SELECTDB | 阿里云

Apache Doris 城市行 2024

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

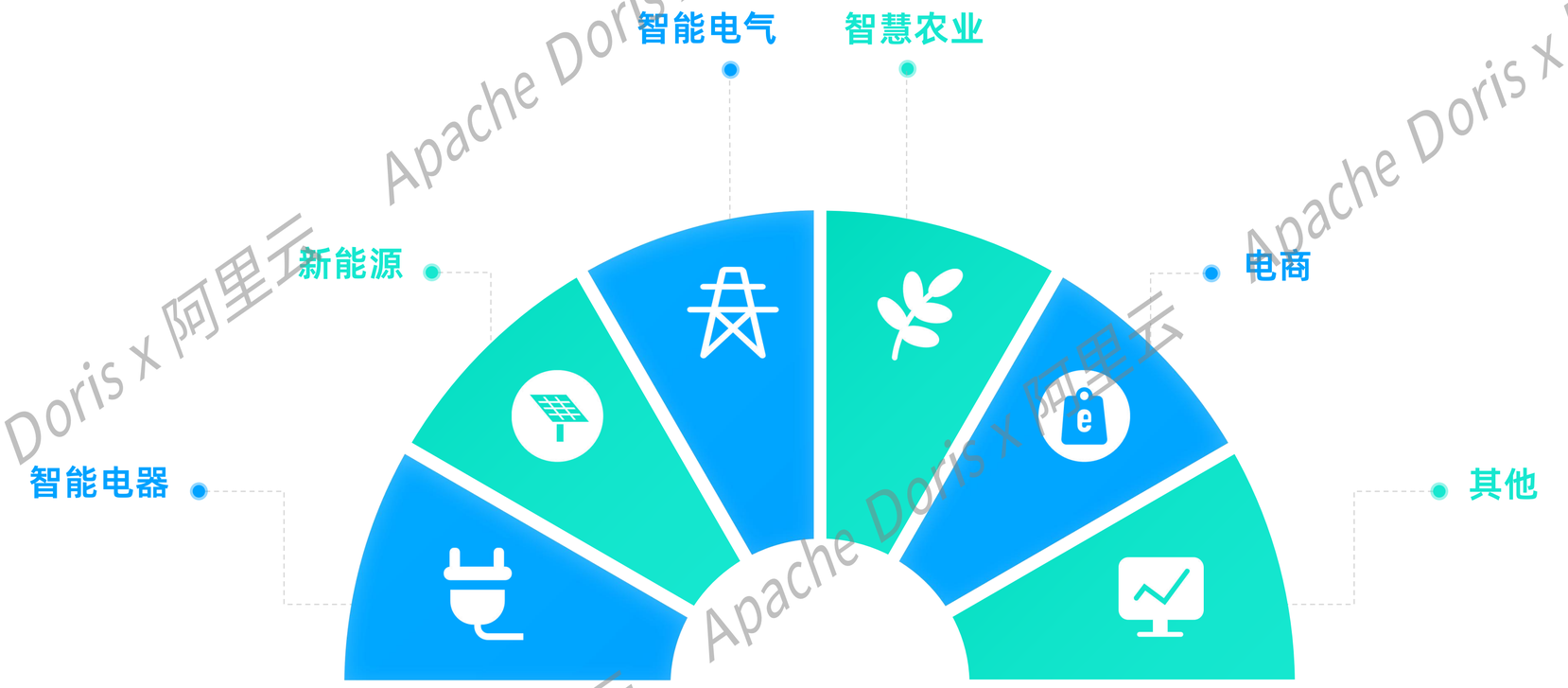
02 数仓架构演进

03 数据中台基于 Doris 的应用

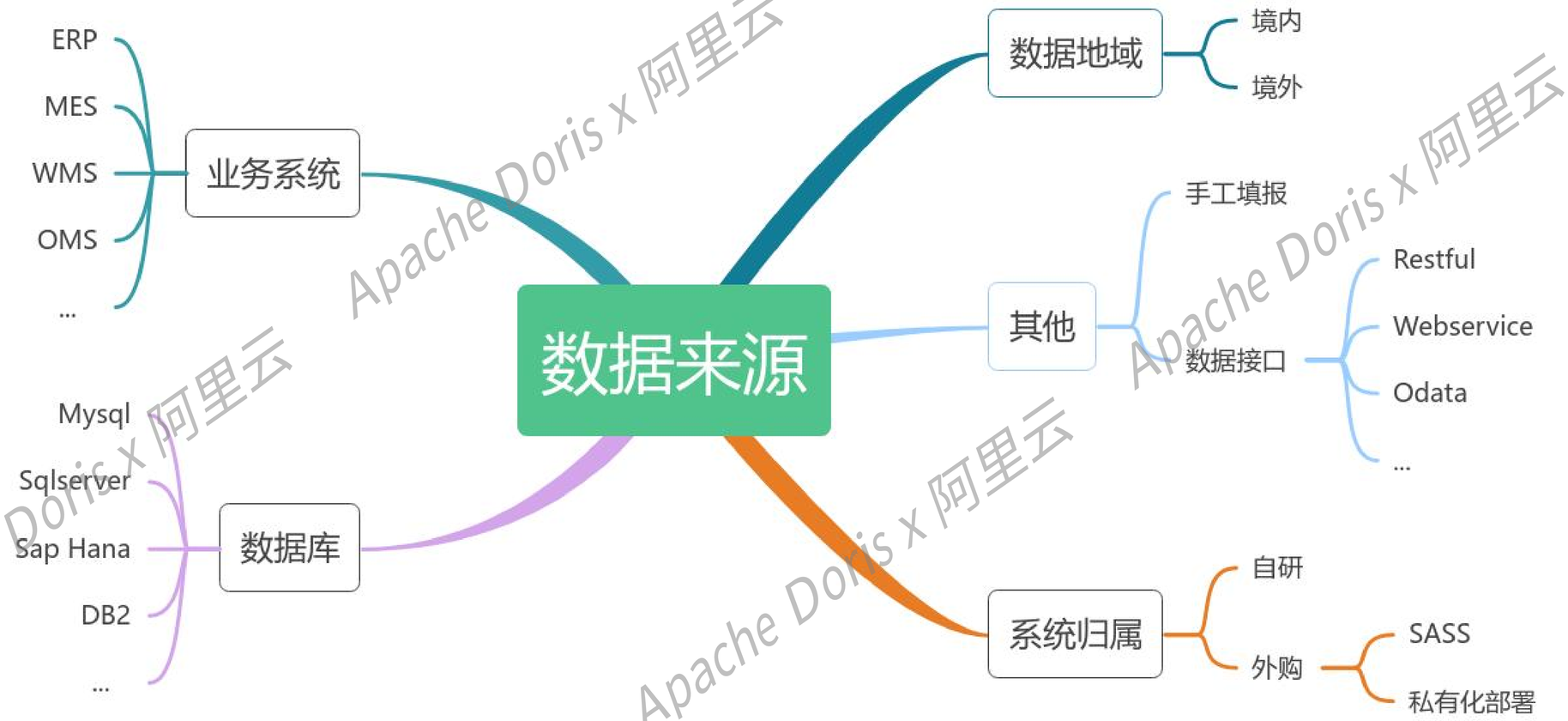
