

Apache Doris × 阿里云联合 Meetup

🕒 10月26日 (周六) 13:30-17:15



正泰集团数据中台基于 Apache Doris 的应用实践

许继良

正泰集团技术中心大数据架构师



目录

01 背景介绍

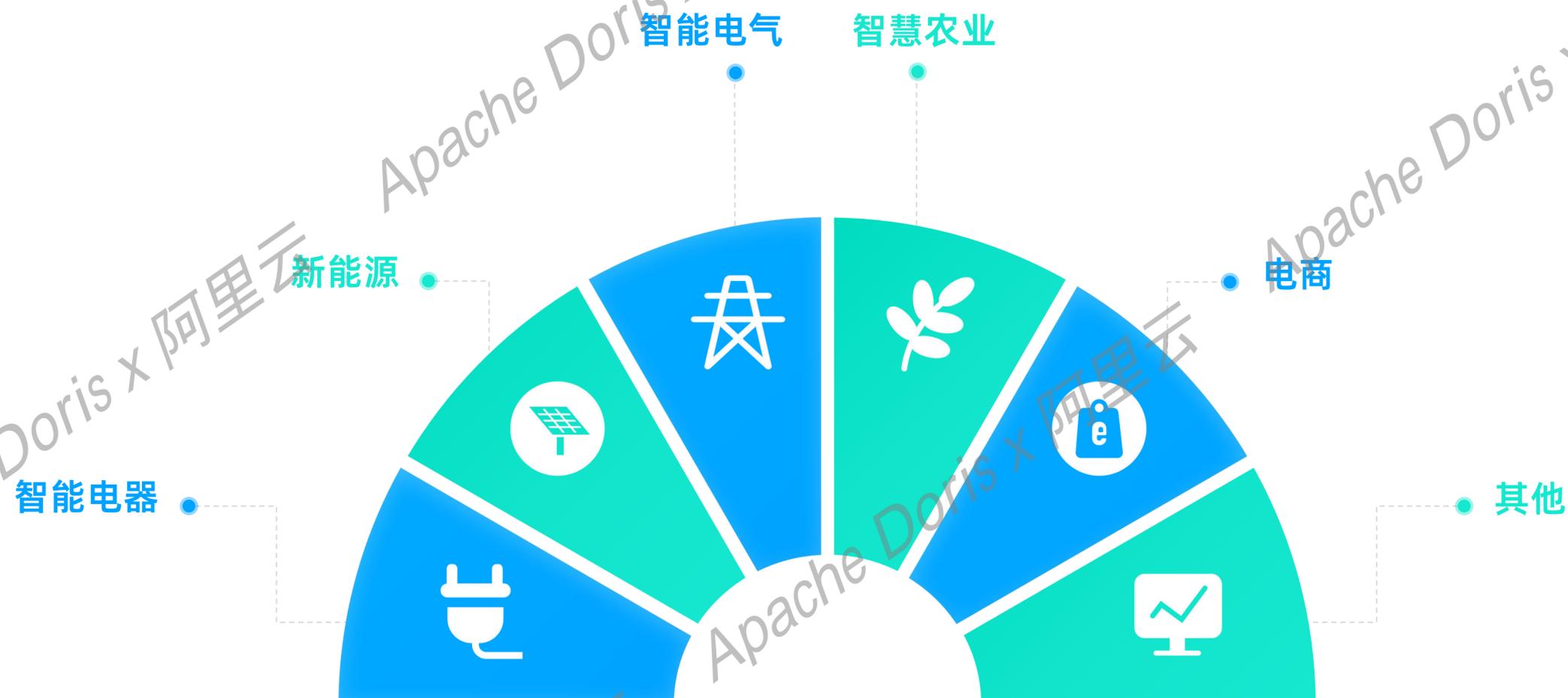
02 数仓架构演进

03 数据中台基于 Doris 的应用

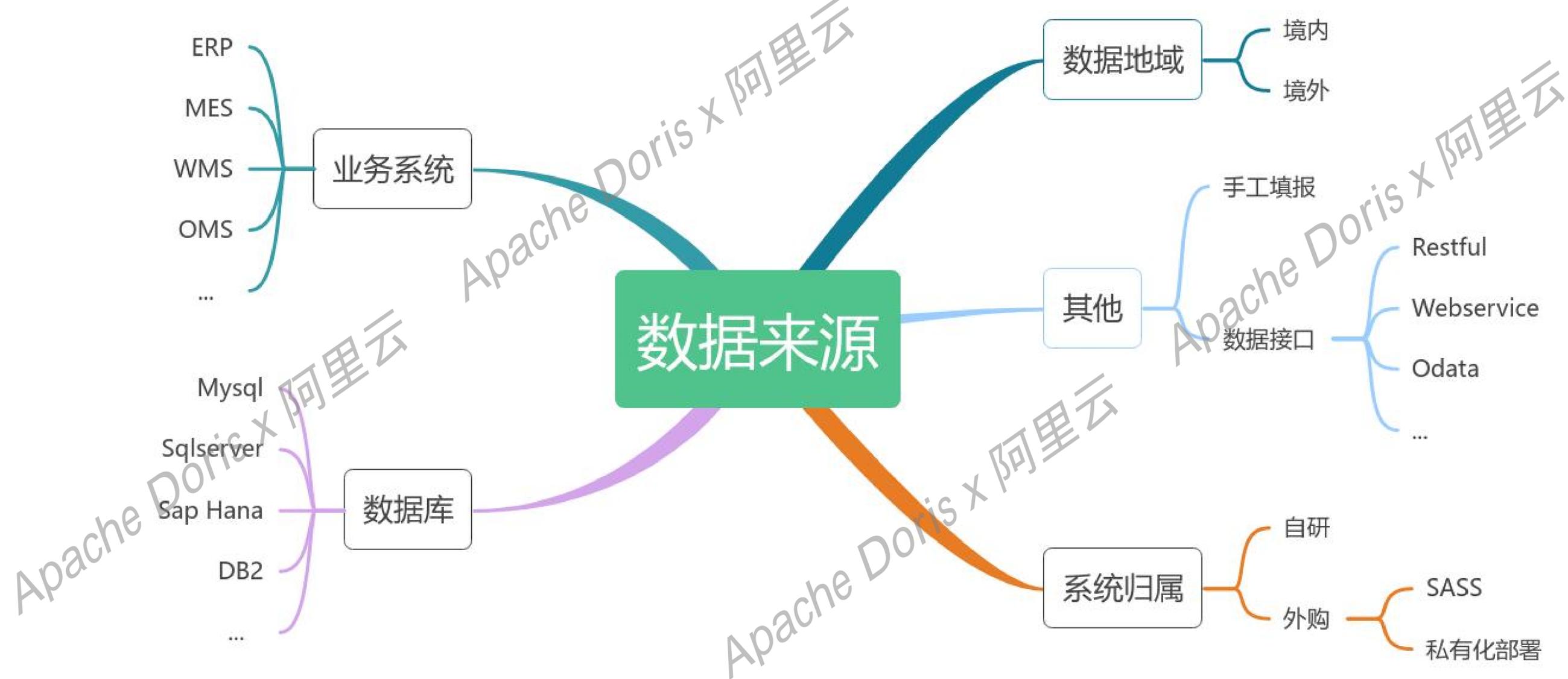
04 未来规划

集团业态分布

正泰集团股份有限公司始创于 1984 年，是全球知名的智慧能源系统解决方案提供商，业务遍及 140 多个国家和地区，全球员工 5 万余名，2023 年集团营业收入 1550 亿元，连续二十余年上榜中国企业 500 强。



