嘉宾｜ 李刘东 蚂蚁集团 数据技术专家

蚂蚁数科数据治理全景解析与实践

01

数据治理概念与框架

为了更好地理解数据治理的概念，可以将其与数据管理相对比。数据管理旨在确保数据能够被有效地创建、维护、使用和处置；而数据治理，其核心在于精细化规则设定与执行力保障，从而提高数据价值，降低数据风险。

因此，可以从三方面来定义数据治理：

- 首先，数据治理是一套管理行为的集合，通过一些管理手段确保数据的规范化处理和高效利用。
- 第二，数据治理是一系列政策与流程的集合，旨在保障数据质量、安全与合规性。
- 第三，数据治理的核心在于提升数据质量，发挥数据资产价值，同时降低数据风险，为组织提供准确的数据支持和决策依据。

数据治理的驱动因素包括：

- 法律遵从性：近年来，频繁的数据滥用与泄漏事件促使监管趋严，合规成为硬性要求，以避免潜在的巨大损失。
- 数据质量与安全：数据管理者肩负重任，需确保数据的准确性，保障信息安全，防范泄露风险，维系信任基石。
- 商业智能扩展：数据作为关键资源，驱动数字化转型与智能升级，赋能业务革新。

相应的，数据治理的目标为：

- 合规性与风险管理：确保数据操作合法合规，满足监管要求，有效规避风险。
- 高质量数据保留：保障数据安全的同时，提供便利的访问途径，支撑深度业务洞察。
- 减少数据孤岛：打破信息壁垒，提升数据共享与流动，充分释放数据价值。

