链上存证、链下传输的可信数据共享平台

张召[1]，田继鑫[2]，金澈清[1]
1. 华东师范大学数据科学与工程学院，上海 200062；2. MCT Technology，上海 200023

摘要
区块链系统可以为分享数据的互不信任的多方之间提供可信的基础设施。但是，将原始分享数据直接上链的方式并不适合大规模的数据分享场景。因此，提出了一种数据共享请求和应答记录上链存证、原始数据链下安全传输的数据共享平台架构，该架构在一定程度上可以缓解系统负载过重以及隐私保护方面的问题。最后总结了随着参与节点的增多，以及每秒需要处理的数据共享请求和应答的增多，已有的区块链技术被应用到数据分享和确权领域时，在分布式存储、共识协议、智能合约执行以及轻客户端查询方面面临的挑战以及改进的方向，以期为已有区块链系统应用于数据共享领域指明需要进一步突破的技术瓶颈。

关键词
数据共享；数据确权；数据追溯；区块链

中图分类号：TP315　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2020047

On-chain witness and off-chain transmission trustworthy data sharing platform

ZHANG Zhao[1], TIAN Jixin[2], JIN Cheqing[1]
1. School of Data Science and Engineering, East China Normal University, Shanghai 200062, China
2. MCT Technology, Shanghai 200023, China

Abstract
Blockchain system can build a trusted infrastructure for sharing data between multiple untrusted parties. However, directly uploading original shared data to blockchain is not suitable for large-scale data sharing scenarios. A data sharing architecture where data sharing request and response records are deposited on-chain and original data is transmitted securely off-chain was proposed. The architecture can alleviate the problems of system overload and privacy protection to a certain extent. Finally, with the increase of participating nodes and the data sharing requests and responses to be handled per second, the limitations in distributed storage, consensus protocol, smart contract execution, and query from light clients, directions for further research were proposed, in order to specify the technical bottlenecks that need to be further broken for the existing blockchain system applied to the field of data sharing.

Key words
data sharing, data property rights, data tracking, blockchain

1 引言

随着互联网技术的发展，以及数据和关键信息的采集、传输、存储和处理的自动化，越来越多的数据信息以电子资源的形式记录和存储，这些基础数据是企事业单位的核心数字资产，可以为各行各业的决策支持和精准营销奠定数据基础。然而，由于不同企业或者不同政府部门之间缺乏互信，鉴于数据泄露以及不正当使用的风险，以及企业之间或者政府部门之间行政利益的不同，很多数据拥有者不愿意共享数据，从而形成一个个数据孤岛[1]，数据的价值无法得到应有的利用，对数据资源造成了极大的浪费。

1.1 数据共享模式

数据流通过程涉及的主体包括数据生产者、数据收集者、数据使用者、数据处理者和数据监管者等。为了打通业务流程，更大程度地发挥数据隐藏的价值，通过数据共享让数据流通起来是一个非常有效的方法。根据数据共享应用的业务场景，数据的共享模式可以分为如下3类。

（1）数据不离开私有域，通过授权实现远程访问共享

该模式下，基于某种业务逻辑，需要访问多个数据提供者的共享数据，其基本特点是按需共享。这一般属于协同业务，对于共享的数据，参与方一般预先签订授权或者法律法规授权，根据业务的需要，随时访问共享数据。如对新型冠状病毒肺炎确诊患者居家隔离的监控及活动轨迹的流控，授权机构可以通过随时访问授权用户的相关数据（如手机用户的移动轨迹、支付平台的消费地点、监控数据等）来实现。

（2）数据离开私有域，通过数据移动汇聚实现在汇聚点上的集中数据共享

该模式下，由多个数据提供方提供的数据经规范化处理、汇总、分析后，形成新的共享数据。比较典型的应用是征信平台，其从各类银行类金融机构、公共事业、保险公司、支付平台获取企业或个人的信贷信息、支付信息、交易信息，经过汇总、处理后形成企业或个人的信用信息，信用信息可供各类授权企业或个人访问，并作为业务的参考依据。

（3）数据离开私有域，并且所有权也随之发生转移，在此过程中需要对数据进行确权

共享数据交易是该模式的典型应用，其基础是数据确权，在确权的基础上，共享数据的某些权利发生转移，同时数据提供方获得经济利益。交易的进行需要双方或多方的认可，并且共享数据的获得方必须在合约规定的权限内使用共享数据。如果数据获得方需要将共享数据交易给其他第三方，必须得到原权利人的许可。

从以上3个典型的数据共享模式可以看到，在数据共享流通的过程中，为了避免数据隐私泄露和数据滥用等问题，在共享过程的多个参与方之间建立互信的协作关系是非常必要的。作为一种由互不可信的多方共同维护的分布式账本，区块链具有防篡改、可追溯、去中心化的特点。区块链技术支撑的数据共享流通可以保证数据从收集到使用、共享乃至销毁的过程都公开透明、有据可查，并可以通过追溯问责的方式来避免数据共享过程中某一参与方的消极怠工问题[2]。而数据存储、处理和共享流通等过程公开透明可查使得企业或者部门赖以决策的数据来源更可信，从而使得决策结果更精确。

