基于区块链的数据市场

汪靖伟，郑臻哲，吴帆，陈贵海
上海交通大学计算机系，上海 200240

摘要
在互联网时代，每天都产生着不可估量的数据，在数据共享过程中，涌现出了数据隐私性和所有权归属等复杂问题。区块链是一种去中心化的分布式数据存储技术，引入区块链能消除集中式数据市场的弊端，但同时分布式数据市场又产生了安全与隐私问题。综述了国内外大数据交易市场的产业现状和研究进展，提炼了基于区块链的大数据共享流通平台应满足的性质。根据这些性质提出了一个基于区块链的数据市场框架，分析和讨论了这个框架中的安全性和隐私性问题及对应的解决方案。基于这个框架，实现了一个数据市场测试系统，并证实了该框架的可行性和安全性。

关键词
数据资产；数据市场；区块链；隐私保护

中图分类号：TP311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020021

Blockchain based data marketplace

WANG Jingwei, ZHENG Zhenzhe, WU Fan, CHEN Guihai
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract
Blockchain is a kind of decentralized distributed data storage technology. Blockchain can solve the disadvantages of the centralized data market, however, distributed data marketplaces also introduce the problems of security and privacy. Firstly, the current status from industry and progress from academia of the bid data marketplaces were reviewed, and then the required properties that a well developed blockchain based data marketplaces should satisfy were proposed. Based on these properties, a blockchain based data marketplace framework was proposed. Then the potential security and privacy problems in this framework were investigated, and the corresponding solutions for these problems were designed. Based on this framework, a data marketplace demonstration system was implemented and its feasibility and security were verified.

Key words
data asset, data marketplace, blockchain, privacy protection

1 引言

当今世界中的数据量正在迅速增加。在线社交网络Facebook自成立以来，已经收集了超过300 PB的个人数据，而这个规模还在进一步扩大。IBM公司的研究人员提出，当今世界90%的数据是在过去2年中产生的，而且随着新的设备和技术的出现，数据增长会进一步加快。在大数据时代，数据不断地被收集和分析，进而引领科技创新和经济增长。公司和组织使用其收集的数据提供个性化的用户服务、优化公司决策过程、预测未来趋势等。在广泛的数据使用过程中，人们开始关心个人数据的安全问题，担忧提供服务、收集数据的互联网公司是否会保护用户的数据隐私，而人们几乎无法控制他们所产生的数据及其使用方式。近些年来，许多与侵犯用户数据隐私有关的事件被报道，其中最著名的例子就是Facebook的5 000万用户数据被泄露，用户的隐私遭到了很大程度的侵害。

为了保证数据的正常流通与使用，充分发挥大数据的价值，近年来兴起了众多关于个人数据共享与交易的新兴机构。除了传统的数据流通方式（即公司与用户之间广泛存在的数据换取服务的模式）外，还涌现出了大数据共享交易市场，通过将数据需求与数据源进行匹配来促成数据交易。这些数据市场已经具备了相当的规模，这些数据市场被估值数百亿美元，并在持续增长。在数据市场中，数据持有者展示他们的数据信息，以吸引潜在的数据消费者；数据消费者搜索、选择他们需要的数据集，并通过支付一定的费用来获取数据使用权；数据市场通过促成数据交易来获得收益。但随着数据共享交易规模、数据价值的增长，共享交易过程中的欺诈和泄密的情况也会逐渐增多。集中式数据市场的架构一般如**图1**所示，在这种架构中，集中式的公司或组织运营的市场平台在系统中起着至关重要的作用。市场中涉及的各方——数据卖家、数据买家和市场平台，能通过串通舞弊、套利购买策略等方式获得更高的收益。此外，集中式的数据交易模式缺乏数据买方与数据卖方之间有效的信息沟通渠道，导致数据交易效率低下。最后，市场平台拥有更多的信息优势，即市场平台知道数据内容，而数据买家在未购买数据之前无法知晓数据内容，因此市场平台可以通过构建信息壁垒并控制信息披露来非法获得收益。

集中式数据市场存在一些不可避免的数据安全隐私、数据版权保护以及共享流通性能瓶颈等问题。首先数据交易的中介（通常情况下是市场平台）必须是安全可信的。市场平台需要具备公信力，确保其不会非法使用交易中的数据，泄露数据持有者的隐私。然而市场平台存在这样的动机，而且即使它违规使用、出售了数据，一般也难以追究。同时，集中式数据市场很容易成为攻击者的目标，用户的敏感信息（例如位置、聊天记录等）被保存在集中式的数据库中，存在隐私泄露和数据丢失的风险。现有大多数的数据市场在集中式服务器上运行，这样的系统存在着单点故障和单点性能瓶颈。有研究[1]表明，现有的集中式数据市场还会控制买家和卖家的互相搜索，导致市场运行效率低下。

为了规避集中式数据市场的弊端，去中心化的数据市场诞生了。去中心化的数据市场架构可以规避依赖可信中介介入数据交易的要求，摆脱单点故障和单点性能瓶颈，并提高透明度和可信度。但是去中心化的数据市场由于缺乏中心的管理，其

系统设计与安全性保证会比集中式数据市场更困难，比如“双重支付”问题一直是分布式系统的难点。近些年来，区块链技术日趋成熟，区块链去中心化的架构可以作为数据市场的底层架构，提供良好的支持。区块链是一种去中心化的分布式数据存储技术，在数据市场系统中引入区块链层，将使个人用户能够直接与数据需求方达成交易，不依赖任何第三方，从而让用户保持对数据的所有权，并确保交易过程的公开透明。

