

Apache Doris 在华为云的实践

鲁光明

华为 大数据高级研发工程师

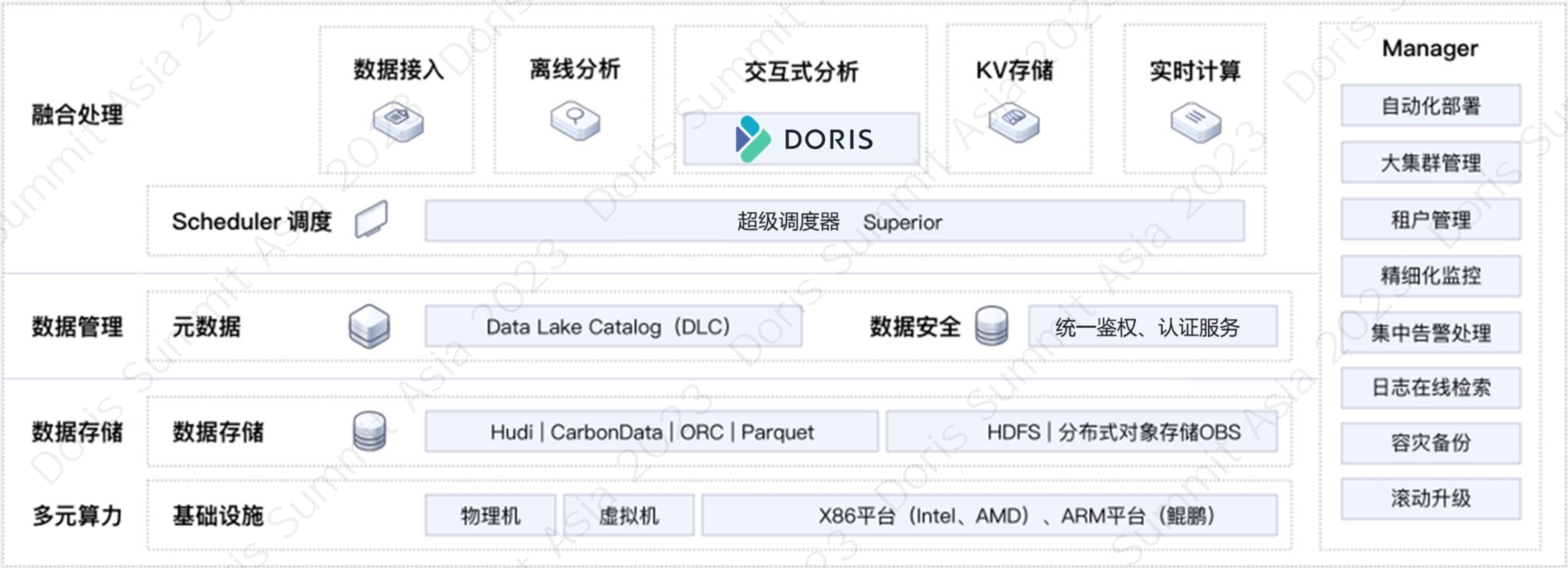
目录

1. Doris 在华为产品中的场景
2. 企业级增强
3. 客户案例分享
4. 未来的规划

1 Doris 在华为云产品中的场景

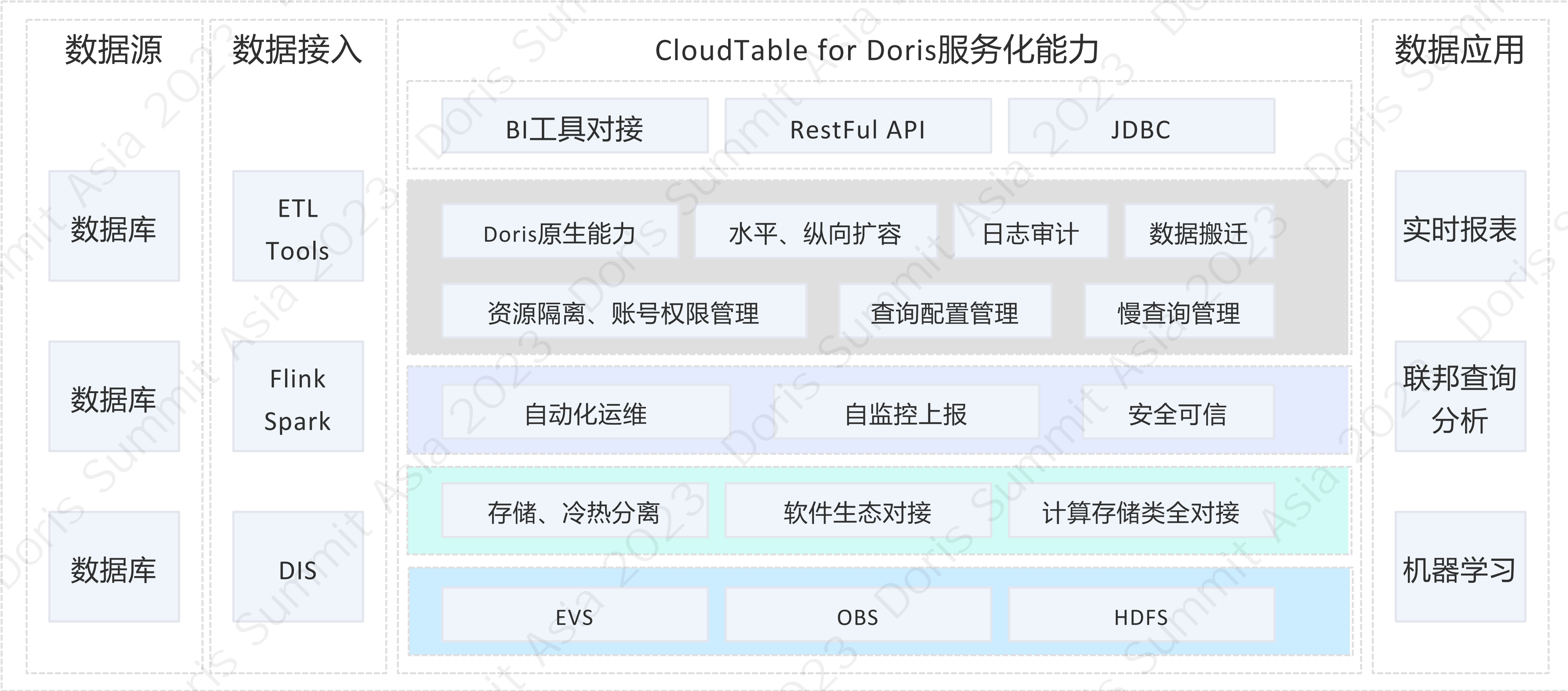
MRS For Doris

云原生数据湖 MRS（MapReduce Service）为客户提供 Doris、Hudi、ClickHouse、Spark、Flink、Kafka、HBase 等Hadoop 生态的高性能大数据组件，支持数据湖、数据仓库、BI、AI 融合等能力。MRS 同时支持混合云和公有云两种形态：混合云版本，一个架构实现离线、实时、逻辑三种数据湖，以云原生架构助力客户智能升级；公有云版本，协助客户快速构建低成本、灵活开放、安全可靠的一站式大数据平台。



