

# 浪潮海岳 inDataX 数据中台 基于 Apache Doris 的传统行业实时数仓建设实践

汪克忠 平台软件研发部 大数据技术架构师

# 目录

01 浪潮海岳数据中台介绍

02 inDataX 基于 Apache Doris 的传统行业实时数仓建设实践

03 未来思考与展望

## 分享嘉宾 — 浪潮数字企业技术有限公司



**汪克忠**

平台软件研发部 大数据技术架构师

# 浪潮海岳数据中台介绍



共创一流企业 建设数字中国

# 数字企业-浪潮海岳



国家重点研发计划“变革性技术关键科学问题”重点专项 拟立项的2017年度项目公示清单							International Standard	
序号	项目编号	项目名称	项目牵头承担单位	项目负责人	中央财政经费 (万元)	项目实施周期 (年)	ISO 5405:2024	
1	2017YFA0700100	甲烷和二氧化碳催化转化及 外场耦合化学键精准重构	清华大学	王训	2810.00	5	<div>Audit data collection extension — Government regulated financial reports and payroll</div> <div>Edition 1 2024-05</div> <div>Reference Number ISO 5405:2024</div> <div>© ISO 2024</div>	
2	2017YFA0700200	微波毫米波数字编码和现场 可编程超材料理论体系 与关键技术	东南大学	崔铁军	2674.00	5		
3	2017YFA0700300	工业园区多能流综合管控与 协同优化	大连理工大学	赵曙	2660.00	5		
4	2017YFA0700400	完整肝脏三维结构与功能信 息的精准介观测量	华中科技大学	张智红	2776.00	5		
5	2017YFA0700500	人体器官芯片的精准介观测 量	东南大学	周忠泽	2811.00	5		
6	2017YFA0700600	面向智能制造的软件自动构 造	浪潮通用软件有限公 司	王兴山	2463.00	5		
7	2017YFA0700700	低温晶界及相界调控实现材 料强化的原理及演示验证	中国科学院金属研究 所	李秀艳	2689.00	5		
8	2017YFA0700800	下一代深度学习理论、方法 与关键技术	西安交通大学	孙刚	2722.00	5		
9	2017YFA0700900	深度神经网络处理器的新原 理、新结构和新方法	上海寒武纪信息科技 有限公司	陈天石	2671.00	5		
10	2017YFA0701000	面向生物医学应用研究的 新型太赫兹辐射源	电子科技大学	方广有	2691.00	5		
11	2017YFA0701100	仿生假肢手感知与控制的神经 信息解码和交互技术研究	国家康复辅具研究中 心	雷宁	2830.00	5		
12	2017YFA0701200	基于自由曲面的共体光学系 统的纳米精度制造基础研究	天津大学	赵继	2812.00	5		
13	2017YFA0701300	人工视觉系统中的基础科学 问题和变革技术	南京大学	蒋锡群	2617.00	5		



## 企业 ERP 产品市场领先

中国 aPaaS 市场竞争力第一  
装备制造 MES 解决方案市场第一  
首创“财务云”概念

## 技术研发实力国家认可

承担国家核高基、863 研发计划，国际先进  
国际标准 3 项、国家标准 23 项  
4 个省部级重点实验室、创新平台

## 创新理念行业先进

自研国内首个低代码开源模型 UBML  
入选 Forrester 国内首份低代码报告  
有效发明专利 417 项，登记软著 713 项



# 浪潮海岳 inDataX

## 企业采存治算用一体化数智平台

提供“采、存、治、算、用”的数据全生命周期管理工具及数据管理解决方案，打造五大核心数据能力，助力企业实现内外部数据高效汇聚，满足企业全方位、多类型数据采集需求，快速变现数据资产价值

**数据应用**

- 灵活数据共享、多维度数据门户
- 多维分析、酷炫大屏呈现，chatBI

**数据计算**

- 机器学习、场景化建模
- 组件化、拖拽式算法模型创建

**数据治理**

- 全链路数据加工、全周期数据治理
- 低代码大数据开发平台

**数据存储**

- 湖仓架构、冷热交换
- 逻辑集中、物理分散弹性架构

**数据采集**

- 实时采集、离线采集
- 结构化、半结构化、非结构化数据

### 大规模数据采集和存储能力

某矿产集团：数据日采集量10亿条，36GB、时序总数据量18TB

### 安全可控的数据共享能力

某省国资委：1.5亿条数据，委内13部门共享

### 百优案例检验的数据治理能力

某动力央企：52条规则、10万条数据问题

三家客户：入选数据治理全国百优

### 丰富多样的智能决策能力

某水务、某能源：供水量预测、井下设备故障预警(省部级奖项)

