Tencent 腾讯 | CSIG 云与智慧产业事业群

释放数据潜能 驱动业务增长

WeData 数据治理介绍及内部实践

腾讯云 史汉发

2022.7

企业网D1Net 2022全国CIO大会

目录

企业网D1Net 2022全国CIO大会

01

数据治理挑战

2022全国CIO大会

数据治理挑战

管理挑战

数据信息分散

- 业务系统不清晰
- 数据资产不明晰

业务挑战

元数据不全

- 需要有业务接口人维护补全业务信息
- 数据多样化，缺少统一标准

技术挑战

数据质量差

- 上游是一张空表，或者无效字段，总靠人肉来判别
- 修改一张表，不清楚带来哪些影响

管理挑战

维护困难

- 数据有问题，总是事后才知道
- 不知该由谁去维护预防

业务挑战

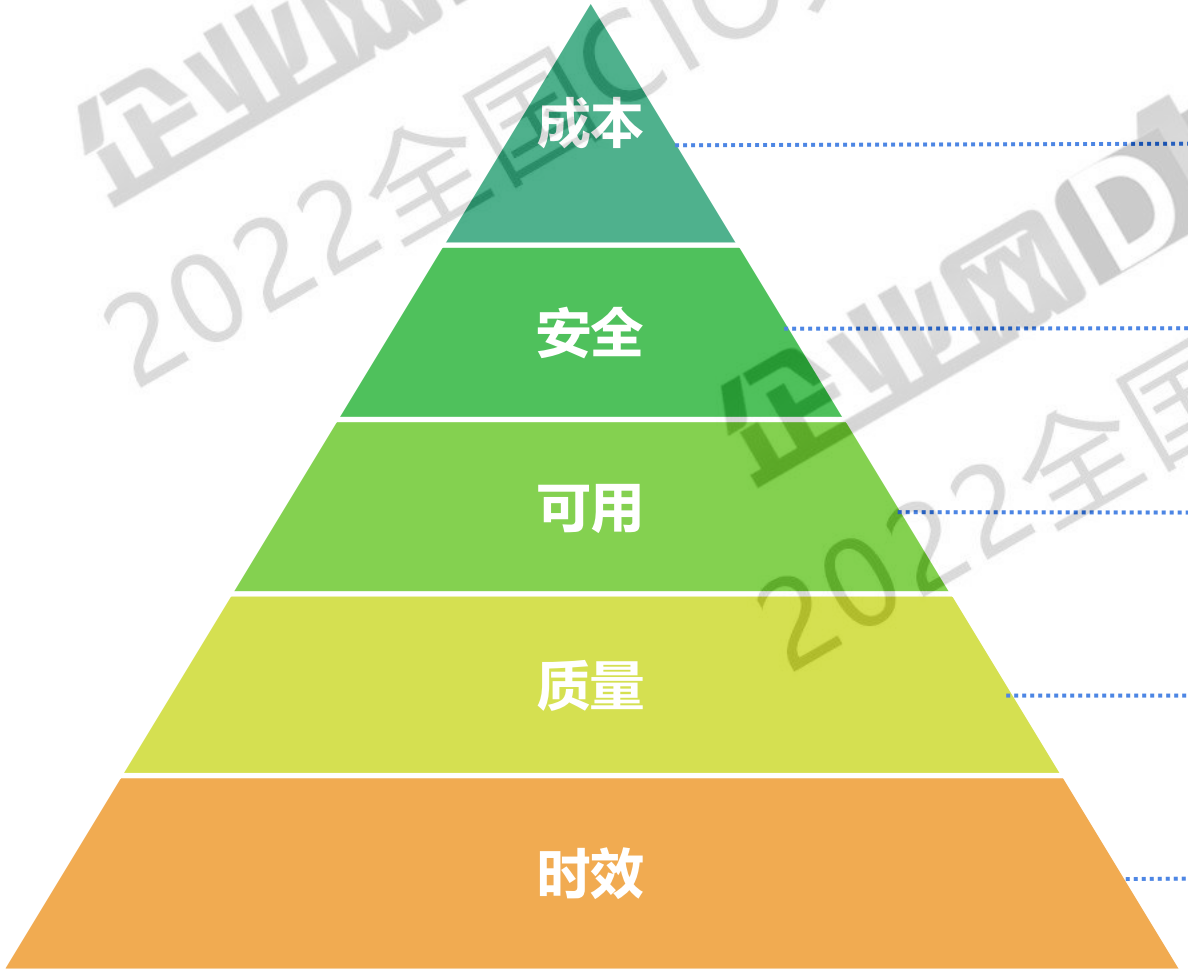
无法审计和度量

- 用数据的人这么多，如何保障公司数据访问是可靠安全的
- 数据成本是什么样

2022全国CIO大会 企业网DINet

数据治理“马斯洛的需求层次”

数字化不同阶段，数据治理关注的核心需求不同



- 数据存储、使用的成本优化和控制
- 数据安全、敏感数据识别、合规保证
- 数据易查找、好理解、可复用
- 数据有保障，准确、完整、有效
- 数据产出及时性

企业网DNet 2022全国CIO大会

02

腾讯内部数据治理实践

2022全国CIO大会

腾讯内部现状

数百产品线

数千数据分析师

数万业务场景

离线计算 千万级/天

实时计算 千万级/秒

总存储 EB级

腾讯内部实践—组织支撑

基于Oteam协同共建，建立全集团大数据统一体系

TEG 技术工程事业群	WXG 微信事业群	IEG 互动娱乐事业群	PCG 平台与内容事业群	CSIG 云与智慧产业事业群	CDG 企业发展事业群
腾讯大数据	微信	腾讯游戏	QQ	腾讯云	腾讯广告
腾讯计费	企业微信	腾讯电竞	QQ空间	腾讯安全	腾讯金融科技
腾讯问卷	微信支付		腾讯文档	腾讯云启	腾讯投资
腾讯乐享	QQ邮箱		Tim	腾讯政务	青腾
腾讯智维	微信读书		腾讯视频	腾讯教育	腾讯公益
腾讯安全平台	微信听书		腾讯微视	腾讯智慧交通	腾讯为村
腾讯数据中心	秒剪		应用宝	腾讯智慧出行	腾讯SSV
Tencent AI Lab	小程序		腾讯体育	腾讯地图	
TENCENT ROBOTICS X	公众号		腾讯新闻	腾讯会议	
	微信搜一搜		腾讯看点	腾讯企点	
	微信看一看		QQ浏览器	腾讯健康	
	视频号		腾讯影业	腾讯医典	
			企鹅影视	腾讯觅影	
			腾讯动漫	腾讯云医	
			腾讯内容开放平台	腾讯优图	
				腾讯多媒体实验室	

