能力

**四、流程规范**

1、Doris业务准入规范

2、Doris使用规范

3、Doris变更规范

# 思考

1. 稳定性的建设是持续的、成体系化的，而**非靠运气**

2. 稳定性的目标实现需要业务方支持，而**非靠单点突破**

# 规划

01

## 稳定

多集群HA、多租户隔离、冷热存储

02

## 易用

OLAP能力平台化，提升易用能力

03

## 高效

紧跟Doris社区，尝试更多的应用场景：高并发点查/文本搜索代替ES/联邦查询


感谢观看