爱玛集团基于 Apache Doris 的数据平台构建实践

胡磊

爱玛科技集团股份有限公司 项目经理

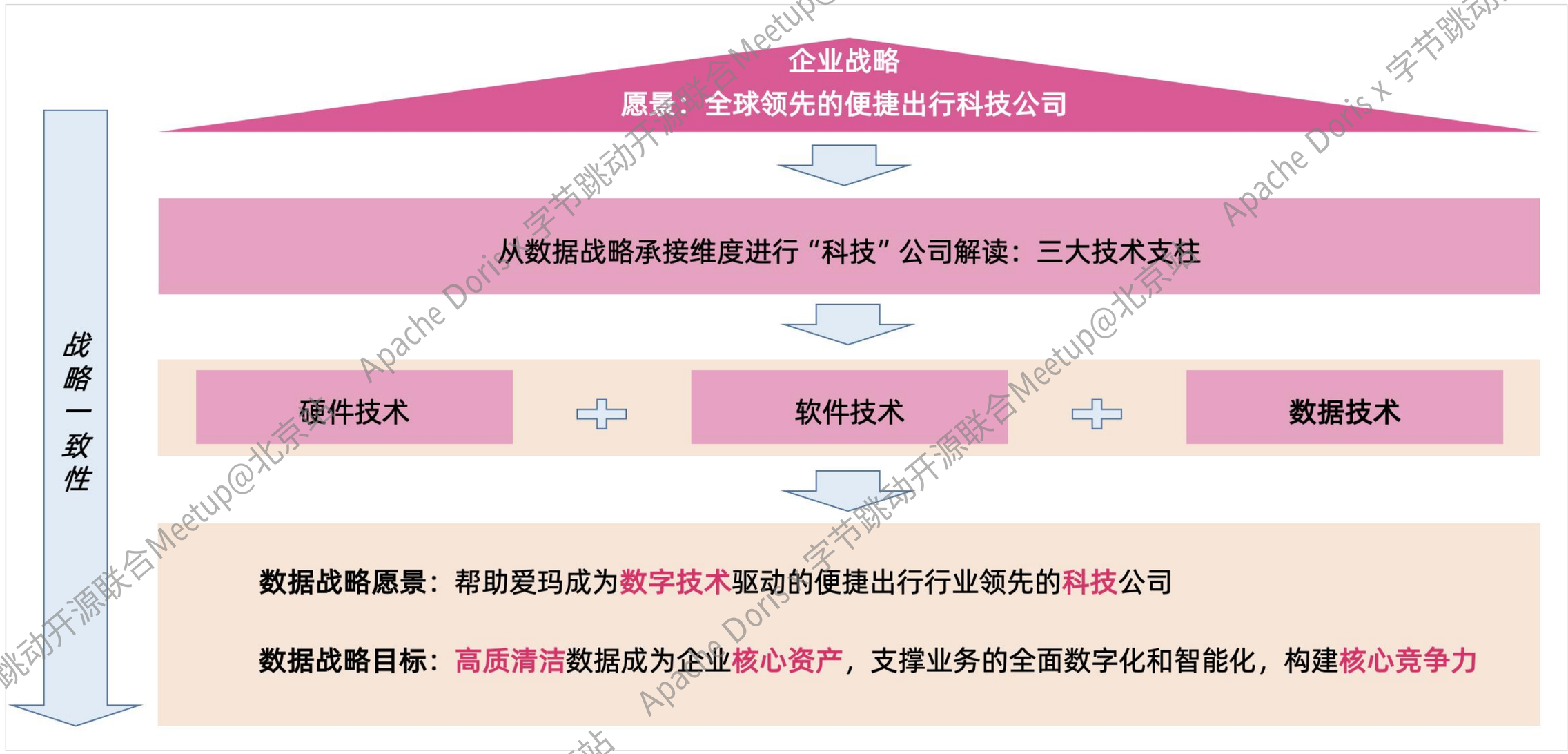
目录

主营业务

目录

数智化转型

数据战略



思考 & 目标

思考

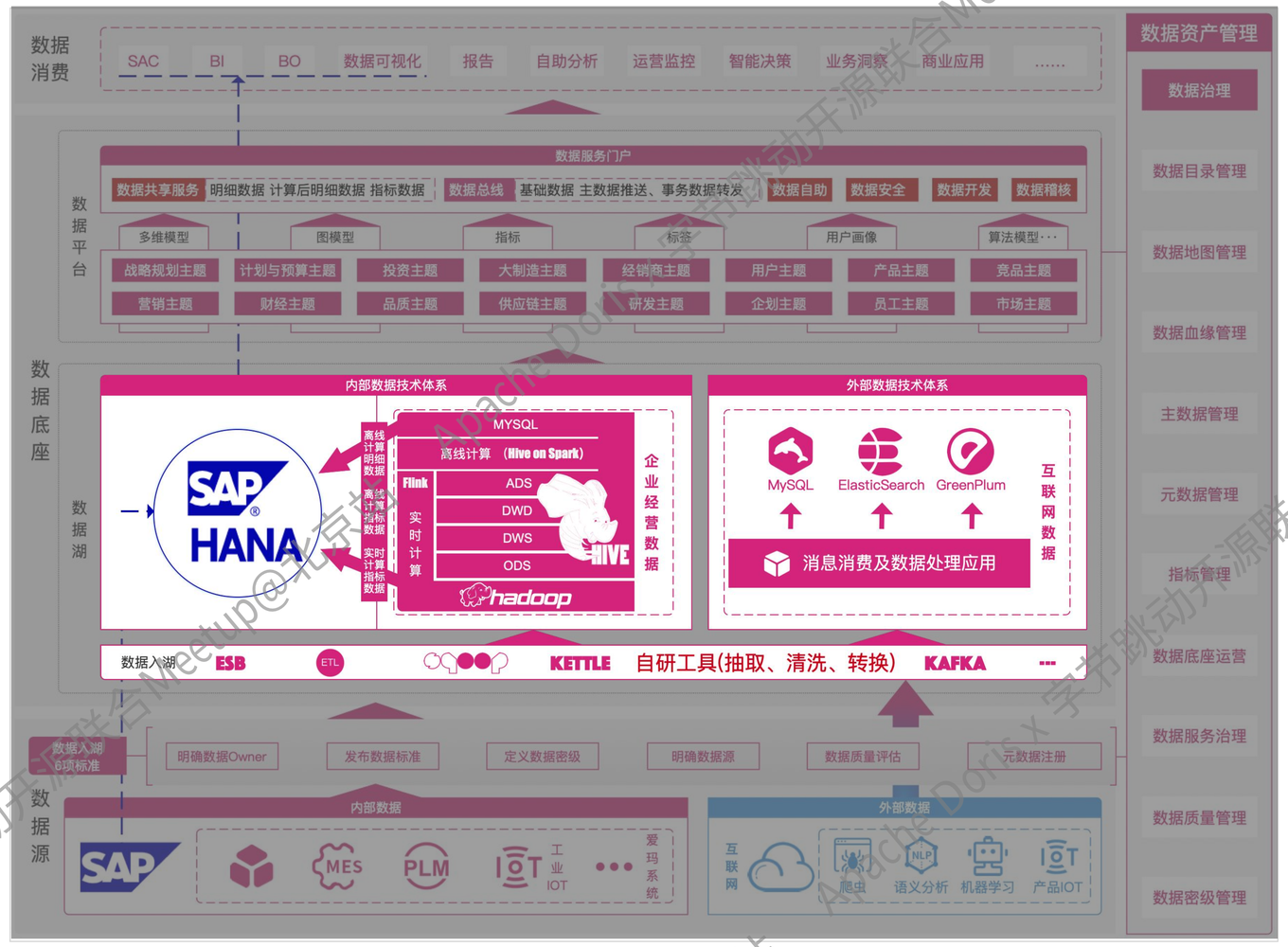
如何低成本快速构建数据技术平台体系，来支撑爱玛数智化转型战略落地？

目标

