数据管道模型：场外流式数据市场形态探索

任洪润[1,2]，朱扬勇[1,2]
1. 复旦大学计算机科学技术学院，上海 200438；
2. 上海市数据科学重点实验室，上海 200438

摘要
当前数据要素市场建设探索主要集中在数据交易场所（场内）建设，而流式数据市场指数据供应商向数据使用者持续、快速地供应特定数据的市场，流式数据并不适合在场内交易，因此需要探索流式数据的场外交易模式。研究了当前流式数据市场的运行现状，指出了市场无序、监管工具不足是存在的主要问题，提出了场外流式数据市场的数据管道模型，包括管道流通要件（数据管道、数据工厂、数据供应链）、市场规范要件（数据计量表、质量抽检器、合规审核仪）等，论证了数据管道模型的技术可行性，以期为场外数据市场建设、规范和监管提供理论和技术支持。

关键词
数据市场；数据管道；流式数据；场外交易

中图分类号：TP311　　文献标志码：A　　doi: 10.11959/j.issn.2096-0271.2023031

Data pipeline model: exploration of the over-the-counter form of streaming data

REN Hongrun[1,2], ZHU Yangyong[1,2]
1. School of Computer Science, Fudan University, Shanghai 200438, China
2. Shanghai Key Laboratory of Data Science, Shanghai 200438, China

Abstract
*T*he current exploration of data element market construction mainly focuses on the construction of data trading venues, while the streaming data market refers to the market in which data suppliers continuously and quickly supply specific data to data users. Streaming data is not suitable for transactions in the data trading venues, so the over-the-counter trading model of streaming data needs to be explored. After studying the current operation status of the streaming data market, it was pointed out that market disorder and insufficient regulatory tools were the main problems, and a data pipeline model for the over-the-counter streaming data market was proposed, including pipeline circulation elements

(data pipeline, data factory, data supply chain), market regulation elements (data meter, quality sampling device, compliance auditing device), etc. The technical feasibility of the data pipeline model was demonstrated, and it provided theoretical and technical support for the construction, regulation and supervision of the over-the-counter data market.

Key words
data market, data pipeline, streaming data, over-the-counter transaction

0 引言

建设数据要素市场是国家大数据战略的重要内容之一，其目的是引导数据市场从自发、无序的状态走向规范、有序的状态，支撑以数据为关键要素的数字经济健康、快速发展。当前的数据交易主要由买卖双方以数据相关的技术服务自发进行，未能真正体现数据交易，数据交易行为尚不规范、政府监管难度较大。“十四五”期间，各地政府将建设数据交易所、交易中心等数据交易场所作为数据要素市场建设的首要任务，通过引导数据产品进入场内交易，实现数据交易的合法性和有序性[1]。然而，从挂牌的数据交易产品来看，当前场内交易主要是数据集产品的交易流通，流式数据交易仍然在场外进行。

随着数字经济的快速发展，数据要素流通日益加快，流式数据市场需求旺盛且流通方式愈发复杂，例如，在证券行情分析、交通导航和人工智能实时训练等典型流式市场中，存在大量的流式数据交易[2]，但这些流式数据交易长期在场外进行，尚未见到在场内进行交易，这说明流式数据难以在场内进行交易。同时，由于当前流式数据的场外交易主要是买卖双方以数据相关的技术服务自发进行，其价格、质量以及合规性等都难以得到政府的有效监管，不利于流式数据市场的健康发展。例如，一次性计价或订阅计价是常见的数据定价模式[3]，但这两种模式本质上都是卖方垄断的定价行为，损害数据使用者利益的情形时有发生，难以形成数据供需双方的利益平衡。以订阅计价模式为例[4]，数据使用者支付供应商设置好的订阅费用对流式数据进行订阅，数据量需求较小的使用者可能不得不为其需求外的数据支付费用。由于缺少可计量的计价依据、缺乏有效的政府监管，合理的数据价格难以形成。流式数据质量、数据内容的合规性同样难以保证。

综上所述，在大力推动场内数据交易市场建设的同时，也要高度重视场外流式数据交易市场的建设，积极探索规范有序的场外流式数据市场，实现流式数据市场的规范运行和有效监管，对计价方式、数据质量、内容合规等进行严格管控。本文分析了流式数据市场的基本特征和基本要求，提出了数据管道模型，为当前流式数据市场在流通、监管等方面存在的各种问题提供可行的解决方案，保障流式数据的安全、快速流通，促进场外数据市场的健康、繁荣、可持续发展。

1 流式数据市场分析

1.1 流式数据市场案例

当前存在多种细分且活跃的流式数据市场，比如证券数据市场、智能导航交通数据市场、人工智能实时训练市场等，这是流式数据流通的3种典型场景，流式数

据市场繁荣但相对无序，且尚无可计量的数据计价依据。

（1）证券数据市场。证券数据领域是流式数据计算非常典型的应用场景之一。证券交易机构的日常运营中会产生大量数据，不仅有结构化数据，也有非结构化数据和半结构化数据，但这些数据的时效性往往较短。在瞬息万变的金融市场中，传统批量数据处理不能及时响应需求，需要对实时数据进行流式处理。通过对这些数据进行流式计算，挖掘并发现数据中的内在特征，有助于数据使用者实时决策[5]。以证券交易所为例，证券交易所产生的实时数据，由相关公司收集整理后出售给证券数据服务商，证券数据服务商再将加工后的数据出售给证券数据使用者，证券流式数据的传输仅由某几家证券数据商完成，存在流式数据计价方式混乱的问题，并且给市场监管带来一定困难。

