

蚂蚁安全领域 智能化数据治理

高级数据技术专家 / 霄元（黄国龙）

个人简介

黄国龙 高级数据技术专家

- 20年加入蚂蚁，目前是安全大数据团队数据智能资产、内容数据资产负责人
- 0-1主导建设蚂蚁集团内容安全数据资产专项
- 关注大数据研发、数据架构、数据治理等领域

目录

数据治理
通用能力

蚂蚁数据治理架构及能力

数据治理
业务案例

安全领域智能化数据治理实践

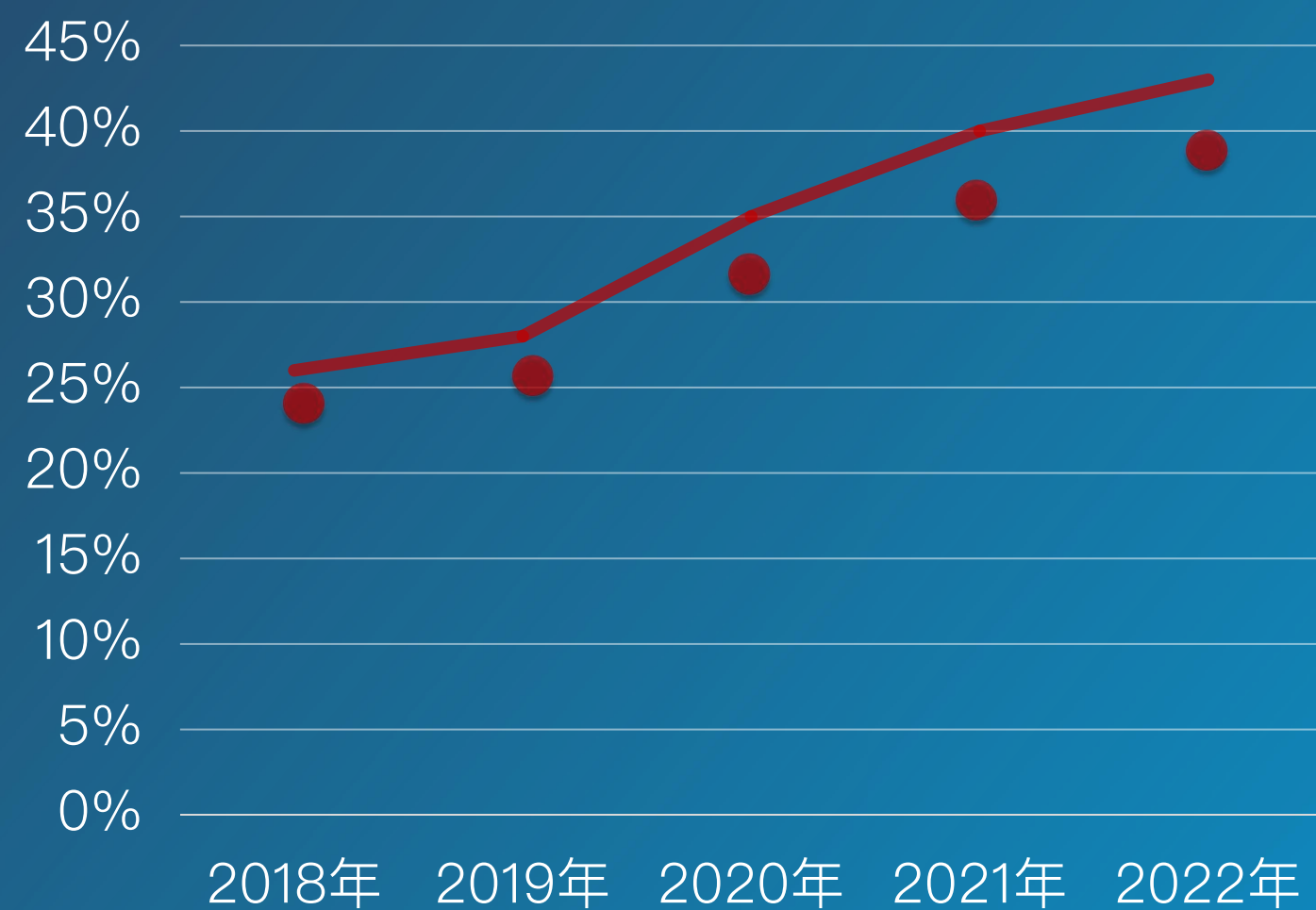
数据治理
创新探索

数据治理探索与创新

蚂蚁数据治理架构及能力_面临挑战

40%+

存储增速



成本增速快

- 存储资源：蚂蚁存储达 EB 级别，年增速40%+；
- 计算资源：计算资源 千KCU/日。

业务需求多

- 新业务资源需求旺盛；
- 人工智能发展快，数据需求呈现爆炸性增长。

成本看不清

- 资源使用细节看不清；
- 成本很难分摊到业务。

蚂蚁数据治理架构及能力_治理思路



能力建设

平台提效 & 技术治理

平台提效：将治理能力产品化服务用户

技术治理：结合引擎能力升级做技术优化

无效资产治理

重复资产治理

TOP 资产治理

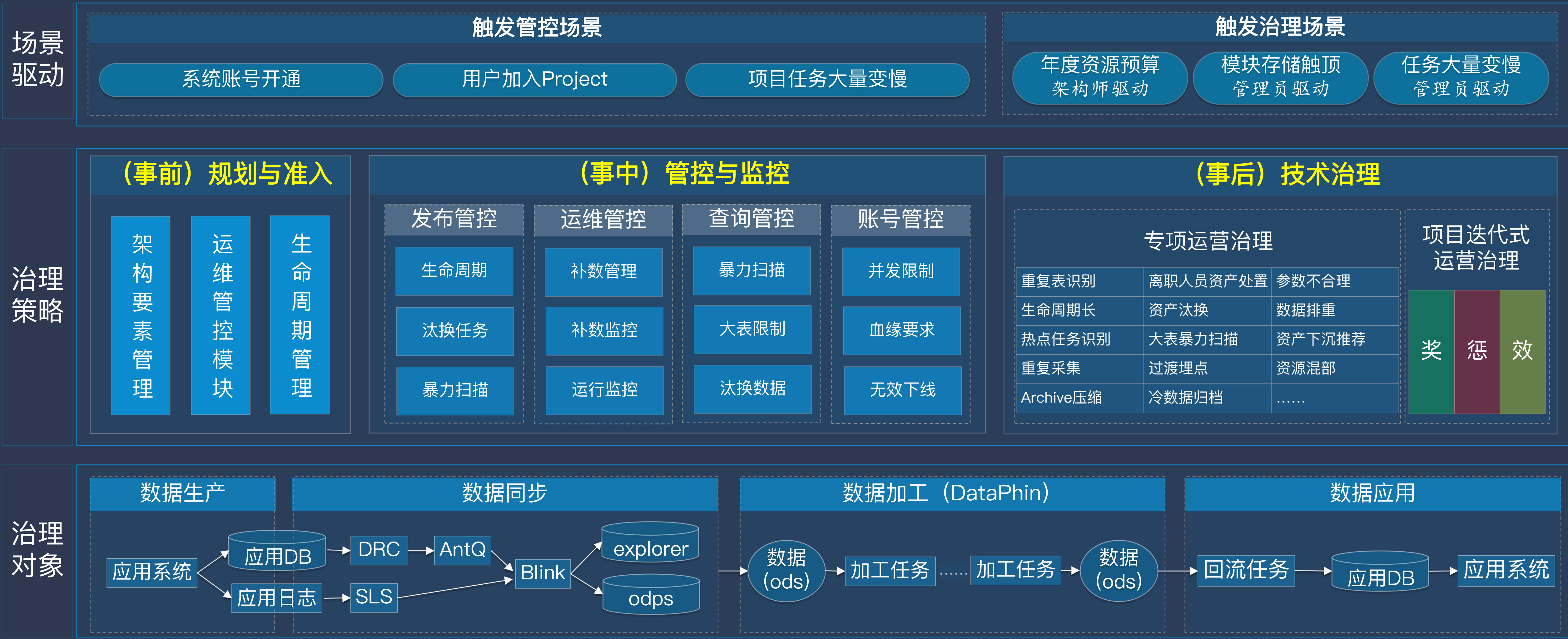
存储治理

计算治理

集群治理

蚂蚁数据治理架构及能力_治理方案

关键能力建设方案：从事前 -> 事中 -> 事后，构建成本治理全链路能力。



蚂蚁数据治理架构及能力_治理架构

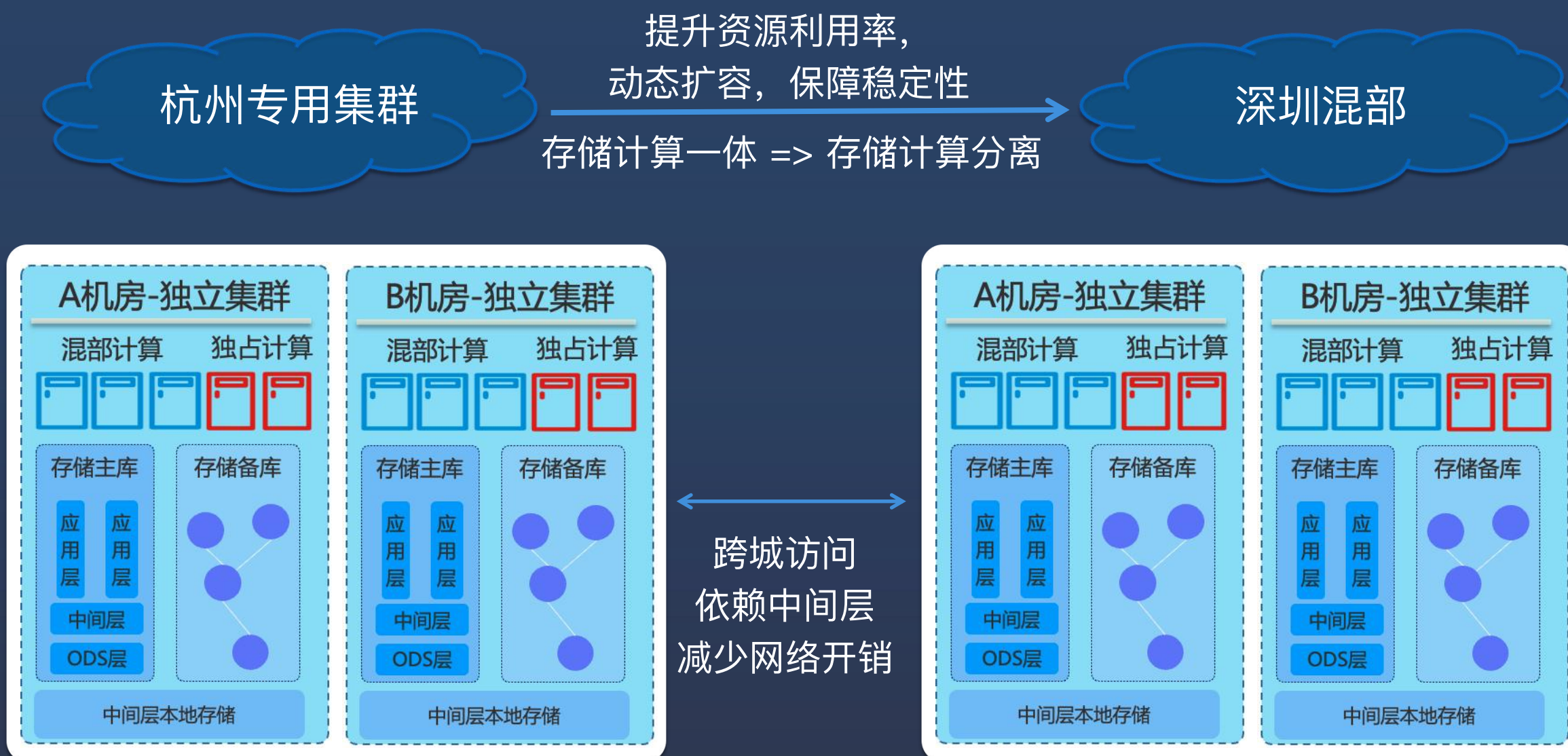


蚂蚁数据治理架构及能力_资源治理领域案例

集群资源混部

推进在离线混合部署，计算算力会提升10%，机器成本降低25%.