在数据接入、数据计算、数据管理、数据呈现方面建立合理的机制，向基、中、高层用户提供高效、高质的数据服务，辅助业务、经营、管理侧进行决策、变革。

目录

【数仓建立初期】整体架构 1.0

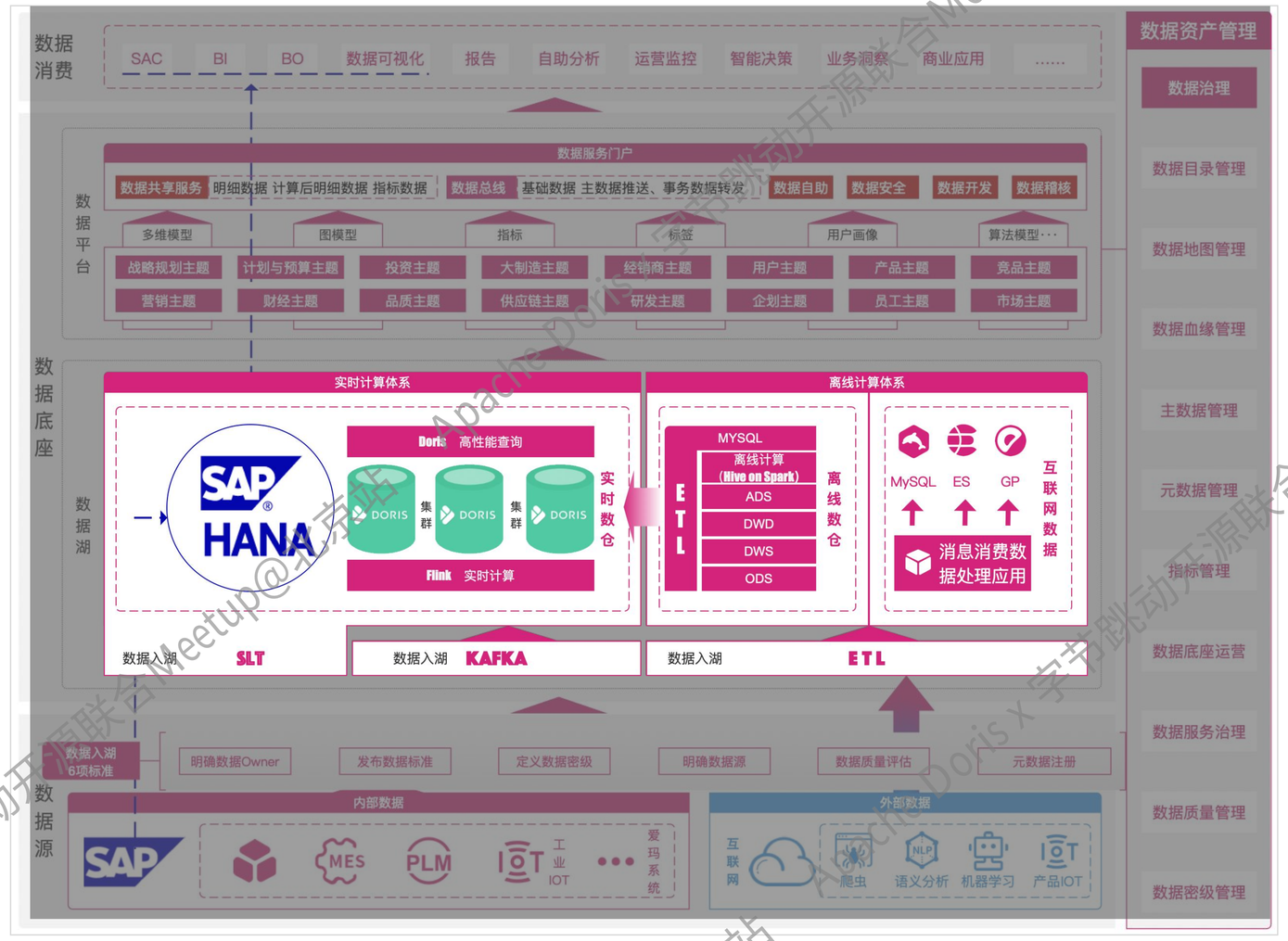


调度数量	调度时间
1.5千	6+ h
故障恢复时间	**最大数据呈现速度**
3+ h	8+ s

痛点

1.调度时间过长
2.故障恢复时间长
3.复杂可视化页面的数据展示过慢

【引入 Doris 实时数仓】整体架构 2.0



