

2021

阿里云 | DataFunSummit

阿里巴巴数据 治理实践

吴永明 阿里巴巴高级技术专家



目录

CONTENTS

01

数据治理概念和 需求层次

- 数据治理的理论参考
- 数据治理的概念和需求层次

02

企业数据治理痛点、 阿里巴巴数据治理 实践

- 企业数据治理的典型痛点
- 阿里巴巴数据治理的挑战
- 阿里巴巴数据治理的成功关键
- 阿里巴巴数据治理的发展实施阶段

01



数据治理概念和需求层次

- 数据治理的理论参考
- 数据治理的概念和需求层次

数据管理&数据治理 理论参考

数据管理协会知识体系 – DAMA-DMBOK2

十大职能领域

DCMM：数据管理能力成熟度评估（2018）

GB/T 36073—2018

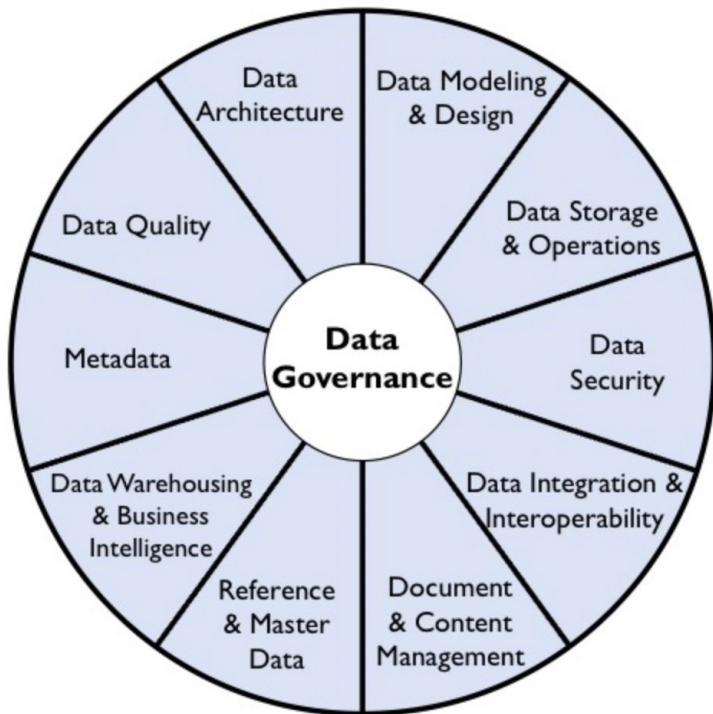
信通院：数据资产管理实践白皮书

理论和实践相结合的落地指南

DataFunSummit

数据治理的范畴

国际：DMBOK2 十大职能领域 [1]



Copyright© 2017 DAMA International

国内：DCMM 八大过程域 [2]