04 未来规划

集团业态分布

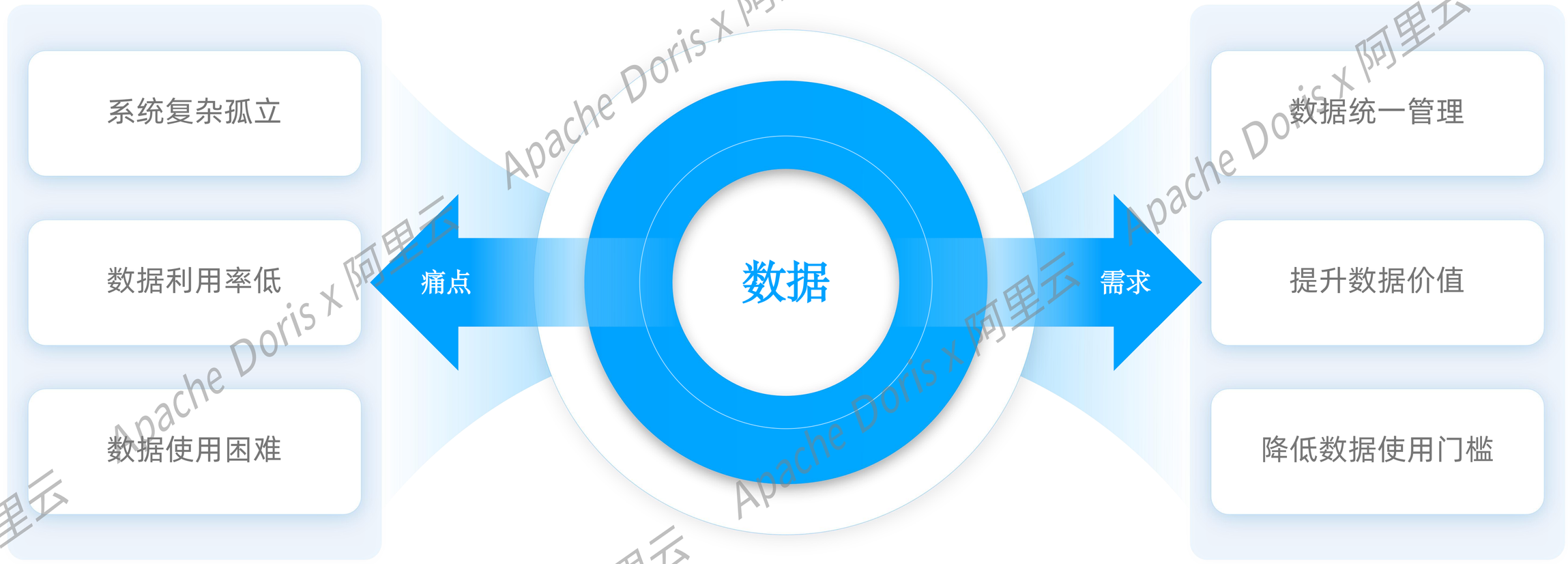
正泰集团股份有限公司始创于 1984 年，是全球知名的智慧能源系统解决方案提供商，业务遍及 140 多个国家和地区，全球员工 5 万余名，2023 年集团营业收入 1550 亿元，连续二十余年上榜中国企业 500 强。