CloudTable For Doris

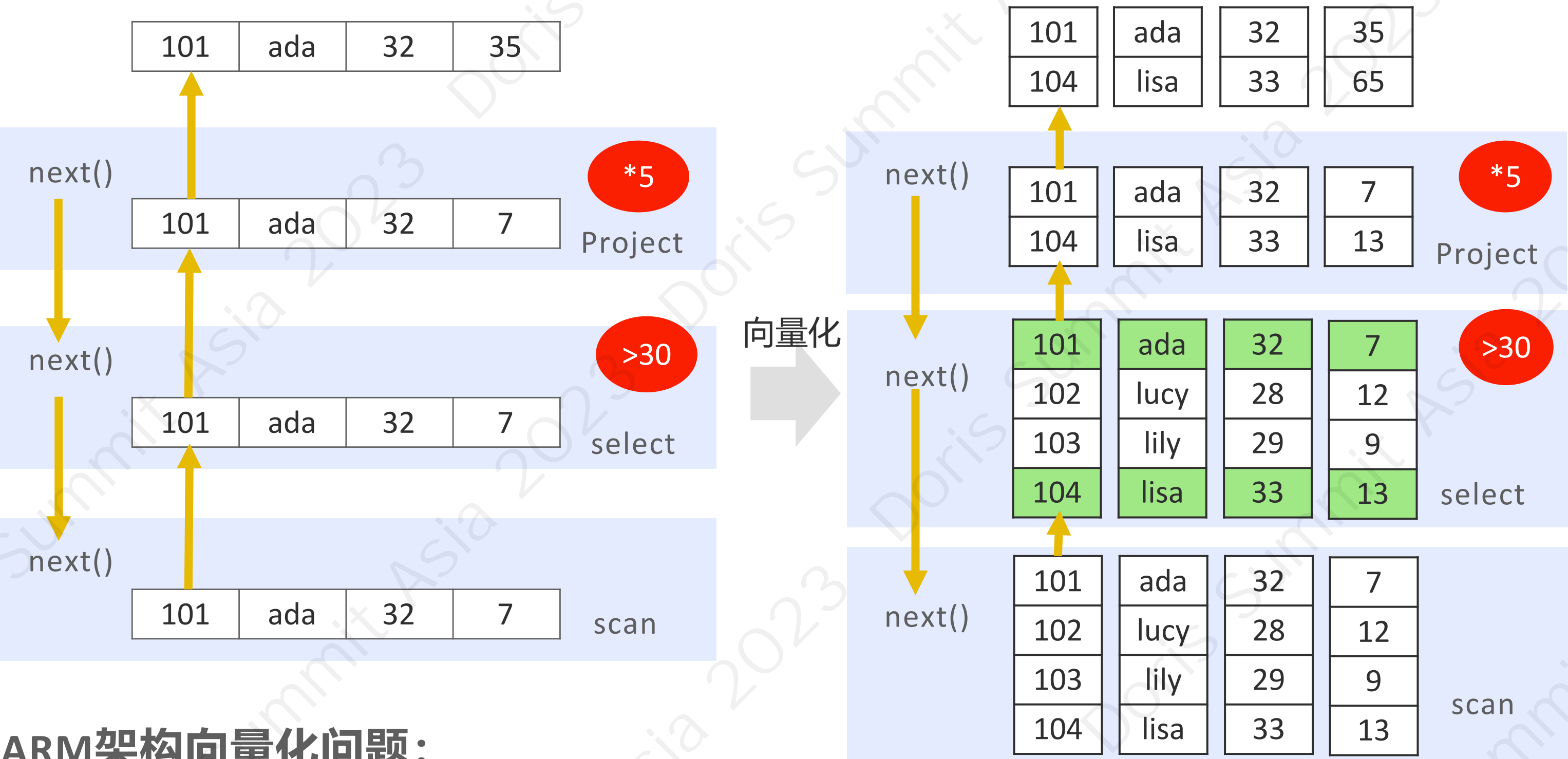
表格存储服务（CloudTable）是一款Serverless化产品为用户提供专属集群，即开即用，适合业务吞吐量大，时延要求低的用户。轻松运行HBase、Doris、ClickHouse等大数据组件。