（2）智能导航交通数据市场。近年来，无人驾驶技术作为未来城市交通的最佳解决方案受到了大众的瞩目。无人驾驶车辆行驶过程中对大量数据的及时处理事关驾驶可靠性和安全性[6]，不仅包括对车辆搭载传感器产生的流式数据的处理，还包括对路基、路面、道路设施等路况信息的处理。因此，无人驾驶的难点之一在于如何将各种与交通相关的因素进行整体协同以处理复杂的实时道路状况。在场内数据交易模式中，数据供应商和数据使用者通过交易所交易数据的场内交易模式难以满足无人驾驶对大量流式数据实时流通、分析和处理的需求，需要探索新的面向流式数据的场外流通形态，这对未来无人驾驶的全面普及意义重大。

（3）人工智能实时训练市场。人工智能内容生成是人工智能实时训练的一个典型市场，主要集中在文字、图像和音频领域[7]，已初步形成市场分工及产业链，并已产生一批代表性企业。对话式AI是人工智能内容生成的主要应用之一，实时流式数据的传输对其模型的训练、新场景中高质量对话的构建至关重要。然而，当前人工智能内容生成市场仍处于相对无序的状态，虽然已经初步形成市场分工及产业链，但针对流式数据市场的流通模式尚未形成，垄断企业存在随意定价的情况，定价方式混乱，市场监管困难。基于流式计算的实时用户画像也是人工智能实时训练的典型应用之一，通过对多源流式数据进行实时计算处理、挖掘用户的特征及需求，实时用户画像系统能够刻画出精准的用户画像。然而，当前市场流通的用于实时用户画像的流式数据质量良莠不齐，可能会影响模型训练效果，从而影响用户画像的精准性和实时性。

1.2 问题分析

当前数据要素市场建设分为场内建设、场外建设两大类。场内交易指交易所、交易中心的模式，这是当前各地政府在大力探索建设的数据市场。场外交易长期存在，之前主要以有明确知识产权的数据产品交易为主，如音频、视频、图片等。但这些并不是当前数据要素市场建设重点关注的，数据要素市场建设关注的是非标准、多类型的大数据产品。当前关注的流式数据指大量连续到达且难以存储在内存中的数据序列项[8]，只能按照到达顺序被数据使用者访问，呈现出实时性、易失性等特性[9]。流式数据流通需要在较短时间段内（甚至实时）由数据供应商将数据可靠地交付给数据使用者，并且使用者希望这种流通能够做到“想用即买、随买随用”，就像使用自来水一样，打开水龙头就实现了对水的购买计量和使用；就像使用电力一样，打开电器就实现了对电力的购买计量

和使用。

目前，场外数据市场客观存在，但处于无序状态，难以监管，给数据的安全利用、合法利用带来隐患。而流式数据基本是在场外市场流通的。就目前情况看，流式数据流通存在如下问题。

问题一：场内交易难以满足流式数据流通的需求。

在当前的数据交易所模式中，数据供应商提供数据集到交易所挂牌，然后数据使用者通过交易所购买所需的数据集，交易所提供数据交易的结算服务，并确保后续的数据交付行为。然而，截止到2022年，在30多家有交易的数据交易所挂牌的产品中，尚未看到在交易所挂牌的流式数据产品，存在如下3点原因：一是流式数据流通需要不断地从卖方传输到买方，数据交易所并不能将买卖双方进行管道连接，数据也无法以流式形式从卖方流向买方；二是由于不能进行数据管道连接，交易所不能对流式数据在流通过程中进行实时计量计价；三是交易所模式难以满足“想用即买、随买随用”这种流式数据使用者的基本需求，这点也是最重要的。例如，上述3种场景中流通的数据均未见到在当前的数据交易机构挂牌流通。因此，交易所模式并不适合流式数据流通，需要探索适合流式数据的场外流通模式。

问题二：场外流式数据市场监管工具不足。

当前流式数据基本是在场外市场流通的，场外市场尚处于无序状态，缺少市场监管工具，难以对价格、质量、合规性等进行有效监管。

● 价格监管：当前流式数据流通的定价方式主要以订阅制[4]为主，按一个时间段（如月/季/年）进行订阅缴费。订阅收费为数据供应商带来了相对平缓的现金流，但这本质上还是一种卖方垄断的定价行为。在订阅制模式下，数据供需双方难以形成利益平衡，其原因包括部分数据使用者可能会被迫接收用不到的数据，并不得不为之支付费用、资源等；同时，部分对数据需求较大的使用者可能以较小的代价就可以获得所需的数据；对于数据供应商而言，传输这部分数据也会给其带来巨大的不必要的资源开销。流式数据具有持续、实时、易失等数据集产品所没有的特性，这些特性是流式数据价值的重要体现。针对这些特性设计一种通过政府监管许可、能对流式数据实时计量定价的新模式，有可能会帮助流式数据达到合理的市场价格。

● 质量监管：对于流式数据而言，数据产品在流通过程中的质量保证是决定其价值能否得到充分体现的重要因素。流式数据产品的质量不仅体现在数据的准确性、一致性、完整性、可访问性等维度，更体现在流式数据能否进行连续流通。通常，流式数据使用者需要对实时数据进行分析，将实时产生的结果提供给业务使用者，维持使用端业务的正常运转。如果流式数据在传输过程中出现数据质量过差、数据断流等严重的质量问题，后续的业务将无法正常运行[10]。例如，自动驾驶系统需要实时收集并分析道路状况数据，一旦采集到的路况数据出现严重的质量问题或者断流，可能会导致自动驾驶系统的决策出现重大失误甚至失效，造成严重事故。设计有效的流式数据流通的质量检测及管控机制，可以使流式数据传输过程中的数据质量得到有效保障。由于缺少相关的技术工具，要做到质量保证与监管并不容易。