1.2 区块链系统的相关知识

起源于比特币的区块链技术是一种按照区块产生的时间顺序将区块以密码学哈希顺序相连的链式数据结构，其中密码学哈希链接方式可以保证数据的难以篡改和难以伪造，拜占庭容错的共识协议保证在存在恶意节点的网络环境下，数据仍能在多节点间达成一致。区块链系统中包含交易（transaction）、区块（block）和链（chain）3个基本概念。其中，交易是指对账本的操作导致的账本状态的改变，可以是一次转账或者一次智能合约调用；区块中记录了一段时间内发生的所有交易和状态结果，是交易执行的基本单元，是多节点间对当前账本状态一致性的一次共识；链是由区块按照生成顺序以密码学哈希链串联而成的，可以被理解为整个账本状态变化的日志记录[3]。

图1是一个节点包含的简单的区块日志数据和状态数据的示意图。系统以追加的方式记录了6条交易日志，假定每个区块只包含1条交易（如**图1（a）**所示，假设A、B和C的账户初始余额来自外部系统），需要注意的是，以未花费的交易输出（unspent transaction output，UTXO）模型为代表的比特币系统如果要查询数据的当前状态，必须重做所有6条交易日志，才能得到A、B、C 3个账户的当前余额（如**图1（b）**所示），即数据的当前状态。而以账户模型为代表的以太坊以及以无账户模型为代表的超级账本Fabric则同时存储和维护区块日志数据和状态数据。

1.3 基于区块链的数据共享模式

作为一种难以篡改、历史数据可追溯的分布式账本，区块链系统虽然可以为不可信的多个参与方提供方便的数据共享服务，但如果直接将需要共享的数据上链，通过区块链系统在多方之间分享，则会带来如下几个问题。

- 在全复制数据分布下，共享数据的存储开销太大。在区块链系统中，为了避免恶意节点对数据的篡改，每个节点都存储一份完整的数据副本，如果将所有共享数据上链，则很快会突破单个节点的存储容量上限。
- 在对等（peer to peer，P2P）网络模式下，网络传输代价太大，影响了系统吞吐量。在区块链系统中，所有节点都采用对等模式组网。在由n个节点组成的区块链网络中，任何一方待共享的原始数据均要通过P2P网络将数据传输至其他$n-1$个节点，节点之间传输太大的数据会导致网络带宽资源的急剧减少，并大大降低系统的吞吐量。
- 共享数据采用明文记录方式，导致

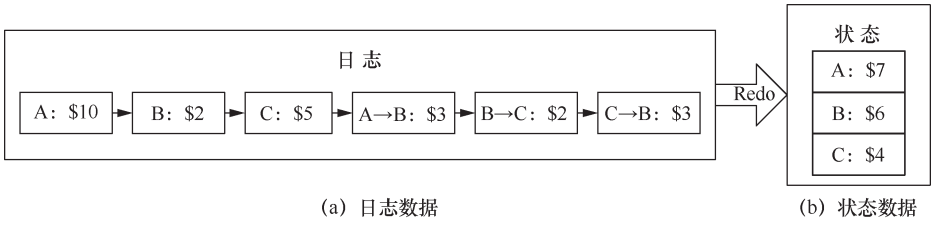

图1 日志数据和状态数据示意

数据的隐私得不到保护。在当前大多数区块链系统（如以太坊和超级账本Fabric）中，仅采用密码学签名来防止交易被篡改，交易内容仍以明文存储，所有参与者都可以看到，但这种依赖区块链的数据共享方式会带来隐私泄露的风险。

因此，考虑到数据共享流通系统的计算、存储、网络开销，以及数据隐私保护方面的问题，基于现有的区块链系统，将全部待共享数据直接上链的方式并不能满足大规模的数据共享需求。本文针对数据共享的应用场景特征，提出了一种链上共享存证、链下数据传输的基于区块链的新的数据共享架构。

2 区块链系统及关键技术

2.1 区块链系统

不同于只支持有限脚本的比特币系统[4]，随着以支持智能合约为代表的区块链2.0平台的出现，区块链技术不仅可以为数字货币领域提供服务，也可以为包括数据共享场景在内的很多传统业务提供服务。其中，任何业务逻辑都可以被编程为智能合约，并以去中心化应用（decentralization application，DApp）的方式公开透明地部署到不同的节点上[5]，所有的数据和智能合约代码通过全复制的方式在不同节点间实现共享，通过共识达成一致。而通常所说的区块中的一笔交易就是对智能合约相关功能的一次调用。

图2所示是一个典型的区块链系统，包含A、B和C 3个全节点（保存完整区块链数据），以及A’、B’和C’ 3个轻客户端（只保存区块头），其中全节点A和B各自有一个隶属于自己的轻客户端A’和B’，而轻客户端C’不隶属于任何全节点。因为各自立场的不同，A、B和C 3个全节点之间并不完全互相信任。全节点既可以发起交易，也可以接收交易（无中心化），交易被操作者签名后在P2P网络中传播，最终交易被共识协议所确定的主节点成批打包成区块，继续通过P2P网络传播给其他验证节点，验证节点确认无误后，在本地以区块为单位记账，其中区块与区块之间通过密码学哈希指针连接[6]。这样，每个节点都能保存一份完整的区块链数据（全复制分布）。这种交易带签名、成批打包进区块，区块以哈希链方式追加存储，且最终采用全复制方式分布的数据存储模式保证了交易数据的难以篡改和可追溯。

