星云零售信贷基于Doris的OLAP演进之路

演讲人：刘建波—腾梭科技—研发总监

腾梭科技 TECHSIGHT | DataFun.

目录

产品演进

1、互联网网贷业务

2、助贷+联合贷导流

3、纯自营业务

I)、业务架构

网贷 -> 联合贷 -> 新零售

II)、技术架构

单体应用-> 分布式应用 -> 微服务

OLAP选型困扰

1、生态复杂且门槛高、技术劝退

2、投入产出比极低、成本劝退

1、诉求：需要的不是大数据平台，而是一套高性能的OLAP工具

2、原则：数据规模、灵活性、成本可控

演进之路

第一阶段：基于kettle 离线ETL

插件丰富、链路长、难管理

第二阶段：基于trino的统一查询

异构数据源、联邦查询、大内存

第三阶段：基于doris的存储分析

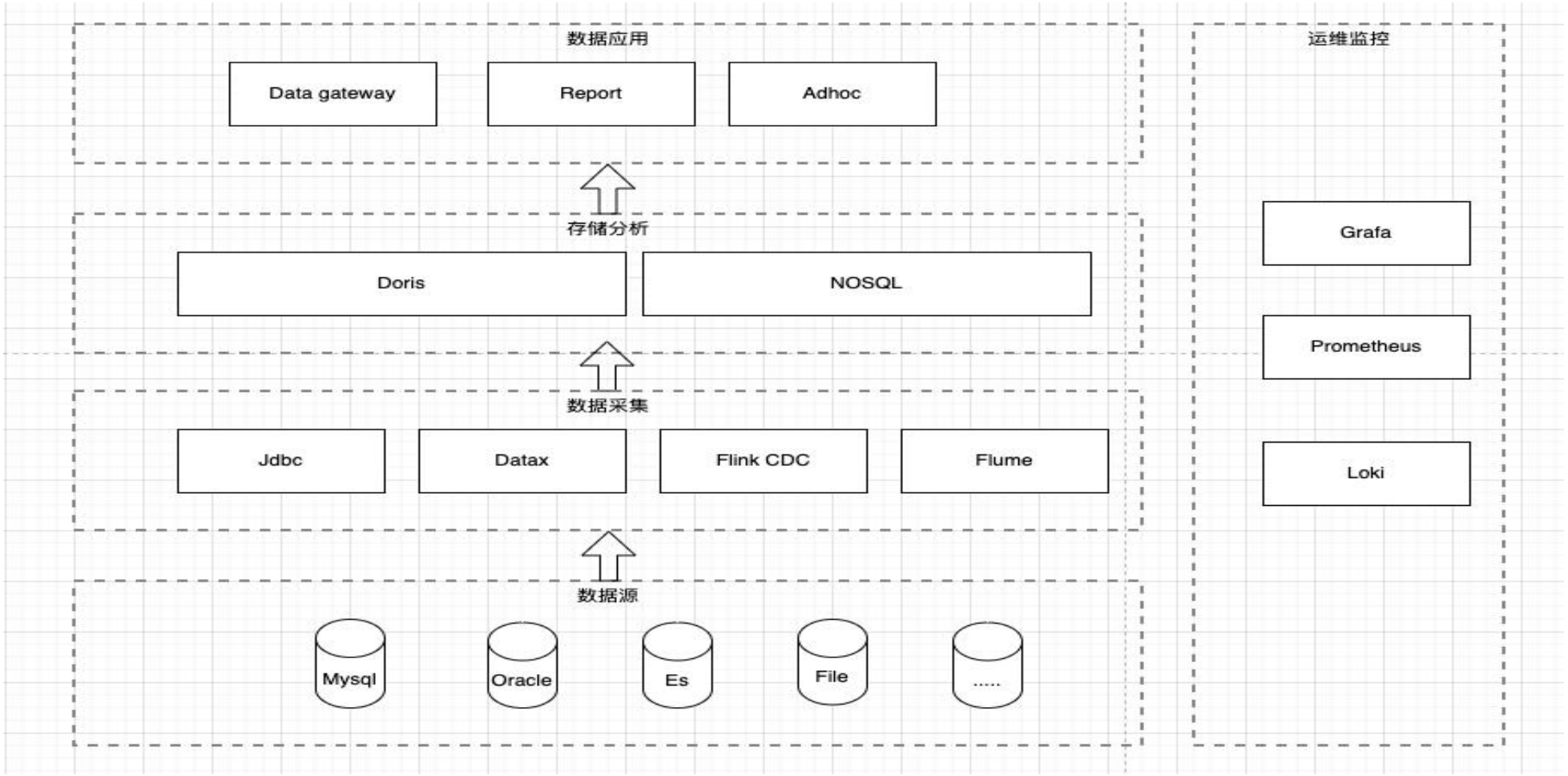
简单易用、极速查询、实时统一

Apache Doris实践

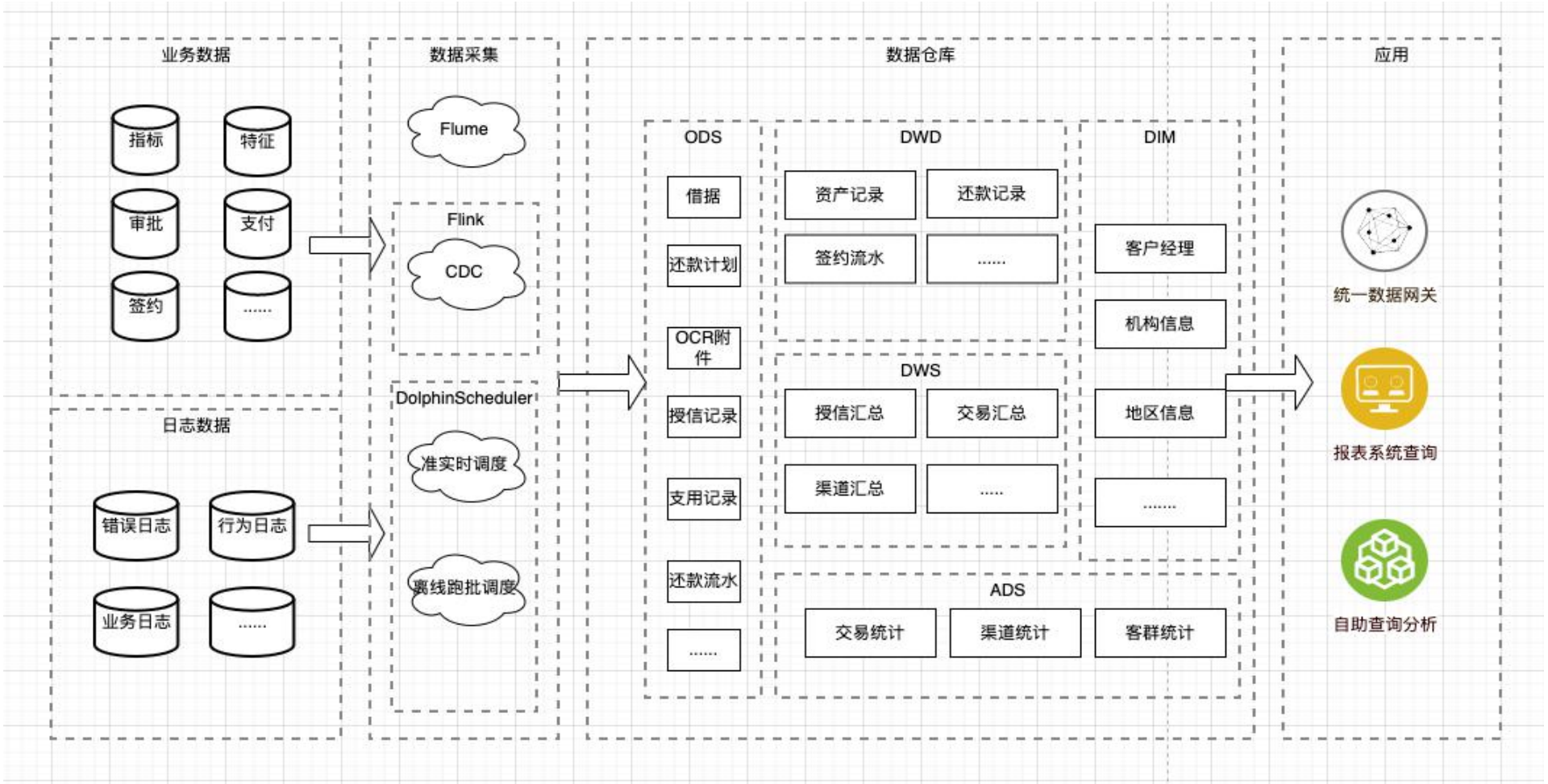
1、并发查询加速

2、数仓底座建设

运行架构


数据应用
Data gateway
Report
Adhoc
存储分析
Doris
NOSQL
数据采集
Jdbc
Datax
Flink CDC
Flume
数据源
Mysql
Oracle
Es
File
.....
运维监控
Grafa
Prometheus
Loki

业务模型


业务数据
指标
特征
审批
支付
签约
.....
日志数据
错误日志
行为日志
业务日志
.....
数据采集
Flume
Flink
CDC
DolphinScheduler
准实时调度
离线跑批调度
数据仓库
ODS
借据
还款计划
OCR附件
授信记录
支用记录
还款流水
.....
DWD
资产记录
还款记录
签约流水
.....
DWS
授信汇总
交易汇总
渠道汇总
.....
ADS
交易统计
渠道统计
客群统计
DIM
客户经理
机构信息
地区信息
.....
应用
统一数据网关
报表系统查询
自助查询分析