数据源分布



数据痛点与需求



数据中台能力框架

中台应用

数据分享		数据可视化		数据质量	数据挖掘		数据应用	
API	数据集	数据地图	血缘追溯	数据比对	指标/标签	预测/推算	经营分析	销售分析预测
数据推送	数据查询	自助报表	自助看板	数据校验	模拟/仿真	预警/告警	智能制造	企业赋能

中台系统

数据集成		数据开发		数据服务	
数据源管理	结构化数据	数据建模	作业与任务编排	查询接口	推送接口
离线数据采集	非结构化数据	抽取、转换和加载	实时计算引擎	跨租户数据分享	BI集成
实时数据采集	数据存储	数据清洗和处理	机器学习	资产应用	应用管理

数据治理

资产管理	元数据	数据标准	主数据	数据质量	数据安全

中台数仓

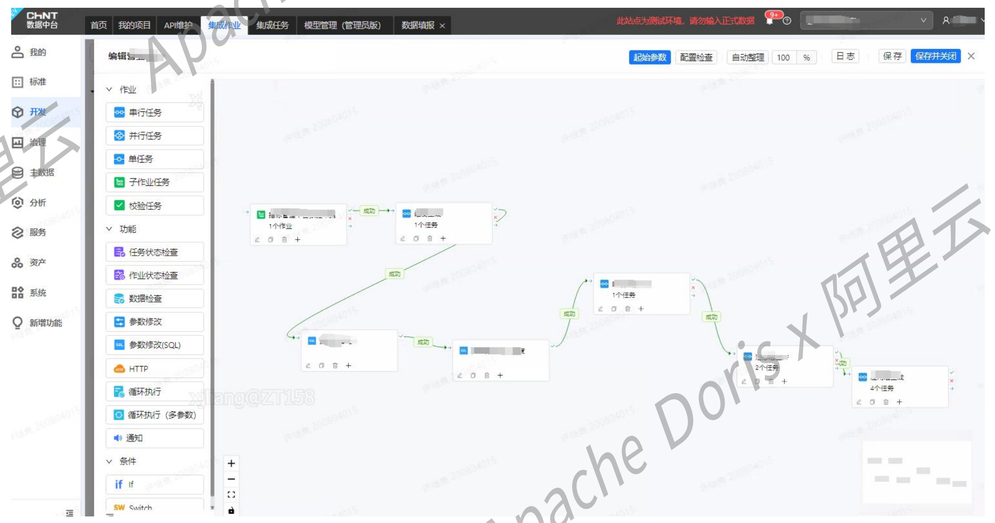
湖仓一体化		
结构化数据 二维表/图数据	半结构化数据 XML/JSON	非结构化数据 音频/视频/图片

