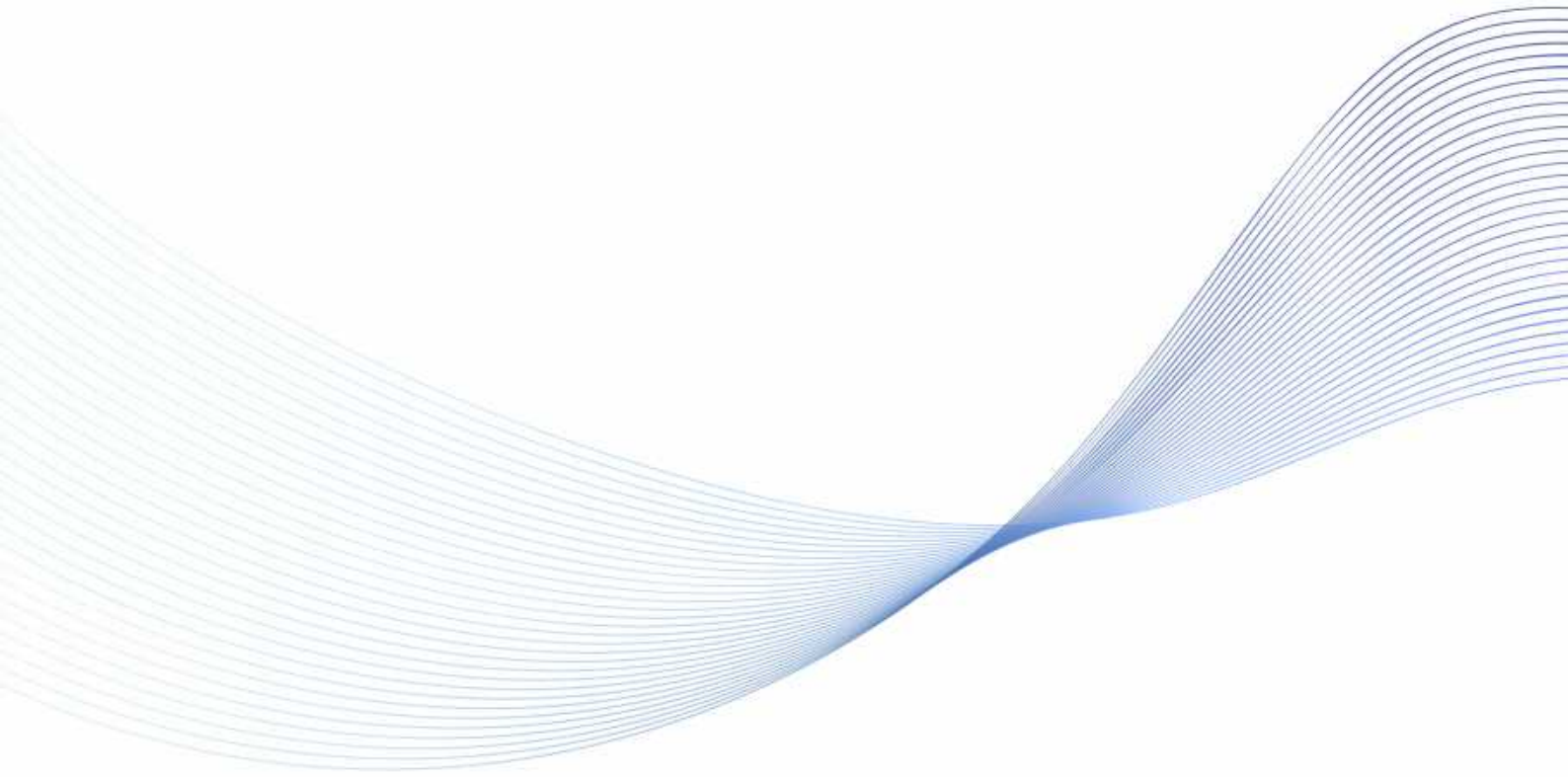
网易数帆 | CSDN

网易数据治理白皮书

开　发　治　理　一　体　化



网易数帆大数据团队出品

目录

第一章 数据资产

1.1 数据分类

对于企业来说，数据的产出、应用和管理无处不在。在数字化的大浪潮下，无论是企业的管理者，还是企业的基层员工无时无刻不在和数据打交道，如何应用好数据、更好的挖掘数据价值是每个企业都面临的问题。企业使用数据的前提是了解数据，我们将常见的企业数据分为三大类：主数据、业务数据以及分析数据。如果将企业比作大树的话，那么主数据是树上的树干，业务数据是树干上的枝叶，分析数据则是长在枝头的果实。

主数据作为树干用来承载业务数据和分析数据，它是有关业务实体（如雇员、客户、产品、地点等）的数据，为业务交易和分析提供了语境信息，因此离了主数据的业务数据和分析数据都是没有意义的，就像叶子和果实离了树干无法独自生长。业务数据用来描述企业的经营活动，如产品出入库、财务应收、采购、销售等活动产生的数据。业务数据是分析数据的基础，分析数据来自业务数据的加工，就像叶子通过光合作用为果实制造营养物质，没有叶子就没有果实的存在。分析数据用来反映企业的业绩，被用于企业的决策分析，在一个企业中，数据的真正价值在分析数据时得到体现，高质量的分析数据就像成熟饱满的果实，能为企业带来巨大价值。因此，主数据是业务数据和分析数据的基础，业务数据为分析数据的产生提供了环境，分析数据是企业数据的价值体现。

1.2 数据资产定义

对于一个企业来说，并不是所有的数据都值得去管理、去维护甚至去分析的。数据的生成、汇聚、存储、分析、共享等阶段都会因为数据管理的不当、数据治理手段的缺失，从而产生低质量的数据。低质量的数据不仅没有价值，它的存在还会导致错误的决策，如同不健康的树干会影响叶子的生长，不健康的叶子无法提供足够的营养进行果实的孕育，不健康的果实无法食用一样。

因此，在企业当中数据和资产是不等价的，中国信息通信研究院联合多家企业于2021年12月发布了《数据资产管理实践白皮书5.0》，其中将数据资产定义为“由组织(政府机构、企事业单位等)合法拥有或控制的数据资源，以电子或其他方式记录，例如文本、图像、语音、视频、网页、数据库、传感信号等结构化或非结构化数据，可进行计量或交易，能直接或间接带来经济效益和社会效益。”

上述定义分别从数据主体、数据资源以及数据价值三方面对数据资产进行了描述。数据主体表明了数据是有主体的，可以来自政府机构、企事业单位等组织。数据资源表明了数据的存在形态，以电子或其他方式记录的结构化或非结构化数据。数据价值则反映了在组织中，数据可直接或间接带来经济效益和社会效益，是一种数据资源。

1.3 数据资产与数据治理

根据对数据资产的定义，我们了解到数据资产体现数据的价值和数据的应用。通过对数据资产盘点及价值分析，找出有价值的数据并展示其价值和应用，也就是说并非所有的数据都是资产，只有对企业有价值的数据才被认作是数据资产。因此，这里就存在一个关键性的问题，如何将企业数据变为资产，从而进一步实现价值变现就显得尤为关键，而数据治理是解决这个核心问题的钥匙。

数据治理在整个数据体系中主要解决的是人与人、人与数据之间的事，在整个治理过程中体现了数据的管理以及数据价值的呈现。如果说数据是企业信息化的原料，那么数据治理便是企业信息化的基石，数据资产则基于数据治理的数据，挖掘数据的价值，通过数据运营、数据分析的手段，为企业赋能，助力企业的信息化建设，完成数字化转型。

第二章 数据治理解决了什么问题?

网易作为一家互联网公司，很早就在生产活动中应用数据的分析结果，助力业务的增长。随着业务规模的扩大，如网易云音乐、网易有道、网易新闻、网易严选等多个业务线的孵化，同时也诞生了大量的集群，内部对于集群统一管理的呼声也日益变高。2018年以前，网易还没有将数仓的建设提升到组织架构的层面去规划，导致各个业务部门的不同团队都有一些零散的数据开发和分析人员承载本团队内的数据分析需求，这样的一个组织架构导致的结果就是很多零散分裂的小数仓存在，烟囱式的开发对业务带来了严重的影响。到了2018年，因业务规模的快速扩大，数据量的急速增长，相应的数据问题终于爆发，例如数据使用率低、数据经常违反常识、数据成本指数增长导致投入产出比低、数据安全风险日益突出等等，数据治理迫在眉睫。我们将上述问题进行了归类，总结出了四个数据使用过程中的问题，分别是找不到、看不懂、信不过、管不住。

找不到

除了数据量的不断增大，数据的发现效率成为使用数据的门槛之一。在网易内部，严选的业务线约有8万张表，音乐的业务线约有4万张表，对于数据分析而言，越靠近应用层，越会存在很多大的宽表，一个表有上百个字段是一个非常正常的事情。对于数据使用者而言，从几万张表中找到自己需要的数据，犹如大海捞针，谁也不清楚系统中到底有哪些数据，也不知道如何去快速准确的找到这个数据。对数据地图的用户进行分析，发现居然有90%以上是IT人员，而原本作为产品目标用户的业务人员却几乎无人使用。

看不懂

即便业务人员找到数据，我们发现，他也很难看懂数据。据统计，高达78%的表都存在元数据缺失，尤其是管理元数据和业务元数，而业务元数据和管理元数据，是业务人员了解数据业务含义最重要的信息。通常来讲，技术元数据的完整度一般都是最高的，可以通过系统化的采集获得；而管理元数据和业务元数据，与业务相关性较高，是需要业务配合来补充完善的，因此相对的，其缺失度更高。

信不过

质量是数据的生命线，没有质量保障的数据，不仅没有价值，还会产生错误的决策。我们在严选就曾经出现过，因为开发修改了一个上游任务的数据计算逻辑，影响了下游一张涉及资损的表的数据正确产出，结果导致红包超发，产生了几十万的资损。这些血淋淋的教训反复告诉我们，保障质量对于业务团队对数据的信任有多重要。

数据违反常识是数据质量问题的一种表现，开发人员往往不理解数据背后的含义而无法从开发结果上判断数据是否满足业务方要求，导致数据质量的问题最后都在业务方使用过程中暴露出来，久而久之业务方对于开发团队的数据不再信任。我们曾对popo群（网易内部工作通讯工具）里面每日反馈的问题进行统计，平均下来，每周就有10个数据质量问题被反馈，“数据违反常识”是当时我们听到的最多的一句业务部门的吐槽。更为严重的是，这里面90%的问题，都是数据使用方先于数据开发方发现的，对于我们数据团队来说非常的被动，往往出现问题我们自己都不知道。

管不住

企业业务的高速发展，导致业务上的数据量不断增加，相应的数据的成本也呈指数级增长。而在企业众多成本当中，数据的成本往往是最容易被忽略的。事实上，数据的成本不仅仅是钱的问题，还是资源没有最大化使用的问题。我们曾对内部某事业部的数据进行分析，发现78.39%的表占据了21.63%的存储空间，这些数据都是无人访问的，造成了大量的计算资源和开发资源的浪费。

另外，资源的滥用还会影响集群的稳定性，据内部的记录，每个月都有5次事故跟资源滥用有关，例如一个5层嵌套的大SQL瞬间把集群打挂等等。公共资源的管理不当、缺少科学的数据资产沉淀方式，最终导致存在大量无人问津的数据，数据成本居高不下。

随着数据量的不断增大，除了资源管理上的问题，企业还面临着因为使用数据的人越来越多而导致的数据安全问题。企业既想要员工多用数据，又要确保不同密级的数据被对的人所使用。不合理的数据权限及功能权限的分配会导致数据误删、数据泄密的问题经常发生，对于企业来说是无法接受的。此外，复杂的权限设置以及频繁的授权申请都会对数据的使用效率造成影响。一个业务运营，想要使用这张表，他首先要能够找到这张表的负责人，然后联系对其授权。这个过程往往要花费一天的时间。同时，作为授权方，数据负责人往往也不清楚，到底该不该把权限授予申请人，这就造成一个很尴尬的情况，就是权限流于形式，只要你能找到表的负责人，就可以获得这个表的访问权限。

第三章 到底什么是数据治理?

3.1 数据治理宏观政策

目前，全世界已经进入数字经济时代，数字经济已经成为支撑当前和未来世界经济发展的重要动力。自十八大以来，党中央就高度重视发展数字经济，并将其上升为国家战略。19年11月首次将数据列为生产要素，20年4月在《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》中正式提出将数据作为与土地、劳动力、资本、技术等传统要素并列的第五大生产要素，数据要素是实施国家大数据战略、加快建设数字中国、深化数字经济化发展的核心引擎。

党的二十大报告中明确指出“坚持把发展经济的着力点放在实体经济上，推进新型工业化，加快建设制造强国、质量强国、航天强国、交通强国、网络强国、数字中国。”在此背景下，数字技术作为企业数字化转型的核心动力，赋能企业帮助企业完成数字化转型，提升企业竞争力开辟第二条增长曲线。而企业数字化转型过程中需要数据先行，以数据治理为肯綮，通过对数据进行规范化、标准化以及流程化的治理，提炼企业数据资产，激发数据要素潜能，实现企业数据资产化、价值化、智能化，从而助力企业完成数字化转型。因此，数据治理是企业在数字化转型过程中关键而又绕不开的一个环节。

3.2 数据治理的概念

那么到底什么是数据治理？数据治理的内容又包括哪些？这些都是企业在准备开始进行数据治理之前需要考虑清楚的问题。此外，企业在进行数据治理之时也经常容易走入误区，比如在项目初期就希望进行大而全的数据治理，往往因为缺失重点而导致事倍功半，又比如知道元数据、数据标准、数据质量在数据治理过程中的重要性，但是在实际交付过程中却发现难落地等等。所以在做数据治理之前，首先要了解数据治理，正所谓以汤止沸，沸乃不止，诚知其本，则去火而已矣。

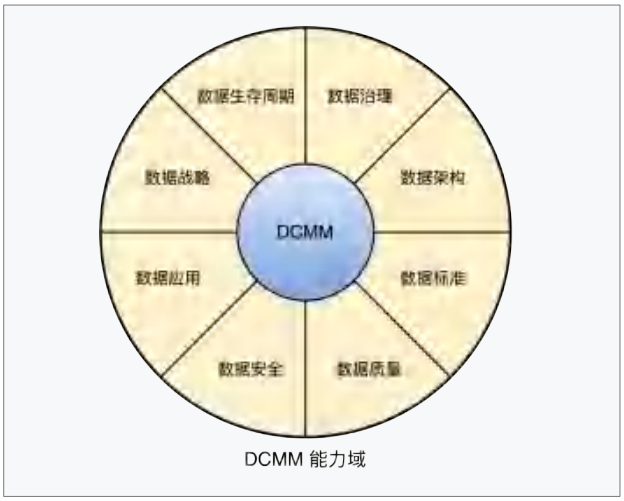
DAMA

国际数据管理协会（DAMA国际）在其《DAMA数据管理知识体系指南（DAMA-DMBOK2）》一书中将数据治理进行了定义，即在数据资产管理过程中行使权力和管控的过程，称为数据治理。并将数据治理作为数据管理十大知识领域的中心，负责知识领域的平衡和一致性。DAMA对于数据治理的定义显得较为抽象，但实际上去了解其中对于数据治理主要工作内容的定义，不难发现它是从数据战略，数据制度，数据架构、建模和设计等标准，以及数据监管合规、数据资产估值等维度进行描述。同时，DAMA-DMBOK2一书中还给出了数据治理实施和度量指标的相关指导和建议。可以说，DAMA从数据治理的定义、活动、工具和方法、实施指南以及度量指标给出了比较全面的解释，但是距离企业可落地的数据治理还是距离较远，更像是纲领性的介绍，因此对于如何进行数据标准的制定以及如何进行数据资产的评估都缺少具体的描述。