2 企业级增强

Doris ARM架构向量化优化

背景：Doris 通过向量化计算提升查询性能，向量化计算具备以下优势：cache 更亲和、减少虚拟函数调用、函数计算 SIMD 化。

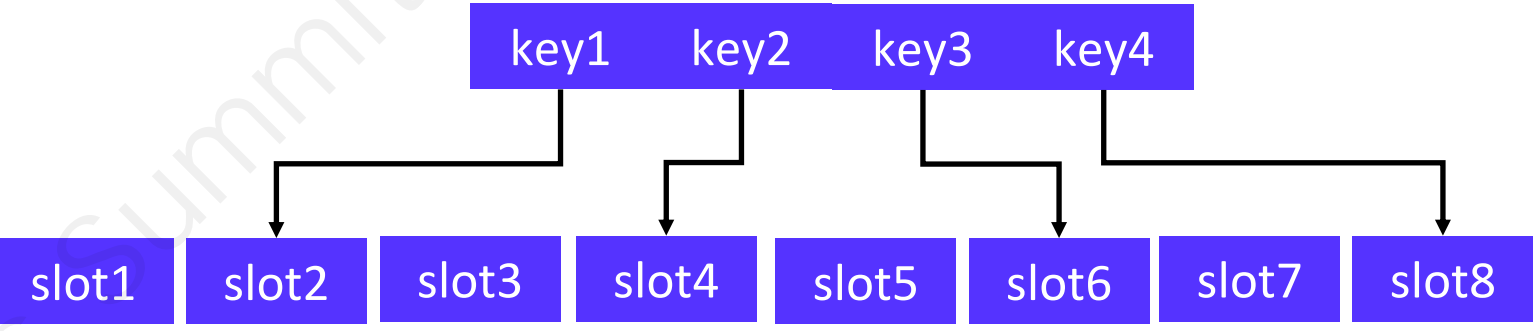


ARM架构向量化问题：

- 1. 实现效率低，ARM 的向量化实现通过对X86向量化代码的语义翻译，未考虑到 ARM 指令的特征
- 2. 未结合 ARM 下一代 SVE 指令
- 3. 部分第三方依赖包向量化实现亦存在上述问题，导致出现负优化

ARM架构向量化优化方向：

- 1. ARM架构向量化实现优化，充分考虑 NEON/SVE 指令特点和约束，充分发挥指令性能
- 2. 第三方依赖包整改，通过性能分析工具，找出对性能影响较大的第三方依赖包，查看其是否需要进行向量化整改
- 3. 向量化 HashTable，实现多路并发 probe，加速常用算子 HashAgg 和 HashJoin 性能



安全增强加固



敏感信息泄露加固

- 部分catalog和Doris明文密码打印、传输。
- AKSK、keytab等认证敏感信息明文打印、传输。
- 第三方报DEBUG日志泄露敏感信息
- 原生UI密码打印



通信安全加固

- FE、BE部分数据传输通道非加密、Cipher非安全等
- 服务全网监听端口
- 非安全加密算法的引用BASE64/AES128等



第三方依赖漏洞加固

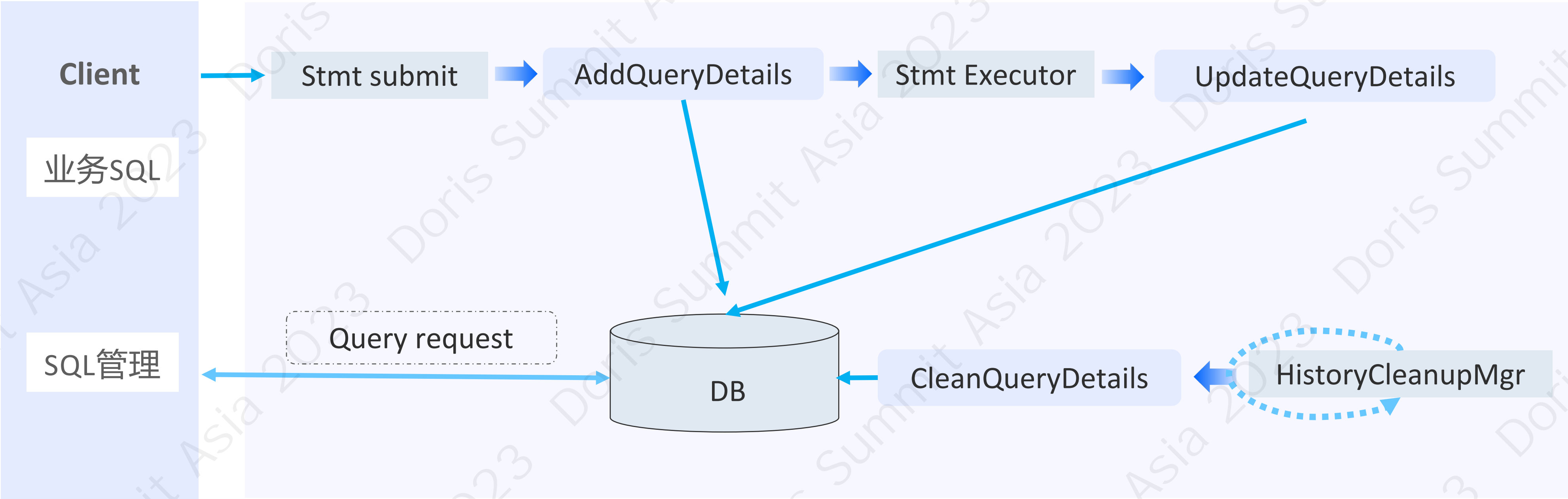
- 第三方存在安全漏洞的jar包的引用
- 第三方存在安全漏洞的BE二进制包的引用
- 版本过低的有安全问题的系统包引用



