

企业级区块链系统设计课程读书报告



题 目 区块链的数据存储

作者姓名 秦清澳

作者学号 22151174

所学专业 软件工程

所在学院 软件学院

提交日期 二○二一 年12月

## 摘要

区块链这一颠覆性技术已经引起了广泛的关注。在大多数情况下，区块链被视为一种用来存储数据的分布式数据库。然而，随着时间的推移，区块链的数据存储面临着严重的数据膨胀问题。另外，应用区块链存储数据也带来了不可编辑、性能低等问题。目前，学者们针对这一方面的挑战做了许多研究，但缺乏全面的概括。对相关文献进行了筛选，分析。选取了65篇与研究主题高度相关的文献，并进行了定性分析。通过分类的方式介绍了当前的研究现状，总结了目前的挑战以及现有的解决方案，并对未来的研究方向进行了展望。

**关键词**：系统文献综述；区块链；数据存储；存储架构；分类；挑战

## 背景

2008年，有学者化名中本聪在《Bitcoin: A Peer-to-Peer Electronic Cash System》一文中提出了一种名为比特币(Bitcoin)的数字货币[1]。比特币巧妙地解决了困扰数字货币已久的拜占庭将军问题以及双重花费问题。作为其底层技术，区块链也应运而生，但此后几年并未掀起波澜。2015年，以太坊（Ethereum）横空出世，赋予了区块链图灵完备的可编程特性[2]。区块链技术也因此开始受到了学者们的广泛关注，直至今日已然成为了最热门的研究领域之一。

区块链作为一种全球性的分布式公共账本，有着去中心化、可追溯性、防篡改性等优点。大多数情况下，区块链作为一种分布式数据库，已被成功应用到医疗[3]、供应链[4]、教育[5]、物联网[6]等领域。然而，由于区块链要求节点必须存储全部历史数据，在这个每天产生海量数据的时代，节点的存储压力会越来越大。以比特币为例，截止到2021年12月31日，其所占存储空间已达到约383.0GB。这导致当新节点加入区块链网络时，初始同步过程耗费大量时间。另外，使用区块链存储数据，在带来便利的同时，也会带来安全性问题以及查询效率低下、审计计算成本高昂等性能局限性问题。不管从区块链本身的角度，还是从区块链应用的角度，区块链的数据存储问题都极为重要。许多学者认识到了这一点，进行了各种相关的研究。有的学者面对挑战，提出了新的存储架构；有的学者则运用了新的编码技术；有的学者将相应问题建模，提出了新型算法。对现有研究进行系统分析，明确问题所在以及现有解决方案，从而指出未来的研究方向有着重要的意义。本研究介绍了关于“区块链的数据存储”的研究现状。

## "区块链的数据存储"的研究分类

本节通过描述提出的分类法来组织分析研究现状。为了使读者可以有的放矢地了解研究进展，分析了文献的创新点。创新点是一篇文章的核心价值，可以帮助研究人员了解目前解决挑战的各种方式，研究人员可以改进已有方式，也可以另辟蹊径，探索新的可能。对选择的文献进行了分析，得出了现有研究创新点的4个类别，分别为：存储架构、编码技术、密码学策略、算法/方法论.大多数的文章提出了独特的存储架构(46篇，70%)，一部分文章采用了新的编码技术(11篇,17%)，或者使用了新颖的加密算法/数字签名策略(6篇,9%)。只有极少数文章进行了算法(3篇，4%)/方法论层面的探索研究.

### 2.1 存储架构

#### 2.1.1 区块链+IPFS

为了应对区块链的数据存储所占空间日益增大的问题，多项研究提出了使用区块链存储关键信息，使用分布式文件系统这种链下方式存储真实数据的存储架构。星际文件系统(Inter Planet File System, 简称IPFS)可以存储结构化和非结构化文件，不仅提供了基于内容寻址和基于位置寻址两种交易访问方式，还可以处理数据冗余以及优化副本文件。因此，大多数工作(15篇，35%)使用IPFS作为存储真实数据的分布式文件系统[7]。

为了尽可能多地降低区块链数据的存储空间，学者们采取了只用区块链存储用户秘钥和文件关键字索引[8]、交易有关信息(包括地址签名)[9]或者IPFS返回的哈希值[10-13]的方式，并结合基于属性加密机制(Attribute Based Encryption，简称ABE)、智能合约等实现了额外的功能。文献[8]利用区块链结合ABE实现了对数据的细粒度访问控制，并使用智能合约提供了在密码文本上进行关键字搜索的功能。文献[9]利用IPFS的分布式哈希表(Distributed Hash Table)技术帮助区块链减少了响应时间，并通过智能合约引入了RESTful API与区块链通信的协议服务，提高了区块链的效率。IPFS的哈希值仅包含46字节，文献[10-13]利用这一特点，在大大减少了区块链存储空间的同时，解决了区块链网络规模增长的限制，并保证了数据一致性以及可追溯性。

