

基于数据全生命周期的数据资产价值评估方法及应用

李冬青, 刘吟啸, 邓镭, 李铭洋

阿里巴巴集团, 上海 200120

摘要

数据资产价值评估是现代数据资产管理和运营以及数据流通的基础。基于数据全生命周期理论, 从第一性原则出发, 通过评估单张数据资产表的成本、数据管理以及数据应用价值, 实现对单张数据资产表的系统性评估。利用数据仓库和图算法等技术, 以层为单位, 每层分摊, 血缘路径继承, 精确计算得到单张数据资产表的成本价值; 然后利用层次分析法得到数据资产非经济因素权重, 进而得到数据资产阶梯价值; 最后通过实例分析验证了新方法的合理性和可行性。

关键词

数据资产; 数据资产价值评估; 数据仓库; 数据血缘

中图分类号: TP3

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2023044

Data valuation approach and application in view of data full lifecycle

LI Dongqing, LIU Yinxiao, DENG Lei, LI Mingyang

Alibaba Group, Shanghai 200120, China

Abstract

Data asset valuation is the foundation of modern data asset management, operation, and data circulation. Based on the theory of the data full lifecycle, starting from first principles, single data asset table was evaluated by assessing their cost, data management, and data application value. Using technologies such as data warehousing and graph algorithms, the cost value of a single data asset table was accurately calculated by using a layer-by-layer allocation method and inheriting the lineage path. Then, the non-economic factor weight of data asset was obtained by using the analytic hierarchy process, and the value of data asset was obtained through a ladder evaluation. Finally, verifies the rationality and feasibility of the new method was verified through an example.

Key words

data asset, data asset valuation, data warehouse, data lineage

0 引言

《“十四五”数字经济发展规划》中指出,数字经济是继农业经济、工业经济之后的主要经济形态,是以数据资源为关键要素,以现代信息网络为主要载体,以信息通信技术融合应用、全要素数字化转型为重要推动力,促进公平与效率更加统一的新经济形态^[1]。数据要素是数字经济深化发展的核心引擎,需要有序开展数据确权、定价和交易活动,并探索建设与数据要素价值和贡献相适应的收入分配机制。数据资产价值评估是实现数据流通和应用的重要基础,数字经济的发展迫切需要人们对评估数据资产这一课题进行深入研究。

在国家大力推动数字经济发展的同时,企业也在积极响应并开展大规模的数字化转型。目前,科技的高速发展使数据在工业生产中的体量日益增大,同时各种技术也日新月异。不论是内部管理方面,还是外部交易方面,企业都需要一套合理的数据资产价值评估方法。数据资产的合理估值对内可以衡量企业数字化建设进程和数据运营效果,对外可以提升数据资产的流动性,如数据交易,给企业提供新的收益来源。从公司价值层面来说,数据资产价值将在企业的会计3张报表中展现或在附录中披露,这将直接影响未来企业的市场估值。目前,世界级科技公司基于大量用户数据进行挖掘和分析以创造商业盈利。然而,数据资产价值的衡量仍缺乏实际的解决方案。

广为人知的IBM大数据4V特性意味着大数据的价值评估一定是一个难题。尽管各种关于大数据的研究、挖掘、分析、实践和应用等热门技术都已经取得了显著的成

果,但是客观且科学的数据资产价值评估体系和数据交易研究仍处于初期。只有经过科学管理,并且能够被运营转化为应用价值的数据才能算作真正的数据资产。在价值评估方法论方面,传统领域包括无套利定价、收益最大化定价、公平和真实定价。同时,也有一些涉及机器学习的动态数据定价、在线定价以及联合和协作学习中的定价方法^[2]。综合目前的数据估值发展研究,总结得出,同时考虑数据的经济因素(如数据成本、市场收益等)和非经济因素(如数据质量、时效性等)是更可行且可操作的方案。

迄今为止,数据资产价值评估问题尚未有成熟的解决案例和类似计算器的数据资产价值评估操作系统。本文基于国内互联网行业通用的数据技术,设计并开发了一套数据资产评估模型,旨在解决这个问题。

1 研究现状

数据资产价值评估属于交叉学科,涉及计算机科学、经济学、市场营销学以及新兴的数据科学等多个领域。由于数据具有多面性并且价值评估的目的不同,其原理和侧重点也有一定差异。近年来,随着信息和数字化时代的发展,该领域的研究逐渐受到重视。

姚建国等人^[3]研究了基于熵的数据价值衡量与定价方法,仅依赖数据交易平台收集到的数据集的浏览点击次数和获得该数据集支付的成本费用信息来对数据进行定价。信息熵定价法充分考虑了数据资产的稀缺性^[4],但该方法缺乏对数据的本质的讨论,没有考虑到实际数据源获取和加工等问题的复杂性。

2019年,中国资产评估协会制定了

《资产评估专家指引第9号——数据资产评估》^[5], 阐述了成本法、收益法和市场法3种方法。成本法适用于对个人数据的隐私补偿定价^[6]; 收益法主要用传统金融学模型对未来现金流和收益做折算, 直接量化数据效用, 体现买方市场增收; 市场法主要基于有效率的交易价格(类似二级市场的股票)的供需关系进行定价。对于难以量化的数据资产来说, 市场法(如拍卖和交易)是最公正的方法。然而现实情况是数据交易所的机制并不是对所有的企业都适用的, 并且尚未进行规模化发展。

闭珊珊等人^[7]基于成本法提出了一种数据资产评估的CIME模型, 即成本费用、固有价值、市场供求和环境约束4个因素分别对应4种简单的方法: 成本评估、层次分析法(analytic hierarchy process, AHP)评估、市场法评估和收益法评估。但其更侧重于对系统框架的构建和工具的设计, 并没有对方案的算法和实现技术进行进一步的精细化设计, 同时缺乏落地的结果的合理性检验。熊巧琴等人^[8]总结了数据资产的特性、流通方式、交易方式以及不同的数据估值方法和局限性, 同时对数据产品作为交易对象和区块链技术如何完善交易体系进行了讨论。但是其仅从理论和研究现状方面进行了总结以及客观评价, 并没有给出具体可以进行实际操作的办法。

与经典的金融领域中的资产抽象的资产评估问题不同, Babaioff等人^[9]认为数据资产具有协同性, 即不同的数据集组合可以带来不同的价值; Kerber^[10]指出数据资产具有先验不确定性, 即如果买方了解该数据资产的详细信息, 则数据带来的效用价值难以确定。Demchenko等人^[11]认为数据只有满足了6个重要属性, 才可以进行价值评估, 分别为独立、可靠、可复用、可互换、可操作、可衡量(sovareign, trusted, reusable, exchangeable, actionable,

measurable), 即STREAM原则, 这为未来的数据交易提供了一些标准和参考。Pei等人^[12]对数据资产评估背后的动机、基础原理和相关方法进行了总结, 但该文章仅从理论层面进行探讨, 缺乏实际案例。

