DORIS SUMMIT

Apache Doris 在区域医疗影像平台中的应用

国家健康医疗大数据（东部）中心

王建 大数据研发经理

目录

01

背景介绍

背景介绍

2016.10	获批试点	国家健康医疗大数据中心及产业园建设**国家级试点**
2019.03	专业团队	运管分离架构设计，常州国资与中国电子合资负责规划、建设、运营
2019.12	新型基建	江苏省卫生健康云（常州区域）试运行，数字化、国产化、数据服务同步推进
2020.01	试点服务	常州市“医疗废物服务（监管）系统建设试点、“基层医疗机构信息化提档升级省级试点”、“常州市医学影像云”等多个项目依托“云大脑”开发、服务、管理
2020	基地荣誉	长三角一体化联盟智慧城市应用示范基地；省、市网信办联合启动，建设**大数据和云服务安全保障试点；列入省发改委重大产业服务平台**
2021.01	全省覆盖	影像平台项目列入“大数据+产业链”三大省级大数据应用示范重点项目，由省领导挂钩联系
	数据治理	2021常州市健康医疗数据开放创新应用大赛，推动三医数据联动治理开放应用
2021.09	百日攻关	影像平台项目百日攻关，省属三级医疗机构、南京市三甲医院、宿迁第一人民医院以及常州全市公立医疗机构全面接入卫生健康云
2022	全省推进	影像平台项目计划年内覆盖全省70%公立医疗机构
2023	全省实现	实现全省100%公立医院以及部分私立医院接入影像平台
2024	数据服务	在常州市率先提供数据服务，实现全市‘无胶片化’

华东云计算基地

占地82.1亩，总建筑面积9万平方米，包括4栋云计算数据中心楼、1栋综合楼及1座**22万伏变电站**。基地存储基础扎实，拥有两个省级重点数据机房，可容纳**1.2万个机柜**，提供**16000PB的存储能力**。

数据特殊性

及时性要求高

- 影像的结构化数据及非结构化数据上传后，在临床需要进行及时的应用。但是因为影像数据复杂度高，需要多源数据进行关联并对多质量指标进行稽核，并进行预警、处理，保证不对临床的数据服务产生影响。

数据重复性高

- 影像数据存在大量的重传、补传的动作，对多场景的数据操作进行兼容，保证数据唯一性及可靠性。同时需要对上传记录进行回放。

数据采集点多

- 影像平台涉及医疗机构 2000 余个，每个医疗机构的状态以及条件千差万别，需要对各医疗机构的多种数据情况进行兼容。

行业特殊性

关联难度大

- 同一个检查的数据分批上云，时间差不确定。同时，单一放射检查的涉及的数据类型较多，需要对多数据体进行关键计算，JOIN 难度大。

指标口径多

- 以数据质量模块为例，需要对 27 个核心指标，共 140 余个质量指标进行监控。

数据服务场景多

- 需要对数据质量预警、数据质量看板、数据质量监控、质量代办、BI 大屏、报表以及数十个业务系统进行数据支撑。

02

架构演进

Hadoop体系特点

- **组件多**：因为 Hadoop 架构的特殊性，导致实现从采集、治理、存储到服务的全链路数据流程所需要的组件庞杂，在平台中，拥有超过 20 个开源的组件以及 30 余个自研的平台管理服务；
- **运维难度大**：组件多导致运维成本高，甚至单独一个组件的兜底要求也非常高；
- **部署成本高**：一套完整的 Hadoop 集群，需要的管理资源较多。在集群规模不够大的状态下，计算节点的边际成本较高；
- **较难对新场景进行兼容**：随着业务的发展，数据实时性的要求愈加的高，Hadoop（Hive）的体系，无法满足实时性的需求；
- **拓展性较差**：体系内的单一组件只面向单一的能力。面相新的业务需求，只能拓展新的组件进入集群。带来极高的维护成本。

