IBM Storage

为何区块链需要新的链下存储

文档版本号：4.1

IBM

目录

图表目录

引言

在考虑区块链技术的存储问题时，测试团队碰到了一个反复出现的主题。即，无需为链下数据提供新的存储，因为大多数企业已经在使用区块链会用到的数据。由于企业已经在使用这些数据，因此数据已经保存在存储环境内的某个位置，只需要通过区块链应用编程接口 (API) 的引用即可使用这些数据。本报告将探讨这一主题是否正确。

1 什么是区块链？什么不是区块链？

区块链是一个供相关用户联合体使用的分布式账本。这些用户可能是一个银行联合组织、一个贷款机构联合组织或者一个食品提供商联合会。每个用户都有分布式账本的完整副本，并且有连接到任意链下数据的链接。链下数据必须能够共享和访问，这样才能真正贯彻分布式账本的概念。

什么是区块链？

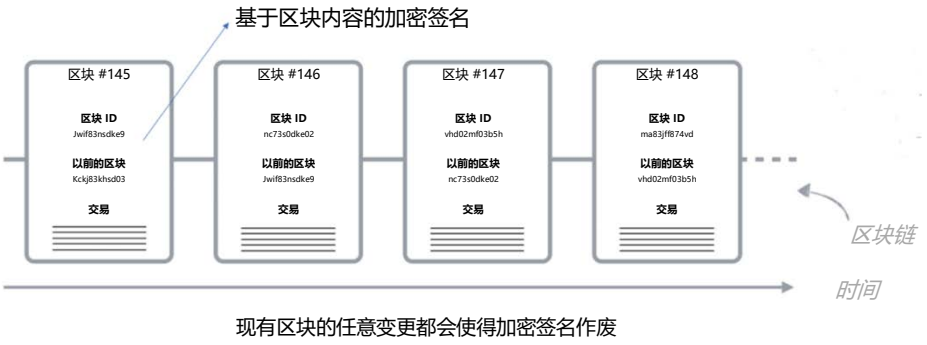

图 1：区块链结构示例

1.1 区块链技术的成本

您可以采用与计算存储容量相同的方式来计算不同区块链技术的成本。使用针对比特币和以太坊发布的每笔交易的成本数据，您可以将此成本与五节点 IBM Hyperledger 系统的成本相比较。

1.1.1 Hyperledger 成本

IBM Hyperledger 当前的基础企业级成本为 1000 美元/月，再加上每个主动节点 1000 美元的费用，这个五节点 IBM Hyperledger 系统每月的总成本为 6000 美元。其中不包含项目可能产生的其他软件即服务 (SaaS) 成本，因为这类成本可能因项目而异。

1.1.2 非许可区块链成本

在撰写本报告之时，比特币的交易成本为每笔交易 1.3 美元。但是，这种交易成本可能会随着基础比特币加密货币成本的变化而变化。以太坊的交易成本为每笔交易 0.25 美元，此交易成本也取决于基础以太币加密货币成本。

1.1.3 结果

非许可区块链的成本因交易速度和基础加密货币的当前成本而异。许可区块链（如 Hyperledger）的成本取决于节点数量和应用中使用的 SaaS 数量。这种情况下，许可区块链每月的成本是固定的，而非许可区块链每月的成本可能会差异巨大。根据每笔交易的费用，您可以按当前的每笔交易成本，轻松计算出使用不同区块链平台的成本。

图 2：区块链每月成本

如图 2 所示，不同的非许可区块链（如比特币和以太坊）的成本可能会是一个庞大的数字，即使在每秒交易速度很慢的情况下。

1.2 那么，什么是链下数据?

链下数据指的是因规模太大而无法高效存储在区块链上、或者需要能够变更/删除的任意非交易数据。图 3 展示了一些链下数据类型示例。

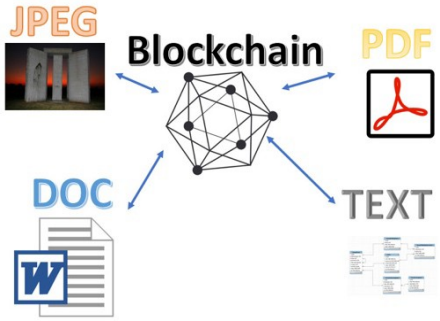

图 3：链下数据示例

1.3 所有这些数据都必须共享

比如，区块链不是只允许 A 公司看到 A 公司数据的孤岛式应用。区块链似乎运用了重复使用现有系统中数据的理念。若要将数据存储在区块链上或者让数据可以被区块链引用，数据必须在联合体的其他成员之间共享。图 4 用图形展示了由不同对等节点上的链下数据组成的扩展型数据集。