元数据管理企标　数据质量管理企标　数据安全管理企标　数据价值管理企标　数据共享管理企标　数据模型管理企标　数据运营管理企标　数据标准管理企标

协同共建

（18+大数据协同共建组织，涉及大数据全栈技术）

统一数据接入层　统一数据存储层　统一资源管理层　统一计算引擎层　统一作业调度层　统一数据应用层

2022全国CIO大会 企业网DINet

腾讯内部实践—方法论

表现	
5级（卓越）	体系非常成熟，可快速接入业务；具备智能化设施，科学的分析设施能够自我改进完善；
4级（优秀）	有体系，数据驱动
3级（可控）	有平台，管理可复制 平台接入2个以上业务
2级（基础）	有方法、有工具、有流程
1级（初始）	依赖人工，使用Excel等工具

2022全国CIO大会

腾讯内部实践—方法论

企业级治理标准

规范体系

- 元数据规范
- 数据质量规范
- 数据安全规范
- 数据共享规范
- 数据模型规范
- 数据运营规范

一站式数据治理工具平台

资产生产

- 数仓建模
- 数据开发
- 数据发布
- 数据运维

资产评估

- 数据定级
- 资产分
- 资产认证
- 成本账单

资产运营

- 数据地图
- 血缘分析
- 资产市场
- 责任归属

全域元数据

- 元数据模型
- 技术元数据
- 业务元数据
- 血缘元数据

大数据平台

- 天穹
- Hermes
- SuperSQL
- Flink

企业级测评体系

测评体系

- 元数据成熟度
- 数据质量成熟度
- 数据安全成熟度
- 资产管理成熟度
- 数据模型成熟度
- 数据运营成熟度

2022全国CIO大会

腾讯内部实践：如何进行标准制定？

资产健康分评估

资产指标维度	指标名称	指标引导目的	指标计算口径	不及格分标准(0分)	60分标准	80分标准	90分标准	100分标准	单项权重
规范性评估	1)表名命名规范性指标	表命名更加规范	a)表模型层级前缀不符合规范 b)表后缀命名不符合规范 c)表主题命名不符合规范	命中a,b,c	命中b,c	命中c或者b	命中c	无任何命中	x%
	2)表和字段注释规范性指	表注释更加规范易懂	a)表没有注释 b)表有注释，但是注释太短 c)字段没有注释 d)字段部分有注释，或者有注释的字段占比（目前定90%）	命中a	未命中a，但命中b或c	未命中a,b,c，只命中d	-	无任何命中	x%
	3)字段安全等级性指标	表权限管理符合规范	a)该表字段有安全等级配置占比<10% b)该表字段有安全等级配置占比<40%，但>=10%; c)该表字段有安全等级配置占比<60%，但>=40%; d)该表字段有安全等级配置占比<80%，但>=60%;	命中a	命中b	命中c	命中d	无任何命中	x%
	4)责任人符合规范	表责任人合理有人管	a)表归属离职人员 b)表BU和owner的BU不一致 c)表owner和任务生产owner不一致 d)表owner和最近一次DDL的owner不一致（目前后者元数据不准确，暂废弃）	命中a	命中b，无a	命中c，无a,b	命中d，无a,b,c	无任何命中	x%
	5)依赖加工规范	引导表依赖和计算加工更加规范	a)逆层依赖，违背了数据模型主体流程依赖的顺序，ods->dwd->dws->ads\dm的顺序，出现直接或者跨多级逆序的； b)底层依赖，出现除dwd以外依赖ods层的，如dws依赖ods，ads\dm依赖ods的； c)加工链路过长，数仓模型中，从源头到该表的任务层架>N的（此扣分项识别但不扣分）；	命中a	未命中a，命中b；	未命中a，b, 命中c；（调整为不扣分，即此条规则作废）	-	无任何命中	x%
保障性评估	1)有监控保障指标	引导任务有监控管理	a)表没有任何监控 b)表仅有非强制性监控，且监控数=1 c)表仅有非强制性监控，但监控数>1 d)表有强制性报错监控，且监控数=1 e)表有强制性报错监控，且监控数>1	命中a	命中b	命中c	命中d	命中e	x%
	2)有DQC报警保障指标	引导资产产出有DQC保障	a)表没有报警owner配置 b)表没有短信、企微报警配置 c)表没有电话报警配置 d)表没有产出承诺配置	命中a	没有命中a,命中b	没有命中a,b，命中c	没有命中a,b,c,命中d	无任何命中	x%
	3)有基线保障指标	资产生产有资源或承诺保障	a)表没有任何基线 b)表有普通任务基线 c)表有高优先级任务基线	命中a		命中b	-	命中c	x%
准确性评估	1)DQC达标指标	引导数据产出准确	a)最近一天出现数据为空报警 b)最近一天出现数据重复报警 c)最近一天表资产出现强阻塞报警						