而在《信息经济学》^[12]中, 资产估值工作被分为3个阶段: 质量衡量、价值衡量和经济效益衡量。资产价值衡量指标如图1所示。其中, 质量衡量指标比较可靠, 而价值衡量指标和经济效益衡量指标则更多是理论指导, 不太具备实际参考价值。

现有大多数研究还停留在理论层面, 仅对数据的价值评估因素进行描述, 如从数据使用者、数据生产者、数据管理者等视角进行分析, 或者根据某个学科专业, 在该学科的背景下解释数据资产价值问题。然而, 数据资产价值问题涵盖的领域非常广泛, 目前还没有一套完整的数据资产价值评估方案, 能够全面考虑数据从生产到消费的价值链, 并能够在实际生产实践中落地, 取得明显的效果和成果。

本文结合数据的生产和使用路径, 基于数据全生命周期的框架, 提出了一个具有实际应用意义的企业数据资产价值评估解决方案。该方案使数据从采集阶段便可进入价值评估系统中, 同时重点考虑数据存储和加工的价值转化过程, 让资产价值从数据源沿着数据仓库加工链路流动到数据实际应用层。数据资产价值可以从单张表的维度进行计算、查询和监控, 使企业可以在内部对运营情况进行量化评估, 在外部为数据在交易市场上流通提供价值参考。本文的创新性体现在3个方面: 首先, 从数据加工角度结合数据仓库的理论和大数据的特征, 提出了数据成本和数据血缘的成本继承思想, 并通过图算法解决其中相关路径问题, 开创性地对单表数据资产进行价值评估; 其次, 设计了非经济因素和专家打分机制, 将行业相关性重点且灵

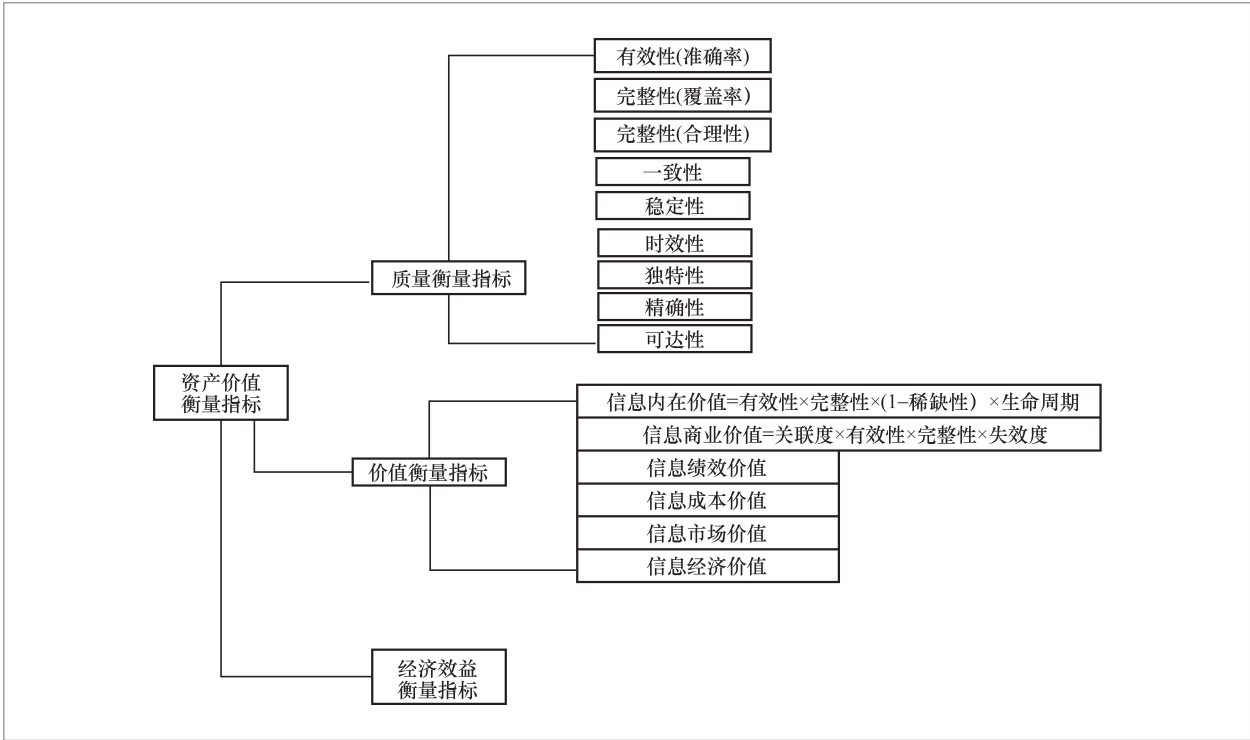


图1 资产价值衡量指标

活地反映在关键指标上,使用期望回报率代表数据的本身效能和使用价值。最后,本文提供了实证结果验证,并研发了数据资产价值计算器产品,有利于方案后续效果评估数据的积累和更多行业的覆盖性应用。

2 数据资产价值评估理论框架和方法

2.1 数据资产价值评估理论框架

2.1.1 基本概念界定

在数据资产价值评估中,需要明确数据交易范围、数据隐私和数据归属这几个基本假设。具体而言,在数据交易方面,需

要交易已经被清洗和分析过的应用层数据,而底层数据尚未被处理;在数据隐私方面,数据需要存储在数据管理系统中,并采取适当的权限保护措施,以避免因共享而导致的数据泄露;在数据归属方面,交易双方需要确保数据的归属权,并确保使用方向数据表归属方付费。

数据的全生命周期是指数据采集、传输、存储加工,以及数据后续使用的从生产到消费的全周期。该方法沿着该链路讨论各个环节的重要影响因子。数据规范加工和生产是数据资产价值评估的基石。在数据成本分摊计算中,依照数仓维度建模准则,数据加工方法如下。

数据仓库由数据仓库之父比尔·恩门(Bill Inmon)于1990年提出。数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据系统。范式建模是一种基于特定范式的建模方法,从数据源到公

司级数据仓库再到部门级数据集市,可以减少数据冗余,提高数据的一致性和稳定性。维度建模采用的是从数据集市、数据仓库到分散的异构数据源自下而上的建模方式。维度建模允许将维度信息适度冗余到事实表中,以提高易用性和查询效率。目前,传统行业(如银行等)普遍使用范式建模,而互联网公司则普遍采用维度建模。