混合部署方案



迁移混部（四个阶段）



开源

蚂蚁数据治理架构及能力_资源治理领域案例

关键技术

提升治理自动化率，实现自动识别、归因分析、自动清理，形成常态化管控能力。

引擎优化		模型优化		代码优化		资产管理优化							
参数优化	调度优化	数仓模型	代码设计	计存设计	数据格式	JOIN优化	UDF优化	数据倾斜	聚合优化	生命周期	无效表	计算浪费	重复资产
• Split Size • 小文件合并 • Reducer instantce • CPU数 • Dynamic parallelism	• 任务归并 • HBO优化 • 集群混部 • 错峰运行 • 冷热分层	• 业务领域建模 • 抽象公共层 • 通用应用层 • 配置化指标系统 • 大宽表设计	• 全量改采样 • with替代tmp表 • 视图化改造 • 避免数据膨胀 • 执行顺序优化	• 渐进计算 • 累计计算 • Zorder • Shuffle优化 • Bitmap索引 • 全改增 • 极限存储	• 重排压缩 • Cube预计算	• Map join • Hash join • Skew join • Dynamic Filter	• 内置替换 • 提前计算 • 参数调优 • 本地缓存	• Map端 • Reduce端 • 热点值	• Grouping sets • UDF转UDTF • Count(distinct)	• Map join • Hash join • Skew join • Dynamic Filter	• Map端 • Reduce端 • 热点值	• 临时表 • 系统表 • 长周期表 • 回收站优化 • 大字段生命周期	• 同链路相似表 • 相似任务节点 • 分区不更新 • 缓慢变化维表