数据集成		
ERP	OA	SF
MES	IoT平台	其他系统

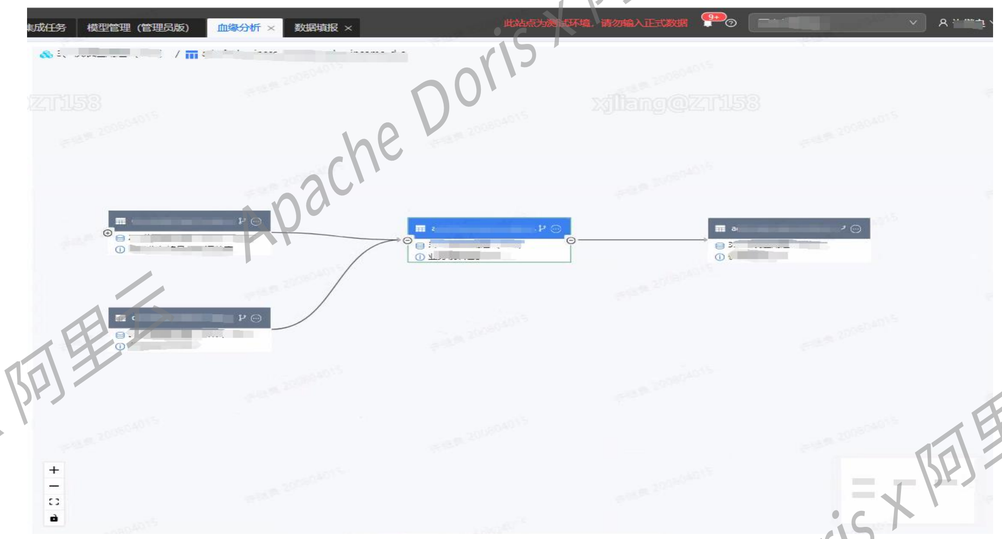
数据标准		
统一命名规范	标准数据模型	入湖入仓标准

数据中台功能

首页



集成作业



数据血缘

API服务

目录

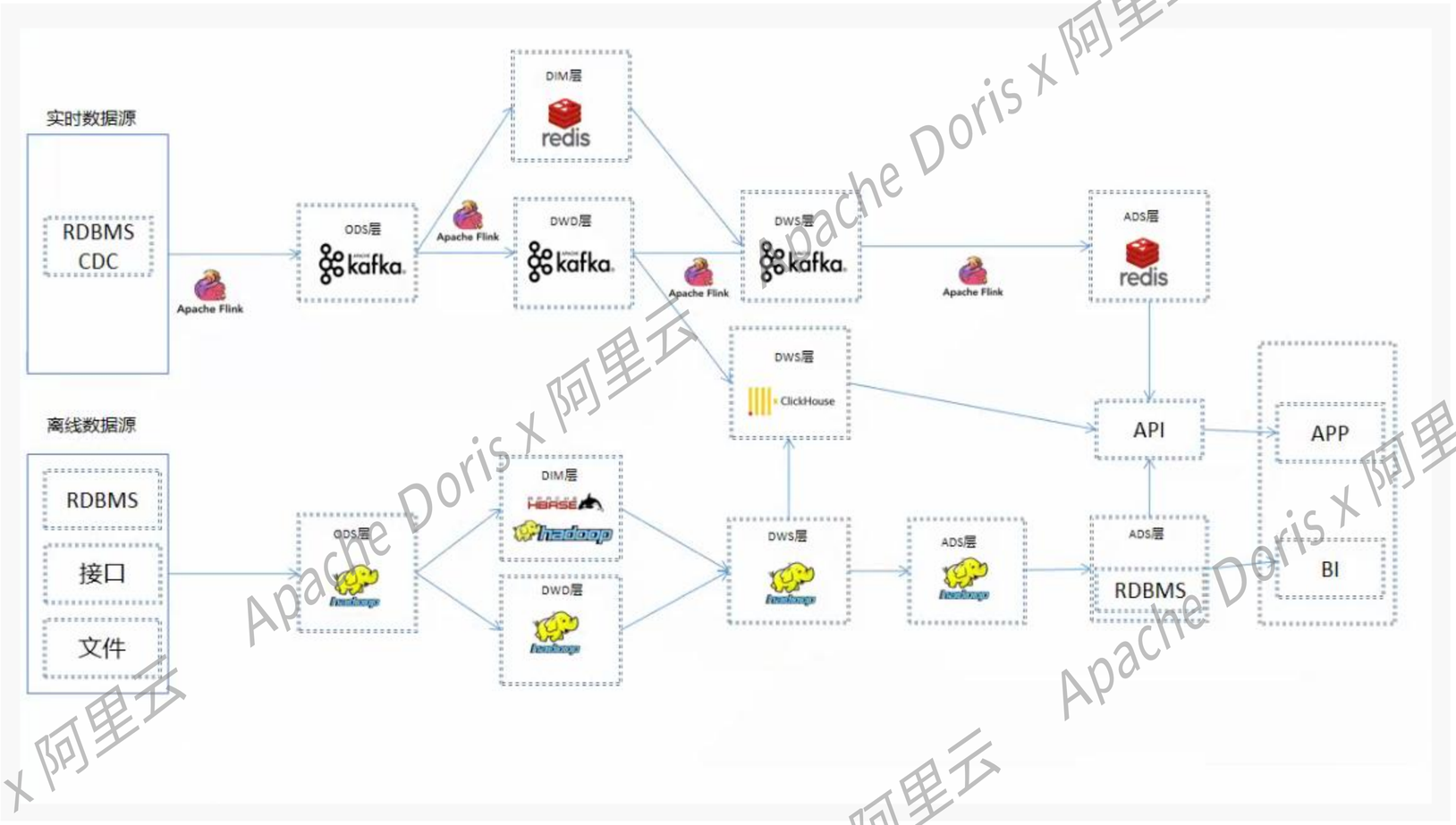
01 背景介绍

02 数仓架构演进

03 数据中台基于Doris的应用

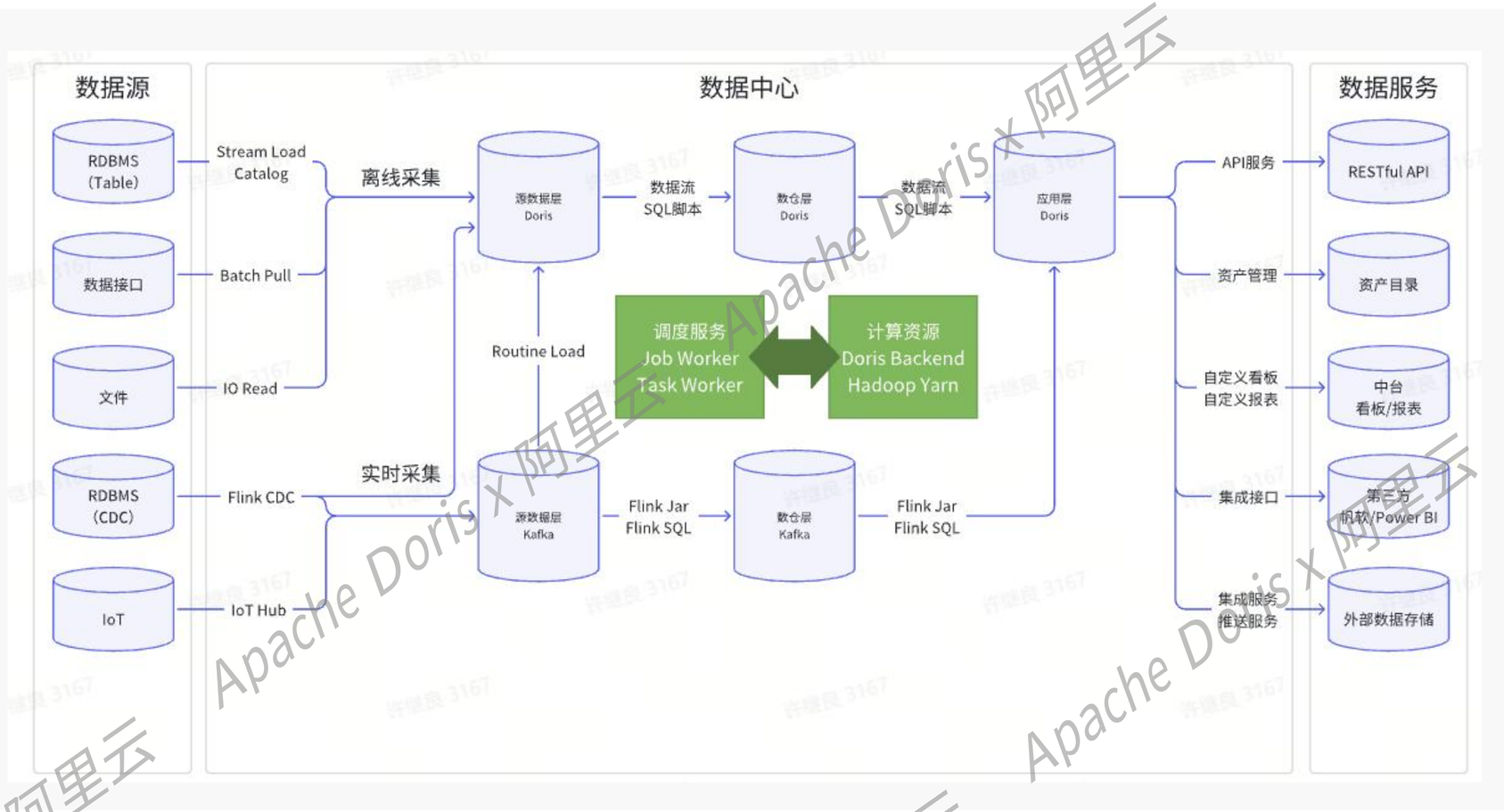
4 未来规划

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

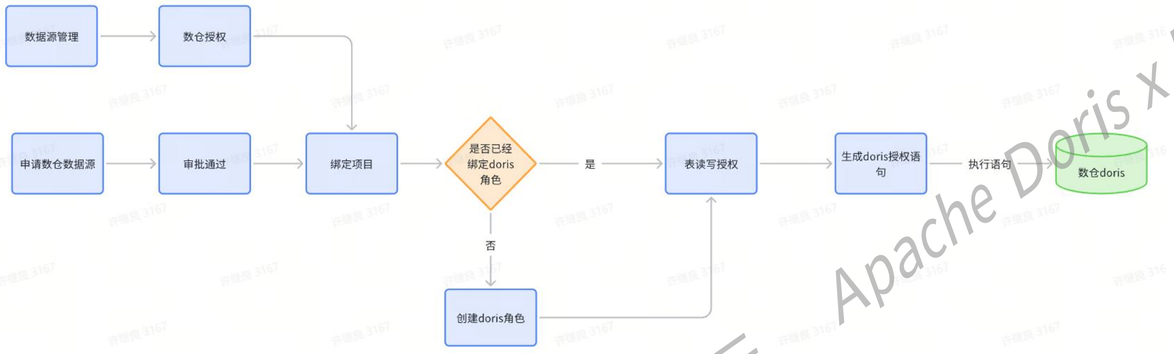
背景介绍