场景痛点分析

业务痛点	改进方向
数据质量反馈周期长 T+1的反馈周期，无法及时反映整改措施的有效性，极大拉长了工作周期	**打造高效的大表交叉查询** 在对院端数据进行实时质量计算的同时，提供能将指定时间段的检查数据和存储日志数据进行交叉查询的能力，解决最关键的影像完整性检测问题。
缺乏实时监控能力 普通的流式计算难以参照历史数据，无法实现对应用数据监测的多维度分析。	**构建实时-历史数据比对能力** 在不仅对应用日志进行实时存储，也可以将实时日志按照日、周、月等较大时间维度进行统计分析，有效支撑各种评估维度。
指标开发过程长 需要针对大量指标进行定制化开发，牵涉人员多、流程长、工作量大。	**提升指标实时计算性能** 基于明细数据对指标进行实时计算，无需开发多层数据源，保证数据出口固定，支持指标统计维度的自由设计。
难以支撑分析业务 数据即席分析与数据开发共用平台，操作难度大、数据复杂，且容易对开发业务进行干扰。	**提供数据分析查询专用入口** 将明细数据提供给数据分析业务使用，通过运维手段保证资源占用情况，提升数据分析师的工作效率和工作体验。

期望的数据底座

简单

架构轻量化 开发便捷 维护简单

强大

具备强大的计算引擎，实现快速写入、快速查询，特别是在当前即席及实时的场景，能给与更多的支撑

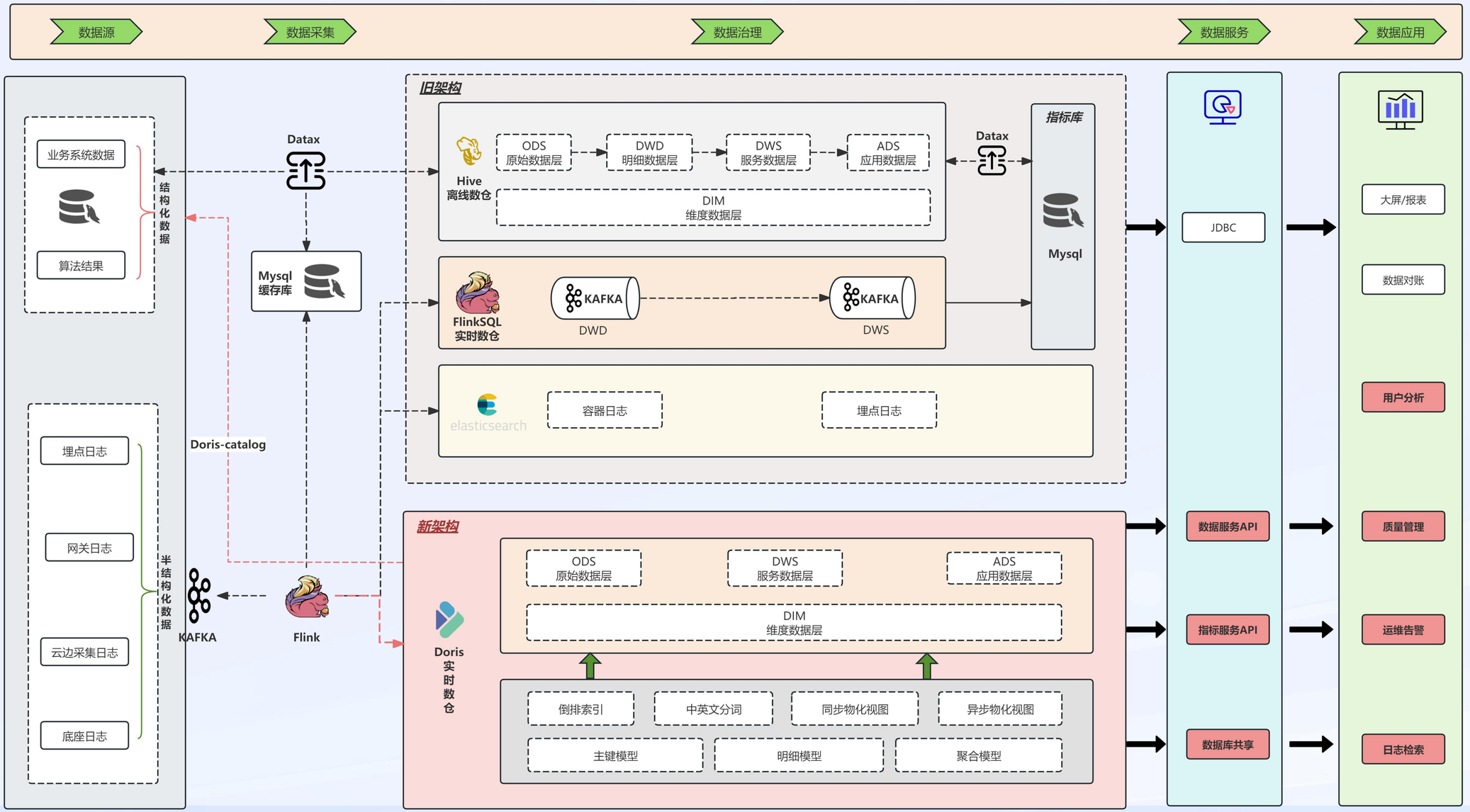
全面

覆盖采、存、治、用的数据生命周期

稳定

稳定运行、故障修复、资源管理

新老架构图


数据源
数据采集
数据治理
数据服务
数据应用
业务系统数据
算法结果
结构化数据
Datax
Mysql 缓存库
旧架构
Hive 离线数仓
ODS 原始数据层
DWD 明细数据层
DWS 服务数据层
ADS 应用数据层
DIM 维度数据层
Datax
指标库
Mysql
FlinkSQL 实时数仓
KAFKA
DWD
KAFKA
DWS
elasticsearch
容器日志
埋点日志
JDBC
大屏/报表
数据对账
用户分析
埋点日志
网关日志
云边采集日志
底座日志
半结构化数据
Doris-catalog
KAFKA
Flink
新架构
Doris 实时数仓
ODS 原始数据层
DWS 服务数据层
ADS 应用数据层
DIM 维度数据层
倒排索引
中英文分词
同步物化视图
异步物化视图
主键模型
明细模型
聚合模型
数据服务API
指标服务API
数据库共享
质量管理
运维告警
日志检索

03

案例分享

案例背景

在省级影像平台中，需要对从数千家医疗机构采集的数据进行质量评估及预警，以保证良好的数据质量，实现影像数据服务环节的稳定性以及提升用户的体验。

数据维度多、数据量大、对实时性要求高等问题成为质量改善的阻碍。