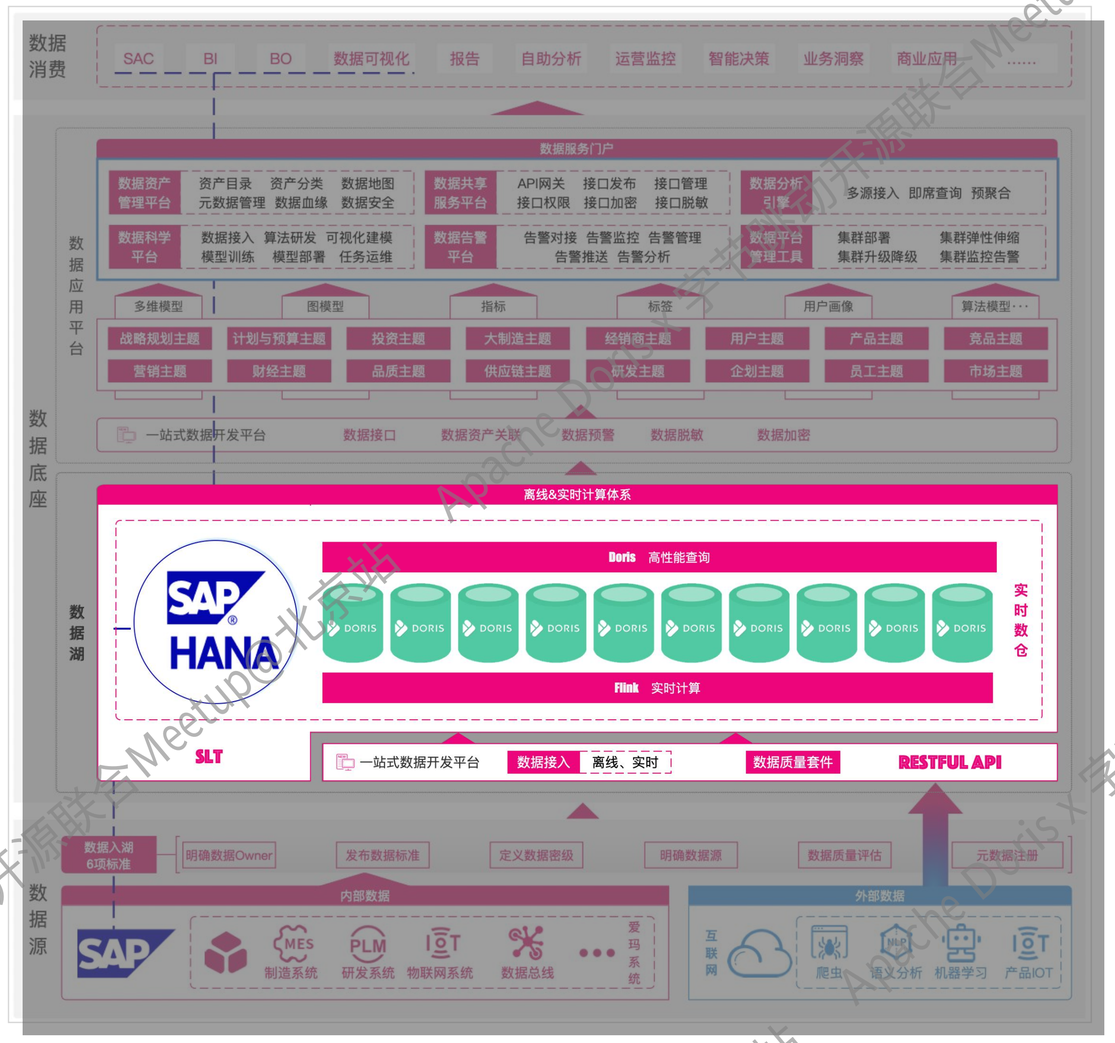
Doris Version：0.13 - 1.0

调度数量	调度时间
2 千	7+ h
故障恢复时间	**最大数据呈现速度**
3+ h	8 s

痛点

1.调度时间过长
2.故障恢复时间长
3.复杂可视化页面的数据展示过慢
4.不支持实时大屏

【数仓以 Doris 为基座】整体架构 3.0



Doris Version : 2.0.11

调度数量	调度时间
4 千	2+ h

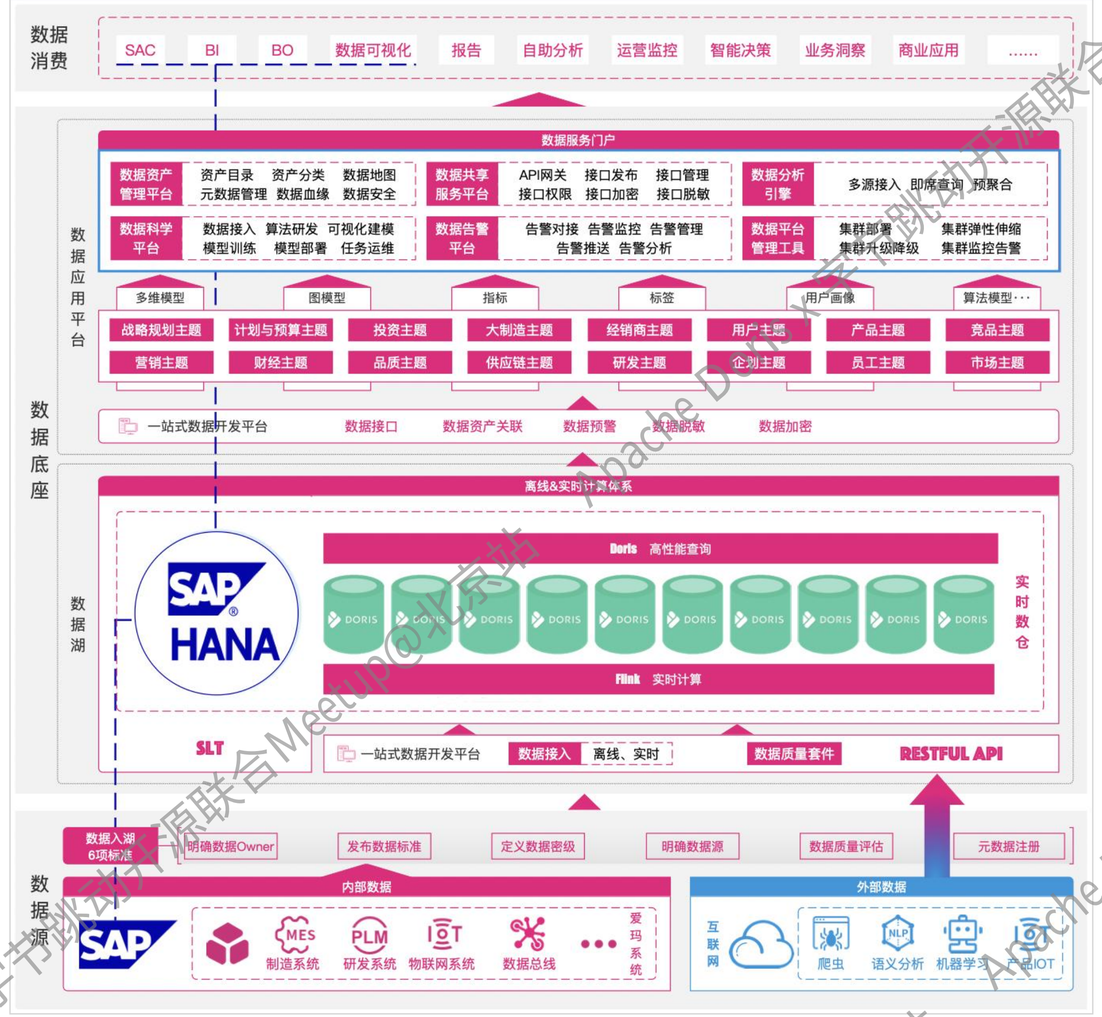