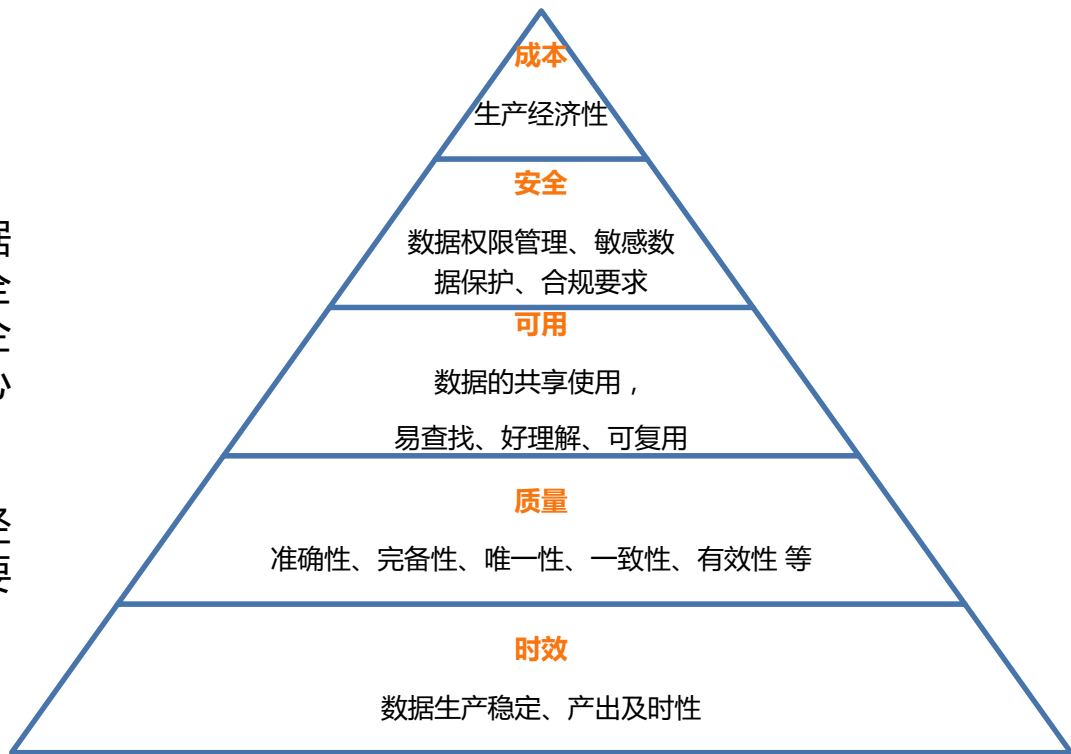


【1】：DAMA国际 <https://dama.org/content/dmbok-2-wheel-images>

【2】：DCMM：<http://www.dcm.org.cn/DCMMIndex/index.html>

数据治理的概念和需求层次

- 数据治理的内涵逐步泛化是业界共识
- 企业数据治理，涵盖数据发现可用、数据及时稳定产出、数据质量保障、数据安全合规和数据生产的经济性等**多个层次**。企业数字化转型阶段不同，治理关注的核心需求存在差异
- 在数据管理过程中，要保证一个组织已经将数据转换成有用信息，这项工作所需要的**流程和工具**就是数据治理的工作^[1]



【1】来源：DAMS，作者 Jelani Harper。有用的信息：数据资产

02



企业数据治理痛点、 阿里巴巴数据治理实践

- 企业数据治理的典型痛点
- 阿里巴巴数据治理的挑战
- 阿里巴巴数据治理的成功关键
- 阿里巴巴数据治理的发展实施阶段

企业数据治理典型痛点

数据治理成效进展缓慢，数据问题依旧严重，缺少系统化的工具平台支撑治理落地和成效展现是关键原因之一

数据治理咨询成果落地不足

数据治理产出成果，比如各类规范和管理办法，包括数据字典，多以“纸面文件”的形式流转与企业中，与实际业务和数据没有紧耦合，能满足“我有”，但是没能做到“我执行”

数据治理成效可視度低

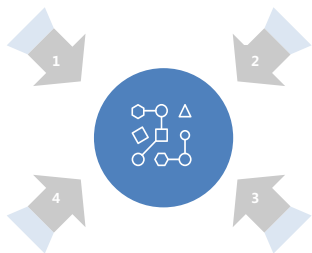
缺少量化方式来评估数据治理成熟度水平，数据治理工作的推动成效无法体现，变成了纯手动的脏活累活，严重影响数据治理工作的开展推进

自动化服务程度不高

业务人员使用数据更多需要数据和技术人员的贴身服务，按照IT建设的模式提出数据加工需求或者取数需求，以被动支持的方式满足业务需求，没有形成数据资产目录，数据服务目录