02

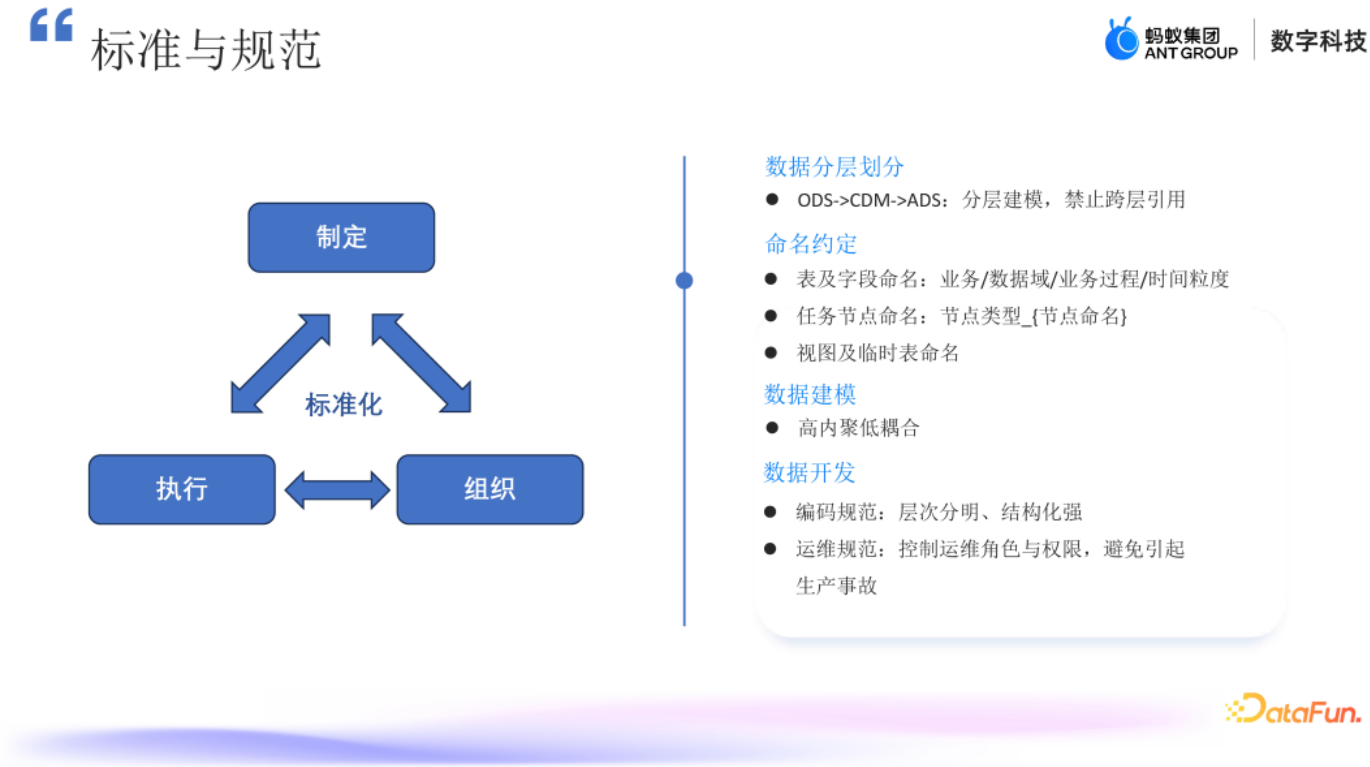
数据治理实践策略

数据治理不是单一的技术点，而是综合性的解决方案和策略的集合，需要跨领域融合，涵括数据架构、研发准则及平台工具等诸多方面，以构建全方位的立体支撑体系。

1. 数据架构

构建合理数据架构，乃数据治理与应用成功之基石，不仅关乎成本控制、效率提升与数据品质保障，更为数据开发指引方向，确保治理规范有序实施。统一的数据集成、建模与指标管理体系，是达成有效数据治理的前提，否则，将难以在治理流程中寻得平衡策略。