故障恢复时间	最大数据呈现速度
0.5 h	1+ s

痛点

1.调度时间过长
2.故障恢复时间长
3.复杂可视化页面的数据展示过慢
4.不支持实时大屏

目录

整体架构 3.0



平台组件&应用组成

数智门户

- 个性首页
- 可视化大屏
- 可视化移动端
- 数据分析报告
- 数据小品
- 自助分析

一站式数据管理平台

- 资产管理
- 数据标准
- 元数据
- 数据血缘
- 数据沿袭
- 数据安全
- 数据服务
- 数据质量

一站式数据开发平台 基于Dolphin

- 项目管理
- 任务管理
- 资源中心
- 监控中心
- 源中心
- 告警中心
- 巡检排班
- 接口管理

SAP HANA Ver:2.0 SP5

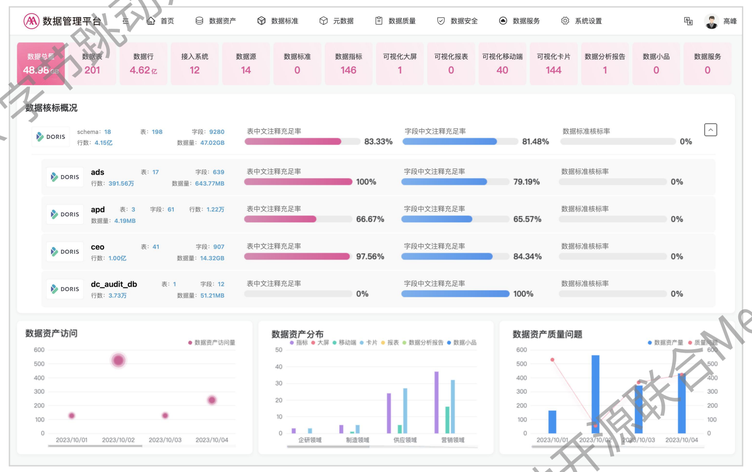
Apache Doris Ver:2.0.2

目标用户

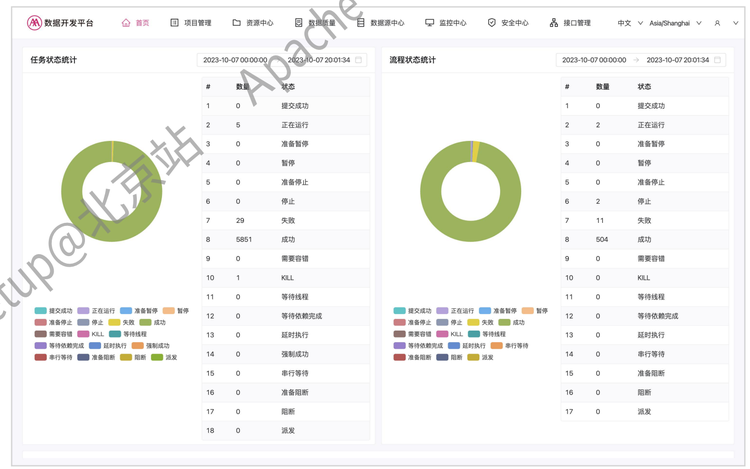
- 各层级用户
  数据分析师
- 数据管理员
  数据分析师
- 数据技术人员