数据治理在线管理能力不足

缺少灵活友好的数据治理在线管理工具，来支持数据治理全流程工作
数据治理与数据原仓之间没有打通，“数据的描述”和“数据的记录”两张皮



企业数据治理新模式

从传统架构思维向DT架构思维转变，围绕**数据资产化**、**数据价值释放**的核心目标开展工作

变思维：转变传统思维定式，IT思维向DT思维转型

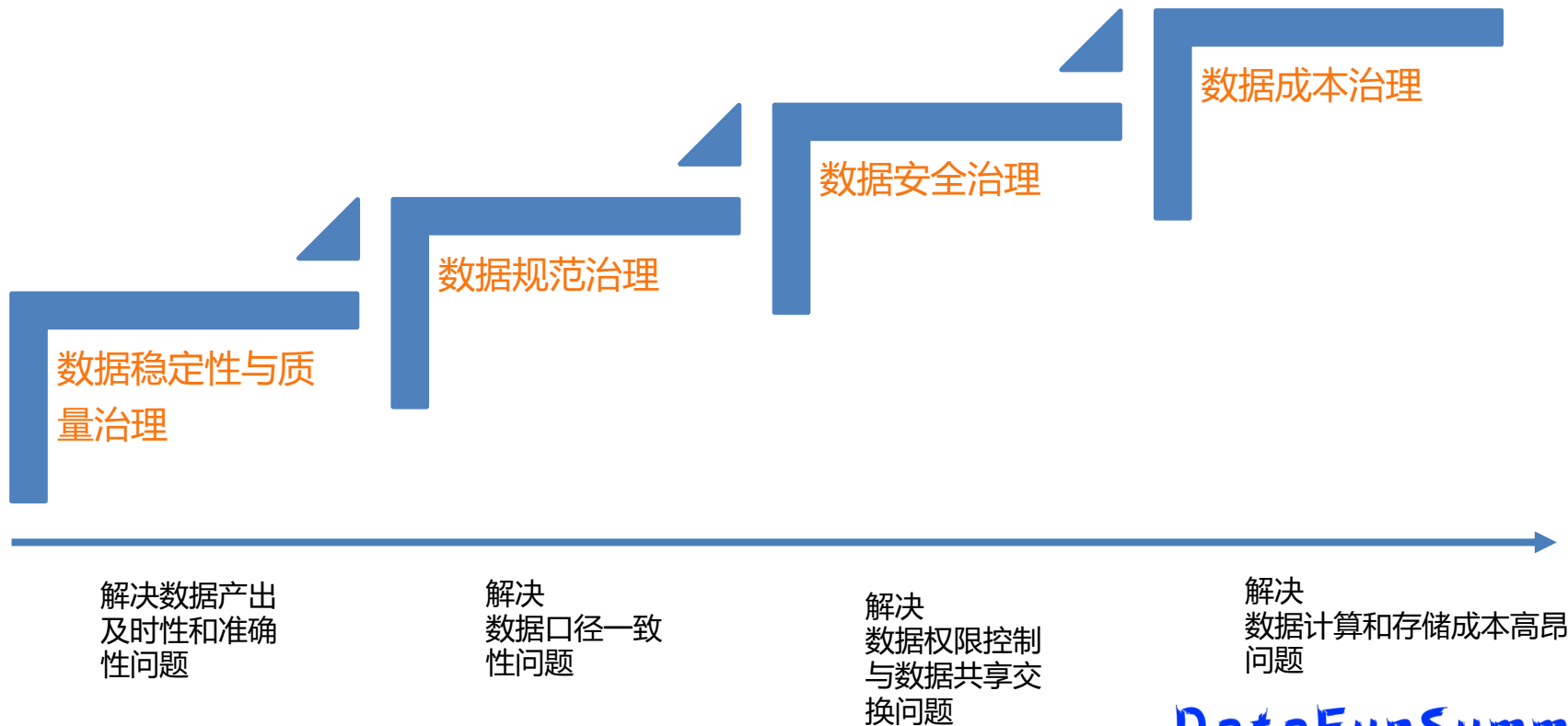
变模式：工具和技术是生产工具，数据才是核心，IT流程不是核心

变定位：摆脱成本中心泥潭，通过运营数据资产，探索如何成为利润中心



阿里巴巴数据治理发展实施阶段

特定阶段专注解决主要矛盾



阿里巴巴数据治理实践：数据稳定性

千万级任务的调度情况下，调度依赖关系复杂程度远超过人工处理程度，独有智能基线监控机制确保高优先任务高保障产出

监控告警的痛点

监控数量

监控所有任务是不现实的

配置难度

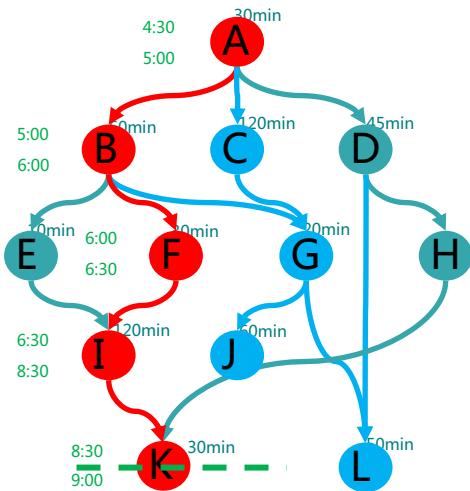
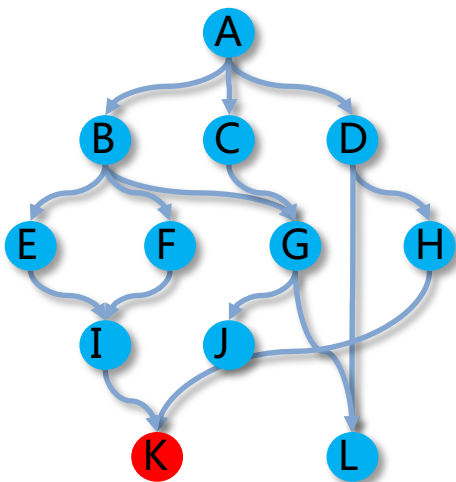
为每个任务配置监控规则极为繁琐