在维度建模基础上发展了OneData建模规范:将公共数据划分为操作型数据仓储(operational data store, ODS)和通用数据模型(common data model, CDM)两层。ODS主要完成基础数据引入开放数据处理服务(open data processing service, ODPS)(一种阿里云自主研发的分布式处理服务),CDM主要完成公共数据加工与整合,建立一致性的维度,构建明细事实表和公共汇总事实层。明细粒度事实层将业务过程作为建模驱动,基于每个具体的业务过程特点,构建最细粒度的明细层事实表,将明细层事实表的某些重要维度属性字段进行适当冗余,也就是宽表化处理。公共汇总粒度事实层将分析的主题对象作为建模驱动,基于上层应用和产品的指标需求,构建公共粒度的汇总指标事实表,并采用宽表化手段物理化模型。

在公共层的基础上,还有应用数据(application data service, ADS)层,存放数据产品个性化的指标数据,计算CDM中间层针对业务的数据产出。通常情况下,越靠近数据源的层的数据加工难度越大。CDM层要求数据设计和数据加工严谨,而ADS层对数据的要求会随着业务的建设和需求变化更加灵活和随意。这种层层隔离的设计既保障了数据的严谨性,又保障了业务的灵活性,同时也给复杂数据的处理留足了空间。所有可售卖和调用的数据都

会放在ADS层进行统一管理。

2.1.2 数据资产价值评估整体框架

数据资产的价值受到多种因素的影响,涉及广泛的知识和技术领域。通过对数据全生命周期的透明化分析和管理,针对不同阶段的数据生产到消费过程,可以使用各种技术来衡量相关的要素影响。在数据管理领域,学术界和产业界的研究人员提出了不同的数据生命周期理论。阿里巴巴基于多年在大数据领域的实践,从“混通晒、存管用”^[13]逐步演进到“汇通管用”^[14]。其中,数据的全生命周期透明化管理对数据资产运营和数据资产价值发挥着重要的作用,数据的“汇通管用”全生命周期如图2所示。

数据的“汇通管用”4个阶段被用于数据资产的管理和应用。在这4个阶段中,“汇”代表汇聚异构数据的全过程,“通”代表数据治理的完备和连贯,“管”指全量数据资产的管理,“用”则强调数据应用服务的作用。基于这4个阶段相关的因素,进一步拆解构成数据资产价值评估模型,如图3所示。

- “汇”:包含数据成本的采集和追踪,包括对数据成本的录入和维护模块。
- “通”:基于数据仓库理论的数据血缘继承成本分摊模块。
- “管”:从非经济因素如数据质量管理要素和数据应用要素的角度来评价数据资产的价值。
- “用”:考虑公司整体情况和所处行业,给出对应的数据资产行业回报率。

整体数据资产价值评估流程如下。

- 首先,对数据生产成本进行全面梳理和归纳,参考《资产评估专家文件》,对数据的各项成本归纳。本文将数据成本总结为建设成本、机器成本、知识发现、软件

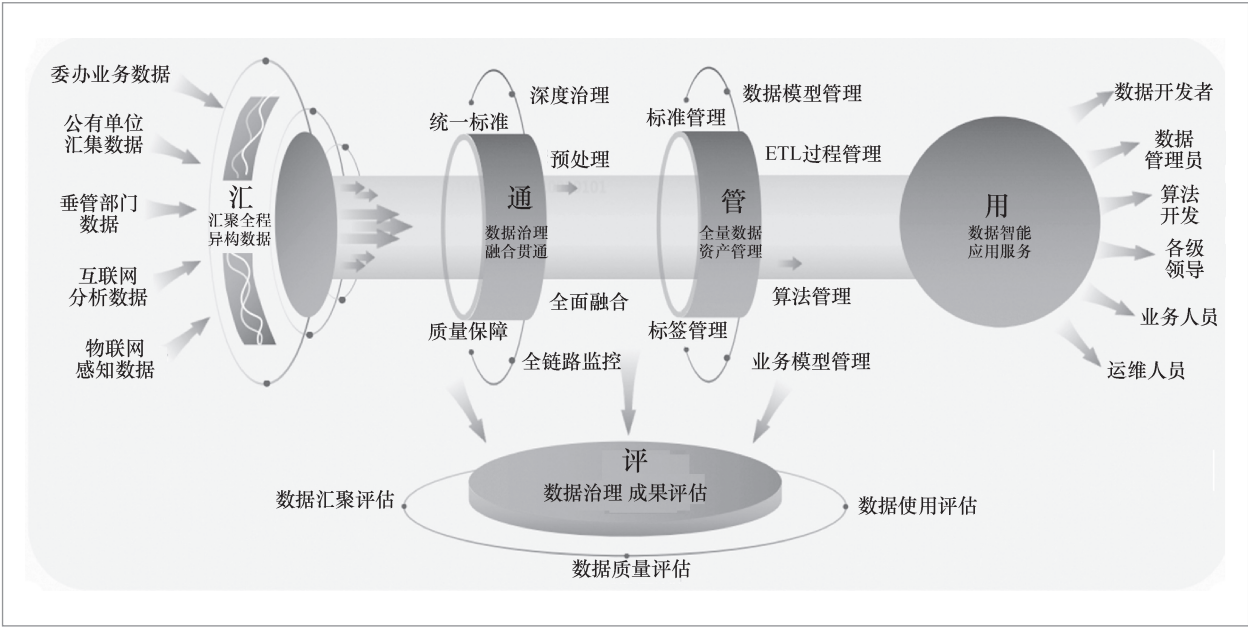


图2 数据的“汇通管用”全生命周期

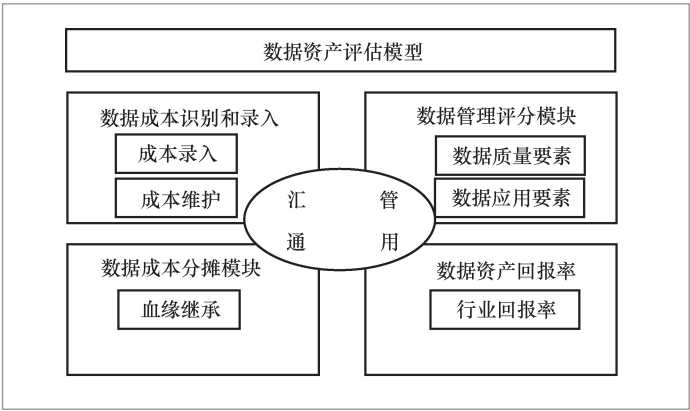


图3 数据资产价值评估模型

成本、运维成本。

- 其次，追踪数据成本，将可收集到的成本全部分摊给数据资产表的根节点。通过数据的血缘传递关系，沿着数据加工血缘传播数据成本。
- 然后，基于已有数据管理评分维度的总结，本文建立了数据管理评分体系，包括数据质量要素和数据使用要素。