「轻量级」数据平台

数智门户



一站式数据管理平台

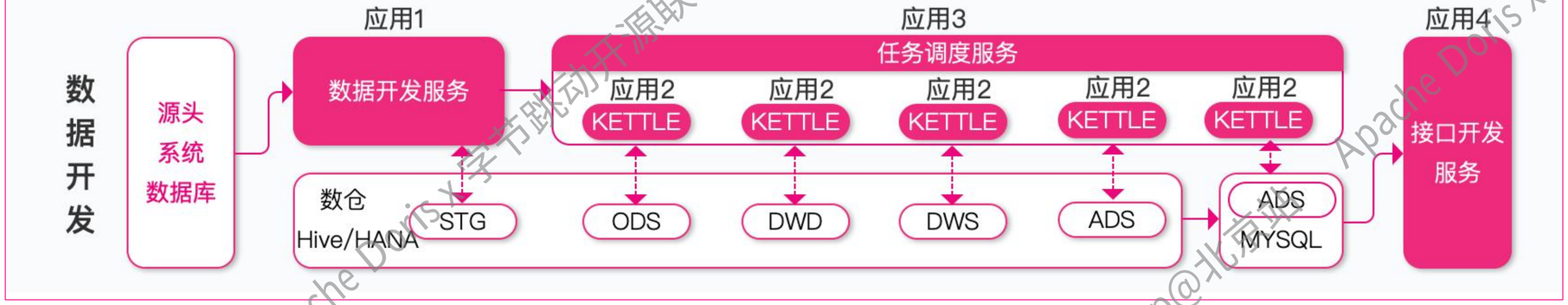


一站式数据开发平台

新数据计算基座 – Apache Doris

一站式数据开发平台

原数据开发过程



现数据开发过程

一站式数据开发平台

01 结合 Catalog，源头至 ODS 层快速接入

- 只需要选择源头库表和创建好的 Catalog，通过配置 CRON 即可定时同步 ODS 层数据

02 可配置化 API 管理

- 在海豚中增加接口配置的功能，开发人员只需选择 Doris 源，通过表或 SQL 配置的方式即可完成 API 的暴露

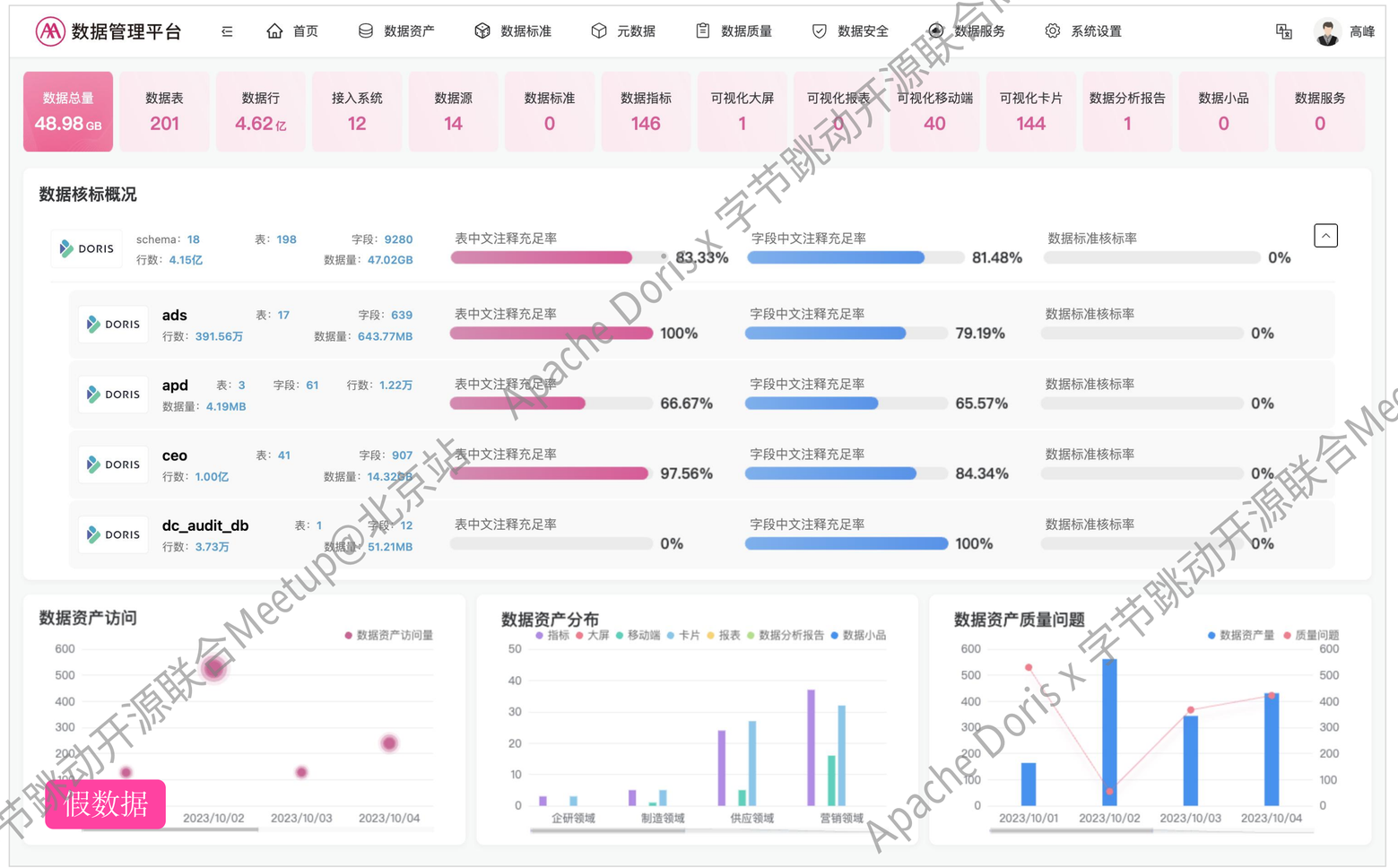
03 工作日巡检排班

- 通过配置爱玛工作日及巡检人员，可动态排班每日夜间数据调度巡检人员

04 接入阿里语音告警

- 调度任务出现异常时，及时电话语音告警对应巡检人员

一站式数据管理平台



01 数据标准化管理

- 依据数据治理形成的标准和规范，导入系统对数据仓进行扫描核标，及时发现问题，不断提升数据质量。

02 数据沿袭

- 对于已经形成的数据资产，例如大屏、报表、移动端、数据分析报告、数据小品、表、字段之间的全链路血缘追溯跟踪。

03 数据安全

- 定义数据资产密级，通过可配置化方式，当对外提供数据时，可进行脱敏或加密。

04 数据服务

