金融级数据研发治理一体化平台实践

廖晓格

廖晓格 长期大数据平台、AI平台建设经验

- 平安银行数据资产管理及研发中心团队负责人，负责大数据基础平台、数据中台、BI及AI中台能力的建设
- 曾就职于PPTV、ebay、携程、华为，负责大数据平台应用的研发工作

目录

金融数据的特点及治理挑战

大数据服务应用数据质量**缺乏必要的监控和告警**

各业务的数据存在**孤岛**　　数据多份存储，加大**数据成本**

PB级别大数据海量存储和计算，造成**极高的负载**，影响系统**稳定性**，**批量时效**难以有效保障

大数据测试数据难造，生产数据脱敏到测试环境又**有安全隐患，敏感数据多，安全管控难**

数据流量洪峰不断刷新记录，如何**提升实时化能力**

虽然提供了各种线上平台和工具，但**思维**和动作还未全面**数据化**

安全　　**提效**　　**降本**

金融数据治理的价值

包袱？  资产！

存储成本剧增

计算负载极高

数据安全隐患

不可靠决策

……

数据治理

数据治理

数据治理

A 降低成本

B 提升数据加工效率

C 改善数据质量

D 控制数据风险

E 赋能管理决策

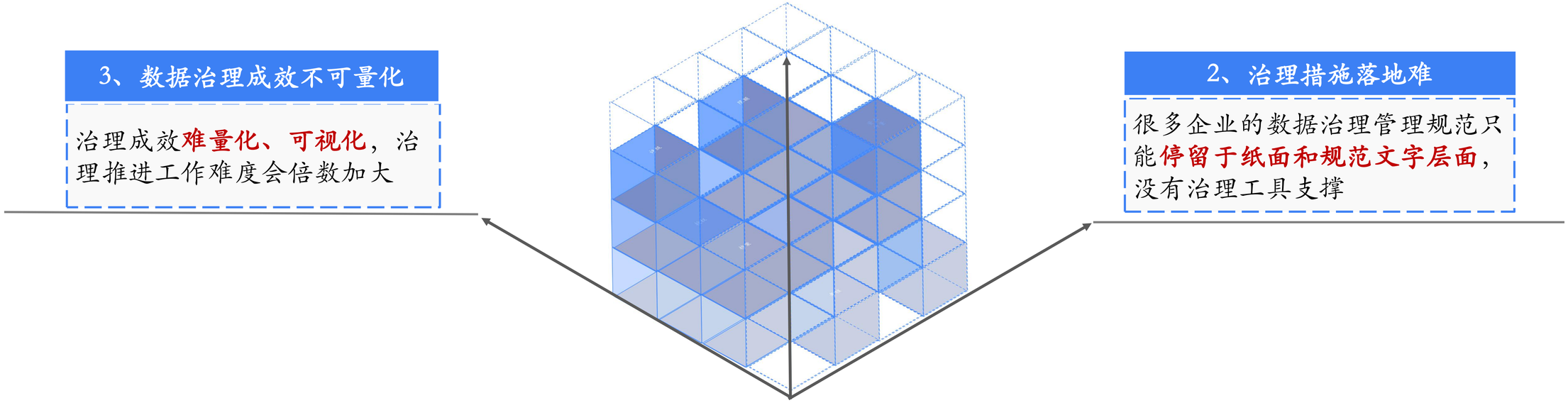
数据治理价值

数据治理传统模式的痛点

- 传统模式下的数据治理工作更多参考DAMA或者DCMM理论体系推行，但公司内部从哪里入手、以什么样的路径推进目前并没有结合自身企业和行业特点制定数据治理的标准体系，涉及效率、技术、管理、安全等，导致治理效果一直不佳，主要体现在以下几个方面：

1、运动式治理

通过**事后治理**，在短期内能看到一定成果，但因为没有融入到日常数据生产流程中，导致治理效果不可持续，不能长久解决治理痛点



3、数据治理成效不可量化

治理成效**难量化、可视化**，治理推进工作难度会倍数加大

2、治理措施落地难

很多企业的数据治理管理规范只能**停留于纸面和规范文字层面**，没有治理工具支撑

目录

数据治理的目标是什么

- 金融行业数据治理核心目标在于**兼顾安全、成本并最大化数据价值**，因此数据治理需要解决四个使命：

治理DAMA方法论与工具结合，将治理方法论以及行内所有规范，通过平台工具结合，提供工具化的治理能力，实现治理线上化；

全周期治理解决方案，数据治理涉及多个流程、多个平台、多方不同角色，整合各方在平台提供统一治理能力，实现治理标准流程；

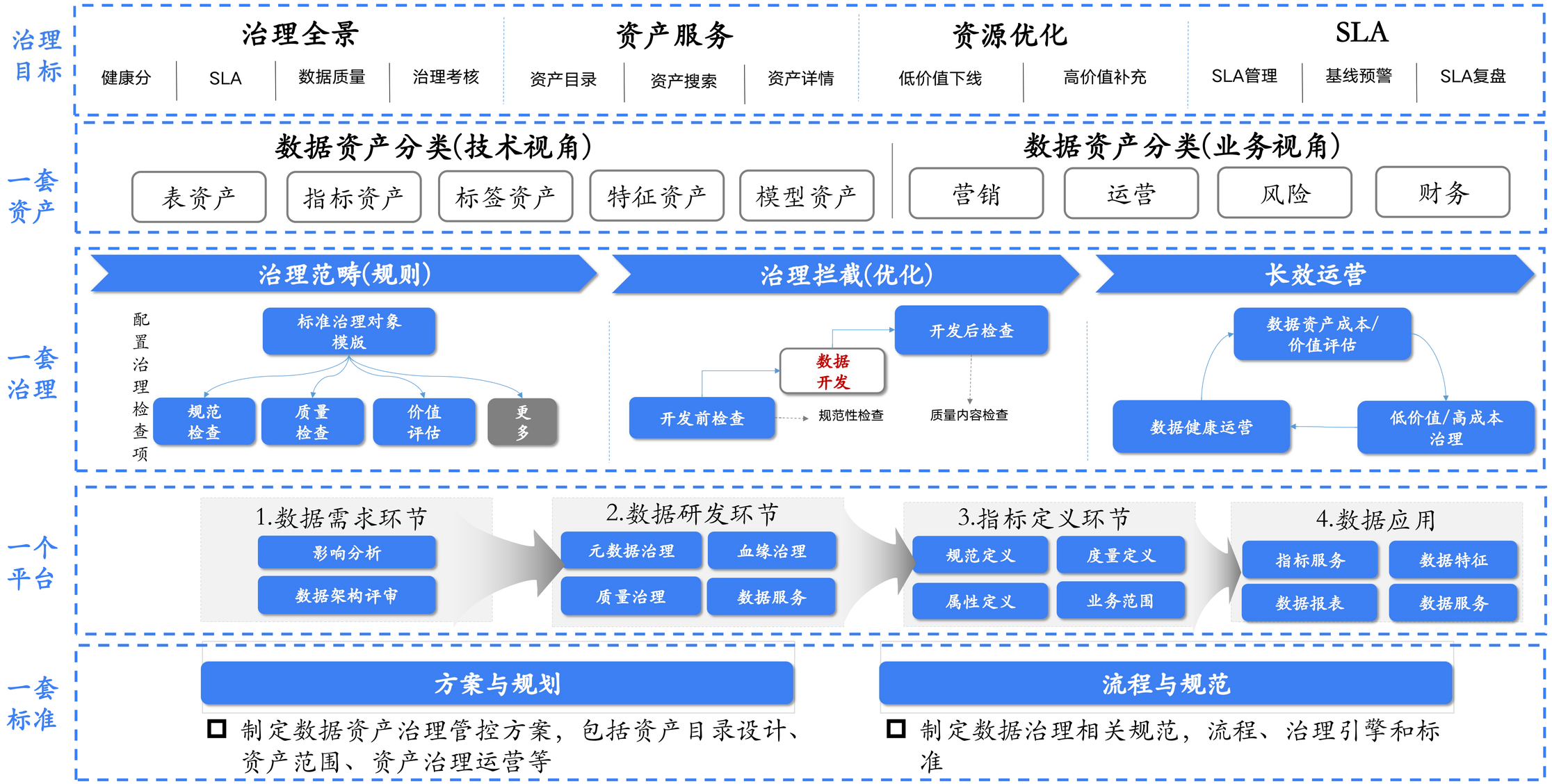
集成规则策略，可以通过自动化治理能力识别安全风险、敏感数据，通过内置规则和策略提升治理效果；