● 合规监管：当前的流式交易难以实现对数据合规性的有效监管。在流式数据流通中，很可能会发生身份证号码、手机号码、银行账户等个人隐私的泄露，虚假广告、暴力等不良信息[11]的传播，对涉及政治、宗教、种族、性别等敏感话题的讨论。

如果不加以监管和限制，不仅个人隐私难以保证，不良信息的传播也可能会影响公共安全和社会稳定。另外，如果在数据交易中没有及时发现并屏蔽这些话题的相关内容，可能会有违反法律法规的风险。需要设计有效的数据监管方式，保障数据交易的合法合规。

由此，当前迫切需要探索场外流式数据市场形态，对流式数据进行合理的计量计价，同时保障数据在传输流通中的质量管控、内容合规审查，使得流式数据在场外市场健康发展。

2 流式数据管道模型

针对流式数据市场的基本要求，本文设计了数据管道、数据工厂、数据供应链、数据计量表、质量抽检器、合规审核仪等，其中，数据管道、数据工厂和数据供应链是管道流通要件；数据计量表、质量抽检器和合规审核仪是市场规范要件。这6种要件的连接构成一种流式数据流通模型，被称为数据管道模型。

2.1 流式数据规范有序流通的基本要求

针对交易所模式难以满足流式数据流通，以及当前场外交易模式存在的监管困难问题，需要探索场外流式数据市场形态。一个规范有序的流式数据流通市场应该具备持续性、快速性、高可靠性、可监管等特征，以满足流式数据流通的需要。

● 持续性。流式数据可以抽象为一个无穷的数据序列，只要数据源处于活动状态，数据就会一直产生和持续增加，流式数据的到达、处理和输出是持续不间断的，因此规范有序的流式数据流通市场首先要能保证数据的持续流通。

● 快速性。在保证流式数据流通市场可持续性传输的基础上，还要保证其数据流通的快速性。流式大数据是实时产生、实时计算的，其数据价值会随着时间的流逝不断减少。保证流式数据的快速传输，对保障数据处理结果的及时反馈和数据产品的实时更新至关重要。

● 高可靠性。流式数据流通市场的可靠性建设主要考虑两个方面：一是设计安全备份，通过状态备份和故障恢复策略保障数据管道的传输安全、不断流；二是设计传输的并发响应机制，传输过程中出现用户访问量高并发时，采取并发响应机制保证数据流式传输过程中的负载均衡和稳定运行。

● 可监管。流式数据市场的规范流通离不开对数据价格、质量以及内容的有效监管。合理的数据价格应该得到政府监管许可且满足数据供需双方的利益平衡。流式数据流通市场的数据质量和内容同样需要受到政府监管。

2.2 管道流通要件

流式数据是边生产、边供应、边流通的，因此流式数据市场需要有数据管道、数据工厂和数据供应链等基本内容，这些被称为管道流通要件。

（1）数据管道

数据管道是指供应商到使用者之间的数据连接，用于供应商向使用者提供所购流式数据，可以通过公共网络或专用网络实现。为了实现流式数据流通的持续性、快速性、高可靠性等要求，数据管道需要有较好的带宽资源作为基本保障，为数据管道的快速流通提供时间和数据量方面的基本保障。

另外，备用管道的建设也必不可少。当故障发生时，根据预先定义的策略进行数据流的重放和恢复，备用管道技术可以保

障主管道故障时流式数据的正常流通。

（2）数据工厂

数据工厂是指流式数据市场中的一个数据产品生产加工点，分为核心数据工厂和供应数据工厂。核心数据工厂是流式数据管道模型中的最终数据产品生产者（集成工厂或总装工厂），实现最终数据产品的生产加工并将其传输给最终的市场用户；供应数据工厂是相对于核心数据工厂而言的，除核心数据工厂外的其他工厂都被称为供应数据工厂。供应数据工厂接收来自上一级数据工厂的流式数据，联合自行采集的数据，经过加工生产形成新的流式数据，并将其传输给下一级数据工厂。供应数据工厂既是数据使用者（相对于上级数据工厂），也是数据供应商（相对于下级数据工厂）。数据工厂如**图1**所示，其中，当右侧为最终用户时，其为核心数据工厂；当右侧为下级数据工厂时，其为供应数据工厂。

（3）数据供应链

随着流式数据市场规模的不断扩大，流式数据市场分工逐渐细化，逐渐形成了由流式数据生产商、数据采集商、数据加工商、数据产品总装商、数据最终使用者（即市场用户）等角色构成的流式数据供应链。其中，数据产品总装商作为数据供应链的核心数据工厂，将数据生产商、采集商、加工商等各级数据工厂供应的数据进行总装以形成产品，并将数据产品出售给市场用户。例如，在人工智能内容生成市场中，已初步形成由数据供给商、数据标注商、模型算法开发商、智能芯片制造商、创作者平台开发商等上游数据商，内容设计商、内容制作工具提供商、数据分析梳理商等中游数据商，以及内容终端生产厂商、第三方内容服务机构等下游终端数据商角色构成的市场分工及产业链，产业链上各级数据工厂之间关系密切。

因此，在数据管道建设中要充分考虑数据在管道中的流通过程，建设以核心数据工厂为中心、上下游数据工厂密切相关和配合的流式数据供应链，把各级数据工厂通过数据管道进行连接。流式数据供应链如**图2**所示。数据管道中存在一个核心数据