- 提供多种对外提供数据的方式，并结合安全及权限配置，控制对外输出合规数据。

数智门户

01 个性化首页-千人千面

- 每位用户关注的大屏、报表、指标等内容都不一致，他们希望我们提供可配置的展示的方式为他们提供不同的首页信息展示内容。

02 数据权限的精准控制

- 不同职位的用户能查看对应权限的数据，保护每一个人的数据隐私。

03 全种类的数据资产呈现

- 用户都可以在门户上找到所有依据数据形成各类资产，比如大屏、报表、表、字段等，都通过门户的形式对外提供服务。

04 一站式高效数据获取方式

- 数据在线查看、数据在线下载、数据在线分析、接口在线申请及调用等。

05 个人数据工作台

- 数据的各类通知及告警，数据资产的申请、审批，数据资产的修正等等与数据类相关的核心工作均可在个人工作台完成。

一站式数据管理平台

通过建设一站式数据管理&开发平台，**统筹管理数据资产和数据开发**，而不是像以前较为分割的状态。

01 数仓规范升级

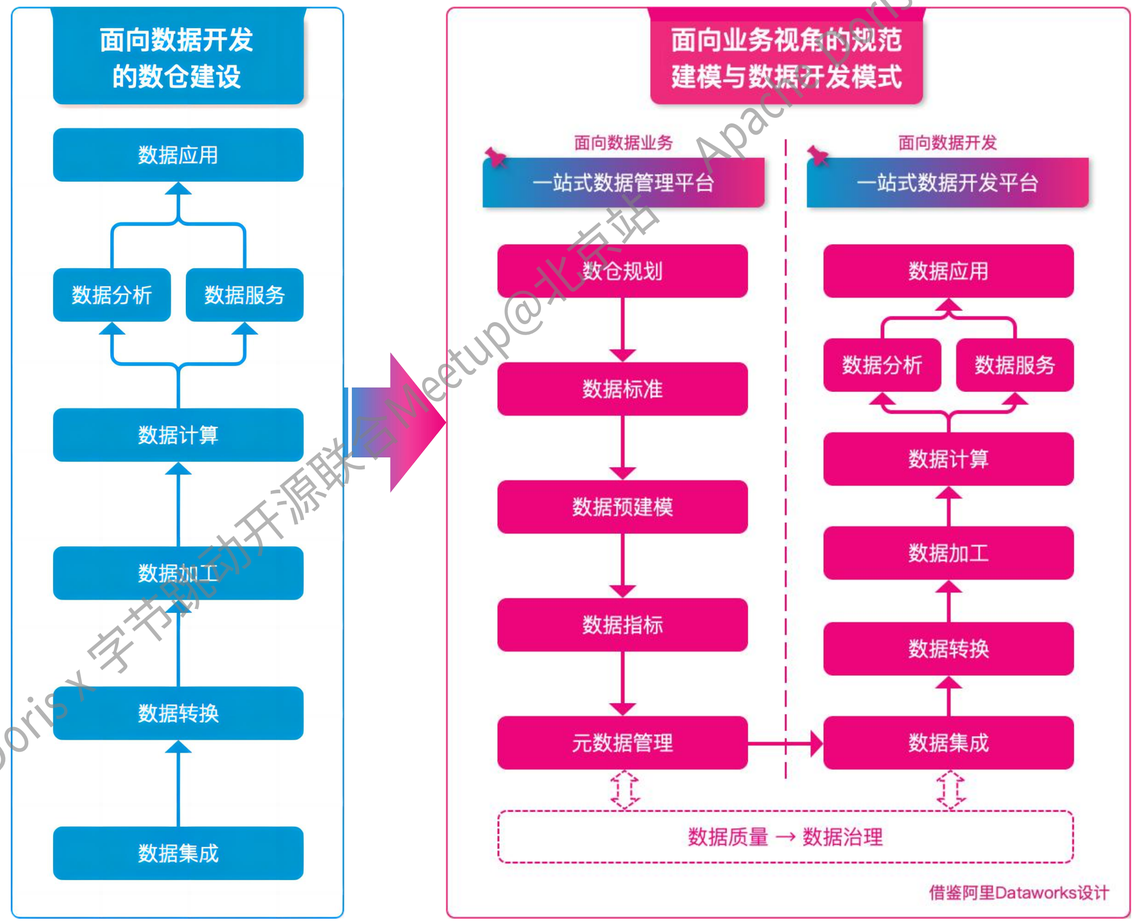
- 从业务领域维度，建立数据标准，建模过程全面执行
- 数仓分层规划与模型设计结合，规范数仓架构
- 模型设计与数据开发相融合，快速生成规范的表与代码

02 预构建数据模型

- 业务依据数据标准设计，并预构建数据模型
- 按照业务预构建数据模型进行开发
- 同时支持可视化 / Excel / 代码等多种建模方式

03 元数据管理