数据价值最大化，包括通过数据生命周期、成本/价值评估逆向推动成本治理，释放数据价值、降低数据应用成本

沉淀数据资产

提供高可用的数据服务

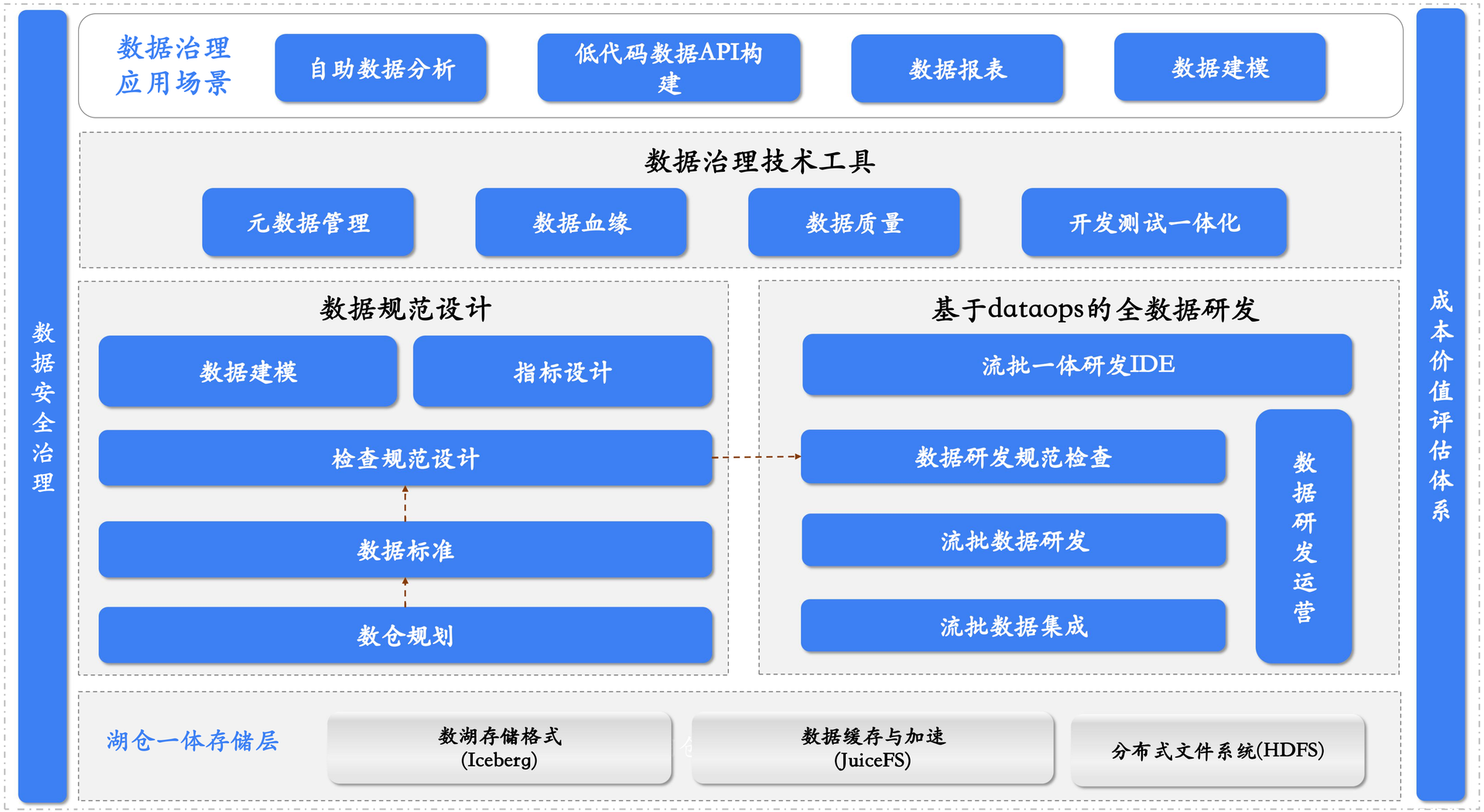
数据治理体系建设


治理目标
治理全景
健康分
SLA
数据质量
治理考核
资产服务
资产目录
资产搜索
资产详情
资源优化
低价值下线
高价值补充
SLA
SLA管理
基线预警
SLA复盘
一套资产
数据资产分类(技术视角)
表资产
指标资产
标签资产
特征资产
模型资产
数据资产分类(业务视角)
营销
运营
风险
财务
一套治理
治理范畴(规则)
治理拦截(优化)
长效运营
配置治理检查项
标准治理对象模版
规范检查
质量检查
价值评估
更多
开发后检查
数据开发
开发前检查
规范性检查
质量内容检查
数据资产成本/价值评估
数据健康运营
低价值/高成本治理
一个平台
1.数据需求环节
影响分析
数据架构评审
2.数据研发环节
元数据治理
血缘治理
质量治理
数据服务
3.指标定义环节
规范定义
度量定义
属性定义
业务范围
4.数据应用
指标服务
数据特征
数据报表
数据服务
一套标准
方案与规划
制定数据资产治理管控方案，包括资产目录设计、资产范围、资产治理运营等
流程与规范
制定数据治理相关规范，流程、治理引擎和标准