2.2 从数据管理的角度看区块链系统涉及的关键技术

从第2.1节的叙述可知，区块链系统是一种全新的分布式基础架构与计算范式。在单节点上，区块链系统使用密码学哈希链串联的链式数据结构来验证与存储数据；在节点间，每个节点独立保存完整的区块数据，利用分布式共识协议使对数据的修改达成一致；利用密码学方法保证数据传输和访问的安全；利用可编程的智能合约来灵活操作数据。而从数据管理的角度来看，区块链的本质是一个网络上节点独立对等，数据以日志方式记录，并通过全复制分布实现数据记录共享，采用哈希链数据结构保证数据难以篡改，采用共识算法实现不同节点间数据副本一致性的分布式数据管理系统[7-9]。

从数据管理的角度看，与传统的数据库管理系统相比，区块链系统主要涉及的关键技术包括以下4个。

- 开放透明，数据全复制分布。节点间采用全复制的数据分布，即每个节点保存一份完整的数据副本。在区块链系统中，

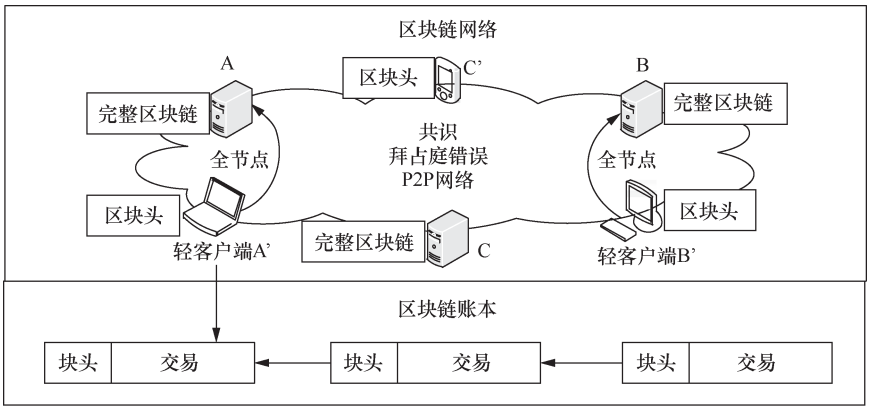

图 2 典型的区块链系统

单个节点存储的数据包括两类，一类是区块数据，另一类是状态数据。其中区块数据就是通常所说的记录一批交易的链式区块数据，一般被存储在原始文件或者key-value数据库中。状态数据则保存以区块为单位的一批交易执行后的最新世界状态（world state），也是执行智能合约时要访问的数据。因此，高效的数据存储和组织是区块链系统的关键技术之一。

● 存在恶意节点，节点间数据的一致性需要拜占庭容错共识协议来保证。由于区块链系统中的节点间数据是全复制分布的，因此在分布式环境下必须要有共识算法保证位于不同节点的数据之间的一致性。因为在区块链系统中，节点可能会主动作恶（拜占庭节点，多方互不可信），包括发送假消息、给不同的节点发送不同的消息等主观恶意行为，所以拜占庭容错的共识协议是区块链系统中的关键技术之一。

● 可编程智能合约，支持去中心化应用。智能合约为区块链系统提供了灵活的编程语义，支持用户在区块链系统上搭载自定义的应用程序，使得区块链技术可以被应用在数字货币以外的很多其他领域。以智能合约编写的DApp的运行逻辑、状态都会经过共识机制的协商确认，保证了执行过程的完整性。智能合约的安全高效运行也是区块链系统的关键技术之一。

● 数据防篡改，轻客户端数据查询可验证。与比特币系统中的简单支付验证节点一样，绝大部分区块链系统中存在一种只保存区块头部信息的轻客户端，轻客户端的查询请求往往会被转发到具有全部区块数据的全节点上执行。对来自轻客户端的查询进行响应，以及对查询结果的完整性进行验证也是区块链的关键技术之一。

3 区块链技术在数据共享流通中的应用

打破数据孤岛、实现数据共享可以发挥数据更大的价值，然而在实际的业务场景中，由于在需要对数据进行共享的不同参与方之间缺乏互信，导致数据确权困难，使得数据共享过程产生很多不必要的壁垒。数据确权就是对数据所有权和使用

权的确认。其中，所有权的归属可以是个人（如隐私数据所有者），也可以是机构（数据提供者），所有权和使用权都可以交易，或者通过法律法规转让和授予。具有去信任化、去中心化以及防篡改、可追溯等特点的区块链系统可以便捷地为参与数据共享的多方之间建立互信，并通过其上的智能合约来实现灵活多样的数据共享规则，为公开透明、可信、无争议的数据确权提供技术平台。