在节省存储成本的同时，学者们通过创建二级区块链[14]、冷热数据分类[15]、数据交易并行[16]等方法，解决了区块链吞吐量低、初始同步时间慢等性能问题。文献[14]将原始数据块通过加密相互链接，从而创建二级区块链形成了双链架构。与比特币核心相比，该架构在节省了1600多倍的链上存储空间的同时，提高了20多倍吞吐量。文献[15] 将区块链的数据按照使用频率分为热数据和冷数据，将热数据存储在本地数据库（LevelDB）、冷数据存储在IPFS集群，大大缩短了区块链初始同步时间，并节省了大量的存储空间。文献[16]将节点分为两类，分别是为获取激励而存储其他节点生成文件的存储节点和提供计算能力来加密文件的代理节点。这种架构实现了在区块链上数据和交易并行，减轻了节点的存储负担并且大大提高了区块链的性能[16]。对于大文本数据的存储，文献[17]提出根据业务需求将数据按日期、功能点、数据大小等粒度分块的方案，大大提高了交易效率，减少了网络的冗余传输。

由于物联网设备规模庞大，每天产生巨量数据，基于IPFS的链下存储也出现了存储成本过高以及定期删除的问题。为了解决这些问题，学者们提出了定期回收机制[18]、激励机制[19]、网络分层机制[20]。结合食品过期的特性，文献[18]定期回收存储空间，减少了IPFS链下存储空间的增长速度。文献[19]出于IPFS节点在一定的时间间隔后删除数据的原因，介绍了一种激励IPFS的节点存储数据的激励机制，保证了区块链的可追溯性。文献[20]则将区块链和IPFS网络分为高低两层，分别存储不同交易者之间的交易及数据，使得基于区块链的系统适应工业4.0下的物联网环境。

智能合约可以保护数据隐私和安全，是用来存储用户的隐私数据的好方法[21]。然而，由于区块链存储的不变性，智能合约的迭代更新成为了一大挑战。文献[22]提出了一套热交换智能合约架构，基于合同地址来管理去中心化应用(Decentralized Application，简称DApp)的用户数据。该机构将通过智能合约分别实现后台服务功能，智能合约通过合约地址相互调用，从而构建了后台服务功能的热插拔架构，减少了后台的迭代更新和维护成本。

#### 2.1.2 区块链+云存储/边缘计算

如今，利用云存储、边缘计算存储海量数据是技术发展的大势所趋。一些学者(9篇,23%)也提出了区块链+云存储/边缘计算的存储架构来缓解区块链的数据存储成本太大的问题。

在采用区块链+云存储架构的同时，一些学者提出了结合去中心化存储以及两者混合存储[23]或者结合密文策略-基于属性的加密（Ciphertext Policy Attribute Based Encryption，CP-ABE）模型[24]的方法，缓解了区块链的高交易处理时间、权力管理计算成本过高等效率问题。 除了采用云存储外，文献[23]另外给出了去中心化存储和两者混合存储的方案，根据需求选择链下存储，可以降低交易处理时间、提高可伸缩性。文献[24] 通过扩展CP-ABE模型，提出了角色分配和属性树的概念。这种新型的由用户控制的链上权限管理方案，解决了权限冗余、权限管理计算资源浪费的问题。

为了保证区块链的存储规范性、可用性，学者们采用了智能合约[25]、存储证明协议(Proof of Storage,简称PoS)[26]、新建查询层[27]等手段。文献[25-26] 使用云存储加密后的教育档案，而档案的哈希信息存储在区块链上，保证了保证数据存储的安全性和可靠性。另外，文献[25]使用智能合约来规范存储和共享的过程。文献[26]则结合副本和PoS机制，保证数据的高可用性和完整性。不管是对于开发者还是用户，区块链的可用性都可以定义为轻松访问(包括处理和分析)区块链数据的能力，而这正是传统的区块链系统被诟病的地方。因此，文献[27]新建了一层查询系统，提供了三种主要的查询功能:(1)基于多个搜索参数查找区块链数据(检索查询)，(2)区块链数据集合的简单统计分析(聚合查询)，(3)根据其区块链分量对区块链数据进行排序(排名查询)。

为了减少数据冗余备份、提高错误数据定位效率，学者们采用了隐私/共享数据分离[28]、BLS同态线性验证器(BLS-Homomorphic Linear Authenticator，简称BLS-HLA)结合计数布隆过滤器表（Counting Bloom filter Table，简称CBT )技术[29]的方案。文献[28]将数据隐私部分与共享部分分离，使用存储区块链存储完整的索引、共享区块链存储共享部分的索引的双链架构，避免了数据的冗余备份。另外，文献[28]还结合了CP-ABE与对称加密技术，既保证了数据在云中的安全存储，又实现了数据更新的一致性和便捷性。运用BLS-HLA、CBT技术，文献[29]介绍的区块链+边缘计算+云计算的存储架构提供了存储数据快速定位的功能。该架构下，边缘计算实时地给终端设备提供数据处理、存储和计算服务；云计算收集并存储边缘计算中碎片化的数据，实现更强大且长期的数据处理、备份和存储功能。值得一提的是，BLS-HLA技术，可以在不检索数据副本的情况下对分布式存储进行检查。另外，当数据存储检查失败时，边缘服务器可以在CBT上执行二分查找，准确定位系统中的错误数据块。

在提出本地+区块链+云存储新架构的同时，一些学者则采取了系统分层[30]、周期性回收存储空间[31]的方法来应对区块链的可扩展性问题。该新型架构将每部分数据存储在一定比例的节点中，而不是全部节点上。文献[30]将系统视为本地、区块链和云三层，并且提供了两个连接器（即区块链连接器和云连接器），以无缝连接这三个层。而且连接器提供了功能可扩展性、提高了不同区块链存储结构层间的效率。文献[31]介绍了一种基于循环队列的回收闲置空间的方法，避免了存储空间的浪费。另外，通过重用区块链以太网存储的智能合约字节码，节省了将近46%的智能合约数据存储空间,提高了区块链的存储可扩展性。