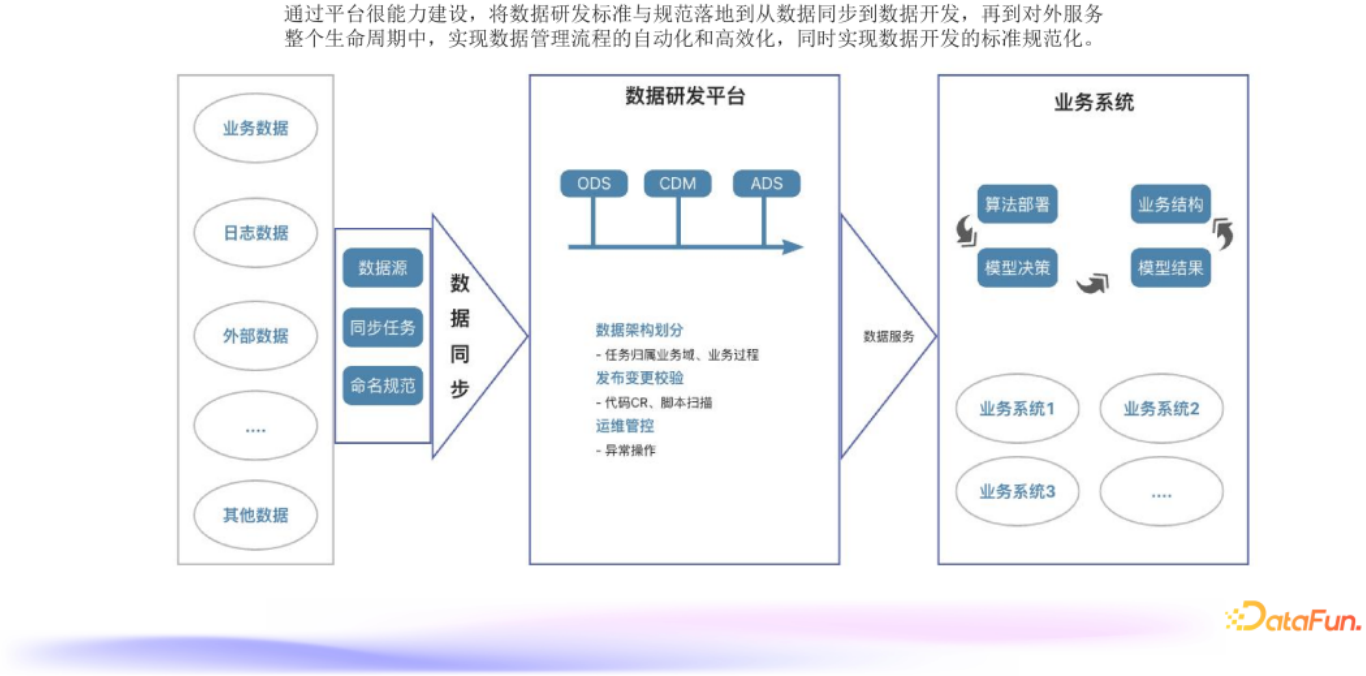
2. 标准与规范



标准与规范不仅对数据来说至关重要，对于应用开发侧也是不可或缺的。华为、阿里、谷歌等行业巨头在其 Java、Python 等后端开发领域均建立了详尽的编码规范文档，涵盖了函数命名、类命名、结构设计等各个方面。

在探讨数据领域的规范制定前，须思考数据开发者角色定位与价值。数据开发者不仅是组织的取数工具，而更是数据资产的建设者。数据资产的真正价值在于流通与利用。为促进数据流通，标准化与规范化必不可少。合理的分层、命名规则及开发规范，能够帮助下游用户准确理解数据内涵，明确用途，从而降低消费门槛，加速数据流转。

3. 平台与工具



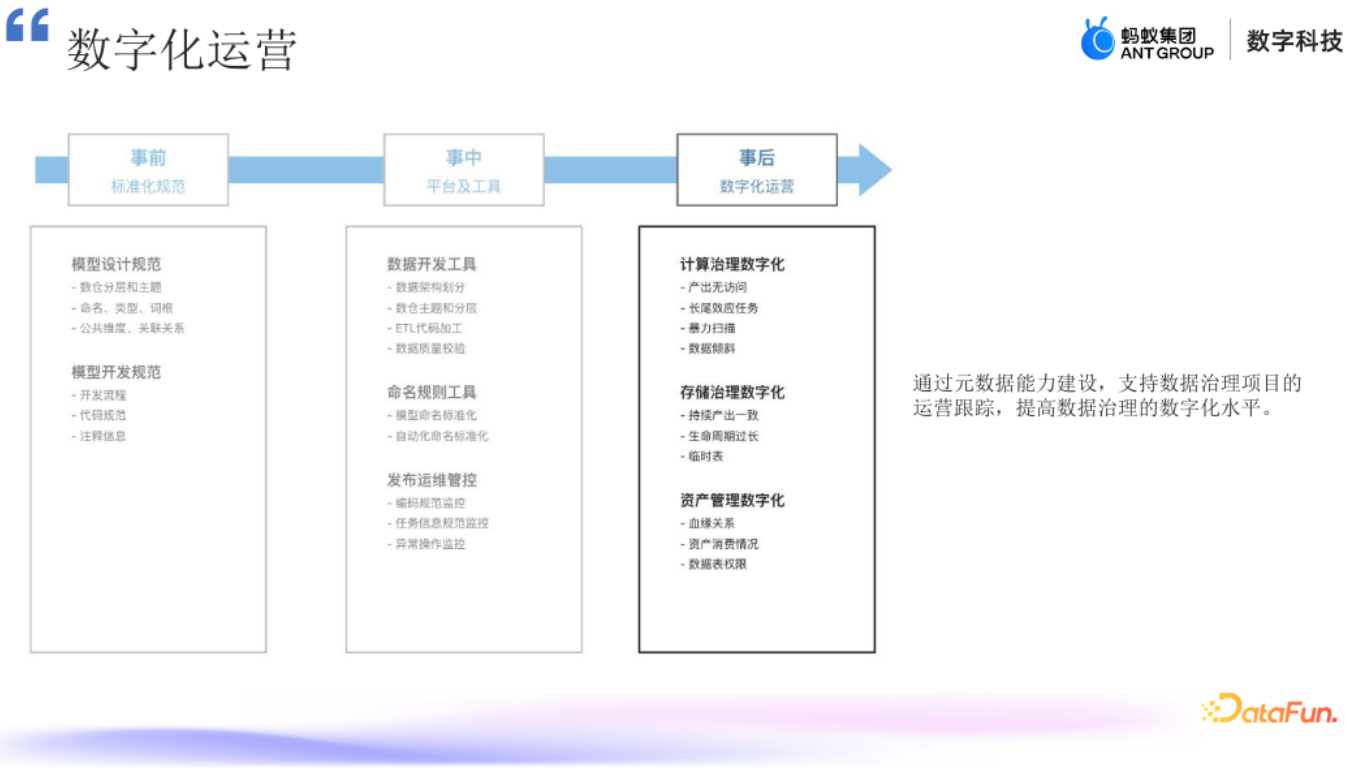
数据治理规范仅停留在文档层面难以贯彻，个体差异性使得统一指导难行。理想状态要求整合平台与工具，贯穿数据全周期，确保标准落实无碍。平台与工具的关键作用就显现出来，为增强标准执行力度，覆盖数据同步、研发至服务全程。

