

# Apache Doris

## 在联通多业务场景下的应用与实战

田向阳

联通西部创新研究院 研发一部 数据研发专家

# 目录

1. 背景介绍
2. 业务场景
3. OLAP 选型
4. 平台架构
5. 未来规划

# 1 公司简介



联通西部创新研究院是中国联通集团在西部地区布局的创新主体，由联通数字科技有限公司出资设立，陕西省分公司协同运作，并与西安交通大学、西安电子科技大学等高校开展深层次联合创研。

联通西部创新研究院将原数科公司总部西安研发中心和原云数据事业部西安研发中心整体纳入，锚定“做成规模专业化高水平的研发机构”高定位，内部设立多个研发部门及技术部，分别承担数科公司各事业部的重点研发项目及任务，肩负“先行者”“排头兵”使命感，持续推动公司研发集约化。



# 「技术布局」





# 业务场景

「丰富的业务场景」



多行业

生产制造  
经营分析  
海量日志  
智慧政企

...



集群规模

无配套  
小集群  
大集群  
超大集群



数据量

GB级  
TB级  
PB级



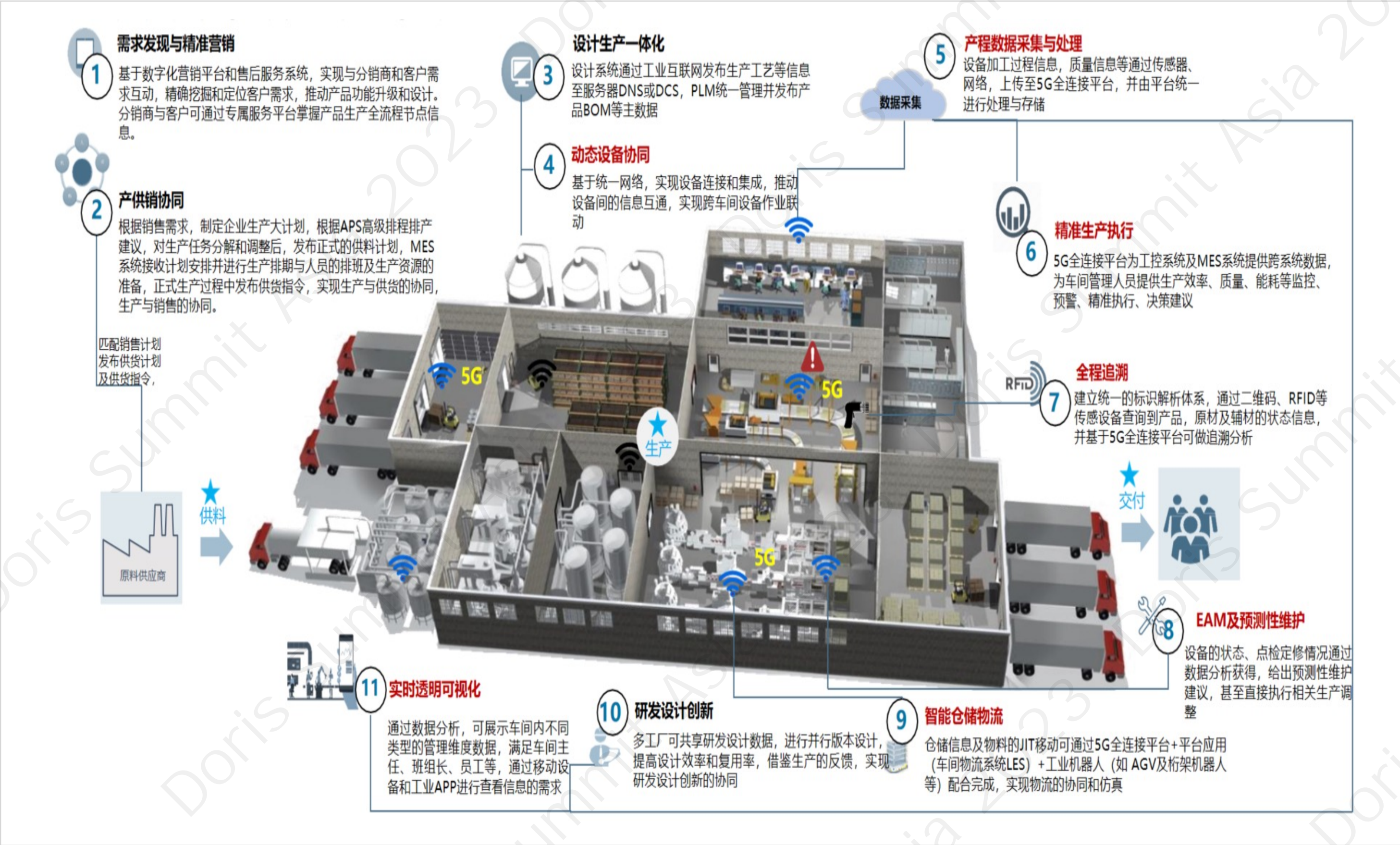
数据需求

数据资源目录  
可视化大屏  
自主分析  
海量数据检索  
算法模型

...