由于当前各个领域（例物联网，人工智能等）海量的数据需求，选择哪种存储技术以降低成本成为了一个困扰人们良久的问题。文献[32] 联合考虑了云、多云和去中心化存储技术，针对成本最小化的存储选择问题进行了建模，提出了基于动态规划(Dynamic Programming)和贪婪风格(Greedy Style)两种启发式算法。在此基础上，实现了一种智能选择存储技术的自适应中间件。

#### 2.1.3 其他存储架构

除了区块链+IPFS、区块链+云存储这两种被广泛使用的存储架构之外，学者们还提出了许多新颖的存储架构(例如,双链[33]、复合链[34])。文献[33]利用链式数据结构存储区块链交易的哈希值，与区块链的链形成了一种双链存储架构。这一架构巧妙地解决了区块链技术难以实现自动化存储和基于身份批量获取区块链存储的哈希数据的问题。而文献[34]则构建了一种结合私有链与联盟链的复合链区块链架构。该架构根据待存储的实体数据的特点和实体本身的存储需求，在每个实体内构建一条私有链。在此基础上，结合区块链技术不同应用模式的特点，在不同实体之间构建联盟链。该存储架构实现了在复杂或分类场景下的自适应数据关联存储，从而解决了单一区块链存储结构带来的存储效率低的问题。

为了降低区块链的存储空间，学者们提出了一些基于分片[35]、半全节点[36]、快照[37]、时间序列数据[38]以及虚拟区块组(Virtual Block Group，简称VBG)[39]等新思想的存储架构。文献[35]使用了一种基于分片技术的存储架构，主要分为三个模块：基于分片的区块链模块、特征提取模块和分类器模块。 特征提取模块会提取出块的特征值。分类器模块根据特征值使用极限学习机(Extreme Learning Machine，简称ELM)方法将区块划分为热块和非热块。分类后，每个节点保存最相关的热块，从而显著地降低存储空间，并且提高查询效率。文献[36]则在全节点和轻节点之间设计了一种半全节点，半全节点只需要同步和验证区块链上的部分数据。在该架构中区块链会被划分为若干区块链段，每个半全节点随机存储一个或多个区块链段，并将每个区块的块头存储在主链中。另外，该架构还使用MCMC(Markov Chain Monte Carlo，简称MCMC)随机算法公平删除区块。通过使用该存储架构，可以大大地减小区块链存储容量。文献[37]采用了快照的思想,快照包含完整的未花费交易输出集(UTXO集)和所有区块头。基于快照的存储架构会定期创建区块链的快照并存储在数据块中。这些存储了快照的区块被链接在一起，形成了与主链链接的第二个区块链。显然，需要节点下载和存储的数据量被大大地缩减，区块链的规模甚至因此缩小了93%。文献[38]提出了一种新颖的时间序列数据存储架构，有区块存储、索引存储(LevelDB)和时间序列数据存储三个模块组成。通过适当的参数配置，该架构可以保留一定的时间窗口内的数据，自动清除过期的数据块，释放用于动态增长的存储空间。文献[39] 介绍了一种VBG的思想，所有区块根据区块高度分为若干个VBG。一些VBG存储完整的区块数据，称为全VBG；而其他VBG只存储元数据信息，称为简单VBG。在该架构下，每个节点只需要存储部分区块数据，大量节省了节点的存储空间。

为了加强区块链存储数据的审计功能，学者们给出了链上链下合作审计[40-41]、BLS签名[42]的方案。文献[40] 提出了一种基于区块链交易用户存储资源的架构。租用空间、释放空间等交易都在区块链上进行，而文件的传输在链下进行。该架构不仅可以充分利用用户的存储资源，还提供了比现有方案更强的审计（链上审计、链下审计），可以更早地检测和惩罚恶意行为的一方。由于区块链的存储架构，其不可变性和公共可访问性带来了安全性和效率问题。而链下存储的外包行为，使存储提供商能够破坏存储审计的完整性。因此，文献[41]介绍了一种基于链上链下合作的存储审计方案，采用了带有链上检查机制的零知识存储审计方案解决了安全问题，并优化了审计原型，提高了整体审计效率。文献[42]使用了一种区块链+双服务器的存储架构,通过引入重复数据删除协议，有效地节省了存储空间。为了避免重复审计，利用BLS签名设计了轻量级认证器生成算法和相应的更新协议，减少了元数据冗余。因此，该架构既能节省存储资源，又能减少审计的计算资源。

