密码学原语如何应用？解析单向哈希的妙用

原创 廖飞强 微众银行区块链 4月29日

来自专辑

WeDPR隐私保护周三见

隐私数据如何验明真伪？区块链数据何以可信？如何快速检验海量数据是否被篡改？单向哈希在其中起到了什么作用？

隐私数据的价值很大程度上源自其真实性，如何防止数据被恶意篡改，是隐私保护方案设计中不可忽视的关键目标之一。为此，密码学领域提出了一系列基本组件，即密码学原语（Cryptographic Primitive）来实现这一目标，其中最常用的便是**单向哈希**。

在区块链中，单向哈希能够链接多个区块数据，形成可信的链式数据结构，在弱信任环境下，提供防篡改且经过多方共识的可信数据源。

这一特性对隐私保护方案的设计意义重大。隐私数据往往以密文形式表达，需要快速检验

海量隐私数据的真伪，查验是否被恶意篡改。此时，单向哈希作为一项关键技术，大有用武之地。

为何单向哈希如此神奇？其常见的用法有哪些？又能具体解决哪些问题？以下将据此一一展开。

单向的哈希算法

哈希算法是信息科学中的基础算法组件，“快速实现数据比较和效验”是其设计初衷之一。

现实业务场景中，可能会涉及海量隐私数据，逐一比对数据原文，在很多场景中非常不现实，尤其是需要通过网络传输的数据，会大大增加网络带宽的负担。

哈希算法的出现，使得高效的数据验证成为了可能。

哈希算法的核心功能为，将任意长度的输入m映射为固定长度的输出H(m)，H(m)常称为哈希值、散列值或消息摘要。

一个精心设计的哈希算法具有以下特征：

- **输出确定性**：同一种哈希算法，相同的输入，其输出固定不变。
- **输出长度不变性**：同一种哈希算法，针对任意长度的输入，其输出长度不变。
- **输入敏感性**：同一种哈希算法，即便输入数据有微小的改变，其输出哈希值也会发生巨大变化。

因此，只要比较数据的哈希值是否与预期的一致，就能大概率地判别隐私数据原文是否被篡改。其典型的实现有：各大主流编程语言中，HashMap数据结构所使用的哈希算法。

然而，只是大概率，在密码学协议中是不够的。我们需要更强的哈希算法，将实际的检验概率提升至接近100%。

与之对应的一个重要概念是『哈希碰撞』。哈希碰撞是指，存在两个不同的数据原文m1和m2，其哈希值完全相同，即H(m1) = H(m2)。

容易出现哈希碰撞的哈希算法在密码学协议中不安全，同时，密码学还进一步引入了单向性的要求。

一个密码学安全的哈希算法，在传统哈希算法的基础上，还需满足以下特性：

- **单向性**：根据数据原文计算哈希值很容易，但要求难以根据哈希值计算数据原文，提供计算上的不对称性，以此防止攻击者轻易地从哈希值反推出可能的隐私数据原文，保护哈希值的机密性。
- **抗碰撞性**：给定任意两个不同的数据原文，要求它们经哈希算法计算后得到相同哈希值的概率极低，以此防止攻击者轻易地为篡改之后的隐私数据原文构造出合法的哈希值，确保数据检验的有效性。

以上两个特性，赋予了密码学安全的哈希算法对数据内容**公开可验证的约束能力**。这一约束能力使得经过单向性转换获得哈希值，在一定程度上可以作为**隐私数据原文的等价信息**。