某省国资委：企业健康度评价模型

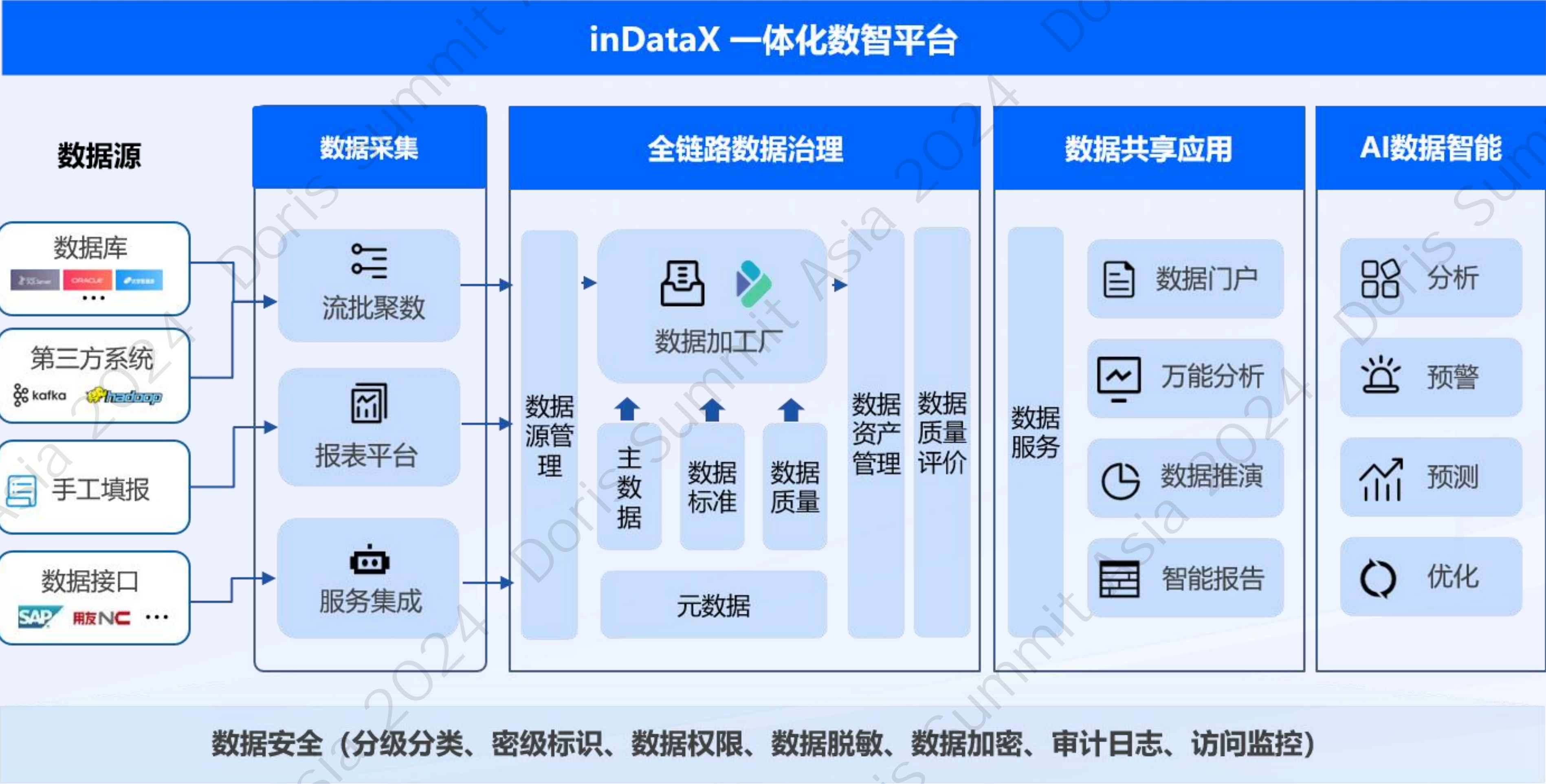
### 智能化数据呈现





# 浪潮海岳inDataX

## 企业采存治算用一体化数智平台



# inDataX基于Doris 的传统行业实时数仓建设实践



# 传统行业实时数仓建设的难点

## 线下业务流程多

业务流程审批过程多，线下审批多，流程慢  
业务流程涉及的权责部门多  
问题矫正周期长、依赖自上而下推动

## 信息化进程慢

团队技术能力较弱，技术能力欠缺  
信息化起步晚、转型慢，新技术引进较为谨慎  
技术架构规划弱、治理能力弱，运维能力不足  
信创要求高，技术栈多样化，硬件、数据库种类多  
技术安全需求：不联网离线安装部署要求多



## 组织架构和权责管理严

影响面大，涉及国计民生、国家安全，权责管控严格  
条块化管理架构，行政上级和业务上级双重管理  
组织架构分层明显、层级划分细

## 应用和数据架构复杂

不同时期不同厂商建设，数据口径不一致  
基础设施较弱，硬件、网络资源少，性能相对不足  
行业专业化知识多，数据指标计算逻辑复杂  
流量低，查询 QPS 低，RT 要求低  
场景复杂，设备、系统种类多，数据种类多、数据存储、共享方式不一致