在关注存储架构的同时，学者们也注意到了使用区块链存储的性能问题，并提出了底层数据分类[43]、数据价值密度[44]的手段。文献[43]将联盟链的底层数据分为两类:连续数据和状态数据。对于连续数据，设计了一个专门的基于索引的存储引擎，区块链存储连续数据对应的唯一键值。而连续数据作为文件，采用双层索引的方式以便快速访问。对于状态数据，提出了一种多级缓存机制，设计了特殊的读写缓存，结合持久化策略，可以保证海量状态数据的安全性和完整性。通过这种分类处理方式，满足了企业级业务场景下海量数据存储的高效数据访问需求，缓解了主流联盟区块链的存储架构表现出明显的性能局限性。文献[44]中的节点根据价值密度选择性地存储数据，高价值密度的数据被存储在本地，低价值密度的数据会根据选择性存储方法存储在其最近的k个节点中。该架构由数据语义信息模板、滞后数据片和历史数据存档三种机制来实现。数据语义信息模板机制通过存储简化的语义信息到本地缓存来进行共识。滞后数据片机制将数据分片，根据数据价值密度和节点存储容量动态地将不同的切片存储到不同的节点中，从而降低每个节点的存储成本。历史数据存档机制是指如果数据的价值密度很低，并且相关节点的存储容量不允许存储更多的数据，那么低价值密度的数据将被归档到集中式数据库中。该架构在可以减少数据的存储的同时，保持着数据的安全性和高效率。

基于秘密共享的区块链，在共享数据的同时，保证了数据的安全性和隐私性[45]。文献[46]提出了一种新型的基于后量子点阵的安全门限可验证的多秘密共享方案(verifiable multi-secret sharing scheme,简称VMSS)。该方案不仅具有验证和私密通信功能，还可以降低恢复通信成本，提高以往分布式存储区块链(Distributed Storage Blockchain,简称DSB)系统的鲁棒性，并且显著降低了传统区块链系统的存储成本。

### 2.2 编码技术

编码技术作为一种提高系统容错性的冗余手段，有着比基于副本策略更低的存储成本。因此，许多学者在编码方面进行了探索，包括Reed-Solomon(简称RS)编码等。

#### 2.2.1 Reed-Solomon编码(RS)

分布式存储系统往往采用冗余策略来提高系统可靠性，而擦除编码(Erasure Coding，简称EC)就是一种有效的冗余策略。EC将原始数据块们转换成一个编码块集合(即chunk)，使得任何具有足够可用chunk的子集都可以重建原始数据。在相同的容错级别下，EC的存储开销比基于副本方式低得多。作为一种被广泛使用的EC码，RS编码可以使每个数据块的存储成本从O(n)减少到O(1)，极大地降低了存储空间。

在减小存储空间的同时，为了实现存储节点的负载均衡、保证数据的不可篡改以及隐私性，文献[47]介绍了一个结合RS编码、参考前向模型（The reference forward model，简称RFM）以及智能合约等技术的系统。RFM确定存储数据的节点，智能合约保证数据不被篡改。该系统还对所有数据访问进行实时审计，保护了隐私数据。

RS在提高了存储效率的同时，也带来了高修复损失的问题。为了修复故障数据，RS往往需要读取比实际故障数据数量还要多的额外数据。与全副本方式相比，由于解码时节点之间的数据交换，区块链的读取性能也会大幅下降。另外，当新节点加入区块链网络时，重新编码历史块也会消耗巨大的网络传输通信成本。

为了增强区块链的存储可扩展性、提高读取性能，文献[48]使用了领导者驱动的在线重新编码(Online Re-encoding Protocol，简称ORP)协议和基于冷热数据的缓存结构。领导者驱动的ORP协议采用RS重新编码历史块，既保证了所有块的可用性，又降低了新节点加入时的通信成本，从而提高了存储可扩展性。另外，文献[48]采用基于冷热数据的块缓存和请求缓存结构，保证了块的访问效率，提高了系统的读取性能。

#### 2.2.2 再生编码

BCH(Bose–Chaudhuri–Hocquenghem，简称BCH )编码是一种流行的基于擦除编码的极大距离可分码(maximum distance separable,简称MDS）。然而，利用BCH码恢复丢失数据会耗费巨大的带宽。针对BCH码的问题，提出了优化固定带宽的恢复编码-再生编码(Regenerative Code)。作为改进的BCH码，再生编码不仅保留了MDS擦除编码的特性，而且通过结合网络编码(Network Coding,简称NC)的原理来修复出错的数据，大大降低了所需的网络带宽。再生编码在不影响数据整体完整性的情况下，大幅地减少了每个节点的存储空间。

为了尽量降低数据修复成本，学者们使用了再生编码[49]、本地再生编码[50]的方法，不仅降低了存储资源成本，而且最小化了分布式存储体系结构中恢复数据的延迟。文献[49]在基于区块链的副本机制的情况下，使用再生编码修复错误的文件时，明显减少了修复带宽和磁盘I/O。每个节点节省了高达91.18%的存储空间，从而使将区块链应用到容量较小的物联网设备网络中成为了可能。另外，再生编码可以使降低87.19%的延迟。文献[50]则介绍了基于本地再生编码技术(local regenerative code )的区块链系统，该系统保证了网络数据的可靠性和隐私性。当发现存储的数据被损坏时，由矢量编码(vector code)构建的多个本地修复组可以同时且高效地修复多个分布式数据存储节点，从而使系统有良好的局部可修复性和实时性。