其他

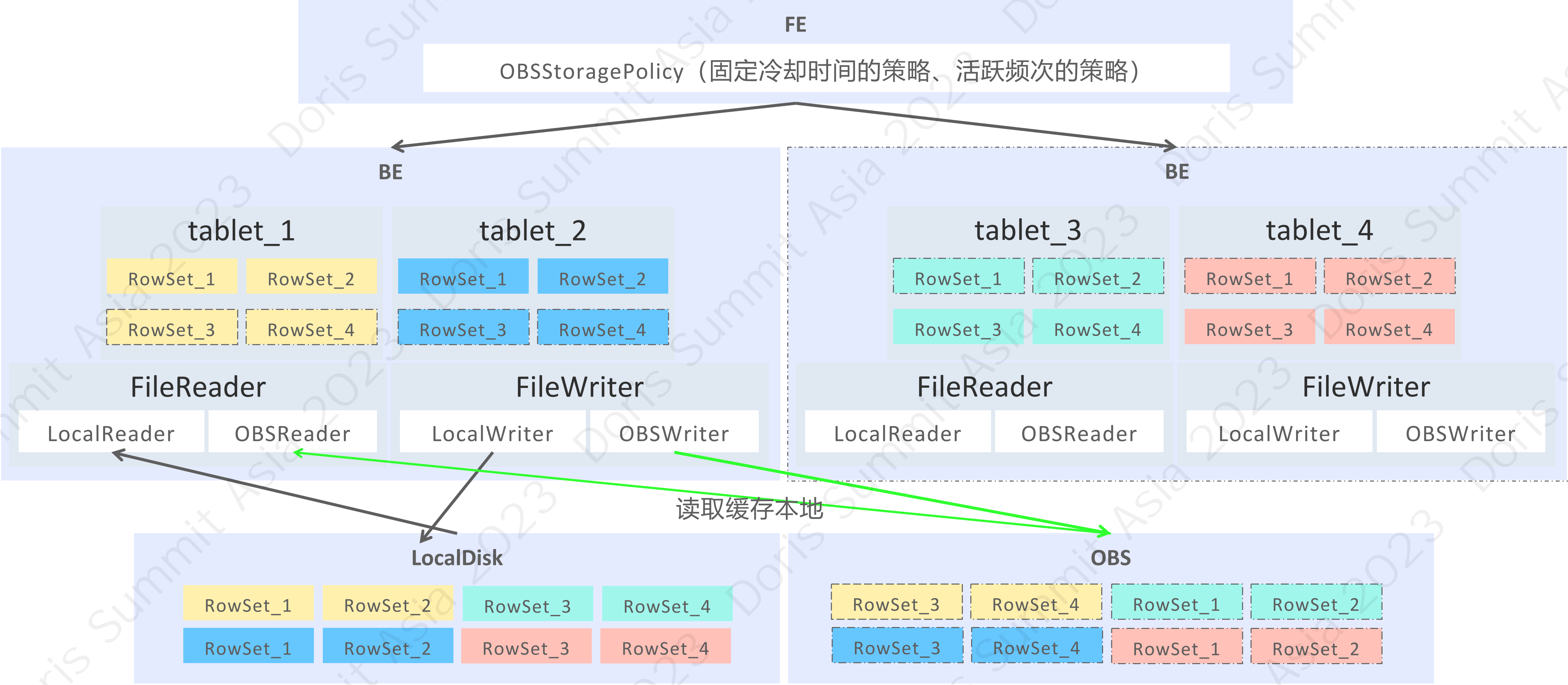
- Flink on Doris非安全传输、明文密码等
- Spark on Doris非安全传输、明文密码等
- Web登录暴力破解
- cookie默认非安全设置