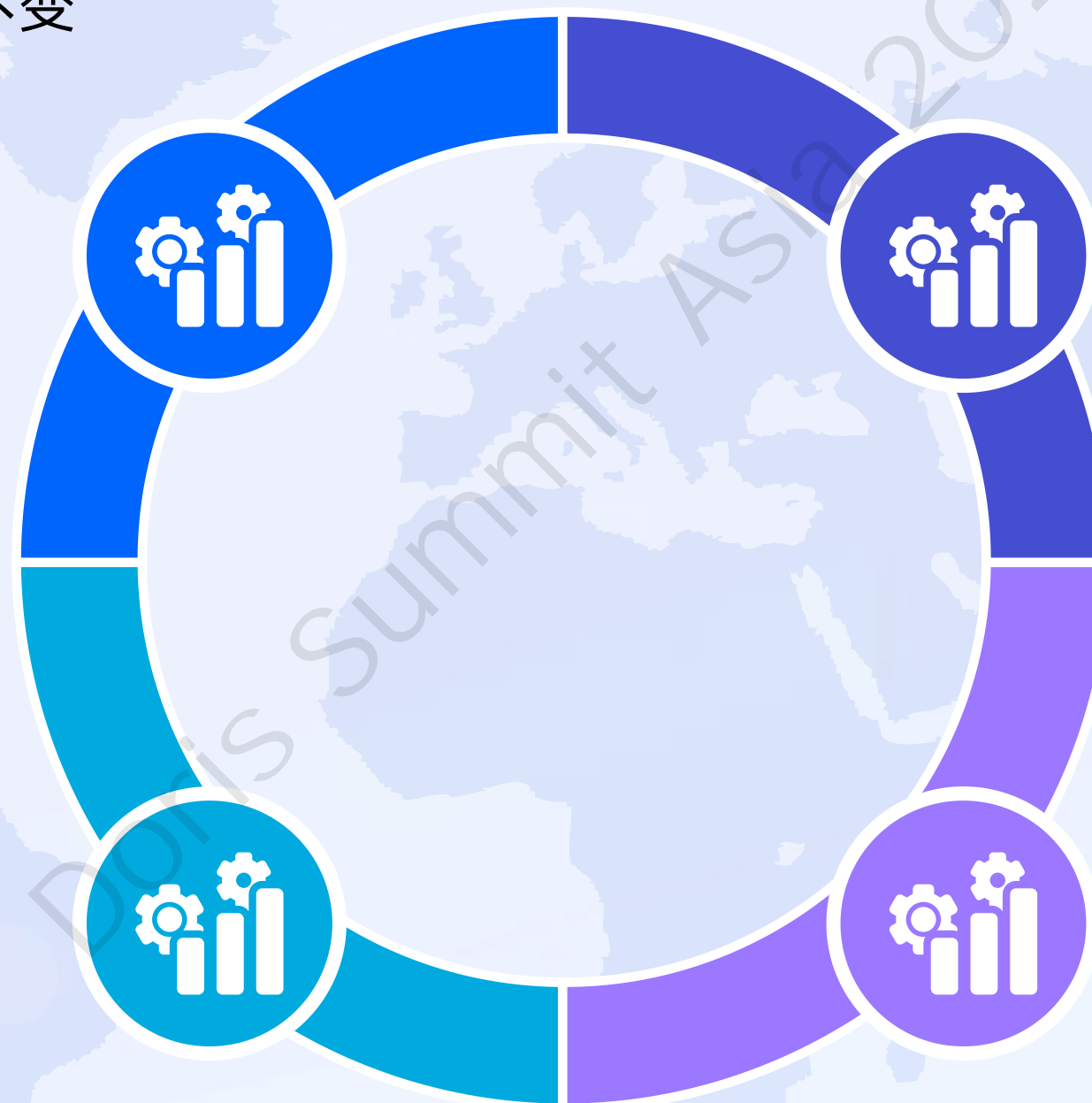
# 为什么选择 Doris

## 功能契合度高

联邦查询能力，直连，保持企业现有数据架构和设施不变  
数据种类支持多，支持半结构化、文件分析  
高维数据分析，适合专业化部门多(维度多)的场景  
运维相较于 CK 等简单便捷

## 架构栈兼容性强

提供便捷的 Multi Catalog 兼容已有库表  
兼容 MySQL 协议，系统切换方便  
多种数据格式转储方式便捷高效  
支持信创服务器、CPU 架构



## 技术栈自主可控

依赖 Java 和 C++，技术栈自主可控，不受政经因素影响  
开源代码可根据不同信创场景进行适应性改造，国产化支撑能力  
可按照国产 ARM 硬件进行自主编译、离线安装，适合不联网的高安全需求企业