在 2023 年初，团队在数据质量管理的场景探索质量改善的技术方案。

数据质量管理-数据流向图

代码示例

```
CREATE TABLE `dws_xxxx_xxxx` (
  `k_1` varchar(255) NULL COMMENT '主键1',
  `k_2` varchar(255) NULL COMMENT '主键2',
  `k_3` varchar(255) NULL COMMENT '主键3',
  `k_4` date NULL COMMENT '主键4',
  `v_1` date NULL COMMENT '值1',
  `v_2` int(11) NULL COMMENT '值2',
  `v_3` int(11) NULL DEFAULT "值3",
....
  `time_1` datetime NULL COMMENT '计算时间'
) ENGINE = OLAP UNIQUE KEY(`k_1`, `k_2`, `k_3`, `k_4`) COMMENT '质量原子指标结果表'
PARTITION BY RANGE(`K_4`) ()
DISTRIBUTED BY HASH(`k_1`) BUCKETS x
PROPERTIES (
    "xxx":"xxx",
    "function_column.sequence_col" = "time_1"
    );
```

```
INERRT INTO dws_xxxx_xxxx
SELECT ...
from (
--增量数据
        WITH xxx_schema AS (
            SELECT ...
            FROM (
                ...
              WHERE ...
                ) a)
        select ...
  FROM A
  LEFT JOIN B
  LEFT JOIN C
  ...
  ON ...
  WHERE ...
```

```
INSERT INTO dws_xxxx_xxxx
(k_1,k_2,k_3,k_4,time_1,__DORIS_DELETE_SIGN__)
SELECT
    k_1,k_2,k_3,k_4,time_1,true
FROM (
    SELECT
        ...,
        rank() over(partition by ...
order by ... desc) as rank
    FROM dws_xxxx_xxxx
    WHERE ...
    ) t
WHERE t.rank>1
```