数仓规范体系

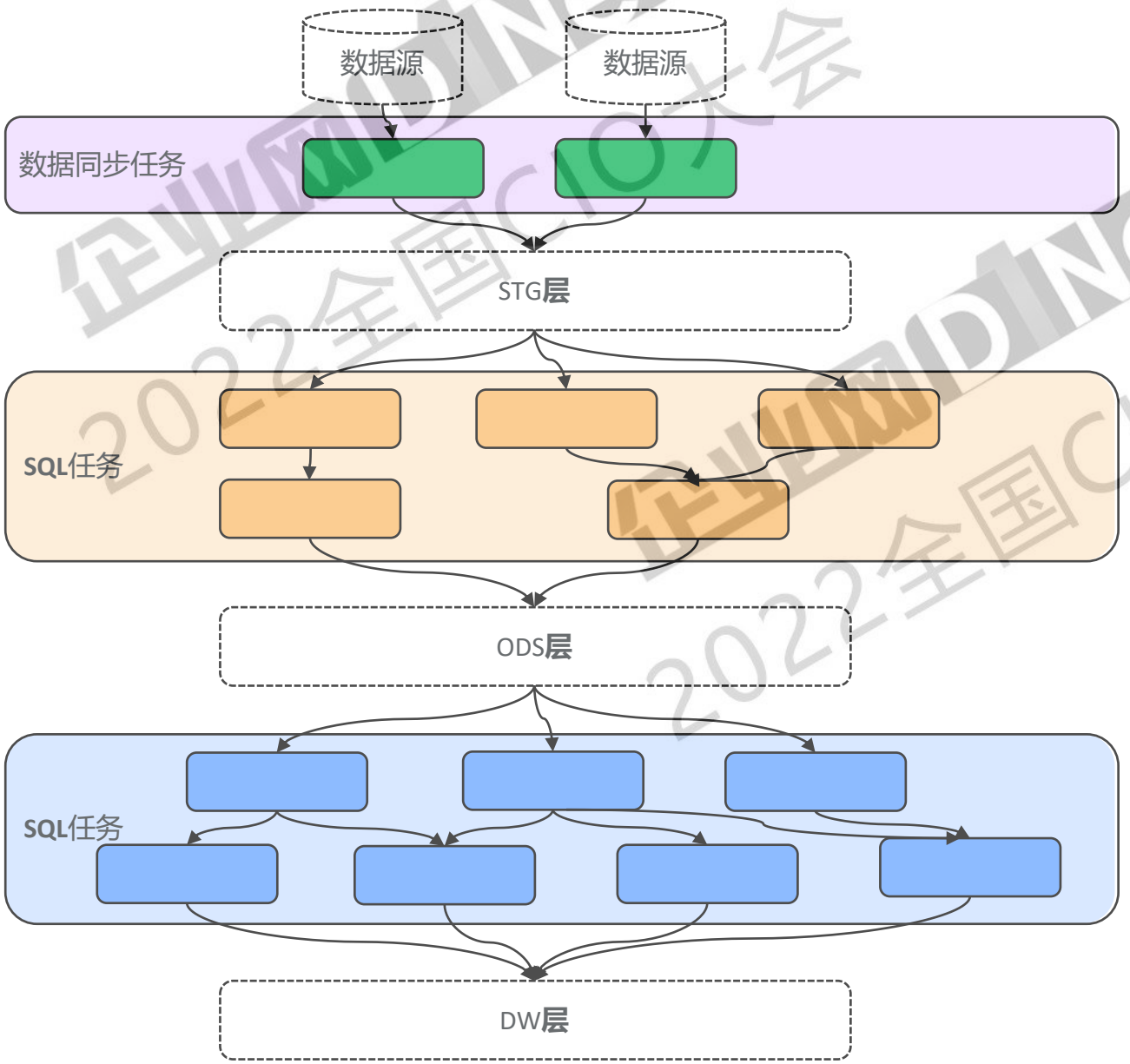
信息安全部-用户资料表-天全量

微信增值业务部-游戏中心点击表-天级增量

- 荣获2019年度中国信通院大数据“星河奖”
- DQMIS 2019中国《数据质量卓越实践奖》
- 首家互联网企业通过DCA大数据管理平台基础能力测评
- 参与制定国家《大数据资产管理实践白皮书》行业标准
- 信通院《数据标准管理白皮书2019版》
- 企业标准建立《腾讯数据治理 元数据管理规范》
- DAMA数据治理最佳实践