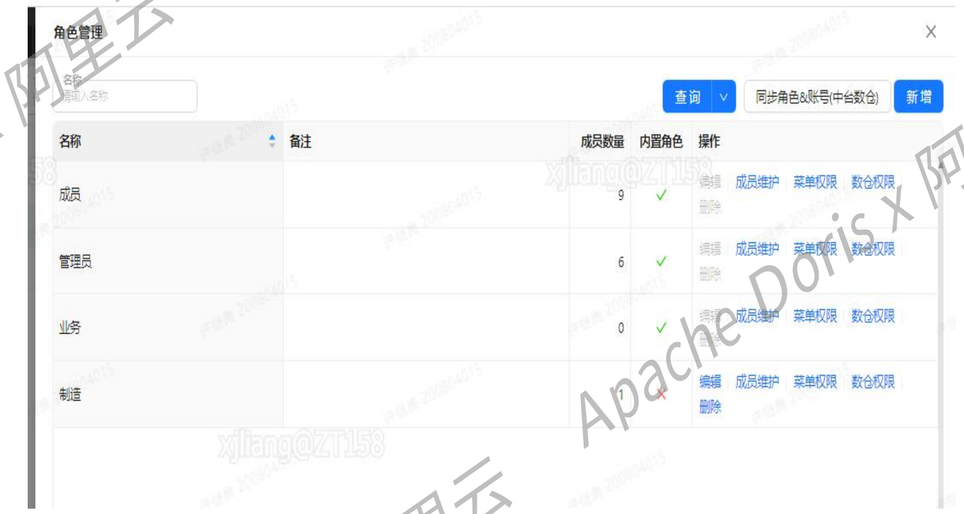
数仓架构演进

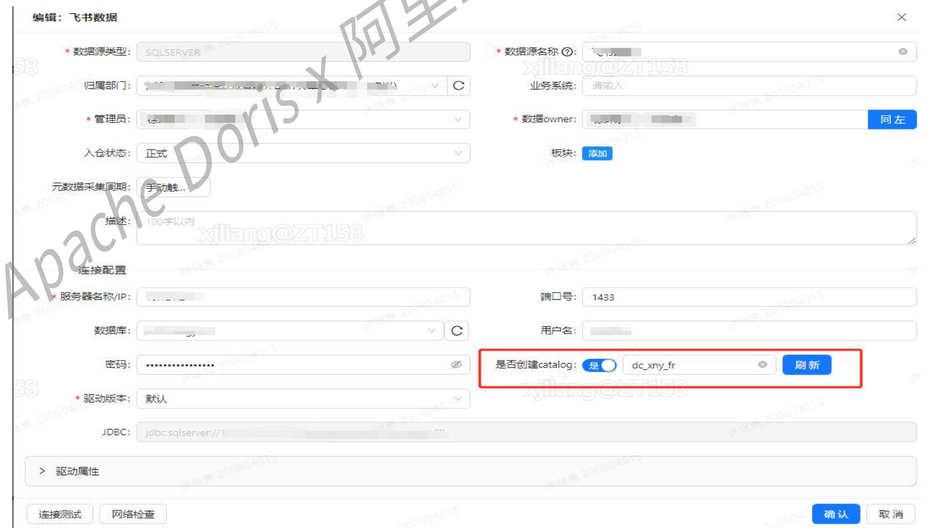
数据中台基于Doris的应用

未来规划

角色绑定

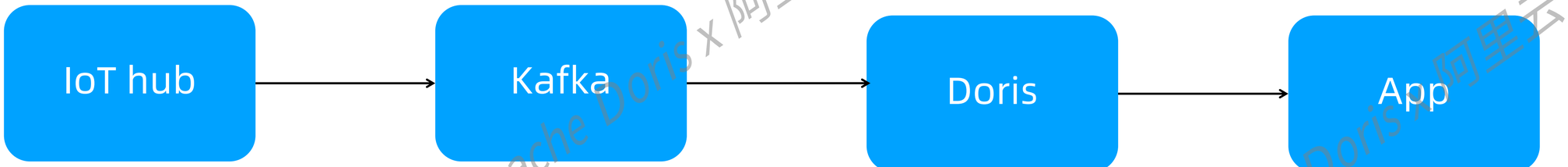






通过 Doris 2.0 数据角色与中台项目角色绑定，实现了 Doris 表级别控制(包含 catalog)，使数据权限更方便更安全。

IoT数据分析



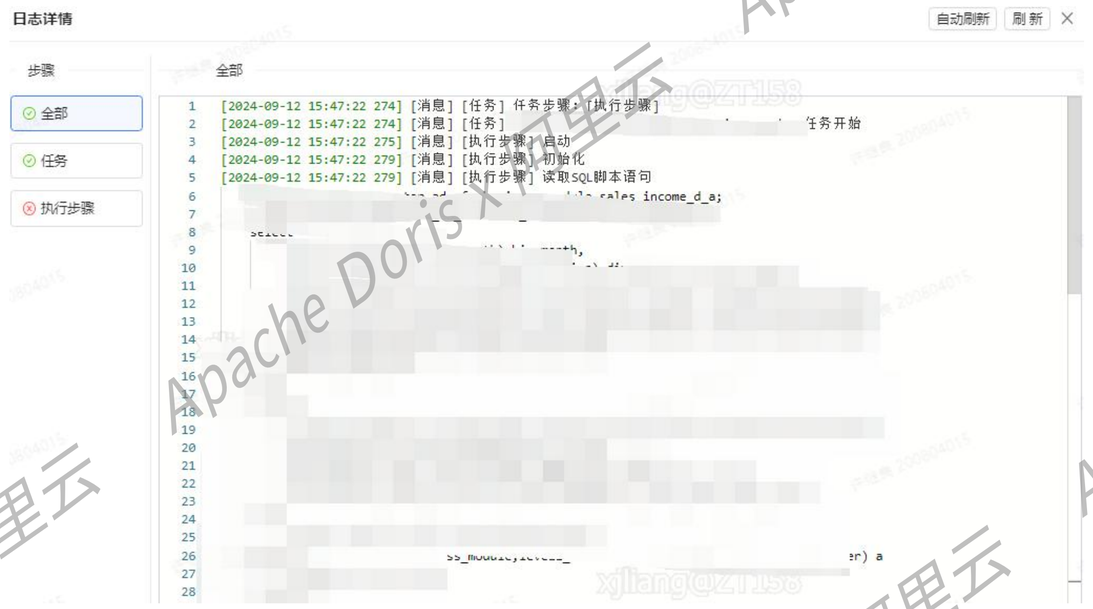
通过 Doris 存储全量 IoT 数据，进行准实时数据修正、分析，通过分区、倒排索引、算法优化，提升 90% 性能。

```
  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