数据管理因素主要解释数据研发后的运维和运营情况。本文采用数据质量评价指标和数据应用评价指标，并通过层次分析法对数据的非经济因素进行评价，从而得到单张数据资产表的数据管理分数。

- 最后，通过分析公司的经营情况得出整体的行业回报率，并将数据管理分数拟合合成每张表的数据资产回报率。企业的整体资产回报率与所处行业以及该行业的数字化建设能力相关。目前缺乏相关数据或有效的交易市场，因此，本文用企业期望值代替。

在工程上，该数据价值评估方案分为3个模块：数据源接入、数据ETL建模和算法模型。数据源接入包括采购数据、将原始信息导入系统和API接入。一般将数据存储到对象存储云服务中，然后由大数据产品进行数据的抽取、加载、清洗和转换，并以数据任务的形式进行开发、调度、预警和运维。数据加工主要包括数据源的

清洗和转换,各个业务域的数据处理和加工,加工层包括ODS、CDM/公共维度模型层(DIM)和ADS等;算法模型主要包括基础成本血缘分摊和阶梯价格两个算法。数据资产表的最终价值由这两部分的加和构成。

2.2 数据资产价值评估方法

基于传统资产评估理论的指导,本文采用成本法和收益率法,对数据在生命周期所处的每个阶段进行价值评估。同时,本文融合了大数据特有的数据仓库分层处理架构和数据血缘特性。在实际操作层面,本文横向基于业务应用划分数据域,纵向基于技术原理划分数据层的方式。

为了进一步将数据的实用性和质量等非经济因素量化为资产行业回报率,本文提出了单资产价值的核心计算式:单资产价值=基础成本价值+阶梯价值。基础成本价值由两部分组成:通过业务定义直接输入的财务发票,以及通过图算法计算的血缘继承成本价值,主要应用于ODS层,而其他上游层的成本价值基本上是通过血缘继承获得的。阶梯价值通过层次分析法衡量数据的管理和使用价值。

2.2.1 基础成本的计算

数据资产的基础成本指在特定时间点系统开发委托合同或实际支出发票能够追溯到的成本。参考《资产评估专家指引第9号——数据资产评估》^[5],数据基础成本分类如图4所示,主要包括建设成本、机器成本、知识发现、软件成本和运维成本。

其中,建设成本包括数据建设规划、采集获取、数据建设实施3个方面。数据

建设规划是指数据实施之前的调研、方案设计等,按照数据域的权重摊派一次性建设费用。最常见的输入形式为人天;采集获取是指数据采买费用,或者数据获取费用。从财务记账处获取,成本以单表计入,若出现一次采买多张表的情况(捆绑式售卖),建议平摊费用,由资产管理者手工录入;如果是API的调用费,则用数据接入的流量成本表示;数据建设实施是按照业务数据域的预定权重摊派一次性数据建设的费用,如数据工程师加工数据中台表的整体人天。

机器成本主要包括使用的硬件对表的存储和计算的费用,可以直接使用云产品的官方网站售价或合同价计入。

知识发现指通过算法工程师/大数据分析师挖掘带来的数据资产沉淀的成本,建议按照指标维度进行分摊。

软件成本包括数据平台和应用平台的使用费,例如,使用阿里云一系列的产品的购买费用,按照表的张数均摊。

运维成本包括外包维护费、运维人员维护表的费用,以及数据管理人员费用。建议按照表的张数均摊。

资产管理人员需要将总体数字输入产品页面,并通过系统对资产总体数量按照数据域的拆分比例进行分摊。数据域是建设数据仓库时重要的原始指标,它是一类相同主题表的集合,费用会最先被分摊到ODS层,再依据血缘继承价算法向应用层传递。根据数据成本的业务来源,基础成本=数据建设成本+机器成本+知识发现+软件成本+运维成本。