# 「工业互联网-5G全连接」



## 项目特点

5G全连接平台是助力生产制造企业数字化转型的能力底座平台，具备工业数据采集、工业数据治理、工业业务分析、低代码大屏等能力（应用集成+数据集成）

### 1. 数据源类型丰富

主要来自于各个业务系统数据、物联网设备数据、日志数据，采集方式分为实时和离线批量采集

### 2. 业务诉求

数据需求主要是提供大屏、在线自主分析、经营报表

### 3. 数据量中等

### 4. 存储方式

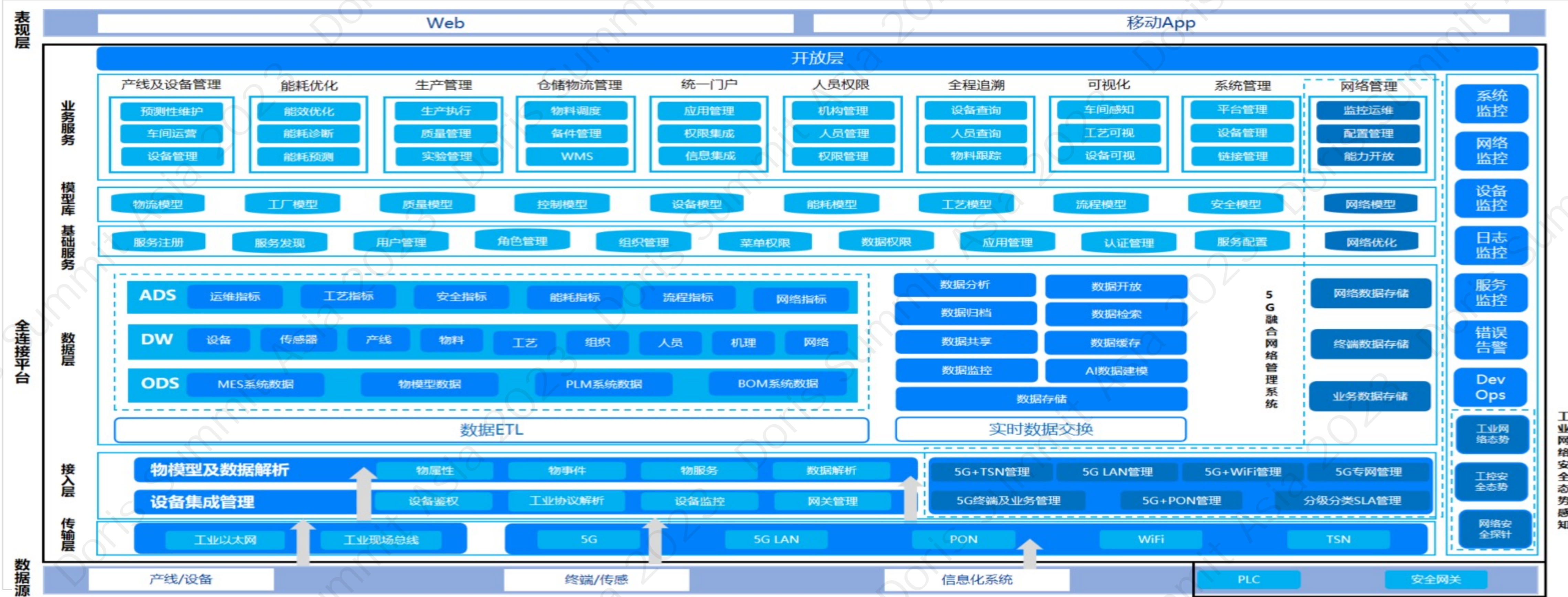
Doris+时序数据库



# 5G 全连接功能架构



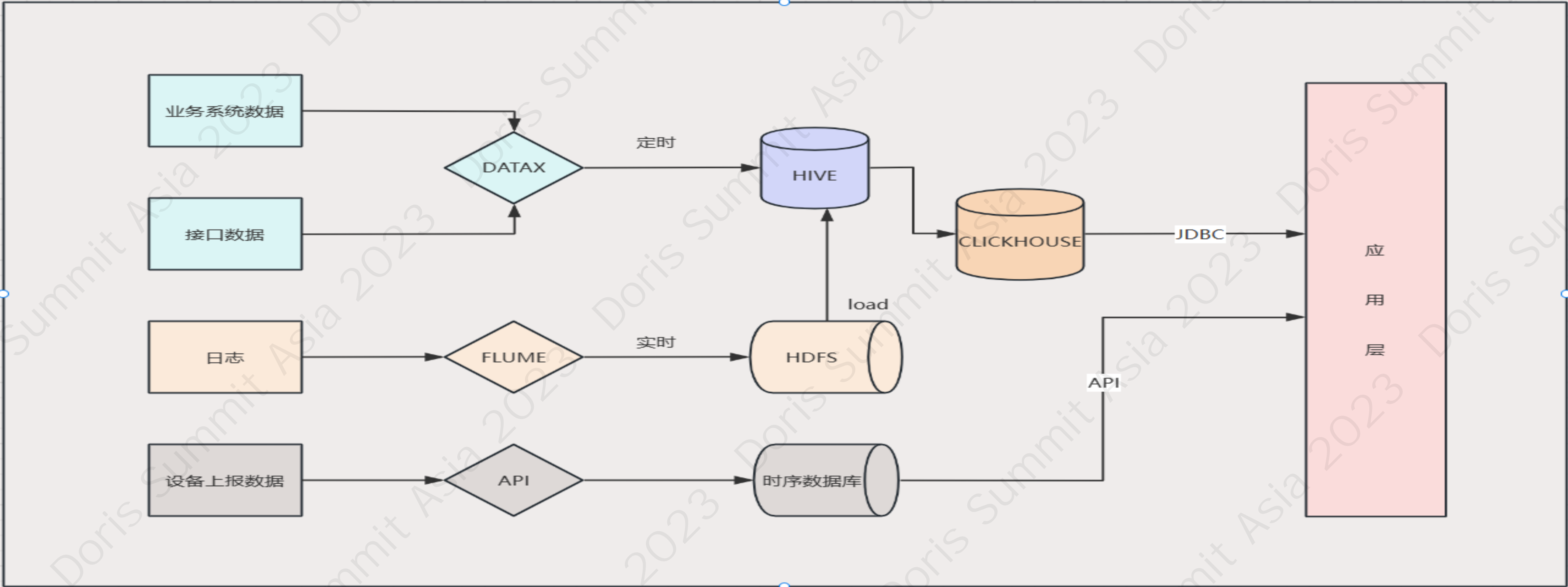
「5G 全连接」