在隐私保护方案设计中，哈希算法的单向性是最常用的特性之一。相应地，密码学安全的哈希算法也常被称之为**单向哈希**。

目前主流的单向哈希有如下算法标准：

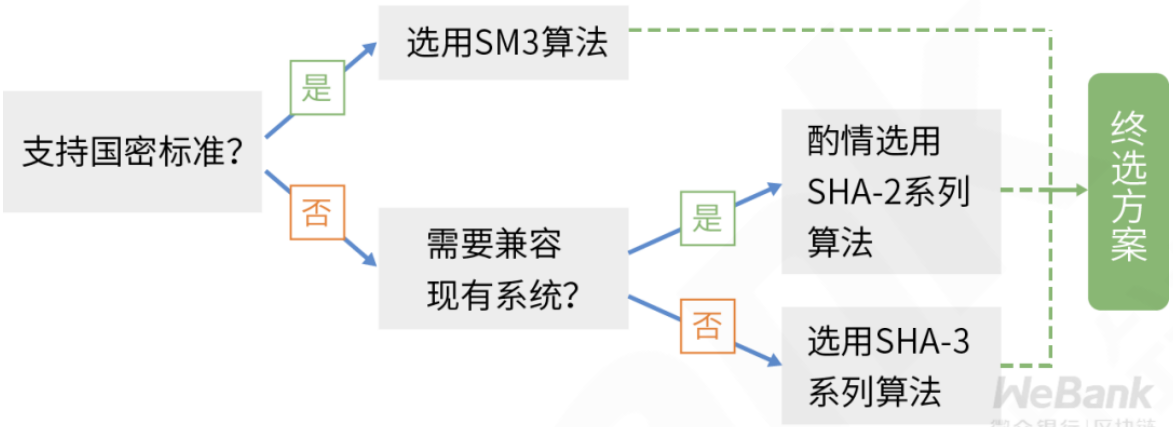
- **NIST标准**：SHA（Secure Hash Algorithm）系列是由美国国家标准技术研究院（NIST）制定的SHA哈希算法系列，主要包括SHA-1、SHA-2和SHA-3三个系列。**SHA-1由于安全问题已不再使用**。

  SHA-2系列包括SHA-224(表示哈希值长度为224位)、SHA-256、SHA-384、SHA-512等算法，**其中SHA-256是目前使用最广泛的单向哈希。**

  SHA-3是最新算法标准，源自参与SHA-3竞赛的keccak算法。2015年，NIST在完成SHA-3的标准化时，调整了keccak的填充流程，因此，标准的SHA-3算法与原始的keccak算法并不兼容。

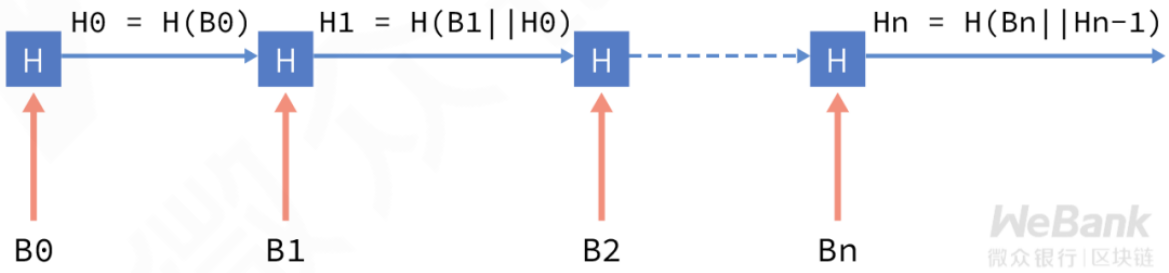
- **国密标准**：国密哈希算法SM3，是我国制定的单向哈希算法标准，由国家密码管理局于2010年12月17日发布，其安全性和效率与SHA-2系列的SHA-256相当。

单向哈希的选型可以参考业务部署的地域性要求，建议在SM3和SHA-3之间做出选择，如果需要与现有系统进行兼容，也可酌情选用SHA-2系列中的SHA-256。



0.2. 链式哈希结构

单向哈希的用途很广泛，最直接的应用就是构造链式哈希结构，即大家所熟知的区块链，提供难以篡改的可信数据源。



由于单向哈希的单向性，从结构上可以看出，从前一个数据块原文，很容易计算下一个数据块所用的哈希值输入，但已知一个哈希值输入，难以反推出所有可能的数据块原文。

区块链技术结合单向哈希和共识算法，当某一区块的数据共识确认后，下一区块将会记录

前一区块数据的哈希值，从而实现整条链上所有数据块的难以篡改。

在隐私保护方案设计中，以区块链为代表的基于链式哈希结构的可信数据源，可以起到简化协议设计的作用，尤其对于第4论中提到的恶意模型特别有效。恶意模型下的密码学协议，为了防范内部参与者不遵守协议、随意篡改数据，不得不引入复杂的多方交互验证过程。