2022全国CIO大会 企业网DNet

腾讯内部实践：如何进行质量监控？



基于【数仓分层】做质量监控，让数据有保障

入口层/数据引入层/基础层

- 业务主外键是否唯一
- 周期性数据波动是否过大/特殊类（如日志等）
- 无周期性则判断、数据是否大于固定值

数据清洗层/整合加工层

- 增加一些对清洗逻辑的监控
- 对于整合的数据判断其数据唯一性、重复性

轻度/高度汇总层

- 根据汇总逻辑做平衡值监控

维表层/事实表层

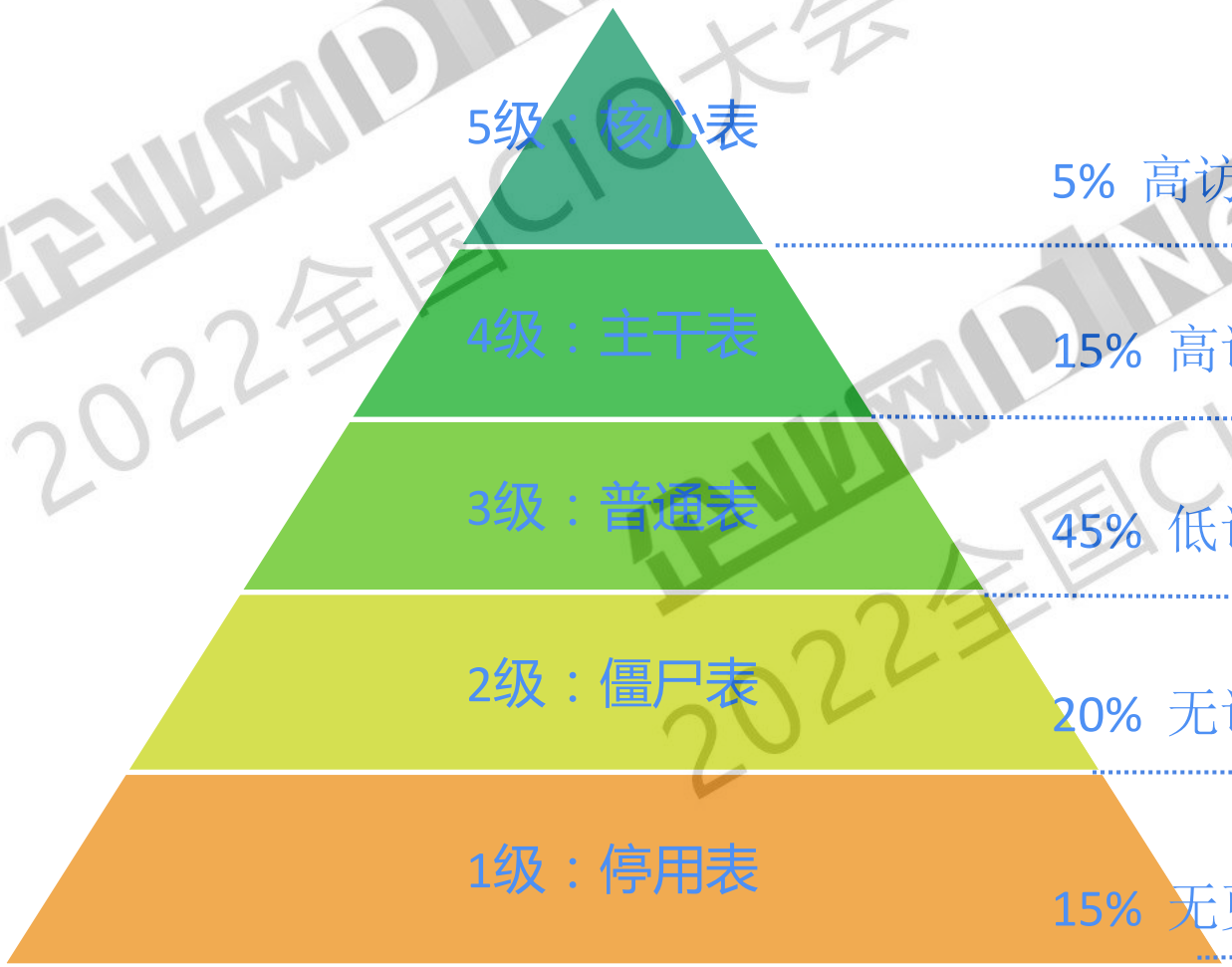
- 主外键一致
- 维度值增加/减少监控

出口层/应用层/报表层

- 逻辑主键
- 多表之间的平衡关联
- 特定业务逻辑监控

2022全国CIO大会

腾讯内部实践：如何进行资产评级？



5% 高访问且高扇出 扇出数>15,访问频率>20

15% 高访问或高扇出 扇出数>5,访问频率>10

45% 低访问且低扇出 扇出数>1,访问频率>10

20% 无访问且无扇出 6月内数据无访问且无上下游依赖

15% 无更新 6月内数据无更新

企业网DINet 2022全国CIO大会

腾讯内部实践：如何进行成本优化？

Bi、数据透明

红黑榜形成

健康度周报

资源消耗账单

多维度量化指标评分

稳定性评分 | 计算性能评分 | 存储成本评分 | 规范性评分 | …

任务标签

30日成功次数 | 30日失败次数 | 运行时长 | 子任务数量 | …