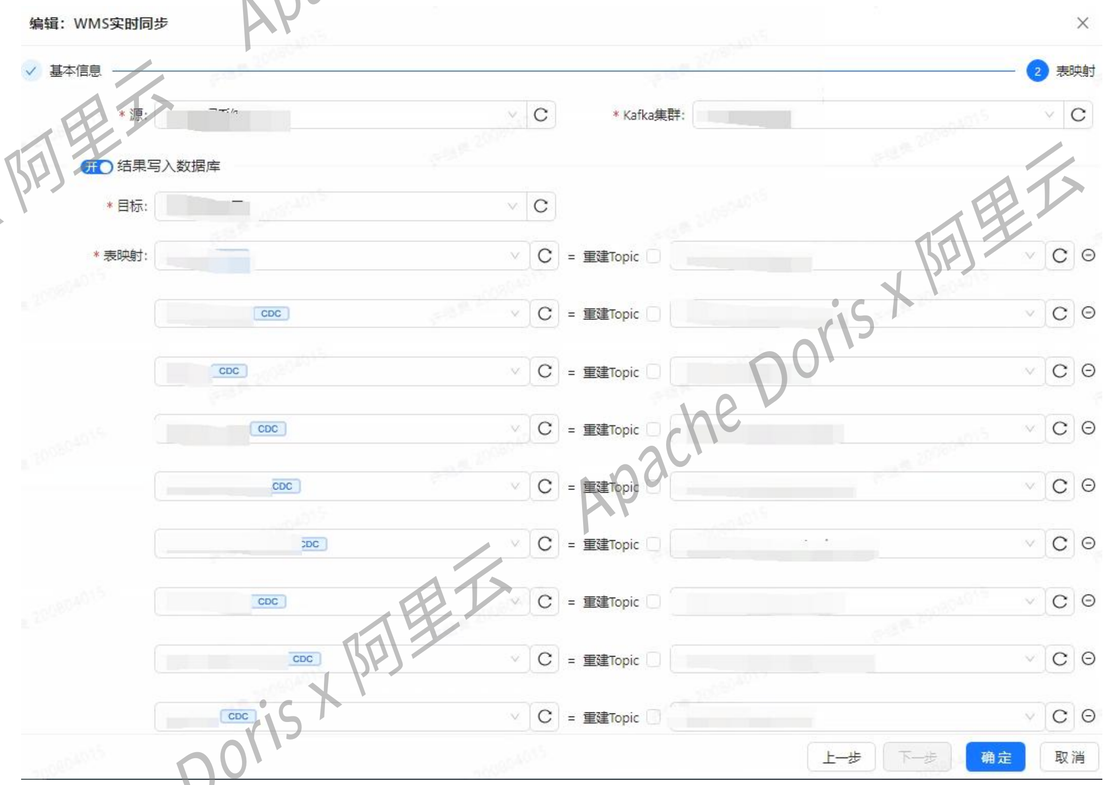
elasticsearch → DORIS

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。



数据实时同步

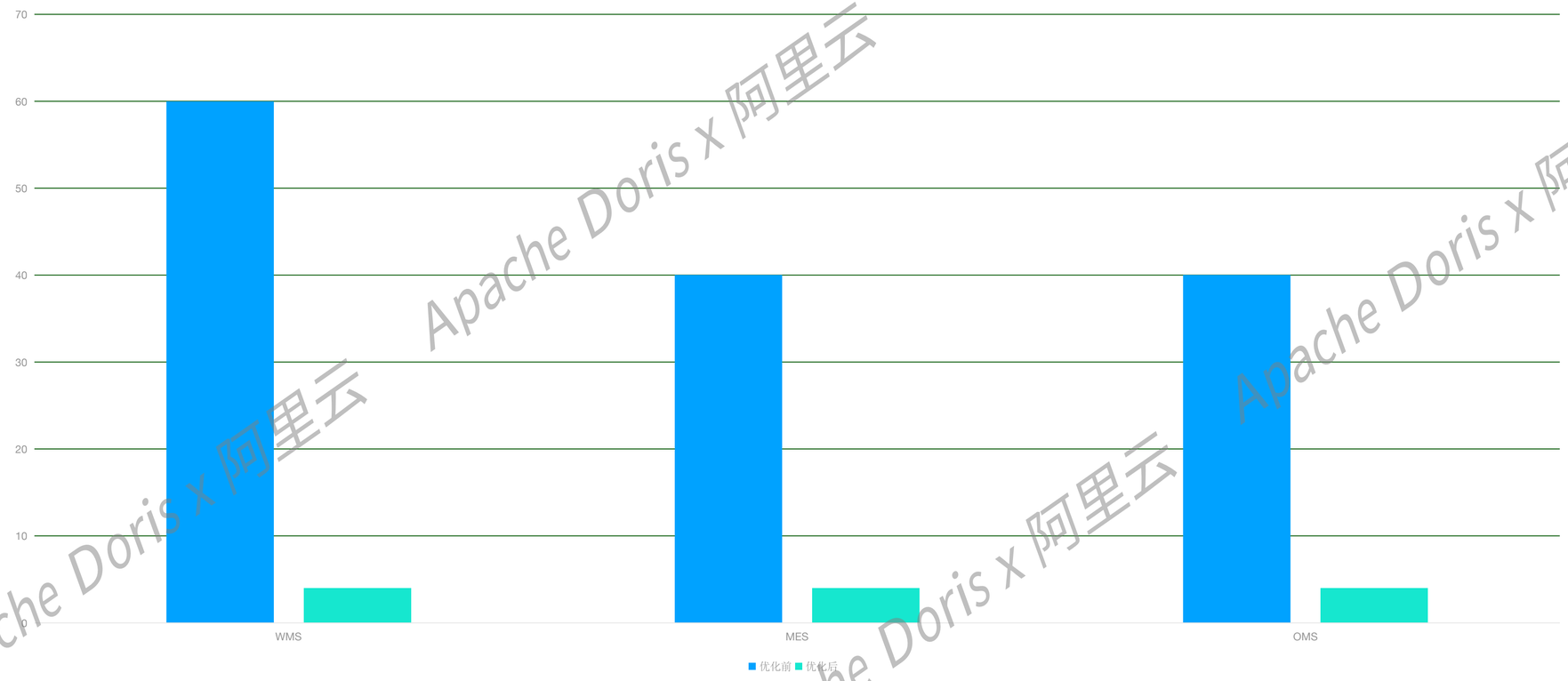
```
{
    "jobcode":"test",
    "plugin": "mysql",
    "sinkType":"doris",
    "sourceDb": {
        "url": "",
        "host": "xx.xx.xx.xx",
        "port": "xx",
        "database": "dbname",