通过链式哈希结构，在现实系统中引入一个可信数据源，可以对关键的中间流程数据进行存证和溯源，一旦有参与方作恶，便能在第一时间检测出，且定位到对应责任方，有效保障隐私保护方案全流程的正确性。

哈希树

单向哈希不仅仅能构造简单的链式哈希结构，还能根据业务需要扩展为更复杂的数据结构，其经典的形态之一便是哈希树。

哈希树常称为Merkle Tree，最早由Ralph Merkle在1979年的专利申请中提出，为大数据量的完整性验证提供了高效灵活的解决方案。

这里的**完整性验证**是指，核实原始数据在使用和传输的过程中没有被篡改。

在真实的隐私保护业务中，隐私数据多为高价值数据，而且多以密文的形态保存和使用，一旦被篡改，在不知道明文的前提下，难以通过常规技术手段来有效识别真伪。

对于涉及多方协作、联合计算的隐私保护业务，隐私数据密文交换和共享通常是其中的核心流程。所以当这些隐私数据密文跨越系统边界时，数据接收方会有两方面数据检验需求：

- **整体完整**：验证隐私数据中任意部分都未被篡改。
- **篡改定位**：如果存在攻击，能够有效定位被篡改的数据位置，便于开展应对流程。

为了体现哈希树的设计优越性，我们以举例的形式展示其效果。

为了简化说明，以下分析假定：

- 发送方与接收方之间存在一个带宽有限的可信信道，如区块链上经过共识的数据，可以将简短的哈希值安全地传递给对方。
- 隐私数据相关的原始文件由于数据量过大，不得不通过低成本低密级信道传输，如公共网络，因此可能被攻击者篡改。

方案1：整体单次哈希

本方案中，发送方在发送原始文件之前，将所有的原始文件数据作为哈希算法的输入，计算哈希值，然后将哈希值与原始文件均发送给接收方。

当接收方收到哈希值和原始文件后，重复发送方计算哈希值的操作，然后将新计算得到的哈希值与从网络上接收到的哈希值进行比较，如果相同，就可以判断原始文件在传输过程中未被篡改。