新增慢查询管理



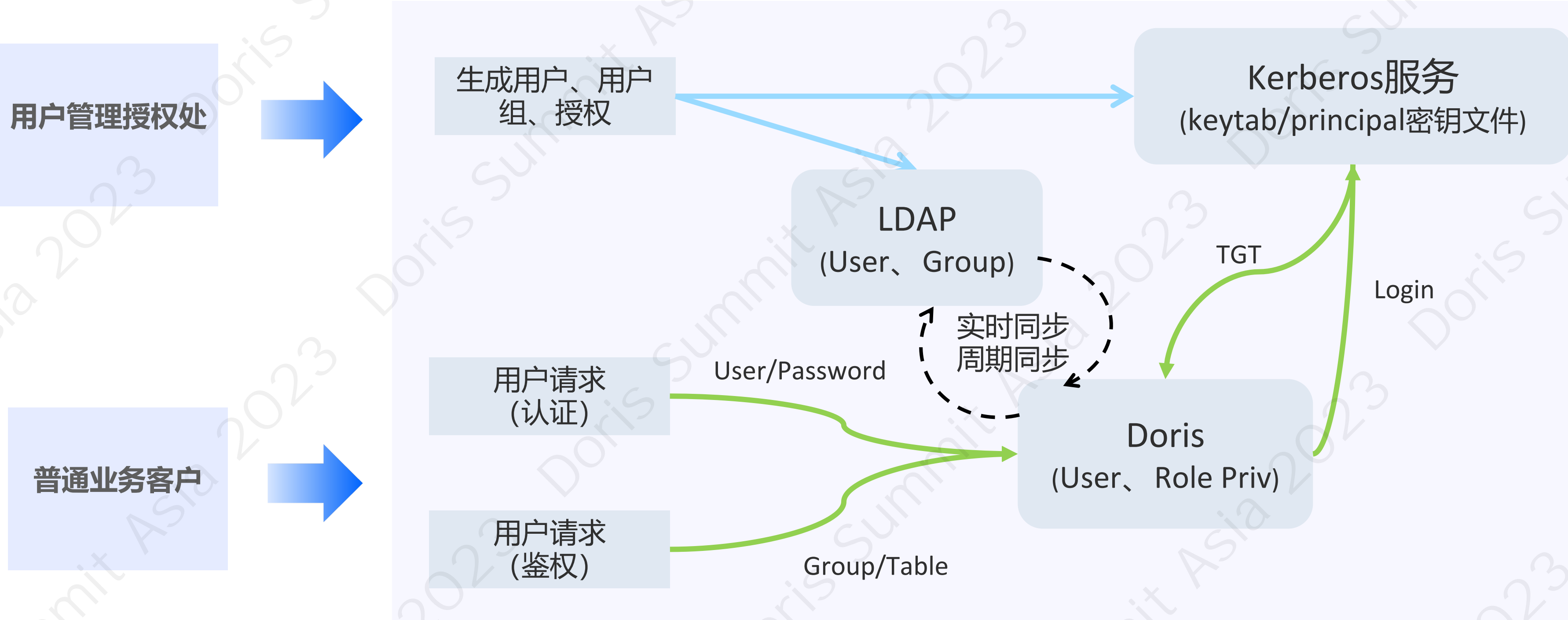
- 1. 慢查询任务是客户热频痛点场景，慢查询管理可以帮助用户很好的治理此类任务
- 2. SQL运行生命周期中，将时间状态等信息序列化QueryDetail对象存储到内置库的一张管理表中。
- 3. Doris客户端可以查询信息表，用于查杀或监控SQL状态的信息。
- 4. 周期任务清理过期时间和超长阈值存储的慢查询记录。

基于OBS的冷热分离增强



- 1. OBS SDK接口深度调用优化、AKSK/OBSA认证、Label细粒度权限，数据更高效安全
- 2. OBS冷数据读取后支持缓存本地、重置冷数据分层时间
- 3. OBS冷热分离策略支持固定冷却时间的策略以及根据数据活跃度设置冷却时间的策略