DAMA-DMBOK2 数据管理框架

DCMM

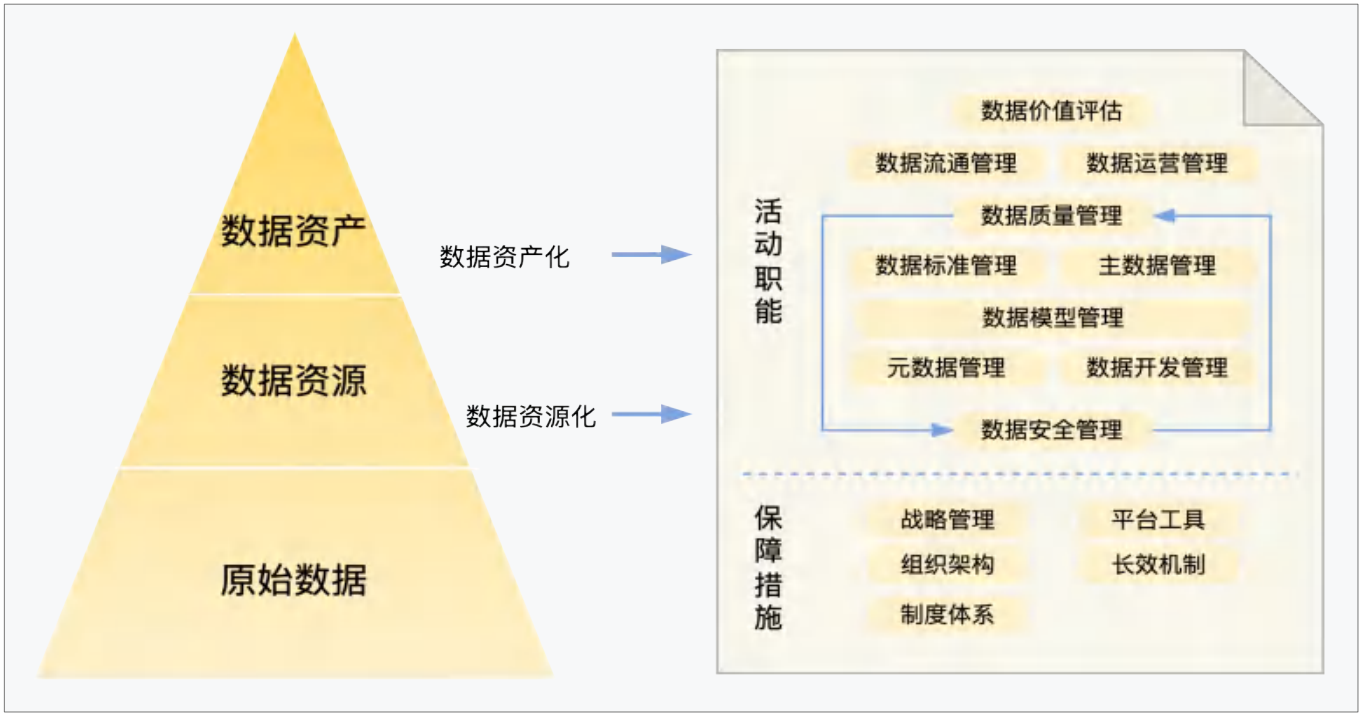
DCMM（Data Management Capability Maturity Assessment Model，数据管理能力成熟度评估模型）是我国首个数据管理领域国家标准。数据管理能力成熟度评估模型给出了数据管理能力成熟度评估模型以及相应的成熟度等级，定义了数据战略、数据治理、数据架构、数据应用、数据安全、数据质量、数据标准和数据生存周期等8个能力域。相较于DAMA，DCMM将数据标准作为数据管理中的独立一项内容，进行了明确的定义和能力等级说明。此外，还将数据开发、数据应用（包含数据服务、数据分析）等内容进行了相应定义和规范化说明。但是从整体上看，DCMM并没有提及数据资产相关的内容，缺少数据资产的评估。



DCMM 能力域

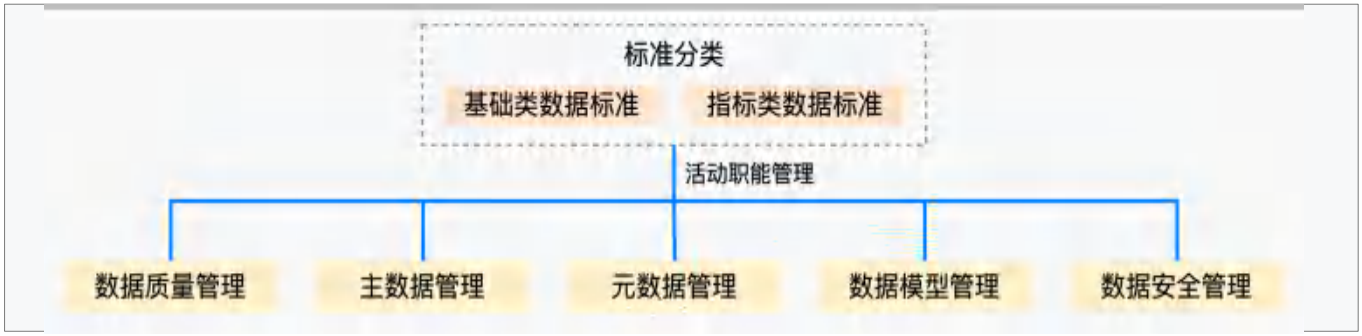
数据资产管理实践白皮书

《数据资产管理实践白皮书》是大数据技术标准推进委员会、中国信通院云计算与大数据研究所联合业内知名企业共同编写的关于数据资产管理实践的白皮书。该白皮书聚焦于数据资产的管理，有别于DAMA和DCMM，更加强调数据的资产属性以及价值，给出了数据价值的广义定义以及数据价值的评估方法。



数据标准管理实践

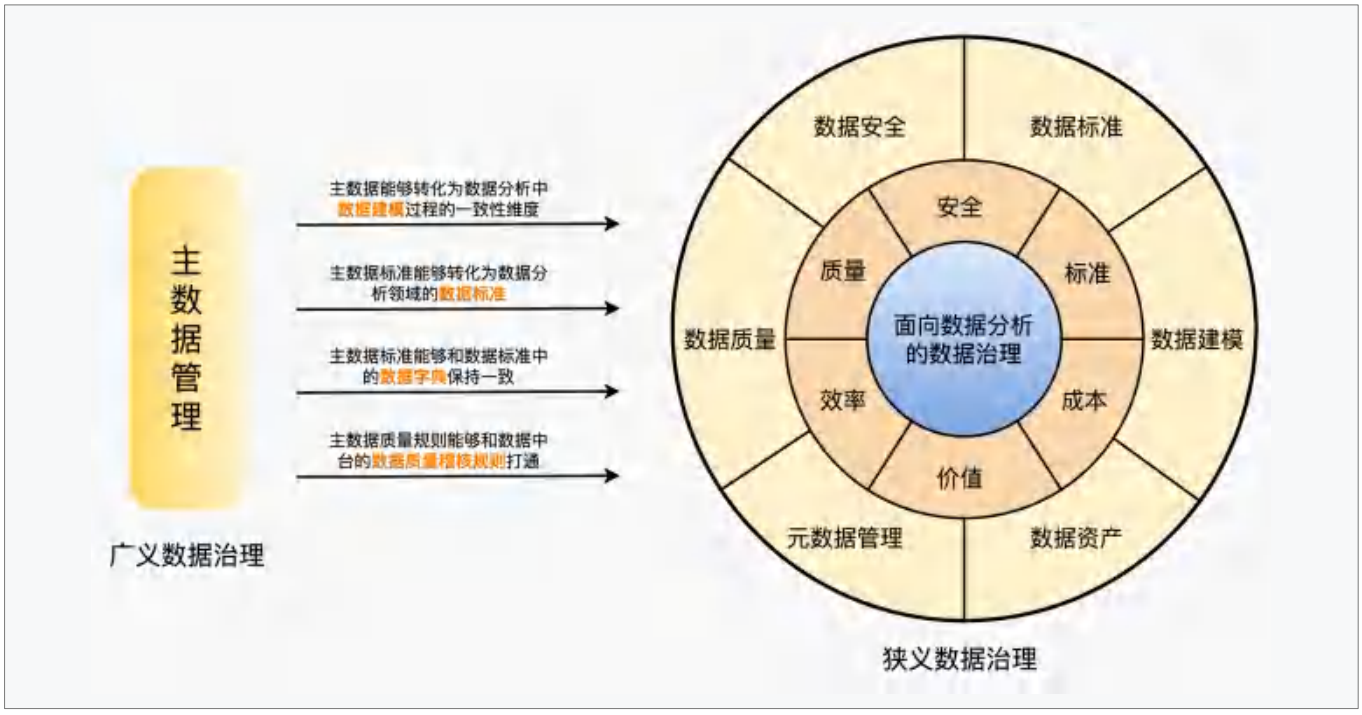
《数据标准管理实践白皮书》也是由大数据技术标准推进委员会、中国信通院云计算与大数据研究所联合业内知名企业共同编写的关于数据标准管理实践的白皮书。该白皮书聚焦于数据标准，提出了数据标准是数据资产管理多个活动职能的核心要素，主要体现在数据质量管理、主数据管理、元数据管理、数据模型管理和数据安全管理几个方面。



3.3 网易对数据治理的定义

网易认为，数据治理是对企业全域数据资产实施有效管理的活动，根据数据治理的对象不同，数据治理可以分为面向业务系统的数据治理和面向分析系统的数据治理。主数据管理就是典型的面向业务系统的数据治理，它核心要解决的问题是跨业务、跨系统和跨流程的企业核心数据的一致性、正确性和权威性的问题。面向分析系统的数据治理，主要解决的是数据分析过程中，指标数据计算的口径一致性，数据质量，标准规范、成本优化以及安全管控问题。因为数据本身还是来源于业务系统，所以面向业务的数据治理是面向分析的数据治理的前提，如果业务系统的数据治理没做好，分析系统的数据治理就很难从根本上解决。

数据中台其实跟传统的数据治理的概念并不相同，但是数据中台的核心思想是构建统一的指标管理体系和企业级公共共享的数据模型层，打破烟囱式的数据架构，本质还是面向分析系统的数据管理，所以从数据治理是企业所有数据管理活动的定义来看，也可以将数据中台的构建方法纳入面向分析系统的数据治理。



第四章 传统数据治理面临的挑战

传统数据治理包括三大件，分别是数据标准，元数据管理和数据质量。数据治理的一般流程是从制定数据标准开始的，简称定标。然后通过元数据管理的采集、注册、扫描以及发布完成数据标准和数据模型之间的连接，这个过程称为落标。最后通过数据标准关联的数据质量稽核规则，对数据模型进行稽查，发现质量问题形成质量报告，推动相应的业务部门进行整改，提升数据质量。

对于传统数据治理来说，其更加强调对业务系统存量数据的治理，此外，对于数据长效治理体制的建设也不够重视，所以传统数据治理在企业数据管理实践中也遇到了一些新的问题。

4.1 开发与治理脱节

传统数据治理面临的第一个挑战便是开发与治理的脱节。由于传统数据治理更加关注存量数据的治理，而忽视了新增数据的长效治理，导致企业需要通过不断的数据治理项目维持数据治理的效果。但是，对于企业来说，相比于存量数据其增量数据价值更高、也更为重要。此外，传统的数据治理是一个反向治理的过程，并不会融入到数据生产的整个过程当中，与数据开发、建模、运维、安全等环节都存在脱节现象，对于企业来说进行传统的数据治理就需要对其现有的系统和流程进行改造，因此必然面对高昂的成本。基于上述情况，我们需要将数据治理的活动前置，从数据的生产环节加入数据治理的活动。

数据质量与数据开发脱节

如何确保数据开发的结果符合业务逻辑并能被业务方所使用，就需要通过质量稽核规则对开发结果进行监控。但是，我们发现在实际开发过程中质量稽核规则的覆盖率只有10%。其次，由于不同的开发人员对数据的了解程度以及业务的熟悉程度不同，对于相同数据项的质量稽核规则设置也不尽相同，在早期的网易内部有70%的相同数据项，其稽核规则存在不一致，阈值设置也不一致。这就导致了数据开发的结果大多不符合业务方的预期，长此以往，业务方不再相信数据。究其原因，首先是质量稽核规则缺少统一的标准，其次开发人员对于数据质量的重视程度不够，导致数据质量和数据开发严重脱节。

数据标准与数据建模脱节

数据标准一般会包括标准规划、标准制定、标准发布、标准执行、标准检查等流程。一个企业会根据自身的情况结合国家标准、行业标准制定自身企业的数据标准。但是标准制定之后如何让开发人员贯彻执行却是大多数企业面临的问题。标准和数据建模的脱节，就会导致开发出来的表的命名无法统一、缺少规范，相同字段的名称也会因为开发人员开发习惯的不同而出现不同的命名方式，从而导致数据的理解成本和管理成本上升。

元数据与数据开发脱节

在数据开发过程中，任务之间往往存在依赖关系，下游任务运行依赖于上游任务的实例产出，因此，需要将有关联的上下游任务进行依赖关系配置。在现网环境中，客户的任务数量往往能够成百上千甚至达到万级，要在如此多的任务当中完成依赖关系的配置，就非常考验开发人员对任务的熟悉程度，而且这种通过手工配置的方式极易出错，一旦依赖关系的漏配就会造成任务的空跑，导致下游产出数据的异常。因此，在任务依赖配置中，能够自动推荐上游依赖任务就显得尤为重要，而要实现自动推荐上游依赖任务的关键便是元数据，通过元数据获取任务间的血缘关系，根据血缘关系推荐上游依赖任务。元数据和数据开发的脱节导致在开发过程中，任务间的依赖配置更多的是通过手动维护，这就大大增加了出错的概率。

元数据与任务运维脱节

任务运维过程中，当大量任务运行出现了异常情况，此时运维值班人员需要知道任务处理的先后顺序从而保证重要任务被优先恢复。此外，可能有些任务的故障是因为上游任务的异常所引起的，因此知道故障的源头任务也非常重要，而这些都离不开元数据。元数据与任务运维脱节导致运维值班人员无法识别重要任务，也无法快速定位故障的源头。

数据标准与数据安全脱节

随着数据量的增大，使用数据的人数变多，数据管理的难度呈指数上升。数据的权限如何设置？数据的安全等级如何定义？哪些数据需要脱敏？哪些人可以被赋权？这些都是对企业的考验。数据标准是制定数据安全策略的依据，数据标准与数据安全的脱节使得企业对数据安全管理时失去了抓手，变得盲目。

4.2 烟囱式的数据开发

在网易，早期的数仓建设更多的是业务部门各自内部进行维护，并没有上升到组织架构的层面进行规划。各业务部门内部都存在数据开发和分析人员，承载本团队的数据分析需求。这样的一个组织架构导致的结果就是存在很多的零散分裂的小数仓，对于企业内部的数据管理、数据共享造成了极大的影响。

烟囱式的数据开发容易造成指标口径不一致、数据重复开发等问题。中台建设前，因为缺少统一的规范建模标准，各业务部门存在大量名称相同但是口径不一的指标，导致业务人员看不懂数据造成了极大的困扰。此外，还因为建模混乱，超过40%的表都没有分层，每当开发人员接到新的开发需求时，因为无法高效复用已开发的表，导致超过50%的任务直接读取原始数据，每次需求都要重新发开，开发效率低下。烟囱式的开发还会造成资源的浪费，系统中存在大量的长时间未曾访问且重复的数据。

4.3 不同平台之间缺少统一的管控

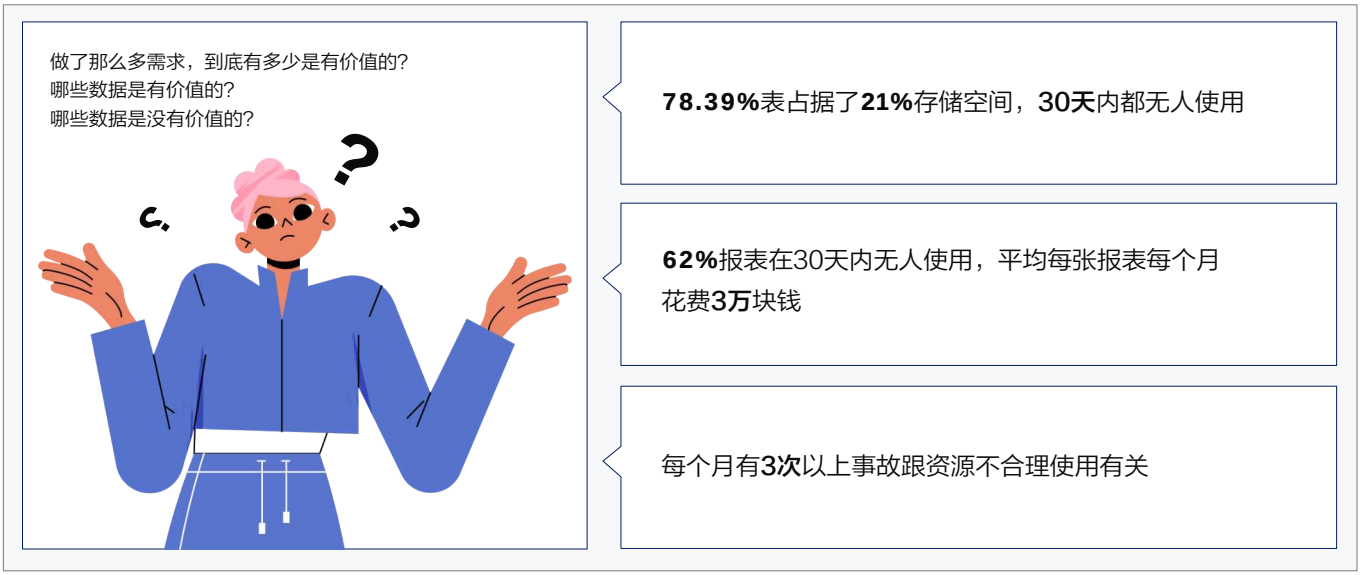
大型企业的IT架构往往存在不同的平台系统，如关系型数据库系统、分析型数据库系统、数据湖系统等，各系统从后端到前端相互独立、紧耦合开发，导致系统臃肿、建设效率低、无法快速响应业务，且存在大量重复建设工作。因此，建立统一标准的大数据开发与治理平台成为刻不容缓的集团战略。而各个独立系统中存在的大量历史数据及任务成为了中台建设的阻碍，不仅需要承担高昂的数据迁移成本以及面对迁移过程中必然出现的数据遗失问题，同时还要培养人员掌握新的开发、分析工具，这些问题都会让企业对中台化望而却步，所以需要一个能够统一管控不同平台的开发与治理平台。