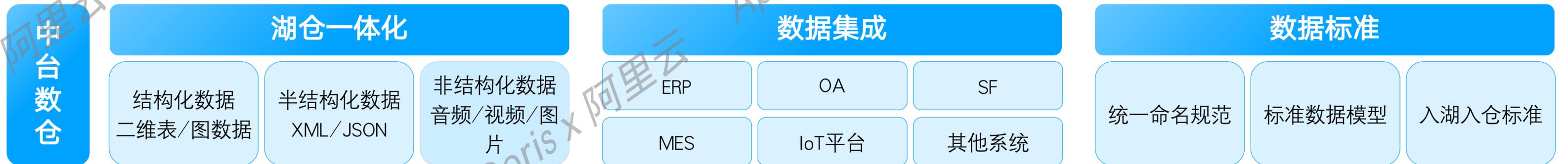
数据源分布



数据痛点与需求



数据中台能力框架



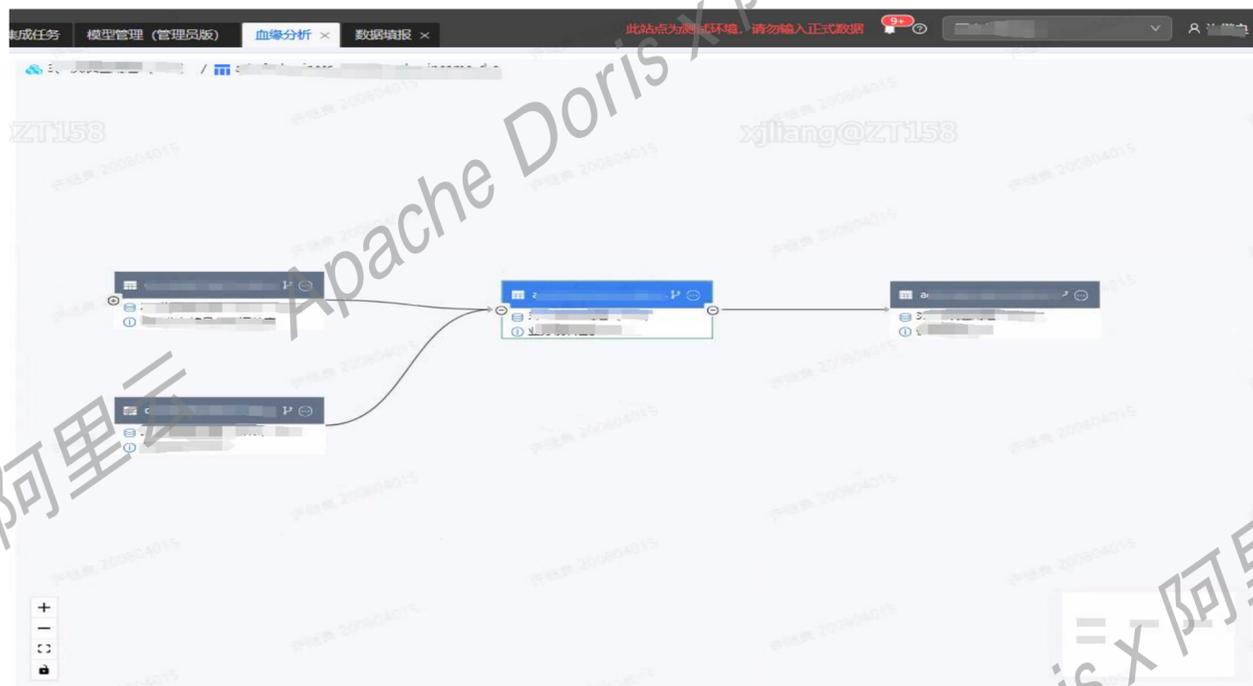
数据中台功能



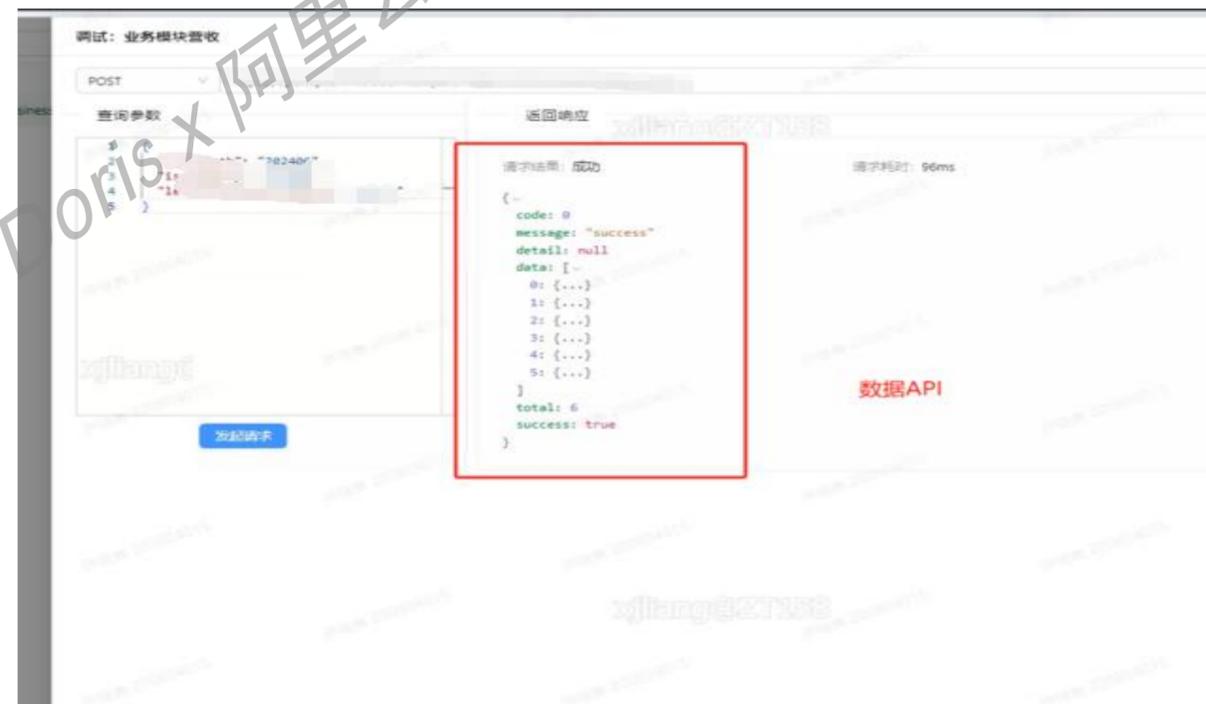
首页