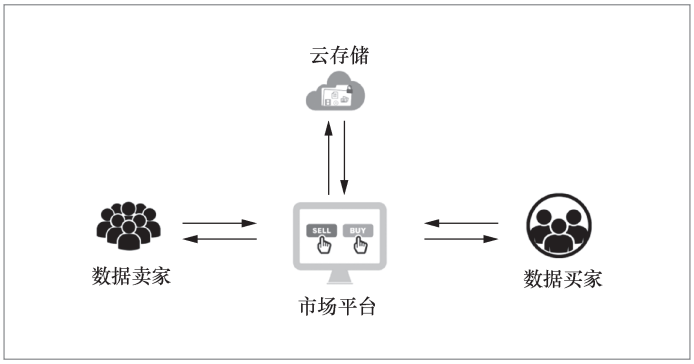


图1 集中式数据市场架构

2 市场调研与相关研究

2.1 现有数据交易市场调研

由于数据有优化决策和提供服务的功能，各个组织和机构都开始关注数据的流通和交易。比如，Datashift、Gnip、NTT DATA等公司转售来自Twitter等社交网络的数据，Xignite公司出售金融行业的数据，Factual公司则关注地理位置数据的交易。同时，还涌现出了大数据共享交易市场，通过将数据需求与数据源匹配来促成数据交易，比如Infochimps、AWS Dataexchange、Qlik Datamarket、Here等。Datacoup是一个集中式数据市场平台，允许用户出售各种类型的个人数据（包括财务数据和社交账户数据），其客户端应用程序允许用户从第三方应用程序（如Facebook和Twitter）导入数据。由于Datacoup从用户手中收集原始数据，因此用户必须在数据存储和数据管理上完全信任Datacoup。与Datacoup类似，People.io是一个集中式平台，其最大的特点是不会将个人数据直接出售给其他组织。它使用机器学习算法分析用户的个人数据，然后向用户推送个性化广告。用户虽然不会因提供其个人数据而直接获得奖励，但是他们可以通过接收个性化的广告来获得收益。

国内数据市场的发展尚不成熟，还未形成完整的数据流通交易的产业链条。比如有以互联网企业为主导的大数据共享平台，它们的数据大多来源于旗下应用软件收集的数据，如阿里云、京东万象等；还有数据堂、数海、浪潮天元、数多多等数据共享与交易的平台，这些平台以多种途径收集来自第三方的数据，实现大数据资源的在线交易。此外，还有由政府主导的大数据交易中心，这些中心多为政府/国企独资，或国企与民企合资，如贵阳大数据交易所和上海数据交易中心等。然而这些数据交易中心一般比较封闭，其具体的数据市场架构技术还比较模糊。

近几年，基于区块链的分布式数据交易市场引起了业界的极大关注。IOTA是专门针对物联网（IoT）设计的加密“货币”，利用区块链技术已经搭建了针对物联网数据的交易市场。类似的公司还有DataBrokerDAO、Datum、Datapace和Wibson等，这些公司有的直接出售其收集的数据集，有的向大众收集个人数据，并将其出售给个人用户。国内也有一些使用区块链技术构建数据市场的例子，如上海

数据交易中心采用联盟链将与交易有关的信息存储在区块链节点中，以确保数据交易安全、高效、可信。现有的区块链数据市场都只保证了数据市场构建的某些要素，没有全面地考虑构建数据交易市场应该达到的目标：去中心化、公平性、隐私性、有效性以及经济学性质。在第3节中，笔者将详细讨论这些性质。

2.2 相关研究工作

在研究界中也有许多工作涉及区块链、数据市场及其相关问题。Balazinska M等人[2]讨论了新兴数字数据市场的意义，并列出了这一方向的研究机会。Zyskind G等人[3]使用区块链保护个人数据的隐私，将区块链转变为一个不依赖受信第三方的自动访问控制管理器，以明确数据的所有权，确保用户控制他们的数据，但该工作只讨论了数据的存储和共享问题。Zheng X等人[4]讨论了使用区块链共享个人健康数据的系统，使用户以符合通用数据保护法规的方式安全地控制和共享其个人健康数据，这项工作着眼于数据的采集和存储，重点引入了一个控制数据质量的方法。Goldfeder S等人[5]研究了使用区块链交易实物商品过程中的公平性、安全性和隐私性问题。The AdChain Registry[6]是一个基于以太坊的广告发行商注册表，为广告投放者提供被推荐的广告发行商，为特定的数据买家提供一组被推荐的数据源。FairSwap[7]则主要考虑在智能合约中实现公平交换的有效协议，它主要着重于数字商品公平性的实现。

也有一些文献研究使用区块链构建分布式的数据市场中的问题。Missier P等人[8]考虑了物联网的特征，构建了一个去中心化系统，并分析了其性能。Cao T D等人[9]则构建了一个实时人类感知数据的市场。然而这些系统往往没有考虑合适的定价机制，仍然以某种形式依赖于具有公信力的第三方，或者只是停留在理论的分析上，并没有完整的系统[10]。针对集中式数据市场的研究工作侧重于探讨定价机制的问题，对数据的收集、处理、拍卖各个过程有更细致的设计。Mun M等人[11]提供了一个个人数据保管库，提供了管理数据策略的机制，让个人用户可以细粒度地控制访问和共享数据。CrowdBC[12]是用区块链构建的群智感知系统，作者着重讨论了图片数据在交易中的特性和处理办法。Banerjee P等人[13]构建了一个基于区块链的数据市场，在买卖双方的交易中引入了一个可信中间人，这虽然使得买卖双方的交易变得更加简单，但也会使系统的安全性降低很多。虽然Gupta P等人[14]提出的交易仍需要第三方的介入，但是其在交易过程中设置了多个分布式的中介参与交易，限制了中介的垄断能力。Ramachandran G S[15]的创新在于其数据的传输和支付都在链下进行，这能节省区块链昂贵的存储空间。Liu K等人[16]则针对基于区块链的数据市场中的定价问题，设计了一种自动化的定价谈判机制。DDV[17]是一个出售个人病历的分布式数据交易框架，当数据卖方想要出售其病历时，他必须将其加密的数据上传到云存储服务提供商，并将其数据信息提交给区块链智能合约。DDV仍使用第三方云存储服务来存储数据卖方的数据。尽管在将数据上传到云存储之前已进行了加密，但是数据购买者和数据出售者仍必须信任第三方云服务，以实现数据的持久性和数据传递。

2.3 区块链概述

区块链概念于1991年由W.Scott-Stornetta提出，他在论文中描述了一种称为

“区块链”的数字体系结构系统。2008年，中本聪提出了比特币这一新型的“数字货币”[18]，其背后的区块链技术也得到了许多研究领域的广泛关注。

在区块链网络中，所有参与者本质上是一组不相互信任的编写者，他们共享着一个没有可信中间人的数据链。为了防止分叉现象在这个分布式环境中爆发，区块链设计了共识协议。区块链节点可以作为矿工提供计算资源，以竞争将事务记录到区块链中的权力，获胜者将得到经济激励。这是区块链达成共识的一种机制，叫作工作量证明（proof of work，POW）。