4.4 治理过程缺少可量化的监控

在数据治理过程中存在这样一个现象：好像做了很多又好像什么都没做。治理过程的难衡量、忽视可视化成果的展示导致领导或者客户不易感知数据治理的成果，从而无法认同治理团队的工作。最终导致治理项目难验收、员工工作成果难展现。因此，在数据治理过程中除了要有阶段性的目标还需要可视化效果的呈现，例如：管理发布了多少元数据，这些元数据在哪里能够被看到？制定了多少数据标准，这些标准引用的情况如何？又有多少标准处在发布状态？构建了多少资产目录，资产在目录中挂靠情况如何等等。通过可视化的展现同时结合阶段性的目标就能很好的反应出治理阶段性的成果。

4.5 对数据的成本和价值缺少精细化的管理

随着企业业务的高速发展，数据量呈指数增长，相应的数据成本也极具增加，因此企业需要识别有价值的数据、去除无用数据、沉淀数据资产。但是对于企业来说，因为缺少对数据成本和价值精细化的管理，导致如何在日常数据开发、运营过程中发现无用的数据成为了一个非常棘手的问题。没有做好公共资源的复用、没有去基于ROI的方式沉淀数据资产、成本的日益增长而交付效率却无法提升，对数据团队是极大的考验。网易内部统计过，78.39%表占据了21%存储空间，而这些表在30天内都无人使用；报表亦是如此，62%的报表在30天内无人使用，平均每张报表每个月需要花费3万的成本费用。从上面的数据可以发现，数据团队日复一日的响应需求，但是实际上真正有价值的表只是其中一部分，这就是缺少对数据成本和价值精细化管理导致的。



4.6 数据治理缺少闭环

很多企业在谈到数据治理时，认为只要将数据标准制定好、质量规则都配置完成、数据资产都上线便可以。但在实际的治理过程中会发现，业务人员、技术人员完成新系统的搭建后很快便会在这上面遇到新的问题，比如配置好了质量的稽核规则，也通过规则找出了一大堆质量问题，但是然后呢？往往结果便是不了了之，质量问题得不到落实，不该出现的问题反复出现，质量规则形同虚设。又比如数据资产的消费者在查看数据资产的时候发现了问题，因为缺少相应的反馈机制，导致数据资产得不到及时有效的治理，长此以往，有问题的数据资产便会越来越多。

事实上，数据治理是一个长期、可持续的过程，因此需要在治理活动的各个环节做到闭环，保证治理的结果切实落地，而传统数据治理却在治理活动的展开过程中缺少相应的闭环机制。例如，当数据资产的消费者在资产目录中浏览时发现数据有问题时，可以直接通过申请数据治理工单将问题反馈给数据治理部门，由数据治理部来对问题进行初步判断。如果发现是业务相关的问题，就将工单派发给数据的业务负责人；如果是技术问题，就发给技术的负责人。问题修复后，工单重新流转给数据治理部，数据治理部审核通过后进行重新发布，同时通知申请治理的数据资产消费者。对于上述的例子，传统数据治理只考虑如何治理数据，缺乏数据治理过程中发现问题、解决问题的流程和方法，自然而然也无法对治理进行闭环。质量问题亦是如此，而发生这种现象的原因是没有形成数据质量问责的闭环，对于质量问题查找出问题环节、定位数据问题、最后实行问责机制，实现问题的闭环。

4.7 忽视了开发过程中效率和质量的问题

开发人员需要对已开发的任务配置充分的质量稽核规则，从而确保开发结果的准确性。但是在实际开发过程中，往往因为质量稽核规则缺少标准化、开发人员的经验参差不齐等因素导致任务整体的质量稽核规则覆盖率较低，存在少配、漏配甚至没有配置的现象，最终导致开发事故频发。经统计数据开发任务变更导致的生产环境数据问题占比达到65%，例如网易内部某电商活动，因为上游任务变更，导致下游涉及资损数据计算异常，红包超发，造成几十万的生产事故。

在数据治理过程中，会引入很多的标准和规则去约束开发流程，在此情况下如何保证开发人员效率的同时提升数据的质量是一个必须重视的问题。优秀的产品能够有机的将数据治理和数据开发过程结合起来，两者相辅相成，可以在开发过程中融入数据治理过程而不影响开发效率。

第五章 网易数据治理 2.0

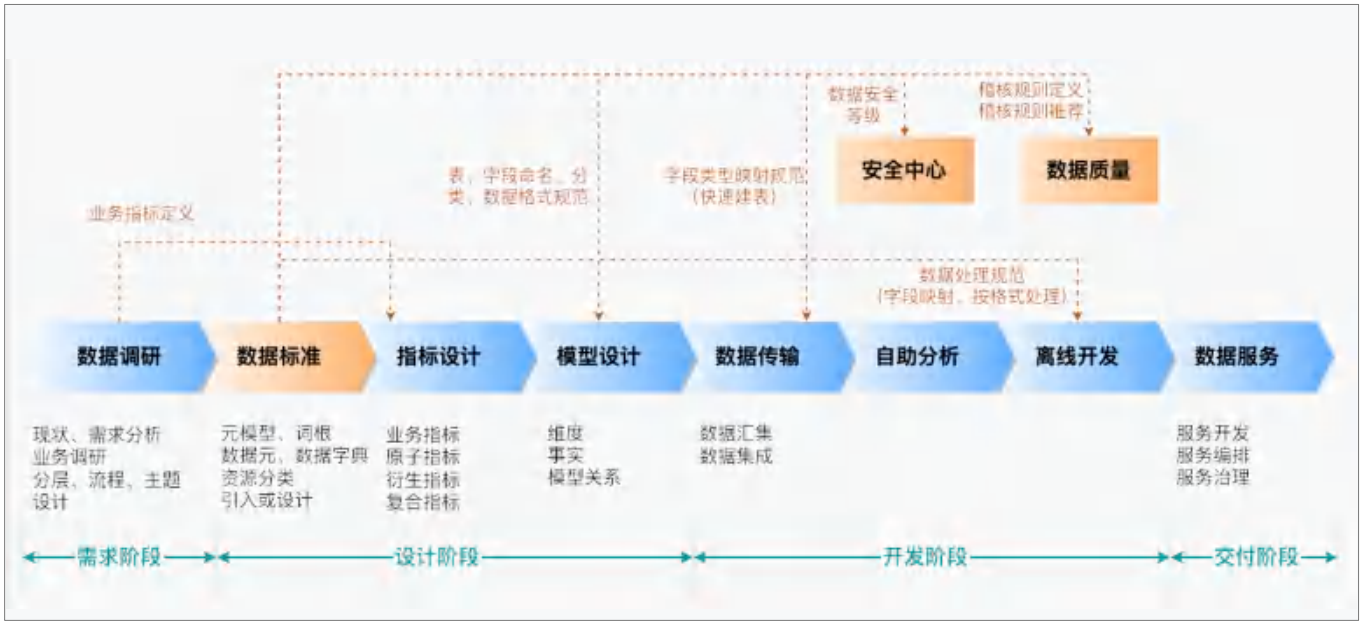
在传统数据治理的基础上，网易提出“开发治理一体化”的核心思想，将数据治理的过程前置到数据开发环节，确保生产出来的数据就是符合标准和规范的，同时，结合数据中台的构建思想，打破烟囱式数据架构，构建统一的指标管理体系和企业公共数据模型层，通过服务化的方式对外提供服务。

数据治理2.0的核心特色：

- ✓ 开发与治理一体化
- ✓ 数据中台架构
- ✓ 基于DataFabric的逻辑数据湖
- ✓ 基于DataOps开发底座
- ✓ 数据治理的持续闭环
- ✓ 数据治理360
- ✓ 基于ROI的数据资产精细化管理

5.1 开发与治理的一体化

数据开发与治理一体化是将数据治理的过程融入到数据开发的全生命周期中，强调“先设计、后开发、先标准、后建模”的原则，其目标就是将数据治理的流程与数据开发的全生命周期相融合，在数据开发过程中，完成数据治理。通过指标和数据标准的定义实现“规范即设计，设计即开发，开发即治理”的开发治理一体化理念。



我们将整个开发治理流程分为四个阶段，分别是需求阶段、设计阶段、开发阶段以及交付阶段。

需求阶段

在需求阶段，我们需要对当前企业的业务现状进行分析，了解客户的业务诉求，完成数据和业务的调研。数据调研是对客户业务及数据盘点的过程，由于在进行标准设计时需要进行元模型的构建，同时完成词根、数据字典、数据元的制定和录入，因此需要进行表和字段的调研。表调研包括表范围、存储位置、中英文名、来源系统、优先级、更新方式、业务场景以及数据的条数和增量条数等。表级别的数据调研主要是为了数据接入做准备，根据存储位置、来源系统、优先级制定接入计划以及确认哪些资源需要实时接入、哪些资源需要离线接入。字段调研则包括表结构、字段类型格式、样例数据、有值行、有值率、是否需要制定标准、是否挂载已存在标准、值域分布等，字段级别调研可以增加数据标准化的效率。此外通过业务的调研，对指标、模型分层、主题进行初步的梳理为后续的指标和模型设计做准备。

设计阶段

标准设计

标准的构建流程主要包括标准规划、标准制定、标准发布、标准执行、标准检查五个步骤，其中标准规划、标准制定、标准发布属于标准的设计环节，标准执行属于标准的落标环节，标准检查属于标准落标的验证环节。

首先来看标准规划，标准规划有多种方式，可以收集现行的国家标准或者行业标准，也可以结合企业自身业务特点根据实际需要进行标准的梳理，但是不管哪种方式，规划标准时尽可能的围绕提升企业在业务协同、监管合规、数据共享开放、数据分析应用等各方面的能力进行展开，对于非交互、非公开的数据，其标准化的优先级便没有那么高。

其次是标准制定，在该阶段主要的活动包括词根的制定和录入、标准字典的制定和录入以及数据元的制定和录入。词根作为企业维护的标准词库，用来确保统一含义的中文词能够被翻译成相同的英文名称，解决字段名称含义不明确的问题。如果企业已维护了词根列表，可直接导入到平台中。如果没有维护词根列表，可以根据需要进行词根的制定，词根制定的过程包括词根的拆解、去重、定义。词根拆解是将原始数据的字段名称进行拆解，拆解方式可按照具体的业务诉求进行。对于拆解完成的结果进行去重，去重后的词根需要重新进行标准化的定义，按照网易数据标准词根的管理要求，需要完成中文全称、英文全称、英文简称等内容的制定，对于完成定义的词根便可在平台中进行录入，录入后的词根将应用于数据元命名、数据字典命名、字段命名等。

标准字典是维护某个枚举类型字段的标准代码集。以“物料类型代码”为例，描述的则是5代表采购物料，10代表制造物料这样的关系。由于企业当中不同的业务系统可能使用不同的枚举范围，因此需要再中台层面进行统一，可以直接引用国家标准或行业标准，也可以根据多个来源系统的统一类型字典的数据进行合并、去除、修改等操作形成一套标准字典，尽可能满足各业务系统的需求，同时也满足标准化的要求。标准字典制定步骤分为定义、录入以及审核，定义主要是根据标准规划内容中原始数据里的枚举字典（即原始字典）进行标准化制定。

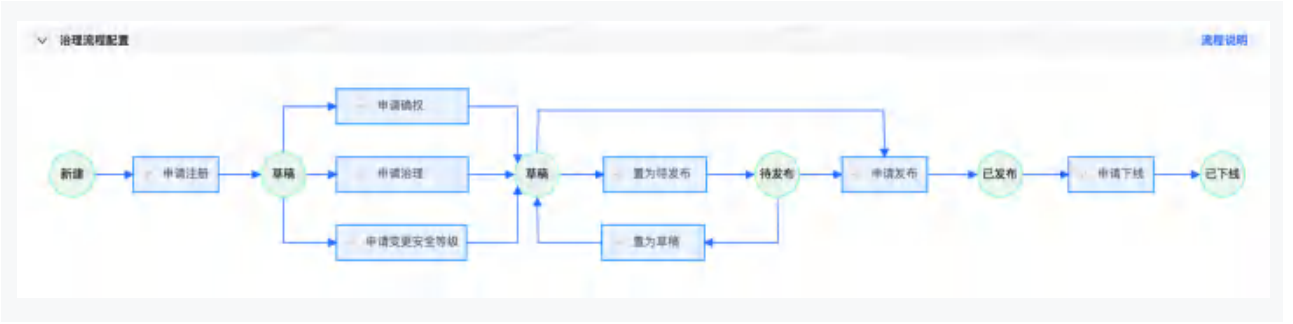
标准制定的第三块内容是数据元制定。在GB/T 18391.1-2002中将数据元定义为用一组属性描述定义、标识、表示和允许值的数据单元。数据元是基础类数据标准的具象化体现，也是数据标准管理的核心。数据元制定的步骤和数据字典类似，包括定义、录入和审核。数据元的定义是对原始数据进行结构化提取的过程，按照数据元的管理要求，完成数据元中文名称、英文名称、数据类型、数据格式、值域等属性的描述。

标准设计的第三阶段是标准发布，标准发布用于将处于具备发布条件的数据元（标准态）、标准字典（标准态）进行整包发布，审核通过后便可应用于整个系统。

指标设计

指标设计阶段是针对已经规划好的指标内容进行制定，主要包括指标的定义和录入步骤。首先是指标的定义，将业务指标进行分析和拆解，得到原子指标、派生指标以及复合指标，同时明确数据域、业务过程、修饰词、衍生词等相关内容，例如派生指标最近1天PC端支付金额，通过拆解可以得到其业务过程为支付、修饰词为PC端、时间周期为最近1天、原子指标为支付金额，同时规划其指标域为订单域。

指标录入的过程是指标治理的流程，整个流程包括指标的新建、注册、确权、治理、发布等步骤，如图所示。



当在指标的使用过程中发现指标存在问题则可以通过申请治理，提交问题工单，交由数据治理专员进行初审，如果是业务口径存在问题，则将工单指派给负责维护指标业务口径的业务部门进行确认修改；如果是技术口径存在问题，则将工单指派给数据开发部门进行确认修改，修改完成后的指标通过申请发布重新上线，并生成相应新的版本，方便管理。

模型设计

在模型设计阶段，需要完成维表、事实表以及模型关系的定义和构建，同时通过标准在该阶段的落标以及与指标系统的挂钩，将字段命名、分类、数据格式等进行规范化的定义。

将标准导入后，系统会自动完成字段名称、字段描述、数据元、数据字典等信息的填写，规范了表结构。表构建完成后，通过数据地图查看表的信息，了解字段和数据标准的绑定情况。此外，在模型设计建表过程中提供标准化标签支持字段绑定已制定好的维度和指标，从而确保指标、维度的统一。

开发阶段

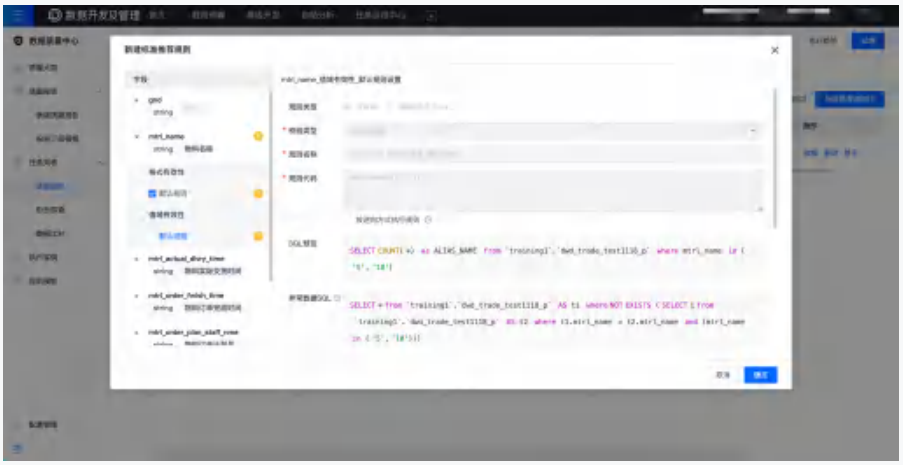
数据传输、自助分析、离线开发

在开发阶段，将设计阶段通过规范化构建的数据模型进行物理实现，与技术元数据（血缘、质量、调度任务信息等）和标准规范相结合，实现模型设计与数据开发的协同，真正意义的完成了元数据的标准化落地。

数据传输作为数据的采集模块，经常需要对接不同的数据源以及业务系统。在数据的采集过程中，要根据不同的行业标准和业务需求将字段类型完成转换，例如在金融行业对数据小数点后保留的精确位数和制造业的要求就不一样。因此对于不同的企业来说，需要一份可自定义的字段类型标准映射规范，应用于企业内部不同业务系统的数据集成，确保入湖后的数据字段类型统一，符合监管部门规定。此外，该标准化体系的制定，还能减少开发人员手动调整字段类型的频率，减少人为失误带来的开发问题，提高开发人员的开发效率。