目录

数据开发治理一体化解决方案

- 将数据研发与数据治理方法论结合，提供开发治理一体化解决方案平台，目标实现数据安全可控、高质量，最终驱动数据在业务场景释放更大价值



开发治理核心能力

DataOps全周期数据研发

将数据研发过程标准化，引入CI/CD方法融入数据研发流程；

数据治理嵌入研发过程

改变以往先产生后治理的流程逻辑，将治理规范融入数据研发流程；

先设计再开发服务

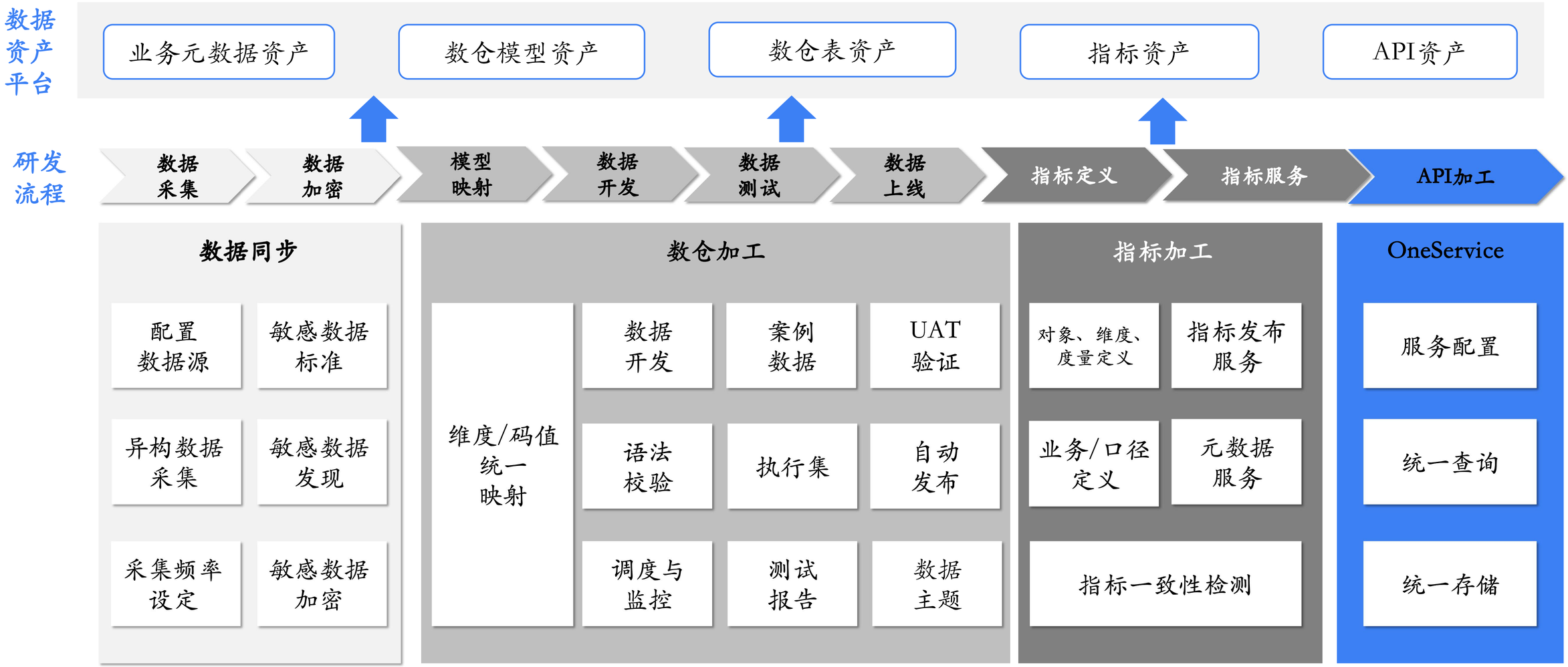
以数据服务和数据指标驱动数据研发过程，遵循先设计再研发的治理设计理念；

数治理技术工具

面向数据自动校验端，提供数据质量/数据血缘/元数据管理/规范检查/开发测试一体/能力服务，实现线上数据的自动检核；

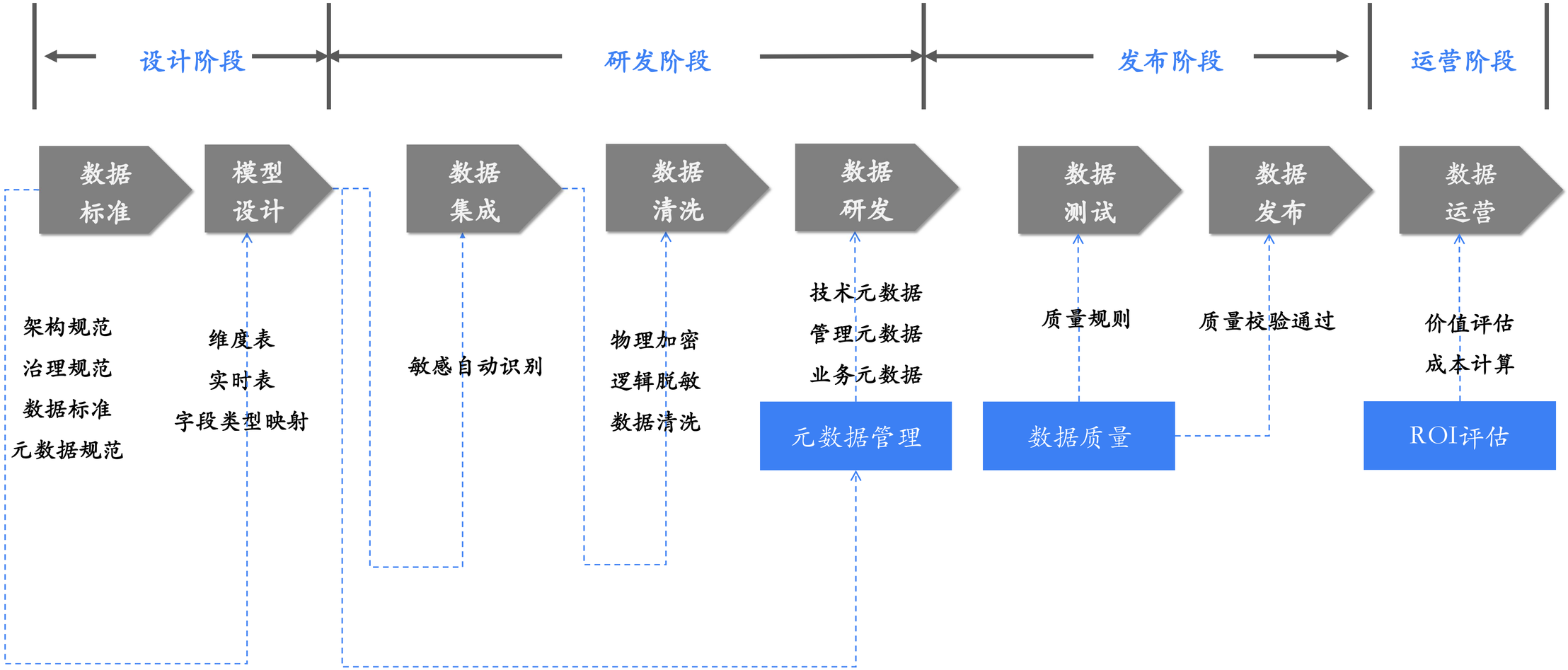
数据研发治理一体化平台全流程

- 统一数据研发全周期流程，标准化数据建模过程，降低模型研发过程中的人为风险同时，提升整个数据研发效率



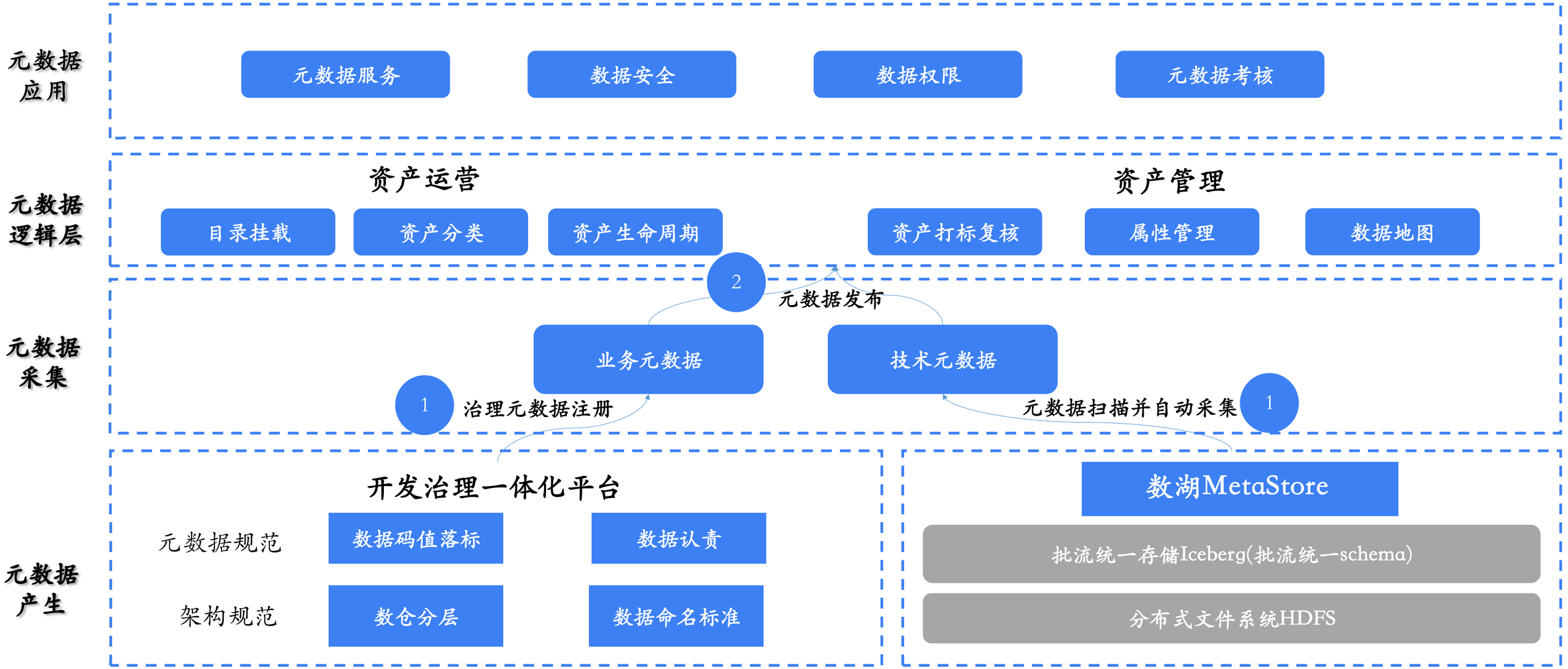
数据开发治理一体化解决方案-DataOps全周期开发治理能力

- 将数据治理融入数据研发的全生命周期流程中，在数据开发过程中完成全域数据治理工作,最终实现数据开发过程中自动化治理的管控目标



数据开发治理一体化解决方案-元数据治理

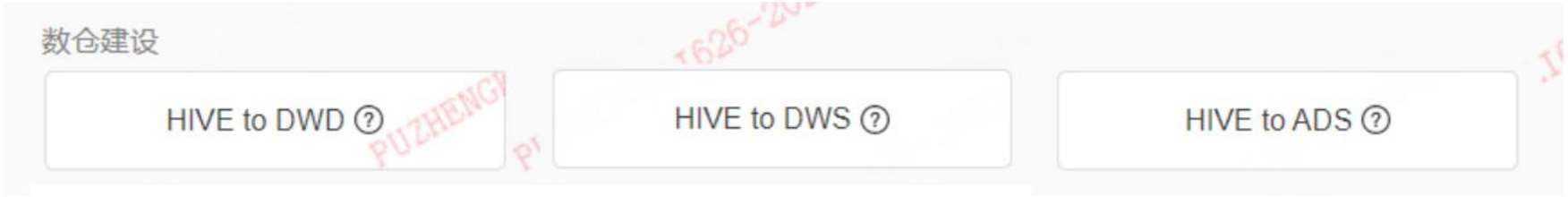
- 数据模型设计阶段，元数据治理是核心治理对象，遵循**数仓层级、命名规范、数据标准落标**等通过开发治理工具执行，开发治理一体化平台针对事前、事后的自动盘点运营；



数据开发治理一体化解决方案-元数据治理-强制检查项(举例)

- 针对于数据治理中基础元数据管理，基于行内统一数据标准治理规范，在开发过程中实现对于元数据管理的各项自动落标，确保元数据可用、可管、可控；

检查1



基于数仓建设标准，平台层面强制约束DWD/DWS/ADS分层的数据研发团队，通过平台约束分层落地

检查2

数据研发过程中明确分配归属部门管理

检查3

字段英文名称	字段中文名称
BECIF_CUSTID	12位客户号
PRE_LMT	预审额度标签
EXP_DATE	额度有效期
PRODUCT_TYPE	产品类型
CHINESE_NAME	中文姓名
BIRTHDAY	生日
TEL_NUM	电话号码

命名规范自动化，包括表名和字段名

检查4

码值	码值说明	码值排序
0	不详	码值排序
10	标准身份证	码值排序
11	临时身份证	码值排序
12	户口簿	码值排序
13	解放军军官证	码值排序
14	武警警官证	码值排序
15	护照	码值排序
16	港澳居民往来内地通行证/回乡证	码值排序
20	其他身份证明文件(个人)	码值排序
22	解放军士兵证	码值排序
23	解放军文职干部证	码值排序

码值落标线上化，开发过程中实现码值落标

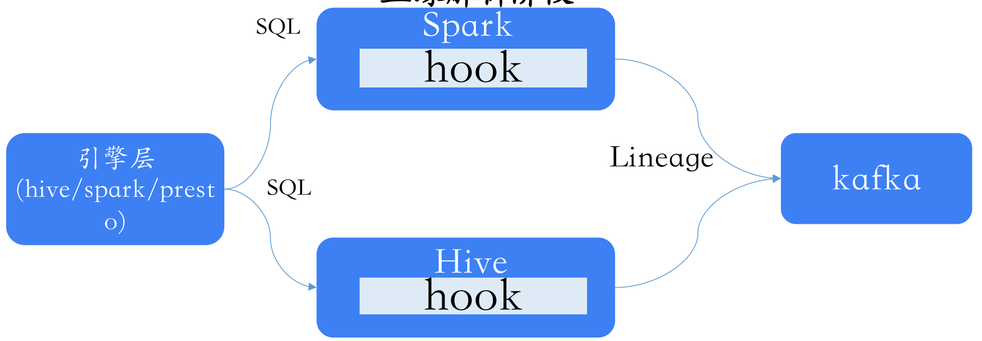
数据开发治理一体化解决方案-血缘治理能力

- 数据研发人员只需关注将自身需求的业务逻辑转换为开发逻辑，由开发治理平台自动进行脚本解析并生成血缘关系，同步实现**血缘链路、血缘层级依赖、数仓分层依赖**等治理事项自动化检查，确保数据血缘健康运营；

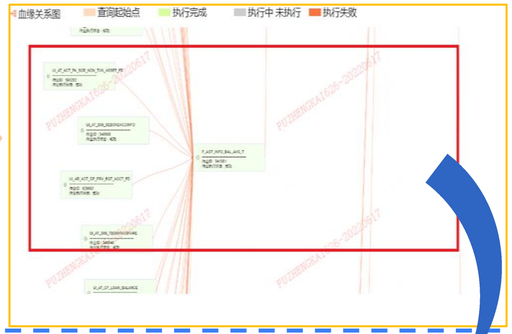
编码阶段

```
,A.NEW_RISK_LEVEL
,A.IS_HIS_FLAG
FROM VIEW.T_IF_P_CUST_ATTRIBUTE A
-- LEFT JOIN FACT_RPSM.F_ORANGE_TRADE_SUM H
--   ON A.CUST_NO = H.BECIF
--  AND H.DT='${start|yyyyMMdd}'
--  LEFT JOIN FACT_RPSM.F_CARD_INFO_ALL Y
--    ON A.CUST_NO = Y.CUST_NO
--   AND Y.IS_FIRST='Y'
--   AND Y.DT='${start|yyyyMMdd}'
LEFT JOIN FACT_RPSM.F_CUST_CARD_LVL_INFO J
  ON A.CUST_NO = J.CUST_NO
 AND J.DT='${start|yyyyMMdd}'
LEFT JOIN FACT_RPSM.F_DF_CUST_TRANS_SUM O
  ON A.CUST_NO = O.CUST_NO
 AND O.DT = '${start|yyyyMMdd}'
LEFT JOIN FACT_RPSM.F_CUST_ASSET_LEVEL_M L
  ON A.CUST_NO=L.CUST_NO
 AND L.DT=[('${start|yyyyMMdd}'=='${mend|yyyyMMdd}')
          '${start|yyyyMMdd}'
```

血缘解析阶段



血缘应用阶段



血缘治理阶段

开发过程中血缘治理

- **层级依赖检查：** 数据研发作业提交之后，依据自动计算的血缘分析与DWD层血缘层级，层级太深禁止上线；
- **分层依赖层面：** 依据ODS-DWD-ADS分层规范，禁止进行跨层依赖，同时ADS内私有域集市层禁止互相依赖；

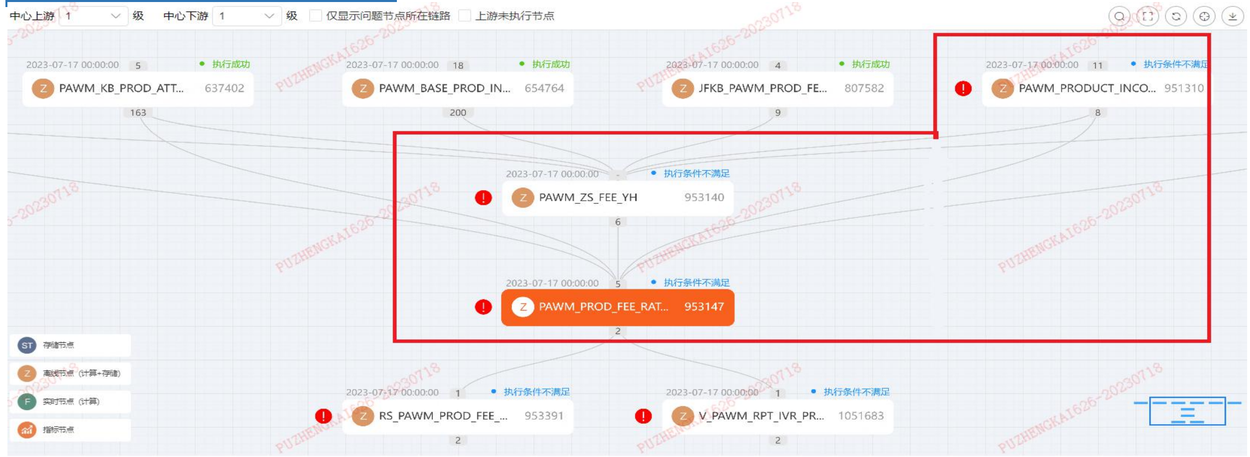
运营过程中血缘治理

- **运营时效检查：** 实时分析层级依赖作业的调度运行时间，根据高保作业的时效要求，线上分析延迟影响；
- **运营成本治理：** 依据作业互相依赖以及访问热度，自动针对冷作业进行识别并进行下线，降低集群存储和计算成本；