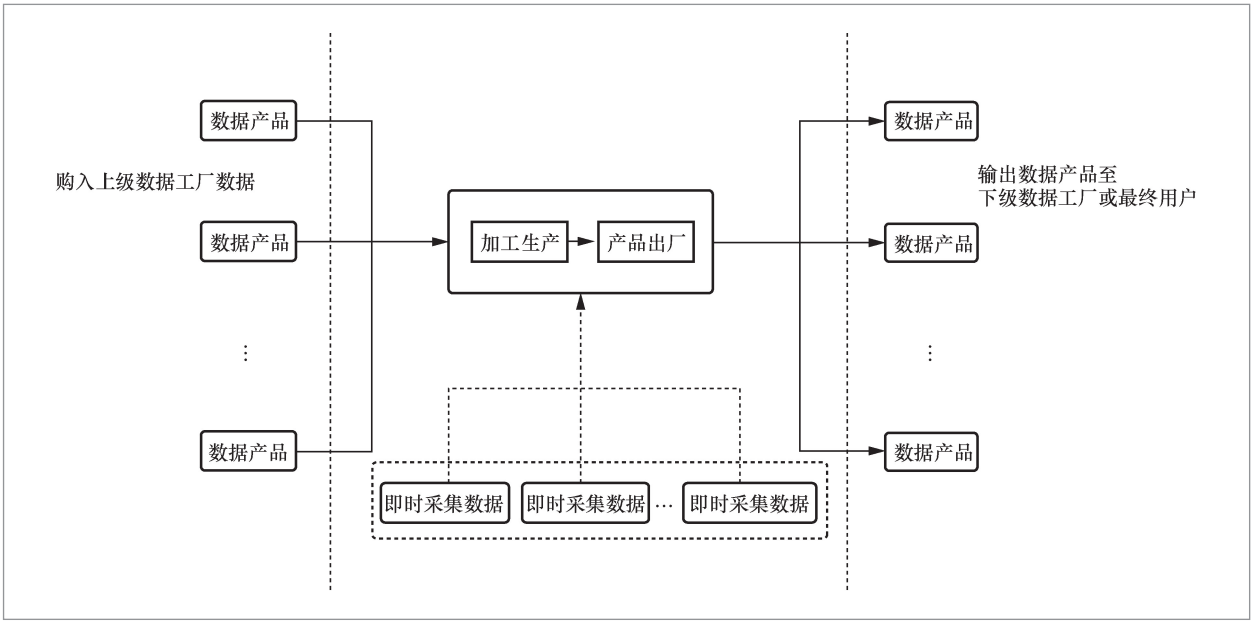

图1 数据工厂

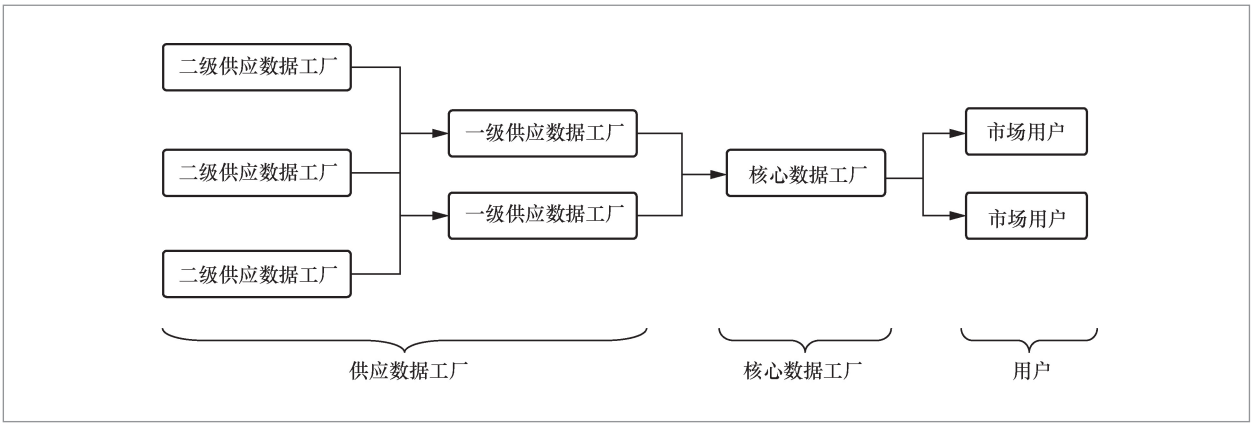

图 2 流式数据供应链

工厂，其上游是供应数据工厂，负责为核心数据工厂供应数据，下游是市场用户。核心数据工厂、供应数据工厂和市场用户共同构成了流式数据管道供应链的3类市场角色。

2.3 市场规范要件

市场规范要件指政府规范数据市场所需要的监管工具。市场规范要件都是在政府监督下按照政府标准生产的仪器仪表类器件（作为市场监管工具），可以是硬件设备也可以是特定软件，在数据管网建设期间安装到位。其中，质量抽检器和合规审核仪安装在数据管道的公共部位，分别由质量监督部门和网信部门使用，数据计量表则安装在用户端，主要用于计价付费，类似于水表、电表之类的仪器仪表。

（1）数据计量表

在基于数据管道的数据流通市场中，数据计量表是流式数据资金结算的依据。从流式数据满足用户需求及其相应的资金结算方法视角来看，需要设计满足政府监管许可的、新的流式数据的计量计价方式——通过数据计量表计价，为价格的形成和监管提供依据。

参考水表、电表的计量方式，考虑流式数据实时、持续、无限等特性，数据计量表可能是一种更加合理的计价方式。过去的订阅制计价方式无法对流式数据的实时传输进行计量，不适用于流式数据计价，还可能损害数据供需双方的利益。在数据计量表定价的情况下，数据使用者可能会更加关注数据量、流速、时长等指标，采用数据计量表对这些指标进行计量并将其作为定价标准，可能会改善结算方式，使得数据供需双方达成更加合理的平衡。