## 数据分析性能强悍

基于 MPP 架构超强的数据查询分析能力  
允许用户随意即席查询，响应快，不怕折腾  
向量化执行引擎、物化视图等多种内置优化策略



# inDataX 实时数仓建设案例一：某能源集团数据治理

## 系统异构化严重、分散，链路长

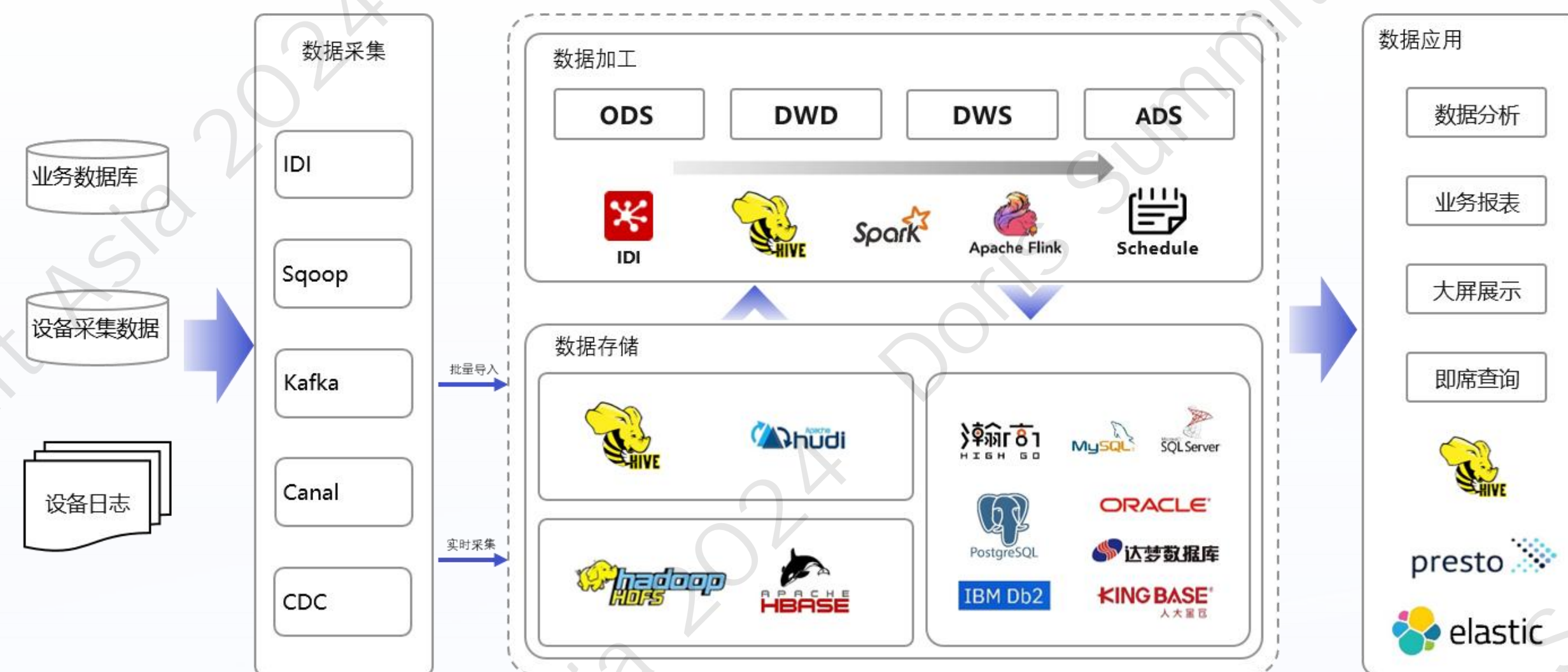
数据架构复杂，工具多，维护难，故障多  
原始指标简单，数仓数据加工链路长  
业务系统由不同厂商构建

## 数据来源复杂

有边缘端(煤矿)设备数据(MySQL、XML文件)、有集团端数据仓库(HDFS、Hive)、有分公司数据(集团同城数据，瀚高)

## 数据计算需求复杂，查询入口多

业务专业化强，公式复杂，计算逻辑复杂  
数据质量依赖 Hive SQL，执行效率慢  
在 Hive、ES、MySQL 多处提供查询，散乱



## 客户需求

- 兼容已有 Hadoop 数仓
- 集成边缘端数据、分公司数据
- 提供即席查询能力，统一入口
- 复杂 SQL 分析能力(计算复杂逻辑)