方案2：分块多次哈希 + 哈希树

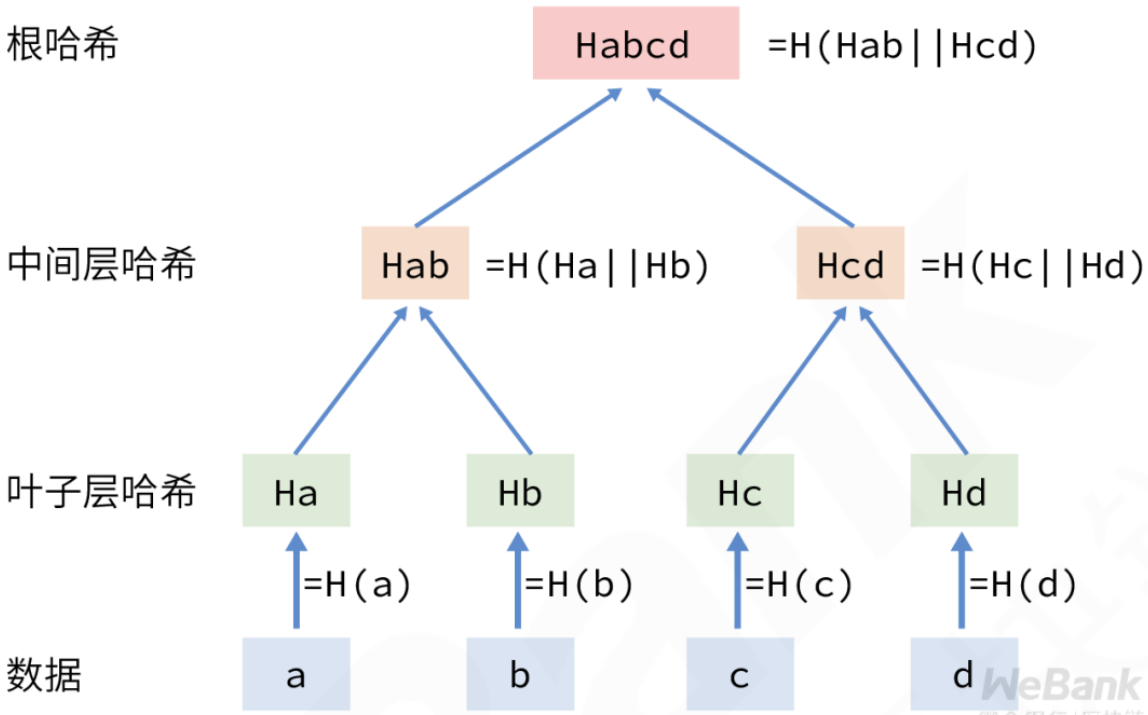
方案1对于满足整体完整需求十分有效，但对于第二条篡改定位需求就无能为力了。

基于哈希算法的输入敏感性，接收方可以知道至少有一个比特的数据被篡改了，但不知道具体在哪里。发送方不得不对所有数据进行重发，在这种情况下，攻击者很容易对隐私保护方案实施拒绝服务攻击。

为了解决这一点，本改进方案中，将原始文件分成一系列数据块，为每一个数据块分别计算哈希值。接收方验证的过程与方案1相似，区别在于可以对具体的数据块进行验证，一个数据块被篡改，导致的哈希值不匹配不会影响到其他数据块的验证，由此实现了篡改定位需求。