运行效率：1H+ 提升至 30s
依赖组件：6 个降低至 3 个
数据模型数：15 个降低至 2 个，另外增加 6 个视图
质量反馈周期：由原来的 T+1，提升至准实时（分钟级）

在数据去重，多表 JOIN，即席查询等场景，展现了强大的能力

应用示例

质量代办

机构名称	异常类型	异常明细	待办跟进人	业务时间	任务开始时间	任务结束时间	跟进状态	持续天数	操作
[illegible]	数据质量异常	检查影像完整异常	[illegible]	2024-11-27	2024-11-28 08:25:06	2024-11-28 10:14:06	已完成	1	查看详情
[illegible]医院	未上传影像	上传影像量为0	--	2024-11-27	2024-11-28 08:25:06	--	未跟进	1	查看详情
[illegible]	未上传影像	上传影像量为0	--	2024-11-27	2024-11-28 08:25:06	--	未跟进	1	查看详情
[illegible]保健院	数据质量异常	证件号异常	[illegible]	2024-11-27	2024-11-28 08:25:06	--	已跟进	1	查看详情
[illegible]口腔医院	数据质量异常	检查序列匹配异常; 检查影像匹...	[illegible]	2024-11-27	2024-11-28 08:25:03	2024-11-28 10:14:09	已完成	1	查看详情
[illegible]医院	数据质量异常	检查影像完整异常	[illegible]	2024-11-27	2024-11-28 08:25:03	2024-11-28 10:13:55	已完成	1	查看详情
[illegible]区域PACS	数据质量异常	检查报告匹配异常	[illegible]	2024-11-27	2024-11-28 08:25:03	2024-11-28 10:13:46	已完成	1	查看详情
[illegible]PACS	数据质量异常	检查序列匹配异常; 检查影像匹...	[illegible]	2024-11-27	2024-11-28 08:25:03	2024-11-28 10:13:40	已完成	1	查看详情
[illegible]	数据质量异常	检查序列匹配异常; 检查影像匹...	[illegible]	2024-11-27	2024-11-28 08:25:03	--	未跟进	1	查看详情
[illegible]	未上传影像	上传影像量为0	[illegible]	2024-11-27	2024-11-28 08:25:03	--	未跟进	1	查看详情

质量监测

机构名称	所属市	所属区	更新时间	未上传影像(天)	质量异常(天)	异常明细	操作
[illegible]医院	[illegible]	[illegible]县	2024-08-28 ~ 2024-11-28	--	1	检查影像完整异常	查看
[illegible]医院	[illegible]市	[illegible]县	2024-08-28 ~ 2024-11-28	--	1	检查序列匹配异常,检查影像匹...	查看
[illegible]医院	[illegible]市	[illegible]区	2024-08-28 ~ 2024-11-28	--	1	检查序列匹配异常,检查影像匹...	查看
[illegible]医院	[illegible]市	[illegible]区	2024-08-28 ~ 2024-11-28	--	1	检查序列匹配异常,检查影像匹...	查看
[illegible]医院	[illegible]市	[illegible]县	2024-08-28 ~ 2024-11-28	--	1	检查序列匹配异常,检查影像匹...	查看
[illegible]	[illegible]市	[illegible]县	2024-08-28 ~ 2024-11-28	--	1	检查序列匹配异常,检查影像匹...	查看
[illegible]区域PACS	[illegible]市	[illegible]区	2024-08-28 ~ 2024-11-28	--	--	--	查看
[illegible]心)	[illegible]市	[illegible]区	2024-08-28 ~ 2024-11-28	--	--	--	查看
[illegible]口腔医院	[illegible]市	[illegible]区	2024-08-28 ~ 2024-11-28	--	--	--	查看
[illegible]卫生院	[illegible]市	[illegible]市	2024-08-28 ~ 2024-11-28	--	--	--	查看