集成作业



数据血缘



API服务

目录

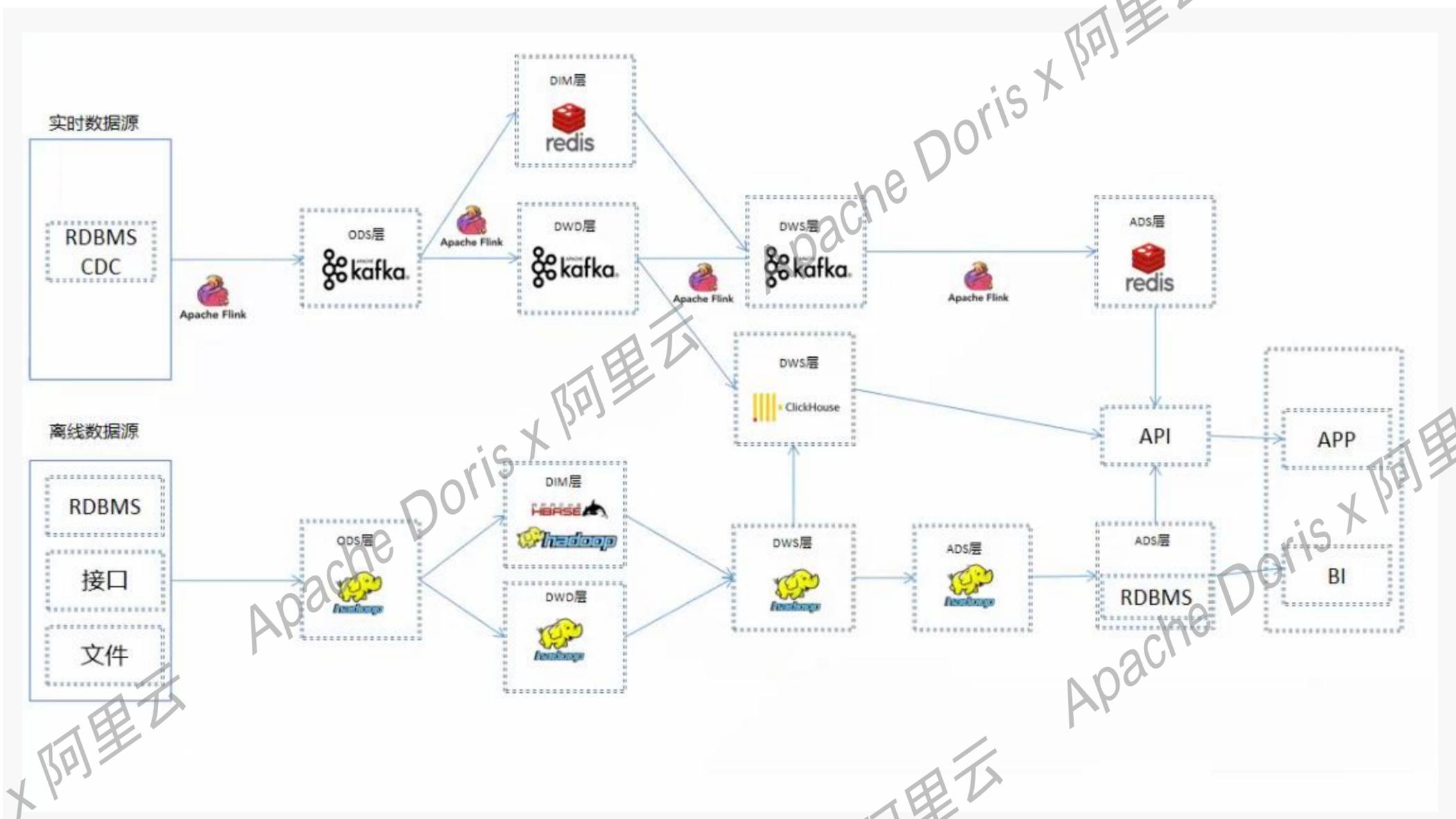
01 背景介绍

02 数仓架构演进

03 数据中台基于Doris的应用

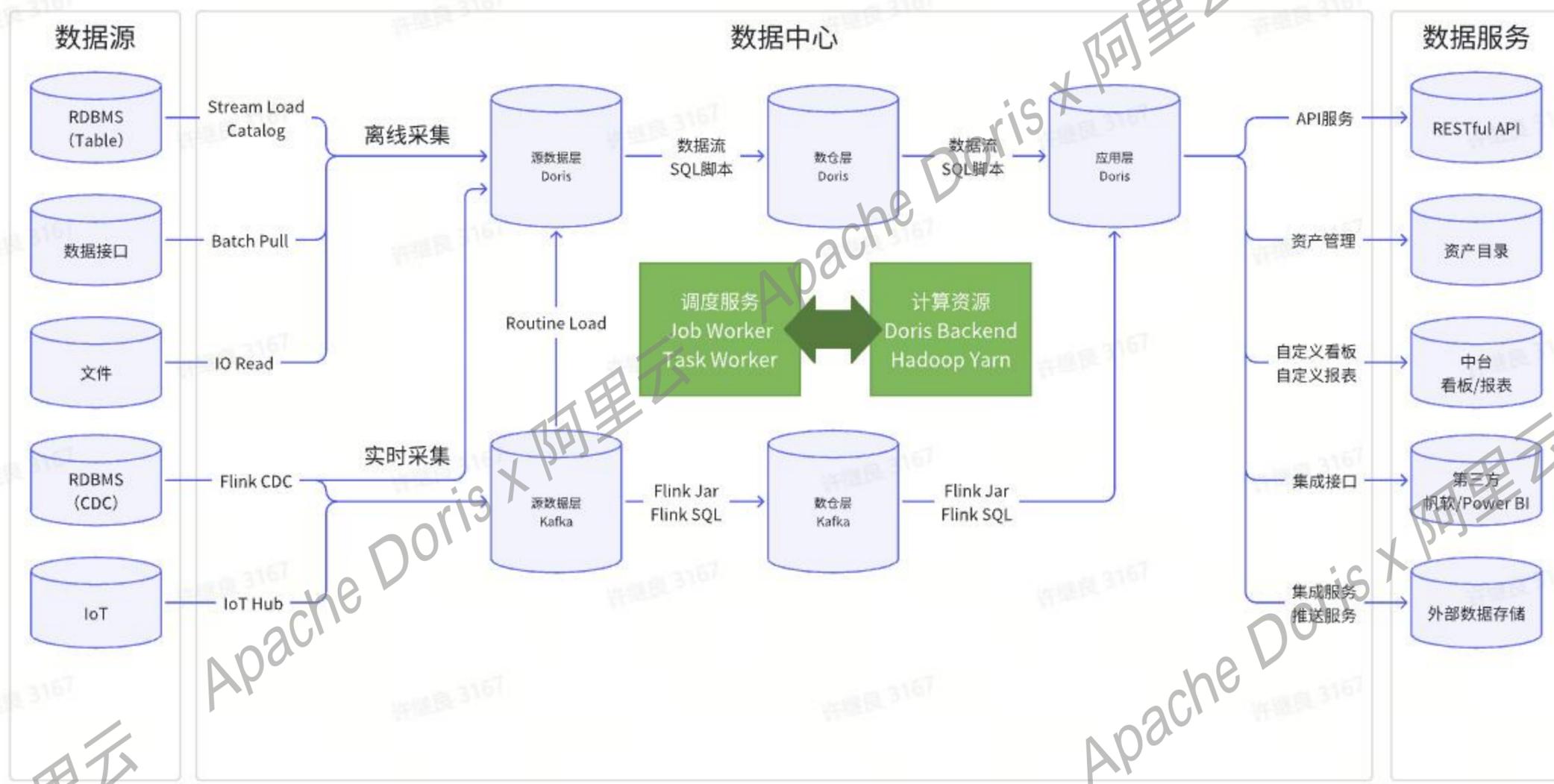
4 未来规划

数据仓库架构V1.0