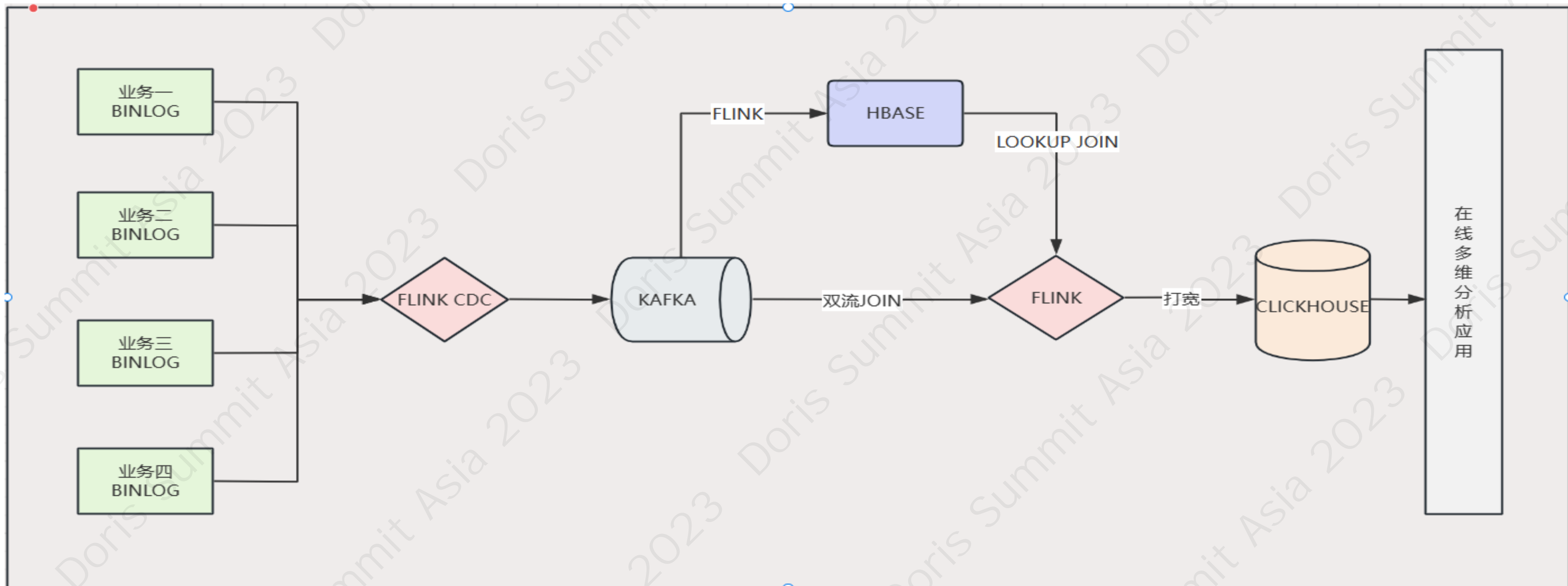
# 5G 全连接数据架构

# 「老架构离线数据链路」





## 「老架构实时数据链路」



# 「业务痛点」



## 数据链路长且复杂

- 离线产出报表需要同步到MYSQL提供查询
- 实时链路计算复杂度高



## 数据准确性问题

- 实时计算采用多流join由于业务更新复杂性导致join结果正确性有3%-5%偏差，每天通过离线数据修正



## 维度变更问题

- 宽表模型下，业务维度变更需要对历史数据离线重刷

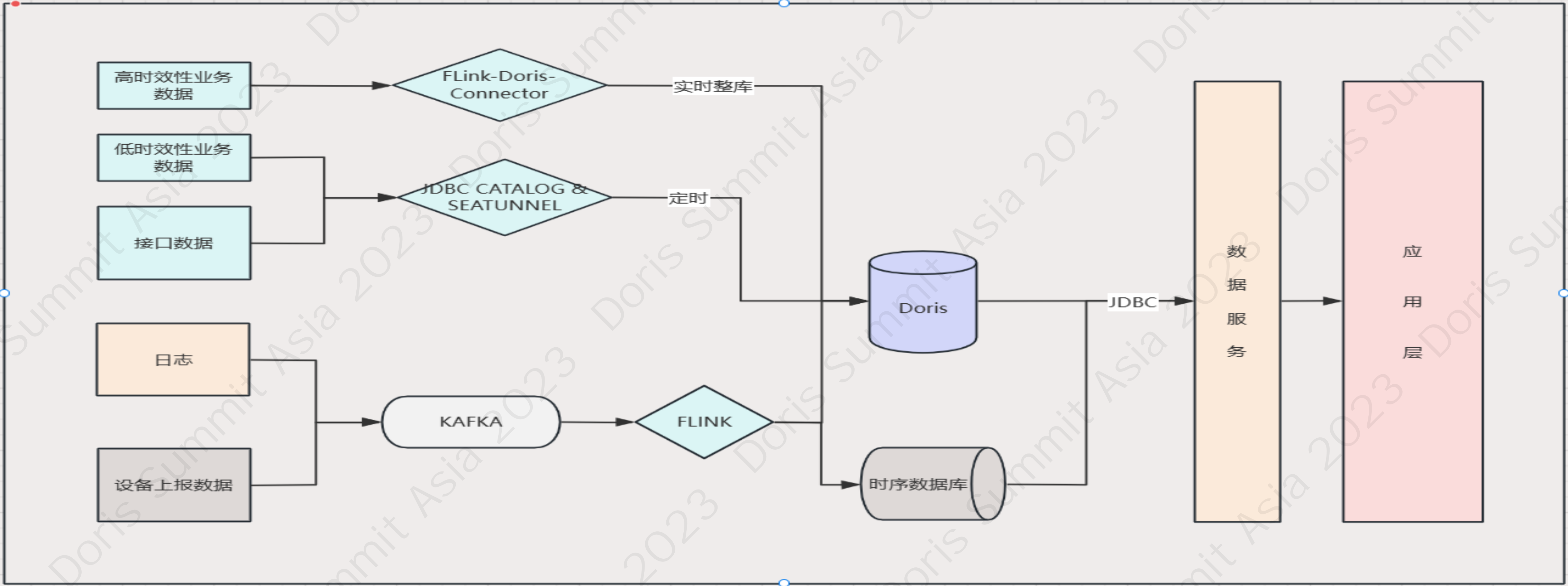


## 维护成本高

- 整个架构涉及技术栈和组件较多，维护成本高