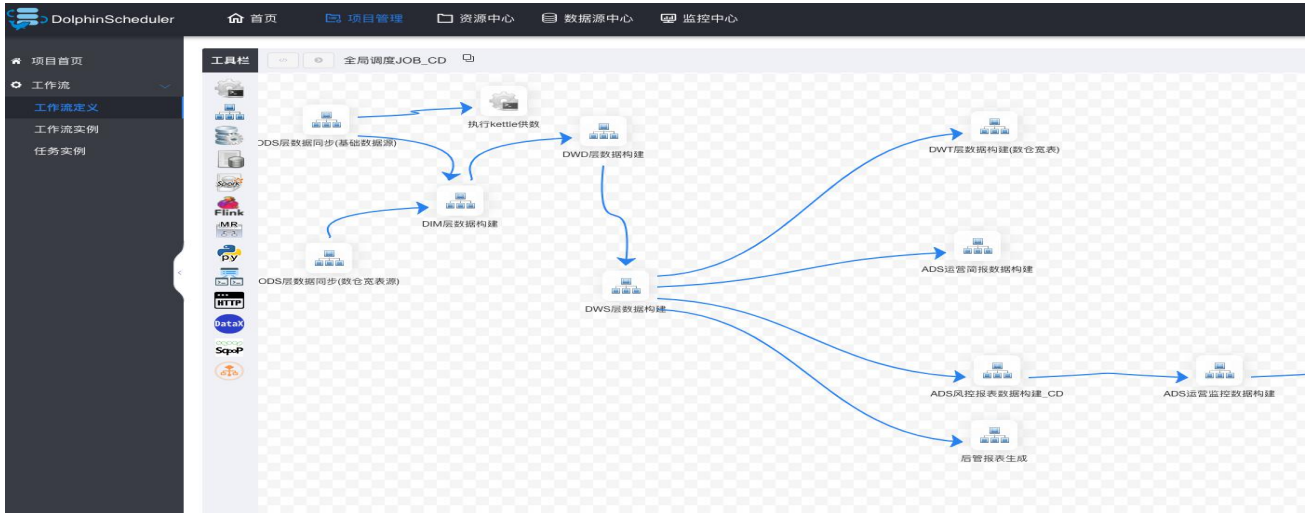
业务场景及技术方案

1、风控大数据报表平台

2、统一日志存储与分析

3、用户行为日志存储分析

任务编排


DolphinScheduler
首页
项目管理
资源中心
数据源中心
监控中心
项目首页
工作流
工作流定义
工作流实例
任务实例
工具栏
全局调度JOB_CD
执行kettle供数
ODS层数据同步(基础数据源)
DWD层数据构建
DWT层数据构建(数仓宽表)
DIM层数据构建
ODS层数据同步(数仓宽表源)
ADS运营简报数据构建
DWS层数据构建
ADS风控报表数据构建_CD
ADS运营监控数据构建
后管报表生成


当前节点设置
节点名称
ods_db_biz_asset
运行标志
正常
禁止执行
描述
请输入描述
任务优先级
MEDIUM
Worker分组
default
失败重试次数
3
(次)
失败重试间隔
3
(分)
超时告警
脚本
1 bash sql_job_shell/doris_cli.sh ods_sql_job/ods_db_biz_asset.sql
资源
sql_job_shell/doris_cli.sh
ods_sql_job/ods_db_biz_asset.sql
自定义参数
取消
确认添加