- 基于 Hadoop 生态构建
- 架构复杂，运维成本高
- 数据链路长，容错率低
- 数据权限控制不便

数据仓库架构V2.0



- 基于 Doris 构建
- 轻量级架构
- 实现流批一体
- 数据权限控制方便快捷
- 对比 Imapla, 响应时间提升 50%+

目录

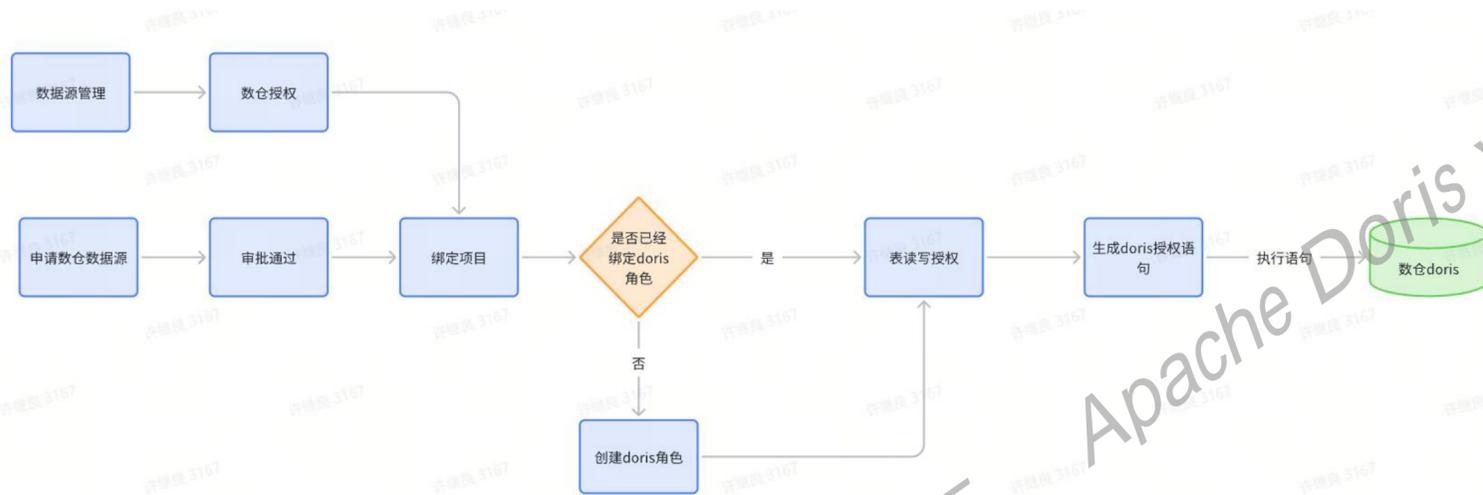
背景介绍

数仓架构演进

数据中台基于Doris的应用

未来规划

角色绑定



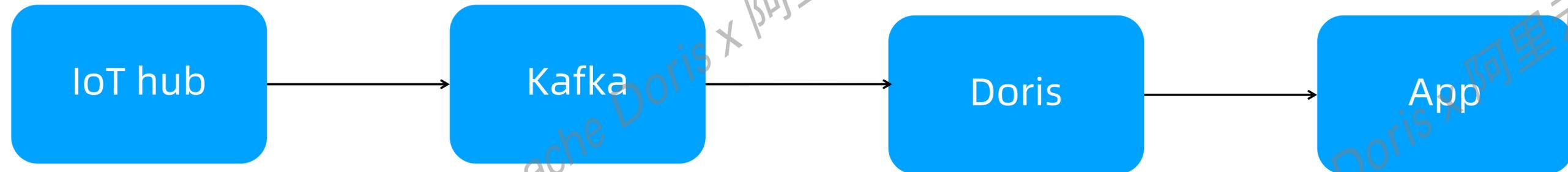
通过 Doris 2.0 数据角色与中台项目角色绑定，实现了 Doris 表级别控制(包含 catalog)，使数据权限更方便更安全。