数据开发治理一体化解决方案-自动调度能力

- 开发治理一体化平台基于研发作业的依赖血缘，同时支持数据研发人员人工添加自定义依赖，实现对于调度的整体自动化平台管控，屏蔽人为控制影响，提升数据运行的自动性

1.血缘依赖调度



2.并发度调度管控

最新执行信息

时间	运行状态	详情
2022-06-17 19:21:42	CHECKING:执行条件不满足	目标集群(PSA_SAMS_ES_CLUSTER)并发性已经达到上限[10/10]:2022-06-17 19:21:42,[235988605, 235975790, 235981828, 235987730, 235975796, 236099323, 236006737, 235975721, 235988907, 235975789]，当前实例需要排队等待

历史实例状态

下载jit_header_history

时间	运行状态	CHECK_CODE	LAST_CHECK_REASON
2022-06-17 19:21:42	CHECKING:执行条件不满足	CHECK_CODE_13008	目标集群(PSA_SAMS_ES_CLUSTER)并发性已经达到上限[10/10]:2022-06-17 19:21:42,[235988605, 235975790, 235981828, 235987730, 235975796, 236099323, 236006737, 235975721, 235988907, 235975789]，当前实例需要排队等待
2022-06-17 19:21:42	CHECKING:执行条件不满足	CHECK_CODE_13008	目标集群(PSA_SAMS_ES_CLUSTER)并发性已经达到上限[10/10]:2022-06-17 19:20:37,[235988605, 235982874, 235975790, 235987730, 235975796, 236099323, 236006737, 235975721, 235988907, 235975789]，当前实例需要排队等待
2022-06-17 19:21:41	CHECKING:执行条件不满足	CHECK_CODE_13008	目标集群(PSA_SAMS_ES_CLUSTER)并发性已经达到上限[10/10]:2022-06-17 19:20:37,[235988605, 235982874, 235975790, 235987730, 235975796, 236099323, 236006737, 235975721, 235988907, 235975789]，当前实例需要排队等待
			目标集群(PSA_SAMS_ES_CLUSTER)并发性已经达到上限[10/10]:2022-06-17 19:20:37,[23598

3.自定义条件调度

最新执行信息

时间	运行状态	详情
2022-06-17 19:27:28	CHECKING:执行条件不满足	用户配置的Check条件不满足，{22:05=22:05}

历史实例状态

下载jit_header_history

时间	运行状态	CHECK_CODE	LAST_CHECK_REASON
2022-06-17 19:27:28	CHECKING:执行条件不满足	CHECK_CODE_15001	用户配置的Check条件不满足，{22:05=22:05}
2022-06-17 19:27:27	CHECKING:执行条件不满足	CHECK_CODE_15001	用户配置的Check条件不满足，{22:05=22:05}
2022-06-17 19:27:27	CHECKING:执行条件不满足	CHECK_CODE_15001	用户配置的Check条件不满足，{22:05=22:05}
2022-06-17 19:22:25	CHECKING:执行条件不满足	CHECK_CODE_15001	用户配置的Check条件不满足，{22:05=22:05}
2022-06-17 19:22:25	CHECKING:执行条件不满足	CHECK_CODE_15001	用户配置的Check条件不满足，{22:05=22:05}
2022-06-17 19:22:24	CHECKING:执行条件不满足	CHECK_CODE_15001	用户配置的Check条件不满足，{22:05=22:05}
2022-06-17 19:19:22	CHECKING:执行条件不满足	CHECK_CODE_15001	用户配置的Check条件不满足，{22:05=22:05}

执行调度检查

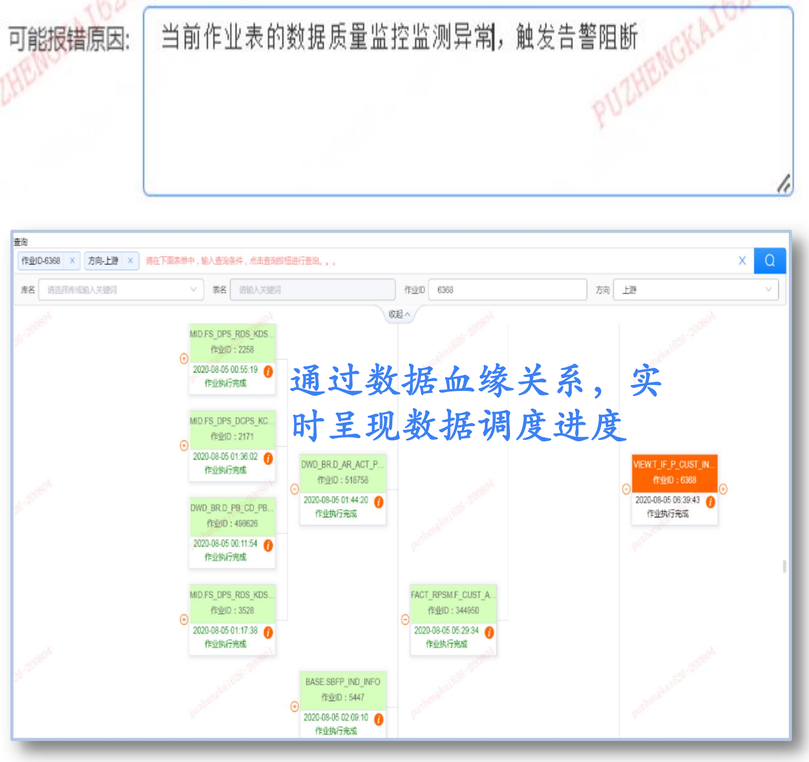
数据开发治理一体化解决方案-质量治理能力

- 数据质量已经成为银行数据治理的核心组成部分，从治理视角而言，建立完整全流程的数据质量体系，及时发现质量问题->实时预警属主修复->事后复盘增强测试发布环节检测、提升银行数据整体质量，提供更精准的决策分析数据；

事前-质量核验规则

事中-质量核验

实时质量异常阻断(依据血缘依赖阻断下游自动调度)