非共享型数据存储

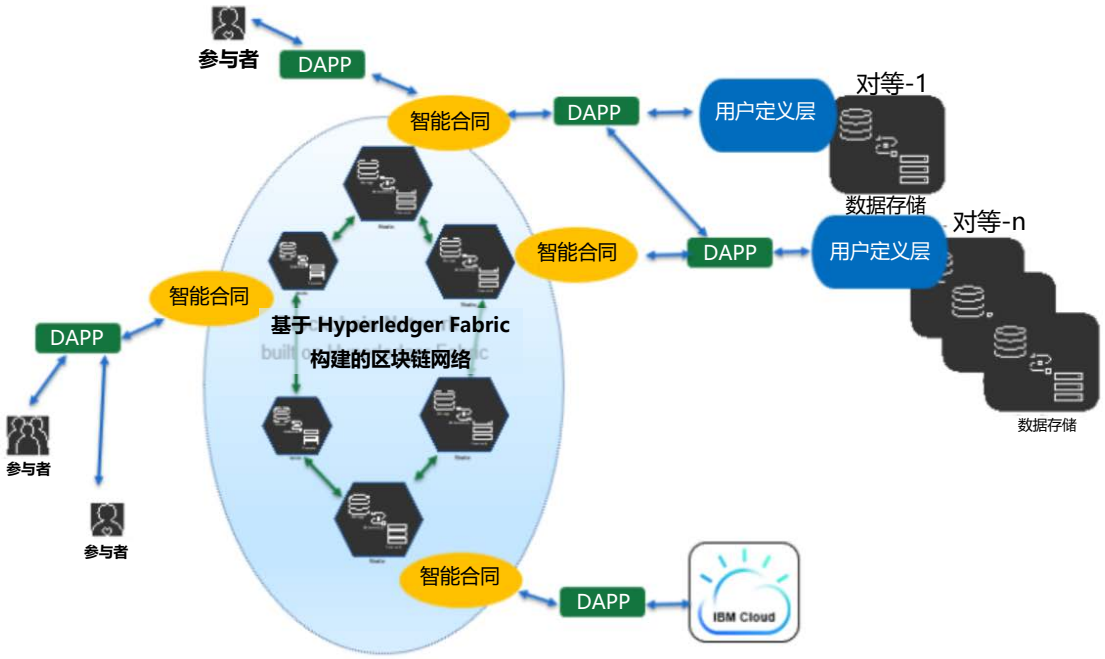

图 4：区块链数据流示例

1.4 链上存储计算

在大致了解存储技术以及区块链如何存储数据后，您就可以粗略地计算出 Hyperledger 区块链需要多少存储容量。

1.4.1 链上区块链假设

以下假设基于链上存储计算结果提出：

1. 所有计算结果都将是实际的字节，即，1024 B=1 KB
2. 所有 Hyperledger 区块大小为 1 兆字节 (MB)
3. 区块链中只存储哈希值、签名数据或密钥数据
4. 企业每天的工作时间为 8 小时
5. 企业一年大概有 240 个工作日，每秒的交易数仅为这些工作日（包括节假日）的平均值。
6. 目前，比特币每 1 MB 区块存储了约 1400 笔基础交易。Hyperledger 有更大的标头和交易规模，因此，它在每个区块上存储了 1000 笔交易。区块链交易目前每次运行的规模为 5 KB，相当于 205 笔交易/秒 (TPS)。

1.4.2 计算

根据上述假设，按以下公式计算每个 TPS 所需的存储容量：

(1 TPS/1000 TB)*1024 KB*3500 秒/小时*8 小时/天*240 天/年=

7,077,888 KB 的数据/笔交易/年

6,912 MB = 6.75 GB = .00659 TB/笔交易/年

鉴于交易规模可能随着区块链上存储的交易类型的变化而变化，您可以在上面的公式中填入不同的交易规模。

2 使用传统数据存储所带来的问题

在新的区块链项目中使用传统数据存储会带来几个严重的问题。

2.1 访问问题

如果每位联合体成员都必须提供传统数据存储的链接，而非提供一个隔离的数据存储来保存链下引用数据，情况将会怎样？每位联合体成员必须出资为所需的各种数据连接器、加密和安全应用编写程序。您可能允许其他联合体成员查看 XYZ 文档，但是如果 XYZ 文档没有存储在区块链上，那么他们肯定无法查看 XYZ 文档。