数据质量

我们通过标准规范了模型的设计，通过模型明确了开发的内容，通过开发完成了标准的真正落地，而落标之后就需要标准检查来验证落标的实际情况。数据标准指导数据质量中心进行监控，形成基于标准的通用稽核体系。在标准的执行阶段，因为表关联了数据元，且部分字段还关联了标准字典，因此能够根据标准来自动生成质量稽核规则，如下图所示，在质量稽核规则任务创建过程中引用通过标准生成的格式有效性和值域有效型规则。



数据安全

在数据标准中还包含了业务敏感数据的对象和属性，从而实现对数据安全管理相关规则的定义。将安全中心的敏感类型、安全等级、脱敏规则、脱敏算法以及脱敏配置同数据元进行关联，快速生成字段级加密或脱敏规则。

当在模型设计过程中引用了数据元作为表的字段，则会根据该数据元的敏感类型和脱敏规则在数据传输中实施静态脱敏；在离线开发过程中会根据该字段敏感类型以及脱敏规则实施动态脱敏，同时根据脱敏配置对不同的用户实施不同的脱敏策略。

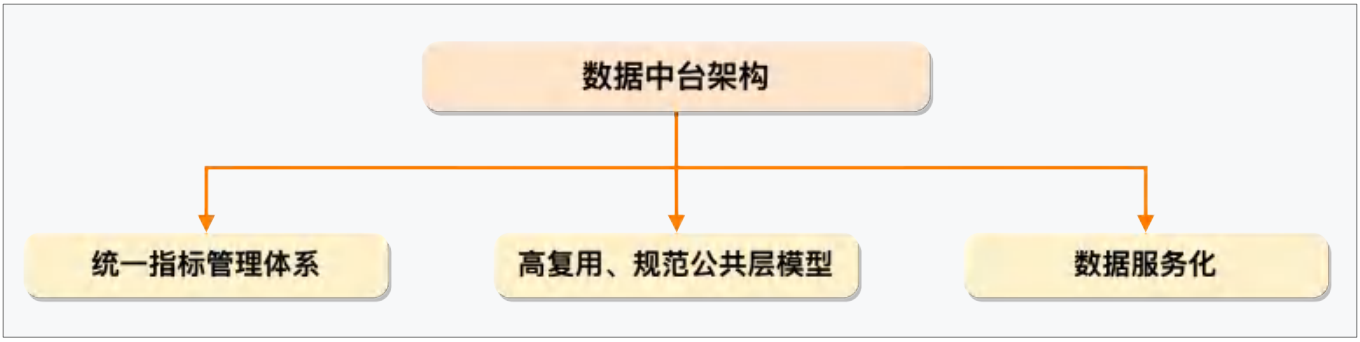
此外，对于引用了数据元的表，会根据相应的安全等级提供分级分类的依据，对于不同密级的数据提供不同的资源申请流程，从而确保数据的安全性。

交付阶段

数据服务
在交付阶段，数据开发人员通过可视化的配置方式对已加工好的数据进行服务的开发和编排，数据消费人员通过API集市查看已发布的API的调用说明并根据需要申请API的使用权。此外，通过权限、熔断、限流技术极大改善了数据交付过程中的质量、安全问题。平台支持对API使用情况的进行统计，了解API的调用次数，同时结合平台的治理模块对API资产进行盘点，挖掘闲置API和异常API并通知给负责人进行治理整顿。

5.2 数据中台架构

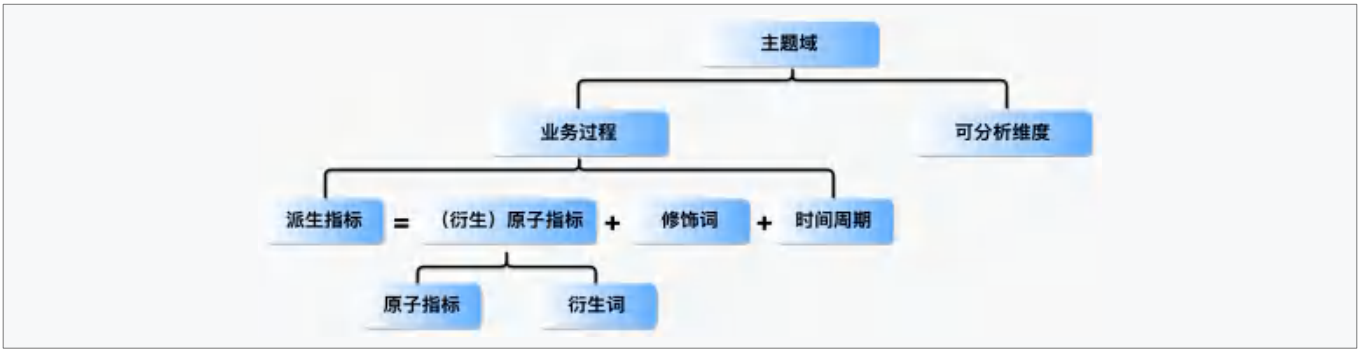
数据中台架构需要包括统一的指标管理体系、高复用、规范的公共层模型以及可交付的数据服务。



统一的指标管理体系

指标是数据和业务的交汇点，是数据分析需求的载体。如果指标口径定义不一致，看数据的人就无法正确的理解数据。长此以往，指标数据的可信度降低，严重影响问题的分析决策，最终导致数据失去分析价值。为了确保指标口径一致，就必须要实现指标的统一管理。指标统一管理需要组织架构、流程规划和工具产品的三者结合。首先，要有能够统一管理指标的组织，这个组织必须是跨业务部门的，一般就是数据中台部门。其次，要有统一的指标管理流程规范，包括指标的规范化定义、指标分类管理以及审批流程等。最后，指标的管理必须还要有与规范相配合的工具产品。

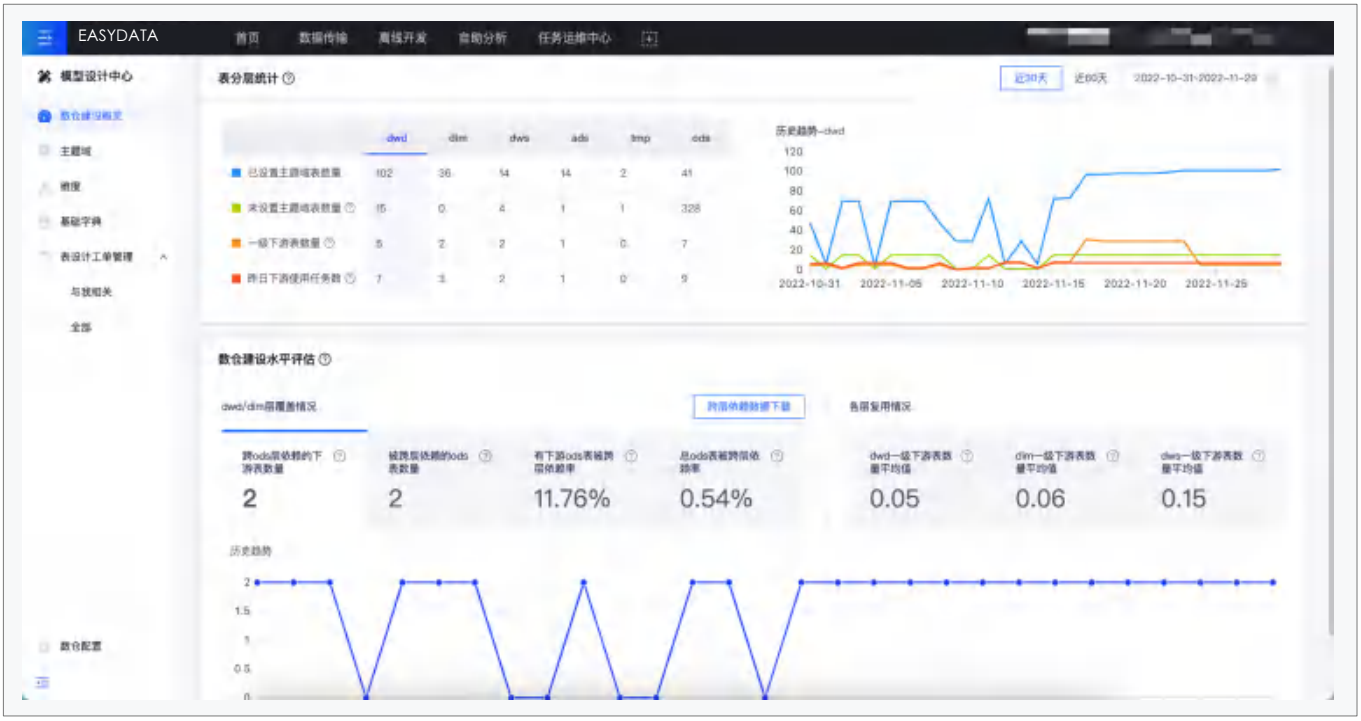
网易经过多年的实践结合指标管理方法论打造了统一的指标管理体系，对业务进行分析，划分和定义主题域、业务过程、维度、修饰次、时间周期、原子指标、派生指标、复合指标。同时，将涉及口径的原子指标再进行细分，得到主原子指标和衍生原子指标，衍生原子指标由主原子指标和衍生词构成，衍生词不同于修饰词带有计算口径，因此衍生词的构建需要进行审批从而保证口径的一致性。



此外，系统通过指标血缘关系的展示解决指标来源不清晰、难追溯的问题。根据指标加工逻辑构建指标血缘关系，当业务端质疑指标异常或需要确认指标口径时，可以基于血缘关系找到指标的源头以及相关的表，从而提升问题查找的效率。

高复用、规范的数据模型

网易同样认为数据模型是构建数据中台的基石，一个面向数据中台的模型设计，必须有一套可以量化的衡量标准，能够评价当前数据模型设计的质量。网易推荐的数据中台的建设方式是采用迭代式构建，所以必须要对建设过程中模型的设计质量进行持续跟踪，确保模型的设计符合数据中台建设的规范和高复用的设计目标。为此，网易提出了业界首个面向数据中台的模型设计标准，提出通过跨层引用率、模型引用系数等指标评价模型设计质量。



完善度

对于模型建设的完善度，网易引入跨层引用率的概念，旨在通过对未遵守模型建设规范、直接跨层依赖ods层的情况进行分析和统计。对于模型构建过程中，我们认为应该根据模型架构分层，逐层引用进行建表，直接跨过dwd层或dim维表直接对ods层的表进行引用，不仅无法体现dwd层表的使用价值，更为主要的是跨层建表的方法会造成烟囱式的开发，不利于数仓的搭建。因此，网易针对建表过程中跨层引用的情况给出了相应的度量指标，分别是跨层依赖ods层的下游表数量、被跨层依赖的ods层表数量、有下游ods表被跨层依赖率、总ods表被跨层依赖率，其含义具体如下：

- 跨层依赖ods层的下游表数：该指标反映了有多少非dwd、dim表直接引用了ods层的表，将这些表的数量进行统计，数量越接近0，说明数仓建设越好。
- 被跨层依赖的ods表数量：该指标反映了有哪些ods表被非dwd、dim表直接引用，将这些ods表的数量进行统计，数量越接近0，说明数仓建设越好。
- 有下游ods表被跨层依赖率：该指标指的是被跨层引用的ods层表与有下游表的ods层表的比值，反映了被跨层引用的ods层表的占比情况，占比越低说明数仓建设越好。

完善度

- 总ods表被跨层依赖率：该指标指的是被跨层引用的ods层表与ods层所有的表的比值，反映了被跨层引用的ods层表的占比情况，但是因为ods层表数量会不断增加使分母变大，因此会导致统计结果波动下降，整体来说也是越低说明数仓建设越好。

复用度

数仓模型构建过程中除了考虑其建设的完善度还要证明其建设的价值，网易从各层表被下游引用的情况入手，考察表的复用程度，制定了复用度指标。通过表的复用度能够让开发人员直观的了解到表是否闲置、是否使用率低下，对闲置的表进行下线，对复用度低的表进行优化，避免烟囱式开发的同时，提高了数据开发效率，加快了对业务部门的响应速度。根据数仓的分层情况构建如下三个复用度指标，分别是：dwd以及下游表数量平均值、dim一级下游表数量平均值、dws一级下游表数量平均值。

- dwd一级下游表数量平均值：该指标指的是dwd层一级下游表数量与dwd层已设置主题域表数量的比值，该指标越高越好，越高说明dwd层被下游复用的次数越多，产生的价值可认为也越大。
- dim一级下游表数量平均值：该指标指的是dim层一级下游表数量与dim层已设置主题域表数量的比值，该指标越高越好，越高说明dim层被下游复用的次数越多，产生的价值可认为也越大。
- dws一级下游表数量平均值：该指标指的是dws层一级下游表数量与dws层已设置主题域表数量的比值，该指标越高越好，越高说明dws层被下游复用的次数越多，产生的价值可认为也越大。

规范度

数仓建设的规范上构建数仓的基础，无论是完善度还是复用度的统计都是建立在规范建模的基础上。规范度的判断主要从表的命名是否规范、是否归属到具体的分层和主题域。在实际的交付场景中，开发人员往往为了加快表的开发速度，绕过建模工具，直接通过SQL语句进行建表，通过这种方式建的表没有归属到具体的分层和主题域，表名也可能是不规范的，因此，需要通过模型管理工具去制定表名的命名规范，并甄别出表名不规范的表，将其整理成清单告知表负责人进行整改。通过表名可将处于游离状态（无分层和主题域归属）的表进行抓取并关联，对于无法设置归属的表可根据表使用情况进行处理，对于无活跃下游表、无数据、无调度产出的表进行治理下线。

基于这套规范，在网易内部治理跨层依赖模型200+，跨层引用率从30.8%降低至9.42%，迁移下线3.4万模型，模型复用率从2.4提高到9.6，基于此不但需求平均交付速度从一周提升到3天，而且对大量明细数据的平均查询也降至21秒，在节省内部资源的同时提高了效率，实现了降本增效。

数据服务化

网易认为数据服务一方面作为服务网关，提供了鉴权、日志、限流、熔断等能力，实现了数据的规范化交付，提高了数据交付的效率。同时，数据服务还必须具备编排的能力，通过API之间的组合，可以创建出不同应用场景的服务。此外，数据服务还能够提升数据管理的效率，数据服务打通了数据应用和数据模型的血缘，实现了数据血缘从源系统到数据应用的全链路覆盖。一方面出现问题的时候，数据中台可以快速评估影响范围，另外一方面，也可以通过数据服务，对产生的脏数据进行快速的止损。数据服务化主要体现在服务交付效率、服务共享与安全、服务血缘分析、服务编排能力。

服务交付效率

数据团队承接业务方的需求，需要讲数据仓库中海量数据通过接口化方式交付给数据使用方，但开发和维护API的链路相对较长，投入成本过高，数据也更容易滞后。此外，API对接不同的业务方，呈现烟囱式开发导致API复用率极低，加大了开发人员的工作负担。平台秉承“配置即服务”的理念，通过可视化的配置方式，开发人员不再需要重复编写代码开发数据接口，只需在平台上进行简单的配置，便可自动生成和发布数据API。此外，通过在线升级，对于小的逻辑改动，不需要将API进行下线，对数据进行热更新，实现了在线API的灰度升级与快速回滚功能，让“边开车边换轮子”成为了可能。

服务共享与安全

API在开发、调用过程中，如果缺乏安全管控，极容易造成数据泄露的风险。平台提供多重管控策略，解决API在对外服务中的安全问题，主要策略包括流控策略、访问策略、报警策略、行/列级权限控制策略。数据服务通过支持流控策略，限制单位时间内最大调用次数保障API的稳定调用；通过黑白名单访问策略，保证仅授权或被禁止的IP地址实现调用API的数据的需求；通过报警策略，时刻监控重要的API的调用状态，满足应用方平稳顺利的数据使用；通过控制行/列级权限，实现API开放的灵活性和安全性，让开放出去的数据能真正的服务于业务。此外，通过数据集市提供统一的服务申请入口，用户可根据自身需要进行申请，由API负责人审批通过后进行授权。

服务血缘分析

在传统的API开发过程中，数据团队更关注于将数据表构建为API开放给业务人员使用。但随着业务的快速发展，需求日益增多，开发成本越来越高，开发人员需要提高API的复用率从而提高需求的交付效率。因此，开发人员要了解已开放出去的API是否还在使用以及使用效果如何。数据服务通过构建表、API和应用之间的血缘链路，支持从任一视角查看上下游关系，同时结合子产品管理统计调用、闲置、异常的API信息并形成列表，为API治理提供依据。结合API血缘信息，让API治理有迹可循，实现API数据应用价值最大化。

服务编排

在复杂的交付场景中，业务人员往往需要对已调用的API进行二次加工才能满足交付需求，无形中增加了工作的复杂度。数据服务通过提供服务编排能力，支持在画布中拖拽API节点、python节点、条件判断节点和UDF节点，实现将API参数进行复杂的逻辑处理，满足业务多种数据使用需求，由原本调用API后进行二次开发转变为只需要调用服务编排API，极大的提高了数据开发效率，简化了业务方使用数据的复杂度，进一步降低了数据API的开发门槛。