任务基础信息 | 任务运行状态 | 任务依赖 | 资源消耗信息 | 元数据信息

任务优化
累计优化任务2,940个

计算优化
累计节省546计算单元，节省成本百万级元/年

存储优化
累计节省存储8PB+

成本优化
每年节省千万级元

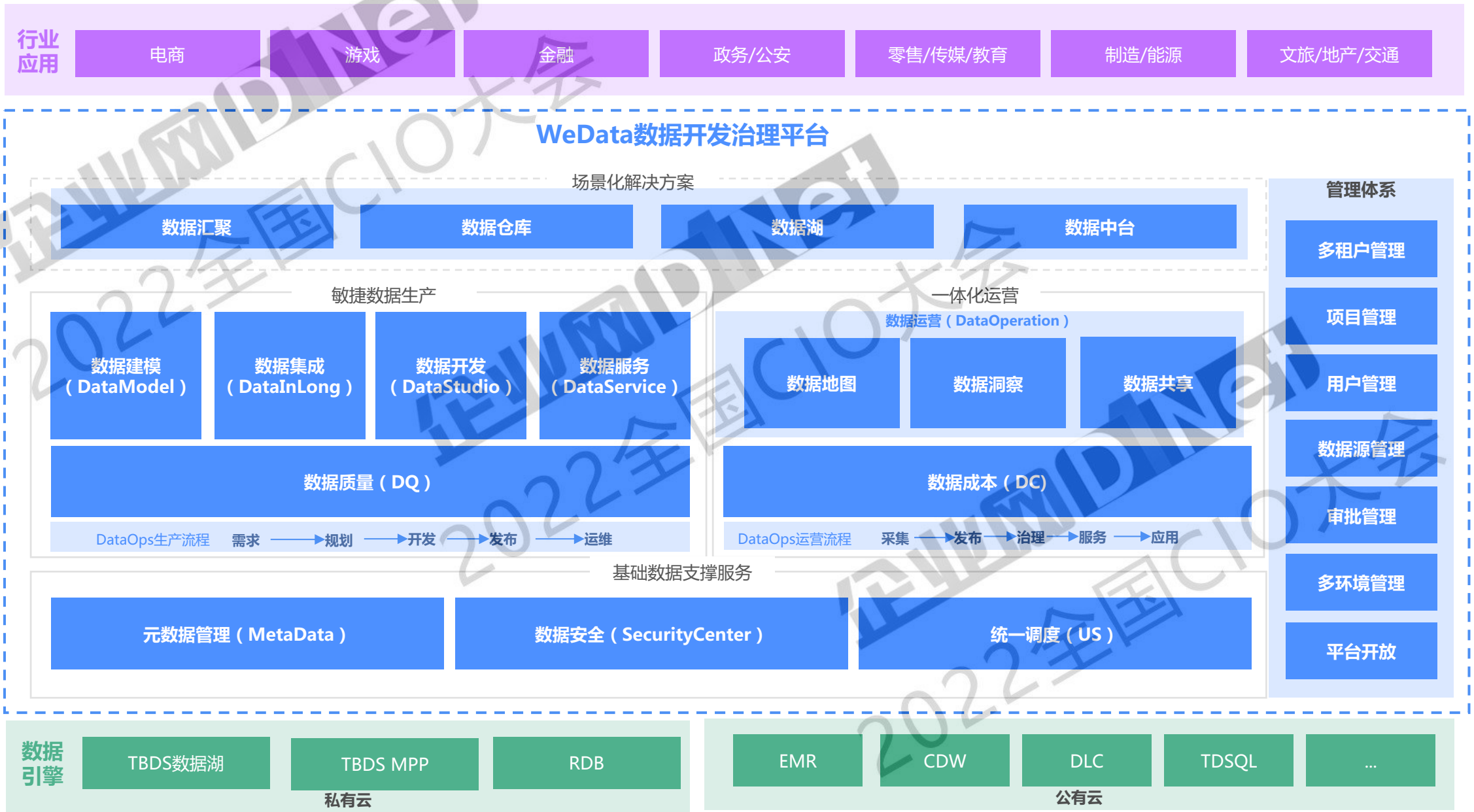
2022全国CIO大会

03

WeData 数据治理平台能力

2022全国CIO大会

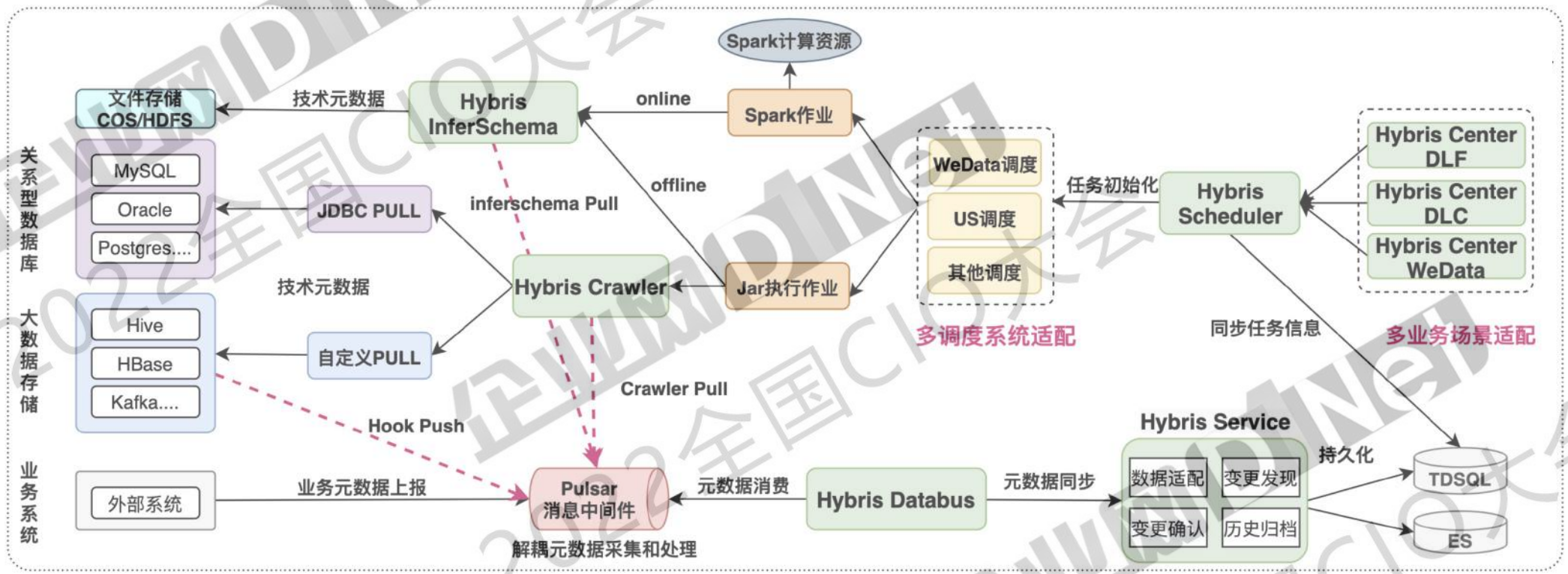
腾讯内部大数据能力的对外商业化输出-WeData



2022全国CIO大会 企业网D1Net

WeData数据治理-时效性：统一元数据

时效 | 质量 | 可用 | 安全 | 成本



DATA

1 多源支持

- 关系型数据库
- 大数据存储
- 业务系统

2 多种采集方式

- 实时Hook汇聚
- 离线周期采集补偿
- 业务元数据主动上报

3 多场景适配

- OLTP在线数据目录
- OLAP数据湖分析

2022全国CIO大会 企业网DINet

WeData数据治理-质量：数据质量

时效 | 质量 | 可用 | 安全 | 成本

1 可视化配置

- 数据监控
- 规则模板
- 运维管理

2 丰富模板

- 53种官方模板
- 自定义规则模板
- 字段级
- 表级

3 全维度规则

- 测试运行
- 事中检测