平台与工具的设计一定要注意避免管理立场过于强硬，忽视用户体验与效率。设计原则应兼容双重视角，兼顾管理者，注重研发者实际需求，提升操作便捷性，将规范融入日常作业，形成自然习惯，而非额外负担。

数据集成乃研发链条首环，尤为关键。涉及日志、外采数据与业务系统数据的收集，过程耗时费力，且效率低下。尤其是安全方面，账户密码、密钥管理不足，增添了后续治理的难度。因此需要构建统一集成平台，自动化处理数据来源，保障安全。

通过统一数据源管理，自动化串联业务系统与研发平台，实现数据无缝同步，缓解手动操作繁琐与安全隐患。此举不仅提升效率，更有利规范执行，如设定字段命名规则，区分不同类型任务（如全量/增量同步），令治理工作有序开展，减轻维护压力。

4. 数字化运营



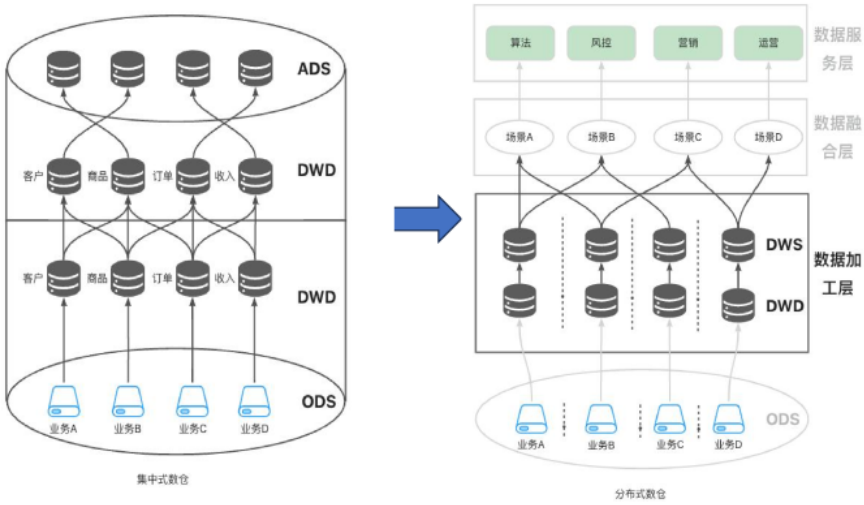
在制定好标准和规范，并有了统一平台后，数字化运营也是至关重要的。随着时间推移，企业数据资产不断积累，需要监控其价值与效用。通过元数据建设，对数据资产情况、治理过程和治理结果进行可视化跟踪，洞察各类数据的访问频次，评判长期保存必要性。从而实现数据治理透明化，追踪项目进展，强化数字化管理水平，确保数据资源合理配置与优化利用。

03

蚂蚁数科的数据治理革新实践

1. 架构转型：从集中式到分布式

分布式数据网格（数据架构）



集中式数仓->分布式数仓

明确数据资产所有权归属，提高数据安全性，组织跨部门数据资产，控制访问权限，满足合规要求。

传统上，众多企业在数据中台构建初期，倾向采用集中式数据处理模式，即各类数据与需求统一由单一数据中台团队管理。此种模式虽便于统一对外提供服务，但也暴露出诸多弊端。不同业务场景与安全级别的数据混杂一处，难以精准治理。加之，物理或逻辑隔离的缺失，使得跨业务线数据混用现象频发，无疑加大了统一管理的难度与潜在风险。