工作量证明使得理论上攻击者只有掌握了超过整个系统50%的算力才能攻破区块链系统。然而共识协议付出了消耗海量算力和资源的代价，近几年，全球用于比特币“挖矿”消耗的年用电量达到每年总电量的0.13%。每个完整节点都必须存储所有事务，以在区块链上验证这些事务的合法性。此外，由于块大小的限制和用于生成新块的时间间隔，比特币每秒只能处理7个事务，这不能满足实时处理数百万个事务的要求。

以太坊的概念在2013年由Buterin V[19]受比特币启发后提出，其最大的特性就是增加了对智能合约的支持，它会在用户的以太坊节点中运行一个虚拟机，其中运行的智能合约采用Solidity编写，该语言能够支持图灵完备性。为了满足在用户客户端的虚拟机中运行的需求，Solidity语言的功能被设计得很弱小，Solidity只以一种特殊的方法实现了JavaScrip函数中的一部分，使得智能合约在以太坊中比较容易出错。另外，在以太坊网络中进行的涉及状态更新的计算都需要消耗“天然气”（gas，采用Wei为单位，这是以太币的最小单位），这使得在以太坊网络中进行复杂计算变得不划算。

区块链技术为打造去中心化的数据市场、降低中间机构在交易双方之间的干预作用提供了新的方向。区块链具有一些特性：一是权力下放，在集中式系统中，每个交易需要通过中央可信的机构进行验证，这不可避免地提高了中央服务器的成本和性能瓶颈；二是其透明性与安全性，交易可以快速验证，诚实的矿工不会承认无效的交易，一旦交易被包含在区块链中，就几乎不可能删除或者回滚交易；三是匿名性，每个用户可以使用其地址与区块链交互，目前很多区块链系统致力于使区块链成为完全匿名的系统，如Monero币等。

由于这些特性，区块链目前具有广泛的应用前景。不同领域（如IoT[20]、智能交通系统[21]、命名和存储系统[22]以及健康记录共享[4]等）的应用都有基于区块链技术的实现。区块链的底层技术——星际文件系统（interplanetary file system，IPFS）[23] 是一种内容可寻址的对等超媒体分发协议，在分布式系统环境中也有广泛的应用价值。

3 区块链数据市场设计目标

在数据交易中，区块链系统替代了集中式数据市场的地位，买卖双方会直接就区块链中智能合约的执行进行交易。设计基于区块链系统的数据市场主要考虑如下几个问题。

- 去中心化：由于集中式数据市场的弊端，许多研究开始讨论建立一个去中心化的数据交易系统，让数据持有者与数据需求者通过安全可信的分布式系统直接进行交易。然而许多现存的相关研究与系统设计仍然在一些模块中依赖可信的第三方实体，而这些第三方实体有动机和能力通过破坏买方和卖方的交易获益。因此对于

去中心化的数据市场，笔者认为应构建一个不依赖任何可信第三方、只有数据买家与卖家参与的系统。交易直接由买卖双方达成，各种安全性、隐私性的需求由分布式交易系统的设计来实现。

- 公平性：公平性指的是买家和卖家在整个交易中的地位应是相同的，他们会对交易的数据和其价格达成共识，他们都具备随时停止交易的能力。最基础的公平性应在智能合约执行结束时实现，要么买家获得有效数据、卖家获得付款，要么买家和卖家均不获得任何收益，要防止数据卖家提供不合法数据、数据买家抵赖数据购买费用等不公平情况的出现。
- 隐私性：隐私性要求系统保护用户的身份隐私和数据隐私。身份隐私即数据市场中用户的匿名性，比特币区块链的匿名能力仅仅是化名的程度，有许多区块链系统致力于提升它们的匿名能力。数据隐私指数据只能由购买了这份数据的用户使用，攻击者无法从存储在区块链中的信息得到数据的任何额外信息。在很多场景中，数据消费者往往是购买一份数据的一次使用权，在这种系统中，数据隐私性要求数据买家在使用完数据后也无法获得数据的任何额外信息。
- 有效性：对于所有广泛实施的实际系统，要能够有效实施，需要保证参与者的使用体验，这对系统的执行提出了运行效率和资源消耗上的要求。在基于区块链的数据市场中，需要考虑区块链系统自身的执行速度能否跟上大数据交易的需求。由于区块链智能合约的特殊机制，应该尽量避免使用智能合约进行复杂计算。
- 经济激励：数据市场的一个主要目标是为所有参与系统的用户谋取利益，以激励他们参与到大数据共享与交易系统中。首先是买家和卖家都会在交易中获益，数据卖家能够通过出售自己数据的使用权得到尽可能多的经济收益，数据买家能够得到符合自身需求的高质量数据。同时，交易产生的手续费将激励其他参与者做好系统平台的维护工作。在定价博弈中，一些卖家和买家可能会串通，以攫取其他用户的利益，因此数据市场在提供经济激励的同时，还需要确保不出现经济套利等非法策略行为。

4 系统架构

本文设计的基于区块链的数据市场系统由以下3个组件构成：以太坊区块链中的一个智能合约、系统参与者持有的客户端和一个点对点的数据传输网络。当一个数据买家需要一些特定的数据来计算一个特定的任务时（如需要所处位置附近的温度传感器的数据计算当地室外温度），他会使用数据筛选模块，并通知智能合约。整个系统通过安全计算的方式定位到一些符合条件的数据，经过数据买家与数据卖家的交互后，数据定价模块会确定出售的数据及其价格。付款完成后，系统又以安全计算的方式运行买家的计算任务，并将结果返回给买家，交易完成。在交易进行的过程中，系统需要确保前文提到的设计目标，后文将详细介绍该系统的各个模块。

4.1 数据筛选

数据购买者通常不需要所有用户的所有数据，而是关心特定用户的特定数据。例如，在众包任务中，数据购买者想知道某位置的室外温度，他可以提出限制条件：卖方提供数据的传感器的位置需要在一定的范围内，并且提供的数据也需要满足一定的时效性。该过程被称为数据筛选，买家可以将自己的数据需求整理成一个逻辑表达或者一个数学函数存入区块链，以供卖家查询判断。