应用示例

质量看板

机构编码	机构名称	市	区											评估结果更新时间	操作
				2024-11-28	100.00%	98.11%	53	2.01	100.00%	100.00%	92.45%	92.45%	88.68%	92.45%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	99.34%	152	10.87	100.00%	100.00%	100.00%	100.00%	99.34%	100.00%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	200	28.08	100.00%	97.50%	97.50%	97.50%	97.50%	97.50%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	23	0.53	69.57%	100.00%	86.96%	86.96%	78.26%	86.96%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	98.02%	303	76.57	100.00%	100.00%	91.75%	91.75%	90.43%	91.75%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	97.26%	73	1.94	98.63%	100.00%	90.41%	90.41%	89.04%	90.41%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	236	38.42	100.00%	100.00%	84.75%	84.32%	83.90%	84.75%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	99.31%	144	18.96	95.14%	99.31%	97.92%	97.92%	95.83%	97.92%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	36	0.03	100.00%	100.00%	5.56%	5.56%	5.56%	5.56%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	0.00%	2		100.00%	50.00%	0.00%				2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	123	39.94	100.00%	100.00%	86.99%	86.99%	85.37%	86.99%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	70	17.41	100.00%	100.00%	100.00%	100.00%	100.00%	98.57%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	98.10%	105	13.81	100.00%	100.00%	74.29%	74.29%	65.71%	74.29%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	92	1.96	100.00%	100.00%	95.65%	95.65%	94.57%	95.65%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	99.05%	423	20.95	99.76%	100.00%	84.63%	84.63%	84.63%	84.63%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	4	2.54	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	353	39.27	44.19%	100.00%	100.00%	100.00%	100.00%	99.72%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	283	20.82	100.00%	100.00%	44.52%	44.52%	40.99%	44.52%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	5	0.04	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	2024-11-28 11:24:00	查看详情
				2024-11-28	100.00%	100.00%	129	30.03	100.00%	100.00%	86.05%	86.05%	86.05%	86.05%	2024-11-28 11:24:00	查看详情

质量推送

时间： 2024-11-27 00:00:00 – 2024-11-27 23:59:59

当前采集共计[redacted] 家医院

[redacted]: 200016

[redacted]): 204705

[redacted]): 201333

[redacted]: 70768951

[redacted]: 33223.099

[redacted]: 14772565.129

[redacted]比: 98.45%

[redacted]率: 98.6%

[redacted]率: 98.49%

[redacted]率: 98.6%

[redacted]率: 97.58%

[redacted]比: 96.97%

[redacted]比: 96.91%

[redacted]占比: 98.23%

[redacted]比: 99.31%

[redacted]占比: 99.31%

[redacted]比: 100%

[redacted]率: 98.43%

质量改善效果图

2023-03之前

在2023年3月份引入 Doris 之前，质量的改善效果缓慢，存在分析困难、实时性查等问题，导致质量问题反馈慢、根因分析困难

2023-03之后

2023年3月份以后，引入 Doris。并在后续半年的时间内，逐渐以 Doris 为底座，构建质量体系，并依托于 Doris 的特性，如聚合模型、物化视图等能力，实现了质量预警、统计以及质量问题溯源等能力。
在半年的时间内，数据质量迅速提升。
并依托于 Doris，搭建实时数仓。

改进成效

3 倍

人员效率提升

70+%

平台组件降低

30+倍

计算效率提升

70+%

物理资源节省

04

未来展望

未来展望

平台级建设

以 Doris 为基础的数据底座，向中台的方向进行建设，向简单、强大、便捷的方向发展。

业务向演进

在部分轻量化且面向数据分析、即席查询的业务场景，以 Doris 为核心，独立承担数据的存储、计算、服务等功能。

管理向发展

依托于 Doris 的能力，在数据管理特别是资产管理、血缘管理等方向，进一步的探究。

DORIS SUMMIT

Thanks for Watching!