近年来，蚂蚁数科转向分布式数仓。其核心变革在于中间加工层，依据数据归属产品线、业务场景及类型，精细划分。例如，按 B 端/C 端、设备数据等类别，建立独立数仓，实现实体间清晰隔离，规避越权使用，强化数据安全与合规性。

具体地：

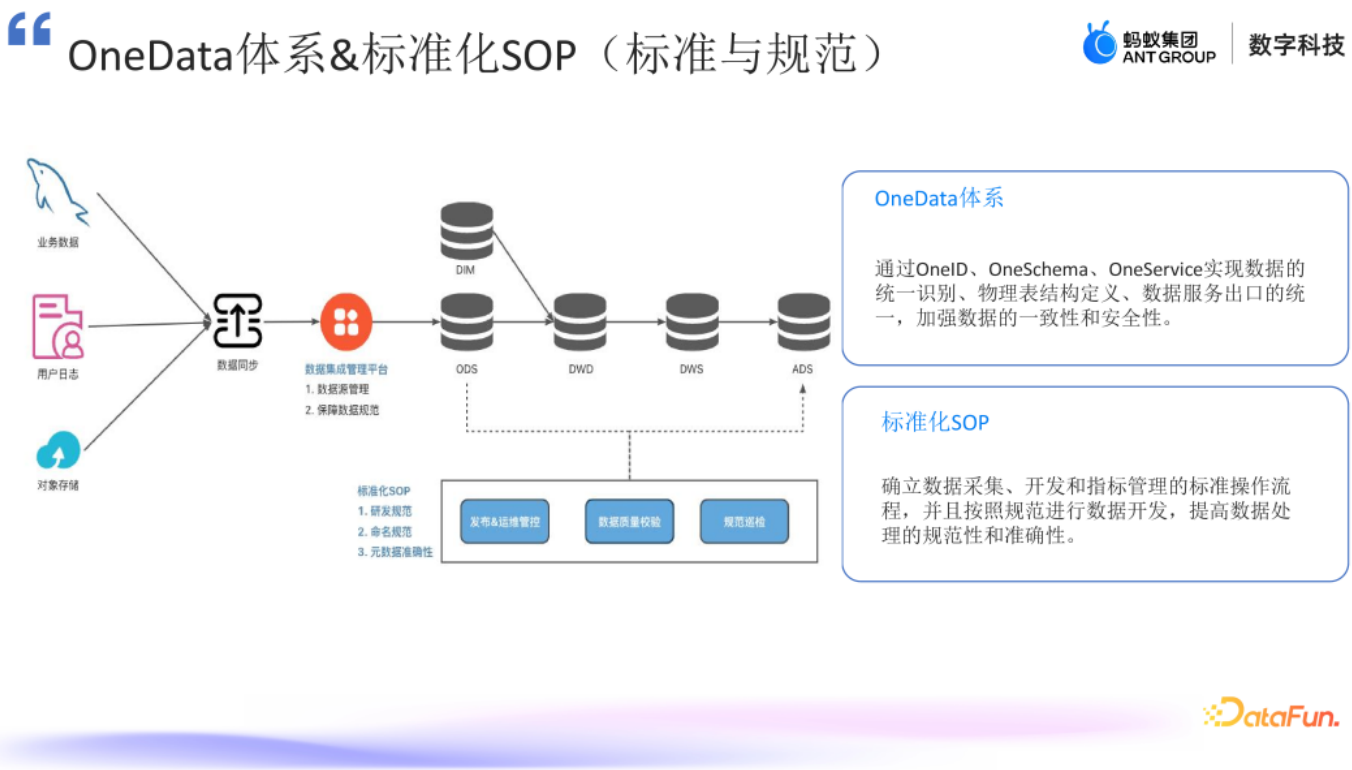
- 隔离机制：针对特定产品场景设立专属数仓，确保数据权限清晰，杜绝混用隐患。
- 权限控制：明确界定数据资产范围，限制非授权访问，强化安全保障。
- 安全合规：通过逻辑隔离，规避未经授权二次加工风险，符合法规要求。