这些数据的筛选条件一般比较简单，因此将买家的数据筛选需求直接上传到区块链是不合适的，这显然暴露了买家和卖家的隐私。根据数据筛选的形式，攻击者很容易推断出买家的数据需求，从而获取买家和卖家的隐私。在上例中，如果买家感兴趣的一个位置被公开，攻击者可能会推断出买家的活动范围，同时，卖家持有的临近该位置的设备也被暴露。因此，为了保护系统参与者的隐私，应该在隐藏数据筛选条件的同时筛选出目标数据。最直观的解决方法是函数加密，即拥有解密密钥的用户（买家）可以获得密文数据的函数值（数据筛选函数），却不会获得有关明文的任何信息。

另外，将买家的数据筛选需求转化成数学逻辑表达有时并不直观。数据的交易有时也针对一些非结构化的数据，例如，买家需要一些“猫”的图片。买家也许能够判断一段数据是否是他想要的，但无法通过简单适当的逻辑来量化这些需求。为了给出买家筛选需求的逻辑表达，系统中需要运行一些复杂的数据处理算法来得到数据的质量以及与筛选需求的契合程度。由于需要对每一份潜在的数据进行运算，这些算法应尽量简化。

4.2 数据存储

数据存储机制是一个通用术语，用于描述如何推送数据以及将其存储在何处。主流公共区块链对区块中的交易数量和空间有限制，可以是区块的大小（比特币）或区块中消耗“天然气”的上限（以太坊）。对于数据交易市场而言，将海量的数据直接存储在区块链上是不可行的。

Quorum和Corda都是受区块链启发针对金融领域的平台，它们提出了一种不将数据公开存储在区块链上的模型。其中，数据由参与的第三方（金融机构）保持脱链状态，并且共识功能旨在确保交互各方达成协议。这种方法对于金融机构可能是实用的，但它违背了去中心化的设计目标。

尽管将完整的数据存储到区块链中是不现实的，但可以上传与特定的数据绑定的“数据摘要”。因此，一般提出的数据市场针对的是非实时和可容错的交易模式，这些系统可以在较长的时间间隔内将数据发送到数据后端，此方法需要分布式文件存储层的辅助。IPFS和Swarm是2个主要的分布式文件存储层。这2种技术都是点对点（P2P）技术，具有分布式的文件传输系统，其中文件通过其内容的哈希值进行寻址。当数据成功存储在IPFS中时，用户将收到一个哈希索引，这将允许用户以后检索该文件。这个索引将代替数据存入智能合约，节省整个系统的负担。这些分布式文件系统也是公开透明的，其中存储的数据应是加密的。同时，如果所有参与系统的用户都维护一个IPFS或Swarm节点，其代价是很高的，可以让一部分系统参与者作为分布式文件存储的服务提供者，向上传数据的用户收取一定的费用。

4.3 数据定价

在数据市场中，数据售卖形式的设计和价格的设定一直是一个活跃的研究领域。本文在博弈论环境中考虑定价机制的设计，每个数据持有者对于他们的数据都有一个私人的估值，即数据持有者隐私被泄露出去的损失；数据购买者对于他将购买的数据也有一个估值，即这份数据对于购买者的价值。其中出价者可能选择不诚实地报告他们对一份数据的估值。这会给交易的机制设计带来麻烦，博弈论中的解决方法是设计激励兼容的机制，让每个出价者都在报告其真实估值时得到最高收益。出价者报告其真实估值能够让设计定

价机制变得简单很多。

多份相关数据的捆绑定价也是数据市场交易中常见的问题。早期研究通常简单地假设对于数据买家，捆绑数据的期望价值等于所有数据的单独价值的总和。后面的研究发现，数据之间相互影响，每份数据的价值依赖于整个交易数据集的内容，数据集价值的体现来自数据的相互关系。

数据作为一种商品有一些独特的性质，这些性质使得数据定价需要考虑额外的一些问题。一是数据的边际成本极低，或者说根本没有边际成本。边际成本指在一个商品生产后，复制一份需要的代价。数据基本为零的边际成本使得一旦数据买家获得了数据卖家的数据，那么他就可以随意处置、随意出售这份数据。二是数据的价值和数据量没有必然联系。比如对于一个需要一些“猫”的图片的人，一堆关于“狗”的图片几乎没有价值。三是数据价值的量化。数据的价值难以量化，数据持有者很难估计出数据的价值，同时不同人之间的估值也大相径庭。

针对这些数据的特性，数据市场中的数据交易模式也发生了变化。传统的数据市场会直接交易用户的数据，而不诚实的买家可以在卖家不知情的情况下转售自己已购买的数据集，从而获得利益。许多工作发现绝大多数数据消费者仅仅需要在大量数据背后的一些统计结果或高级特征，比如计算数据集的平均值，或者为机器学习模型训练数据，而不是需要数据本身。因此，数据市场可以从数据持有者处收集数据，然后为数据消费者手中的计算任务服务。买家提供一个特定的任务，其输入是其购买的多份数据的一次使用权，而输出是买家想要的结果。这样，数据本身就与数据消费者隔离了。

在上述设计中，对数据价值的评判实际上变成了对买家计算任务的结果准确度或者其计算结果对买家的价值的评判。虽然数据本身的价值难以量化，但任务结果的提升容易量化。由于买家需要的往往是多个卖家的数据，所以在交易中还需要区分每份数据单独的价值。在计算数据价值时，Shapley值可以用于计算单份数据的贡献。在博弈论中，计算Shapley值是一种将收益和成本公平分配给参与合作的多个参与者的一种解决方案。Shapley值的计算复杂度随着数据数量的增长量指数级增长，因此在实际使用中常常采用近似算法。

4.4 安全计算

区块链是一种公开透明、去中心化的数据存储技术，所有进入区块链系统的信息都是公开的，所有交易或者脚本的执行都是透明的。同时，区块链智能合约的计算能力非常弱小，由于受区块大小和gas限制，调用一次智能合约能进行的计算很少，但代价很高。因此，无论是在智能合约中直接运行买方发布的任务（智能合约几乎无法承担这样的负担），还是在区块链外交给某个不受信任的个人，都是不安全的，同理，数据筛选过程和数据定价过程也存在这样的问题。比特币、以太坊等区块链公共链中的节点之间互不信任的同时又完全公开透明的特性，使得隐私保护工作产生了新的问题。

本文的目标是在整个数据市场系统和市场系统内的交易过程中，数据市场不泄露任何关于用户及其数据的额外信息。有许多方法可以安全、正确地执行系统中的计算工作，同时保护各方的隐私。安全计算的作用就是在保护隐私性、公平性等特性的同时完成各种计算任务，不同领域的研究都能不同程度、不同角度地达成目标。