The screenshot shows the '角色管理' (Role Management) interface. It includes a search bar, a '同步角色&账号(中台数仓)' (Sync Roles & Accounts (Data Warehouse)) button, and a '新增' (Add) button. The main table lists roles with columns for '名称' (Name), '备注' (Remarks), '成员数量' (Member Count), '内置角色' (Built-in Role), and '操作' (Operations).

名称	备注	成员数量	内置角色	操作
成员		9	✓	编辑 成员维护 菜单权限 数仓权限 删除
管理员		6	✓	编辑 成员维护 菜单权限 数仓权限 删除
业务		0	✓	编辑 成员维护 菜单权限 数仓权限 删除
制造		1	✗	编辑 成员维护 菜单权限 数仓权限 删除

The screenshot shows the '编辑：飞书数据' (Edit: Feishu Data) configuration form. It includes fields for '数据源类型' (Data Source Type), '数据源名称' (Data Source Name), '归属部门' (Department), '业务系统' (Business System), '管理员' (Admin), '数据owner' (Data Owner), '入仓状态' (In-warehouse Status), '元数据收集周期' (Metadata Collection Cycle), '描述' (Description), '连接配置' (Connection Configuration), '服务器名称/IP' (Server Name/IP), '端口号' (Port Number), '数据库' (Database), '用户名' (Username), '密码' (Password), '驱动版本' (Driver Version), '是否创建catalog' (Whether to create catalog), 'JDBC' (JDBC), and '驱动属性' (Driver Properties). The '是否创建catalog' field is highlighted with a red box, showing it is set to '是' (Yes) for 'dc_xny_fr'.

IoT数据分析



通过 Doris 存储全量 IoT 数据，进行准实时数据修正、分析，通过分区、倒排索引、算法优化，提升 90% 性能。

```
INDEX idx_power_station_no (`power_station_no`)
  USING INVERTED PROPERTIES("parser" = "english") COMMENT '电站编号倒排索引'
ENGINE=OLAP
UNIQUE KEY(`pdt`, `power_station_no`, `collector_no`, `product_id`, `device_id`,
COMMENT '逆变器明细表'
PARTITION BY RANGE(`pdt`)
```

日志存储查询



将日志存入从 es 迁移至 Doris，减少中间件，精简架构，节省 70% 的资源成本。

日志详情

自动刷新 刷新 X

步骤

- 全部
- 任务
- 执行步骤

步骤	时间	消息
1	[2024-09-12 15:47:22 274]	[消息] [任务] 任务步骤: [执行步骤]
2	[2024-09-12 15:47:22 274]	[消息] [任务] 任务开始
3	[2024-09-12 15:47:22 275]	[消息] [执行步骤] 自动
4	[2024-09-12 15:47:22 279]	[消息] [执行步骤] 初始化
5	[2024-09-12 15:47:22 279]	[消息] [执行步骤] 读取SQL脚本语句
6		select ... sales income_d_a;
7		select ...
8		select ...
9		select ...
10		select ...
11		select ...
12		select ...
13		select ...
14		select ...
15		select ...
16		select ...
17		select ...
18		select ...
19		select ...
20		select ...
21		select ...
22		select ...
23		select ...
24		select ...
25		select ...
26		select ...
27		select ...
28		select ...

手动触发 定时调度 外部调用 作业调度 补充数据 重新执行 未运行 失败 成功 运行中 已停止 队列中 等待中

状态	调度类型	读入数量	写入数量	待写入数量	错误数量	忽略数量	开始时间	等待时长	启动时间	操作
成功	手动触发	0	0	0	0	0	2024-09-12 16:00:18	3秒	2024-09-12 1	重新执行 日志详情
成功	手动触发	0	0	0	0	0	2024-09-12 15:55:10	0秒	2024-09-12 1	重新执行 日志详情
成功	手动触发	0	0	0	0	0	2024-09-12 15:54:44	0秒	2024-09-12 1	重新执行 日志详情
成功	手动触发	0	0	0	0	0	2024-09-12 15:52:33	0秒	2024-09-12 1	重新执行 日志详情
失败	手动触发	0	0	0	0	0	2024-09-12 15:47:21	0秒	2024-09-12 1	重新执行 日志详情
成功	手动触发	0	0	0	0	0	2024-09-11 11:12:34	0秒	2024-09-11 1	重新执行 日志详情