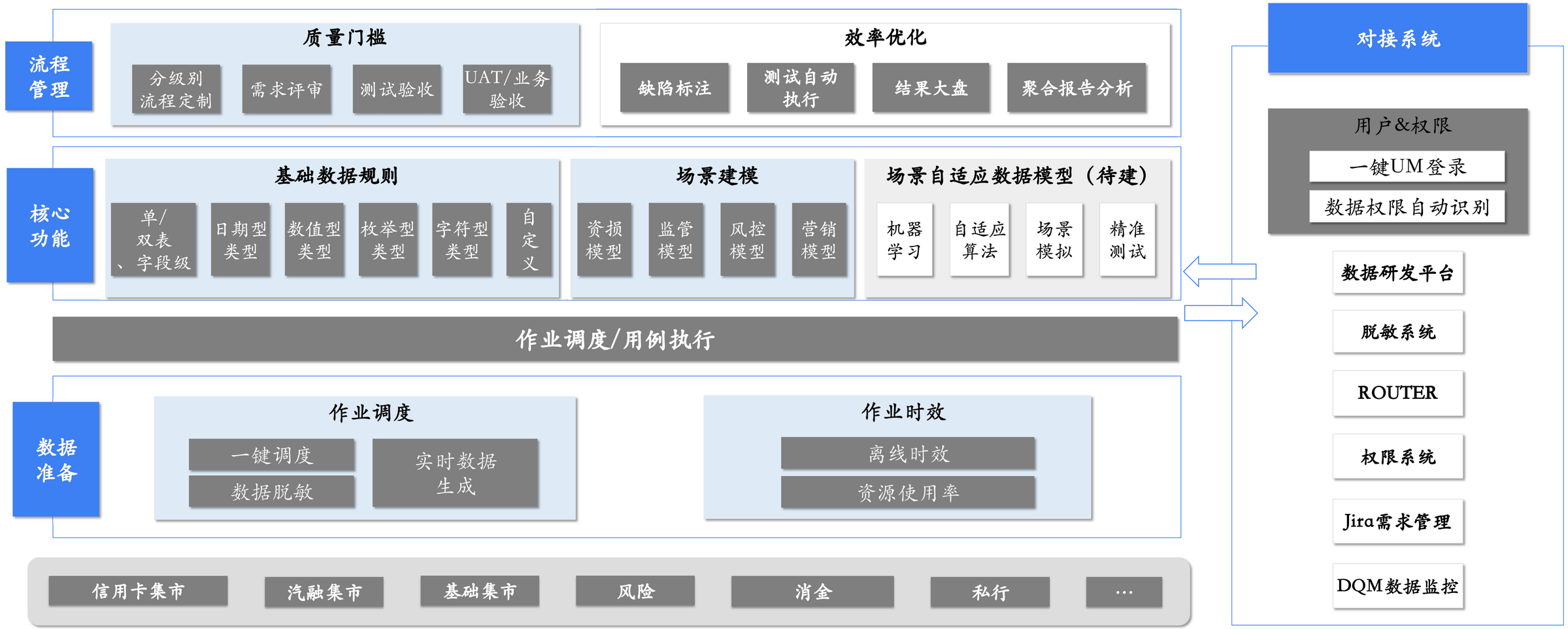


事后-异常质量问题追踪复盘

基于过程质量问题，工单追踪异常整改

数据开发治理一体化解决方案-开发测试一体化

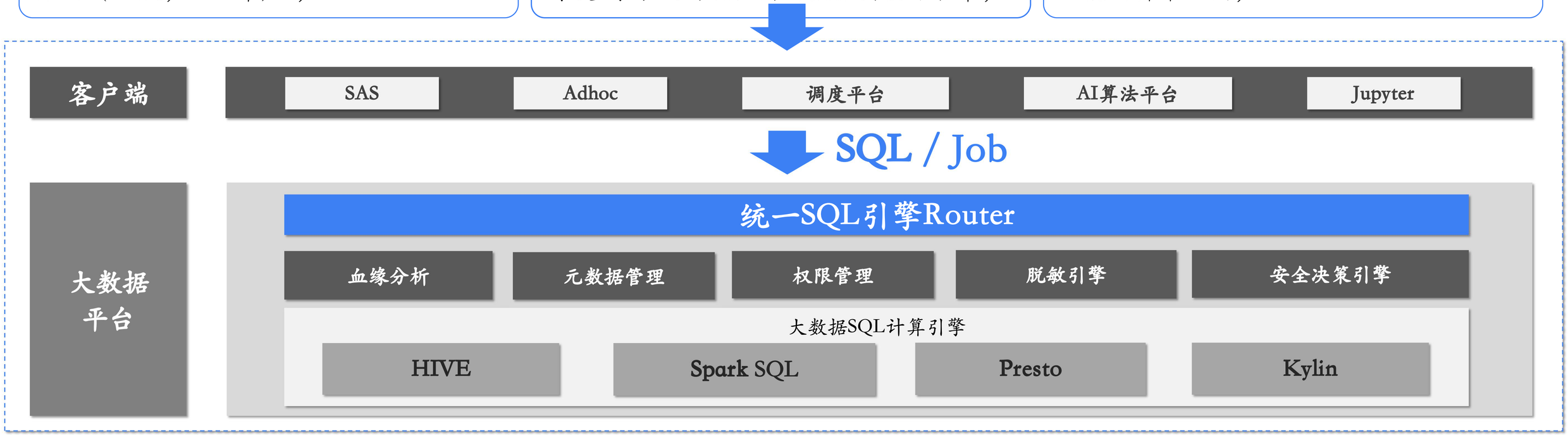
- 为解决大数据数据质量测试痛点，在测试环境无法完全复现生产问题，生产数据脱敏到测试环境仍有安全隐患，因此需要构建数据研发测试一体化平台，完善数据研发流程，满足监控合规的评审需求，数据需求闭环管理，数据开发、测试、变更流程统一管理，并和数据监控规则打通，保证全流程质量闭环



数据开发治理一体化解决方案-数据安全治理

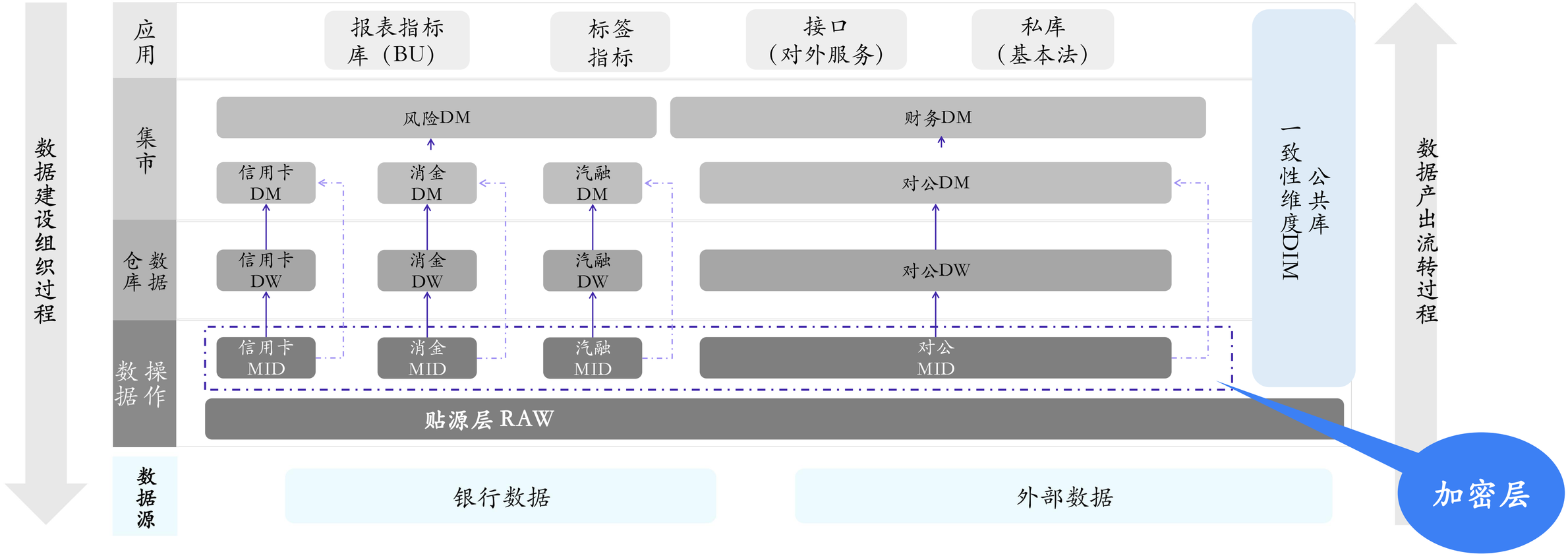
- 从事前、事中、事后分别管控数据安全。以“事中数据脱敏”为例，是通过在SQL/作业埋点用户帐号，分析SQL/Job对应的元数据字段，判断用户权限，返回用户对应的脱敏数据。

事前	事中	事后
事前制度建设：数据安全**“制度”先行**，为此我行修改制定了“平安银行数据安全管理办法（2.0版，2019年）”；	事中技术管控：采用**“数据加密”、“数据脱敏”、“敏感客群保护”、”智能阻断”、“数据外发”**等手段构筑强固的数据安全保护伞；	事后**监控审计**：基于规则引擎建立数据访问审计平台——实时的\自动+人工的识别可能的异常访问；



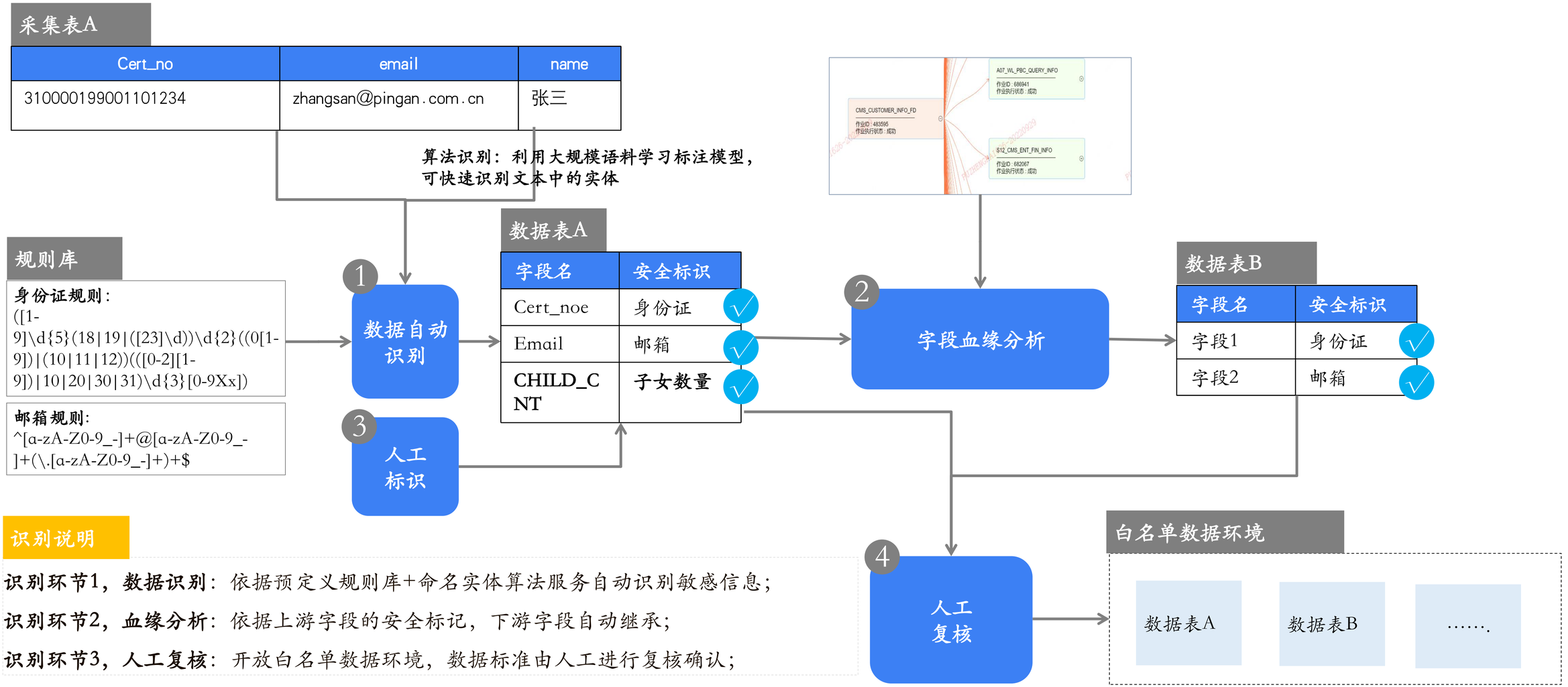
数据开发治理一体化解决方案-数仓分层加密处理过程

- ODS **贴源层(raw)**：敏感字段识别，利用 正则+算法+人工，识别出贴源数据表的敏感字段。
- ODS **加密层(mid)**：高敏感字段加密，将银行卡号，手机号，证件号进行加密储存。
- **数仓、集市等层**：利用字段级血缘关系，标识出每张表敏感字段。
- **数据查询访问**：应用端查询数据时，对统一查询中心(router)，根据访问的敏感字段及敏感脱敏类型进行脱敏处理。