# inDataX 实时数仓建设案例一：某能源集团数据治理

## 统一查询计算中心

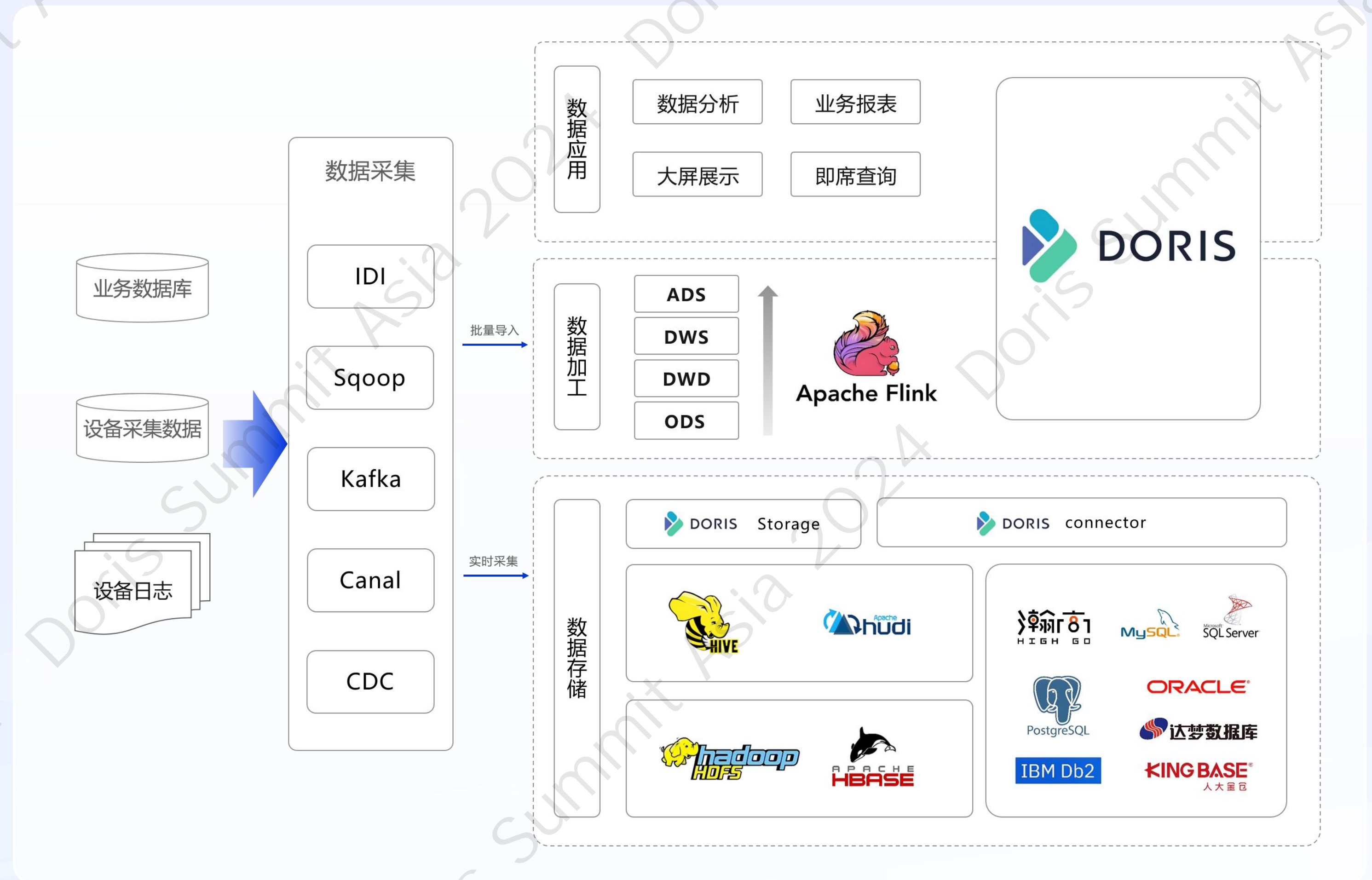
新建 Doris 集群作为统一分析和查询中心  
直连(只读)集团 Hadoop 和子公司异构数据库,  
不改变现有数据架构  
数据质量功能迁移至 Doris 跑批实现

## (准)实时同步边缘数据

采用 FlinkCDC 同步边缘数据至集团侧  
Doris  
ETL 解析边缘 XML 文件数据同步至集团侧

## Catalog连接子公司库

Catalog 连接子公司库进行授权数据访问





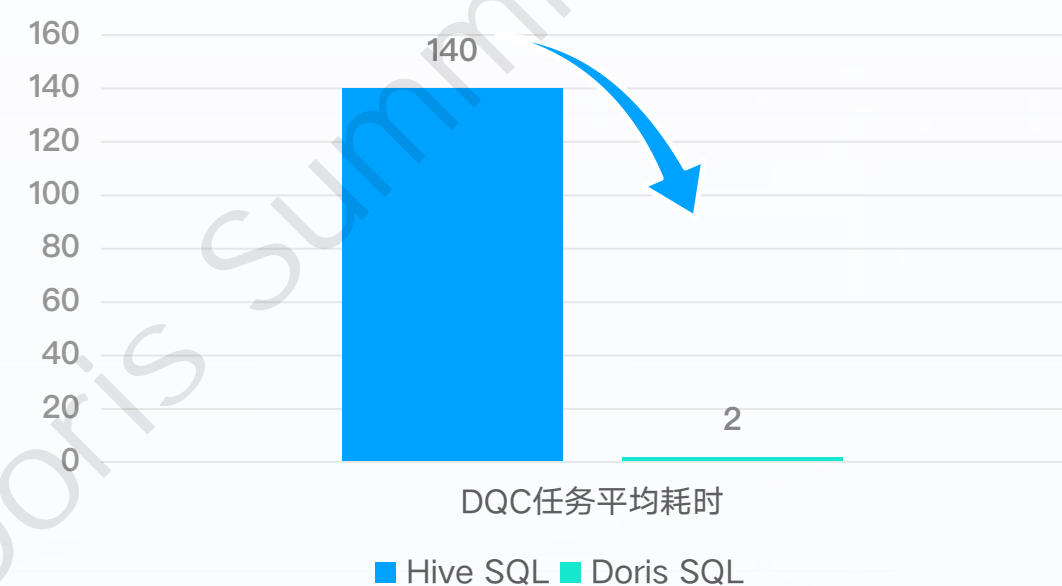
# inDataX实时数仓建设案例一：某能源集团数据治理

## 架构优化效果

- Catalog 读取原有架构，开闭式架构兼容；
- Catalog 替换ETL 链路，架构更加简洁；
- 替换部分老架构组件(Presto 等)；
- 查询入口全部由 Doris 承担、数仓和即席查询库统一。