数据实时同步

```
{
  "jobcode": "test",
  "plugin": "mysql",
  "sinkType": "doris",
  "sourceDb": {
    "url": "",
    "host": "xx.xx.xx.xx",
    "port": "xx",
    "database": "dbname",
    "username": "xxxxxx",
    "password": "xxxxxx",
    "tableList": "dbname.table1,dbname.table2,dbname.table3"
  },
  "targetDb": {
    "url": "jdbc:mysql://xx.xx.xx.xx:xxxx",
    "driver": "com.mysql.cj.jdbc.Driver",
    "username": "xx",
    "password": "xxxxxxx"
  },
  "targetQ": {
    "bootstrap.servers": "ip1:port,ip2:port"
  },
  "parameter": {
    "map": [
      {
        "sourceTb": "table1",
        "targetTb": "table1_sink"
      }
    ],
    "mapQ": [
      {
        "sourceTb": "table1",
        "targetQ": "topic1"
      }
    ],
    "parallelism": "4",
    "ckPath": "xx",
    "ckInterval": "xx",
    "option": "latest"
  }
}
```

编辑: WMS实时同步

2 表映射

* 源: [下拉菜单] C * Kafka集群: [下拉菜单] C

开 结果写入数据库

* 目标: [下拉菜单] C

* 表映射: [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

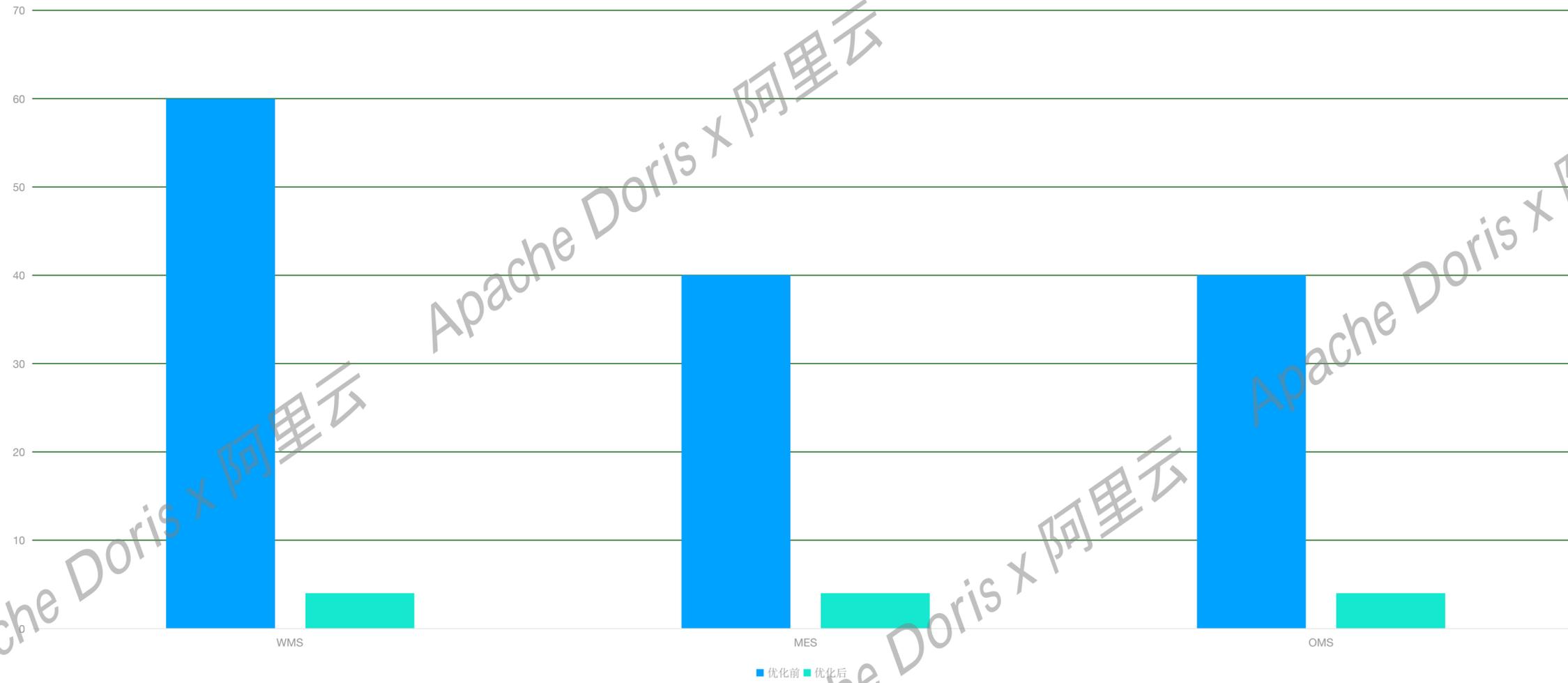
[下拉菜单] CDC [下拉菜单] C = 重建Topic [下拉菜单] C ⊕

上一步 下一步 确定 取消

对 Flink CDC 进行二开，将数据库 CDC 实时写入 Kafka，并通过 Routine Load 同步到 ODS 层，通过启用启用 sequence column 确保数据的准确性、一致性。

```
"enable_unique_key_merge_on_write" = "true",
"light_schema_change" = "true",
"function_column.sequence_col" = "dc_etl_time",
"disable_auto_compaction" = "false",
"enable_single_replica_compaction" = "false"
```