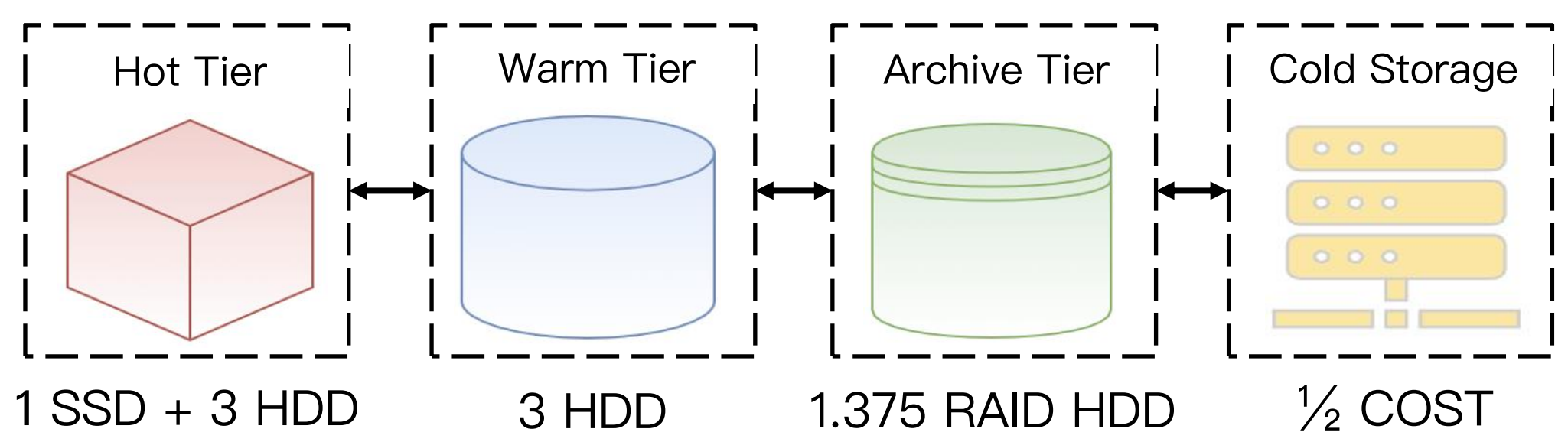
节流

蚂蚁数据治理架构及能力_资源治理领域案例

关键技术-实例

冷存体系

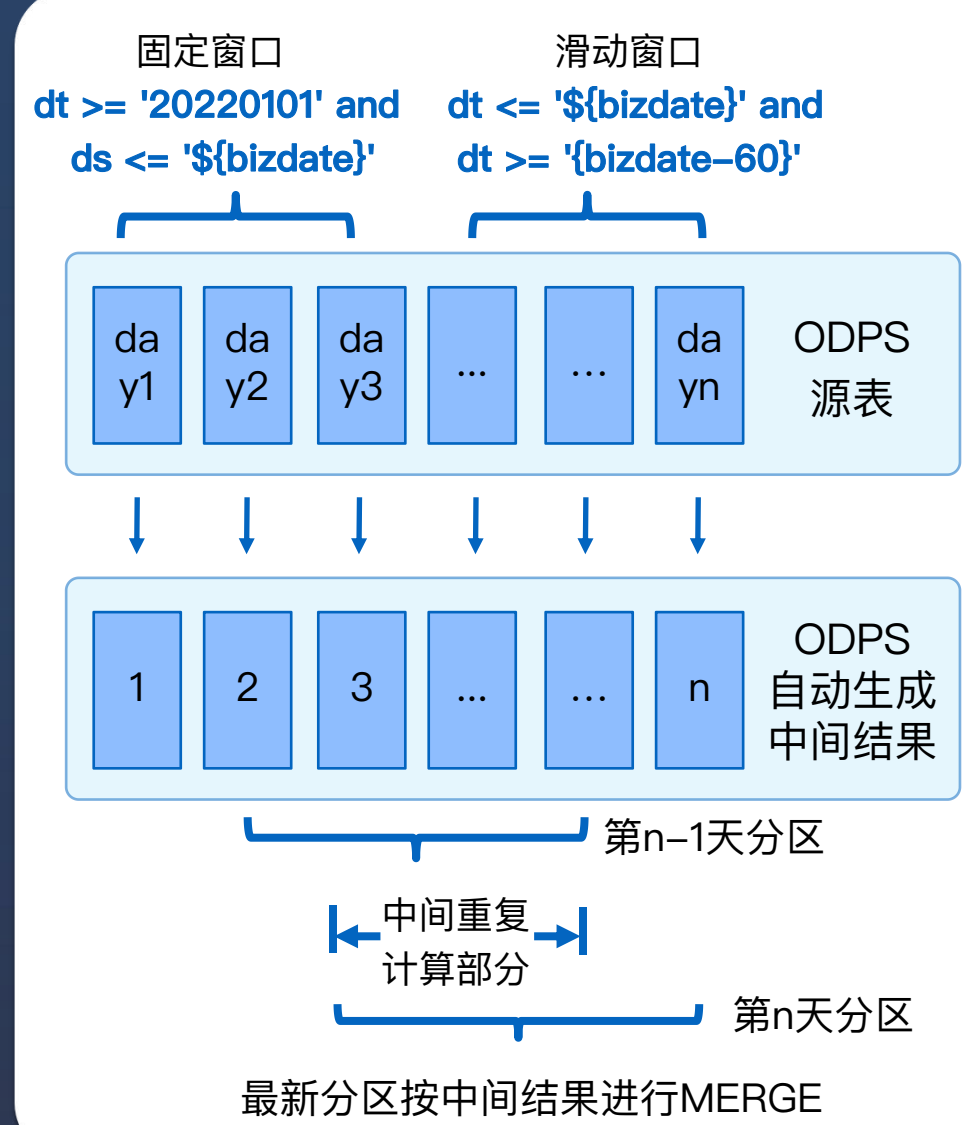
推进在离线混合部署，计算算力会提升10%，机器成本降低25%。



- Hot Tier: 高频消费的热点数据、优化I/O;
- Warm Tier: 热数据、读取频率正常;
- Archive Tier: 数据需长期保留, 访问频次底;
- Cold Storage: 长期保留, 超低频访问。

渐进计算

设置成渐进计算后，每日计算消耗从795CU降到22CU。



原理：空间换时间，自动生成中间表，避免重复计算，其中中间表可采用hash cluster, 提升merge阶段Shuffle效率； Odps支持一键渐进计算、设置一个参数即可。

```
SELECT userid,
       SUM( CASE WHEN **** END) ,
       COUNT(CASE WHEN **** END),
       MAX(CASE WHEN **** END)
from aseccdm.dwd_sec_evt_
where dt > '${bizdate_90}'
and dt <= '${bizdate}'
and ((coalesce(biz_no_1, '') <> '')
or (event_name = 'cashecouponchange')
or (event_name = 'p...')
or (event_name = 'd...it' and length(biz_no) > 0))
group by userid
```

节流

蚂蚁数据治理架构及能力_基础平台能力实例

（事前）发布管控

由平台或业务方事先制定发布管控规则，相关任务上线提交时、平台自动检验结果，如有规则未校验通过，此次上线发布失败。

发布管控规则说明

规则集管控范围和条件说明

规则说明

【D2&DG】表分区命名规范校验

定位：输出表的表分区必须满足管理员设定的表分区名模板，如不满足，则校验为「不通过」

适用类型：SQL任务

【D2&DG】任务命名规范校验

定位：用户提交的任务名称/节点名称必须满足管理员设定的任务名模板，如不满足，则校验为「不通过」

适用类型：全部任务

【D2&DG】禁用关键字校验

定位：提交的ODPS SQL中凡是命中配置的关键词，即校验为「不通过」

适用类型：SQL任务

1.选择校验规则：规范类—命名规范—禁用关键字校验

6

禁用关键字校验 (规范类/命名规范)

阻断

删除

规则参数:
(点击进行修改)

规范类 > 命名规范 > 禁用关键字校验

质量类 > 模型规范 > 禁止节点名称和输出表名不匹配校验

性能类 > 中间层规范 > 禁止跨任务使用临时表校验

一、发布管控产品需求和规则说明

规则集管控范围和条件说明

规则说明

【D2&DG】表分区命名规范校验

【D2&DG】任务命名规范校验

【D2&DG】禁用关键字校验

【D2&DG】禁止跨任务使用临时表校验

【D2&DG】输出表命名规范校验

【D2】JAR包包名规范管控

【D2&DG】输出表名和任务名保持一致校验

【D2&DG】输入表禁止挂依赖

【D2&DG】禁止create特定前缀表

【DG】数据模型定义管控

【DG】发布管控资产判重

【DG】发布管控防穿透

【DG】Aone关联需求发布管控

【D2&DG】表生命周期最大值校验

【D2&DG】任务输出表数量限制校验

【D2&DG】表字段注释长度校验

【D2&DG】禁止SQL只含select校验

【D2&DG】禁止where条件包含==符号校验

【D2&DG】特定表生命周期最大值校验

【D2&DG】SQL中除数为0处理校验

【D2&DG】SQL中NULL参与计算处理校验

【D2&DG】禁止当前节点有对多个正式表的i...