2.2.2 血缘继承成本的计算

“血缘”指数据的起源以及数据随时间变化的位置。它描述了数据的流转过程。血缘的可视化有助于提高分析数据流

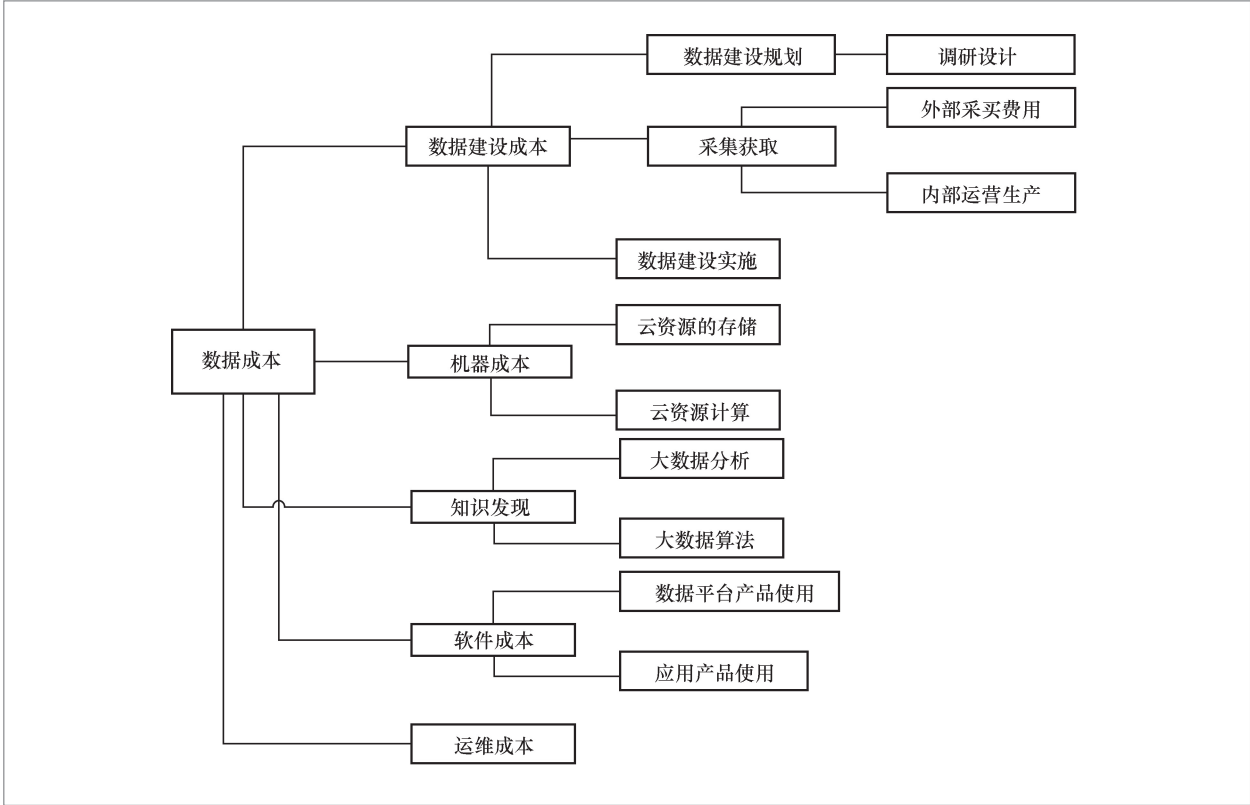


图4 基础成本分类

通的可视性，简化错误回溯的过程^[15]。血缘继承具有时效性和自动更新的特点。如果源数据的加工逻辑和血缘关系发生变化，基于血缘继承的方法可以自动更新，而无需进行重新调优。从这个角度来看，血缘逻辑是体现数据资产表价值的根本标准。在基础价值血缘继承方案中，本文设计了一条血缘继承的规则，即一张数据资产表有多少种路径可以到达下一层。这种方案避免了同层血缘关系的转化讨论，同时能够有效地包含同层空间的转化关系，从而简化了价值守恒问题的复杂性。

层与层之间要遵从的原则是基础成本传递价值守恒。为了解决数据源从某一个层加工到另一个层的分配问题，本文采用了图搜索算法，从一个节点开始，遍历关系直到到达目的地。这种算法可以用于物流

规划、最低成本呼叫、IP路由以及游戏模拟等多个领域。

图检索算法是图算法中的一个领域，它采用树理论对图进行搜索，在找到终点后回溯这一分支，最后获得从起点到终点的路径，图检索算法如图5所示。本文采用广度优先搜索，这种算法层层推进，将表作为起点，先找到距离该起点最近的下一张表，然后从内到外到达下一加工层，在下一加工层搜索记录路径，直至跳出该层。

在路径检索算法中，本文假设表价值继承到下游的每种路径都是无差异的。在加工链路中，数据产出到下游表并没有显著差异的需求侧资源。因此，在基础成本分摊的情况下，本文只讨论数据供给侧的劳力成本。任何实际数据有关需求侧效能指标对数据价格的影响，均体现在非经济因素中。

平均分为A/5；B表有两个路径到达CDM层：B→F和B→F→G，B在每个路径的价值被平均分为B/2；继承价值见表1，其验证了血缘价值守恒。在ADS层，只有一张表H且从ODS到ADS层的应用如下：A表有3个路径到达ADS层，分别为A→E→H、A→E→F→G→H、A→F→G→H，A在每个路径的价值为A/3；B表有一个路径到达ADS层，为B→F→G→H，A在每个路径的价值为B。由此ADS层表的价值为A+B。

2.2.3 阶梯价格的计算

阶梯价格用于衡量除成本外的其他因素的数据价值。其核心是围绕数据质量和数据使用效益进行评估，分别对应数据的固有价值和市场供需关系价值。从传统估值理论上也称之为数据期望回报率，这种价值在专家指导中的数据效用因素中有所体现。

由于大多数的非经济因素是难以比较和量化的，本文采用层次分析法进行分析。具体分为以下3个步骤：①通过大量的访谈定义影响数据估值的非经济因素；②通过专家打分计算所选的非经济因素的权重；③通过权重计算价值分数。

其中，非经济因素是通过业务选择评估目标对业务的重要影响因素来决定

的。除了业务经验之外，本文还参考了国家标准化管理委员会发布的数据管理能力成熟度评估模型DCMM (data management capability maturity assessment model)^[19]。结合专家的意见和方案的可行性，本文认为数据质量和数据应用是该管理框架下可以被量化的重要模块。根据与专家的充分讨论，考虑了工具和产品支持获取的指标统计的完整性和难易程度，本文确定了代表数据质量和数据应用价值的8个指标，数据的指标体系分类及定义见表2。

*n*对应的RI值见表3，专家按照表3对非经济因素重要性进行两两打分：每次打分都会构成对比较矩阵的元素，*a_{ij}*表示第*i*个因素对于第*j*个因素的比较结果。两两比较的因素形成矩阵，进一步进行矩阵一致性检验。如果矩阵中的每个元素均大于0(*a_{ij}*>0)且为正互反矩阵，若正互反矩阵满足*a_{ij}*×*a_{jk}*=*a_{ik}*，则称其为一致性矩阵。步骤如下。

步骤1: 计算判断矩阵最大特征值*λ_{max}*，一致性指标 (consistency index, CI)，*n*是矩阵的维度 (被打分的非经济因素的个数)。

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

(3)

步骤2: 根据*n*的大小，按照表3查找平均随机一致性指标 (rank index, RI)。

表1 继承价值

原始继承价值	继承价值	继承价值说明
E 表	E=A/5	只有一个以A为起点、E为终点的路径
F 表	F=2A/5+1/2B	有一个以A为起点、F为终点的路径和有一个以B为起点、F为终点的路径
G表	G=2A/5+1/2B	有一个以A为起点、G为终点的路径和有一个以B为起点、G为终点的路径
CDM层的总价值	E+F+G=A+B	价值守恒和ODS层一样
H表	H=A+B	有3个以A为起点、H为终点的路径和有一个以B为起点、H为终点的路径

表2 数据的指标体系分类及定义

项	明细指标	明细指标定义
数据质量指标	数据质量规则触发率	取dataphin/dataworks的数据,近30天触发数据质量规则的比率,指标口径为触发规则数量/总规则数量
	元数据属性完整性	数据资产在基础属性、管理属性和业务属性上的信息完整率,指标口径为非空项数量/表中指标数量
	内部引用次数	最近30天资产通过各种方式在数据中台内被引用的次数总和,包含代码调度任务、即席查询、同步任务
数据应用价值	输出次数	最近30天资产通过各种方式输出的次数总和,如数据同步、接口调用
	输出系统数	最近30天资产通过各种方式输出的系统数之和
	输出部门数	资产通过各种方式输出的部门总数
	平台访问次数	最近30天用户在全域资产平台上对此资产的访问、收藏、分享的次数之和
	平台访问用户数	最近30天在全域资产平台上对该资产有收藏、访问、分享的用户数去重之和