#### 2.2.3 其他编码

除了RS编码、再生编码之外，学者们提出了向量量化编码[51]、分布式存储编码[52]、哈夫曼编码[53]等编码技术。文献[51]利用模糊聚类算法和向量量化编码技术，实现了数据的结构重组和特征重构，确定了数据最优存储结构。另外，文献[51]实现了数据存储系统的容错策略优化，提高了数据存储的安全性。针对区块链上不经常访问的已完成交易的记录(即冷存储)，文献[52]结合了Shamir秘密共享、私钥加密以及分布式存储编码技术，使得每个节点只存储每个交易的一部分。另外，分布式存储编码可以降低大规模区块链网络的存储成本，并确保了存储的安全性。为了增强数据完整性，文献[52]还提出了一种通过交易分配节点和分布数据的动态区域分配算法。文献[53]设计了一种基于片段和数据库快照(Database Snapshot，简称DS)的区块存储机制，使用哈夫曼编码（Huffman Code）压缩区块存储的部分数据，显著减少了存储空间。其中，片段是指主链中的区块的集合，DS则指记录的状态(例如，特定时刻的账户余额和全局设置)。另外，通过特别的报酬规则，该系统的数据可以分布式、动态地存储，能够承受大规模节点的突然丢失，有着很强的鲁棒性。

通过引入编码技术，可以显著的降低区块链系统的存储开销，但随之到来的是巨大的计算开销。为此，学者们采取了一系列新颖的编码技术，包括转录部分重复(transcript fractional repetition，简称TFR)编码[54]、流水线编码(Pipelined Code)[55]、新型网络编码（Network Coding,简称NC）[56]、多层次纠错编码（Error Correct Coding，简称ECC）[57]。

为了减少故障节点修复的延迟，一种基于表的非编码修复法-部分重复(fractional repetition，简称FR)编码被提出，它比再生编码具有更低的编码修复成本。但是，FR编码不能适应区块链网络中节点数量的不断变化。为了满足节点数量的变化、增强系统稳定性和减少冗余，文献[54]提出了TFR编码，在FR码基础上添加了复制级别来应对节点数量的波动，从而扩展了FR编码的灵活性。在不改变区块链现有的底层架构的前提下，文献[54]在覆盖网络层使用基于物理模糊距离的算法对节点进行了分组，从而减少了组内节点之间的通信延迟。

为了在存储开销和计算开销之间获得更好的平衡，文献[55]提出了一种基于时间的副本存储机制和流水线编码方案。其中，基于时间的副本存储机制将数据分为用于减少读取时的计算开销的热数据和用于释放存储空间的冷数据，降低了通信带宽，进而避免了网络拥塞。流水线编码则被用来构造存储函数基，降低了存储空间，提高了系统的容错性。

网络编码是一种将编码和路由信息交换的技术,可以对接收的多个数据包进行编码信息融合,从而提高单次传输的信息量,进而提升网络整体性能。为了解决区块链潜在的存储膨胀问题，文献[56]采用NC编码数据块来减少存储空间。基于NC的分布式存储(distributed storage ，简称NC-DS)有两种实现方式，分别是确定速率(deterministic rate ，简称NC-DRDS)，另一种是无速率( rateless，简称NC-RLDS) 。在NC中，使用一种随机线性网络编码(random linear network coding，简称RLNC)编码数据，而该种方式解码复杂度太高。为获得较低的解码复杂度的方法，NC-RLDS采用二进制域内随机移位(binary field random shift encoding)法降低了解码复杂度。

为了在减少存储空间的同时，降低延迟和解码复杂性，文献[57] 介绍了一种基于拓扑感知的多层次ECC编码方案。该方案支持在任意给定拓扑的异构中去中心化存储网络(Decentralized Storage Network，简称DSN)的节点之间进行分层协作的数据保护。与现有的网络编码方法相比，该方案有着更快的恢复速度，并且可以纠正更多的擦除模式。

### 2.3 密码学策略

密码学策略面对当前区块链数据存储空间太大、不可更改等困境，一些学者也在密码策略上提出了一些新想法，例如聚合签名(2篇，34%)、群签名（1篇，16%）、密码累加器(1篇，16%)等。

为了节省区块链的存储空间，文献[58]采用改进的RSA密码累加器替代Merkle树的方案，并且赋予了区块链成员证明、批量更新等的功能。密码累加器具有提供非成员证明、随意删除成员、减少数据存储内存等优点，但存在陷门。因此，文献[59]提出了利用RSA算法构造无陷门批量更新累加器的方法，提高了区块链的可扩展性。

学者们采用顺序聚合签名[59]、无证书聚合签名[60]，在减小区块链数据存储大小的同时，提高了区块链的安全性，赋予了区块链可编辑性。文献[59]提出了一个具有指定验证者的顺序聚合签名方案(sequential aggregate signature scheme with a designated verifier，简称DVSSA)，通过对所有人签名的顺序聚合压缩了区块链存储空间，并解决了非法访问问题。为了使区块链可编辑、提高验签效率，文献[60]重构了区块链结构，并结合了一种固定签名长度的匿名无证书聚合签名算法。只要需要编辑的用户数量达到一定阈值，便可以修改或删除区块链上的数据。另外，该方案还可以对恶意行为进行问责，并被证明数据编辑的有效性和安全性。

为了提高数据存储和监控的效率，文献[61]利用区块链和智能合约，建立了基于同态加密算法的系统。采用完全的基于同态加密的身份认证协议，系统可以实时验证区块链中任何变更操作。由于采用分布式身份认证和实时可扩展存储，提高了系统数据安全性和可扩展性。该算法降低了传输成本，提高了数据存储和监控的效率。