告警时间

每个任务所需告警的时间都不同

智能监控核心功能

- 智能识别关键路径，合理设定告警阈值
- 任务异常产生事件，自动评估事件影响范围，通知相应人员
- 灵活告警方式配置，支持钉钉群机器人、电话

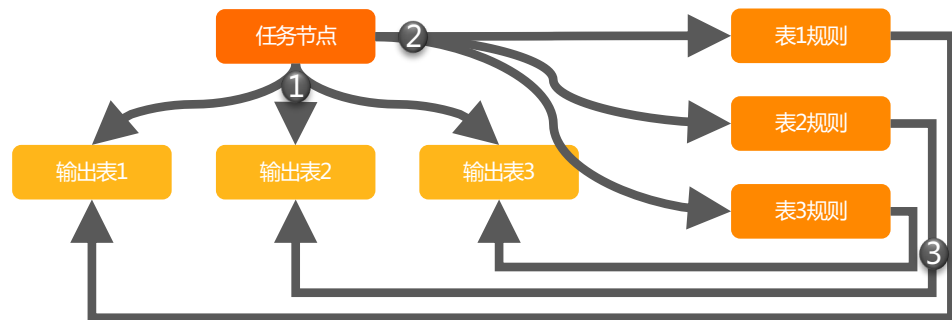


DataWorks独创、荣获国家专利的智能基线监控技术

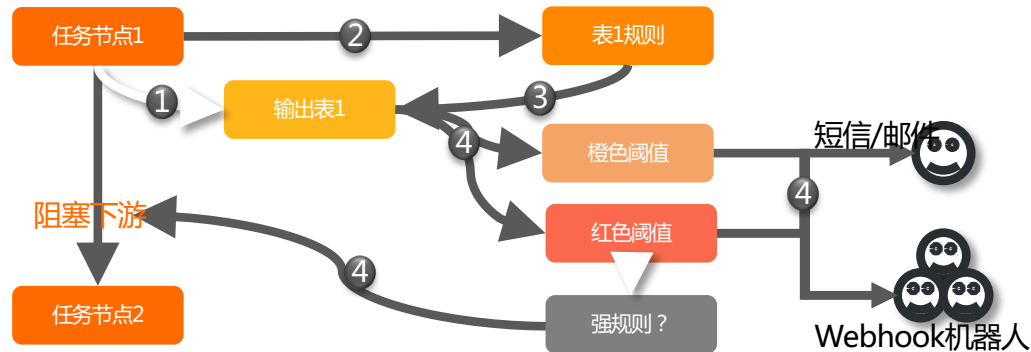
DataFunSummit

阿里巴巴数据治理实践：数据质量治理

通过完整性、有效性、准确性、唯一性、一致性、合理性的全面评估，产出可信的、高价值密度的数据资产



① 执行任务 ➡ ② 触发规则 ➡ ③ 执行规则 ➡ ④ 告警/阻塞

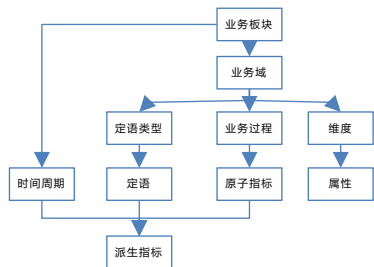


- 质量监控与调度挂钩，第一时间发现问题
- 40+规则&自定义规则，精细化质量控制
- 无需设定阈值，算法自动判断异常值
- 故障快速恢复

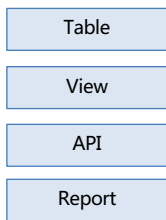
阿里巴巴数据治理实践：数据规范治理