表3 n 对应的 RI 值

n	1	2	3	4	5	6	7	8
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41

一致性比例 $CR = CI / RI$,如果 $CR < 0.1$,则可认为判断矩阵的一致性可以接受;否则需要对判断矩阵进行修正,让专家重新打分;校验完矩阵一致性,进一步得到权向量:

$$w_i = \frac{1}{n} \sum_{j=1}^n \frac{a_{ij}}{\sum_{k=1}^n a_{kj}} \quad (4)$$

其中, w_i 表示各个非经济因素的重要性系数。基于前文AHP (Analytic Hierarchy Process) 算法整合所有信息得到非经济因素的重要性系数,得到单资产综合评价的分数。 S 表示某个非经济指标归一化后的评分, $Coef_i$ 表示该因素通过AHP计算后得到的系数:

$$Assetvalue = \sum_{i=0}^N S_i \times Coef_i \quad (5)$$

为了将不同数据的类型进行融合,需要进行归一化;对数值进行对数处理后进

行线性归一化,这样可以避免中长尾案例的分布过多而导致缺乏区分度的问题,同时再转化为0~100的分数。对数值归一化处理的计算式为:

$$x'_j = \frac{\lg(x_i + 1)}{\max(\lg(x_i + 1))} \times 100 \quad (6)$$

进一步根据价值分计算阶梯价格。为了避免数据过于分散,以及便于对应的业务人员的使用和管理成本,本文在价值分的基础上进一步聚类,选取价值分作为唯一有效特征,采用经典的Euclidean方法度量每次分之间的相似度:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (7)$$

本文采用K-means算法进行聚类。通过观察和业务需求聚类后会形成 k 个分数段,将 k 个分数段基于最高回报率和最低回报率,线性映射到 K 个回报率段中,最终得

到单表回报率,反映非经济因素对数据价值的效用侧和需求侧的影响。同时基于血缘分摊过后的成本价,计算最终的单资产表的价值。单资产表价值等于血缘成本分摊价乘以单表业务回报率,其中血缘分摊成本价就是该资产单价的阶梯价格。后续业务人员需要进一步基于单资产表的价值评估结果和市场期望制订售卖策略,以符合实际业务需求。这些商业收费策略可能会包括产品分润、数据公益、协议共享合作和按次付费等。

3 实证分析

本文的数据来源于某交通领域的数据资产平台,该平台已经完成了数据中台的建设。笔者团队为该公司提供数据估值服务,为公司内部数据的使用和未来对外售卖提供可靠的数据依据。根据数据资产价值的评估结果,业务会依据商业目标对数据进行进一步定价。数据的整体营收需要满足回报率可控和整体方案可解释性强的两个需求。

笔者团队使用了3 654张表进行实证分析。这些表分布在数据仓库模型的不同层中,彼此之间存在较为复杂的血缘关系。其中,应用层表数量为2 460张,根据部门将数据域属性归属划分为道路、交通和客户等。

3.1 基础成本价

笔者团队通过数据成本参数模块来收集初始的数据源成本,通常这部分数据由财务根据历史发票或者人力估算得出。基于该前提与假设,本文对广州交通的3 654张表进行数据实验,进行了小样本试算,以验证该算法在更大规模的数据上的可行性。

这些表的咨询、设计、实施、运维和软件费用比例为2:3:2:4:3,并以万为单位表示原始成本。这些表的总体成本为14万元,参数输入见表4。

除分摊基础成本的营收之外,还需要输入另外两个假设,即AHP打分的聚类个数和表收益预期期望。需要满足客户对4种不同类型的收益的需求,以使价值浮动的分布更加中心化。同时符合实际需求的收益率预计最低为基础价值的20%,而最高的表的价值上限不应超过80%。

3.2 血缘成本价

输入一系列的成本假设后,按照成本血缘路径法进行分配,成本分配价的直方图如图7所示。

从血缘路径的分摊结果来看,数据表可以分为0~50、100~200和200~400共3个区间,但数据分布中心并不是特别集中。这表明血缘路径更真实地反映了数据加工的情况,在总输入一定的前提下,成本会按照具体表加工的血缘情况进行分摊,最终的分摊结果可以很明显地反映各个表在纯数据加工方面的重要性和经济效益。然而,数据加工部分并不能全面地分析和论证数据表的价值,需要引入阶梯价值进行更严格和全面的试算。

3.3 阶梯价值结果

通过AHP对血缘分摊价格理论结果的调整,将数据资产的价值进行中心化收敛的平衡,同时基于价值回报率的调优符合客户期望的收益价值。

通过与相关行业专家对相关非经济因素进行充分讨论,本文列出了数据规则质量触发率、元数据属性完整性、血缘关系解析完整性、输出次数、输出系统数、输出

部门数、平台访问次数、平台访问用户数共计8个因素。其中数据质量规则触发率,元数据属性完整性和血缘关系解析完整性与数据采集和加工的质量有关,其他5个由公司内部的运营机制决定。专家给出的排序打分和访谈记录如下。

在挑选出来的8个重要的因素中:①前3个比后5个重要;②系统和部门数类似;③平台访问次数和平台用户数一样重要,客户打分见表5。

每张数据资产表的元素的归一化分数是对3 654 张表进行排序后的百分比归一化处理,数据产品会自动计算并每日更新。例如,随机选取某设备维度表的非经济因素的AHP结果见表6。

通过对专家对非经济因素的打分进行数学转化,可以得到非经济因素矩阵。对该矩阵进行分解并检验一致性,验证了专家打分的有效性(一致性比例:0.0054<0.1)。将非经济因素的归一化分数与AHP矩阵分解结果相乘,得出该表的分数为61.25分。

接着,采用K-means进行聚类,对即将售卖的表的分数进行聚类,并转化为相应的阶梯回报率。在ADS层的价值分布如图8所示,将该分布按照原始输入分成4个区间:15.88分~40.26分(回报率:36.9%),40.26分~54.12分(回报率:56.3%),54.12分~62.48分(回报率:75.2%)和62.48分~70.59分(回报率:80%)。

下面以一张具体的数据资产表来阐述如何从价值分数到最终数据资产收益率。价值分和最终价值分布如图9所示,某设备维度表的成本价为764.94元,阶梯分是62.87分,在62.48~70.59区间,对应的收益率为75.2%,则最终价格为764.94×(1+75.2%)=1 339元。这意味着如果其他部门要使用该表或将该表对外售卖,每年需要向该表所属部门支付1 339元。