密码学是实现安全计算最直观的方法。安全多方计算（secure multi-party

computation，MPC）是密码学中的一个子领域，也是安全计算问题的直接解决方案。MPC的目标是为各方创建联合计算函数的方法，同时保护这些输入的私密性。与传统的加密方法不同，密码学技术保证了通信或存储的安全性和正确性，而这种模式的密码学侧重于保护参与者之间的隐私。MPC现在可以被看作各种现实问题（特别是那些只需要简单的线性共享秘密的问题）的实际解决方案，例如分配投票、私人竞标和拍卖、共享签名或解密功能、私人信息检索等。许多购买者的简单任务可以通过MPC轻松地执行。然而，MPC方法不能用于深度学习任务。主流MPC框架的核心使用2种加密技术：加密电路和不经意传输。MPC将买家的任务函数转换成一个乱码电路，然后以不经意传输的方式发送出去。在复杂繁重的计算任务中，将深度神经网络转换成乱码电路势必会增加计算量，并失去一定的精度。同时，在MPC中，深度学习可能导致不可接受的通信复杂度。此外，同态加密和零知识证明也可以用于一些简单的任务。

实现安全计算的另一种方法是使用可信硬件，如可信执行环境（trust execution environment，TEE）。TEE是一个通用概念，是主处理器的一个安全区域，它保证装载在其中的代码和数据在保密性和完整性方面受到保护。假设一部分用户拥有一些TEE硬件，则将这些用户视为安全用户。卖方将发送他们的数据到安全用户的TEE设备，买方的计算任务将在TEE中进行计算，并以安全的方式将结果返回买方。只有在TEE中运行的受信任的应用程序才能访问设备的主处理器、外围设备和内存的全部功能。硬件隔离保护数据和计算内容不受在主操作系统上运行的用户安装的应用程序的影响。支持TEE实现的典型硬件技术是ARM TrustZone和Intel SGX Software Guard扩展。SGX是Intel体系结构的扩展，可以在硬件级别保护应用程序的执行。例如，SGX可用于云计算环境中保护购买者数据，示例包括VC3[24]和Haven[25]。SGX技术的核心是在内存中隔离出一块特殊区域（称为“飞地”），指定程序可以在该区域中创建“安全区”，并将关键代码和数据存储到“安全区”中。只有CPU或程序本身可以访问“飞地”中的代码和数据。这些现有的硬件技术都被声称存在漏洞。而一些开源项目和大公司也正致力于使TEE更强大。

联合学习是谷歌公司提出的一种不需要集中训练数据的新型协同机器学习方法。它的工作流程是：数据卖方从买方提供的云上下载当前的模型，并通过从卖家数据中学习来改进模型，然后将更改聚合为一个小更新。模型的这个更新使用加密通信发送到云，在云中更新的模型立即与其他用户的更新进行聚合，以改进共享模型。因此，所有的训练数据都只保存在卖方的设备上，没有单独更新存储在云中。联合学习通常针对用户拥有多个数据或一个数据集的情况，不适用于单个卖家仅提供少量数据的情况。如果卖方只提供了一条数据参与训练，恶意买方可能有能力将卖方的真实数据从模型更新中进行逆推计算，且数据的训练过程需要频繁的通信，效率非常低。

5 系统实现与验证

5.1 系统实现框架

根据第4节设计的数据市场框架，笔者实现了一个基于以太坊私链的数据市场系统，包括系统参与者持有桌面应用客户端、以太坊网络中的智能合约以及数据传输网

络。桌面客户端采用JavaScript编写，以太坊智能合约使用Solidity编写，数据传输则直接使用IPFS的JS接口。笔者设计的系统架构以及各部分组件间的交互如图2所示。

笔者实现了预期系统的一个简化版本，以便于进行系统正确性与可行性的验证与分析。笔者将系统简化为一个买家和多个潜在的卖家进行交易，一个买家只需要一份数据，定价机制采用第二价格拍卖，即成交的数据采用卖家出价最低的数据，而数据按出价第二低的价格付款。买家的计算任务被设置成简单的形式，可以使用同态加密或安全多方计算来保证安全计算。笔者在系统中实现了2种具体的数据形式的交易，非结构化的图片数据和结构化的GPS行程数据。

数据市场的交易过程如图3所示。数据卖家在智能合约中添加一份新的数据信息，数据买家添加一个包含数据需求的订单，卖家根据订单中的条件，在匹配到合法数据后给出一个报价，系统会运行定价机制并计算出获胜的数据，这份数据会通过安全计算的方式计算买家的任务，同时交付数据。在系统中，用户可以进行的操作如下。

● 注册账号。要在数据市场中买卖数据，用户首先必须注册一个以太坊账号，客户端中也包含一个简单的以太坊账号管理功能。同时用户还需要在笔者编写的智能合约（以下简称DDM）中注册账号，比如用户可以在智能合约中注册一个传感器设备的账号，其中需要公开传感器设备的类别、型号等信息。

● 添加数据。数据持有者在生成了一些他想要出售的数据后，可以将数据加密上传到IPFS，同时在智能合约账号中注册这份数据，其中包含数据的存储地址、数据的哈希值和注册时间等。在笔者的系统设计中，鼓励数据买家在添加数据订单以购买数据时设置数据注册时间要求，只有满足数据注册时间要求的数据持有者才有资格竞标该订单，以提升数据的时效性和可靠性。

● 发布订单。若数据买家想要在市场中购买特定数据的一次使用权，那么他会在DDM中添加一个订单，订单中包含买家对数据的需求（包括数据类型、数据选择函数、数据价格限制、数据数量限制等）和买家使用数据的计算任务。

● 数据卖家提供其出价的哈希值。数据卖家选好一份或者多份符合数据买家需求的数据，并定好一个打包的价格。由于区块链完全公开透明的性质，直接在DDM中出价会导致套利发生，后报价的卖家可以看见其他卖家的出价，并通过控制自己的出价来获得更高的利润。笔者在系统中设置了一个时间窗口，卖家需要在窗口内提交其出价的哈希值，这样相当于实现了一个密封拍卖。同时这样的设计有助于实现公平性。

● 数据买家通知DDM获取市场数据卖家的真实出价。数据买家通知DDM停止接受用户参与交易，并开始接受已参与交易的用户公开他们的出价。智能合约无法主动执行命令，需要区块链用户触发，同时数据买家可以将一些卖家列入黑名单，拒绝他们的出价。