群签名是隐私保护过程的一个很好的加密工具，确保匿名性、认证性和可问责性。群签名使组的成员能够代表组生成签名，并且不能追溯到其签名者，除了特殊的打开方，它可以打开任何有效的群签名。群签名用于隐私身份信息保护，为了提高效率，我们选择Clarisse等人的短群签名方案来设计协议，因为它的签名大小和计算成本是标准模型中最好的方案。

为了高效地验证存储数据和保护数据隐私，文献[62]提出了基于边缘计算的简化群签名的概念。群签名可以保护隐私，确保匿名性、认证性和可问责性。而在边缘计算中，边缘设备和终端设备可以很容易地通过云服务建立长期的信任关系。因此，文献[62]提出的简化群签名舍弃了可追踪性，使得数据验证的时间复杂度大小仅为O(1)。采用简化群签名，极大地降低了数据验证成本。

传统区块链系统的数据存储方式公开透明，无法保证用户的隐私。 文献[63]将多权限CP-ABE和DMA-ABS(decentralized multi-authority attribute-based sig-atures，简称DMA-ABS)与区块链结合进行数据的存储和共享，保证了数据的完整性、机密性和可访问性，并且解决了基于CP-ABE的云存储的所有权问题以及基于区块链系统的多用户授权问题。

### 2.4 算法/方法论

目前，在算法层面的研究较为稀缺，只有极少数学者对数据压缩算法[64]、针对区块链存储不平衡问题的算法[65]进行了探索。为了防止汉字数据编码中隐藏的语义信息丢失，文献[64]利用改进的LZW(Lemple-Ziv-Welch)算法对医学文本数据进行有效无损压缩存储，缓解了区块链的存储压力，提供了高效的访问服务。LZW算法是一种基于字典的压缩算法，但不需要将字典存储在区块链中，最大程度地降低了存储空间的大小。

为了解决区块链数据存储不平衡问题，文献[65]使用了启发式算法（FMA）、定制遗传算法(Genetic Algorithm，简称GA)和禁忌搜索算法(Tabu Search Algorithm，简称TSA).其中，FMA用来有效地获得存储平衡问题的近似解、GA和TSA则用来处理非局部数据的存储问题。通过保持存储平衡，提高了区块链网络的可用性与稳定性。

在方法论层面，文献[66]首次提出了基于区块链的分布式存储技术的分类方法。文献[66]对基于区块链的最新分布式存储技术进行了全面调查，并根据提出的六个类别（成熟度、区块链用法、安全性、共识机制、冗余/容错机制和分权级别）进行了相互比较。

## 总结与展望

目前，区块链数据存储最大的困难便是数据膨胀问题。由于区块链的可追溯性，要求存储所有历史记录，这需要巨大的存储空间。针对这一问题，目前学者解决方案包括采用新的存储架构，例如结合链下存储系统或者双链、复合链等；采用编码技术，例如擦除编码、网络编码等；采用新颖的密码学策略，例如群签名、聚合签名等数字签名技术。

但当前的解决方案存在着相应的问题，链下存储会破坏审计的完整性，降低了区块链的安全性。编码技术的解码复杂度太高，导致修复效率低下，甚至会占用大量的网络带宽，严重时会导致网络拥塞。另外，由于区块链的数据存储方式，其不可变性和公共可访问性会引起安全性和存储效率以及数据查询可用性问题。未来学者们可以针对这些问题提出新的方案。

参考文献

[1] Nakamoto S, Bitcoin A. A peer-to-peer electronic cash system[J]. Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf, 2008, 4.

[2] Wood G. Ethereum: A secure decentralised generalised transaction ledger[J]. Ethereum project yellow paper, 2014, 151(2014): 1-32.

[3] Guo H, Zeng Y. Medical data storage management based on blockchain technology[C]//BASIC & CLINICAL PHARMACOLOGY & TOXICOLOGY. 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY, 2021, 128: 38-38.

[4] Zhao R, Bal J, Ma X. Creating Lean and Agile Supply Chains with Blockchain[C]//2021 The 3rd International Conference on Blockchain Technology. 2021: 135-146.

[5] Albeanu G. Blockchain technology and education[C]//The 12th International Conference on Virtual Learning ICVL. 2017: 271-275.

[6] Ali M S, Dolui K, Antonelli F. IoT data privacy via blockchains and IPFS[C]//Proceedings of the seventh international conference on the internet of things. 2017: 1-7.

[7] Grabis J, Stankovski V, Zariņš R. Blockchain Enabled Distributed Storage and Sharing of Personal Data Assets[C]//2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW). IEEE, 2020: 11-17.

[8] Wang S, Zhang Y, Zhang Y. A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems[J]. Ieee Access, 2018, 6: 38437-38450.

[9] Alizadeh M, Andersson K, Schelén O. Efficient Decentralized Data Storage Based on Public Blockchain and IPFS[C]//2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE, 2020: 1-8.

[10] Pham V D, Tran C T, Nguyen T, et al. B-Box-A Decentralized Storage System Using IPFS, Attributed-based Encryption, and Blockchain[C]//2020 RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE, 2020: 1-6.

[11] Kumar R, Marchang N, Tripathi R. SMDSB: Efficient Off-Chain Storage Model for Data Sharing in Blockchain Environment[J]. Machine Learning and Information Processing: Proceedings of ICMLIP 2020, 2021, 1311: 225.

[12] Zheng Q, Li Y, Chen P, et al. An innovative IPFS-based storage model for blockchain[C]//2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). IEEE, 2018: 704-708.