「新架构」



# 「基于 Doris 新架构的收益」

## 数据时效性、准确性（实时场景）提高

相比于之前老架构的小时级数据延迟，新的架构支持数据写入秒级延迟。且实现了流批一体，计算结果的一执行

## 数据处理链路缩短

老架构在基于hive数据建模加工之后将数据同步到clickhouse提供给应用查询，新架构直接doris通过Doris提供查询，避免了额外的数据同步链路

## 业务灵活性准确性提高

老架构olap查询使用了宽表模型，固化了维度，而新的架构直接将dwd事实表和维表开放给应用，直接实现olap查询，提供给业务更高的灵活性

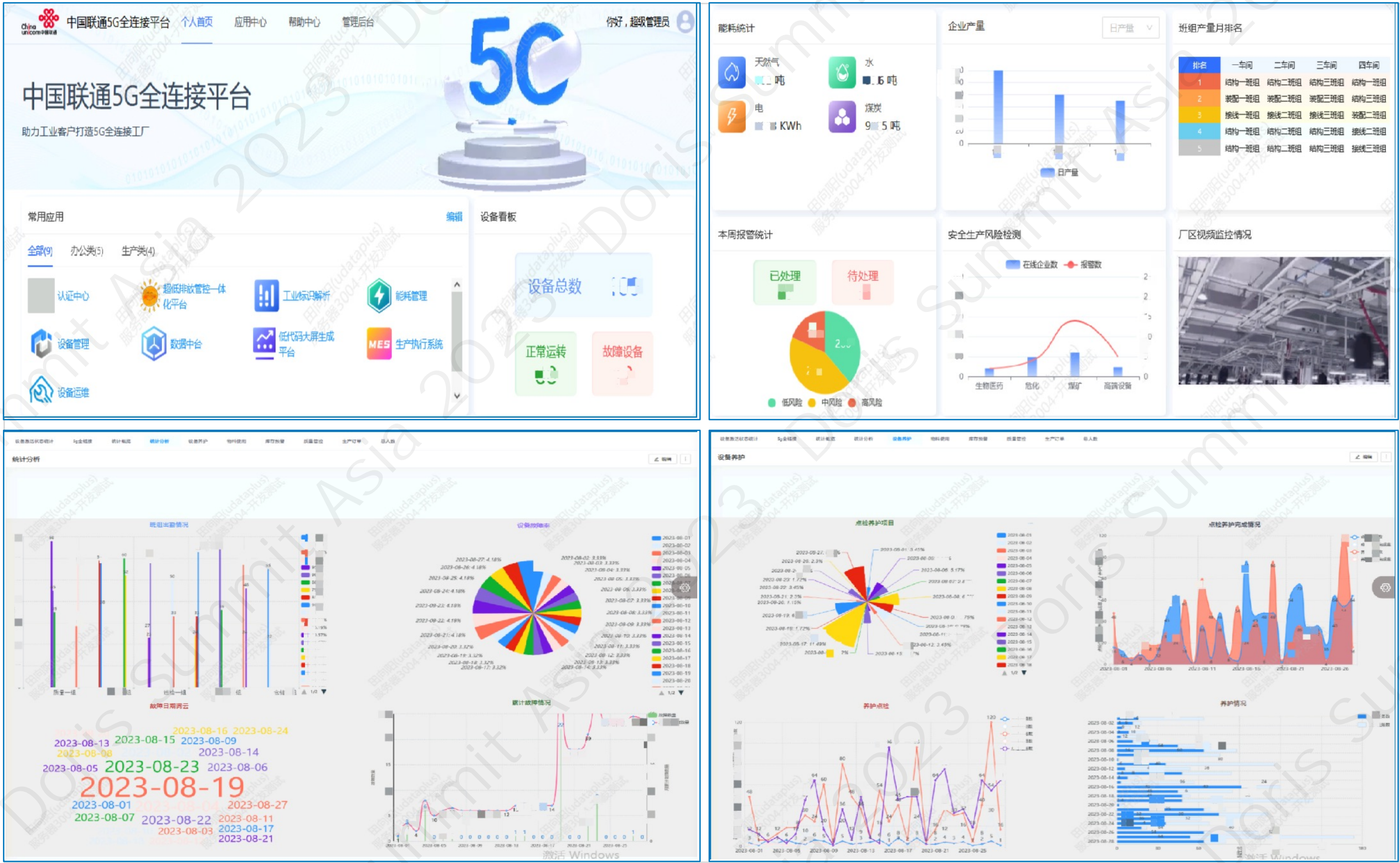
## 架构灵活性提高

老架构完全依赖于hadoop生态，新的架构可根据客户的数据规模对hadoop生态进行取舍，架构上有了更灵活的选择





「5G 全连接应用」



实现业务效果

1. 70%的核心业务实现准实时

(核心业务Flink CDC + Doris) 实时链路+  
(边缘业务JDBC catalog)定时批量抽取

2. 报表毫秒级响应

JDBC+Doris 维度建模定时报表数据产出

3. 在线近实时分析多维分析

DWD明细表基于时间分区+分桶join



「智慧政企-数字乡村」



项目特点

- 1. 数据源类型单一  
主要来自于各个业务系统数据库采集，采集方式分为实时和离线批量采集
- 2. 业务诉求  
数据需求主要是提供大屏、可视化报表、数据服务
- 3. 数据量小
- 4. 存储方式：Doris