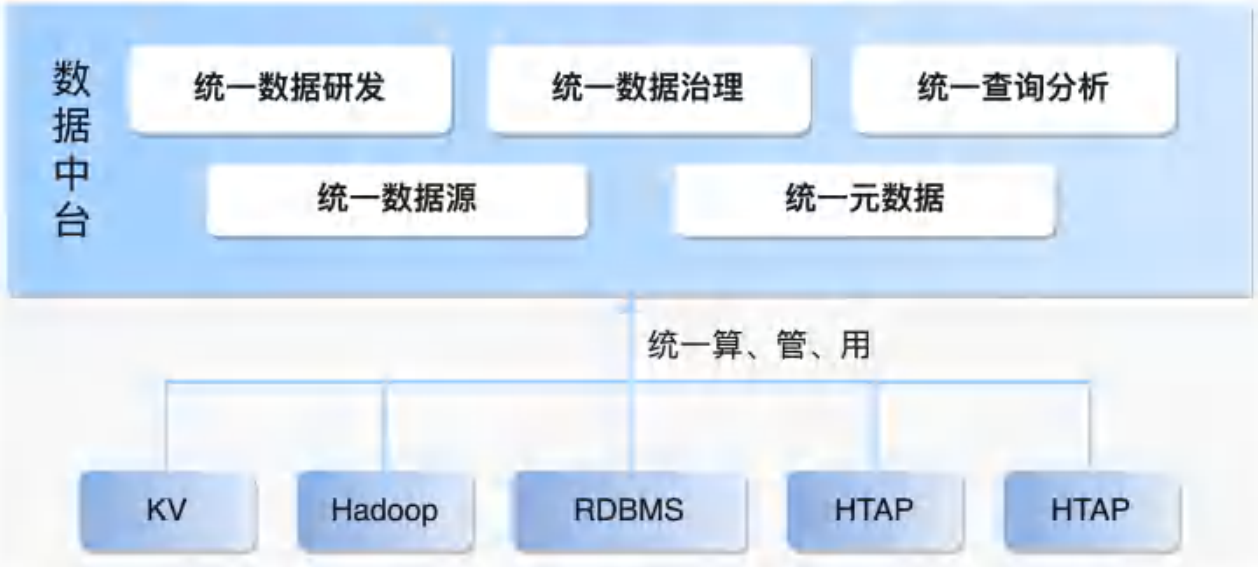
5.3 湖内、湖外数据统一治理

我们在调研外部用户需求的过程中，经常会碰到的问题：每个企业用户的技术建设情况不同，业务复杂度也不一，很多传统企业已有的IT系统已运行了很多年，只是无法再支持日益增长的数据需求，他们在大数据技术体系的经验几乎空白，往往会觉得过于复杂和难以掌握，对落地成效心存疑虑。还有部分用户的业务在现有技术框架上（比如MPP）运行良好，出于对未来发展的前瞻性考虑，需要提前进行大数据的基础技术建设，这部分用户对于大数据未来的必要性是肯定的，但是会特别关心其适用的场景、业务覆盖度以及如何平滑地进行业务的迁移。

数据湖&Hadoop解决的是数据统一汇聚的问题，而统一元数据则是解决数据连接、资产、管理的问题，对于相当部分的用户而言，当前最大的痛点不是海量数据的存储，而是如何将散落到各个子数据系统的数据孤岛统一管控起来。因此通过构建一个逻辑层面的数据湖，实现统一的元数据+分散的物理存储，避免不必要的物理数据入仓（湖），从而将产品上层功能比如主题域构建、数据地图等等及早给用户使用才是解决问题的根本之道，逻辑数据湖方案，依然可以使用物理湖&Hadoop，同时提供通过虚拟表直连数据源的方案将其他类型的数据源也纳入平台的管控中，用户可以根据实际的需要选择适合的存储方案。

我们的构建方法论主要分为如下三个大的层面：

- 数据源支持类型：除了Hadoop（Hive）体系，MPP、RDMS、HTAP、KV、MQ等都需要支持，并且一视同仁，都可以作为具体逻辑数据湖具体对象的物理存储；
- 统一数据源 & 统一元数据：统一数据源要做的是规范每种数据源的登记注册，包括数据源URL格式、数据源Owner、唯一性校验、账号映射、联通性校验、支持的版本、特定的参数等；统一元数据，则是将数据源的技术（物理）元信息和业务元信息进行关联，提供统一的查询修改接口；
- 统一数据开发、治理和查询分析：这三个属于构建在统一元数据&数据源基础之上的应用层。统一的数据开发，包括不同物理数据源之间的交换、离线&实时开发、同源&跨源查询；统一的数据治理，则包括数据主题建设、权限管控、数据生命周期、资产地图等；统一查询分析，则是在完成数据主题建设、数据开发产出以后，提供同源&跨源的模型分析能力。



上图中最底层就是各种类型的数据源，通过统一元数据和统一数据源层，完成上层应用和资源层的解耦。对上层提供统一的计算、管理、应用的功能。

统一元数据

构建逻辑数据湖构建方法论的第一要务就是构建统一的元数据。无论是物理湖还是逻辑湖都需要通过元数据中心对湖内所有元信息进行管理。将构建统一元数据的关键点进行归纳，主要包括如下活动，分别为数据源信息的管理、元模型的设计、元信息的连接管理以及统一的接口调用。

统一数据源

除了统一元数据，数据源的统一也非常关键，直接决定了逻辑数据湖好不好用、能不能用。统一数据源的第一个关键点便是确定数据源注册登记的流程，需要针对不同种类的数据源明确其登记信息。其次是统一的账号管理，需要针对不同场景下，由于数据源认证方式不同、用户账号管理模式不同等特点，提供统一的账号管理方案。网易平台通过源系统账号映射功能，将平台账号同源系统账号进行绑定，绑定后的平台账号对源系统的库、表权限将和所绑定的账号一致，方便用户对账号的管理。

统一应用

在统一数据源和元数据的基础上，构建统一的应用从而实现各个数据源的算、管、用，其具体实现主要体现在模型设计、数据开发、自助分析、数据质量、任务运维、资产地图。

模型设计

在模型设计阶段支持对湖内、湖外数据提供统一的数仓建模，提供统一的表设计规范、表分层管理、模型建设水平评估等功能。

数据开发

数据开发支持用户直接在源系统上进行开发，比如在MySQL、Oracle、Greenplum、Vertica等数据源上进行SQL任务的编写。用户选择对应的数据源进行开发，开发完成后提交上线并进行调度和报警信息的配置。数据开发对不同种类数据源的支持，不仅保留了用户原有的开发习惯，也方便了后续任务的迁移。

自助分析

自助分析和数据开发类似，根据登记的数据源信息以及已经完成关联的账号，提供直连数据源的取数能力。此外，针对某些专门提供自助查数的业务系统（比如用户自研的脱敏平台）也可以作为逻辑数据源录入平台，从而实现建模取数。

资产地图

通过将各个来源的数据对象元信息进行串联和整合，提供快速查找数据的能力。通过解析逻辑数据源的表元信息，与主题、指标、标签等信息关联以后写入ES等检索系统，从而提供多维度的库表检索能力，包括表详情产出任务、血缘信息、变更DDL等内容的展示。

数据质量

为了保证产出数据的质量，平台对逻辑数据源提供质量稽核规则的保证，支持不同种类数据源的质量规则制定，也支持通过已有规则模版和标准推荐规则模版进行质量稽核任务的创建，从而对任务产出结果进行监控。

任务运维

对于已开发完成的逻辑数据源任务，支持任务运维中心统一管理，和湖内任务一样支持重跑、置成功、补数据、任务血缘，同时还支持基线运维的设置，帮助运维人员的提前感知异常任务，降低任务故障率，提升运维效率。

5.4 数据治理360

企业在数据治理的过程中，经常面对这样的问题：首先是数据的不规范管理，比如将外部表分区目录位置设置为表的目录，在进行表的生命周期管理时，如果选择删除目录，那么文件就会有被误删的风险。其次，数据团队人员疲于应付日益增长的业务需求，对数据治理缺乏动力，处于只开发不治理的状态。此外，人员的更替频繁造成系统内遗留大量历史数据，想要治理也无从下手。最后是治理效果的指标模糊，无法有效量化，对于哪些负责人下线了哪些数据、为项目节省了多少资源、具体节省了多少费用这些都缺少有效的评估数据。因此，在和领导汇报治理项目时往往说做了很多工作，但是领导缺无法感知，让领导产生你们好像都做了又好像都没做的感觉，导致最后的反馈并不高，从而造成数据治理人员的消极心态。

数据治理的工作往往是一个长期的、持续性的过程，这就需要我们在治理初期坚持以业务价值为导向，把数据治理的目的定位在有效地对数据资产进行管理上来，明确治理的范围，要让决策者能够看见和关注。其次，基于治理范围构建一套完整的度量体系，让一线用户能够看到问题、解决问题，从而将治理的结果进行量化，呈现治理的价值。最后，将治理过程中的方法论落地到产品功能上，通过短期业务线的宣传运营、培训交流和长期治理制度的建设，形成体系化的治理方案，实现治理的闭环。

明确治理范围

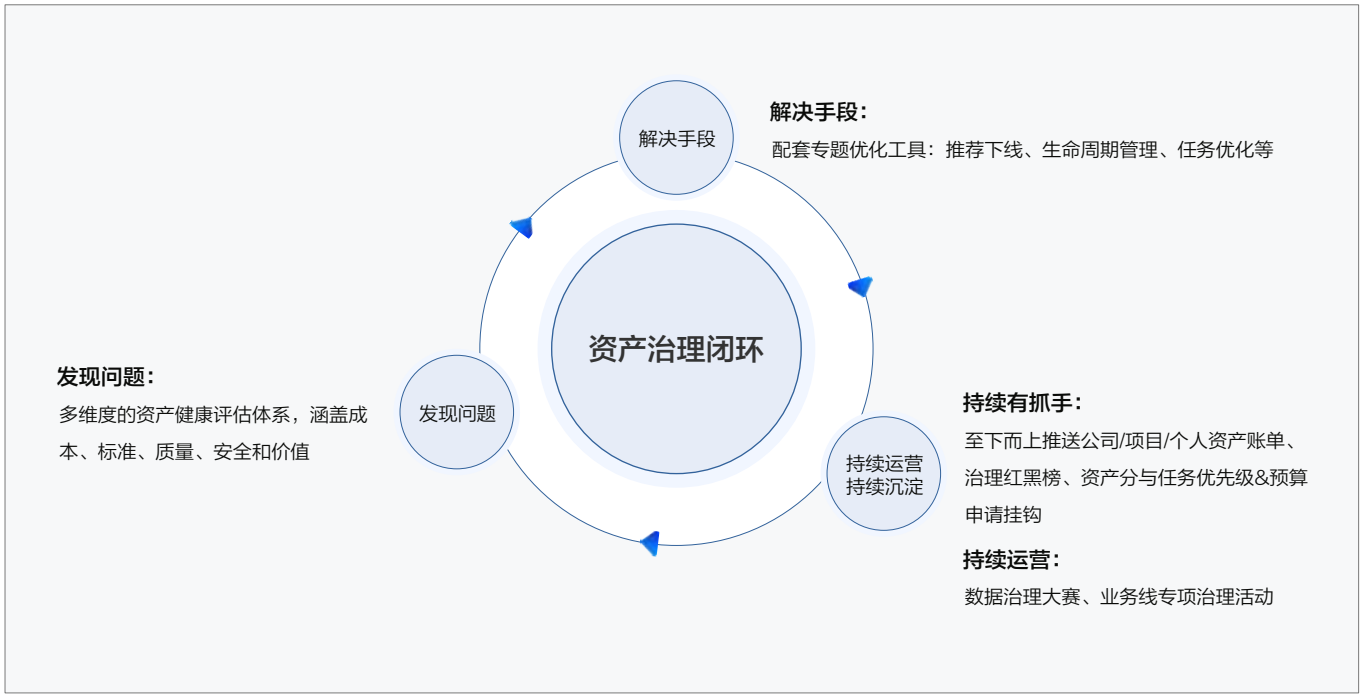
数据治理围绕数据的全生命周期进行展开，因此在数据生产阶段，需要对需求进行分析，明确业务口径，对数据进行规范采集、任务开发和监控运维；在数据消费阶段，涉及到快速的查找数据，对数据的分析和对数据质量的探查；在数据管理过程中，包含权限和成本管理等。整个流程涉及到成本、标准、质量、安全和价值，各个阶段都会面临对数据的治理工作。

量化数据价值

明确了数据治理范围后，需要对于数据治理价值的量化。基于数据的全生命周期，包含了成本、质量、安全、标准和价值五个方面，针对每个方面都要有可量化的指标项。对于成本，包括计算和存储成本的费用量化，对无用数据的下线治理等；对于价值，需要能够评估每个数据模型、数据报告和API的价值；对于质量，包含监控任务覆盖了多少稽核规则、涵盖了多少强弱规则；对于标准，需要对指标和模型进行规范化定义；对于安全，包含数据安全登记和数据权限的治理等工作。

体系化的数据治理

数据治理不是一个临时性的工作，从数据生命周期的全过程到治理体系的健康运行，需要一个长效的治理机制来保证，最后就是体系化的数据治理。体系化数据治理的第一个环节是发现问题，从成本、标准、质量、安全和价值五个方面，明确需要进行治理的内容；然后基于需要治理的内容配套专题的治理工具，比如对无用数据的推荐下线、对表生命周期的管理、对计算任务的优化等；最后，对治理持续运营、持续沉淀，包括推送整个项目、个人的资产账单，数据治理的红黑榜，并将资产健康分和个人的任务优先级或资源申请等挂钩，此外，企业内部可定期举办数据治理、业务线专项治理等活动来持续运营产品功能。整体通过发现问题-->解决手段-->持续运营和持续沉淀形成资产治理的闭环。



5.5 基于ROI的数据资产精细化管理

对于企业来说，一方面随着业务的发展，内部业务线的计算和存储达到瓶颈，但业务方很难判断，是应该继续扩容增加资源，还是对劣质数据进行治理来降低资源危机，在这个过程中，如何定义劣质数据，定义了劣质资源之后又要怎么对其进行治理，都是亟待确定和解决的问题；另一方面，数据本身的加工链路长，数据的加工处理没有统一的标准，整个团队内到底有哪些数据、数据的负责人是谁、这些数据是通过哪些任务产出的、这些数据有没有被有效的使用、数据的存在是否有意义，这些都是管理者比较关心的问题，但数据团队都很难回答。

因此针对上述问题，首先要将表和任务具体化到责任人，由责任人进行资产梳理，对于没有人认领的资产，比如没有负责人、负责人离职等情况，就由各个业务线指定专门的治理负责人进行专项治理；对于存储资源这个环节，首先要对无用数据进行规则定义，明确到底达到什么指标可以作为无用数据，比如近90天访问次数均为0等，确定好无用数据的规则后，将项目内无用数据进行识别，让业务方对无用数据进行二次确认后进行下线操作，同时提供下线列表或累计下线数据等，进行闭环分析；对于计算资源，会对离线任务、自助查询任务消耗的成本进行分析，包括任务的执行时长、预估消耗的费用、消耗的CU等内容，便于业务进行优化，也给任务的下线治理提供依据，成本分析会默认保留近半年的数据，在对任务进行优化后，也可以进行对比分析；最后，对于存储和计算优化后，治理效果要可量化和评估，比如确认下线的存储有多少，对任务优化后节约的资源有多少，便于管理者和治理负责人对治理成果进行评估。

我们将上述的治理方案结合内部的实践沉淀了一套基于ROI的数据资产精细化管理方法，通过基于Hadoop的元数据分析服务，精准计算出每个任务消耗了多少计算、存储资源，同时打通数据生产和消费的全链路的数据血缘，按照任务引用进行下游分摊，最终可测算出每个应用（数据报表、数据API）消耗了多少资源以及使用情况（PV/UV/重要程度），找到没有使用却消耗大量资源的应用。最后，通过采用“剥洋葱”式的数据下线方式，从上层数据应用开发逐层推动数据下线归档。

5.6 数据治理的持续闭环

数据治理是一个发现问题、解决问题、沉淀问题的过程，因此在数据治理的各个环节中都要尽可能的做到闭环，比如元数据的管理，完整的元数据管理流程除了包括正常的发布流程之外，还需包括相应的治理流程，实现流程上的闭环。同样，质量问题也需要闭环，完整的数据质量治理流程应该包含事前、事中、事后三个环节，即事前需求和规则定义、事中质量监控以及事后量化分析和问题追溯。最后是标准流程的闭环，通过标准的发现、管理、执行以及反馈，不断优化标准，保证数据的规范化管理及产出。

元数据治理的持续闭环

对于企业的湖外数据和湖内数据往往需要不同的治理方法，因此针对这两种数据我们给出了不同的数据治理流程。首先是针对湖外数据的治理流程，由数据治理团队发起治理需求，数据中台团队协助完成数据源的登记，之后便有数据治理团队完成元数据的采集、发布工作，期间业务部门完成元数据的补充，当数据资产的消费者在资产门户消费资产时发现资产存在问题则可提交工单申请治理，由治理团队指派工单给相关业务部门的治理专员完成资产的修正，最后重新发布上线，形成元数据治理的闭环。