表4 参数输入

参数名称	参数简写	假设值
咨询	Consult	30 000元
设计	Design	20 000元
实施	Implement	20 000元
运维(当年)	Maintenance	40 000元
软件(一次性投入)	software	30 000元
预计调用次数	Callnumber	500次
预计最高收益率/次	High_income	80%
预计最低收益率/次	low_income	20%

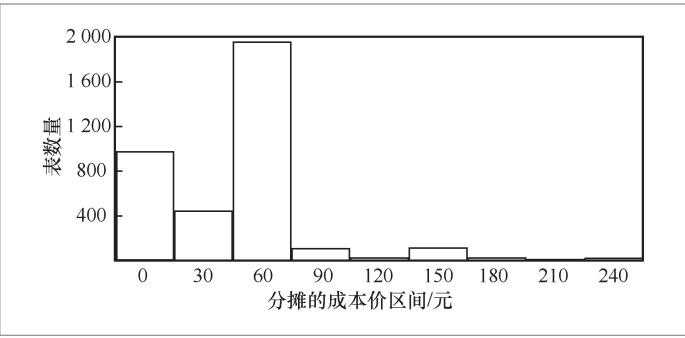


图7 成本分配价的直方图

表5 客户打分

项	Factor2	分值
数据质量规则触发率	元数据属性完整性	1.5
	血缘关系解析完整性	2
	输出次数	3
	输出系统数	2
	输出部门数	2
	平台访问次数	6
元数据属性完整性	平台访问用户数	6
	输出次数	4
...	...	...

3.4 阶梯价值结果

对所有的2 460张存储在ADS层可以

表 6 非经济因素的 AHP 结果

选取元素	归一化分数	分解结果
数据质量规则触发率	80	25.51%
元数据属性完整性	100	15.22%
血缘关系解析完整性	30	18.66%
输出次数	60	7.89%
输出系统数	50	11.83%
输出部门数	30	13.23%
平台访问次数	90	3.94%
平台访问用户数	50	3.72%

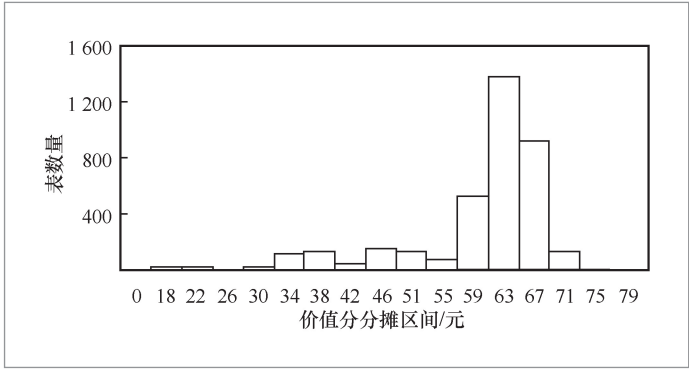


图 8 价值分的分布图

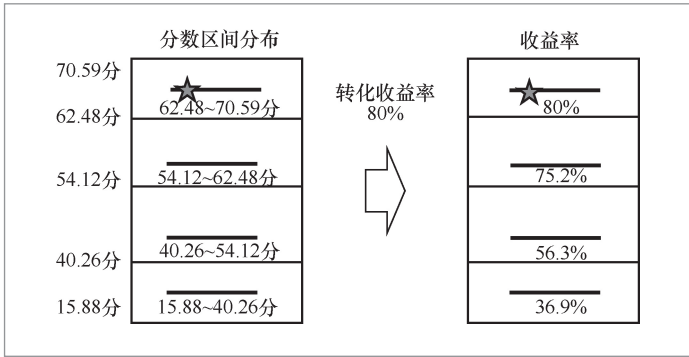


图 9 价值分和最终价值分布

对外售卖的表进行分析，可售卖表价格如图10所示，初始假设成本为14万元；这2 460张表呈现左偏度的双峰分布，25%的表价格在71元以下，75%的表格价格低

于210元，有25%的表格处于210~3 043元区间的价格，最高价值为3 043元，平均价格为200元，可售卖表价值试算见表7。

通过AHP的阶梯价值试算，将血缘分摊的成本价加入期望收益率的偏移影响部分，使数据资产定价更加合理并符合实际分布。其中，双峰分布的中心化数据体现出数据收敛的特性，而并不会因为AHP的调优而导致发散的情况。目前该交通企业通过该数据资产评估方案使每张表的对外交易都有价格可依，且整体资产回报率符合预期。

3.5 实际应用

前文的成本核算和分摊方法清晰呈现了大规模数据的完整生产链路以及成本流向。基于这一成本分摊结果，可以实现对重点使用数据(高成本、高使用量数据)的重点保障，确保数据使用的业务安全以及数据使用合规，同时也可以对无效数据生产链路进行发现与治理，减少公司的无效数据建设。对于有内部结算机制的公司，也可以通过这一方案对数据生产成本进行有效分摊。

在基于数据中台和数据资产管理平台的建设的场景服务中，笔者基于本文提出的数据价值评估方法，为数据部门提供数据资产价值评估服务，为内部的数据管理以及未来的数据对外售卖提供一定的数据资产价值评估依据，并满足了数据的整体营收需要回报率可控，且整体价值评估的方案可解释性强的两个重点需求。

4 结束语

本文从数据资产的生产和消费出发，提出一种数据全生命周期资产价值评估方

法,并在大数据平台和实际业务数据上进行了实践。该方法在模型层面上创新,详细描述了数据资产价值评估的计算过程。这一方法是对原有相关行业研究中^[20]提出的关于数据定价模型的进一步详细设计,目前已经在阿里巴巴集团内部及外部相关场景上得到了实际应用。同时,本文提出的方法可以较为精准的计算被应用和售卖的数据资产的理论成本,但目前还无法计算出数据资产损耗、数据资产的使用回报率等指标。另外,本文也指出了忽略数据资产在应用过程中增值再生的可能性,需要进一步探索和改进的问题,参考文献[21]还有相关框架外流程和要素未考虑。对于难以量化的数据资产,市场法和公开拍卖可能是比较公正的方法,但缺乏有效的数据交易市场 and 成熟的交易机制,评估结果无法进行市场化验证,未来需要密切关注数据交易结果对评估的效果,并进行方案改进。

参考文献:

- [1] 国务院. 国发〔2021〕29号. 国务院关于印发“十四五”数字经济发展规划的通知[S]. 2022.
The State Council of the People's Republic of China. Guofa [2021] No. 29. Notice of the state council on printing and distributing the “14th Five-Year Plan” digital economy development plan[S]. 2022.
- [2] PEI J. A survey on data pricing: from economics to data science[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(10): 4586–4608.
- [3] 姚建国, 李希君, 管海兵. 基于熵的数据价值衡量与定价方法: CN106815743A[P]. 2017-06-09.
YAO J G, LI X J, GUAN H B. Data value measuring and pricing method based on entropy: CN106815743A[P]. 2017-06-09.