通过规范设计和开发来预防问题的发生。统一公共层来减少重复建设和确保口径一致性

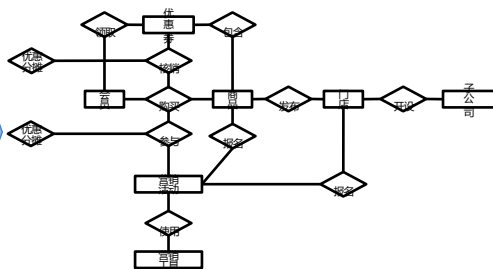
数据规范设计



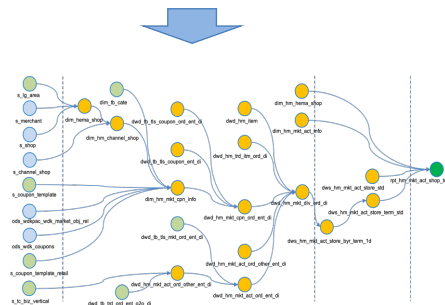
指标体系设计



数据服务开放



数据模型设计



数据处理任务开发

数据公共层建设

强管控

一条门槛线

- 1) 确定标准、流程及规范
- 2) 筛选核心公共层监控范围并持续更新

轻约束

核心公共层

其他

- 核心公共层数据资产：
- 1) 做规范管控，架构评审，发布管控
 - 2) 评估建设水平
 - 3) 发现短板，持续改进

阿里巴巴数据治理实践：数据安全治理

数据分类分级与权限控制

敏感数据发现与脱敏

可信计算环境

数据风险审计

制定分类分级规范

数据自动打标

打标人工调整

更合理管理和使用

判断依据

字段名

字段描述

字段值

匹配规则

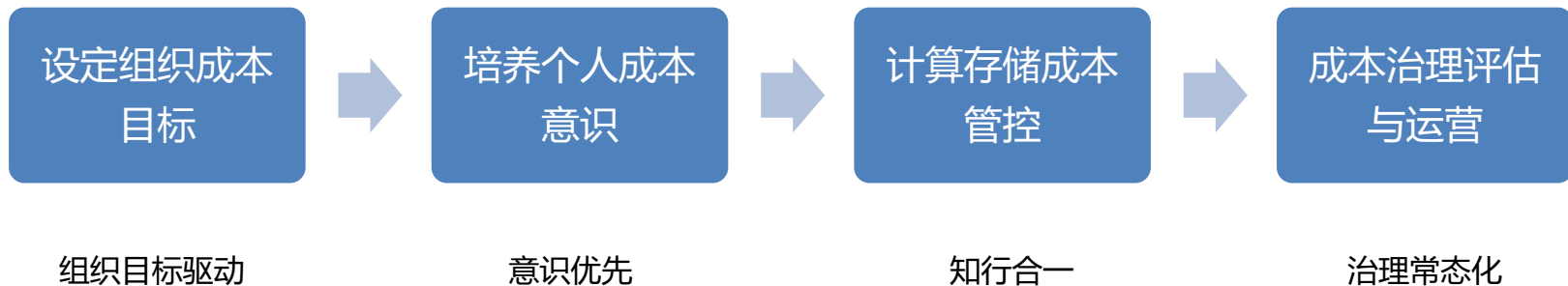
关键字匹配

正则表达

算法模型

(阿里：根据分级差异化审批流)