流式数据价值的有效实现离不开数据流量、数据流速和时长三大要素。首先，数据规模增大到一定程度之后，隐含于数据中的知识的价值也随之增大，大量的数据有助于数据特征和规律的有效识别、挖掘。其次，流式数据是实时产生和计算的，其结果也需要实时反馈，流式数据的价值随着时间增长而急剧减少，这对数据传输和计算速度提出很高的要求。良好的带宽可以提供良好的基础设施，可以保证“大流量”的流式数据以“大速度”传输，保障数据管道传输的畅通和高效。最后，数据流过数据管道的时间不仅是数据使用者比较关心的计量指标，同时也可作为数据计费的依据。由此，如**图3**所示，设计出由数据流量、数据流速、时长3个部分组成的数据计量表，既满足了数据管道的设计要求，同时为流式数据计价提供了计量依据。

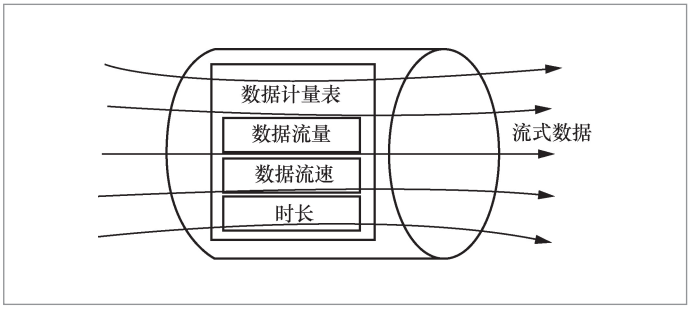

图 3 数据计量表

● 数据流量：指特定时间段内的数据流量。按数据流量进行计费可以较好地反映资源变化对价格的影响，具有较好的资源优化性能，缺点在于计费算法复杂，用户不易控制预算。

● 数据流速：是计量流式数据的重要技术指标之一，指数据在管道中的传输速度，即在单位时间内（通常为1 s）传输的比特数。通过分析数据流速做相应调整，管道可以运行得更加有效，并可以在高负荷时预防处理带宽限制。

● 时长：指两个数据工厂之间建立数据管道连接的持续时间。

（2）质量抽检器

数据质量检测是确保数据准确性、完整性和一致性的重要步骤，数据质量的好坏直接影响数据分析和决策的准确性和可靠性。准确的数据是正确决策的基础，数据的不正确以及必要数据的缺少都会导致不准确的结果和不可靠的决策[12]。

流式数据在数据管道中持续流通，流通过程中很可能会出现质量问题，需要对流式数据的质量进行检测。然而，如果不断地对流式数据进行检测，会影响数据流速、流量等，进而影响流式数据的使用效果。因此，类似流水线的商品抽样检测法，采用质量抽检器对流式数据进行检测，随机抽取给定时间窗口内流过数据管道的所有数据，然后采用数据集质量检测的相关技术对流式数据进行检测。

（3）合规审核仪

随着数据规模和价值的不断增长，数据安全和隐私保护已经成为一项越来越重要的任务。数据合规审查可以检测企业和组织在数据使用、传输等方面是否遵守相关的法律法规和行业标准，对数据市场进行规范。

在数据管道中采用合规审核仪审查数据可以检测数据工厂输出的数据是否符合法律法规的要求，例如是否违反与国家数据安全[13]和个人信息保护相关的法律法规等。此外，数据合规审查还可以帮助防范内部和外部的数据风险，如数据泄露、数据丢失、数据滥用等。合规审核仪能够对数据进行分类、检测，为实现市场对数据合规性的全面管理和控制提供有效工具。

2.4 模型框架

数据管道、数据工厂、数据供应链、数据计量表、质量抽检器、合规审核仪6个要件的有机连接构成了数据管道模型。数据管道模型框架如**图4**所示。

在流式数据管道模型中，在上下级数据工厂的数据传输过程中，均需使用数据计量表对流式数据进行计量，使用质量抽检器检测产品质量，使用合规审核仪监测数据安全以及过滤不良内容等，并将此作为市场主体资金结算的计量依据和政府监管的实现手段。数据工厂中的数据传输及加工流程如下。

● 数据工厂获取上一级数据工厂传输的N个数据产品，经由数据管道传输数据时，首先对数据产品进行质量检测、合规审核等，然后使用数据计量表计量数据产品的数据流量、数据流速、时长等，保障数据产品的质量和安全，并为计价提供依据。