【D2&DG】禁止空查询校验

【D2&DG】禁止in或not in存在子查询校验

【D2&DG】禁止使用select *校验

任务发布状态详情

预提交状态

规则名	规则说明	支持者	校验结果
冒烟测资源耗费和运行时长限制	运行资源 ?		未通过
禁用关键字校验--不校验注释	禁用关键字校验--不校验注释?t ?		未通过
禁止DG任务外部依赖D2	DG节点禁止通过check方式依赖D2任务，需要通过影子链路的形式，有问题请咨询雨淞。 ?		通过
禁用关键字校验	禁用关键字校验 ?		通过
禁网发布规则（新）	禁网发布设置 ?		通过
禁用关键字校验	禁用关键字校验 ?		通过

OK

ArchSummit
全球架构师峰会

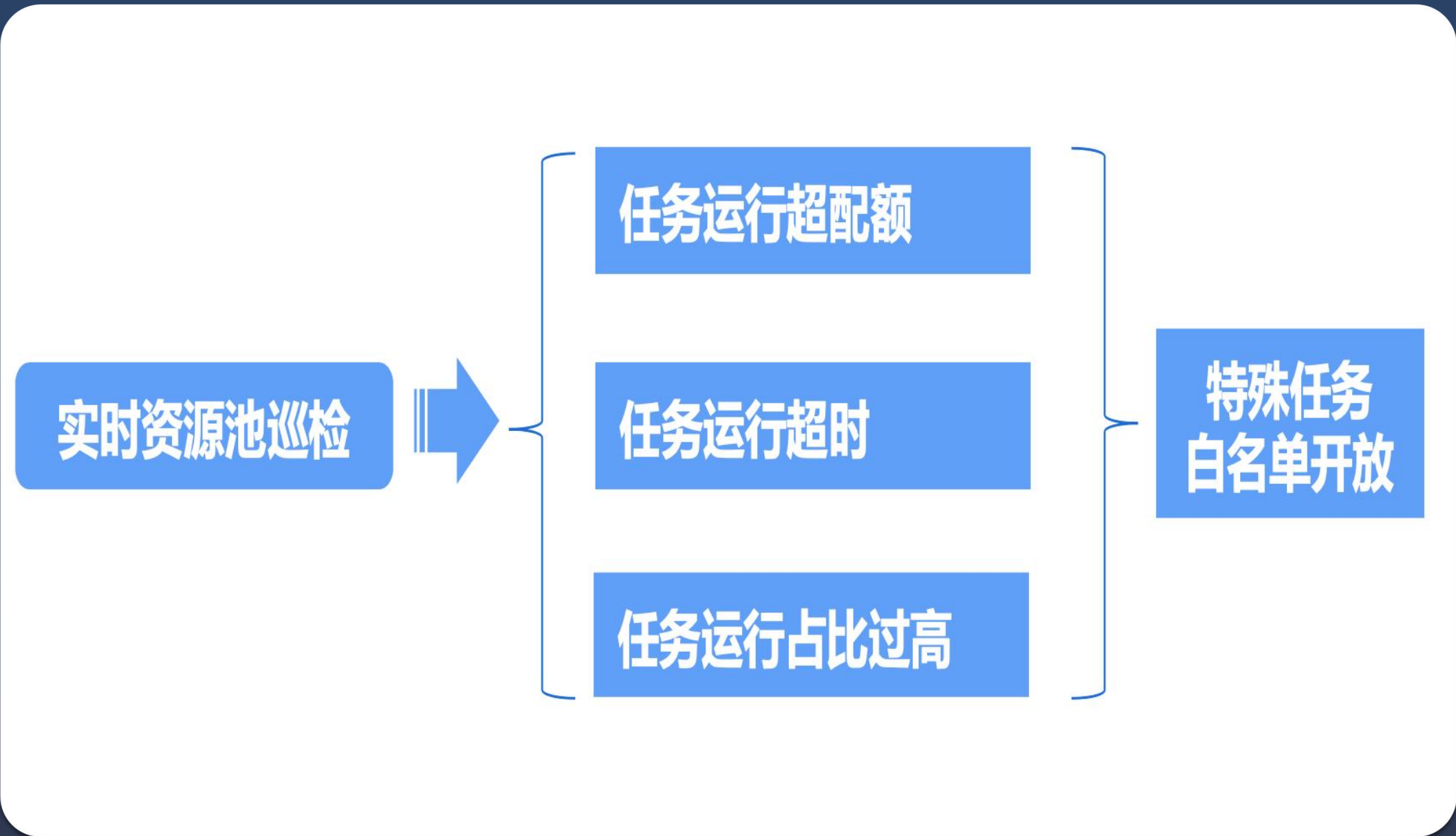
InfoQ 极客传媒

蚂蚁数据治理架构及能力_基础平台能力实例

(事中) 实时巡检

禁止随意提交超大任务导致整个资源池的堵塞和打满，进而造成高昂的成本消耗，同时也兼顾效率，仅对异常使用进行管治。

异常大任务自动查杀方案



异常大任务自动查杀实例

项目名称: a-实例

20221125151538929gebd2g6fn375累计资源使用超告警阈值

- 告警名称: 环境禁止20CU以上的任务
- 项目名称: a
- 作业提交人:
- instance_id: 20221125151538929gebd2g6fn375
- 快照时间: 2022-11-25 23:46:07
- 告警时间: 2022-11-25 23:48:23
- cpu累计使用: 20.0(CU), 当前阈值: 20(CU)
- gpu累计使用: 0(CU), 当前阈值: (CU)
- 内存累计使用: 26.0(GB), 当前阈值: (GB)
- 作业自动查杀结果: 作业已被自动查杀

[查看logview\(代码\)](#)
[历史告警事件](#)
[一键查杀作业](#)

项目名称: ev

- 告警名称: 资源消耗超过90%
- 作业提交人:
- instance_id: 20221125125009480gcnliffc2465
- 快照时间: 2022-11-25 20:56:07
- 告警时间: 2022-11-25 20:57:34
- cpu利用率: 395.30%, 当前阈值: 90.00%
- gpu利用率: 0.00%, 当前阈值: 90.00%
- 内存利用率: 123.53%, 当前阈值: 90.00%
- 作业自动查杀结果: 作业已被自动查杀

[查看logview\(代码\)](#)
[查杀作业](#)
[历史告警事件](#)

阈值有疑问请联系告警添加人:

查杀异常或者优化请加群联系值班同学处理: 11799056

[detail](#)

蚂蚁数据治理架构及能力_基础平台能力实例

(事后) 成本健康分

基于累计浪费和时间加成，设计成本健康分算法。通过健康分管理数据平台使用权限。

成本健康分算法

$$\text{个人健康分} = \begin{cases} f(\sum_{\text{发现日期}}^{\text{统计当天}} \text{治理项每日实际浪费}) & \text{可量化治理项} \\ f(\sum_{\text{发现日期}}^{\text{统计当天}} \max(\text{治理项上游每日实际浪费, 时间加成})) & \text{不可量化治理项} \end{cases}$$

设计理念：基于累积浪费和时间加成，越久治理，扣分越多

健康分产品运营实例

您的健康分因连续8天低于85分，现已被限制研发平台查询权限

存储治理

健康分：83.5分

待治理任务：5个

累计浪费存储：49.54TB

计算治理

健康分：92.5分

待治理任务：2个

累计浪费计算：9.85CU

质量治理

健康分：99.48分

待治理任务：2个

去治理

成本优化指南

治理工作台

治理优化手册 管理我的数据表 我的白名单

存储 86分 待治理对象 3个 累计浪费 15.62 TB 健康分 ① 已治理对象 3个

计算 92分 待治理对象 1个 累计浪费 11.04 CU 健康分 ① 已治理对象 4个

质量 99.4分 待治理对象 2个 白名单 0个 健康分 ①

合规 待治理对象 0个 已治理对象 0个 白名单 0个

模型 暂无数据

表(3) DI报表(0) 标签(0)