通过区块链来共享数据，最直接的方法是将共享数据直接上链，具体如**图3**所示。如果D节点的数据需要共享给A、B和C，则通过区块链网络直接上链，并完成同步。同样，如果其他节点的数据要应答数据共享的请求，也只能借助区块链网络对数据进行移动，从而使得数据分享流程公开透明，数据确权无争议。由于现有的区块链系统采用全复制数据分布的方式，这种共享数据直接上链的方法会使得A、B、C、D 4个节点保存通过区块链网络分享的所有数据，造成系统中存储、计算和网络资源的极大消耗和浪费，并严重影响数据共享的交易吞吐量。更重要的是，数据直接上链也会引起隐私泄露的问题，这是由于区块链仅使用签名的方式防止数据被恶意篡改，而共享数据仍以明文的形式保存。但是如果以加密的方式进行数据上链传输，又会影响链上智能合约的数据正常读取。

因此，这种数据直接上链的方式虽然能保证共享公开透明，但数据隐私得不到应有的保护。考虑到系统吞吐量以及数据隐私保护的问题，本文提出了一种链上存证、链下数据传输的数据共享方式。在这种模式下，只有对共享数据的请求和应答会被记录在区块链上，而真正的共享数据通过链下传输，具体链下共享数据的传输可以通过可信的云服务器中转，也可以直

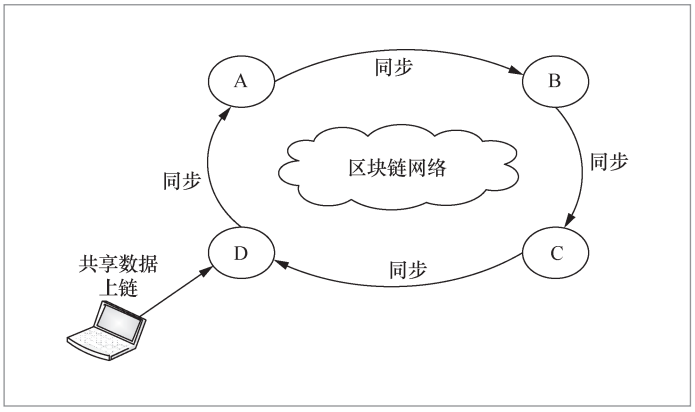

图3 共享数据直接上链

接通过点对点传输。由于所有对共享数据的请求和应答响应都被记录在区块链上，而区块链又能保证其难以篡改、公开透明、可追溯，因此任何参与方如果对数据确权有异议，都可以通过查询区块链记录来对数据的使用权和所有权的迁移过程进行追溯，并且追溯过程不受任何一方人为干扰。另外，共享数据的传输通过链下进行，一方面保护了数据隐私，另一方面也减轻了链上负载，从而提高系统的吞吐量。如果对共享数据的隐私保护有很高的要求，可以采用加密和点对点共享密钥的方式来实现链下共享数据传输的安全和隐私保护。

图4展示了一个链上存证、链下数据传输的应用场景，在该场景中共有A、B、C、D 4个参与方，各参与方拥有属于自己的数据，其中原始数据以密文的形式存储在云服务器上，原始数据的元数据（即数据目录信息，包括数据的基本描述、类别、拥有者等）及其哈希摘要则存储在区块链上。假设某些业务需要D节点获取数据集R，D节点的客户端通过查询链上的目录信息发现B节点拥有数据集R。此时，D节点的客户端向B节点发起数据获取请求，B节点根据预先定制的智能合约，检查数据请求的合

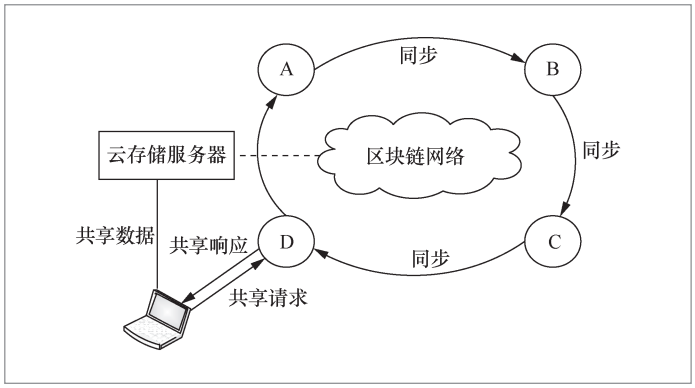

图 4 链上存证、链下数据传输的应用场景

法性，并按照预先制定的规则将数据集R在云存储服务器上的链接、数据集R的哈希摘要以及对数据集R解密的密钥等，以加密的方式发送给D节点的客户端，最终D节点的客户端根据接收到的信息到云服务器下载数据，并验证其哈希摘要，验证通过后，合法使用数据集R。需要注意的是，为了保障数据拥有者和数据使用者各自的权利，来自D节点的客户端的数据请求和B节点的应答均被记录在区块链上，其中A、B、C、D 4个节点都是本次数据请求和应答的见证者。值得一提的是，从功能的角度看，该平台可以部署在任何成熟的联盟链系统中，如超级账本Fabric或者Quorum等。但从性能的角度看，已有的联盟链系统很难满足本文要解决的数据请求和应答的高吞吐需求，需要对已有的开源系统进行进一步改造，第4节将就这一点进行讨论。