- 事后检测

4 质量报告

- 综合质量分
- 维度质量分
- 表质量分明细

2022全国CIO大会 企业网D1Net

WeData数据治理-质量：数据质量

时效 | 质量 | 可用 | 安全 | 成本

1. 规则模版定义

质量规则模版

增加数据质量规则集

2. 基于元数据配置数据质量监控

质量监控配置

离线周期检测

ETL流程监控

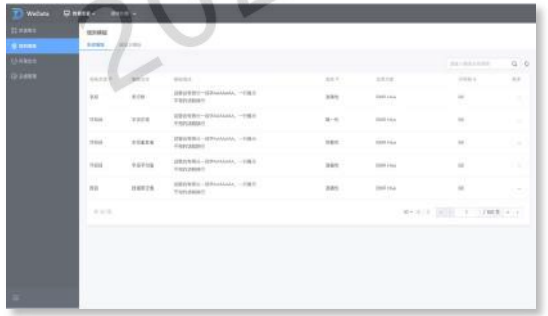
3. 质量问题跟踪处理

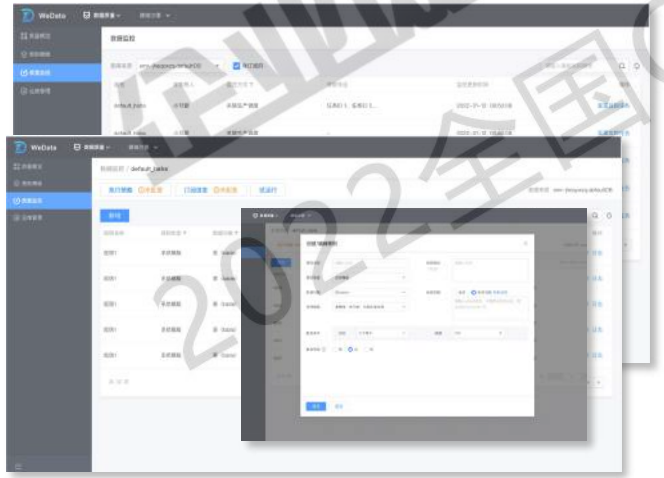
质量任务运维

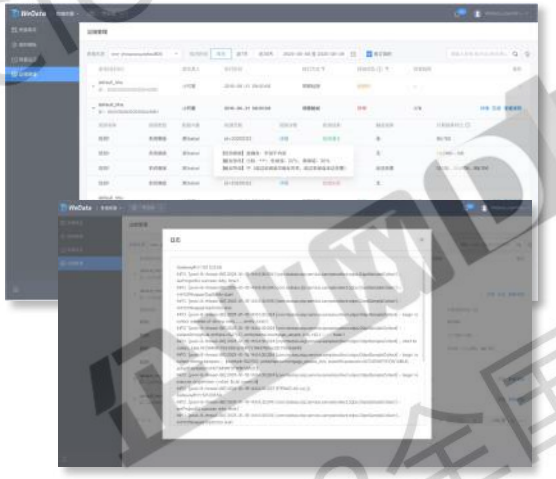
4. 定期数据质量分析

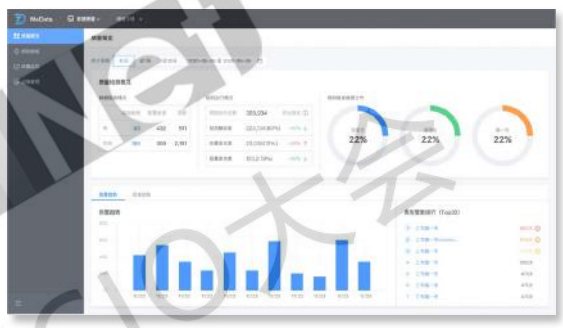
质量概览：质量核心指标日常关注

质量报告：数据质量考核评价









「规则模版」

- 支持系统内置规则模版 40+，包含固定值比较，和波动率对比较

「数据监控」

- 可针对表、字段的进行规则配置
- 同时支持关联到调度实现事中监控阻断、以及离线周期检测固定值比较，和波动率对比较

「运维管理」

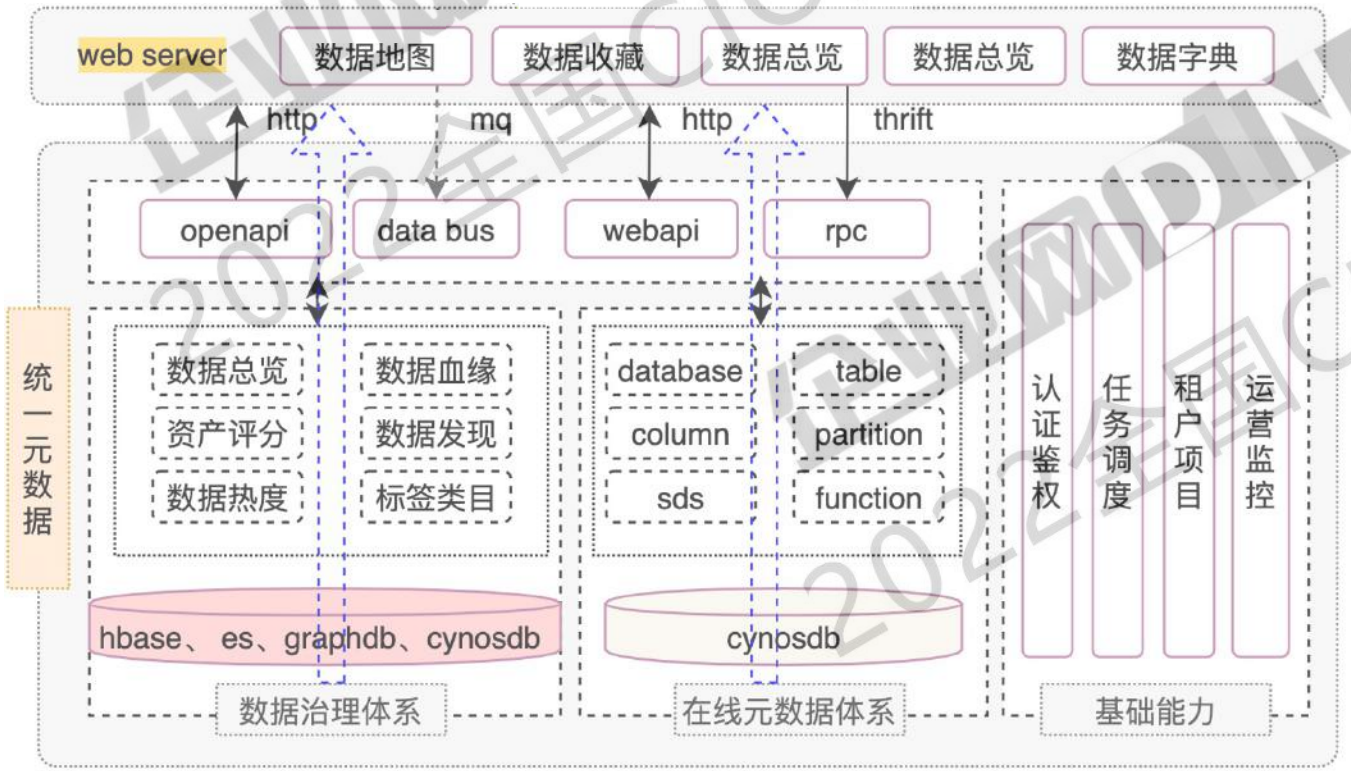
- 任务执行情况的运维管理，可查看检测结果，日志问题追踪
- 支持进行问题分级，并提供不同程度的告警触达（短信、微信等）