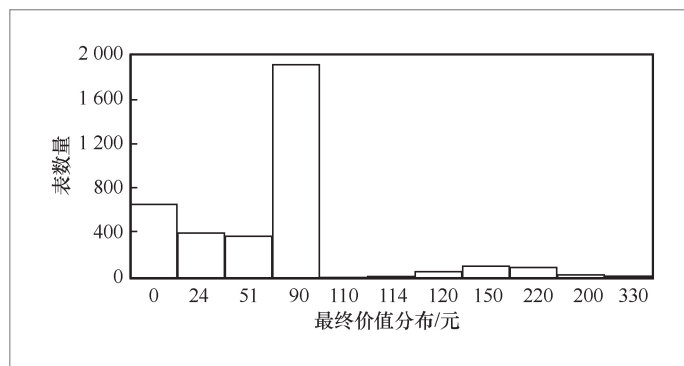


图 10 可售卖价值

表 7 可售卖表价值试算

项	基础价值	阶梯价值	最终价值
Count	2 460	2 460	2 460
mean	125.46	75.41	200.87
std	152.06	96.83	244.55
min	0	0	0
25%	42.16	21.07	71.82
50%	57.35	43.16	100.51
75%	126.48	95.19	221.67
max	1 933.72	1 306.78	3 043.20

- [4] LI X J, YAO J G, LIU X, et al. A first look at information entropy-based data pricing[C]//Proceedings of 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2017: 2053–2060.
- [5] 中国资产评估协会. 资产评估专家指引第9号——数据资产评估[S]. 2019.
China Assets Appraisal Association. Assets appraisal expert guidelines No. 9——data asset appraisal[S]. 2019.
- [6] GHOSH A, ROTH A. Selling privacy at auction[J]. Games and Economic Behavior, 2015, 91: 334–346.
- [7] 闭珊珊, 杨琳, 宋俊典. 一种数据资产评估的CIME模型设计与实现[J]. 计算机应用与软件, 2020, 37(9): 27–34.
BI S S, YANG L, SONG J D. Design

- and implementation of cime model for data assets assessment[J]. Computer Applications and Software, 2020, 37(9): 27-34.
- [8] 熊巧琴, 汤珂. 数据要素的界权、交易和定价研究进展[J]. 经济学动态, 2021(2): 143-158.
XIONG Q Q, TANG K. Research progress on the right delimitation, exchange and pricing of data[J]. Economic Perspectives, 2021(2): 143-158.
- [9] BABAI OFF M, KLEINBERG R, PAES LEME R. Optimal mechanisms for selling information[C]//Proceedings of the 13th ACM Conference on Electronic Commerce. New York: ACM Press, 2012: 92-109.
- [10] KERBER W. A new, intellectual, property right for non-personal data? An economic analysis[J]. MAGKS Papers on Economics, 2016(11): 989-998.
- [11] DEMCHENKO Y, LOS W, LAAT C D. Data as economic goods: definitions, properties, challenges, enabling technologies for future data markets[J]. ICT Discoveries, Special Issue, 2018(2): 23.
- [12] DOUGLAS B. 信息经济学: 如何对信息资产进行定价、管理与度量[M]. 曹雪会, 扈喜林, 朱琰敏, 译. 上海: 上海交通大学出版社, 2020.
DOUGLAS B. Information economics: how the information asset pricing, management and measurement[M]. Translated by CAO X H, HU X L, ZHU Q M. Shanghai: Shanghai Jiao Tong University Press, 2020.
- [13] 车品觉. 决战大数据: 升级版[M]. 杭州: 浙江人民出版社, 2016.
CHE P J. The decisive battle of big data[M]. Zhejiang People's Publishing House, 2016.
- [14] 阿里云, 毕马威. 数生万物, 转型之本—2021 数据资产运营白皮书[R]. 2021. Alibaba Cloud, KPMG: digital life, the foundation of transformation — 2021 Data Asset Operation White Paper[R]. 2021.
- [15] 李春梅, 张星, 耿慧拯, 等. 基于数据血缘构建数据分析方法[J]. 中国新通信, 2020, 22(20): 50-51.
LI C M, ZHANG X, GENG H Z, et al. Building a data analysis method based on data lineage[J]. China New Telecommunications, 2020, 22(20): 50-51.
- [16] GHORBANI A, KIM M P, ZHOU J. A distributional framework for data valuation[C]// The International Conference on Machine Learning. New York: PMLR, 2020.
- [17] JIA R X, DAO D, WANG B, et al. Towards efficient data valuation based on the shapley value[C]//The 22nd International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2019.
- [18] GHORBANI A, ZOU J. Data shapley: equitable valuation of data for machine learning[J]. arXiv preprint, 2019, arXiv: 1904.02868.
- [19] 国家标准化管理委员会. 数据管理能力成熟度评估模型: GB/T 36073-2018[S]. 2018.
Standardization Administration. Data management maturity assessment model: GB/T 36073-2018 [S]. 2018.
- [20] 尹传儒, 金涛, 张鹏, 等. 数据资产价值评估与定价: 研究综述和展望[J]. 大数据, 2021, 7(4): 14-27.
YIN C R, JIN T, ZHANG P, et al. Assessment and pricing of data assets: research review and prospect[J]. Big Data Research, 2021, 7(4): 14-27.
- [21] 叶雅珍, 刘国华, 朱扬勇. 数据资产化框架初探[J]. 大数据, 2020, 6(3): 3-12.
YE Y Z, LIU G H, ZHU Y Y. An initial exploration on framework of data assetization[J]. Big Data Research, 2020, 6(3): 3-12.

作者简介



李冬青 (1995-), 女, 阿里巴巴集团数据科学专家, 澳大利亚精算师, 主要研究方向为精算学、大数据、人工智能。



刘吟嘯 (1989-), 男, 阿里巴巴集团高级算法专家, 主要研究方向为大数据系统优化、人工智能。



邓镭 (1988-), 男, 阿里巴巴集团算法专家, 主要研究方向为人工智能和大数据。



李铭洋 (1998-), 男, 阿里巴巴集团算法工程师, 主要研究方向为量化和大数据。

收稿日期: 2023-01-05