基于链上对数据共享的请求和应答进行存证、链下对待分享数据进行安全传输的基本思想，笔者设计了一个如图5所示的基于区块链的可信的数据共享平台。该平台自底向上一共4层，包括存储层、共识层、智能合约层和应用层。

存储层包括链上的区块数据和状态数据，以及为链下数据传输做准备的链外数据云平台。链上的区块数据主要为数据共享请求和应答做存证，而状态数据是为各种数据共享和确权规则编写的智能合约准备的。对于大数据共享来说，往往数据量巨大，需要较大的带宽、存储空间、处理能力等资源，本平台采用云存储的方式将待共享的原始数据存储在云端。

共识层可以根据不同的应用场景，选择不同的共识协议，比如应用于公有链的工作量证明（proof of work，PoW）、权益

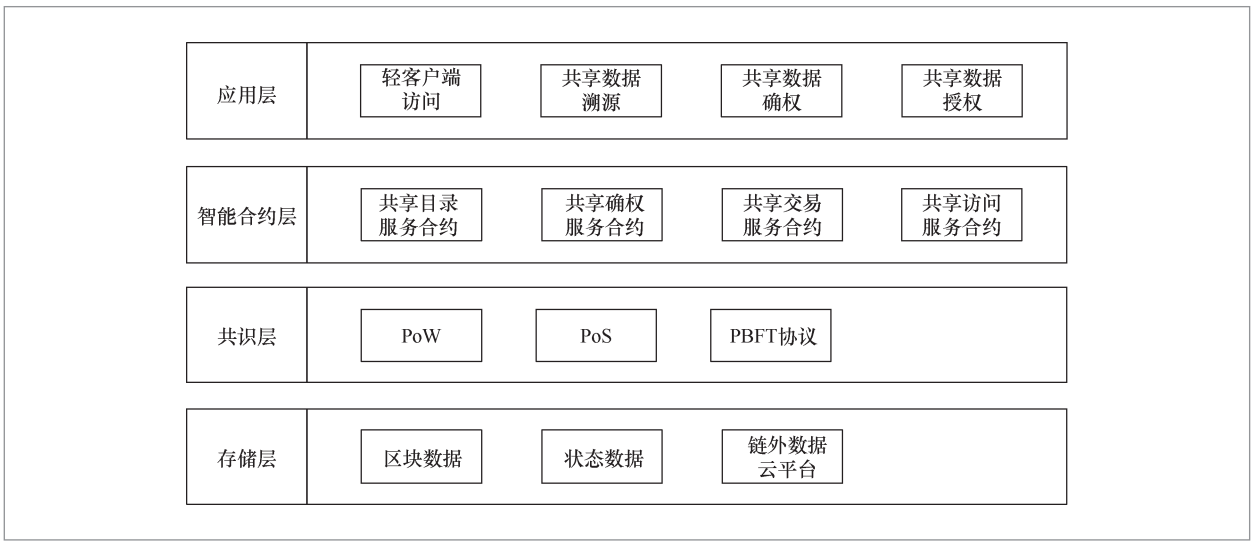

图 5 基于区块链的可信的数据共享平台

证明（proof of stake，PoS），应用于联盟链和私有链的实用拜占庭容错（practical byzantine fault tolerance，PBFT）协议。如果是企事业内部或者企事业之间的数据共享，为了避免算力的浪费，一般采用基于投票机制的PBFT协议；如果是在更广的范围内的缺乏行政约束力的主体之间的数据共享，为了能抵御各种网络攻击，也可采用PoW或者PoS。

智能合约层主要用来定义各种与数据共享和确权有关的业务流程和规则。这一层与应用场景密切相关，是最重要的一层，其中包括共享目录服务合约、共享确权服务合约、共享交易服务合约以及共享服务访问合约。其中共享目录服务合约是关于元数据管理的，主要让各参与方之间互通有无，知道从哪里可以获取需要的数据，每个参与方有哪些数据可以共享。共享确权服务合约则主要负责在数据流通过程中根据合同以及相应的法律法规对数据进行确权，以及对权利转移过程产生的费用进行计算和支付。共享交易服务合约主要用来定义数据买卖和交易的规则。共享服务访问合约主要用于访问控制验证和权限验证。

最上层是应用层，包括轻客户端访问、共享数据溯源、共享数据确权和共享数据授权。应用层主要为应用程序提供服务和访问接口，以及查询和跟踪数据共享的进展和结果。其中轻客户端访问要保证查询结果的完整性，因为数据来自不可信的全节点，需要保证查询结果的保真和完备，即返回的数据没有被恶意篡改，并且返回的数据既不能多也不能少。共享数据溯源主要负责对有争议的数据确权进行过程溯源，要求给出数据权限迁移时间线及相关的证据。共享数据确权用来查询当前某个已共享数据项的拥有权或者使用权归属。共享数据授权用来为共享数据设置访问控制和权限管理规则。