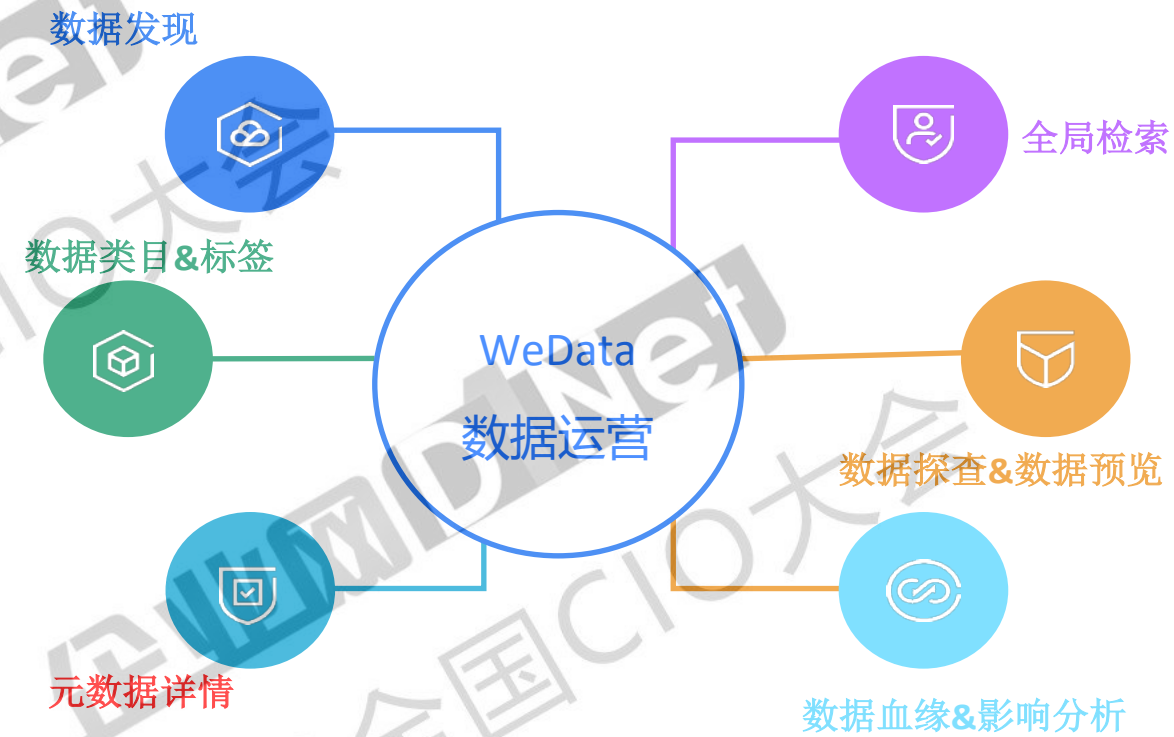
「质量概览」

- 不同周期下 数据检测、规则运行、质量维度问题情况分布等核心指标的统计
- 查看周期内告警和任务阻塞趋势，并快速定位到“搞频”