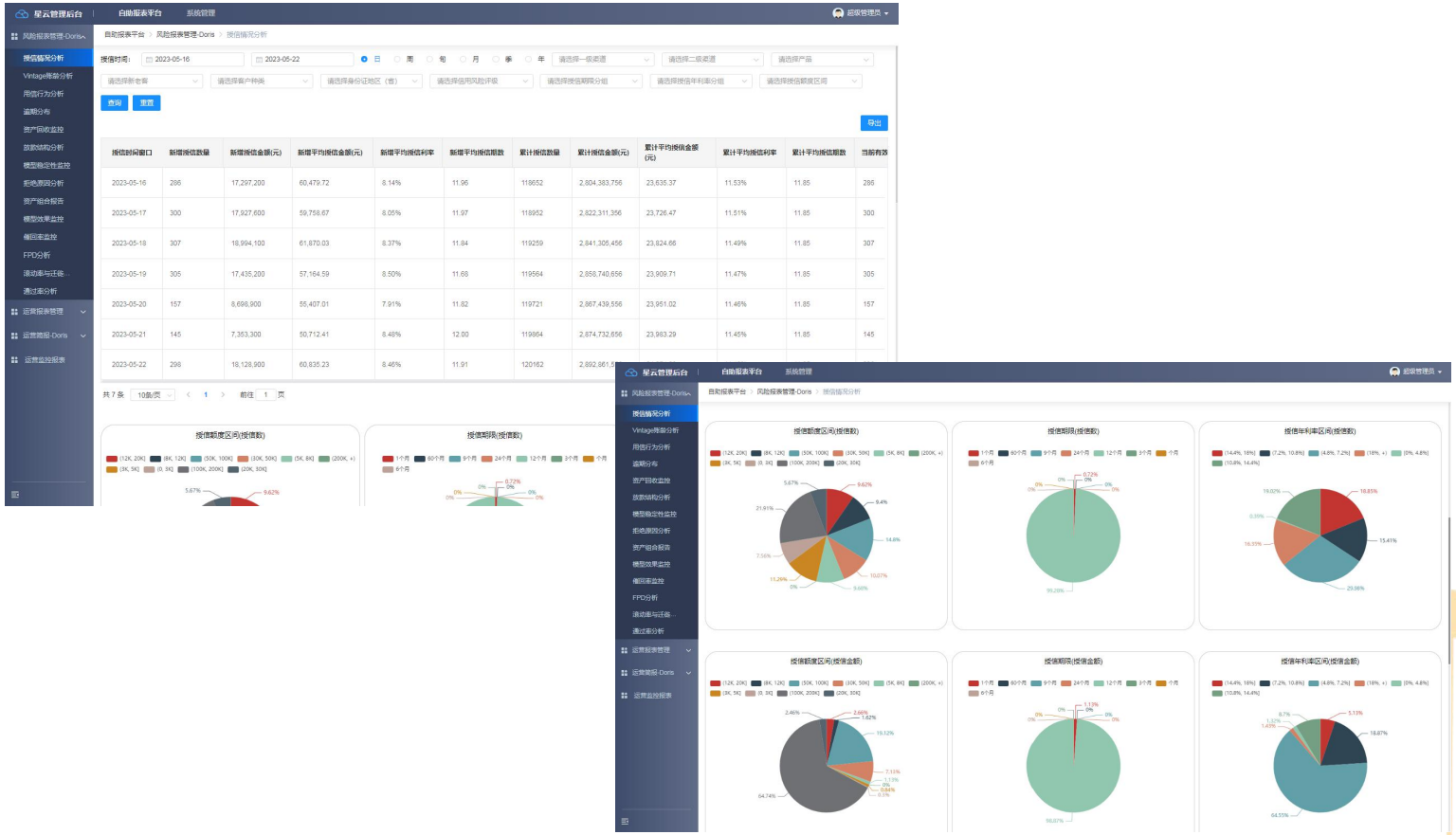

DolphinScheduler
首页
项目管理
资源中心
数据源中心
监控中心
项目首页
工作流
工作流定义
工作流实例
任务实例
工具栏
ODS层数据实时同步(基础数据源)-分钟级
当前节点设置
节点名称
ods_db_risk_t_apply_risk_sql
运行标志
正常
禁止执行
描述
请输入描述
任务优先级
MEDIUM
Worker分组
default
失败重试次数
0
(次)
失败重试间隔
1
(分)
超时告警
超时策略
超时告警
超时失败
超时时长
30
分
数据源
MYSQL
doris_net
sql类型
非查询
sql语句
1 -- 设置完全无法容忍错误
2 SET enable_insert_strict=true;
3 -- 设置查询超时时间为30分钟
4 SET query_timeout=1800;
5
6 INSERT INTO ods.ods_db_risk_t_apply_risk
7 SELECT apply_risk_id
8 ,apply_risk_apply_id
9 ,apply_risk_apply_code
10 ,apply_risk_source
11 ,apply_risk_product_id
12 ,apply_risk_biz_product_id
13 ,apply_risk_individual_id
14 ,apply_risk_type
取消
确认添加

XX银行—风控大数据报表平台

1、T+1离线跑批ETL

2、分钟级准实时ETL

3、Flink CDC实时ETL



XX银行—统一日志存储与分析

1、自定义Flume Sink采集

2、未使用倒排索引

3、基于大文本模糊匹配

```
CREATE TABLE `error_log` (
  `event_id` varchar(32) NULL,
  `log_date` date NULL,
  `app_name`  varchar(10) NULL,
  `create_time` datetime NULL,
  `log_level` varchar(10) NULL,
  `class` varchar(200) NULL,
  `method` varchar(100) NULL,
  `msg` varchar(4000) NULL,
  `exception` text NULL
) ENGINE = OLAP
UNIQUE KEY(`event_id`, `log_date`) COMMENT 'OLAP'
PARTITION BY RANGE(`log_date`) ()
DISTRIBUTED BY HASH(`event_id`, `log_date`) BUCKETS AUTO
PROPERTIES (
  "replication_allocation" = "tag.location.default: 1",
  "dynamic_partition.enable" = "true",
  "dynamic_partition.time_unit" = "DAY",
  "dynamic_partition.time_zone" = "Asia/Shanghai",
  "dynamic_partition.start" = "-7",
  "dynamic_partition.end" = "3",
  "dynamic_partition.prefix" = "p",
  "dynamic_partition.replication_allocation" = "tag.location.default: 1",
  "dynamic_partition.buckets" = "10",
  "dynamic_partition.create_history_partition" = "false",
  "dynamic_partition.history_partition_num" = "-1",
  "dynamic_partition.hot_partition_num" = "0",
  "dynamic_partition.reserved_history_periods" = "NULL",
  "dynamic_partition.storage_policy" = "",
  "in_memory" = "false",
  "storage_format" = "V2",
  "disable_auto_compaction" = "false" );
```