2.2 安全问题

在使用传统数据存储时，意外泄露机密数据的几率很高。此外，您还必须考虑一点：区块链环境的共享特性可能导致恶意攻击者利用区块链访问路径，蓄意入侵联合体成员自己的数据存储。

2.3 加密问题

即使您可以使用传统数据存储，您也需要调整这些数据存储，才能存储所需的哈希值、加密代码和密钥，才能向区块链提供证据，证明自上次访问以来数据没有任何修改。对于频繁修改的活动数据存储，哈希值和加密密钥的重新计算和存储是一项繁重的任务。

2.4 性能问题

使用传统数据存储还会带来另一个问题：整个联合体使用不同的数据库技术、存储硬件和服务器功能，使得确保多个独立的数据存储提供统一的响应时间变得困难重重。因此，联合体需要采用统一的硬件、软件和数据存储技术，才能实现所需的服务级别。

2.5 成功问题

有时，一个项目最大的问题是它太成功了。近期 Hyperledger 环境中的一个区块链开发项目上线，并且成功实现了 300% 的预期增长，于是性能和存储问题随之而来。显然，成功的项目需要弹性（可扩展且安全）的存储。如果使用传统数据存储，依赖数据存储的区块链应用的成功将对现有的应用及其性能产生何种影响？涌入的新数据（这些新数据的互联方式与外键和其他关系的互联方式可能有所不同）会对传统数据存储造成何种影响？这些问题必须得到解决，以免因区块链数据的侵入而导致现有应用出现故障或给出错误的结果。

下面，我们通过一个简单的计算，来看看使用链下存储将对成功的区块链项目的存储容量产生什么影响。我们对成功的定义将取决于每秒交易数量的增长，即，交易数量从 1 增长到 10、100 甚至更多，交易数量越多，项目越成功。在这个例子中，我们将假设每三笔交易就会生成一个链下文档。我们将取文档的平均规模，在这个例子中，文档类型无关紧要。同时，我们还将假定企业并非全年全天候 24 小时营业，但是，如果您希望计算全年全天候营业的企业的结果，您只需进行简单的外推即可。

因为非交易数据（如图片、合同、PDF 和个人信息等）不应该存储在区块链上，因此，您需要某种形式的链下存储或 sideDB 存储。通常，链下数据为非结构化数据。您应该生成链下项目的哈希值或签名，并将这些哈希值或签名存储在区块链中。链下项目本身则存储在云端或近云存储中。预计，市场对链下数据存储的需求将超过对区块链存储的需求。

2.5.1 链下数据假设

以下假设基于链下存储计算结果提出：

1. 每笔交易不会生成一个文档。因为每 3 笔交易生成一个文档，所以在计算中使用参数 0.3。
2. 企业在工作日一天营业 8 小时。
3. 一年大概有 240 个工作日。
4. 索引和其他存储要求未涵盖在内。