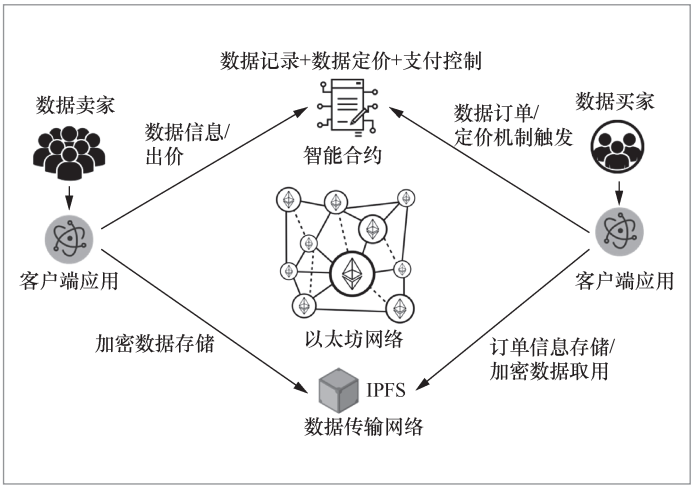


图 2 系统架构以及各部分组件间的交互

● 数据卖家公布真实出价。数据卖家公布其出价，该出价需要与之前的哈希值吻合。如果数据卖家没有在规定时间内公布真实出价，说明卖家后悔了，则系统会降低该卖家的信誉值，以作为惩罚，成功交易数据的卖家的信誉值会提升。信誉值过低的用户将无法参与交易，信誉值体系可以抑制系统参与者的作弊行为。

● 定价机制。笔者采用经典的维氏拍卖（Vickrey-Clarke-Groves auction）机制用于数据挑选和价格决定。VCG拍卖机制保证了真实性，因此每个投标人都有动机对自己的个人数据进行真实估价。

● 安全计算/交付数据。系统会采用同态加密的方式计算买家的计算任务，然后将加密的结果交给数据买家，订单完成。由于现在安全计算仍处于不成熟的阶段，同态加密的计算复杂度非常高，笔者仅测试了使用加法同态算法来计算买家的计算任务。在未来，笔者会支持第4节中安全计算方法中其他的方法。数据买卖方也可以直接交易数据，采用混合加密的方法，将数据的对称密钥用买家提供的公钥进行加密，这种方式可以在公开透明的环境中实现无交流的数据传递。

5.2 系统分析与验证

笔者从构建区块链数据市场的设计目标来讨论实现的演示系统的优势，包括公平性、隐私性和有效性，同时，还考虑了系统的可扩展性。笔者发现，该系统基本实现了区块链数据市场定义的设计目标，整个系统运行流畅，可用性较高，安全隐私性得到了很好的保证。

5.2.1 公平性

在交易结果达成之前，即定价机制开始执行前，交易双方都可以终止交易，虽然这可能会导致参与者的信誉值降低，但在“后悔”不常发生的情况下，能够确保交易的随时终止。同时，在定价机制开始执行后，由于区块链和安全计算的特性，交易双方都无法阻止交易的继续进行，以太币转移和买家任务计算就一定会执行，从而确保了数据交易的不可抵赖性，保证了交易公平。

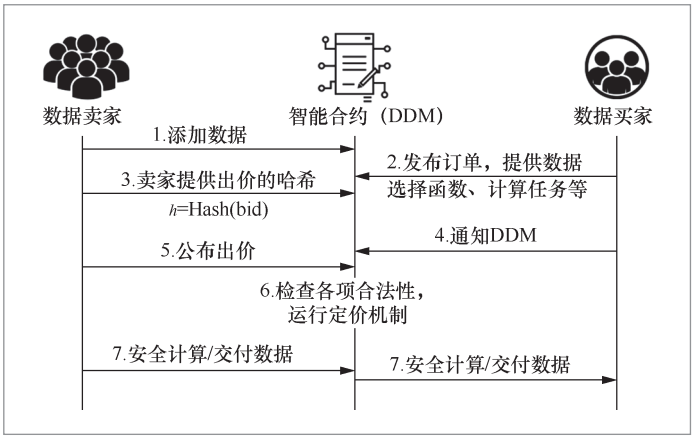


图 3 数据市场交易流程

5.2.2 隐私性

笔者在系统中保护了系统参与者的身份隐私和数据隐私。身份隐私是基于区块链的匿名能力实现的，而数据隐私则得到了全方位的保护，数据加密、数据选择机制和安全计算使攻击者无法从交易的执行中得到数据的任何额外信息。然而笔者实现的系统中没有考虑暴露卖家对自己数据定价所泄露的隐私，未来的工作中将会考虑出价隐私。

5.2.3 有效性

区块链的运行效率问题一直是其被诟病的地方，比特币的一份交易从完成到最后被写入区块链中需要十几分钟，而这份交易在网络上最后被确认的时间甚至超过一个小时。在客户端中计算数据选择函数，以安全计算的方式计算买家的任务以

及图片的处理的计算，都是比较耗时的，这会降低用户使用该应用程序的体验满意度。经过应用程序的随机点击测试发现，从客户端用户的角度，平均一个按钮的响应时间为4 s左右，这说明该系统目前的设计尚无法达到实时，但可以保证基本使用需求。笔者提出的系统直接基于以太坊实现，提高数据市场系统的可用性可以采用更加高效、轻量级的区块链系统。

在以太坊区块链中部署合同或任何类型的交易都会产生交易费用。在以太坊区块链中，使用气体估算成本。通常，系统中的矿工会确定运行一些指令的价格，而一个事务的发起者（比如数据买家向区块链添加订单、数据卖家向区块链提交出价）需要规定可以消耗气体的最大值。由于矿工喜欢处理提供高激励的事务，因此气体价格较低的交易可能需要很长时间。然而由于个人数据的价值很低，因此在智能合约中复杂的计算会使得买卖双方的收益直线降低。笔者提出的系统通过减少智能合约的计算量来保证系统参与者的利益。

表1 系统中各种操作的气体消耗

操作	气体消耗/Wei
在智能合约中注册账号	110 726
向智能合约中存钱	27 638
从智能合约中取钱	35 753
添加一个图片数据订单	2 228 844
添加一个行程数据订单	715 103
向订单提交出价	29 145
向其他账号转账	21 000
添加一份行程/图片数据	193 852
计算定价机制	459 230
平均值	424 587