实际案例中，考虑到业务间数据融合必要性，诸如 A/B 业务交互，原本集中式模式下权限模糊，安全隐患重重。现行分布式体系，则于更高层级设置融合层，引入法务、合规、隐私保护等部门参与审查，确保数据合规无虞。

分布式数据架构，从中心化向分散化转变。通过精细化管理，不仅提升了数据安全性与合法性，更实现了跨部门资产的可控流通。

此外，面对多样化的应用场景，分布式数仓可以提供更好的灵活性与可扩展性，满足现代企业的多元化需求，助推企业智慧化进程，确保业务健康发展。

2. 标准与规范：OneData 体系&标准化 SOP



OneData 体系聚焦于统一数据架构，包括 OneID、OneSchema、OneService 等部分。

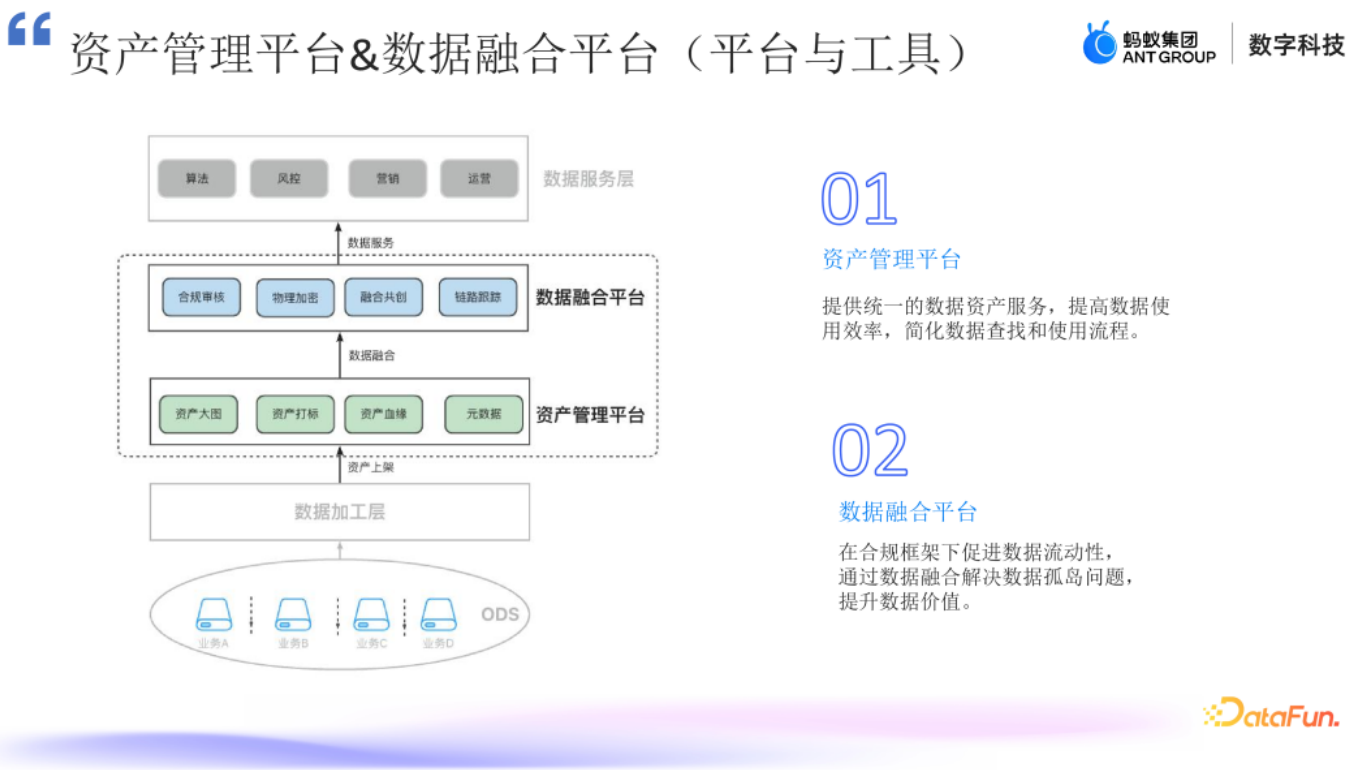
其中，OneSchema 主张标准化表结构、表与字段命名，确保信息一致性，并易于理解。设想构建企业搜索平台时，若各表对企业名称定义各异，将引发识别困扰；反之，统一命名规范如“corporateName”，就可以消除歧义，提升消费效率。