「数字乡村可视化大屏」

实现业务效果



1. 大屏指标毫秒级返回

基于Doris物化视图加速大屏指标查询

2. 用户中心读写分离

Seatunnel Zeta+MYSQL CDC将用户数据实时同步到Doris，提供数据读取服务，实现读写分离。

3. 生成数据资产目录

平台侧根据数据业务域对数据表统一分类形成数据资源目录，通过数据服务模块对外提供数据开放能力



## 「海量日志场景-安全数据分析平台」



### 建模分析

基于网络日志数据和告警数据进行规则或智能挖掘，发现潜在的安全事件，例如钓鱼邮件、非法访问等，实现定向威胁感知。



### 态势大屏

通过多种维度不同监控指标的组合，例如安全事件TOP5等。密切监控当前网络安全态势状况，呈现出攻击威胁的主要分布。



### 追踪溯源

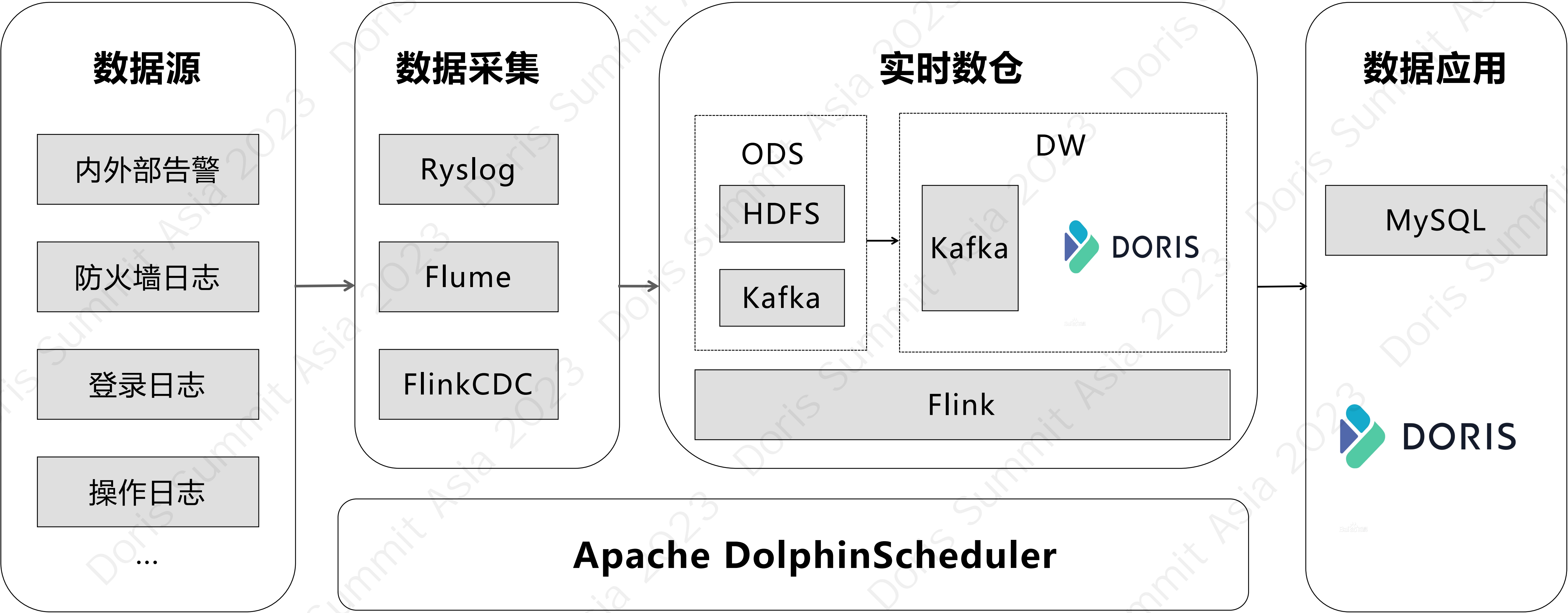
通过对安全事件的快速研判，还原整个攻击链条，进行精准的溯源取证，保障网络和数据安全。



## 「安全数据分析平台业务需求」

- 明细日志秒级查询
- 数据聚合快速响应
- SQL支持
- 实时日志，秒级实时落地
- 正则匹配查询
- JSON数据存储和查询
- ARRAY数据统计计算

# 「安全数据分析平台架构」





安全数据分析平台数据规模

每日入库数据量

30<sub>T</sub>

每日入库数据条数

150<sub>亿</sub>

单表最大数据量(20天)

180<sub>T</sub>

单表最大数据条数(20天)