Run successfully

```
select log_date,count(1) from error_log group by log_date order by log_date;
```

Execution Time: 2850 ms

log_date	count(1)
2022-12-28	115591396
2022-12-29	76440446
2022-12-30	73843088
2023-01-03	85245888
2023-01-04	46749024

Run successfully

```
select * from error_log  where exception like '%153%' order by create_time limit 10;
```

Execution Time: 1227 ms

XX银行—用户行为日志存储与分析

1、自定义Flume Sink采集

2、基于JSONB存储与分析

3、存储极致压缩、降本增效

```
CREATE TABLE `api_log` (
  `request_time` datetime NOT NULL,
  `request_id` varchar(32) NOT NULL,
  `caller` varchar(16) NOT NULL,
  `client_ip` varchar(46) NOT NULL,
  `header` jsonb NOT NULL,
  `request_method` varchar(16) NOT NULL,
  `url` varchar(256) NOT NULL,
  `request_params` jsonb NOT NULL,
  `response_code` varchar(8) NOT NULL,
  `response_data` jsonb NULL,
  `request_microtime` double NOT NULL,
  `spend_microtime` double NOT NULL,
  `log_date` date NOT NULL
) ENGINE = OLAP
DUPLICATE KEY(`request_time`, `request_id`, `caller`) COMMENT 'OLAP'
PARTITION BY RANGE(`log_date`) ()
DISTRIBUTED BY HASH(`request_time`, `request_id`, `caller`) BUCKETS AUTO
PROPERTIES (
  "replication_allocation" = "tag.location.default: 1",
  "bloom_filter_columns" = "request_time, request_id, url",
  "dynamic_partition.enable" = "true",
  "dynamic_partition.time_unit" = "DAY",
  "dynamic_partition.time_zone" = "Asia/Shanghai",
  "dynamic_partition.start" = "-7",
  "dynamic_partition.end" = "3",
  "dynamic_partition.prefix" = "p",
  "dynamic_partition.replication_allocation" = "tag.location.default: 1",
  "dynamic_partition.buckets" = "10",
  "dynamic_partition.create_history_partition" = "false",
  "dynamic_partition.history_partition_num" = "-1",
  "dynamic_partition.hot_partition_num" = "0",
  "dynamic_partition.reserved_history_periods" = "NULL",
  "dynamic_partition.storage_policy" = "",
  "in_memory" = "false",
  "storage_format" = "V2",
  "disable_auto_compaction" = "false" );
```

```
select log_date,count(1) from api_log group by log_date order by log_date;
```

Execution Time: 855 ms

log_date	count(1)
2022-12-30	35098180
2022-12-31	6363808
2023-01-01	6428208
2023-01-03	69945387
2023-01-04	10402381
2023-01-05	10304243
2023-01-06	7160616

Run successfully

```
select  count(1) from api_log where log_date ='2022-12-30' and get_json_int(response_data,'$.code')=101018 group by log_date order by log_date;
```