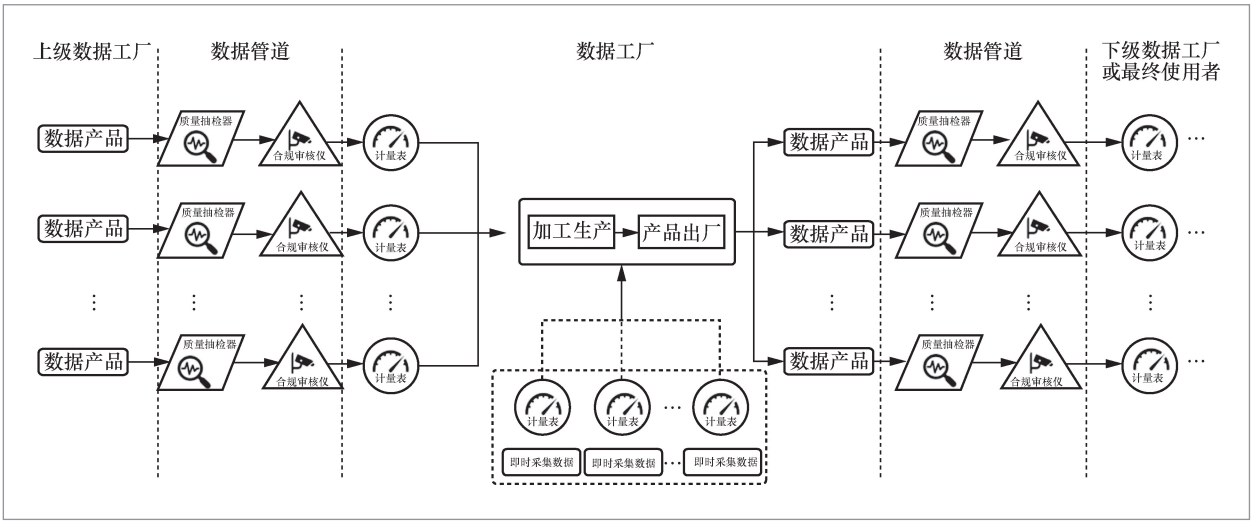

图 4 数据管道模型框架

● 除了获取上一级数据工厂传输的数据产品，数据工厂也可以自行收集流式数据。设置K个流式数据采集端收集流式数据，使用数据计量表对收集的数据进行计量。

● N个数据产品与K个数据采集端收集的数据共同被传输至数据加工生产环节，生产输出新的流式数据产品，经过质量检测和合规审核后，供应给下一级数据工厂或者市场用户。

3 数据管道模型技术可行性

3.1 管道流通要件技术实现

建设具有持续性、快速性、高可靠性的数据管道是管道数据流通的基本保障。根据数据在管道中的流通过程，考虑数据源、数据传输、数据供应端和使用端控制等方面的技术可行性。

（1）数据源——数据接入技术

数据接入技术：数据管道涉及多种类型数据源的接入，包括即时采集的数据以及数据产品等，数据原材料可通过数据接入技术进入数据管道。设计良好的数据接入技术可提供接纳流式数据的能力，将多种数据源进行统一封装后推送至数据消息队列，使用消息队列作为中间件解决了接入数据与后续数据处理之间的应用耦合、异步处理、流量削锋等问题[14]。常见的分布式消息队列包括Java消息服务（Java message service，JMS）、Kafka、RocketMQ、AMQP、RabbitMQ等，其中Kafka[15]是LinkedIn开发的一款消息队列，许多流式数据处理系统使用Kafka作为底层的消息处理模块，保证快速的消息持久化、高吞吐率，以及实现自动的负载均衡。

（2）数据传输——备用管道技术、容错机制

数据管道中各数据工厂之间的数据流通离不开数据传输的相关技术，数据管道中的数据传输方式可分为主动推送方式和被动拉取方式。上游数据商产生或计算完数据后，主动将数据发送到下游数据商的传输方式被称为主动推送方式，其优势在于数据传输的主动性和及时性；只有下游数据商提出请求，上游数据商才会通过数

据管道将数据传输至下游数据商的传输方式被称为被动拉取方式，其优势在于下游数据商可根据自身负载状态、工作状态等进行数据传输[16]。数据管道对流式数据传输的实时性要求较高，数据需要得到及时处理，往往选择主动推送的数据传输方式。当然，主动推送方式和被动拉取方式不是完全对立的，也可以将两者进行融合，从而在一定程度上实现更好的效果。在流式数据传输过程中，需要考虑备用管道、容错机制的建设以保证数据管道流通的安全性。

备用管道技术是当主管道发生故障或数据断流时，根据预先定义的策略进行数据流传输的重放和恢复，保障数据畅通的技术。备用管道技术的备份策略包括定期备份策略和实时备份策略。若采用定期备份策略，当主管道进行数据传输时，备用管道处于待命状态，主管道上的状态会定期同步至备用管道。当故障发生时，通过备用管道恢复数据传输。这种策略的优点是传输开销小，缺点在于所恢复数据的时效性可能不足。若采用实时备份策略，主管道进行数据传输时，同时向备用管道传输一份数据副本。故障发生时，备用管道全部接管主管道的数据传输工作。优点在于当故障发生时，数据的实时性不会受到影响，缺点在于管道传输开销较大。

容错机制指在一定范围内允许或包容错误发生，并进行修正的机制，是进一步改善管道性能、保证数据管道可靠持续运行的重要措施。S4、Puma、Kafka、Storm[17]等流式计算系统实现了对部分容错的支持。Apache Flink[18]提供了容错机制来恢复数据流应用的状态。一旦程序出错（机器故障、网络故障、软件缺陷等），Flink会停止这些分布式的数据流，并将其重置到最近一次正常的检查点，输入流会重置到快照点，确保重启后的数据流状态得到更新。

（3）数据供应端和使用端控制——资源调度技术、负载均衡技术

数据管道中的原始数据可能由多个供应端提供，同时也可能提供给多个使用段，需要将资源安排到不同的数据供应端、使用端以保证最佳利用率，错放资源不仅可能导致资源利用率低，还会导致数据管道不稳定。这个过程中需要保证多个数据供应端、使用端之间的资源调度和负载均衡，确保控制持续、稳定的数据传输，并实现最低的资源消耗。

资源调度的主要功能是根据资源容量、队列等方面的限制条件，将系统中的资源分配给各个数据供应端、使用端。通过系统资源调度可实现对管道中资源的最佳利用，提高资源的利用率，保证完成任务和节省能耗。在TimeStream[19]的系统结构中，集群管理器实现了对系统资源的管理和任务的分配，当出现系统故障或负载剧烈持续变化的情况时，可以通过提前保存的一些数据信息进行系统状态的恢复和实时动态的调整。