数据开发治理一体化解决方案-敏感数据发现

- 源生产系统数据集成过程中，无论实时或者离线采集，开发治理一体化平台基于数据规则自动实现敏感数据发现；



识别说明

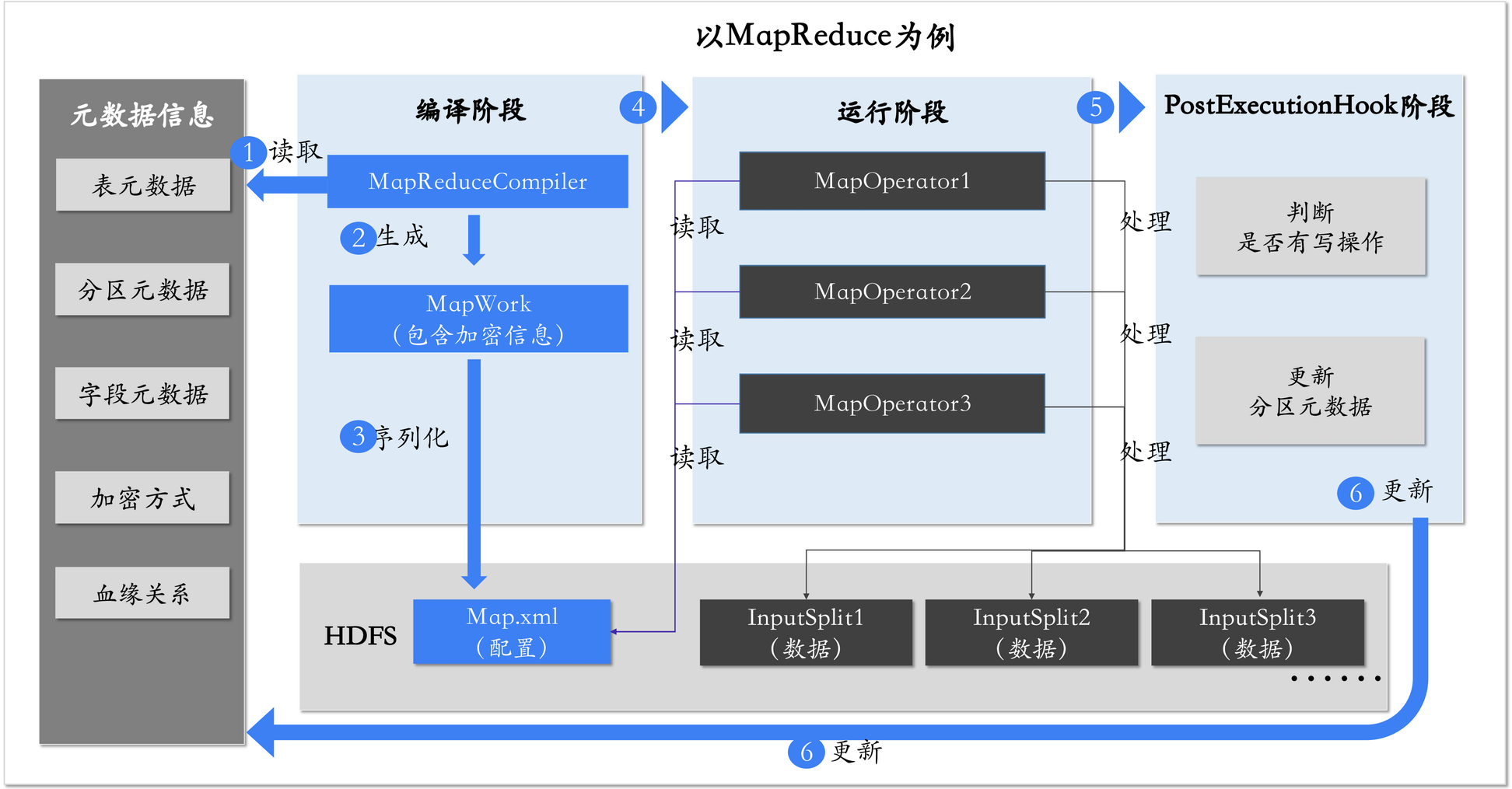
识别环节1，数据识别：依据预定义规则库+命名实体算法服务自动识别敏感信息；

识别环节2，血缘分析：依据上游字段的安全标记，下游字段自动继承；

识别环节3，人工复核：开放白名单数据环境，数据标准由人工进行复核确认；

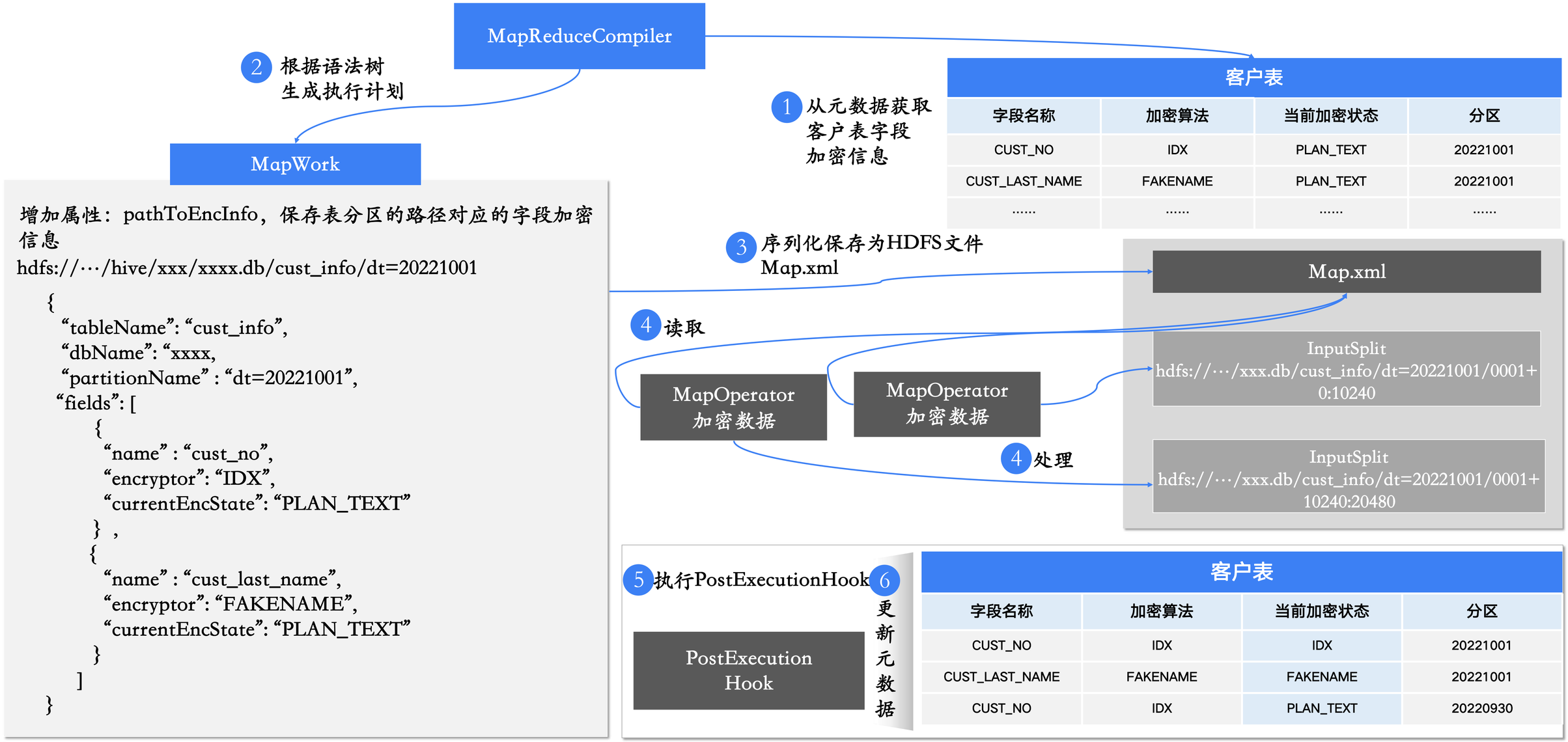
数据开发治理一体化解决方案-基于元数据的加密方案

- 通过元数据标记、底层执行过程中即时加密的方式，兼顾数据安全的同时，提升处理效率。



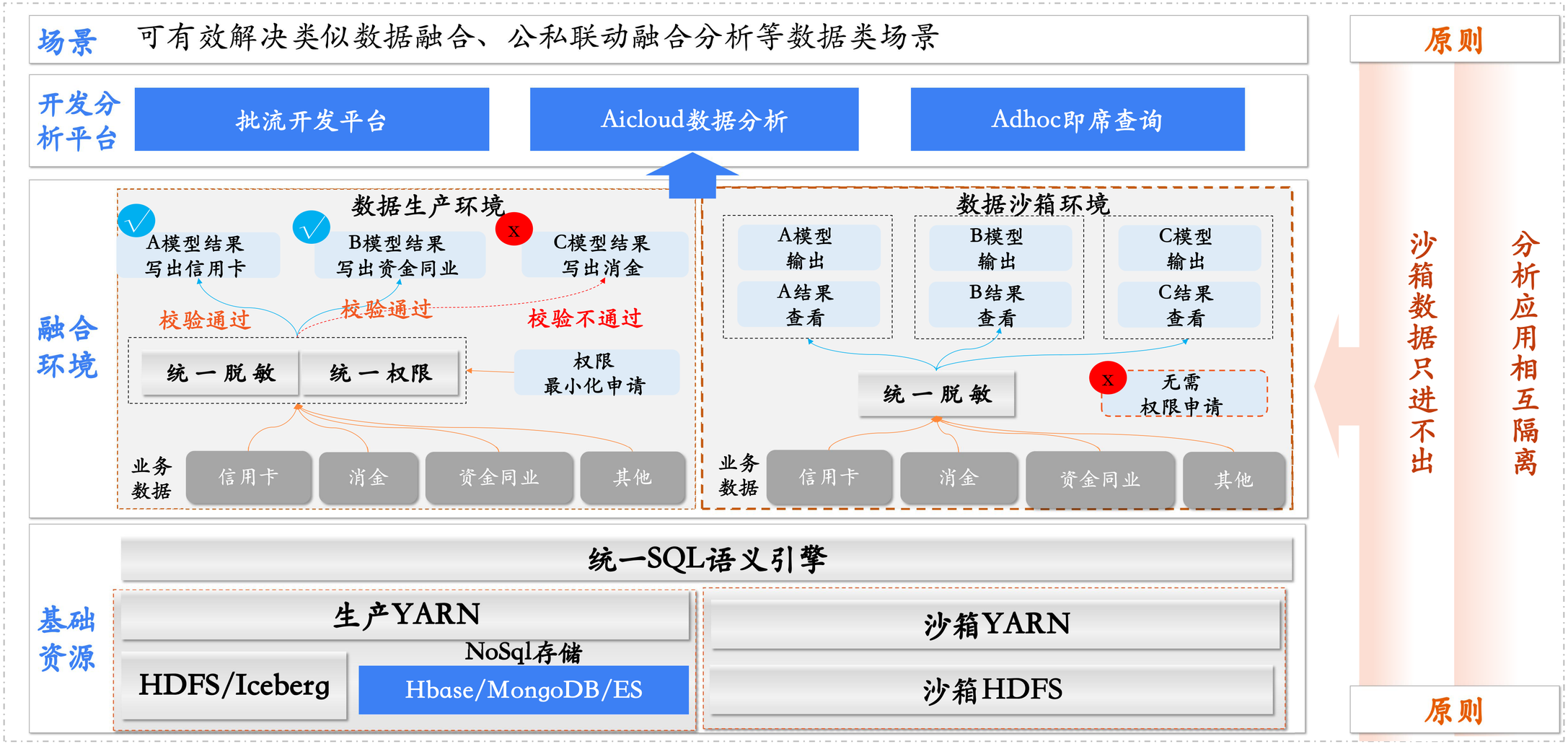
- **编译阶段**，调用元数据获取加密信息，给MapWork增加path到加密信息的映射
- **执行阶段**，MapOperator反序列化map.xml，获取加密信息，并初始序列化工具类，序列化工具根据加密信息**加密数据**
- **任务执行完成后**，根据执行计划，计算字段血缘并更新元数据表分区加密状态