WeData数据治理-可用性：数据运营

时效 | 质量 | 可用 | 安全 | 成本





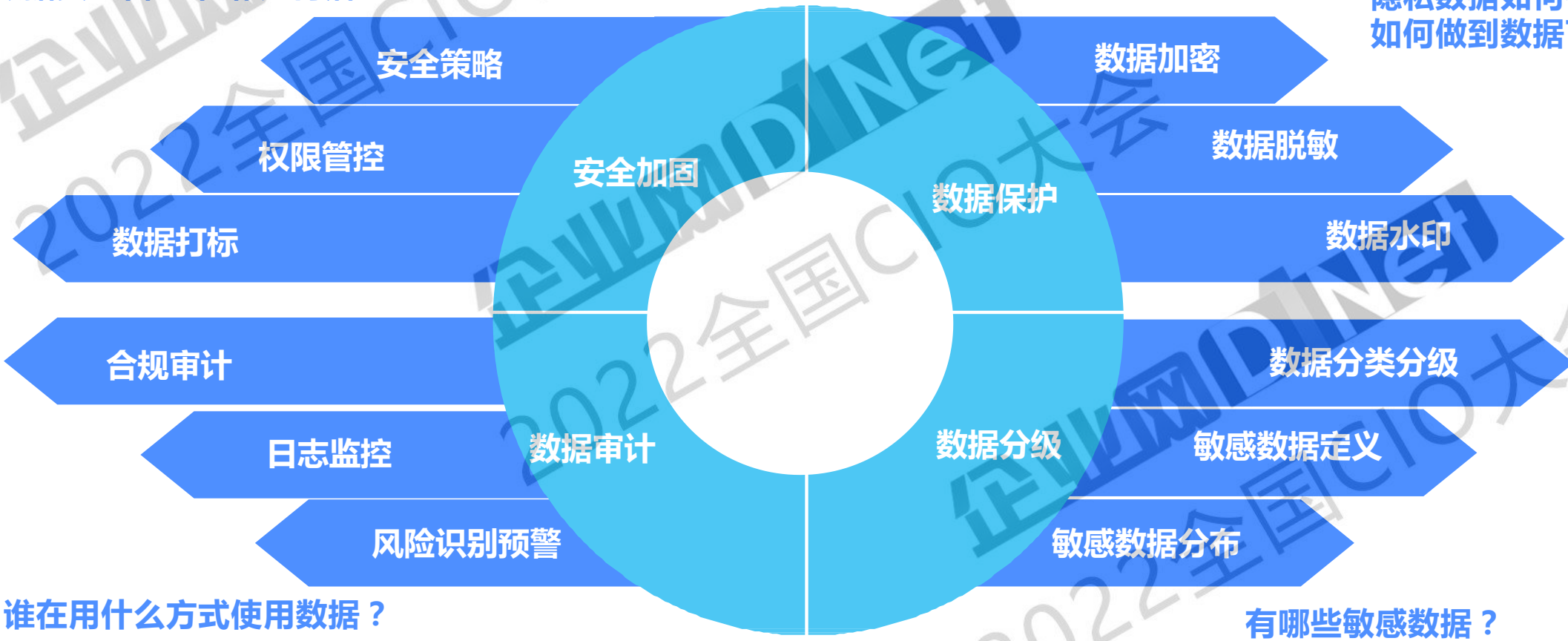
2022全国CIO大会

WeData数据治理-安全：数据安全

时效 | 质量 | 可用 | 安全 | 成本

如何管理敏感 数据的权限？
数据安全管理策略 如何落地？

隐私数据如何保护？
如何做到数据可用不可见？



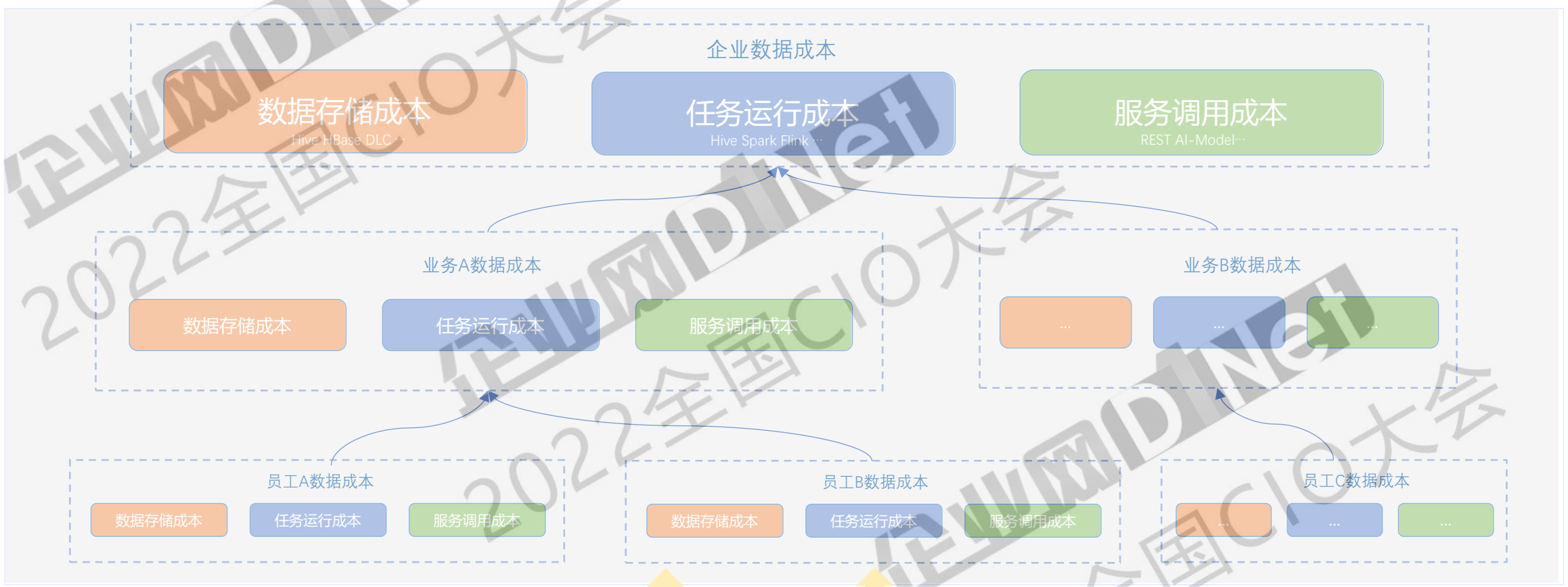
谁在用什么方式使用数据？
如何识别出有风险的数据操作？
如何判断数据操作是否合规？

有哪些敏感数据？
敏感数据分布在哪？

2022全国CIO大会 企业IDINet

WeData数据治理-成本：成本中心

时效 | 质量 | 可用 | 安全 | 成本



数据处理流程主要成本环节

数据集成 → 数据开发 → 数据质量 → 数据服务

2022全国CIO大会 企业网DINet

04 行业案例

2022全国CIO大会

成功案例：某商业银行客户数据能力中心建设

统一开发18000+个数据任务

- 统一对接MySQL/Oracle/DB2/文件等多种数据源
- 统一开发Hive/Spark/Shell/Python等多种任务
- 基于事件与时间的统一任务调度及运维

统一落标1000个标准项

- 数据标准平台建标5000个标准项
- 通过数据开发建模平台事前落标1000个

统一管控19000张数据表

- 数据资产平台展示完整字段级数据血缘
- 统一数据权限的申请、授权、审批等管控
- 精确到人到表的行列权限与动态脱敏控制

客户痛点

- 数据的开发、调度工作分散情况严重
- 数据质量低、数据落标困难
- 大数据组件运维复杂，缺乏集中统一的有效运维中心

业务数据

核心账务，对公信贷，个人贷款，在线贷款，代发代扣，外汇资金，人民币资金，理财，基金，网银，信用卡，柜面，支付清算等

方案架构

- 2021金融业新技术应用创新突出贡献奖
- 2021年度农村中小金融机构科技创新优秀案例

应用场景

零售和网金业务的指标标签 营销集市
反欺诈 智能决策 数据探索

成功案例：某工程机械制造商实时监控与故障预测项目

50万台工程设备实时监控

6.5小时工程设备故障预警

1000+个传感器维度实时分析

87.6%故障预测率

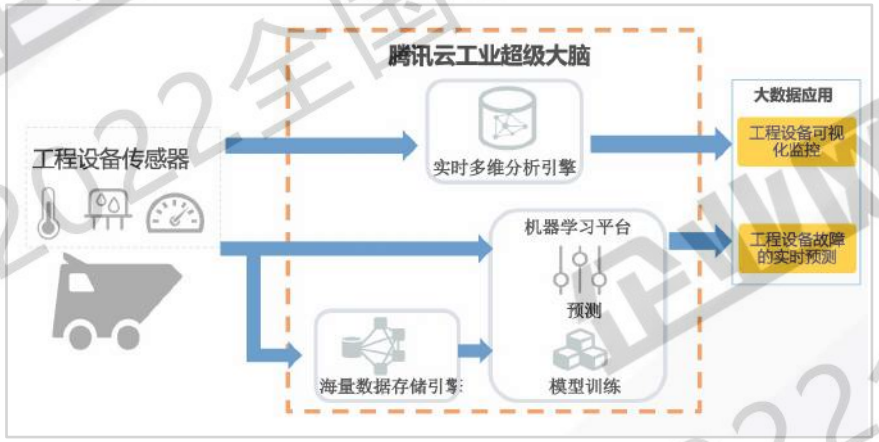
客户痛点

- 物联网设备产生海量数据需要采集
- 实时采集与计算复杂度高
- 基于时序数据进行设备分析

业务数据

工程设备传感器数据、物联网数据

方案架构



应用场景

工程设备实时可视化监控 预测性维护

企业网DINet 2022全国CIO大会

感谢聆听

2022全国CIO大会