使用笔者设计的基于区块链的可信的数据共享平台在不可信的多方之间共享数据，需要注意以下几点。首先，区块链的难以篡改性使得智能合约一旦执行即不可挽回，在数据共享流通的过程中，这一特性要求共享规则和与之相匹配的智能合约的设计需要特别精细，例如，确保转让后的使用权在未经授权的情况下不能再擅自转让给第三方。其次，因为数据共享的过程采用链下传输、链上存证的方式，所以链下数据传输的安全性也需要链上验证，在链下传输的共享数据需多方加密，并将数据摘要通过交易上链，访问时通过智能合约获取公钥和数据摘要，确保共享数据是数据提供者共享的原始数据。再次，由于监管的需要，或者受存储和计算资源所限，有些参与方只保存了区块的头部信息，为了使这类轻客户端用户对共享交易过程及确权结果进行跟踪和追溯，系统需要提供可验证的查询处理，以保证查询结果的可信和完整。最后，数据拥有者需要对共享数据进行控制，只有获得共享数据相关权限的用户才能使用相关的共享数据，并行使相应的权利。

综上所述，与传统的数据共享平台，以及原始数据直接上链的基于区块链的数据共享平台相比，笔者提出的链上存证、链下数据传输的可信的数据共享平台具有如下优点：一是链上负载轻，原始数据隐私方便保护；二是共享流程和权利转移公开透明，不可抵赖；三是共享规则和权利转移智能合约化，避免人为干预，自动完成不可逆转。

4 区块链系统应用于数据共享方面面临的挑战

笔者提出的链上存证、链下数据传输

的方法能在一定程度上避免共享数据直接上链造成的存储资源、计算资源以及网络资源方面的大量消耗和开销，从而缓解系统负载压力，这对于区块链系统中资源受限的单个节点的扩展性尤为重要。

但是，随着平台参与节点和数据分享请求越来越多，系统负载越来越大，用户需求对系统吞吐的要求也会越来越高，已有的区块链技术仍然很难保证系统的正常运行。下面从存储、共识、智能合约执行以及轻客户端可验证查询4个方面阐述该平台面临的挑战和需要进一步改进的方向。

首先，已有的数据全复制分布方式成为数据存储的瓶颈。全复制的数据分布方式使得每个节点都需要保存一份完整的数据副本，系统的存储容量受到单个节点存储能力的限制。TPS（transaction per second）只有20笔/s左右的以太坊在运行了不到三年半的时间以后，存储容量已经超过了1 TB。那么，在交易吞吐率为几百甚至上千笔每秒的联盟链中，系统所需的存储容量会以更快的速度增长，极有可能在短时间内就突破单节点存储能力的上限。因此，全复制的数据分布方式制约了系统存储的可扩展性，也制约了系统的扩展性。

其次，已有的共识算法很难满足系统对交易处理时延和吞吐率的要求。PoW[9]系列共识算法由于需要消耗大量算力，且吞吐率提高以后会引起更多区块链分叉进而导致私自挖矿等安全方面的问题[9]，并不适合上述企业级联盟链的应用场景。而以PBFT协议[10]为代表的面向联盟链的共识协议，TPS也只有400笔/s左右，所能支持的共识节点数目也不能超过20个，因此，已有的共识算法很难满足数据共享请求的高吞吐率的要求。

再次，已有的智能合约以串行执行方式运行，这会成为系统吞吐性能的瓶颈。在面向数据分享和确权应用的区块链系统中，智能合约的业务逻辑往往比简单的转账操作更复杂。尤其是在系统共识算法效率提高、出块速度加快以后，智能合约的执行更有可能拖慢系统的整体吞吐性能。在采用IBFT（istanbul byzantine fault tolerant）共识算法的以太坊平台上做的一组实验表明，当一个区块中包含100笔调用智能合约的交易时，合约执行时间是共识时间的6倍，而当智能合约的交易数目达到200笔时，合约的执行时间是共识时间的20倍。而在相同的系统上采用PoW这种吞吐率很低的共识算法时发现，与共识时间相比，智能合约的执行时间几乎可以忽略。这进一步验证了共识效率提高以后，智能合约的执行效率会成为系统新的瓶颈[11-12]。而现有的智能合约的串行执行方式严重制约了其执行的效率。

最后，对于轻客户端发起的查询，现有的区块链系统无法保证其查询结果的正确性[13]。轻客户端被分为两类，一类隶属于某个全节点，如隶属于某个部门的轻客户端；另一类不隶属于任何全节点，如普通终端用户。来自前者的查询响应比较简单，可以直接在它所隶属的全节点执行；而来自后者的查询并没有一个完全信任的节点，系统需要保证其查询结果的正确性[14-15]。但是已有的区块链系统仅能验证轻客户端发起的交易或者账号是否存在，无法对复杂查询返回的结果集的正确性和完整性予以验证。对于面向数据分享和确权应用的区块链数据管理系统，无信任轻客户端的存在不可避免，对其查询结果正确性的验证也是必须要面对的问题。