这里中间缺了关键的一步，即如何高效灵活地传输这些哈希值，并在原始文件很大时，灵活支持部分数据的获取和验证？

解决这些问题的要点，在于利用好哈希树的特性。

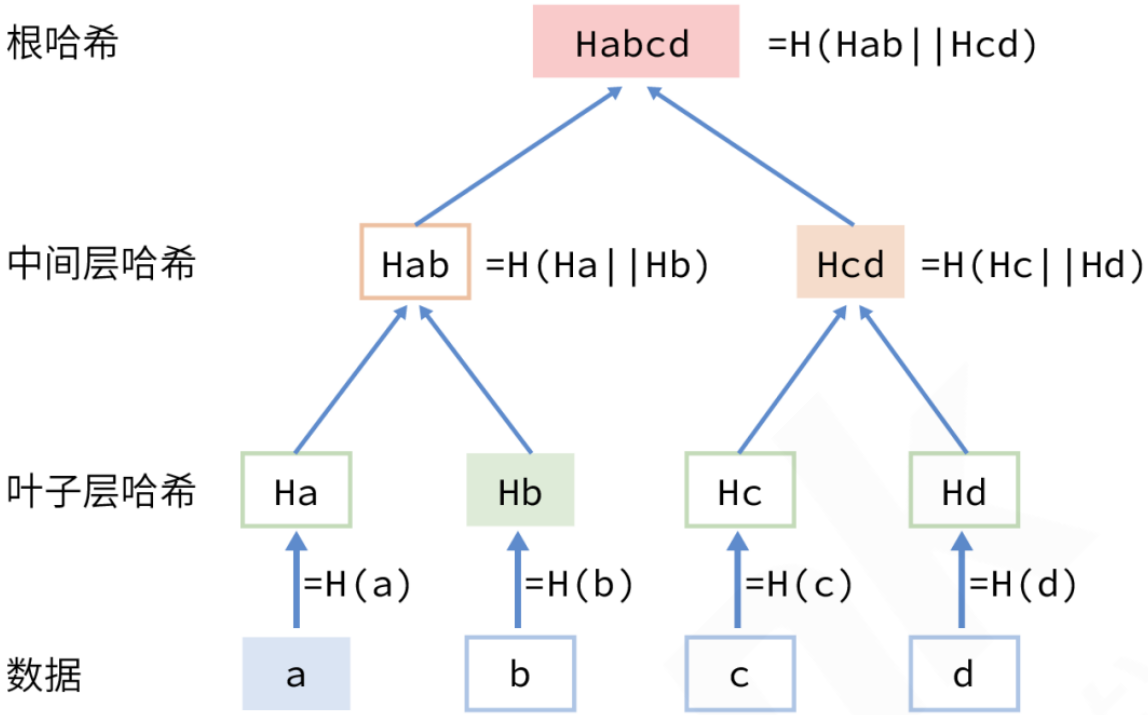


哈希树中，最底层的叶子层是各个数据块的哈希值，往树根的方向迭代哈希计算。即把相邻的两个节点的哈希值串连之后，再进行哈希运算，这样每两个哈希值就生成一个新的哈希值，重复以上计算过程，直到仅剩下一个哈希值（**根哈希**），最终形成一棵倒挂的树。

在哈希值传输方面，接收方只需要通过可信信道下载一个根哈希，其他数据都可以通过低成本低密级信道传输。

在支持部分数据的获取和验证方面，接收方只需要获取所需的部分数据块、根哈希，途经分支节点的哈希值，以O(log(n))的时间复杂度便可完成数据的验证，并实现被篡改数据块的快速定位。

除了哈希树之外，根据业务需求的差异，单向哈希还能用于构造有向无环图等更复杂的数据结构。**一般而言，其作用相当于连接各个数据点的锁扣，为相关数据建立公开可验证的密码学约束，使之难以被篡改，以此保障数据的正确性。**



如果只需要获取数据块 a 的内容，并验证其完整性

接收方只需要获得数据块 a 的原文、根哈希 Habcd 、途经分支节点的哈希值 Hcd 和 Hb ，进行以下验证：

H(H(H(a) || Hb) || Hcd)=? Habcd

正是：隐私数据真假难分辨，单向哈希守正不轻挠！

单向哈希是密码学中处于核心地位的密码学原语，可用于构建难以篡改的可信数据源、高效灵活的数据完整性验证机制等，以此来保障隐私保护方案中隐私数据的正确性。

本论中，我们介绍了单向哈希的基础应用，在往后的文章中，我们还会进一步介绍单向哈希的高级应用，包括构造密码学承诺、零知识证明等。

同时，作为密码学中久经考验的基本组件，除了单向哈希，密码学原语还包括数据编解码、对称加密、非对称加密、数字签名等，基础密码学原语还能进一步组成更高级的密码学组件。在这一系列中，我们将逐一展开与隐私保护密切相关的密码学原语的分享，欲知详情，敬请关注下文分解。

---END---

《隐私保护周三见》

“科技聚焦人性，隐私回归属主”，这是微众银行区块链团队推出《隐私保护周三见》深度栏目的愿景与初衷。每周三晚8点，专家团队将透过栏目和各位一起探寻隐私保护的发展之道。

栏目内容含括以下五大模块：关键概念、法律法规、理论基础、技术剖析和案例分享，如您有好的建议或者想学习的内容，欢迎随时提出。

栏目支持单位：零壹财经、陀螺财经、巴比特、火讯财经、火星财经、价值在线、链客社区

往期集锦

第1论 | 隐私和效用不可兼得？隐私保护开辟商业新境地

第2论 | 隐私合规风险知几何？数据合规商用需过九重关

第3论 | 密码学技术何以为信？深究背后的计算困难性理论

第4论 | 密码学技术如何选型？初探理论能力边界的安全模型

第5论 | 密码学技术如何选型？再探工程能力边界的安全模型

第6论 | 密码学技术如何选型？终探量子计算通信的安全模型

第7论 | 密码密钥傻傻分不清？认识密码学中的最高机密

第8论 | 密钥繁多难记难管理？认识高效密钥管理体系

上下滑动查看更多

长按二维码关注

微众银行区块链

白皮书下载 | 订阅干货 | 进群交流 | 合作联系