负载均衡是一种计算机技术，用来实现流式数据管道中资源的分配负载，以达到最优化资源使用、最大化吞吐率、最小化响应时间以及避免过载的目的。设计良好的负载均衡技术可以保证数据管道实现以下几种特性：一是能实现传输的高可用性，当数据管道运行异常时，能快速实施故障转移并自动恢复负载；二是有良好的内置功能，能自动隔离异常，确保数据的安全；三是能轻松扩展或减少资源，且可以实现灵活扩展在线服务，从而实现对管道中任务的动态、合理分配，动态适应系统负载情况，保证数据管道的任务均衡、稳定地运行。在Kafka消息系统中，发布者和代理节点之间没有负载均衡机制[20]，但可以通过专用的负载均衡器在Kafka代理上实现基于传输控制协议（transmission

control protocol, TCP）连接的负载均衡的调整，以保障整个系统处于良好的均衡状态。

3.2 市场规范要件技术实现

（1）数据计量表相关技术

数据计量表是满足政府监管许可的、用于流式数据计量计价的工具，通过特定软件或硬件设施可实现对流式数据的计量。流量监控软件可以通过对数据流量进行实时监控和记录，实现对数据流量的测量和分析；交换机、路由器等硬件网络设备可以提供数据流量和数据流速的统计信息，这些信息可以用于测量流量和流速。另外，还可采用数据包捕获工具捕获和分析数据包，不仅可以测量流量和流速，同时还可以提供其他有用的性能指标和统计信息。

（2）质量抽检器相关技术

为确保流式数据的质量，可对在特定时间窗口内流过管道的所有数据进行随机抽取并检测，然后对抽取出的数据集进行质量检测，此时可采用与数据集质量检测相关的技术，如数据清洗、数据验证、数据重复性检测、数据完整性检测、数据一致性检测等。蔡莉等人[12]提出的质量评价模型，可以实现数据清洗、数据验证、重复性检测、完整性检测和一致性检测等，帮助识别并修复数据中的错误，确保较高的数据质量。在数据采集和收集过程中，由于各种原因（如数据输入错误、数据缺失等），数据中可能存在不准确、不完整、不一致、不合法或重复等问题。同时，数据中的异常、错误、缺失等问题也会影响数据的准确性和可靠性。数据验证指对数据进行逻辑和合法性的检查，以确保数据符合业务需求和规范要求，数据验证可分为格式验证、范围验证、规则验证等，例如验证数据是否符合预设规范、是否在指定范围内、是否符合特定的逻辑规则等。

由于流式数据来源和数据格式的多样性，同一份数据可能存在多个版本或多个记录，这就需要进行数据重复性检测。数据重复性检测主要通过比较流式数据之间的相似性或差异性，去除重复的数据，以确保数据的一致性和准确性。另外，针对数据的完整性和可靠性可检测数据完整性，例如检查数据是否存在不完整的记录或数据，是否存在空值或缺失的数据等。通过对流式数据进行交叉比对、验证数据之间的逻辑关系和一致性可以完成数据一致性检测，确保数据的准确性和可靠性[21]。

（3）合规审核仪相关技术

流式数据的合规审查可以利用数据安全加密、敏感词过滤、监测和审计等技术实现，通过确保数据的安全性、完整性和可用性，对数据进行分类、管理和监控，实现对数据合规性的全面管理和控制。

首先，数据安全加密技术可以防止未经授权的访问和窃取数据，解密技术可以帮助授权人员在必要时快速获取敏感信息，同时运用数据分析技术对数据进行统计分析、趋势分析和异常检测等，可以发现潜在的合规问题和风险。其次，通过区块链技术可以确保数据的安全性和难篡改性[22]，使用Trie树和确定有限状态机（deterministic finite automaton, DFA）等[23]敏感词过滤算法可以避免敏感信息在数据中泄露和传播，保护用户隐私和数据安全，从而提高数据的合规性。最后，通过监测和审计技术，可以追踪数据的使用和访问记录，及时发现和防止违规行为，保障数据合规监管的实施。

4 总结与展望

数据要素市场探索建设是数字经济发展的基础性工作，当前各地政府着重推进

数据交易场所建设，而对场外数据市场建设不够重视。事实上，当前数据流通主要以场外交易流通的方式为主，因此规范场外数据市场并实施有效监管应该是数据要素市场建设的另一个重要内容。本文分析了场外流式数据交易流通存在的问题，设计提出了数据管道模型，实现了“想用即买、随买随用”的流式数据交易模式。采用数据管道模型，数据使用者可向数据供应商进行限时按量购买、按表计价，不仅解决了订阅制模式不适用于流式数据定价的问题，也可为流式数据定价提供了重要参考。另外，在数据管道中设计有效的质量检测及合规审查机制，可以使流式传输过程中的数据质量得到有效保障。因此，数据管道模型是解决流式数据流通相关问题的一个可行方案，可以保障数据持续、安全、可靠流通。

数据管道模型是对场外流式数据市场形态的一种有效探索，数据管道可能是未来公共数据开放共享和公共数据服务的一种主要方式，类似于现在各家各户的有线电视系统。未来，城市建设除了传统的供水、供电、供气管网建设，还应包括数据管网建设。未来的城市数据管道网络不能只存在一条以核心数据工厂为中心的流式数据供应链，而应该由多个核心数据工厂和错综复杂的各级供应数据工厂共同构成，形成丰富完善的数据供给体系，满足城市数字化转型日益增长的数据需求。

参考文献:

[1] 包晓丽，杜万里. 数据可信交易体系的制度构建——基于场内交易视角[J]. 电子政务，2023：2023.06.003.
BAO X L，DU W L. Institutional construction of data credible trading system—based on the perspective of floor trading[J]. E-Government，2023：2023.06.003.

[2] CHEN J，LI M，XU H. Selling data to a machine learner：pricing via costly signaling[C]//Proceedings of the 39th International Conference on Machine Learning. [S.l.：s.n.]，2022：3336-3359.

[3] CONG Z C，LUO X，PEI J，et al. Data pricing in machine learning pipelines[J]. Knowledge and Information Systems，2022，64(6)：1417-1455.