        "username": "xxxxx",
        "password": "xxxxx",
        "tableList": "dbname.table1,dbname.table2,dbname.table3"
    },
    "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"
    }}
```



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

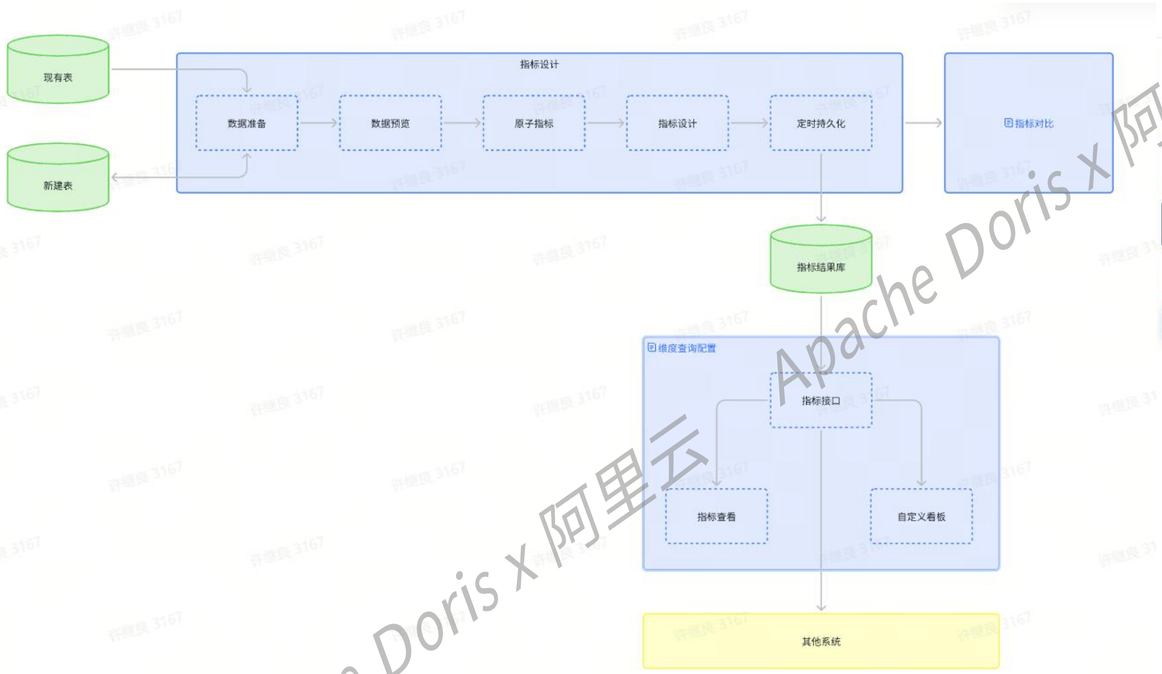
```
create view as ads_mf_lab_test_process_s_i as
SELECT
  `systemType` AS `systemType`,
  `testingType` AS `testingType`,
  `experimentNumber` AS `experimentNumber`,
  `lab` AS `lab`,
  `client` AS `client`,
  `clientUnit` AS `clientUnit`,
  `clientDate` AS `clientDate`,
  `testingProject` AS `testingProject`,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

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。

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 !