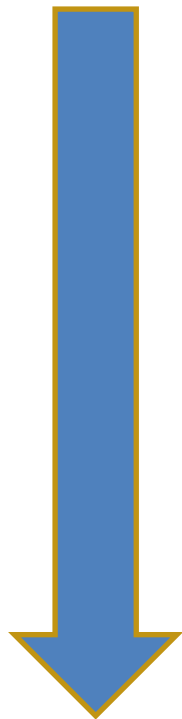
阿里巴巴数据治理实践：数据成本治理



阿里巴巴2020年成本治理的目标：数据成本增速不能超过业务增速

阿里巴巴数据治理成功关键

自上而下

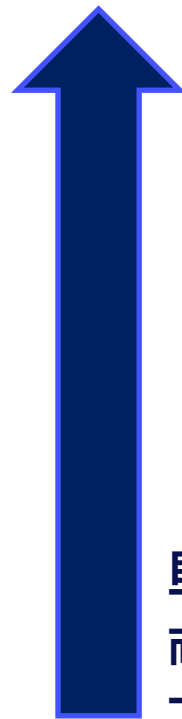


一套组织体系
组织建设、制度保障

一部数据资产治理方法论

产出及时、质量可靠、易找易用、安全可控、生产经济

一组平台工具支撑&运营
阿里云大数据平台/数据中台



自下而上

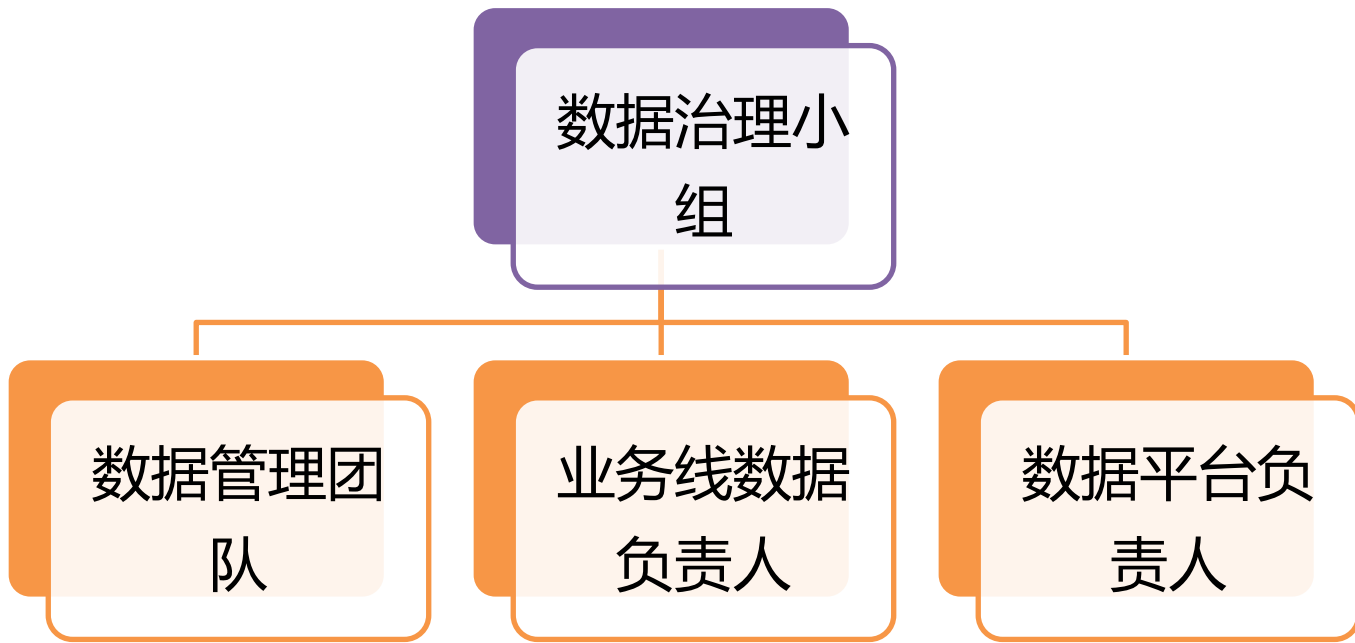
核心目标：
数据资产化、数据价值释放

自上而下：
从公司治理角度入手来解决数据的管理问题，提供足够的授权和支持

自下而上：
以平台技术支撑和完善的运营体系促进治理的切实落地