- 数据开发过程中，元数据管理分散至每个开发环节
- 数据开发完毕，元数据也建立完毕
- 半自动建立数据资产、数据血缘、数据地图



数据平台收益

较低的成本

目前在所有平台实现上的投入大概是3-4个人，全年人天投入约为1000。

数据全局整合

存放在多种数据数仓内的数据全部整合到 Doris，加快数据计算效率、提高故障恢复效率、加强数据易管理。

数据服务化

通过数据整合，用更高效的方式，对外提供数据。

业务新价值发现

业务侧根据获取的数据，结合数据技术提炼、总结、发现所需的，深度激活数据在业务侧的使用。

平台核心目标

低成本

较低的建设和使用成本

高交付效率

较高的数据需求开发交付效率

高计算效率

更高的数据加工、计算效率

高可用平台

更稳定、持续的输出能力

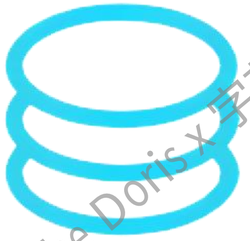
高数据质量

强化数据治理，得到更加清晰的数据

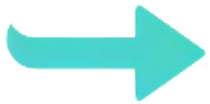
目录

查算分离

计算集群



通过 CCR 将计算数据同步至查询集群



查询集群

计算集群：用于跑批或流式计算的集群，配置更适用于计算的参数或优化方式。

查询集群：用于查询的集群，只存储对外展示的数据，尽量减少该集群中的数据冗余，增加更多适用于高并发查询的配置，和适用提高查询速率的索引或物化视图的功能。

大模型的接入

Thanks !