因此，需要在上述4个关键技术点进行突破，构建高吞吐的区块链系统，以应对大规模节点参与下系统对吞吐率和时延方面日益增长的需求。

5 区块链技术应用于数据共享方面的相关工作

鉴于区块链的可追溯、难以篡改等特性，区块链系统可以解决在供应链数据管理、医疗数据分享等领域进行数据分享时面临的可信问题。但是现有的基于区块链的数据分享系统大多面向某一特定的具体应用，通用的基于区块链的数据分享平台还比较少见。

Tian F等人[16]提出了一种不仅能够实时提供供应链管理系统中食品的追踪信息，同时还能保证信息的可靠公开的系统。Hua J等人[17]提出了一种基于区块链的农产品追踪系统，能够记录生产、存储、运输、加工、分配以及相关供应链的各种信息，同时向第三方（如政府、保险公司、顾客等）公开。Yue X等人[18]提出了一个基于区块链的智能应用HDG，病人不需要通过第三方就能控制和分享自己的病历。Azaria A等人[19]开发了一个去中心化的大规模病历数据管理系统MedRec。在这个系统中，通过区块链记录的综合日志能够实现医疗记录的保密、认证和追溯。Xia Q等人[20]开发了一个可靠的基于区块链的系统MeDShare，专门用来处理在云端存储大量冗余的医疗信息时遇到的数据追溯问题。Rifi N等人[21]讨论了区块链技术应用于医疗数据共享领域的优缺点。

6 结束语

综上所述，在数据共享和流通过程中由于缺乏信任而形成的数据孤岛，使得数据的价值难以最大化。区块链系统虽然能够为在数据共享中涉及的多方构建信任的基础设施，但是共享数据直接上链的方式仍然会面临系统负载重、隐私得不到保护的问题。本文提出了一种数据共享请求和应答记录链上存证、共享数据链下传输的数据共享平台架构。该架构在一定程度上可以缓解系统负载过重的问题，但是随着参与节点的增多，以及每秒需要处理的数据共享请求的应答增多，将已有的区块链技术应用到数据分享和确权领域时，仍然需要从以下4个方面努力：首先，设计高效的可验证、可恢复的数据可扩展存储方法；其次，在确保所有节点都能公平对等地参与共识过程的前提条件下，提高确定性共识协议的效率，并保证协议的安全性和活性；再次，有效利用区块链智能合约的特点，提高智能合约的并发执行效率；最后，以减少网络开销和计算代价为目标，设计精巧的验证结构、高效的验证查询算法来响应轻客户端查询。

参考文献:

[1] 周茂君，潘宁. 赋权与重构：区块链技术对数据孤岛的破解[J]. 新闻与传播评论，2018，71(5)：59-68.
ZHOU M J，PAN N. Empowerment and reconstruction：blockchain technology breaks data isolated island[J]. Journalism and Communication Review，2018，71(5)：59-68.

[2] 苏雄业. 基于区块链的大数据共享模型与关键机制研究与实现[D]. 北京：北京工业大学，2018.
SU X Y. Research and implementation of big data sharing model and key mechanisms based on blockchain[D]. Beijing：Beijing University of Technology，2018.

[3] 杨保华，陈昌. 区块链原理、设计与应用[M]. 北京：机械工业出版社，2017.

YANG B H, CHEN C. Blockchain principles, design and applications[M]. Beijing: China Machine Press, 2017.

[4] NAKAMOTO S. Bitcoin: a peer-to-peer electronic cash system[R]. 2009.

[5] 闫莺, 郑凯. 以太坊技术详解与实战[M]. 北京: 机械工业出版社, 2017
YAN Y, ZHENG K. Ethereum technology and practice[M]. Beijing: China Machine Press, 2017.

[6] 邵奇峰, 金澈清, 张召, 等. 区块链技术: 架构及进展[J]. 计算机学报, 2018, 41(5): 3-22.
SHAO Q F, JIN C Q, ZHANG Z, et al. Blockchain: architecture and research progress[J]. Chinese Journal of Computers, 2018, 41(5): 3-22.

[7] 钱卫宁, 邵奇峰, 朱燕超, 等. 区块链与可信数据管理: 问题与方法[J]. 软件学报, 2018, 29(1): 150-159.
QIAN W N, SHAO Q F, ZHU Y C, et al. Research problems and methods in blockchain and trusted data management[J]. Journal of Software, 2018, 29(1): 150-159.

[8] 钱卫宁, 金澈清, 邵奇峰, 等. 区块链与分享型数据库[J]. 大数据, 2018, 4(1): 36-45.
QIAN W N, JIN C Q, SHAO Q F, et al. Blockchain and sharing database[J]. Big Data Research, 2018, 4(1): 36-45.

[9] O'DWYER K J, MALONE D. Bitcoin mining and its energy footprint[C]// The 25th IET Irish Signals & Systems Conference. [S.l.:s.n.], 2014.

[10] CASTRO M, LISKOV B. Practical byzantine fault tolerance[C]//The 3rd Symposium on Operating Systems Design and Implementation. [S.l.:s.n.], 1999: 173-186.

[11] LUU L, CHU D H, OLICKEL H, et al. Making smart contracts smarter[C]// The 2016 ACM SIGSAC Conference on Computer and communications Security. New York: ACM Press, 2016: 254-269

[12] DICKERSON T, GAZZILLO P, HERLIHY M, et al. Adding concurrency to smart contracts[C]//The ACM Symposium on Principles of Distributed Computing. New York: ACM Press, 2017: 303-312.