针对湖内数据的治理流程，由数据开发团队根据数据标准完成模型的设计和开发，由数据产品团队对元数据进行注册和扫描并提交发布，数据治理部门完成元数据的审核并将其发布到资产门户，发布后的数据资产可提供给业务人员浏览和使用。如在使用过程中发现数据有问题，也可再次发起数据治理或者数据下线。

数据质量治理的持续闭环

数据质量治理闭环的关键点便是要做到数据质量问题的问责，而问责的前提是问题的定责。定责的基本原则是：谁生产，谁负责。数据是从谁那里出来的，谁负责处理数据质量问题。因此，完整的质量治理流程包括事前需求和规则定义、事中质量监控以及事后量化分析和问题追溯三个环节。

事前需求和规则定义

DAMA国际数据管理协会定义了数据质量维度，包括准确性、完整性、一致性、合理性、唯一性、及时性等内容。质量规则除了通用的表级、字段级规则模版外，还要能够根据不同行业的要求灵活的制定质量监控规则。此外，对与企业内部应该具有一套统一的、标准的质量稽核规则，因此还需要和数据标准进行打通。

事中质量监控

将已配置好的质量稽核任务与开发任务进行绑定，确保任务在产出数据过程中能够得到实时有效的监控。对于质量监控的任务，质量异常和质量检测失败都能够灵活的配置告警，支持邮件、短信等多种通知方式。质量问题发生后通过任务中的强弱规则设置，实现对下游任务的及时阻断，达到亡羊补牢的效果。

事后量化分析和问题追溯

质量问题发生后要对已发生的问题进行分析和评估，对质量稽核任务的执行趋势进行分析快速定位结构异常、运行失败的原因。质量问题追溯方面通过定期的质量报告了解数据质量情况，对于已发生的问题落实到责任人令其改正，并进行绩效关联，加强人员对质量问题的重视。

标准的持续闭环

标准的制定也是一个持续闭环的过程，已发布的标准需要根据国家的政策、行业的发展、公司业务的调整不断的进行优化，因此整个标准的流程包括标准的发现、管理、执行、反馈及优化。

标准的发现

标准的发现主要是对已有标准，比如标准文件，或者是无标准的场景进行标准注册，作为后续词根、标准字典、数据元的构建依据，因此，需要事先完成标准文件的梳理。

标准的管理

标准管理环节主要是对核心标准内容生命周期的管理，主要包括元数据标准、数据标准以及标准流程的制定。在这个环节，根据前期调研情况完成词根、字典、数据元、数据项分类等内容的制定。

标准的执行

标准执行环节可将制定完成的标准落实的数据设计开发过程中的各个环节，统一规范。

标准的反馈

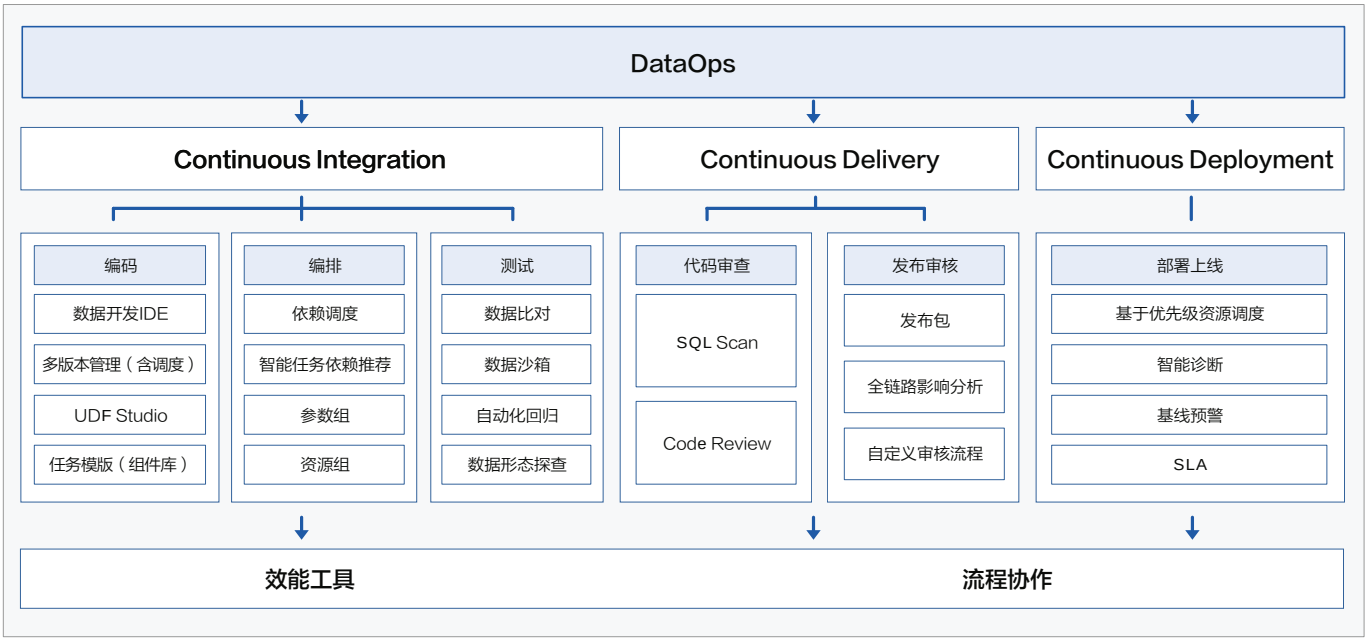
标准反馈环节是对标准资产的盘点以及落标的检查，通过标准资产情况、标准覆盖情况、标准使用情况以及流程操作的统计。

标准的优化

标准优化环节经过标准内容的不断优化，企业对标准进行沉淀形成行业的通用标准，从而实现标准的闭环。

5.7 基于DataOps开发底座

基于DataOps的数据开发方法，是将数据看作是软件，用软件工程的方法去解决数据开发过程中的效率和质量问题，其包括可持续集成、可持续交付、可持续部署三块内容。



六大过程

可持续集成包含了编码、编排以及测试三个过程：

- 编码阶段，主要提供丰富算子的、高效的数据开发环境，同时还支持多版本的管理功能，提供统一的UDF开发和管理工具。
- 编排阶段，在数据的开发过程中支持不同任务间的依赖调度，同时结合数据血缘能够自动识别任务间的依赖关系，减少构建任务依赖的复杂度。此外，引入参数组和资源组功能，能够对公共参数和jar资源进行统一管理。
- 测试阶段，主要包括数据比对、形态探查、数据沙箱、自动化回归等功能。通过数据比对、形态探查、自动化回归提供不同场景下数据测试方法；通过数据沙箱功能解决了测试数据准备的问题，通过沙箱功能可以确保在测试集群直接运行任务，读取生产环境脱敏数据进行测试，同时又只能写入测试环境，避免了污染生产环境的问题，既保证了两个环境的隔离，又解决了跨环境数据测试的难题。

可持续交付包含了代码审查和发布审核两个过程：

- 代码审查，任务发布之前，需要对代码进行审查，提供了系统和人工两种方式，分别是SQL Scan和CodeReview。
- 发布审核，任务发布上线前需要经过发布审核流程，提高任务发布规范性的同时还能再次对发布过程进行审核，在这里包括任务发布包的审核、下游发布影响分析的检测以及根据数据资产等级构建的发布审核流程。

可持续部署包含了部署上线过程：

- 部署上线，在一些金融级的数据平台架构中，除了有开发集群和生产集群，还有测试集群和预发布集群，因此不同环境间任务的发布上线需要基于不同的策略。任务部署上线后还需要提供合适的运维工具，帮助运维人员快速定位异常问题，提高运维效率。

这六个过程共同构建了一个数据发布流水线，在保障质量的前提下，实现了数据开发的敏捷交付。

第六章 数据治理 2.0 最佳落地实践

6.1 某证券公司

背景

国家层面

在当前的数字经济时代，数据已成为经济发展的关键生产要素和核心引擎，数据安全作为国家安全的重要组成部分。近年来，国家陆续颁布了网络安全法、数据安全法等法律法规，完善和数据相关的法律法规体系。2021年9月1日，《中华人民共和国数据安全法》正式施行。该部法律确立了数据分类分级管理，建立了数据安全风险评估、监测预警、应急处置，数据安全审查等基本制度，并规定了相关主体的数据安全保护义务，其中明确了各行各业的主管部门承担本行业、本领域数据安全监管职责。

行业层面

2022年3月银保监会严肃查处了一批监管标准化数据（EAST）数据质量领域违法违规案件，对政策性银行、国有大型银行、股份制银行等共21加银行机构依法作出行政处罚决定，处罚金额合计8760万元。22年1月，银保监会下发通知，要求金融机构切实推进数据治理，提升数据质量和数据专业性，进一步增强数据规范性。此外，对于银行机构在监管数据质量和数据报送中存在的违法违规行为，监管机构不断加大处罚与整治力度，对金融机构数据质量及业务合规性的要求不断提升，又无疑是对该行业提出了更高的数据治理要求。

企业层面

由于证券行业数字化进程相对较晚，系统建设滞后，在非标准化的场外业务数据很多以手工形式存在，数据无法落地或系统化。信息孤岛现象严重，业务系统繁多，不同系统之间数据不一致，系统和数据建设缺乏规划。同时，数据标准的建设的缺失导致致数据口径的不一致、指标口径的不一致、数据质量稽核的规则不一致，容易在对监管机构进行报送时质量通过率不高而引来罚款。除此之外，数据平台未整合、数据无法有效共享、数据资产缺少持续运营、数据流通和安全难以兼顾等问题都是企业常见的数据问题。因此，对于证券公司来说，数据治理势在必行。某证券公司结合自身的实际情况，制定如下数据治理目标：

- 第一阶段，在业务发展同时制定数据管理制度与流程；文化上通过和提高业务部门参与度，鼓励全员参与用数、治数，每个部门有自己的数据治理专员，成立数据治理委员会；定期开展数据治理相关培训和宣贯；
- 第二阶段，依托产品规范流程，纳管外部供应商数据，依靠平台释放数据价值，借助数据地图数据目录让业务部门共享数据建设成果。

数据治理的体系

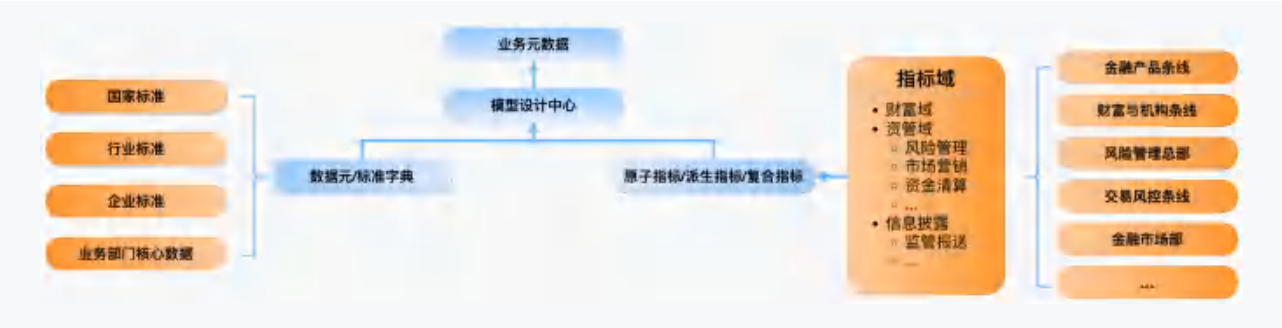
网易根据该证券公司数据治理的需求场景、结合内部数据治理产品工具，将流程建立在工具的基础上，制度建立在流程的基础上，管理建立在制度的基础上，形成全链路的数据治理体系。在实际实施过程中，我们围绕产品工具、流程、制度以及管理展开数据治理。

产品工具方面

首先是产品工具层面，将整个数据治理流程贯穿各个子产品，沉淀一套全链路的数据治理工具体系。如产品架构图所示，展示了各子产品所属位置及关系。数据标准管理会涉及元数据管理、指标系统管理、模型设计管理、数据质量管理、数据安全管理等。各子产品在使用和运行过程中会生成使用数据和监控分析数据，通过这些数据可体现当前待治理数据的现状，再通过资产健康评估体系和治理工作台来不断提升数据治理效果等。

1.规范建模

数据标准是数据在内外部使用和交换时确保其一致性和准确性的规范性约束，在保证数据模型的规范化、提高数据资产化水平、提升数据质量等方面都发挥着不可或缺的作用。数据标准的重要性不可纸上谈兵，其价值需要真正的落实到数据治理流程当中去。因此，数据标准应当是一套由规范要求、流程制度、技术工具共同组成的体系，良好的工具平台才能更好的践行规范要求与流程制度。



首先，该证券公司结合国家标准、行业标准、自身企业标准以及业务部门核心数据打造一套贴合自身业务发展、数据治理、数据开发的数据标准系统。例如，证券行业主要参考的标准是证券期货行业数据模型（SDOM），客户将SDOM模型导入到数据标准管理平台，完成了数据标准的主要载体数据元及数据字典的积累。另外，客户已积攒2000+词根，也均已通过数据标准管理平台词根管理模型进行导入统一管理。

其次，通过对各条业务线的分析，梳理出原子指标、派生指标以及复合指标。通过指标系统对指标进行管理，助力数据规范定义、助力数据模型规范设计等功能，解决指标定义不一致、计算口径不一致、数据来源不一致带来的指标数据可信度低、严重影响分析决策的问题。另外，提供指标的治理流程，协助用户管理好指标。

最后，在标准及指标的规范下构建模型，沉淀业务元数据。

2.数据质量治理

证券行业对于数据质量要求极高，要求数据必须要遵从外部的管理法规、行业标准和内部一些定义的业务规则。因此需要将数据质量与元数据管理、数据标准、治理流程紧密结合，丰富完善企业内部的数据管理和数据治理体系，围绕质量的事前需求和规则定义、事中质量监控、事后量化分析和问题追溯三个阶段进行展开。

前期的需求和规范定义阶段，该证券公司的数据治理团队按照行业规范的完整性、有效性、及时性、一致性、准确性、唯一性六大质量检核维度，梳理了业务侧和技术侧对于数据的质量要求和规范。比如，在客户信息中，客户名称是必填项、手机号码的取值只能是“1”为第一位数字且长度为11，否则不是一个有效的号码等等。

对于庞杂的数据质量要求，业务团队将其抽象成质量规则，分为多表级、表级和字段级规则，对表和字段进行质量监控。并且根据不同的业务场景采取不同的解决方案：

- 针对通用的、复用性强的评价指标，例如客户信息中，客户名称是必填项，不能出现为空的情况；居民身份证号码长度为18位，如果出现“证件类型”为“居民身份证”但“证件号码”非18位的数据，则判定为不满足长度有效性；证件类型+证件号码+姓名三者相同，则代表是同一个客户，则该客户在客户信息表中应该只有一个客户编码等等，可以借助数据质量中心，将这些质量稽核规则抽象为不同的规则模板；
- 通过将质量和元数据及数据标准打通，在元数据注册时选择推荐的规则模板，对于绑定了元数据的字段，质量中心可以实现规则推荐。在数据标准填写数据元信息时，把数据元的质量要求直接配置完成。如“证券市值”的精度取值范围应等于2，如果出现精度大于2的情况，则不满足精度取值范围有效性，不是一个有效的“证券市值”。对于绑定了数据标准的字段，数据质量中心可以自动根据标准定义的质量要求“证券市值等于2”生成监控规则；
- 对于规则模板和数据标准无法覆盖到的个例场景，可通过写自定义SQL实现质量检核。

定义好了质量规则，数据治理团队就需要通过质量监控任务持续地测量和监控数据质量。一个质量监控任务由监控对象、质量规则、执行和调度配置组成。将质量监控任务挂载在离线开发任务节点，还能及时阻断下游任务，避免脏数据造成影响。对于每一次质量监控的执行结果，都可以进行及时通知和报警。质量监控任务支持周期调度，满足数据治理团队定期进行质量盘点的需求。公司通过查看每一次监控任务执行实例中的运行结果，查看表和字段规则的采样结果、期望范围和异常数据的明细，可以清晰定位出数据中的质量问题，并及时纠正。

数据质量管理的第三个阶段就是对质量监控结果进行分析和量化，查找出现质量问题的数据链环节、定位数据问题、实行问责机制。为了更好的衡量质量的好坏，衡量的范围从一条质量规则出发，扩展到表的质量、库的质量、主题域/表分层的质量、表负责人所负责表的整体质量等。细化来看，又可以分别去看唯一性、完整性等每一个校验维度的质量。

该公司的管理者和治理团队，通过质量大屏整体查看关注范围内的整体质量分、监控配置情况、质量规则异常情况等。从而细化每张表，定期可以去查看每张表的质量报告，掌握这些表在最近的执行异常明细，质量分的趋势等信息。此外，还引入了数据质量的问责制度，保障数据生产者对各自的数据质量负责。治理团队完成需求和规则的定义，管理者就要通过事后的分析和追溯责任到人。企业内部以质量分为抓手，将数据质量目标纳入绩效评估中，表质量异常次数过多、质量问