[13] Kumar R, Tripathi R. Implementation of distributed file storage and access framework using IPFS and blockchain[C]//2019 Fifth International Conference on Image Information Processing (ICIIP). IEEE, 2019: 246-251.

[14] Sohan M D S H, Mahmud M, Sikder M A B, et al. Increasing Throughput and Reducing Storage Bloating Problem Using IPFS and Dual-Blockchain Method[C]//2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2021: 732-736.

[15] Chou I T, Su H H, Hsueh Y L, et al. BC-Store: A Scalable Design for Blockchain Storage[C]//Proceedings of the 2020 2nd International Electronics Communication Conference. 2020: 33-38.

[16] Liu S, Wu J, Long C. Iot meets blockchain: parallel distributed architecture for data storage and sharing[C]//2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2018: 1355-1360.

[17] Tang J, Jia T, Chen H, et al. Research on Big Data Storage Method based on IPFS and Blockchain[C]//2020 2nd International Conference on Video, Signal and Image Processing. 2020: 55-60.

[18] Rupasena J, Rewa T, Hemachandra K T, et al. Scalable Storage Scheme for Blockchain-Enabled IoT Equipped Food Supply Chains[C]//2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2021: 300-305.

[19] Ali S, Javaid N, Javeed D, et al. A Blockchain-Based Secure Data Storage and Trading Model for Wireless Sensor Networks[C]//AINA. 2020: 499-511.

[20] Kawaguchi N. Application of blockchain to supply chain: Flexible blockchain technology[J]. Procedia Computer Science, 2019, 164: 143-148.

[21] Shah M, Shaikh M, Mishra V, et al. Decentralized cloud storage using blockchain[C]//2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184). IEEE, 2020: 384-389.

[22] Tang X, Guo H, Li H, et al. A DAPP Business Data Storage Model Based on Blockchain and IPFS[C]//International Conference on Artificial Intelligence and Security. Springer, Cham, 2021: 219-230.

[23] Kumar R, Tripathi R. Implementation of distributed file storage and access framework using IPFS and blockchain[C]//2019 Fifth International Conference on Image Information Processing (ICIIP). IEEE, 2019: 246-251.

[24] Sohan M D S H, Mahmud M, Sikder M A B, et al. Increasing Throughput and Reducing Storage Bloating Problem Using IPFS and Dual-Blockchain Method[C]//2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2021: 732-736.

[25] Chou I T, Su H H, Hsueh Y L, et al. BC-Store: A Scalable Design for Blockchain Storage[C]//Proceedings of the 2020 2nd International Electronics Communication Conference. 2020: 33-38.

[26] Liu S, Wu J, Long C. Iot meets blockchain: parallel distributed architecture for data storage and sharing[C]//2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2018: 1355-1360.

[27] Tang J, Jia T, Chen H, et al. Research on Big Data Storage Method based on IPFS and Blockchain[C]//2020 2nd International Conference on Video, Signal and Image Processing. 2020: 55-60.

[28] Rupasena J, Rewa T, Hemachandra K T, et al. Scalable Storage Scheme for Blockchain-Enabled IoT Equipped Food Supply Chains[C]//2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2021: 300-305.

[29] Ali S, Javaid N, Javeed D, et al. A Blockchain-Based Secure Data Storage and Trading Model for Wireless Sensor Networks[C]//AINA. 2020: 499-511.

[30] Kawaguchi N. Application of blockchain to supply chain: Flexible blockchain technology[J]. Procedia Computer Science, 2019, 164: 143-148.

[31] Shah M, Shaikh M, Mishra V, et al. Decentralized cloud storage using blockchain[C]//2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184). IEEE, 2020: 384-389.

[32] Tang X, Guo H, Li H, et al. A DAPP Business Data Storage Model Based on Blockchain and IPFS[C]//International Conference on Artificial Intelligence and Security. Springer, Cham, 2021: 219-230.

[33] Xie C, Sun Y, Luo H. Secured data storage scheme based on block chain for agricultural products tracking[C]//2017 3rd International Conference on Big Data Computing and Communications (BIGCOM). IEEE, 2017: 45-50.

[34] Wang J, Li S, Ma W. A Composite Chain Structure Blockchain Storage Method Based on Blockchain Technology[C]//Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Springer, Singapore, 2020: 96-103.

[35] Jia D, Xin J, Wang Z, et al. Optimized Data Storage Method for Sharding-Based Blockchain[J]. IEEE Access, 2021, 9: 67890-67900.

[36] Zhao P, Cheng H, Fang Y, et al. A Secure Storage Strategy for Blockchain Based on MCMC Algorithm[J]. IEEE Access, 2020, 8: 160815-160824.

[37] Marsalek A, Zefferer T, Fasllija E, et al. Tackling data inefficiency: Compressing the bitcoin blockchain[C]//2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2019: 626-633.

[38] Yu Z, Cai Y, Hong W. A Storage Architecture of Blockchain for Time-Series Data[C]//2019 2nd International Conference on Hot Information-Centric Networking (HotICN). IEEE, 2019: 90-91.

[39] Yu B, Li X, Zhao H. Virtual Block Group: A Scalable Blockchain Model with Partial Node Storage and Distributed Hash Table[J]. The Computer Journal, 2020, 63(10): 1524-1536.