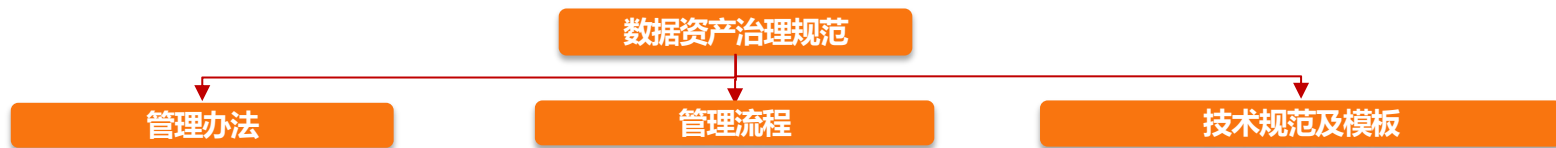
阿里巴巴数据治理成功关键 – 一套组织体系

固定的专业组织、充分赋权，负责数据治理实施的整体推进。制规范 定目标 促落地 保健康



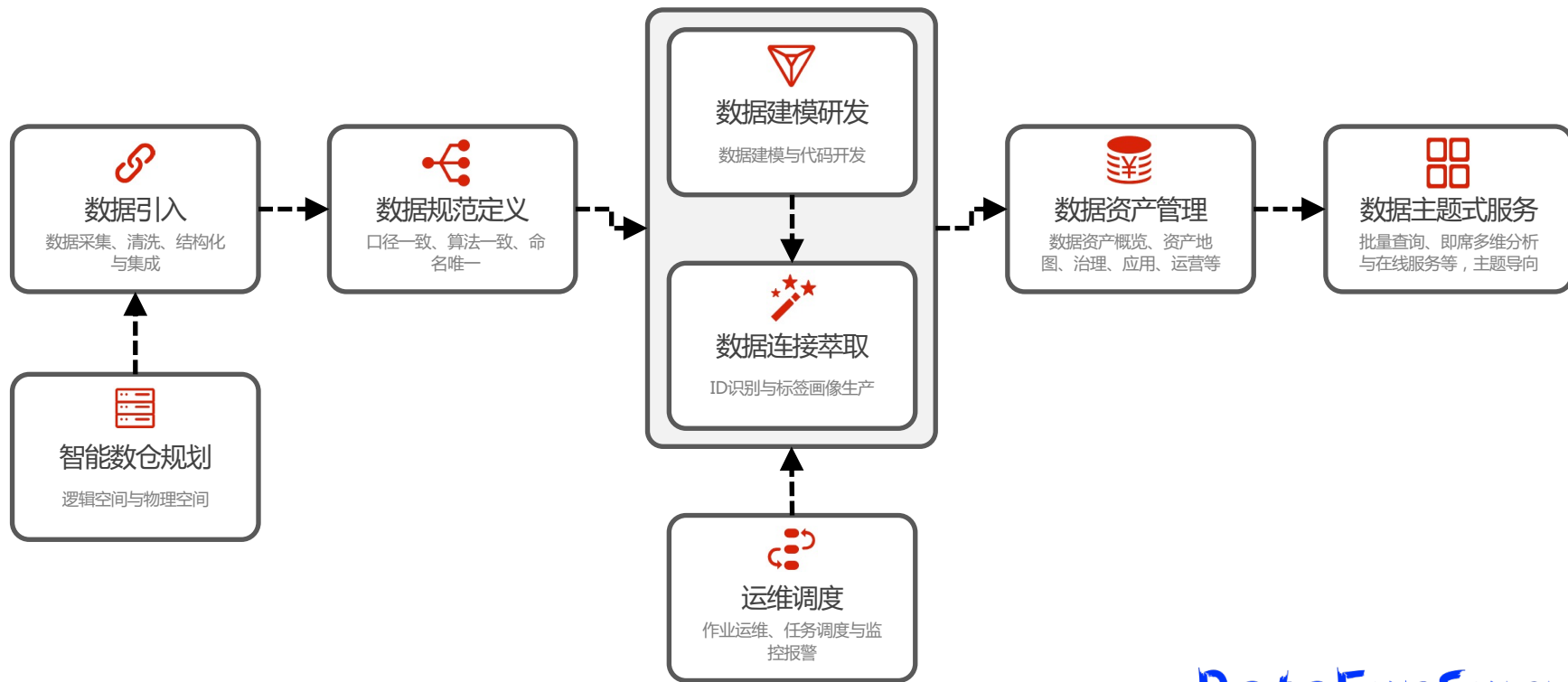
阿里巴巴数据治理成功关键 – 制度保障

从实践中总结制定一系列的管理办法、流程和规范，并及时演进迭代



阿里巴巴数据治理成功关键 – 一部数据资产治理方法论

数据运营思想贯穿数据建设全过程



阿里巴巴数据治理成功关键 - 一组平台工具支撑

DataWorks

一站式 大数据开发和治理平台

阿里自研的大数据平台，各类存储和计算引擎的上层操作系统，提供数据集成、数据开发、数据地图、数据质量、数据安全和数据服务等全方位的产品服务，帮助企业专注于数据价值的挖掘和探索。



MaxCompute