支持Kerberos用户认证

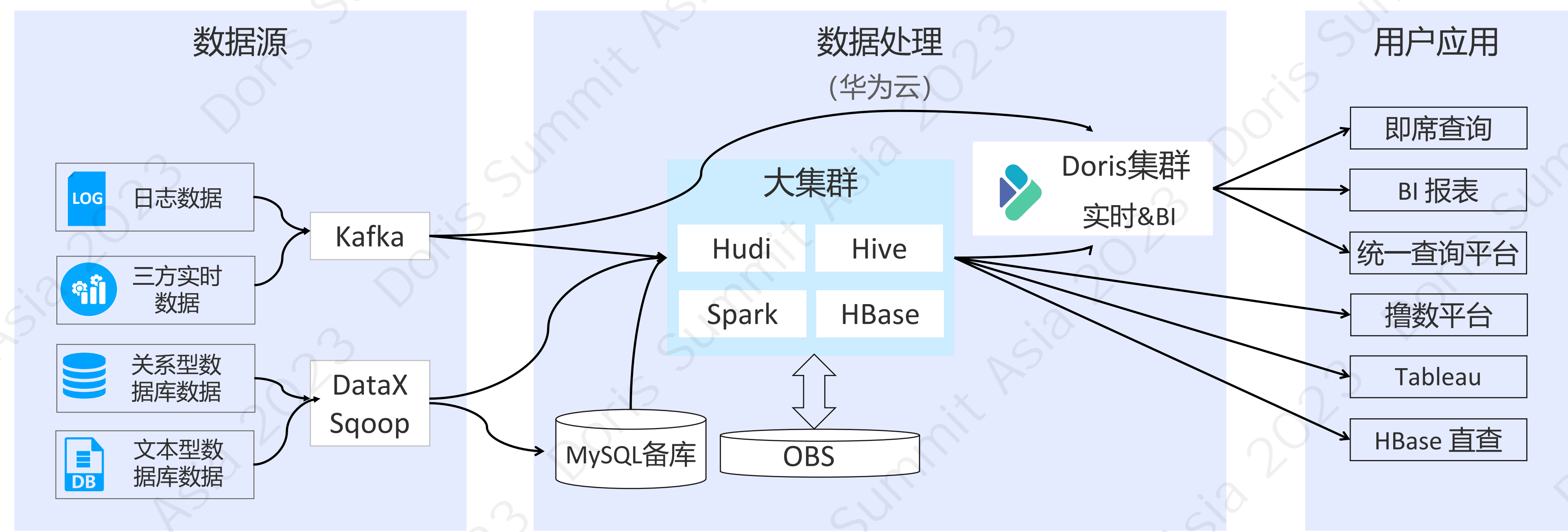


简述:

1. 客户存在三权分立的场景，对管理的统一用户嵌入打通Doris认证授权，将用户权限管理、业务使用、认证服务分离开
2. 生成用户、组时，用户和组存储在LDAP，同时生成Kerberos相同密码密钥的principal用户
3. 业务用户连接Doris时，Doris服务拿用户密码去登录连接Kerberos服务获取有效TGT票据，进行校验用户信息合法性
4. Doris开启和LDAP服务用户同步机制，也周期任务去刷新用户/组信息
5. LDAP中用户组映射到Doris的role，授权管理时是对用户组授权，鉴权时Doris中对role鉴权