2.5.2 计算

基于前文提出的假设，进行以下计算：

1 TPS*0.3 个文档/笔交易*3600 秒/小时*8 小时/天*240 天/年=DPY（文档/年）

= 0.3*3600*8*240=814,301 DPY/TPS

2.5.3 保存文档所需的存储

根据现有参考信息，1,000,000 个（各种格式的）文档需要 333 GB 的存储容量。这意味着，814,301 个文档将需要：

1000000X/(333*814301) = GB/年/TPS

X = (333*814,301)/1,000,000

= 271 GB/TPS/年 = 0.264 TiB/年/TPS

因此，不出所料，您会看到如果每记录 3 笔交易就生成 1 个文档，那么随着每秒交易数量的增加，交易所需的存储很快就会比主区块链所需的存储多几个数量级。

2.6 新兴企业

所有的新颠覆性技术都将催生出新兴企业。新兴企业没有采用传统的数据存储。因此，为了参与区块链革命，他们需要采用新的存储。

2.7 新的业务流程需要新的数据组织

区块链业务应用解决方案通常会引发业务流程的再造。比如，在*支付*（多行业）领域，业务工作流程可能会发生改变，以便在区块链参与者社区中充分利用可用的潜在简化技术。这有助于显著革新机构的支付系统，而很多机构的支付系统已经使用了几十年。为了完成业务流程的再造，您还需要满足新的性能标准，比如达到*近乎实时的目标*。通常，大多数传统的数据存储都是（松散地）基于由相互关联的数据表（通过主键值和辅键值互连）组成的关系型模型。现有的应用能快速、轻松地访问这些表。区块链需要另一种形式的数据模型，这种数据模型更像数据仓库而非关系型数据库，这样，区块链就能充当中央*事实*表，而链下存储将充当*维度*表。区块链只存储事实（总数、集合、交易详细信息），这些事实源自链下存储的数据（文档、图片、PDF 文件和列表数据）之间的交集，或者会导致链下存储的数据产生交集。新的数据组织要求会导致效率下滑，因此，企业需要重组架构，并改变传统数据存储保存信息的方式。

2.8 GDPR 和数据粒化

欧洲的《通用数据保护条例》(GDPR) 以及在欧洲做生意的企业都会推动区块链应用对新的链下存储的需求增长。因为企业最好是将敏感信息保存在链下存储中，这样您可以在必要的时候，删除这些信息。

3 解决方案

要想全面、统一且安全地控制对链下数据的访问，并解决上述所有问题，您的唯一办法是构建一个由存储和服务器资源组成的共享网络，以便为区块链联合体成员提供所需的安全和共享环境。每当数据对象被访问时，它必须利用存储的哈希值进行验证，证明该对象与一开始存储的对象一模一样。每个对象应该存储在不只一个数据存储中，以确保某个节点的丢失不会导致大量数据丢失。此外，一旦节点在恢复后重新加入联合体，您则需要采取一种机制来同步链下引用，并重新平衡链下数据。为了满足这类要求，您需要采用新的区块链专用链下存储资产和新的存储访问模式。

存储计算结果表明，除非区块链是用于基于加密货币的少量交易，否则月费用固定的许可区块链（如 IBM Hyperledger）是您的最佳选择。

IBM United States of America
美国印刷

IBM 可能不会在其他国家或地区提供本文档中讨论的产品、服务或功能。有关在您所在区域提供的产品或服务的信息，请咨询您的当地 IBM 业务代表。任何对 IBM 产品、程序或服务的引用并不明示或默示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以用来代替 IBM 产品、程序或服务。但是，对任何非 IBM 产品、程序或服务的评估和验证均由用户自行负责。

IBM 拥有的专利或未决专利申请可能覆盖本文档中描述的主题。提供本文档并不意味着授予用户使用这些专利的任何许可。您可以用书面形式将许可查询寄往：

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

下列段落不适用于英国，以及本地法律与此类条款不一致的任何国家或地区：

INTERNATIONAL BUSINESS MACHINES CORPORATION “按现状”提供本出版物，不附有任何种类的（无论是明示的还是默示的）保证，包括但不限于暗含的有关不侵权、适销和适用于某种特定用途的保证。某些国家或地区在某些交易中不允许免除明示或暗含的保证。因此本条款可能不适用于您。

本文档可能包含技术方面不够准确的地方或印刷错误。本文所述信息可能会定期更改；这些更改可能会编入到本文档的新版本中。IBM 还可能会随时对本文档中所描述的产品和/或程序进行改进和/或更改，恕不另行通知。

本文档中对非 IBM Web 站点的任何引用都只是为了方便起见才提供的，并不以任何方式充当对这些 Web 站点的保证。这些 Web 站点中的资料不属于本 IBM 产品的资料，使用这些 Web 站点带来的风险将由您自行承担。

IBM 拥有的专利或未决专利申请可能覆盖本文档中描述的主题。提供本文档并未授予用户使用这些专利的任何许可。您可以用书面形式将许可查询寄往：

IBM Director of Licensing
IBM Corporation
4205 South Miami Boulevard
Research Triangle Park, NC 27709 U.S.A.

所有关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

本文档仅用于规划之目的。在所描述的产品上市之前，本文所述的信息可能会随时有所更改。

如果您查看的是本文档的软拷贝，则可能不会出现照片和彩色插图。

商标