3.数据安全治理

数据安全就是保障数据全生命周期中的一切操作符合国家和公司的安全规定，贯穿数据治理始终。某证券公司作为金融行业，业务数据中存在大量涉密的个人身份、金额等敏感信息，迫切需要建立完善数据安全管理制度和技术保护机制。近年来数据安全事件有增无减，无论是国家法律法规要求还是企业自身需求，都驱动其加强数据安全体系的建设。

该证券公司存在大量身份证手机号等敏感的数据，数据分级分类是保障数据安全的基石。参照国家证券行业的分级分类标准，内部对数据进行了分级分类管理。根据实际的业务场景进行了数据分类的划分，并按照法律要求、价值、泄露和修改的敏感性进行了数据分级。公司内部定义了身份证、手机号、邮箱号等常用敏感类型。数据治理团队会定期发起数据识别任务，进行敏感数据的自动发现。基于敏感类型，可以自动推荐相应的安全等级。将安全等级作为判定治理过程中各种流程审批链路及判定风险行为的依据。

对于数据安全的最后一道保障，就是平台操作和敏感访问的监察审计，其核心是什么人在什么时间做了什么事，细粒度的审计可以在让员工在对敏感数据操作时多一份威慑，一定程度上防止事故的发生。同时，在数据泄露等事件发生时，进行及时的告警和风险提示，事后可以进行源头追溯。对于平台操作，管理员定期审计用户数据开发和管理内的所有操作日志，并配置风险行为审计和告警规则。对于敏感数据，从敏感类型和安全等级两个维度配置风险行为审计和告警规则。

4.元数据资产治理

在元数据管理中将业务元数据、技术元数据、管理元数据补充完整，然后根据元数据的治理发布流程将元数据发布上线。同时配合数据资产中心的资产健康诊断以及基于ROI的数据资产精细化管理，对数据资产的健康情况和使用情况进行实时的观察，帮助该公司识别并了解真正有价值的资产。

流程方面

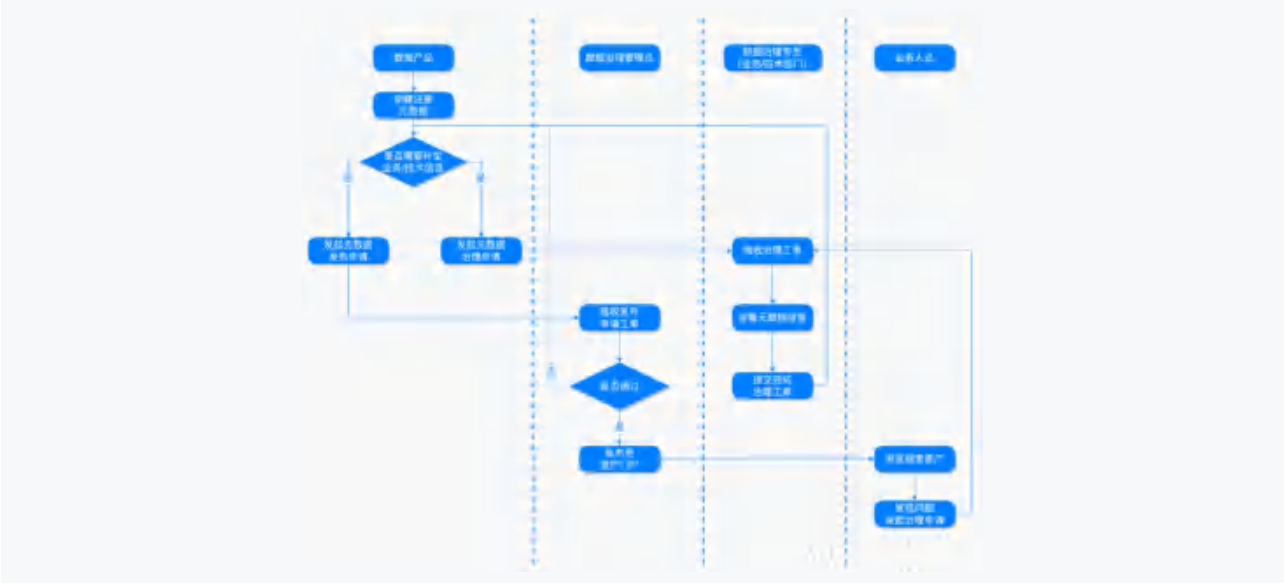
在流程方面，根据某证券公司的业务情况，我们制定了湖内数据和湖外数据的治理流程。

1.湖外数据治理流程

湖外的数据可能来源于业务数据库例如MySQL、Oracle中。针对这类数据，首先数据治理管理员根据治理需求向IT部门发起登记数据源操作，数据源可来源不同的业务源系统。登记数据源后就可进行元数据采集、元数据注册，元数据注册后根据数据的完善度来决定是否需要治理，最终将数据发布为资产，供业务人员浏览和使用。

2.湖内数据治理流程

对于湖内数据主要包括元数据的注册、治理、审批、发布、使用这几个步骤。首先需要进行元数据的注册，注册后经过业务治理专员或者技术治理专员不断完善业务/技术元数据信息，由申请人提交发布申请，最终由数据治理管理员审核发布。发布后的数据资产可提供给业务人员浏览和使用。如果在使用过程中发现有数据问题，也可再次发起数据治理或者数据下线。



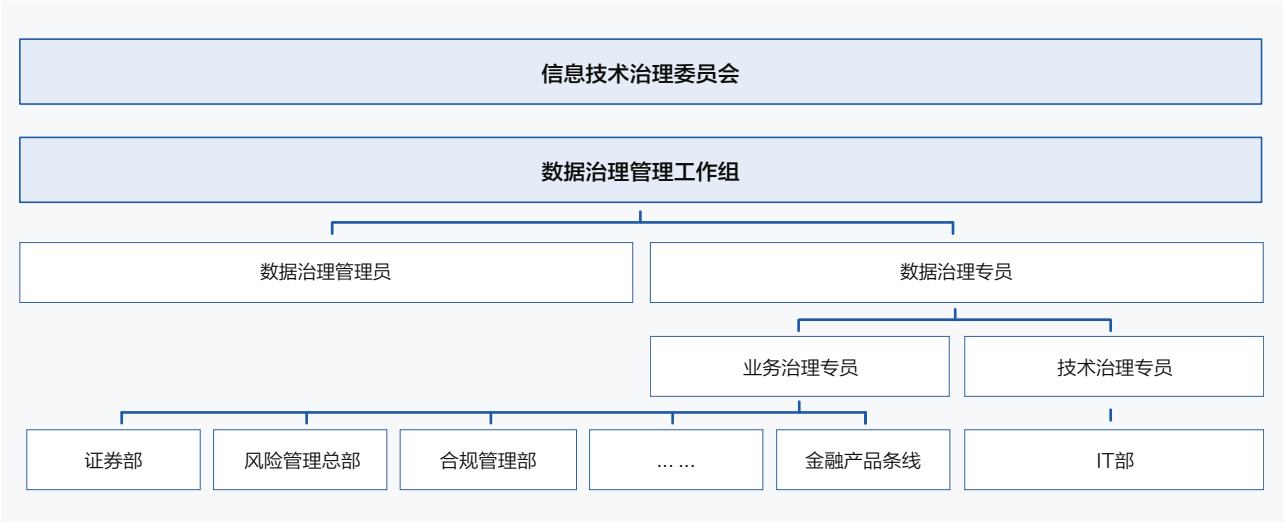
制度方面

在制度方面，基于数据政策的原则性要求，结合该公司其内部组织和业务的特点制定数据治理相关制度，从而作为行为的基本准则为后续各角色的职责问责建立依据。例如数据的确权制度，数据资产发布之前需要交由相关的业务部门进行审核和确认，确认资产的业务口径以及权责归属，当发现资产出现问题时，便由确权的部门进行治理确保资产的准确性。此外，该证券公司还制定了数据源的管理制度，在数据源的使用上要求按项目粒度进行隔离，针对于不同的角色开放不同数量的数据源，同时对数据源的读写进行了使用场景的规定，极大程度上保证了数据源使用的规范性和数据安全性。除了上述的制度外，公司内部还制定的数据治理制度还包括了如指标治理规范制度、质量问题问责制度等，对整个数据资产管理各项活动中所应当遵循的原则、要求和规范进行了详细的描述，确保了数据治理的管理层对准备开展或正在开展的数据治理各项职能活动进行有效控制。

管理方面

在管理层面，我们建立了专门的部门来负责数据治理工作。完善的组织架构、权责分担机制，能够让数据治理工作稳定持久的推进下去。数据治理工作中有2个重要的角色分别是：

- 数据治理管理员：作为集团数据治理工作的管理人员，对集团所有待治理数据负责，推进和协调各部门业务数据的治理，并且对治理的数据能否发布为最终资产有决策权限。
- 数据治理专员：作为集团数据治理工作的一线参与人员，可对本部门数据进行治理。数据治理专员可按照部门属性划分为两类，一类是IT技术部门的技术治理专员，负责技术相关信息的数据治理；一类是各业务部门的业务治理专员，负责业务相关信息的数据治理。技术治理专员和业务治理专员虽然治理内容略有不同，但是治理流程基本类似，都是先接收治理工单，再填写治理内容，最后提交治理工单。技术治理专员和业务治理专员之间没有固定的先后治理顺序，根据实际数据的质量等情况来决定需要哪些治理角色参与。数据治理管理员根据数据的质量情况，若发现业务信息不完善时，可以向对应业务部门的业务治理专员下发治理工单；若发现技术信息不完善时，可以向IT技术部门的技术治理专员下发治理工单，最终数据治理完成后即可发布到资产目录中，供业务人员浏览使用。



某证券公司数据治理实施成效

该证券公司从组织架构、业务流程、产品工具入手通过数据治理专项建设在一阶段初步完成了元数据、数据标准、数据质量、数据安全的治理工作，具体如下：

- 元数据管理：多部门共同建设，携手推进数据治理工作。截止一阶段，完成8个核心系统30,000余张表的全量元数据采集，创建采集任务58项、采集实例163项。当前正在推进对数据资产的质量提升和完善，新增元数据注册47项，累计注册2,814项；
- 数据标准管理：基础数据标准179个，指标数据72个，搭建统一的公共字典合计7个，词根新增159个，制定两套数据标准分类方案，16项主目录细分数；完成金融市场部、风险管理总部等18个部门或单位的CISP报表指标口径填报优化；
- 数据质量管理：为机构CRM上游数据源进行初步质量评估，已累计沉淀400余项数据质量规则；
- 数据安全管理：中台搭建基于数据分级分类的数据服务审批流进度到达80%；敏感数据识别、脱敏规则配置、脱敏任务执行、脱敏白名单等模块投入使用，形成数据过敏规范初稿。

6.2 某电信运营商

背景

管理方面

在数字化浪潮的冲击下，通信行业面临快速演变的内外环境，竞争形势不断恶化，各大运营商均在积极探索数字化转型之路，数据中台的搭建成了转型之路上不可或缺的一步。但传统的烟囱式I T架构又很大程度上制约了运营商行业数字化转型的进程，主要呈现如下:

- 数据建设分散，历史积累尾大不掉。烟囱式的IT架构导致团队内部数据治理分散，历史沉积的大量数据建设成果尾大不掉，成为转型过程中的历史包袱，而若想推倒重来，必然会对当前的业务产生巨大影响，成本无法估计；
- 供应商分散导致各产品间标准规格不统一，数据治理思路各异。运营商行业往往会采购不同厂家的软件工具，导致各厂家工具之间使用脱节，开发、治理、调度各自独立运营，无法形成有效的关联和管理，导致整体作业维护性差、安全管控风险高，平台维护和作业优化缺乏手段；
- 传统数据开发、分析工具效率不高，制约数据应用。对于数据开发人员因为缺少数据可视化的开发平台，Flink、Spark等计算组件的使用门槛又较高，导致数据开发、分析效率低下，制约了数据应用；
- 数据质量运营缺少流程化的支持。数据标准、质量稽核和数据开发脱节，缺少数据建模、开发、测试、发布、稽核的一体化开发管理平台，相关要求停留在规范要求上，无法融入数据生产的流程中，不能有效落地执行和监督，监管力度的缺失则会造成大量的脏数据，最后导致数据的不可用；
- 数据产品建设混乱，数据价值很难彰显。因为数据孤岛现象的存在，加上各区域各部门之间又独立建设，导致数据建设结果良莠不齐，数据价值难以彰显。基于上述痛点，用户希望可以通过数据采集、建模、开发、治理、分析一体化能力，为数据标准化能力提供平台基础，以支持未来3-5年内的数据开发服务能力，提升数据交付质量和效率。

数据中台治理实践

网易数帆经过多年的实践和持续的打磨，在数据生产力DataOps、DataFusion和DataProduct三大方法论的指引下，构建了面向数据生产力的产品技术体系，基于这套体系打造了数据开发及管理平台。针对该运营商的痛点，为其提出了切实可行的解决方案。

构建一站式基于逻辑数据湖的开发运营模式

考虑到业务方数据系统架构复杂，技术栈不统一，历史数据迁移成本较高，为最大限度保护客户历史数据资产，减少数据迁移带来的损耗。逻辑数据湖构建一个“物理分散、逻辑统一”的数据湖体系，用技术把一个个数据孤岛打通，避免了不必要的物理数据入仓(湖)，从而将产品上层功能比如主题域构建、数据地图等等功能及早提供给用户使用，在持续交付中不断纳管历史数据成果。

基于逻辑数据湖的架构，该运营商以中台结合数据开发推动数据作业五统一：统一入湖、统一开发、统一调度、统一治理、统一开放，提升数据交付效率和共享能力。

- 统一入湖：打破部门壁垒，企业数据统一汇聚到大数据平台，通过中台的权限隔离，实现数据分域管理；
- 统一开发：关闭客户端直联数据库的开发工具，数据地图查询、业务数据查询、数据建模、脚本开发都通过中台工具实现，建立数据操作详细审计能力；
- 统一调度：下线多套主要调度平台，所有调度统一至任务调度中心，实现7万个作业的统一调度，按项目隔离权限；
- 统一治理：以谁开发谁负责的原则，用中台工具推动元数据维护、质量稽核要求落地，提升数据质量；
- 统一开放：以BI Saas化应用和API服务为主要模式建立数据统一出口，实现数据不出平台。

同时结合基于DataOps开发底座，对于客户原有系统支持直接在源系统上进行开发，比如各类的SQL任务，用户选择对应的数据源，调度执行节点根据数据源相关的信息、驱动配置等直接连上数据系统执行任务，支持用户保留原有的开发习惯，也方便任务的迁移。

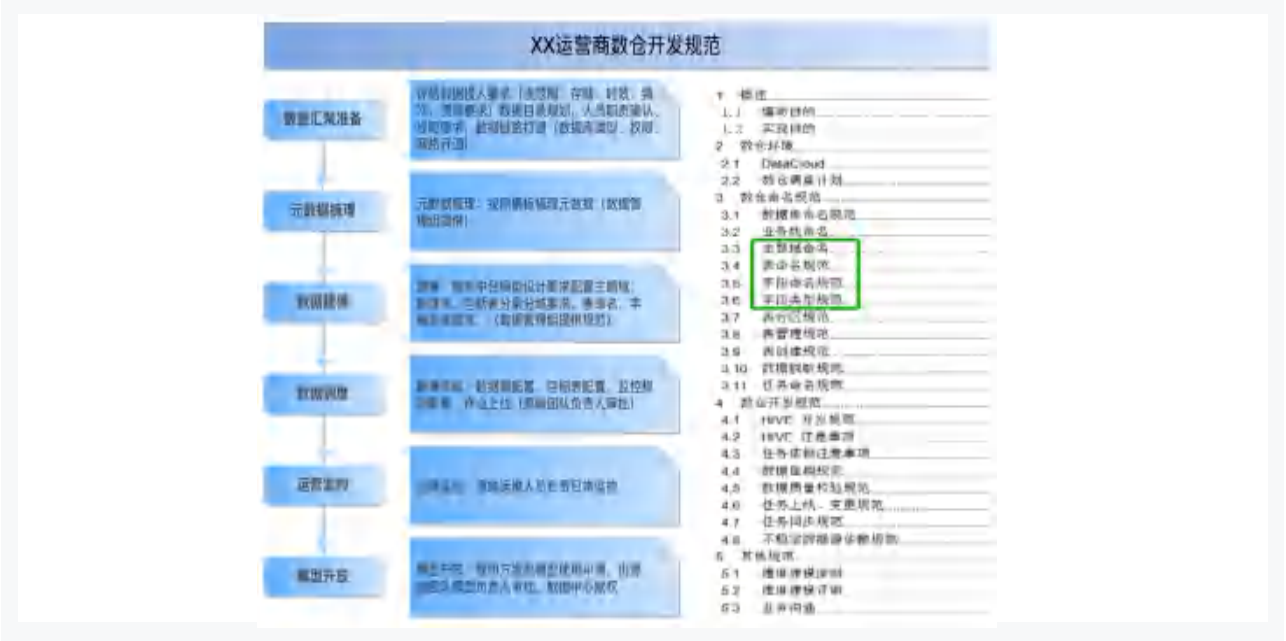
而对于跨源的SQL任务，依托Spark、Flink等计算框架的catalog-manger框架来支持。开发平台会根据任务执行的模式、运行环境，自动关联同catalog下对应的数据源实例和执行账号，动态注册关联的catalog-plugin和相关参数。同时数据稽查、数据探侧等依托于开发框架也同样支持相关数据源。