3 客户案例分享

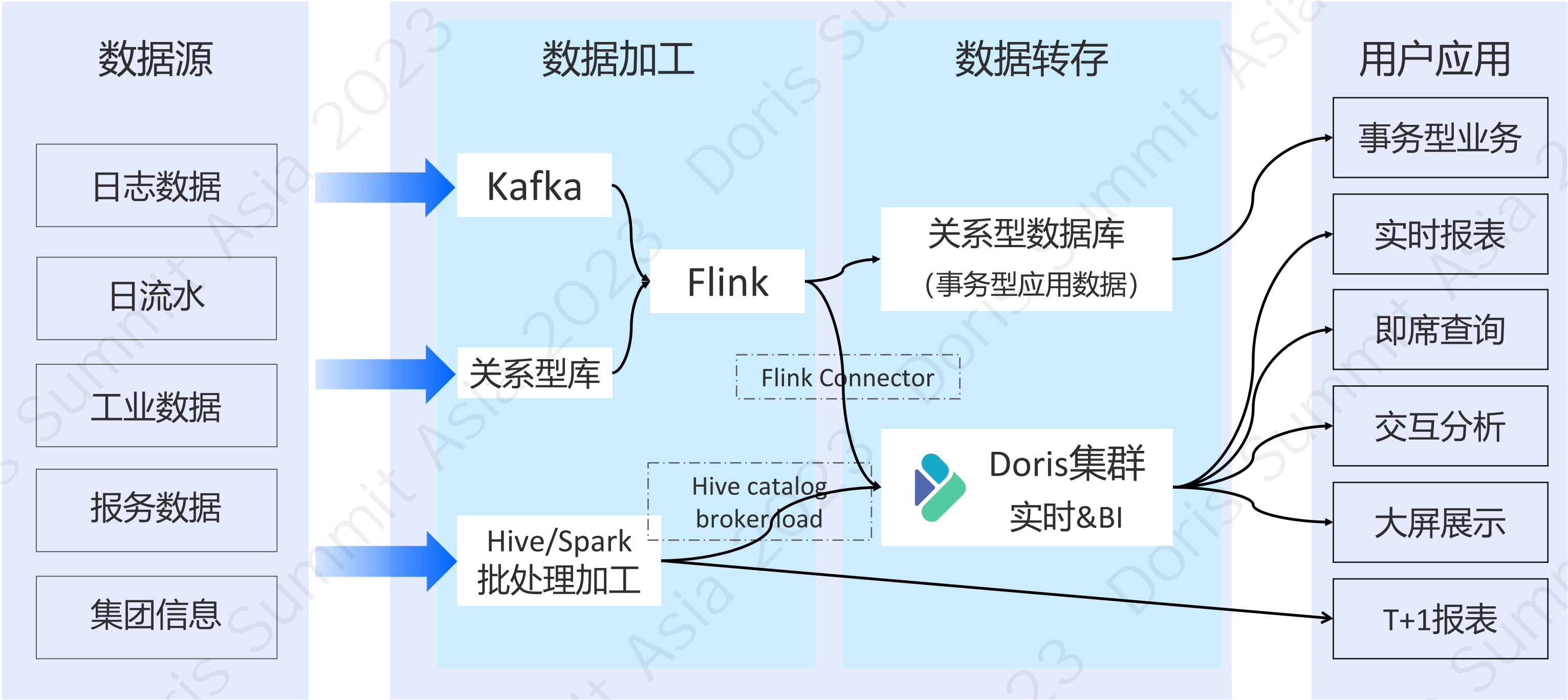
Doris在某互联网客户大数据平台的实践，性能提升200%+



场景说明:

1. 客户大数据平台原先是自建的Doris0.13版本，升级到Doris1.2.3，应用查询的性能大幅提升。
2. Doris在客户场景找那个满足对实时消息或日志等数据快速接入，高速查询后返回给应用，快速响应是客户一直选择Doris的关键考虑。
3. 客户部分存量或T+1的任务数据可以在Doris中做快速交互分析和即席查询，客户的报表有实时的也有T+1的场景。
4. 客户使用业界叫流行的BI报表工具，对接的组件业务不止上述业务流程，Doris具备很好的访问接口，用户无需额外对接开发，生态易用性高。

Doris在某制造业客户的应用实践，小量数据高效负责分析



场景特点:

- 1. 对数据精度要求较高，对多元结果时效性要求高
- 2. 数据批次多，批次量不是很大，单表最大几千万条数据，适用多表复杂查询分析，客户存在几十张表复杂分析
- 3. 存在并发大屏展示查询，小量表并发点查场景。

实践效果:

- 1. 引入Doris后将原本日结流水数据报表业务更改为实时报表，时间极大缩短效果显著。
- 2. 旧的业务系统一些基于数据库的多表join查询性能并不好，适用Doris后，业务增长一个SQL中到最大可跑有16张表的复杂查询，返回时间优于旧的系统。
- 3. 旧系统需要买多个数据库服务，新架构下使用一个Doris集群搭配一个小型关系型数据库即可完成，架构更简洁，客户也节省了很多成本。

4 未来的规划

规划在做的事务

- 基于 2.x 版本深耕竞争力增强、集市层主力演进
- 基于 Hudi 数据湖深度增强
- 基于 OBS 的存算分离增强、深度融合 OBS
- 打通各个云服务之间交互
- 存储过程的探索



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
- Apache Doris GitHub: github.com/apache/doris/

获取更多峰会资料:

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