数据实时同步



通过整库同步大大降低资源同步的消耗，表越多，收益越大。

数据实时计算

```
CREATE ROUTINE LOAD all_manufacture_poc.kafka_ads_device_alter_agg_d ON ads_device_alter_agg_d
COLUMNS(SPEC, ID, FACTORY, EQUIPMENT
,CREATEDATE, GUID, EMAILGROUP
,LEVELENUM, STATUS, WAININGIDDES, RESOURCEGROUP
,dc_is_deleted, dc_etl_time, dt = substring(CREATEDATE, 1, 10)
,alert_count = 1),
```

Routine Load 实时计算

```
create view as ads_mf_lab_test_process_s_i as
SELECT
`systemType` AS `systemType`,
`testingType` AS `testingType`,
`experimentNumber` AS `experimentNumber`,
`lab` AS `lab`,
`client` AS `client`,
`clientUnit` AS `clientUnit`,
`clientDate` AS `clientDate`,
`testingProject` AS `testingProject`,
`status` AS `status`,
`contactNumber` AS `contactNumber`,
`model` AS `model`,
`sampleName` AS `sampleName`,
`createBy` AS `createBy`,
`hasChange` AS `hasChange`
```

创建普通视图

```
CREATE MATERIALIZED VIEW ads_fn_sales_income as
SELECT company_name, sales_income FROM dws_fn_company_sales_d_a ORDER BY company_name;
```

创建物化视图

指标标签



通过主键模型和部分列更新特性构建指标体系，用户通过可视化操作创建个性化指标。

数据填报

TB240564 指标属性

基本信息 排序配置 级联配置 预览 同步 保存 X

字段名	显示名称	数据类型	数据长度	数据精度	固定	显示	可编辑	必填	值唯一	加密	控件类型	基础数据	默认值	操作
i					左固定	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置
		文本	500	0	左固定	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置
su		文本	65536	0		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置
ie		整数	10	0		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	金额输入框			校验配置
ie		整数	10	0		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	金额输入框			校验配置
ie		整数	10	0		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	金额输入框			校验配置
ie		整数	10	0		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	金额输入框			校验配置
form		文本	65536	0		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置
dc_sys_id	主键	文本	500	0		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置
dc_sys_create_man	创建人	文本	65536	0		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置
dc_sys_update_man	修改人	文本	65536	0		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置
dc_sys_create_time	创建时间	日期时间	18	0		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	日期时间选择框			校验配置
dc_sys_update_time	修改时间	日期时间	18	0		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	日期时间选择框			校验配置
dc_sys_position	排序	整数	10	0		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	金额输入框			校验配置
dc_sys_batch	批次	文本	65536	0		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	文本输入框			校验配置

通过主键模型和merge_type，构建填报功能，并且实现行列权限的控制。

编辑角色权限:测试

* 角色名称: 测试

可对记录进行的操作: 可新增 可删除

可编辑和删除的记录范围: 所有记录 指定记录

符合 以下 全部 条件

1 工厂 = 等于

其他记录权限: 仅可阅读 禁止查看

字段内容权限: 可编辑 自定义

字段	可见	可编辑
dc_sys_create_man(创建人)	<input type="checkbox"/>	<input type="checkbox"/>
start_date(开始时间)	<input type="checkbox"/>	<input type="checkbox"/>
dc_sys_update_man(修改人)	<input type="checkbox"/>	<input type="checkbox"/>
dc_sys_position(排序)	<input type="checkbox"/>	<input type="checkbox"/>
dc_sys_id(主键)	<input type="checkbox"/>	<input type="checkbox"/>
dc_sys_batch(批次)	<input type="checkbox"/>	<input type="checkbox"/>
dc_sys_update_time(修改时间)	<input type="checkbox"/>	<input type="checkbox"/>
dc_sys_create_time(创建时间)	<input type="checkbox"/>	<input type="checkbox"/>
now_category(新分类)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

确定 取消

目录

01 背景介绍

02 数仓架构演进

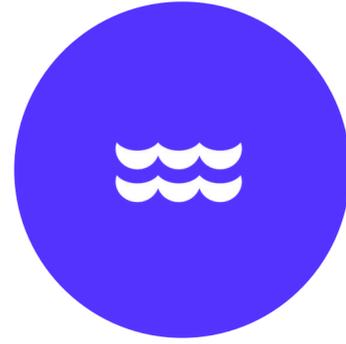
03 数据中台基于 Doris 的应用

04 未来规划

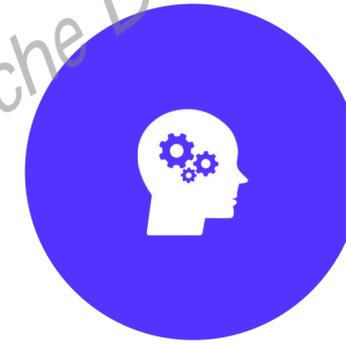
未来规划



架构升级



数据湖扩展



大模型集成

Thanks !



Apache Doris x 阿里云