[40] Ruj S, Rahman M S, Basu A, et al. Blockstore: A secure decentralized storage framework on blockchain[C]//2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA). IEEE, 2018: 1096-1103.

[41] Du Y, Duan H, Zhou A, et al. Enabling Secure and Efficient Decentralized Storage Auditing with Blockchain[J]. IEEE Transactions on Dependable and Secure Computing, 2021.

[42] Tian G, Hu Y, Wei J, et al. Blockchain-based secure deduplication and shared auditing in decentralized storage[J]. IEEE Transactions on Dependable and Secure Computing, 2021.

[43] Chen X, Zhang K, Liang X, et al. HyperBSA: A High-Performance Consortium Blockchain Storage Architecture for Massive Data[J]. IEEE Access, 2020, 8: 178402-178413.

[44] Zhang X, Wang H, Shi P, et al. LS4BUCC: A Low Overhead Storage Architecture for Blockchain Based Unmanned Collaborative Cognition System[C]//2019 IEEE International Conference on Service-Oriented System Engineering (SOSE). IEEE, 2019: 221-2215.

[45] Zhang Y, Ye J. Data Distributed Storage Scheme in Internet of Things Based on Blockchain[C]//International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy. Springer, Cham, 2020: 670-675.

[46] Mesnager S, Sınak A, Yayla O. Threshold-Based Post-Quantum Secure Verifiable Multi-Secret Sharing for Distributed Storage Blockchain[J]. Mathematics, 2020, 8(12): 2218.

[47] Wang J, Sun G, Gu Y, et al. Distributed Electronic Data Storage and Proof System Based on Blockchain[C]//CCF China Blockchain Conference. Springer, Singapore, 2019: 48-67.

[48] Qi X, Zhang Z, Jin C, et al. A Reliable Storage Partition for Permissioned Blockchain[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 33(1): 14-27.

[49] Khalaf O I, Abdulsahib G M. Optimized dynamic storage of data (ODSD) in IoT based on blockchain for wireless sensor networks[J]. Peer-to-Peer Networking and Applications, 2021: 1-16.

[50] Liang W, Fan Y, Li K C, et al. Secure data storage and recovery in industrial blockchain network environments[J]. IEEE Transactions on Industrial Informatics, 2020, 16(10): 6543-6552.

[51] Wang K. Intelligent Agricultural Information Remote Data Storage Method Based on Block Chain[C]//International Conference on Multimedia Technology and Enhanced Learning. Springer, Cham, 2020: 257-268.

[52] Raman R K, Varshney L R. Coding for Scalable Blockchains via Dynamic Distributed Storage[J]. IEEE/ACM Transactions on Networking, 2021.

[53] Xu Y. Section-blockchain: A storage reduced blockchain protocol, the foundation of an autotrophic decentralized storage architecture[C]//2018 23rd International Conference on Engineering of Complex Computer Systems (ICECCS). IEEE, 2018: 115-125.

[54] Qu B, Wang L E, Liu P, et al. GCBlock: A grouping and coding based storage scheme for blockchain system[J]. IEEE Access, 2020, 8: 48325-48336.

[55] Li X, Chen Y, Ba Y, et al. A Distributed Storage System Based on Blockchain and Pipelined Code[C]//International Conference on Smart City and Informatization. Springer, Singapore, 2019: 625-638.

[56] Dai M, Zhang S, Wang H, et al. A low storage room requirement framework for distributed ledger in blockchain[J]. IEEE Access, 2018, 6: 22970-22975.

[57] Yang S, Hareedy A, Calderbank R, et al. Topology-aware cooperative data protection in blockchain-based decentralized storage networks[C]//2020 IEEE International Symposium on Information Theory (ISIT). IEEE, 2020: 622-627.

[58] Wang J, Chen W, Wang L, et al. Blockchain-based data storage mechanism for industrial internet of things[J]. Intelligent Automation and Soft Computing, 2020, 26(5): 1157-1172.

[59] Ren Y, Leng Y, Zhu F, et al. Data storage mechanism based on blockchain with privacy protection in wireless body area network[J]. Sensors, 2019, 19(10): 2395.

[60] Li K, Zheng D, Guo R. An Anonymous Editable Blockchain Scheme Based on Certificateless Aggregate Signature[C]//2021 3rd International Conference on Natural Language Processing (ICNLP). IEEE, 2021: 57-67.

[61] Liang W, Zhang D, Lei X, et al. Circuit copyright blockchain: Blockchain-based homomorphic encryption for IP circuit protection[J]. IEEE Transactions on Emerging Topics in Computing, 2020.

[62] Wang Z. Blockchain-based Edge Computing Data Storage Protocol Under Simplified Group Signature[J]. IEEE Transactions on Emerging Topics in Computing, 2021.

[63] Li G, Sato H. A privacy-preserving and fully decentralized storage and sharing system on blockchain[C]//2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). IEEE, 2019, 2: 694-699.

[64] Du Y, Yu H. Medical Data Compression and Sharing Technology Based on Blockchain[C]//International Conference on Algorithmic Applications in Management. Springer, Cham, 2020: 581-592.

[65] Liu T, Wu J, Li J, et al. Secure and balanced scheme for non-local data storage in blockchain network[C]//2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2019: 2424-2427.

[66] Cangir O F, Cankur O, Ozsoy A. A taxonomy for Blockchain based distributed storage technologies[J]. Information Processing & Management, 2021, 58(5): 102627.