[13] ZHU Y, ZHANG Z, JIN C, et al. SEBDB: semantics empowered blockchain database[C]//2019 IEEE 35th International Conference on Data Engineering. Piscataway: IEEE Press, 2019: 1820-1831.

[14] LI F, HADJIELEFTHERIOU M, KOLLIOS G, et al. Dynamic authenticated index structures for outsourced databases[C]// The 2006 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2006: 121-132.

[15] CHEN Q, HU H, XU J. Authenticated online data integration services[C]// The 2015 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2015: 167-181.

[16] TIAN F. A supply chain traceability system for food safety based on HACCP, blockchain & Internet of things[C]//2017 International Conference on Service System and Service Manage. Piscataway: IEEE Press, 2017: 1-6.

[17] HUA J, WANG X J, KANG M Z, et al. Blockchain based provenance for agricultural products: a distributed platform with duplicated and shared bookkeeping[C]//2018 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2018: 97-101.

[18] YUE X, WANG H, JIN D, et al. Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control[J]. Journal of Medical Systems, 2016, 40(10): 218.

[19] AZARIA A, EKBLAW A, VIEIRA T, et al. MedRec: using blockchain for medical data access and permission management[C]//2016 2nd International Conference on Open and Big Data (OBD). Piscataway: IEEE Press, 2016: 25-30.

[20] XIA Q, SIFAH E B, ASAMOAH K O, et al. MeDShare: trust-less medical data sharing among cloud service providers via blockchain[J]. IEEE Access, 2017, 5:

14757-14767.
[21] RIFI N, RACHKIDI E, AGOULMINE N, et al. Towards using blockchain technology for e-health data access management[C]//2017 14th International Conference on Advances in Biomedical Engineering (ICABME). Piscataway: IEEE Press, 2017: 1-4.

作者简介

张召(1977-),女,博士,华东师范大学数据科学与工程学院副教授,主要研究方向为区块链系统研发、分布式数据管理,多项研究成果发表在VLDB、ICDE和DASFAA等数据管理领域的重要国际会议上。先后主持多项国家自然科学基金项目,作为骨干技术人员,参与开发的“面向大型银行应用的高通量可伸缩分布式数据库系统”项目获得2017年教育部高等学校科学研究优秀成果奖(科学技术)科技进步奖一等奖,“支持互联网级关键核心业务的分布式数据库系统”项目获评2019年度国家科学技术进步奖二等奖。

田继鑫(1977-),男,MCT Technology研发部门负责人,主要研究方向为分布式系统开发、互联网系统后台开发及设计,以及比特币、超级账本Fabric以及以太坊等主流区块链系统的架构、源码及应用开发。

金澈清(1977-),男,博士,华东师范大学数据科学与工程学院教授、博士生导师、副院长,中国计算机学会高级会员,数据库专业委员会委员,已发表学术论文100余篇,研究成果曾获得省部级一等奖和二等奖、霍英东教育基金会青年教师奖,担任《计算机研究与发展》编委,主要研究方向为区块链、计算教育学、基于位置的服务等。

收稿日期: 2020-05-12
基金项目: 国家自然科学基金资助项目(No.61972152)
Foundation Item: The National Natural Science Foundation of China (No.61972152)

银行业金融机构数据治理指引和DCMM的对比分析

代红[1]，张群[1]，芦皓麟[2]，宾军志[3]
1. 中国电子技术标准化研究院，北京 100007；2. 天津大学微电子学院，天津 300072；
3. 全国信息技术标准化技术委员会大数据标准工作组，北京 100007

摘要
近年来，数据治理得到各行各业的普遍重视，国家和行业都发布了相关的标准和政策，通过相关文件明确数据治理的概念和体系，促进数据治理行业的发展。对相关文件进行解读，总结其中的异同之处，帮助人们了解与数据治理相关的管理趋势和应用的重点，同时，提出数据管理能力成熟度评估模型在银行业落地实施的建议，帮助银行更好地满足相关监管要求，提升数据管理能力的成熟度等级。

关键词
大数据；数据质量；数据治理；数据管理能力成熟度评估模型；数据文化

中图分类号：TP30 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020048

Comparative analysis between bank industry data governance guidelines and DCMM

DAI Hong[1], ZHANG Qun[1], LU Haolin[2], BIN Junzhi[3]
1. China Electronics Standardization Institute, Beijing 100007, China
2. School of Microelectronics, Tianjin University, Tianjin 300072, China
3. Big Data Standardization Working Group, China National Information Technology Standardization Technical Committee, Beijing 100007, China

Abstract
Recently, data governance got widespread attention of many industries. Some standards and polices had been published from country and industries, these files defined data governance concept and framework and boost data governance industry development. These files were analysed and some differences between them were identified to help people understand the trends and focus of data governance industry. At same time, some suggestions about how to implement DCMM in the bank industry were provided to help banks to better meet regulatory requirement and improve the maturity level of data management ability.

Key words
big data, data quality, data governance, data management capability maturity assessment model, data culture