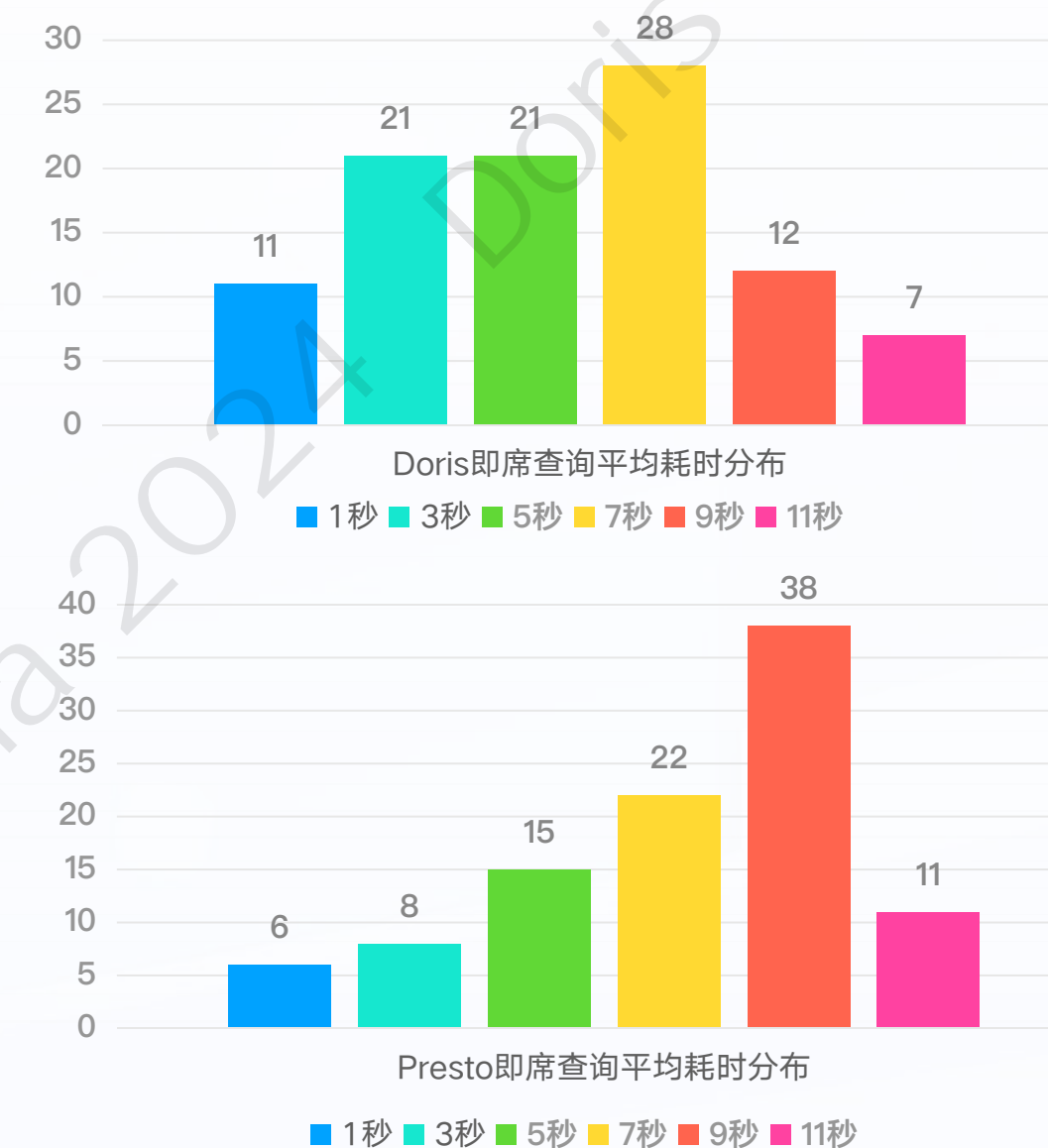
## 数据质量任务跑批性能提升

- 原 DQC 由 Hive SQL 承担，改造后由 Doris SQL 承担，性能平均提升 98%。



## 即席查询响应提升

- 即席查询响应时间分布对比 (100次)：



# inDataX 实时数仓建设案例二：某港口集团数仓升级

## 孤岛系统化，隔离严重

各业务系统隔离，未打通  
不同厂商建设、版本不统一

## 数据链路长

组织架构臃肿、业务流程冗长、审批慢，数  
据在各系统中一致性差，T+N离线同步

## Flink 实时采集、计算

使用 Flink 做多系统 join，大窗口 join  
使用 RDBMS 做数仓供报表查询

业务系统A



业务系统B



业务系统C



业务系统D



Multi CataLog + IDI Connector + CDC