Execution Time: 37637 ms

count(1)
1110981

架构收益

取数导数：业务人员或者开发人员，可以进行自定义的导数和纬度数据分析

运维成本：集群节点做进程保活、几乎自运维

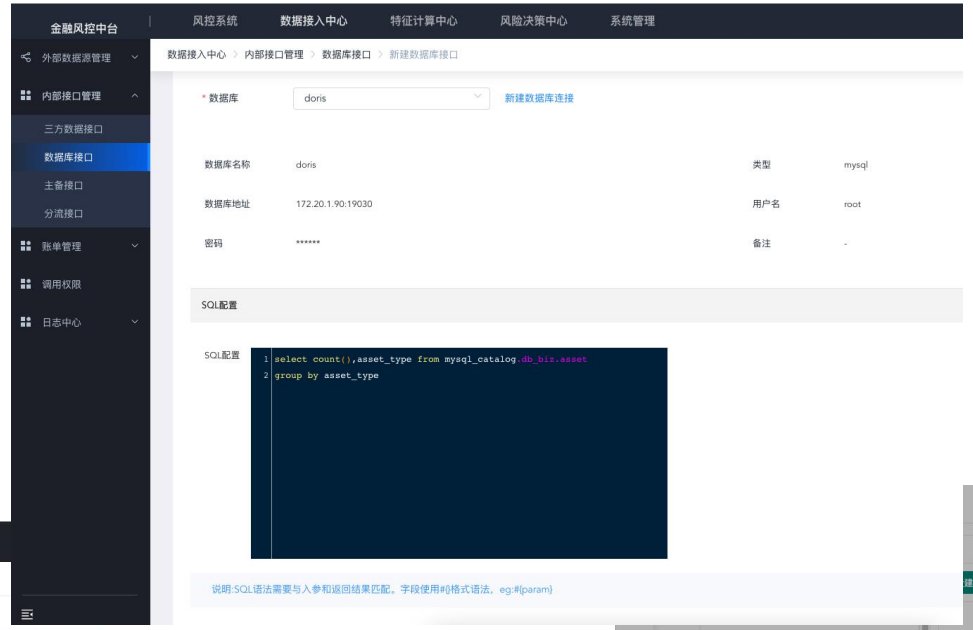
查询延迟：简单查询毫秒级，复杂查询秒级响应

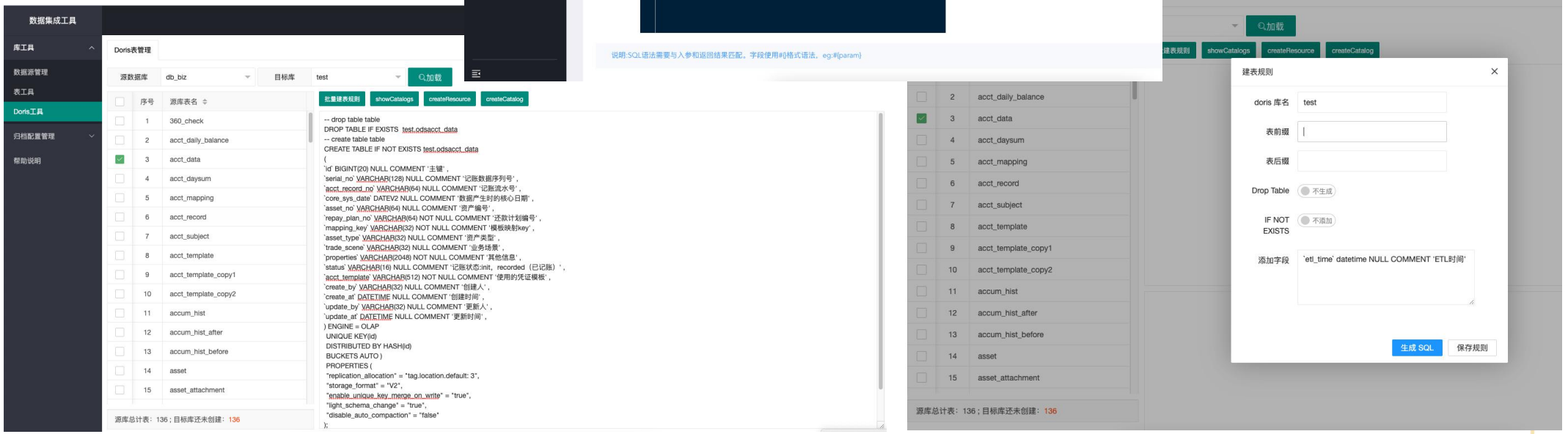
资源节省：Doris极致压缩，将数据压缩了70%，存储成本降低

后期规划

1、智能数据网关

2、统一数据归档





感谢观看