表2 不同数组长度下的气体消耗

数组长度/维	气体消耗/Wei
50	1 466 256
100	2 228 844
125	2 542 012
150	2 885 446
175	3 243 954
200	3 632 920
225	3 991 512

通过多次测试，笔者得到了该系统中各种操作消耗气体的平均值（见**表1**）。根据现在的以太坊公链上以太币的价格（每个以太币价格约250美元）计算，平均每个操作的消耗是$424\ 587\times10^{-18}\times250\approx1.06\times10^{-10}$美元，即便一个数据买家需要成千上万份数据，他参与系统所需要的消耗也非常少。

智能合约中可能存储着许多与交易有关的信息，以太坊的智能合约虽然支持多种复杂数据类型，但实际中稍微复杂的结构会导致请求失效。比如由于要上传的数组长度很长，每一项也是很复杂的小数，1 024维的数组根本无法被成功地上传到智能合约。以太坊还有一个区块气体限制，由以太坊默认设置为8 000 000 Wei，因此事实上无法上传维度过高的数据（见**表2**）。由于以太坊对区块气体量的限制，由以太坊默认设置为最高8 000 000 Wei，如果某一项操作消耗的气体量超过了这个值，这项操作就会被回退，因此在DDM中添加订单就存在被回退的可能。笔者提出的系统将复杂的数据结构（比如数据选择函数）存储在IPFS中，再将其地址放入订单中存入DDM。

由于数据定价是通过智能合约在以太坊区块链上执行的，因此其计算成本不可被忽略。为了评估该系统的数据定价成本，以基于VCG拍卖的数据定价测试大量卖家出价的情况下气体的消耗。**图4**显示了VCG数据定价的耗气量与数据订单收到的出价数量之间的关系。尽管可以进一步优化数据定价算法的实现，但是当大量数据卖方竞标同一数据订单时，数据定价成本将非常高。如果要实现基于以太坊区块链的数据市场，那么高昂的数据定价成本将成为该系统应用的一大障碍。

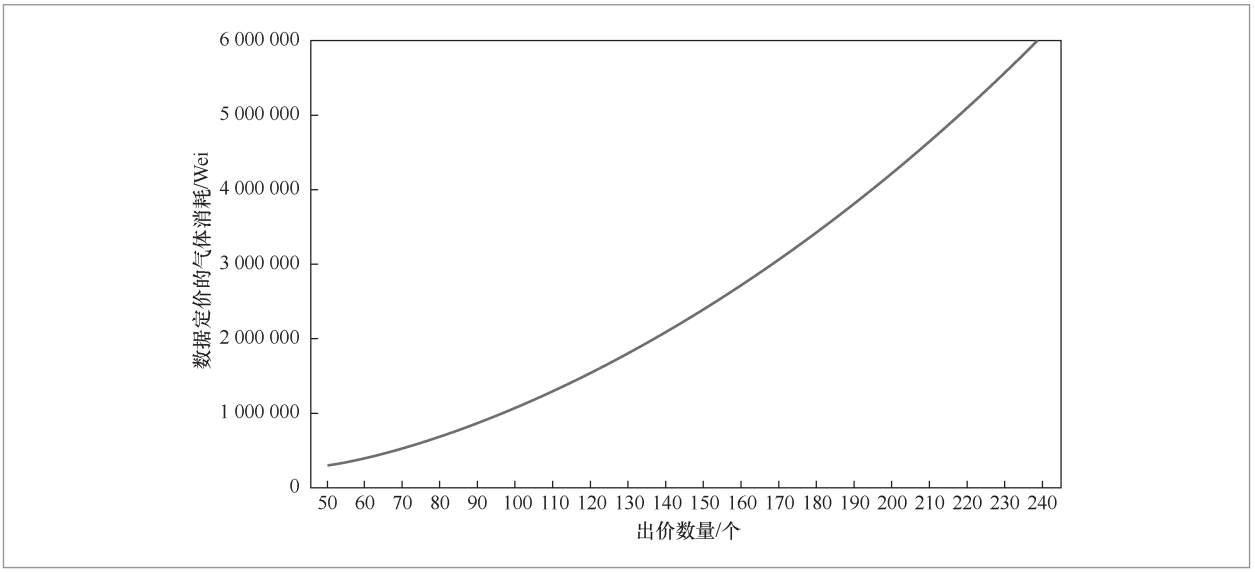


图 4 数据定价的 Gas 消耗与订单收到的出价数量关系

5.2.4 去中心化与可扩展性

“可扩展性三难”是指区块链系统在可扩展性、去中心化和安全性这3个方面的不可避免的矛盾。该系统是一个完全去中心化的系统，不依赖任何可信第三方，因此可扩展性和安全性将受到更大的挑战。为了保证安全性，笔者提出的系统牺牲了可扩展性，如果要支持不一样的数据形式、数据贩卖方式和定价机制，就要重新编写智能合约，而兼容性也将存在问题，这些问题都可以在实际数据市场的构建中被纳入考虑。在未来数据市场的部署中，笔者将进一步就可扩展性、去中心化和安全性这3个方面的权衡做更多的探讨。

6 结束语

数据是数据驱动型经济中的重要资产，它推动了新的数据交易行业的兴起。数据市场是当今数据资产化的一个重要形式，分布式的数据市场具有中心化数据市场不具备的隐私保护能力和交易安全性保障，有巨大的市场前景和研究前景。在本文中，笔者讨论了未来数据市场应满足的特性，分析了基于区块链的分布式数据市场中存在的挑战，提出了初步的解决方案，并探讨了这些技术未来可能的发展方向，为实际数据市场的构建提供了参考性意见。数据市场与区块链技术都处于技术快速发展的阶段，需要研究者根据这些问题进行更深入的探索与研究。

参考文献:

[1] PANG J Z, FU H, LEE W I, et al. The efficiency of open access in platforms for networked Cournot markets[C]// IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2017: 1-9.

[2] BALAZINSKA M, HOWE B, SUCIU D. Data markets in the cloud: an opportunity for the database community[J]. Proceedings of the VLDB Endowment, 2011, 4(12): 1482-1485.

[3] ZYSKIND G, NATHAN O, PENTLAND A S. Decentralizing privacy: using blockchain

to protect personal data[C]// 2015 IEEE Security and Privacy Workshops. Piscataway: IEEE Press, 2015: 180-184.