重复资产_父子重复 (1) 是否扣分: 是

无访问资产 (1) 是否扣分: 是

生命周期过长 (1) 是否扣分: 是

资产策略说明: 系统检测到您的表, 完全可以由您读取的表代替, 建议您直接使用上游表活改成视图, 减少重复拷贝 查看治理方案

表名: 请输入表名 项目名称: 请输入项目 治理进度: 待处理 重置 展开

表名	白名单状态	治理进度	治理方案	发现E	操作
	未申请	待处理	{“重复的表”:“od...	2023 至今	申请白名单 修改生命周期 更多操作

蚂蚁领域数据治理架构及能力_治理成果

安全领域治理成果总结

已全部达成年度目标，预估节约数据成本 **25%+** 。

治理专项	存储用量(PB)	计算用量(KCU)	表数量(张)
累计治理收益	百PB+	百KCU+	百万+

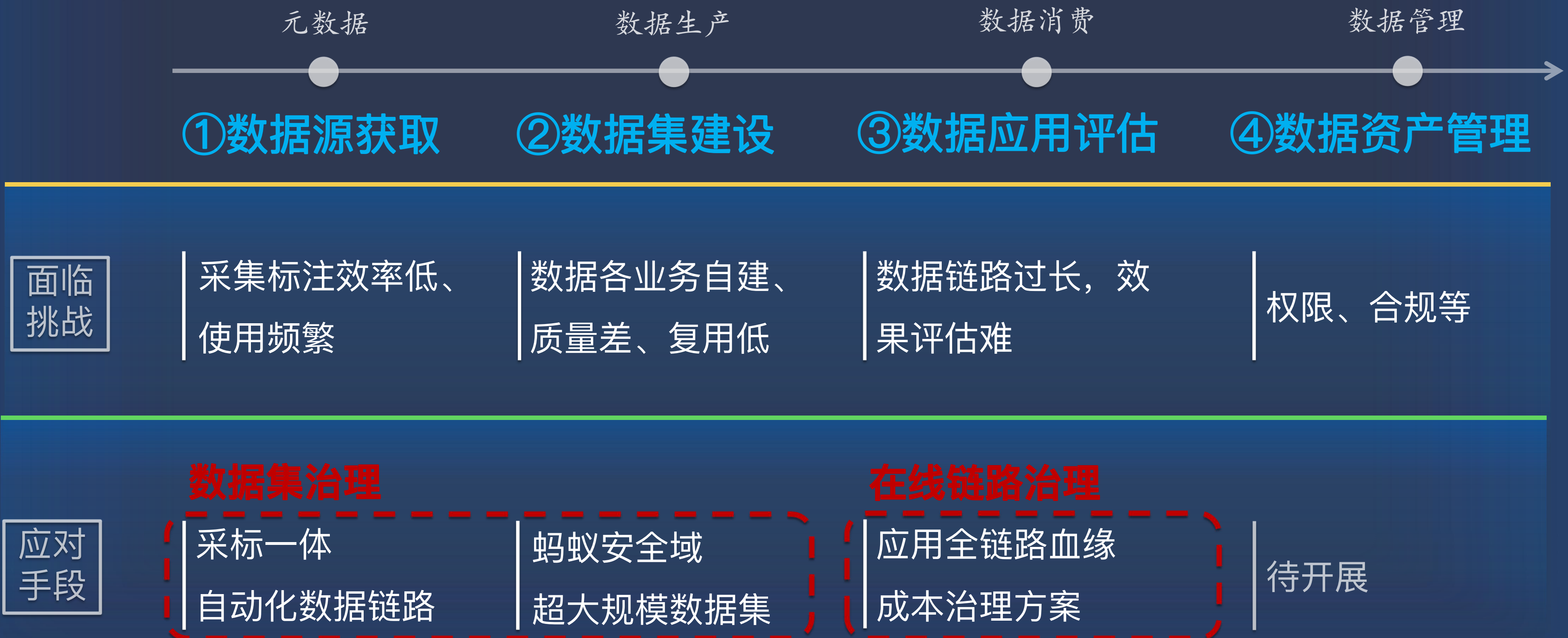
安全领域智能化数据治理实践

数据智能资产建设板块



安全领域智能化数据治理实践

智能化数据治理思路



安全领域智能化数据治理实践



数据集建设阶段，搭建采集标注自动化数据链路降本增效；标准化建设蚂蚁安全域超大规模数据集消除数据孤岛……

数据集治理

采标一体自动化数据链路
蚂蚁安全域超大规模数据集

在线运行阶段，搭建元数据之应用全链路血缘，助力在线模型策略成本治理优化……

在线链路治理

应用全链路血缘
在线模型成本治理方案

安全领域智能化数据治理实践_数据集治理

数据集简介

高品质、多样性、大规模的数据集建设是AI技术应用竞争关键要素之一，在建设大规模数据集过程中，百PB级别的安全数据资产必将带来高昂成本支出、及各种质量风险隐患

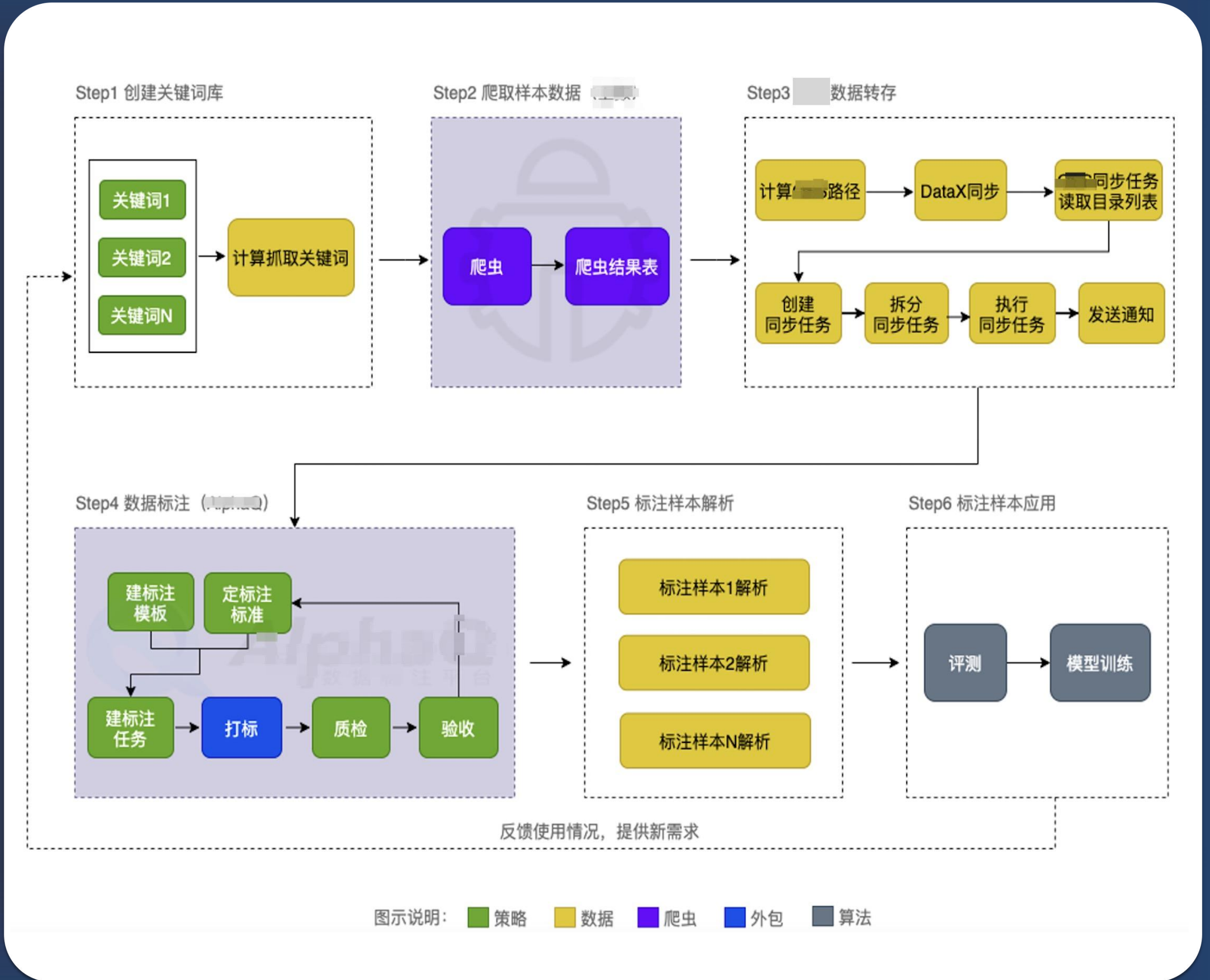
数据集主要有：采集、标注、大规模数据集、训练数据集、评测数据集。

安全领域智能化数据治理实践_数据集治理

采标一体化链路

数据采集、人工标注是数据集建设关键环节之一、也是首要事项，相关事项涉及合作方众多、且时间也不可控，在人工对接中费时费力。

采标一体化自动化数据链路，从关键词计算、对接采集、转存及通知、对接标注、数据ETL全链路实现自动化，降本增效明显、且数据品质也有保障，大大缩减人力成本、将原来采集标注2周以上时效降到5天以内。