OneService 作为另一支柱，倡导统一数据服务出口。在数据资产建设分层（ODS 至 ADS）进程中，中间过程表与临时表泛滥，不仅造成消费者选择迷茫，亦加重开发者负担。通过 OneService，精炼服务层次，剔除冗余，保障数据一致性和安全性。

标准化 SOP 则致力于数据集成、开发、指标管理全流程规范化。从源头至目的地，确立采集、处理与管理规则，遵循既定规程，提升数据处理的规范性和准确性。

3. 平台与工具：资产管理平台&数据融合平台

贯彻 OneData 与标准化 SOP 原则，确保其实践可行，需依靠平台与工具的建设。为此，我们构建了资产管理平台与数据融合平台。

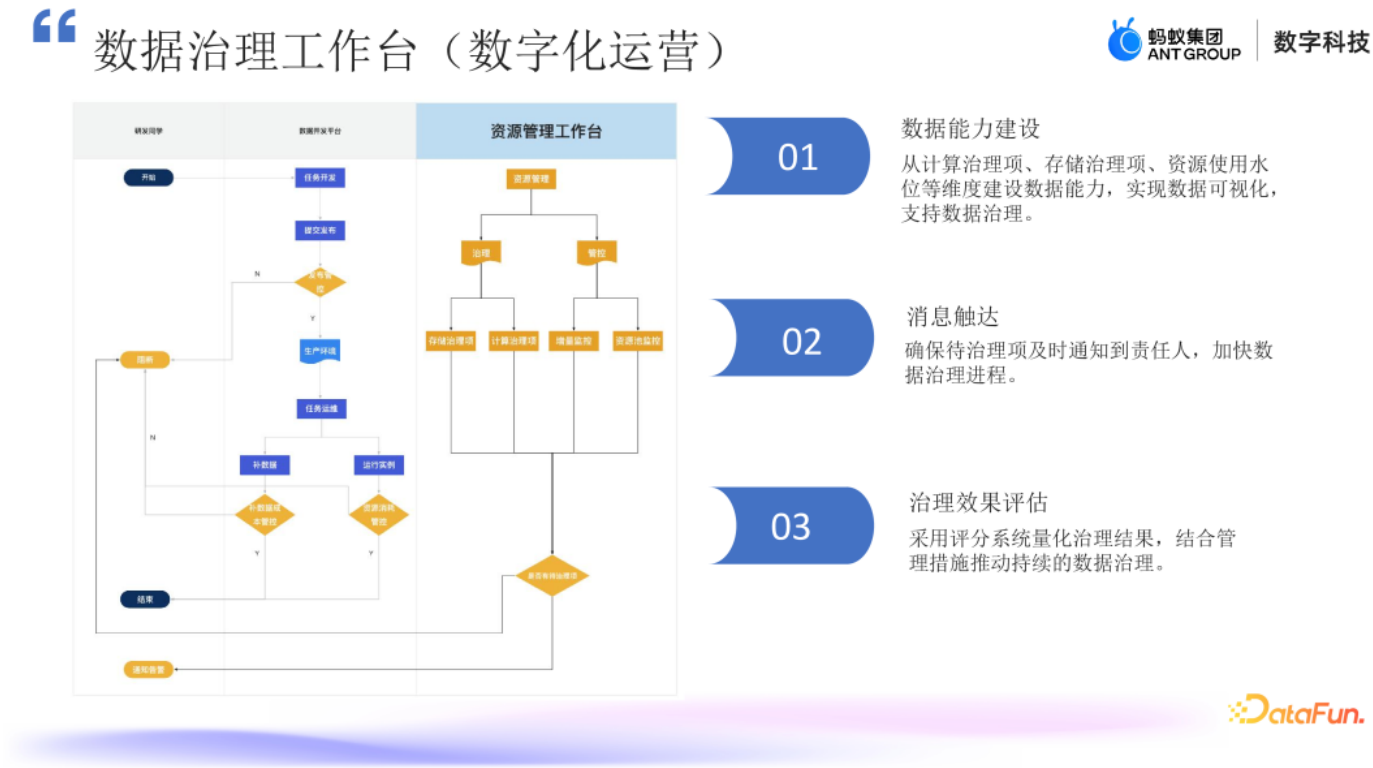


资产管理平台主要承担 OneService 职能，对外提供的数据服务均在资产管理平台中进行注册和管理，在平台中明确加工逻辑、口径变更，并划定资产提供侧与消费侧的边界，从而提升数据治理效率，简化数据查找和使用流程，节省用户成本，同时减轻开发团队维护压力，全面提升数据使用效率。

数据融合平台主要为应对合规挑战。特别是在金融数据领域，合规性尤为敏感。数据融合平台专攻合规问题，连结数据提供方、使用方与法务、合规专业人员，确保在政策框架下使用数据，实现公开透明的授权与决策。通过数据融合可以解决数据孤岛问题，推动数据价值释放。

两个平台分工明确，协同作用，共同织就数据治理的立体网络。资产管理平台保障数据有序运作，融合平台确保合规性，合力促进数据高效、合法流动，驱动业务增长与创新。

4. 数字化运营：数据治理工作台



当前我们的数据治理主要关注于计算与存储资源优化。经年累月，数据积聚呈指数级增长，依循“二八定律”，仅少部分数据能够创造价值，而多数陷于沉寂，却仍耗费高昂成本。鉴于此，我们着手资源治理，从多维度构建数据基础能力。

首要之举是绘制数据全景图，明晰资产现状与关联脉络，透视权限配置合理性。