[4] ZHENG X, MUKKAMALA R R, VATRAPU R, et al. Blockchain-based personal health data sharing system using cloud storage[C]// The 20th IEEE International Conference on e-Health Networking, Application & Services. Piscataway: IEEE Press, 2018: 1-6.

[5] GOLDFEDER S, BONNEAU J, GENNARO R, et al. Escrow protocols for cryptocurrencies: how to buy physical goods using Bitcoin[C]// Financial Cryptography and Data Security. Heidelberg: Springer, 2017: 321-339.

[6] GOLDIN M, SOLEIMANI A, YOUNG J. The AdChain registry[Z]. 2017.

[7] DZIEMBOWSKI S, ECKEY L, FAUST S. Fairswap: how to fairly exchange digital goods[C]// The ACM Conference on Computer and Communications Security. New York: ACM Press, 2018: 967-984.

[8] MISSIER P, BAJOUDAH S, CAPOSSELE A, et al. Mind my value: a decentralized infrastructure for fair and trusted IoT data trading[C]// The 7th International Conference on the Internet of Things. [S.l.:s.n.], 2017: 1-8.

[9] CAO T D, PHAM T V, VU Q H, et al. Marsa: a marketplace for realtime human sensing data[J]. ACM Transactions Internet Technology, 2016, 16(3): 1-21.

[10] SUBRAMANIAN H. Decentralized blockchain-based electronic marketplaces[J]. Communications of the ACM, 2017, 61(1): 78-84.

[11] MUN M, HAO S, MISHRA N, et al. Personal data vaults: a locus of control for personal data streams[C]// The 2010 ACM Conference on Emerging Networking Experiments and Technology. New York: ACM Press, 2010: 1-12.

[12] LI M, WENG J, YANG A, et al. CrowdBC: a blockchain based decentralized framework for crowd sourcing[J]. IEEE Transactions on Parallel and Distributed Systems, 2019, 30(6): 1251-1266.

[13] BANERJEE P, RUJ S. Blockchain enabled data marketplace - design and challenges H[J]. Computer Science, 2018, arXiv:1811.11462.

[14] GUPTA P, KANHERE S S, JURDAK R. A decentralized IoT data marketplace[C]// The 3rd Symposium on Distributed Ledger Technology. [S.l.:s.n.], 2018.

[15] RAMACHANDRAN G S, RADHAKRISHNAN R, KRISHNAMACHARI B. Towards a decentralized data marketplace for smart cities[C]// The 4th IEEE Annual International Smart Cities Conference(ISC2). Piscataway: IEEE Press, 2018.

[16] LIU K, QIU X, CHEN W, et al. Optimal pricing mechanism for data market in blockchain-enhanced Internet of things[J]. IEEE Internet of Things Journal, 2019, 6(6): 9748-9761.

[17] ZHOU J Y, TANG F Y, ZHU H, et al. Distributed data vending on blockchain[C]// The 11th IEEE International Conference on Internet of Things. Piscataway: IEEE Press, 2018: 1100-1107.

[18] NAKAMOTO S. Bitcoin: a peer-to-peer electronic cash system[Z]. 2008.

[19] BUTERIN V. Ethereum: a next generation smart contract and decentralized application platform[Z]. 2014.

[20] CHRISTIDIS K, DEVETSIKIOTIS M. Blockchains and smart contracts for the Internet of things[J]. IEEE Access, 2016, 4: 2292-2303.

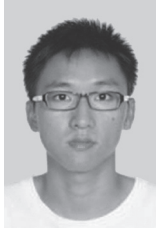
[21] YUAN Y, WANG F Y. Towards blockchain-based intelligent transportation systems[C]// IEEE 19th International Conference on Intelligent Transportation Systems. Piscataway: IEEE Press, 2016: 2663-2668.

[22] ALI M, NELSON J, SHEA R, et al. Blockstack: a global naming and storage system secured by blockchains[C]// 2016 USENIX Annual Technical Conference. Berkeley: USENIX Press, 2016: 181-194.

[23] BENET J. IPFS: content addressed, versioned, P2P file system[J]. Computer Science, 2014, arXiv:1407.3561.
[24] SCHUSTER F, COSTA M, FOURNET C, et al. VC3: trustworthy data analytics in the cloud using SGX[C]// 2015 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2015: 38-54.
[25] BAUMANN A, PEINADO M, HUNT G. Shielding applications from an untrusted cloud with haven[C]// The 11th USENIX Symposium on Operating Systems Design and Implementation, Berkeley: USENIX Association, 2014: 267-283.

作者简介

汪靖伟(1997-),男,上海交通大学计算机系硕士生,主要研究方向为区块链应用与数据交易。



郑臻哲(1989-),男,博士,上海交通大学计算机科学与工程系助理教授,主要研究方向为算法博弈论、移动计算、机器学习。

吴帆(1981-),男,博士,上海交通大学计算机科学与工程系教授,中国计算机学会专业会员,主要研究方向为网络经济学、无线网络、移动计算、隐私安全。

陈贵海(1963-),男,上海交通大学教授、博士生导师,主要研究方向为对等计算、数据处理、传感器网络、路由算法、高性能计算机体系结构、组合数学等。发表的文章被Google Scholar引用12 000余次,SCI引用1 000余次,ESI高被引论文4篇,11次获得国际会议最佳论文奖。2008年获国家杰出青年科学基金,2011年获国务院政府特殊津贴,2015年获教育部自然科学奖一等奖(排名第一),2017年入选中国计算机学会会士(CCF Fellow),2018年获江苏省科学技术奖一等奖(排名第一)。现任中国计算机学会分布计算与系统专业委员会主任,ACM SIGCOMM China副主席。

收稿日期: 2020-01-31
通信作者: 吴帆,fwu@cs.sjtu.edu.cn
基金项目: 国家自然科学基金资助项目(No.61972252,No.61972254,No.61672353,No.61672348,No.61902248);国家重点研发计划基金资助项目(No.2019YFB2102200);装备预研教育部联合基金资助项目(No.6141A02033702);阿里巴巴创新研究计划
Foundation Items: The National Natural Science Foundation of China(No.61972252, No.61972254, No.61672353, No.61672348, No.61902248), The National Key Research and Development Program of China(No.2019YFB2102200), Joint Fund of Ministry of Education for Equipment Pre-Research(No.6141A02033702), Alibaba Innovation Research Program