自研、全托管、EB级 大数据存储和计算引擎

阿里自研的安全可靠、高效能、低成本、从GB到EB级别按需弹性伸缩的在线大数据计算服务，致力于海量结构化、半结构化数据的存储和计算服务，提供数据仓库的解决方案及分析建模服务。



强大的平台能力支撑是治理落地的核心保障；技术的创新和演进是数据治理落地的坚实基础

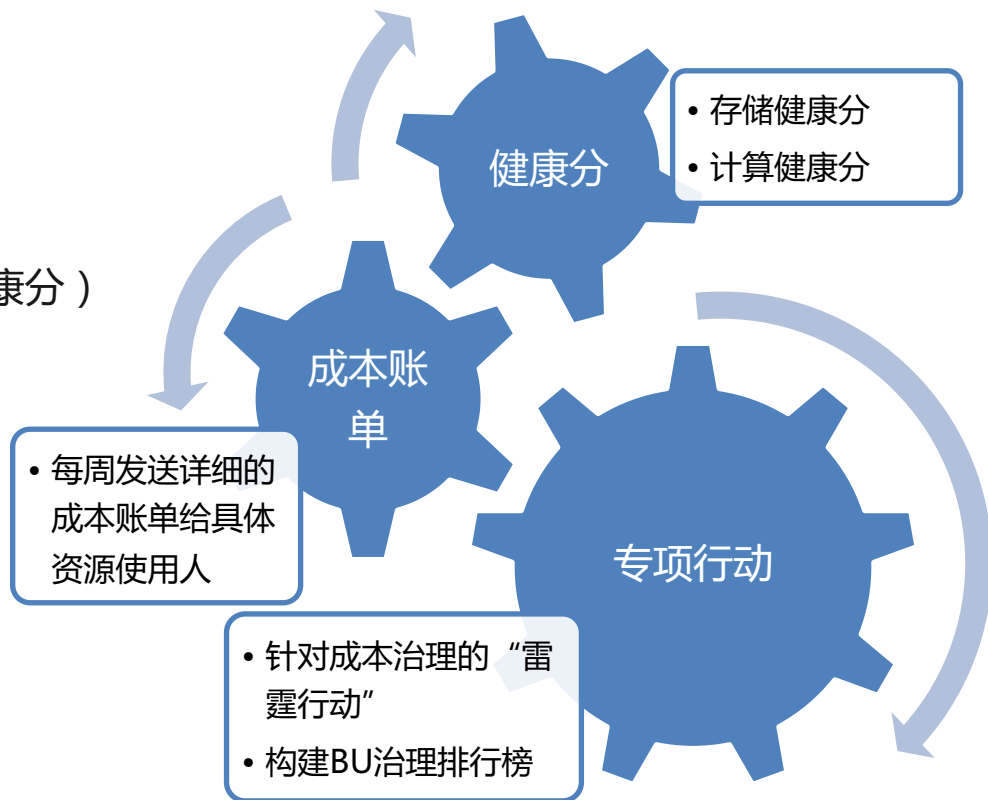
阿里巴巴数据治理成功关键 – 运营落地

构建量化的数据治理评价体系，日常治理运营和专项整治相结合，促进治理工作持续落地改进

治理运营是推动数据治理落地的关键因素

阿里巴巴构建了量化的治理的评价体系（健康分）

日常治理运营推送和专项整治活动密切结合



DataFunSummit

2021

阿里云 | DataFunSummit

THANKS!



Ending