构建数据资产标准化管理

结合网易数帆的数据治理2.0方法论，该运营商根据自身的业务特点通过构建统一企业数据目录、统一元数据标准、统一核心主数据以及统一质量管理标准，使电信内部数据资产做到可见、可懂、可用、可运营。

- 统一企业数据目录：数据目录统一规划，贴源层按部门系统、数仓层按主题域、应用层按专题；
- 统一元数据标准：表命名规范、字段命名规范、字段名称、业务定义、业务解释、技术定义、枚举值说明等等；
- 统一核心主数据：客户、产品、渠道、网格、员工、基站、终端等各业务实体的ID在各系统要保持一致；
- 统一质量管理标准：接口层要配置记录数据稽核，数仓层要配置核心数据一致性稽核，应用层要配置指标波动阈值稽核。

此外，根据数仓开发流程，该运营商从数据汇聚准备、元数据梳理、数据建模、数据调度、运营监控、模型开放六个环节明确开发规范，为企业内部的应用团队、业务团队、中台团队提供开发依据。



数据中台实施成效

通过网易数帆数据中台将开发运营进行一体化建设，由数据中台统一为仓库、经分、网络集群提供数据采集、建模、开发、调度、治理等一体化能力。在生产过程中对于程序上下线、建表等操作实现在线化、流程化操作，一方面减少人工提升效率，一方面完善数据管控的过程。

此外，将数据开发与数据治理有机结合起来，既是对开发过程的管控，也是保障数据质量的有效方法，通过一体化的平台，核心数据的数据标准覆盖率从23%提升到57%，在数据建模过程中就完成了数据标准的落标过程，同时解决了长久以来，数据质量稽核规则覆盖率低，规则不一致的问题，核心表稽核规则覆盖率从17%提升到90%。通过逻辑数据湖功能，在0迁移成本的基础上，5G DPI、DNS访问数据、客户触点、广电EPG服务、电渠业务系统等重要系统的数据完成入湖，累计汇聚业务方86个系统数据。通过统一的数据标准制定和发布，结合制度约束、系统控制等手段，辅以标准自动推荐、自动纠错等标准化服务，实现企业级大数据平台数据的完整性、有效性、一致性、规范性、开放性和共享性管理，提高企业级大数据平台数据治理水平。通过基于ROI的数据资产沉淀，帮助企业了解自身数据资产结构，沉淀高价值数据作为资产，消灭高成本低价值的数据。通过面向数据中台的数据建模，帮助企业内部的内部需求交付速度提升一倍，年平均查询速度为21秒。到目前为止数据中台上日均活跃用户数超过300人，完成迁移上线作业数8000多个，数据质量稽核数超过600个，自助分析累计近30万次，数据开发运营一体化标准流程100%覆盖新建任务。此外，累计梳理数据模型数3000多套并录入中台，形成企业级数据资产目录，实现数据资产的快速查找使用。

6.3 某物流公司

背景

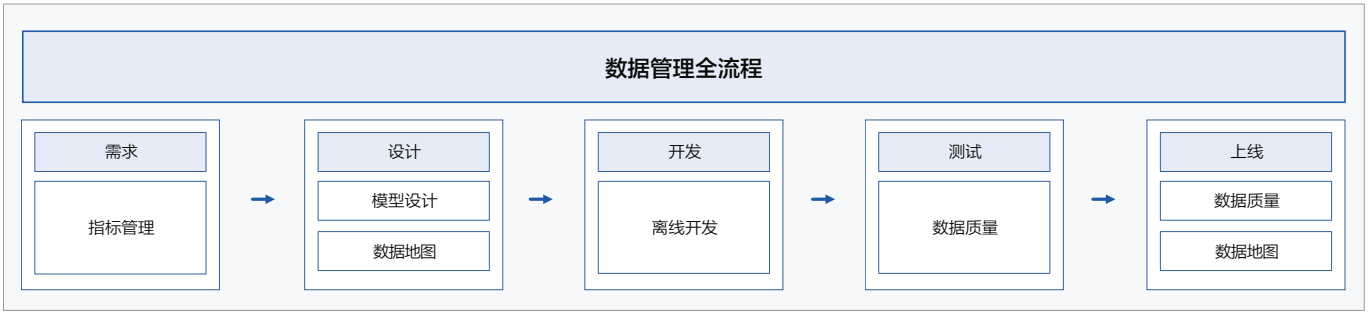
管理方面

某物流公司规模大、业务场景复杂，集团内部对数据非常重视，尤其是在数据质量、数据标准、元数据管理、数据入湖等方面。因此，早在2013年集团内部便开始了数据治理的相关工作，在这期间也经历了数据应用上的问题及痛点，主要体现在流程不清晰、源数据不准确、数据口径不一致、数据不共享、数据无责任人等方面，在数据生产的各个阶段都有涉及。

- 需求阶段：集团发现业务部门经常不清楚系统中已经存在哪些指标，也不清楚去哪里找数，只能根据自身业务需要提出指标需求。其次，各部门之间数据打架，互相之间口径不一致，导致管理低效，高层无法准确决策，例如，在业务上有一个空运指标叫做“时效件收入”，其包含部分“国际收入”，但是因为业务人员不知道线上已存在国际收入，又设计了同样业务含义的指标，造成口径冲突。
- 设计阶段：指标开发人员不清楚系统中现有哪些应用或数仓模型、不清楚模型的计算口径、不敢用，只能竖井式开发，导致效率不高且造成数据口径冲突。
- 开发及上线阶段：存在源系统数据质量不高导致数据分析错误，增加补丁程序，反复刷数及数据验证，用数效率低，例如，同一个车牌号，在资产管理系统和业务系统都存在，既是自有车，也是外请车，导致成本取数重复。其次，源系统数据变更但未充分进行影响分析，导致指标数据错误，通过刷数进行修复，用数效率低。

数据中台实施方案

针对数据管理全流程，通过网易数帆产品可以实现从需求到上线的全流程进行管理，指标管理、模型设计、元数据查询、血缘分析及数据质量监控，提升整体研发效率，并结合数据质量监控，实现全链路数据监控进行实时播报预警，全方位保障数据服务。



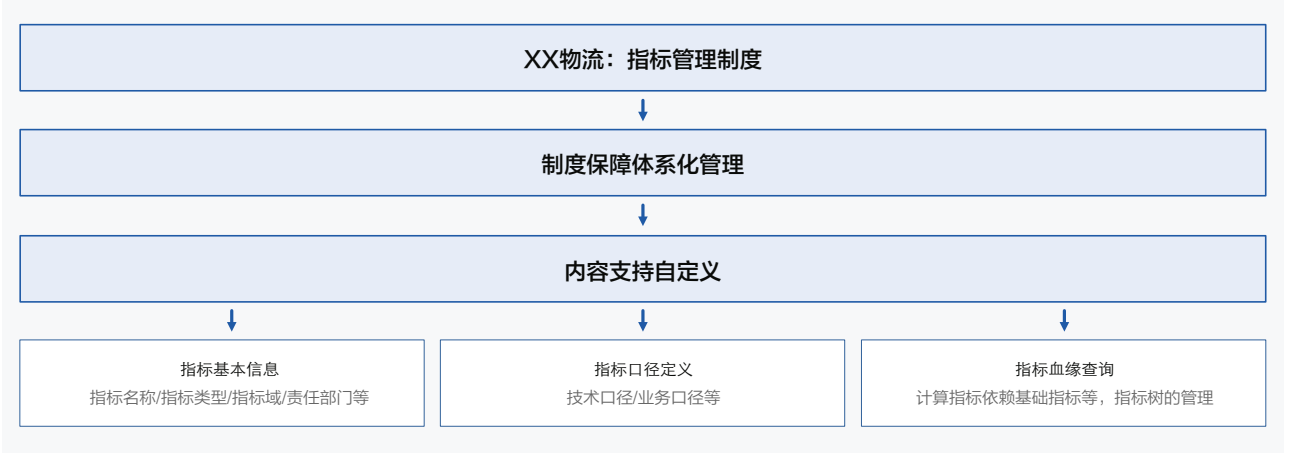
- 需求阶段：指标管理系统作为集团指标库，提供指标的统一管理、统一定义、统一口径。
- 设计阶段：通过模型设计中心构建数仓，通过规范化、统一化模型建设、数据复用，从而提升研发效率。数据地图作为数据导航功能，提供数据血缘、数据资产检索、预览，快速定位用户需要的数据。
- 开发阶段：基于DataOps的数据底座，及提交上线的CI/CD流程控制能力，包括代码扫描、形态探查、数据比对、节点测试检测、产出影响检测、质量监控规则配置检测等。
- 测试阶段：数据质量提供数据测试的能力，数据比对、数据形态探查。
- 上线阶段：通过数据质量提供全链路结果可量化的质量监控、质量大屏、质量报告，结合数据地图的数据血缘功能快速进行问题定位。

数据标准构建

在数据入湖时，推动源端系统进行模型的数据标准的维护，指导数据传输规范，模型设计规范、数据开发规范、数据质量规范等，确认数据传输和使用过程中有数据标准的依据。通过数据标准模块进行数据标准制定与数据模型进行绑定，基于模型的数据标准可以生成对应的数据质量规则，同时通过数据地图可以查询模型的数据标准情况，辅助数据需求设计、开发及质量监控的活动。

指标系统构建

通过指标中心结合指标管理制度实现指标增、删、改、查线上化操作，并实现指标与模型的绑定关系，打通指标共享壁垒，提高指标复用度。



在指标中心结合指标管理制度的管理需求，自定义设计指标录入模板，由业务人员在需求阶段进行指标信息录入，提供给ETL开发人员进行指标开发。指标上线后，可通过数据地图进行指标信息的查看，同时在页面前端指标数据展示时调用指标定义查询接口进行指标口径的查询。通过指标基本信息的查看可以了解到指标名称、指标类型、指标域、责任部门等；通过指标口径定义可以了解到指标的技术口径、业务口径；通过指标血缘查询可以了解到指标间的依赖关系。

模型中心数仓构建

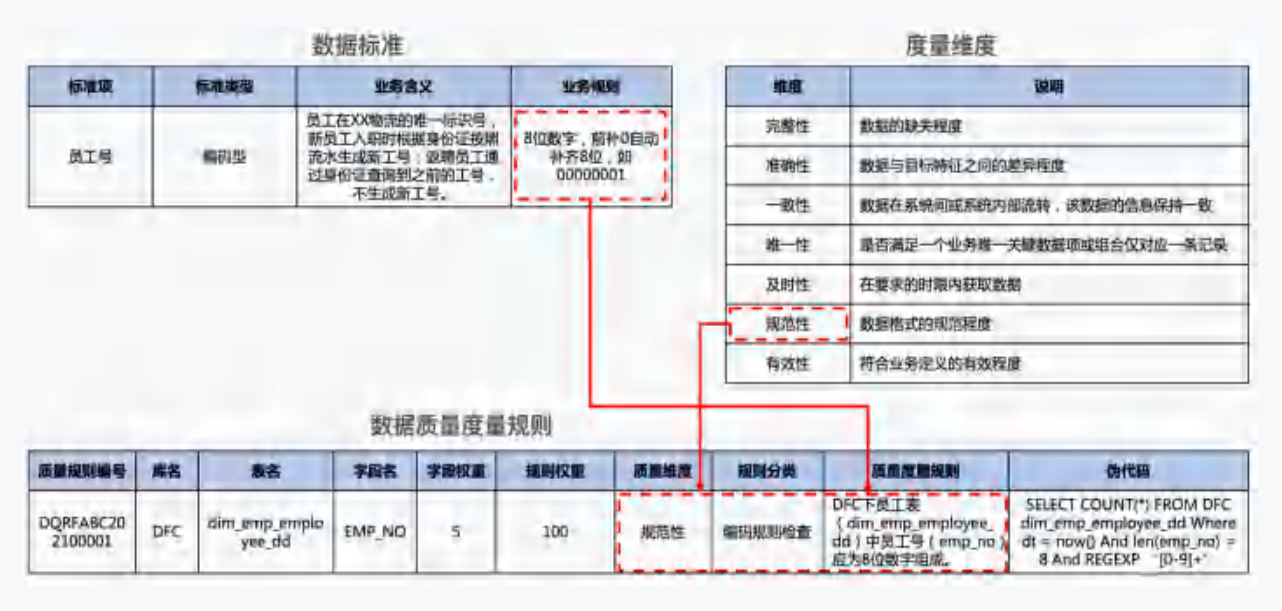
在模型设计中心进行分层模型的设计及维护，DWD层维护表与源表的映射及加工逻辑映射，DWS层与指标进行绑定关系，同时通过模型设计中心结合数据血缘进行模型使用及复用度的监控，进行模型使用的监管及治理。

数据地图构建方案

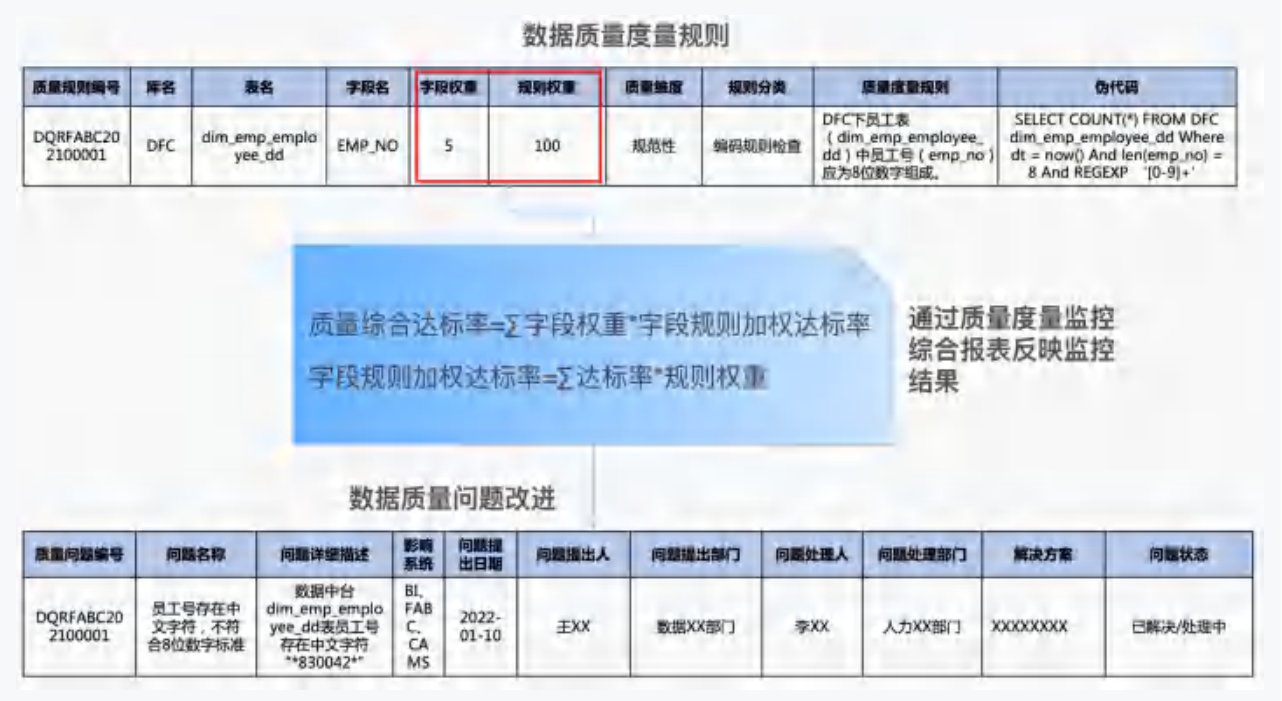
基于平台逻辑数据湖的能力，无论是逻辑入湖还是物理入湖，表的数仓模型、指标、数据标准等都可以在数据地图中进行查询，支持查看模型字段详细信息及血缘信息、辅助需求、开发及设计，从而提升用数效率。

数据质量构建方案

基于数据标准进行数据质量规则制定，结合数据质量度量及改进形成数据质量闭环管理。如员工号为例，在数据标准中制定其业务规则，同时结合度量维度制定出数据质量度量规则，如下图所示：



在数据质量度量规则中指出该规则所涉及的字段权重、规则权重、质量维度、规则分类、质量度量规则，其中字段权重和规则权重作为质量综合达标率以及字段规则加权达标率的依据。最后，根据度量监控综合报表形成质量改进问题清单，落实到个人并结合数据质量播报进行数据质量过程提醒，及时进行数据质量的问题发现及改进，实现质量问题的闭环。



数据中台实施成效

截止到目前为止，集团内部共构建了781条数据标准，分别涉及业务、财务等部门。表的建设总量已达到32000张左右，模型数量16000多个，总ODS表被跨层依赖率为10.95%，DWD一级下游表数量平均值2.92，DIM一级下游表数量平均值为16.60，DWS一级下游表数量平均值为2.00。通过构建质量制度，结合度量维度制定出数据质量度量规则，对数据质量形成了闭环管理。

网易数帆 | CSDN

微信扫码添加在线专家

为您提供企业数据治理全方位咨询

2022年12月印刷