900<sub>亿</sub>

## 2 OLAP 选型

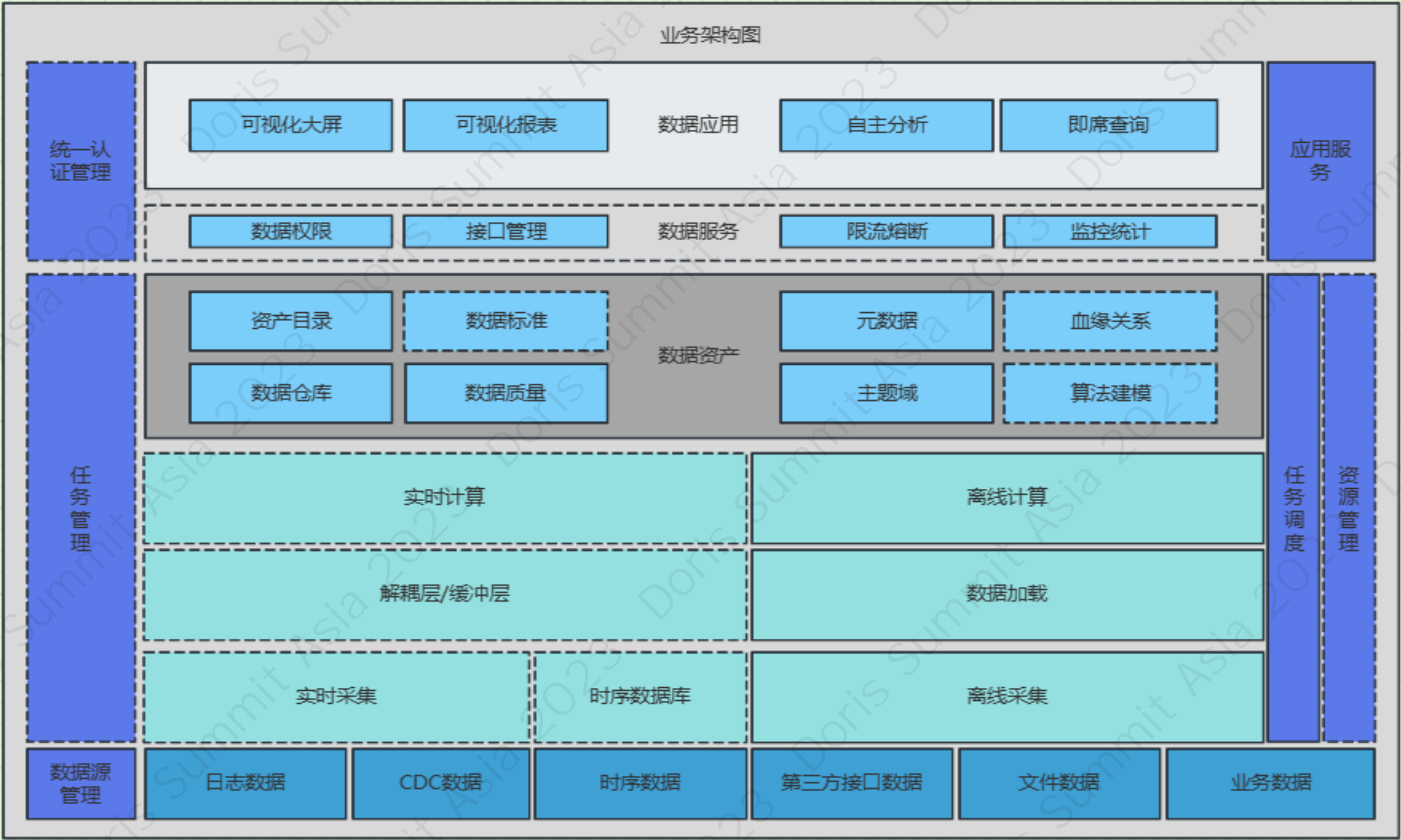


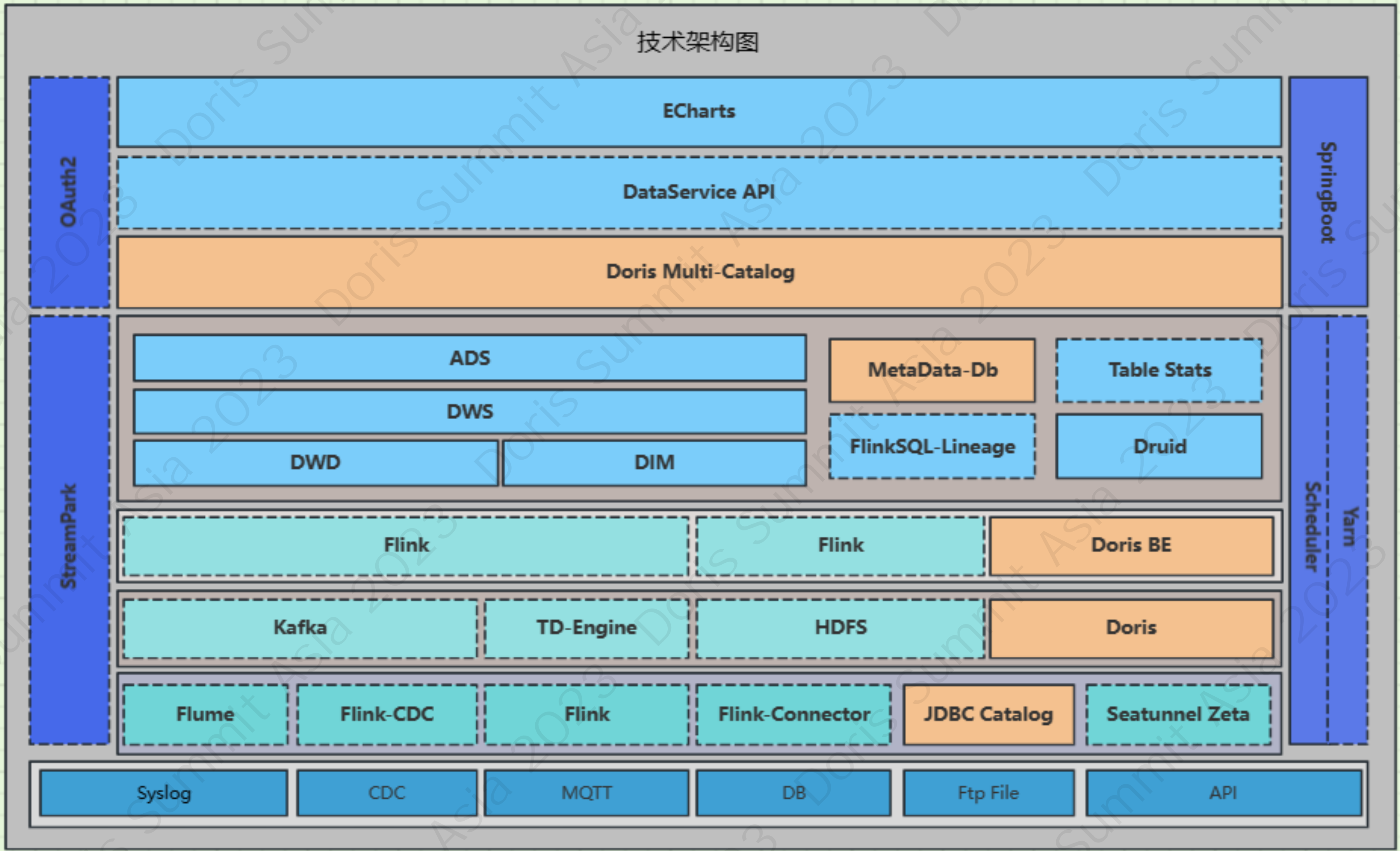
OLAP 选型对比

| 引擎         | 运维成本 | 实时写入 | 实时更新 | 低延迟查询 | 多表关联查询 | 扩展性 | 强依赖<br>Hadoop |
|------------|------|------|------|-------|--------|-----|---------------|
| Doris      | 低    | 支持   | 支持   | 是     | 好      | 好   | 否             |
| ClickHouse | 高    | 支持   | 不支持  | 是     | 不好     | 好   | 否             |
| Hive       | 中    | 支持   | 不支持  | 否     | 好      | 好   | 是             |
| 数据湖方案      | 中    | 支持   | 支持   | 否     | 好      | 好   | 是             |

# 3 平台架构









# 4 架构优势

# 「基于Doris架构的优势」

## 运维成本

基于纯 Doris 生态的轻量级的存储架构，代替了之前的基于 Hadoop 的 Lamda 架构，实现运维成本降低超20%。

## 开发效率

90% 的实时计算，基于 Doris 内部数仓加工体系，代替老的 Flink 多流 join 的复杂场景任务开发，实现开发效率提升超20%。

## 数据准确率及时性

基于 Doris 流批一体数仓建设，重点业务实现 cdc 数据整库高效接入。实现数据准确率及时性提升超5%。

## 交付周期

开发效率提升，运维成本降低带来项目交付周期缩减超20%。



# 5 未来规划

# 「下一步规划」



Doris 2.0

升级版本到2.0，新  
版引擎下高性能查询  
体验



基于 Doris 存  
算分离架构探索

探索HDFS湖存储架  
构下doris计算节点  
部署，联邦查询能力



Doris  
Manager

基于 Doris Manager  
进一步简化运维工作



深入优化探索

根据业务特点进一步  
优化，比如索引加  
速，慢查询监控，资  
源队列划分等等





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](https://doris.apache.org)
- Apache Doris GitHub: [github.com/apache/doris/](https://github.com/apache/doris/)

### 获取更多峰会资料:

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