数据开发治理一体化解决方案-基于元数据的加密方案（举例）



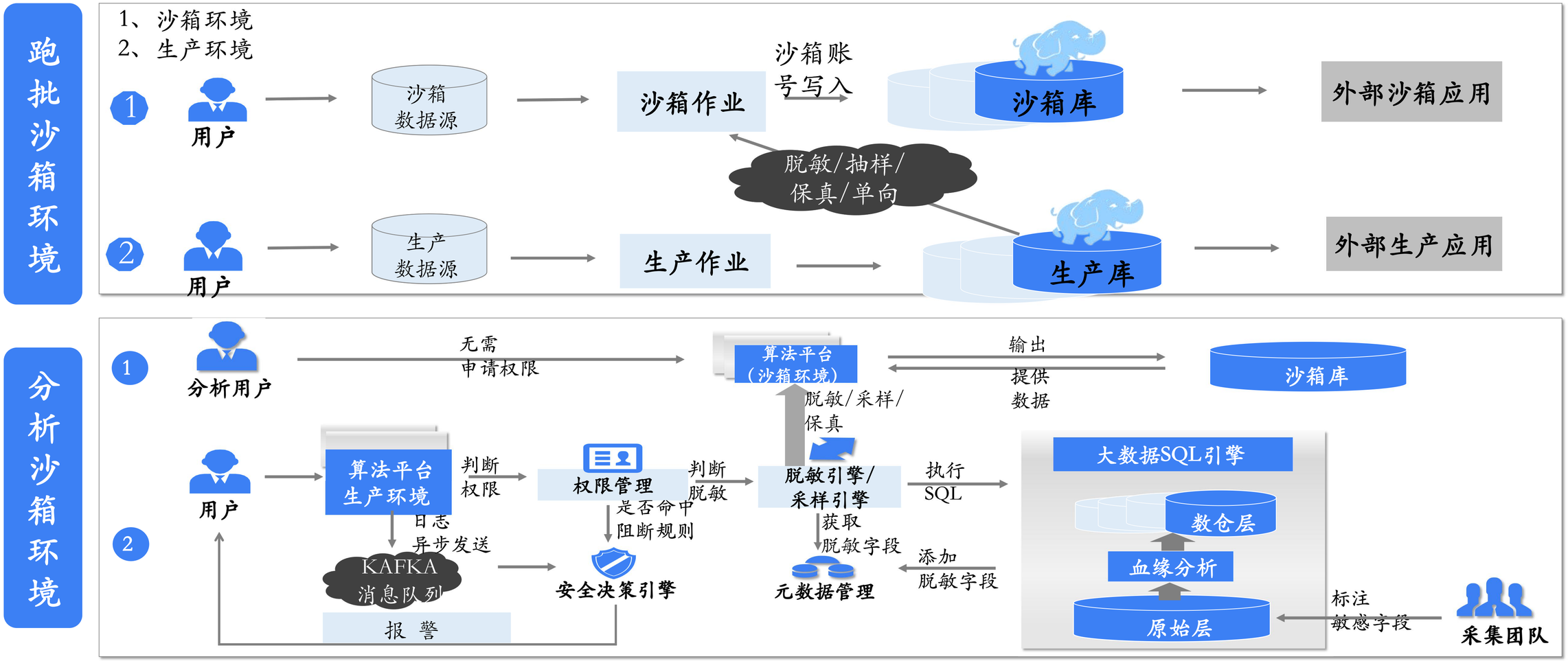
数据开发治理一体化解决方案-数据沙箱实现数据流通、安全共享

- 基于沙箱数据只进不出和分析应用相互隔离两大原则构建数据沙箱环境，差异化数据融合模式，确保安全可控要求下，提升训练和探索环节效率，便捷化数据应用通道。



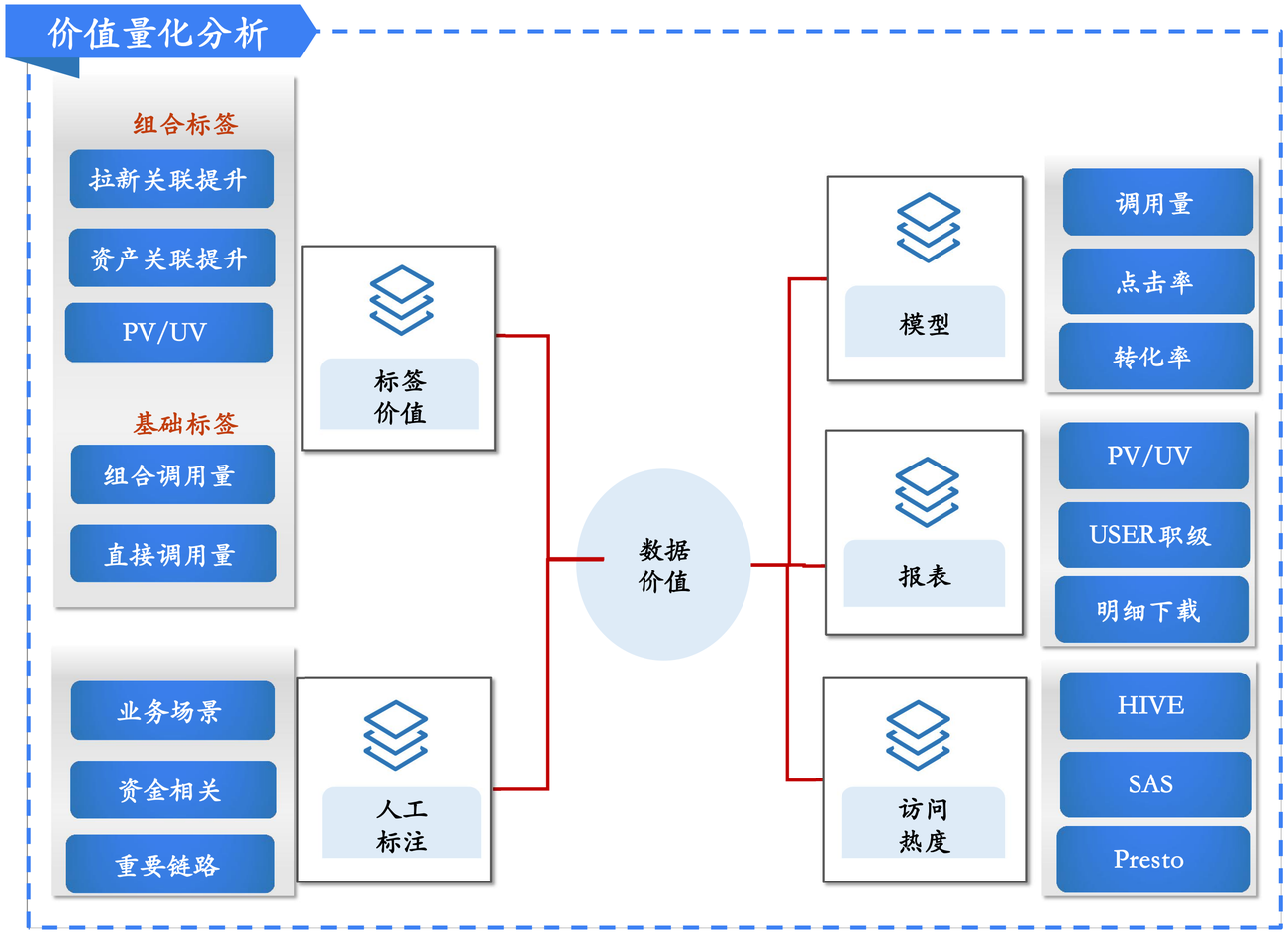
数据开发治理一体化解决方案-沙箱环境数据流程

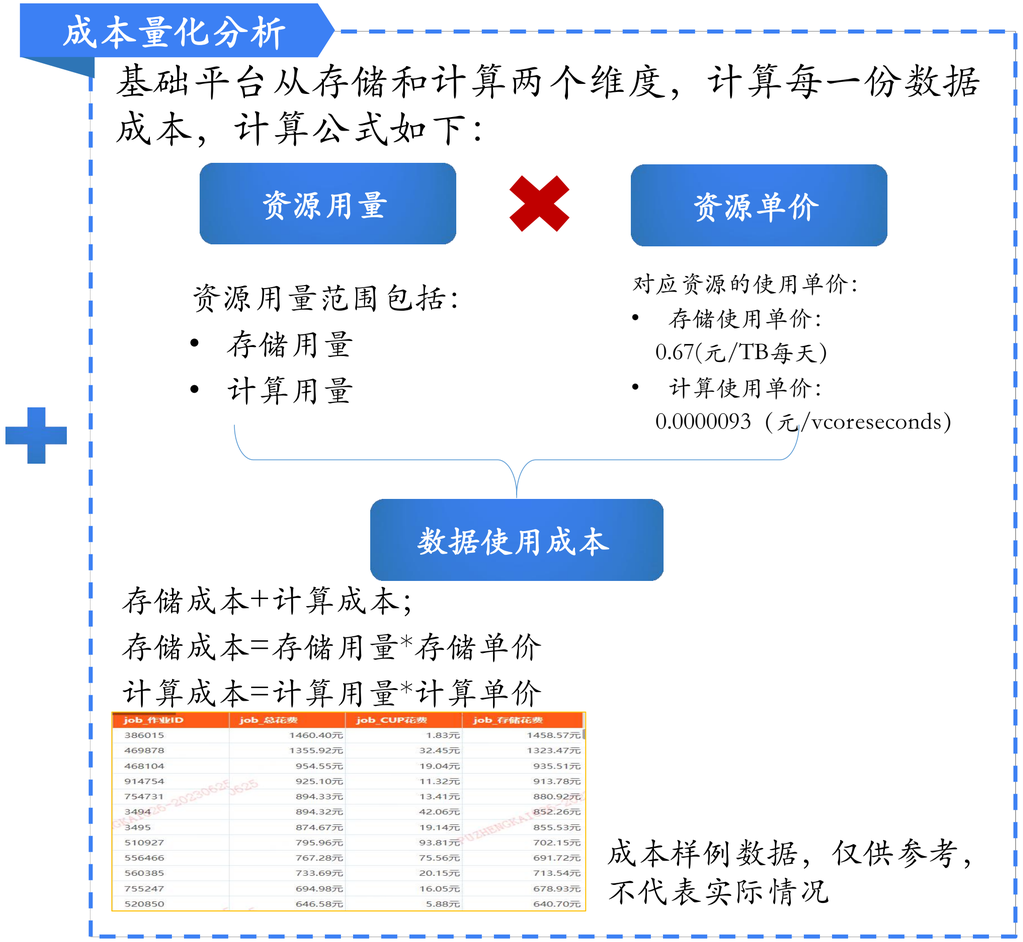
- 构建支持数据开发全流程的沙箱环境，确保与生产库分离，只进不出，数据采样。 既满足应用系统的沙箱环境数据探索需求， 同时提升数据研发使用效率。