inDataX  
数据中台

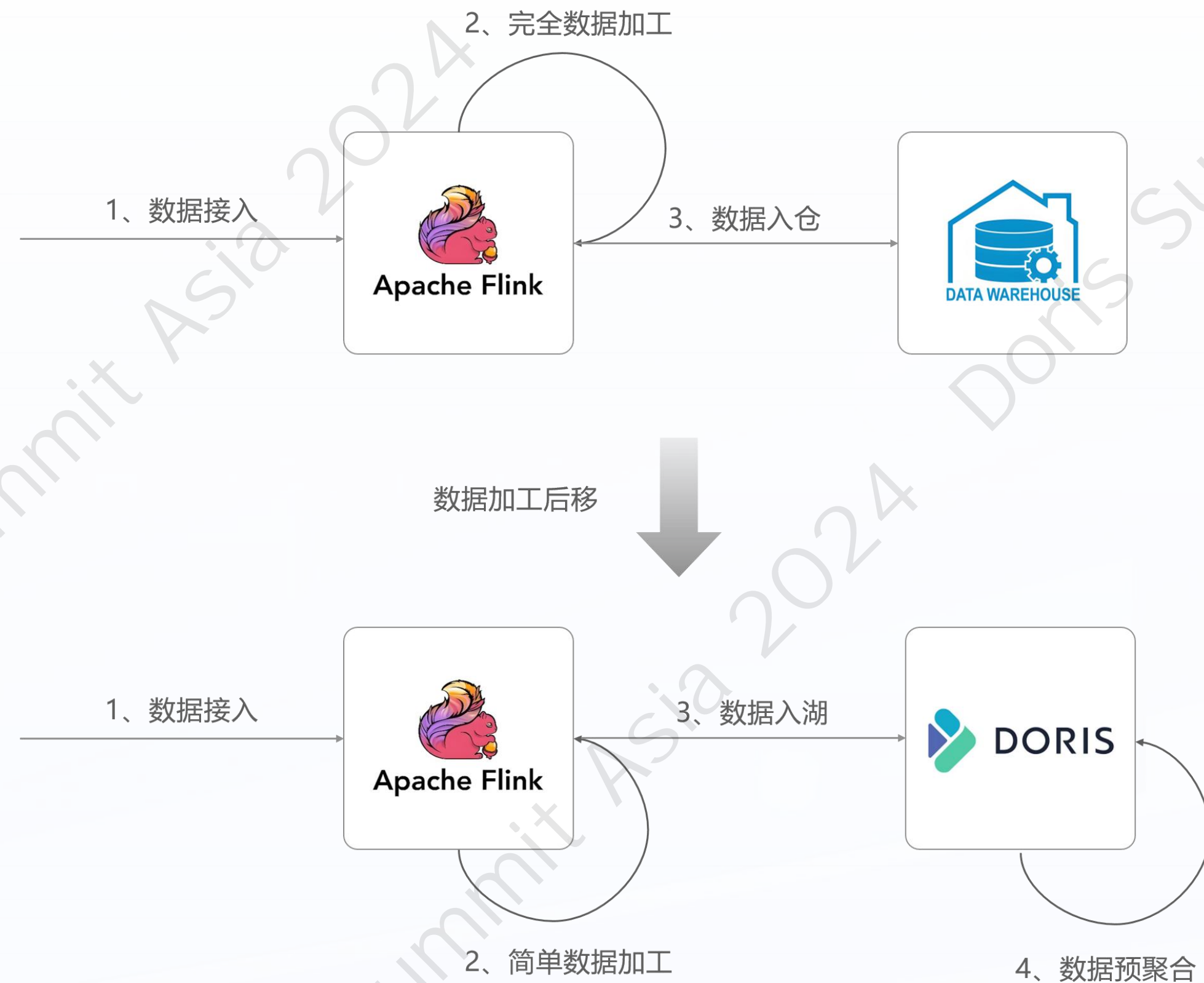


- Flink 流 join 为兼容不同系统时间差，大 TTL
- GC 频繁，Flink 节点宕机概率较高
- GC 影响了数据流的传输，影响全链路时效性
- RDBMS 无法支撑复杂报表逻辑，报表任务卡死频繁，报表应用 OOM 频繁
- 链路 SLA 低，平均 2次/月故障率

# inDataX 实时数仓建设案例二：某港口集团数仓升级

## 引入 Doris，改善架构

- Doris 表与业务系统表映射
- Flink 做一对一表数据同步
- 利用 Doris Aggregate Model 实现数据仓内实时聚合
- Doris 承担报表查询 SQL 执行
- 固定报表业务改造：SQL 由 Oracle 查询形式改为 Doris SQL