其次，深化资源治理，聚焦计算与存储层面，构建中间层，量化分析使用频率与效率。具体而言：

- 计算治理：剖析数据倾斜、无效查询等开发疏漏，根治常态隐疾；
- 存储治理：审视表生命周期、临时表留存与回收策略，防范过度占用；
- 资源使用监控：监测水位，预防资源浪费。

上述措施，辅以智能化预警，及时通知责任方，促动治理行动。反馈治理成效，评估修正策略，迭代治理路径，最终形成闭合回路，持续优化数据生态系统。

以上就是蚂蚁数科在数据治理方面的主要工作。

04

数据治理的未来趋势

人工智能、大模型，成为科技界焦点。大模型依托海量数据集，经深度学习而成，然而伴随应用普及，争议声起，关于模型幻觉与误导信息，以及部分智能系统偶现违法背德回应，让舆论哗然，引人深思。

究其根源，在于模型本质乃算法产物，输出反映输入特性。异常回应源自模型初始阶段摄取不当数据，导致后续生成失真结果。譬如，社交平台负面言论或不当观点的吸纳，潜移默化影响模型判断力，催生争议性反馈。

面对挑战，亟待行业内外反思，如何嵌入数据治理于AI研发，以期技术伦理与社会责任并重。各大企业纷纷投入资源，探索解决方案。在此背景下，制定场景适配的治理策略显得至关重要。唯有严格筛选、净化数据来源，才能从根本上规避偏差，确保技术成果正面、负责任地服务于社会。未来之路，需不断优化数据质量，贯穿研发始终，保障 AI 发展正当且有益。

数据治理非孤立命题，兼顾效率与标准实施乃关键。着眼于此，探索先进技术，如大模型辅助，自动化字段注释、表命名等，可大幅提升数据流通透明度。实践中，开发者常忽视表注释，致使理解难度攀升，影响数据分析。对此，依靠大模型推理智能生成注释，可以解放人力，优化理解体验，同时缓解开发者负担，一举两得。

通过技术创新化解治理痛点，提升效率，无缝融入开发流程，实现质效双升。以此为鉴，未来数据治理不仅满足标准，更添灵活性与人性化，构筑健康数据生态。