数据开发治理一体化解决方案-成本价值管理能力

- 平台层面深化数据价值评估体系探索，实现数据成本与价值的多维度可量化分析，基于成本/价值实现数据资产的ROI分析以及成本治理。





job_作业ID	job_总花费	job_CUP花费	job_存储花费
386015	1460.40元	1.83元	1458.57元
469878	1355.92元	32.45元	1323.47元
468104	954.55元	19.04元	935.51元
914754	925.10元	11.32元	913.78元
754731	894.33元	13.41元	880.92元
3494	894.32元	42.06元	852.26元
3495	874.67元	19.14元	855.53元
510927	795.96元	93.81元	702.15元
556466	767.28元	75.56元	691.72元
560385	733.69元	20.15元	713.54元
755247	694.98元	16.05元	678.93元
520850	646.58元	5.88元	640.70元

数据资产沉淀-全周期数据资产化治理过程

	(1).资产产生	(2).资产认责	(3).资产管理与盘点	(4).资产编目	(5).资产服务
目标用户	数据加工人员	数据加工人员	资产管理人员 资产开发人员	资产运营人员	资产运营人员 资产使用人员
平台工具层	开发治理一体化平台 元数据检查 质量检查 血缘链路核验	开发治理一体化平台 业务属主定义	数据资产平台 数据分类识别 数据盘点	数据资产平台 资产目录运营 资产自动挂载	数据资产平台 资产查找服务 资产链路地图
数据治理层	数仓开发加工 指标加工 API服务加工	资产认定 定义资产业务属主	资产自动打标 (表类型/是否敏感) 资产信息盘点 资产信息变更 资产生命周期运营	数仓目录管理 资产目录挂载	资产全景地图 资产目录导航 资产场景搜索 打通资产场景

(1).**生成**：依赖**元数据治理规范**工具，检测通过的数据(元数据)，接口推送至数据资产平台

(2).**认责定义**：基于推送的数据(元数据)；定义业务属主和认责，将数据责任方界定清楚

(3).**自动盘点**：按照事前定义的业务全景图谱，依赖治理工具实现资产的自动打标，并最终完成分类盘点

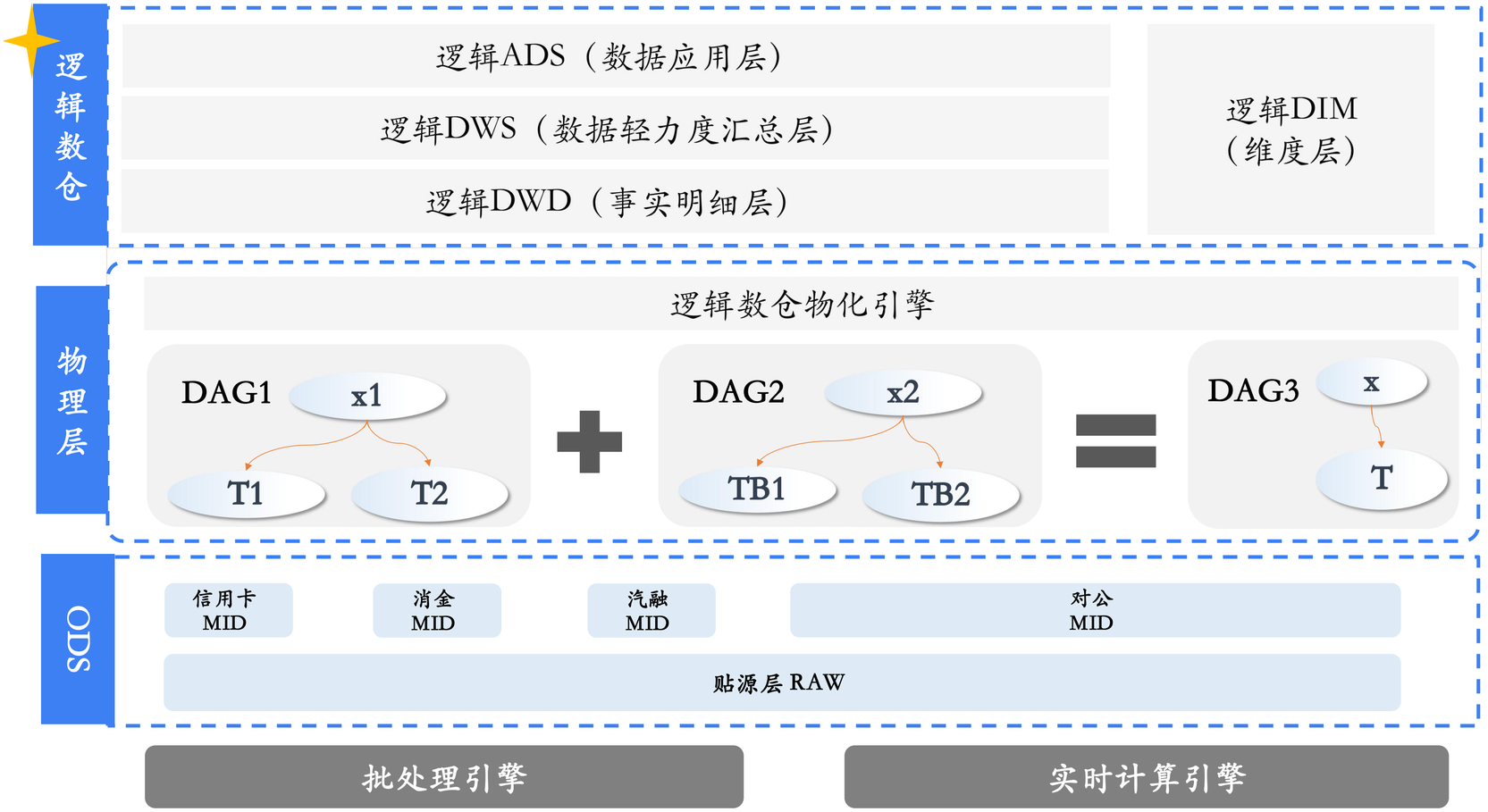
(4).**自动挂载**：基于第三步的自动盘点，完成对于资产目录挂载(事前治理侧需先完成标准资产目录维护)

(5).**资产化应用**：数据资产治理之后，结合数据价值/成本，面向数据用户，提供资产目录和搜索服务，并打通资产与使用场景的平台断点

目录

逻辑数仓- 从用户角度出发，基于数据使用情况自动化构建数据仓库

- 逻辑数仓以用户视角出发，以最大化数据价值和最优成本管控为目标，更敏捷响应用户需求，弱化繁琐的数据流ETL加工链路，让ETL工程师更专注企业通用模型设计，节约存储成本和管理成本



核心能力：

- **逻辑数仓层**

  构建面向用户和下游应用消费的逻辑数仓层，将逻辑表与物理表隔离，将物理表交给系统层优化

- **物理层智能调度**

  透明数据ETL逻辑和物理存储介质，由逻辑层用户行为和需求触发，实现数据生产链路的**智能编排**和**调度**，针对重复、相似计算进行**自动合并**，下线或降权无效、低频、低价值数据生产

- **性能自优化**

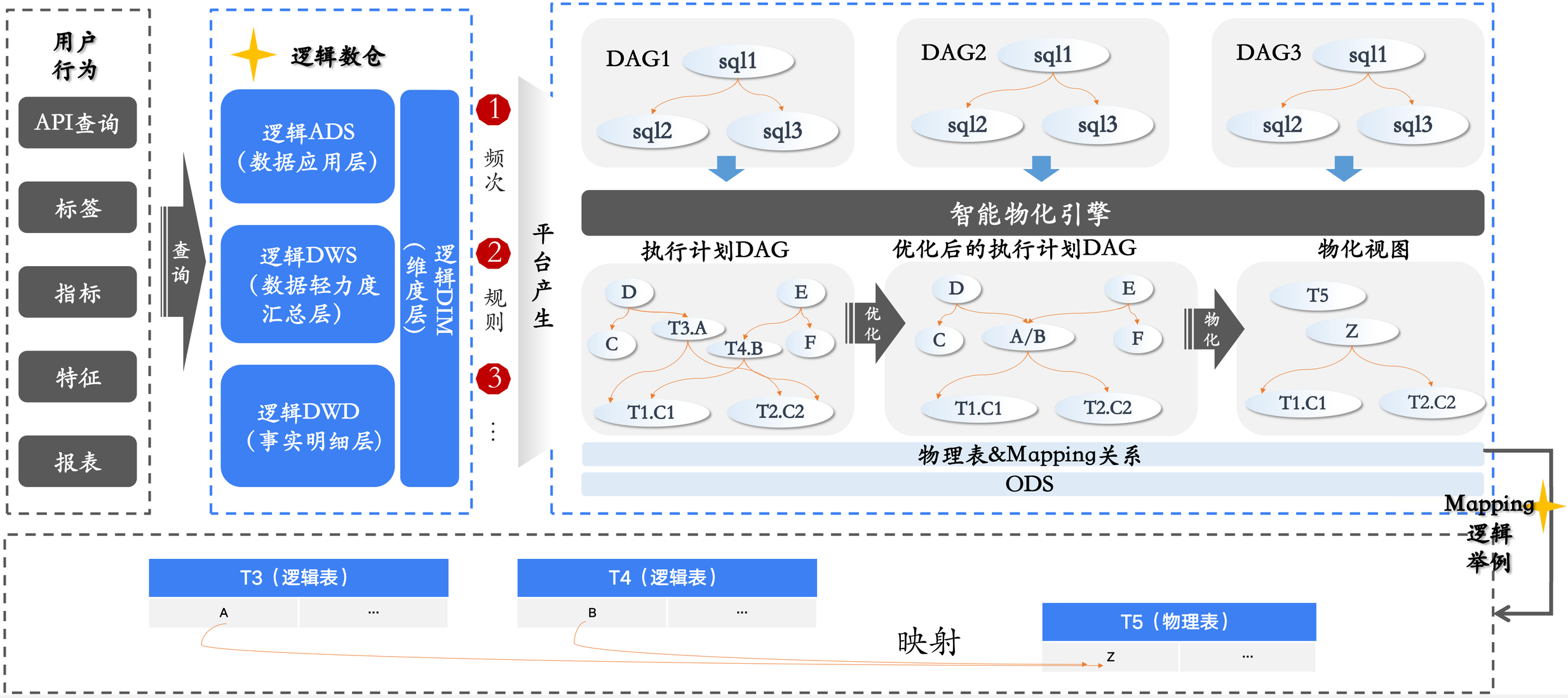
  基于用户查询行为实现自适应的查询性能优化，自动实现物化、缓存或构建Cube/索引

- **从被动到主动的数据治理，实现“数据自动驾驶”**

  逻辑层基于业务需求快速调整，物理层自适应上层调整，识别数据核心资产元数据

逻辑数仓- 整体设计

- 改变数仓开发模式，让数据人员更关注业务开发，解决大数据平台成本暴增问题，让平台做到主动数据治理



想一想，我该如何把这些技术应用在工作实践中？

THANKS