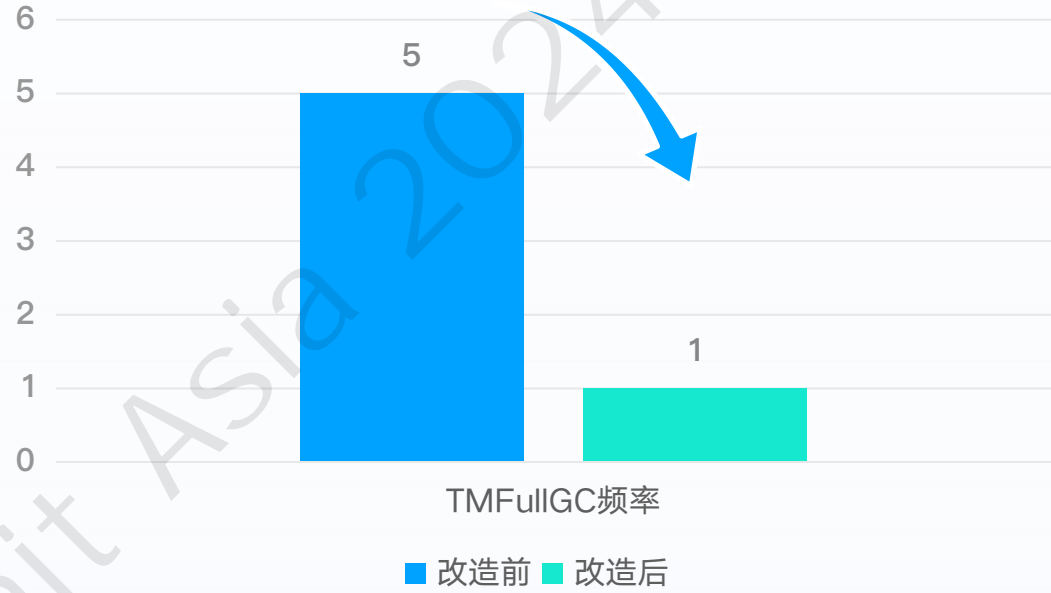




# inDataX 实时数仓建设案例二：某港口集团数仓升级

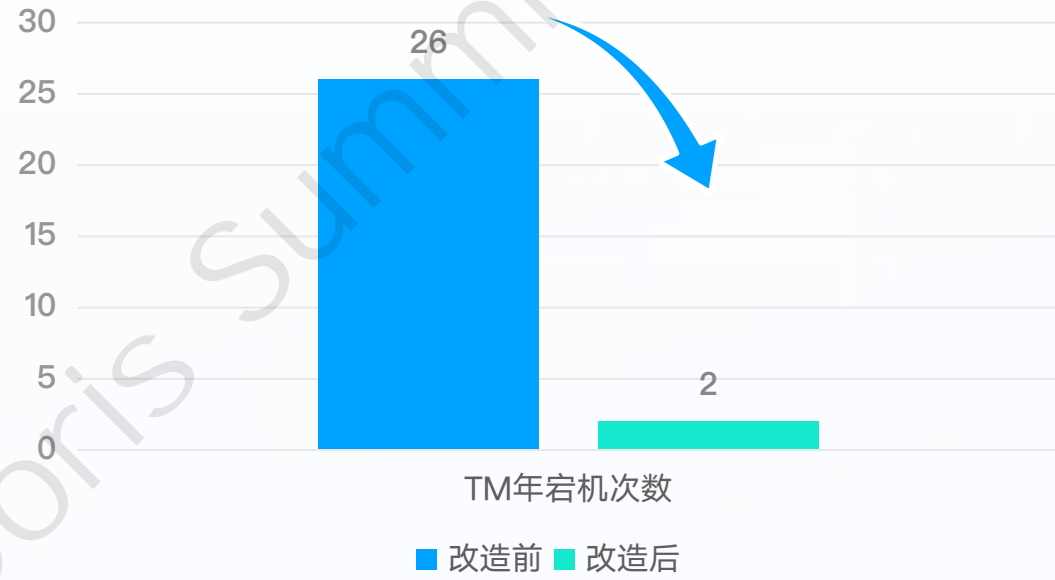
## Flink 集群 FullGC 频率

- FullGC 频率下降80%。



## Flink 集群 TM 宕机频率

- OOM 引起的 taskmanager 节点宕机频率降低92%。



## 全链路数据时延

- 全链路(从业务发生到大屏数据生效平均时长)时效性提高60%。





# inDataX 实时数仓建设案例二：某港口集团数仓升级

```
SELECT
  helpdesk_ticket.id,
  -- 50+其他字段
FROM
  helpdesk_ticket
LEFT JOIN
  (SELECT res_users.id, res_partner.name, res_users.login FROM res_users JOIN res_partner ON
  res_partner.id = res_users.partner_id) AS usertab
  ON usertab.id = helpdesk_ticket.User_id
LEFT JOIN
  helpdesk_ticket_type
  ON helpdesk_ticket_type.id = helpdesk_ticket.Ticket_type_id
LEFT JOIN
  helpdesk_stage
  ON helpdesk_stage.id = helpdesk_ticket.Stage_id
LEFT JOIN
  (SELECT mail_message.res_id, MIN(mail_tracking_value.new_value_datetime) AS firstPlanfixdate
  FROM mail_message LEFT JOIN mail_tracking_value ON mail_tracking_value.mail_message_id =
  mail_message.id LEFT JOIN helpdesk_ticket ON mail_message.res_id = helpdesk_ticket.id WHERE
  mail_tracking_value.field = 'plan_fix_date' GROUP BY mail_message.res_id) AS plan1fixdate
  ON plan1fixdate.res_id = helpdesk_ticket.id
WHERE ...
```

单表千万数据量级场景下多表 Join 查询

Doris VS MySQL

25 倍

查询速度提升

inDataX 基于 Flink + Doris 架构打造了实时数仓链路构建功能，在数据分析中做到多种数据库兼容，实时同步数据查询，将原数据库中的复杂 SQL 查询从 10 分钟级提升到秒级。

# 未来思考与展望

# 对 Doris 数据库的期望

## 数据库兼容

多种数据库 SQL 数据类型  
自动映射

## 方言支持

全面、稳定、可靠的方言转  
换能力

## 生态

多指令集国产环境兼容

# Thanks for Watching!