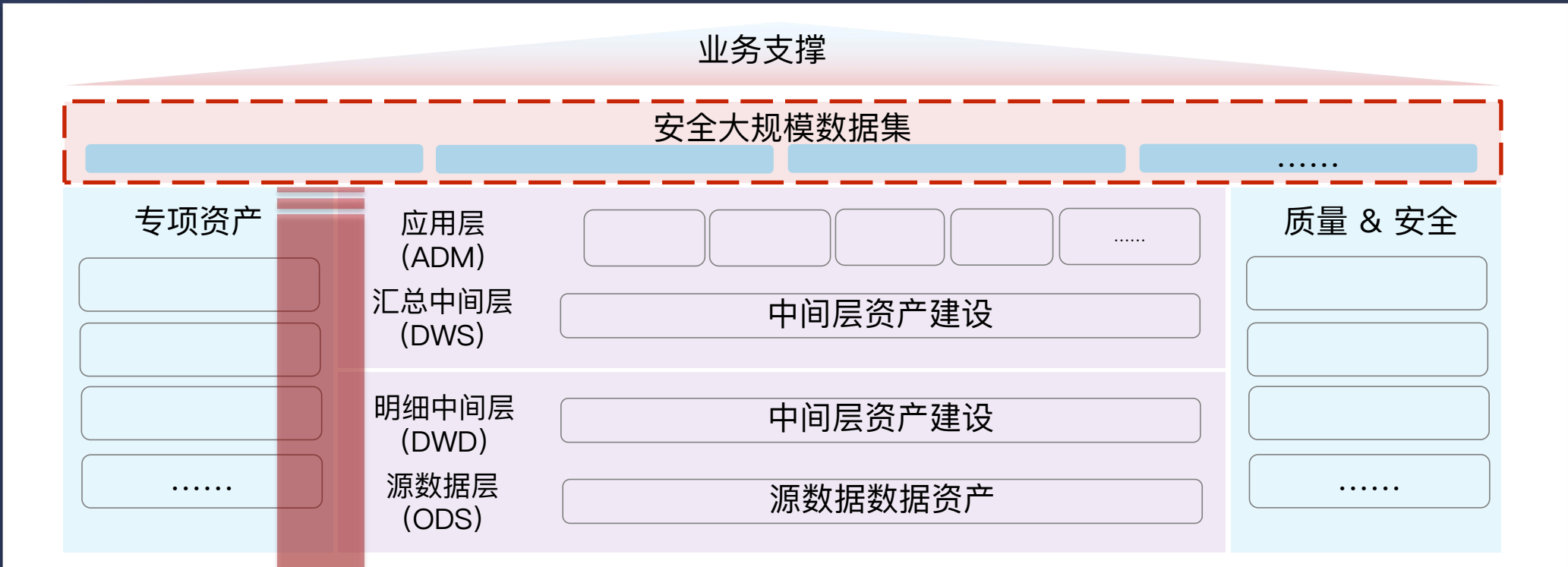


安全领域智能化数据治理实践_数据集治理

大规模数据集

智能数据建模设计架构

标准数仓建模设计确保数据品质，通过大规模数据集支撑业务。



大规模数据集分类体系

规范数据集分类体系，消除数据孤岛、共享数据集资源。

类型	来源	数据集	规模（存储/量级）	类型	特点
内容域	蚂蚁外	MMC4	存储：512G / 文本：23,085 存储：1.8T / 特征：23,085	文本图片对	十亿图像语料库Multimodal-C4 fewer-faces数据
		Laion-5b	存储：24.25T / 量级：3,769,298,024	文本图片对	非人工注释的多模态图文数据集， 训练图文匹配能力
		语音对话数据集	存储：496.9G / 量级：1,982,402.	音频	群聊、对话等场景音频数据
	蚂蚁内	纸币虚拟币数据集	存储：164G / 量级：1,086,253	图片	人民币、美元、比特币等相关图片 样本数据
		蚂蚁图片内容源	存储：***T / 量级：*亿	图片	
资金域	蚂蚁内	生成式(AIGC) 伪造图片标注数据集	存储：***M+ / 量级：***亿+	图片	赌博等图片集
		事件特征数据集	存储：***T / 量级：*亿	文本	