[4] HERNANDEZ D，GAMEIRO L，SENNA C，et al. Handling producer and consumer mobility in IoT publish-subscribe named data networks[J]. IEEE Internet of Things Journal，2022，9(2)：868-884.

[5] KOLAJO T，DARAMOLA O，ADEBIYI A. Big data stream analysis：a systematic literature review[J]. Journal of Big Data，2019，6(1)：1-30.

[6] KARMAKAR G，CHOWDHURY A，DAS R，et al. Assessing trust level of a driverless car using deep learning[J]. IEEE Transactions on Intelligent Transportation Systems，2021，22(7)：4457-4466.

[7] ZHANG P，LIU B X，LU T，et al. A semantic embedding enhanced topic model for user-generated textual content modeling in social ecosystems[J]. The Computer Journal，2022，65(11)：2953-2968.

[8] O'CALLAGHAN L，MISHRA N，MEYERSON A，et al. Streaming-data algorithms for high-quality clustering[C]//Proceedings of 18th International Conference on Data Engineering. Piscataway：IEEE Press，2002：685-694.

[9] 孙大为，张广艳，郑纬民. 大数据流式计算：

关键技术及系统实例[J]. 软件学报, 2014, 25(4): 839–862.
SUN D W, ZHANG G Y, ZHENG W. Big data stream computing: technologies and instances[J]. Journal of Software, 2014, 25(4): 839–862.

[10] LI Y, YANG J C, ZHANG Z, et al. Healthcare data quality assessment for cybersecurity intelligence[J]. IEEE Transactions on Industrial Informatics, 2023, 19(1): 841–848.

[11] WU X B, XU Y H, SHAO Z L, et al. LSM-trie: an LSM-tree-based ultra-large key-value store for small data[C]// Proceedings of the 2015 USENIX Conference on USENIX Annual Technical Conference. New York: ACM Press, 2015: 71–82.

[12] 蔡莉, 梁宇, 朱扬勇, 等. 数据质量的历史沿革和发展趋势[J]. 计算机科学, 2018, 45(4): 1–10.
CAI L, LIANG Y, ZHU Y Y, et al. History and development tendency of data quality[J]. Computer Science, 2018, 45(4): 1–10.

[13] NIU C Y, ZHENG Z Z, WU F, et al. Online pricing with reserve price constraint for personal data markets[C]//Proceedings of 2020 IEEE 36th International Conference on Data Engineering. Piscataway: IEEE Press, 2020: 1978–1981.

[14] 陈纯. 流式大数据实时处理技术、平台及应用[J]. 大数据, 2017, 3(4): 1–8.
CHEN C. Real-time processing technology, platform and application of streaming big data[J]. Big Data Research, 2017, 3(4): 1–8.

[15] KREPS J, NARKHEDE N, RAO J. Kafka: a distributed messaging system for log processing[C]//Proceedings of the NetDB. [S.l.:s.n.], 2011.

[16] MISRA S, REISSLEIN M, XUE G L. A survey of multimedia streaming in wireless sensor networks[J]. IEEE Communications Surveys & Tutorials, 2008, 10(4): 18–39.

[17] TOSHNIWAL A, TANEJA S, SHUKLA A, et al. Storm@twitter[C]// Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2014: 147–156.

[18] CARBONE P, KATSIFODIMOS A, EWEN S, et al. Apache Flink: stream and batch processing in a single engine[J]. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2015, 36(4): 28–38.

[19] ALI M H, CHANDRAMOULI B, FAY J, et al. Online visualization of geospatial stream data using the worldwide telescope[J]. Proceedings of the VLDB Endowment, 2011, 4(12): 1379–1382.

[20] WANG G Z, CHEN L, DIKSHIT A, et al. Consistency and completeness: rethinking distributed stream processing in Apache Kafka[C]//Proceedings of the 2021 International Conference on Management of Data. New York: ACM Press, 2021: 2602–2613.

[21] LIU Q, FENG G Z, ZHENG W B, et al. Managing data quality of cooperative information systems: model and algorithm[J]. Expert Systems With Applications, 2022, 189.

[22] AN J, WU S Y, GUI X L, et al. A blockchain-based framework for data quality in edge-computing-enabled crowdsensing[J]. Frontiers of Computer Science, 2023, 17(4): 174503.

[23] BODON F, RÓNYAI L. Trie: an alternative data structure for data mining algorithms[J]. Mathematical and Computer Modelling, 2003, 38(7–9): 739–751.

作者简介

任洪润（1995– ），女，复旦大学计算机科学技术学院、上海市数据科学重点实验室博士生，主要研究方向为数据科学和数字经济，近期研究重点为数据定价、数据生产模型等。

朱扬勇（1963– ），男，博士，复旦大学计算机科学技术学院教授，复旦大学数据产业研究中心副主任。《大数据》期刊编委会副主任，农业大数据产业技术创新战略联盟副理事长兼首席科学家，大数据协同安全技术国家工程实验室副理事长，中国自动化学会国防大数据专业委员会副主任。国际数据科学倡导者，提出数据界、数据学、数据身、数据自治、数据财政等概念和体系。发表学术论文200多篇，出版《数据学》《旖旎数据》《特异群组挖掘》《数据自治》等专著，并任《大数据技术与应用丛书》（22册）主编、《大数据资源》主编。主要研究方向为数据科学和数字经济，近期研究重点方向为数字化转型、数据财政、数据资产、数据自治与数据跨境等。

收稿日期： 2023-02-27

通信作者： 任洪润，renhr20@fudan.edu.cn

基金项目： 上海市科委发展基金资助项目（No. 22DZ1200704）

Foundation Item: Shanghai Science and Technology Development Fund Project (No. 22DZ1200704)