实例：*数据集数据建模设计

数据集实例，整合资金各业务特征、标签形成全域样本集。



安全领域智能化数据治理实践_在线链路治理

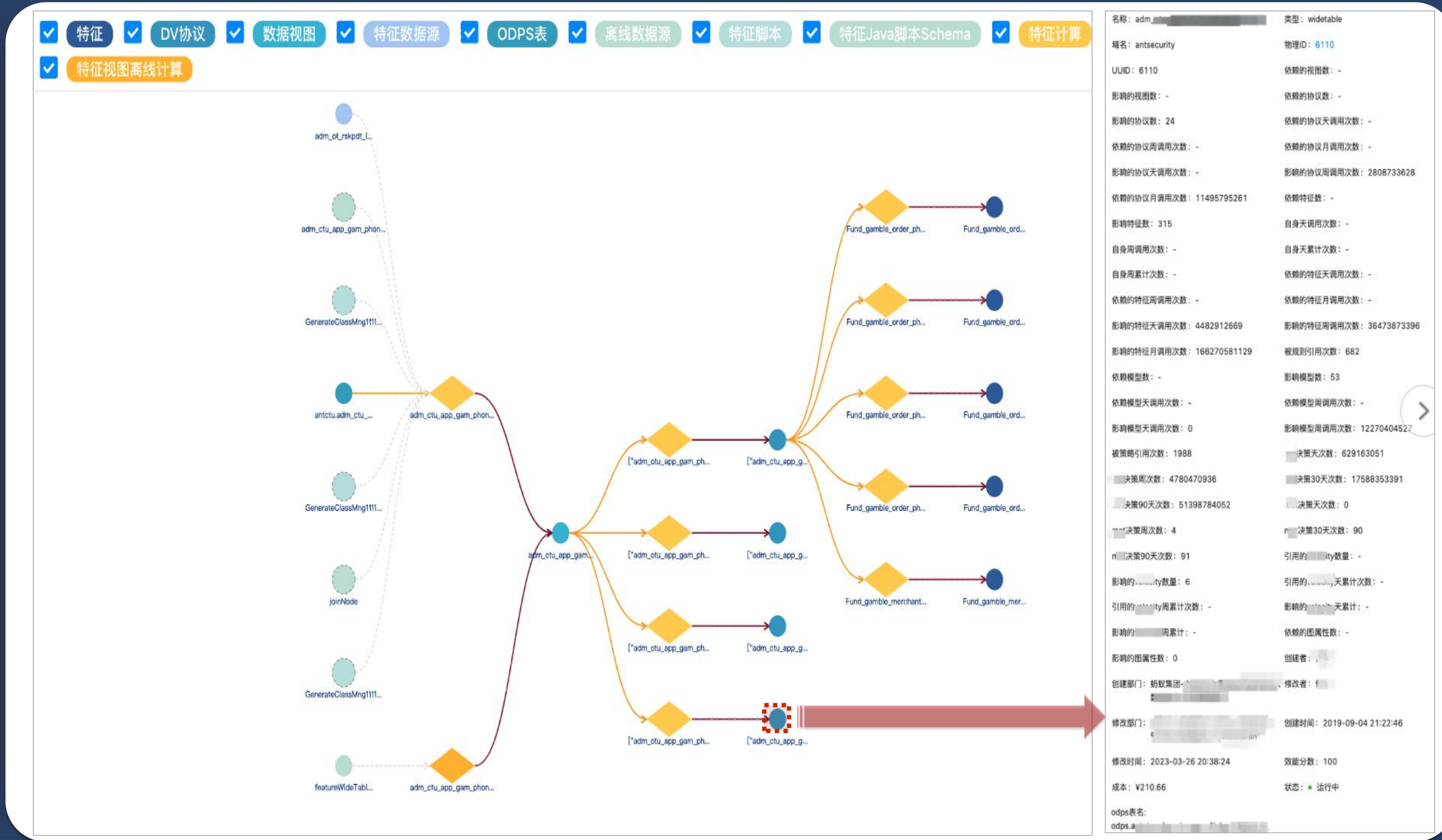
在线链路治理简介

良好的治理离不开对数据资产合理评估，通过对当前安全领域的数据资产进行了效能评估，产出了数据效能分，量化数据资产在风控系统中发挥的作用，从而推动无效资产的治理，计算、保障优先级设定，成本优化等。

安全领域智能化数据治理实践_在线链路治理

全链路血缘

数据资产类型包括策略、特征、模型、协议等90+种资产类型，种类多、数据资产量巨大、关系链路复杂，利用我们的二部图模型，构造了一张全局的资产大图，从连接起各个信息孤岛，打破平台间的血缘鸿沟。



安全领域智能化数据治理实践_在线链路治理

在线链路治理流程

影响因子定义

影响因子分：血缘静态引用量、线上流量调用量以及决策日志量三个层次。

影响因子分层：

- ① 根据数据血缘，计算出该数据资产与其它数据资产在静态血缘上的引用量；
- ② 计算该数据资产线上产生的实际流量，例如特征的调用量、模型的调用量；
- ③ 计算关联该数据资产的决策日志的量级，作为该资产在风控体系中发挥效能的重要特征。

备注：为了更合理的评估数据资产长短期的效能情况，分别从天/周/月/季的维度作为特征。

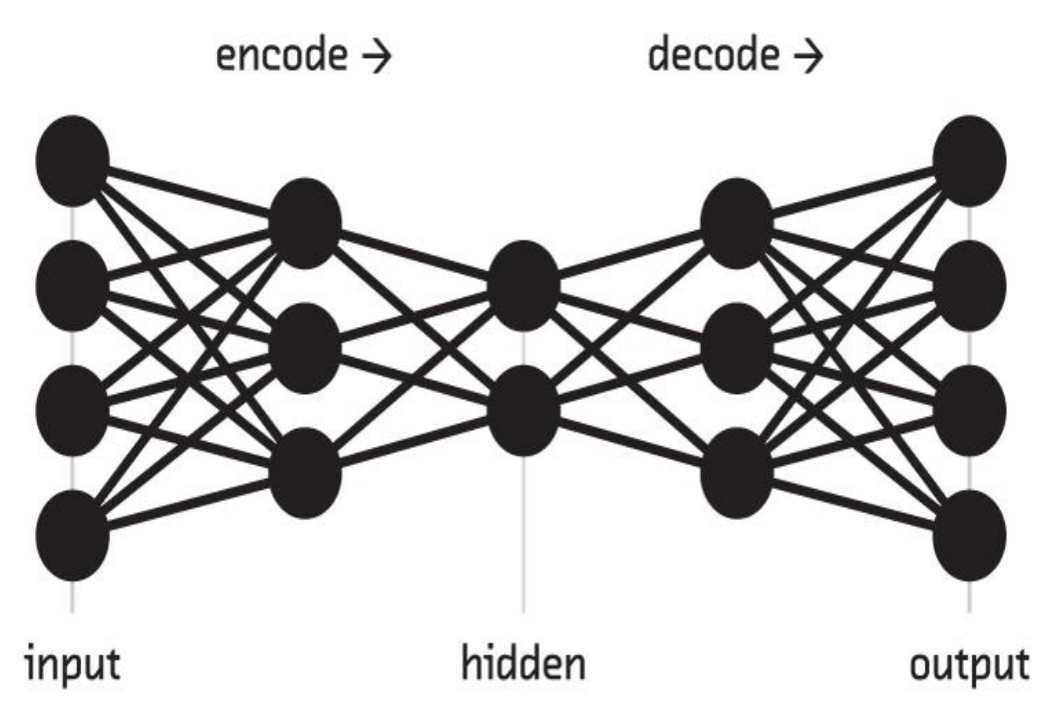
指标数据

基于上述影响因子特征，统计得到了从不同维度评估数据资产效能的源数据。

字段描述	被特征引用次数	被规则引用次数	被策略引用次数	特征快照周调用次数	决策日志周调用次数	
f_1	45	125	67	3,433	120	-
f_2	115	44	-	1,200	4	-
f_3	-	12	100	12	-	-
f_4	86	-	65	10	-	-
f_5	77	10	4	566	1	-
.....	-	-	-	-	-	-

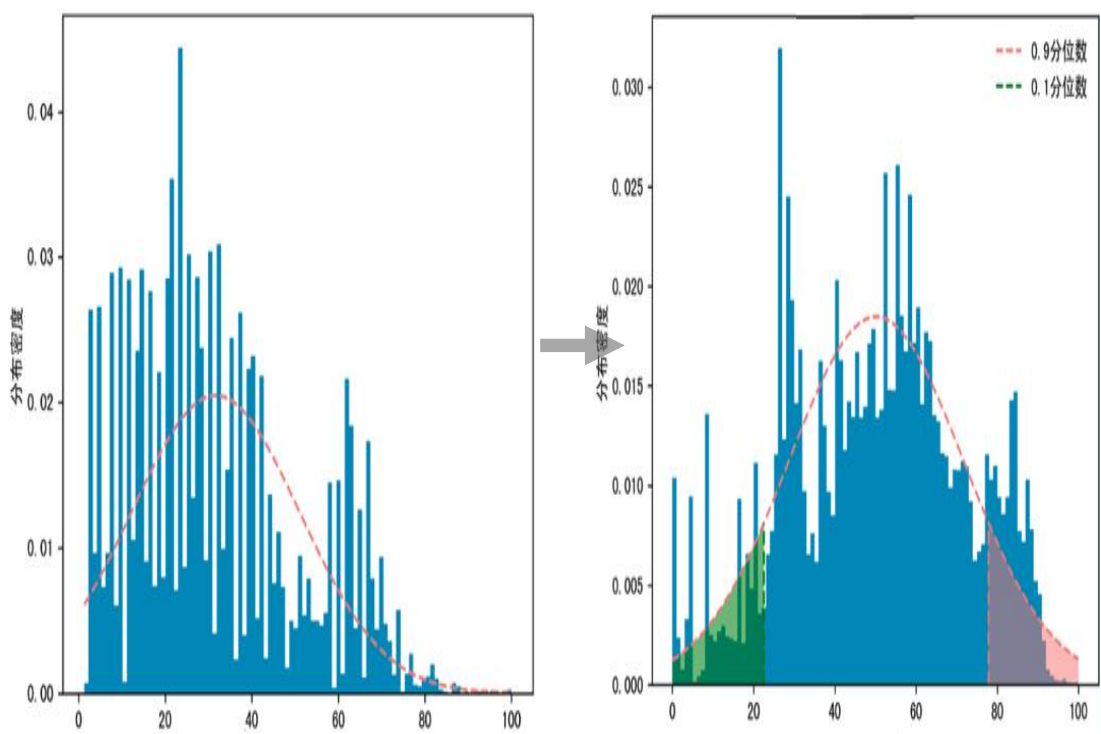
编码器推理

自编码器，对一组特征进行学习，得到有效表征。



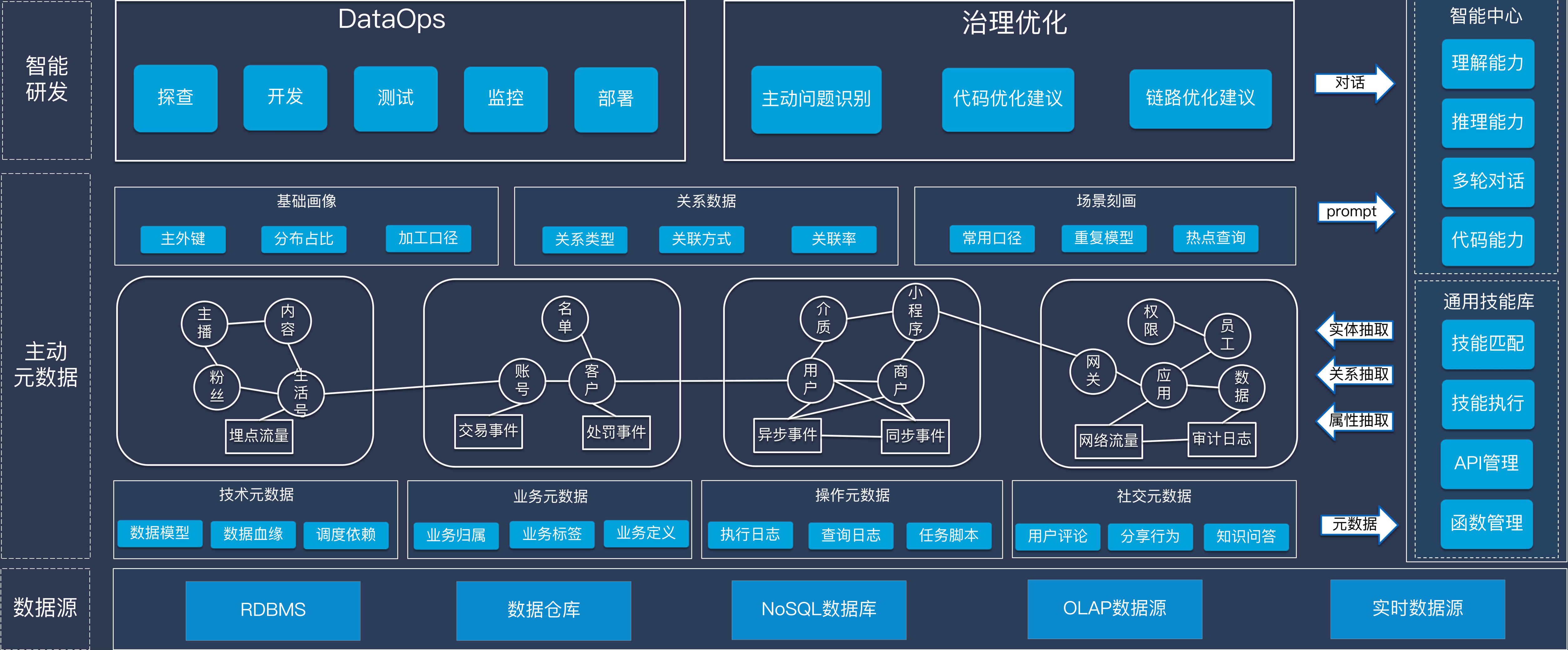
效能分变换

使用回归模型对特征进行回归训练，获各特征重要度；基于特征重要度优化模型和调权重，使效能分产出更合理；通过BOX-COX变换对数据分布进行调整（0~100分正态分布）。



数据治理探索与创新_ ETL AUTOPIPELINE

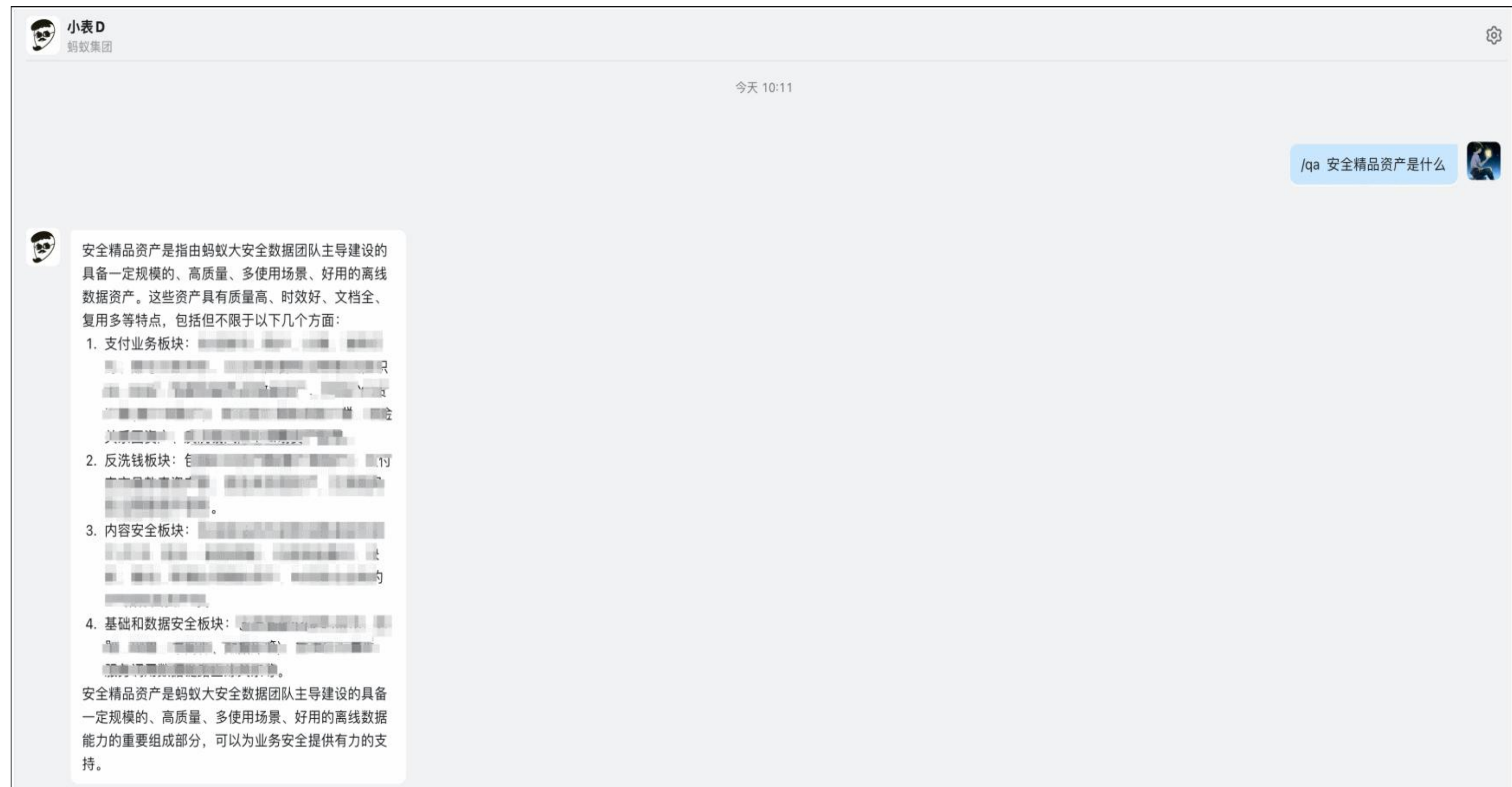
打造一站式智能研发助手



数据治理探索与创新_ 创新案例大模型Copilot

小表D

小表D：安全大数据一站式智能研发助手，结合安全特色，深度整合其他数据类大模型，以小表D为切入口，为用户提供丰富的大模型功能，贯穿用户整个数据研发生命周期，在数据分析、任务研发、任务运维、风险发现等日常生产环节提供一站式数据辅助服务，让数据研发更加智能高效。



想一想，我该如何把这些
技术应用在工作实践中？

THANKS