基于大数据的安全技术分析*

■ 中国人民银行张家界市中心支行 谭正云

摘 要：随着大数据技术的不断发展和应用，许多传统的信息安全技术受到了挑战，金融领域也不例外。事实上，大数据本身就是解决诸多安全问题的一个重要工具，面对挑战，基于大数据的全新安全手段开始出现并得到发展。本文就基于大数据的安全技术展开探讨，阐述了基于Hadoop的大数据安全架构、基于大数据的威胁发现技术以及基于大数据真实性分析等技术要点，并提出只有在完整的安全体系指导下，金融信息安全建设所需的人财物等才能整合并发挥最佳效力。

关键词：大数据；安全技术；Hadoop；身份认证；K-means算法

一、引言

随着信息技术的发展，金融机构通过自己的网络和应用系统收集了大量用户信息，产生海量数据，金融机构可以基于这些数据挖掘出更有价值的信息，提高客户服务水平和经营管理水平。海量数据在采集、传输、存储和分析挖掘的过程中都会涉及信息安全问题。而大数据的“4V1C”特征，也使得大数据在安全技术、管理等方面面临新的安全威胁与挑战，“4VIC”特征如图1所示。

大数据在金融领域的实际应用存在诸多信息安全问题，值得业界思考。由于大数据技术本身就可以提供新的安全技术手段来解决安全问题，下面就对5类基于大数据的安全技术展开探讨。

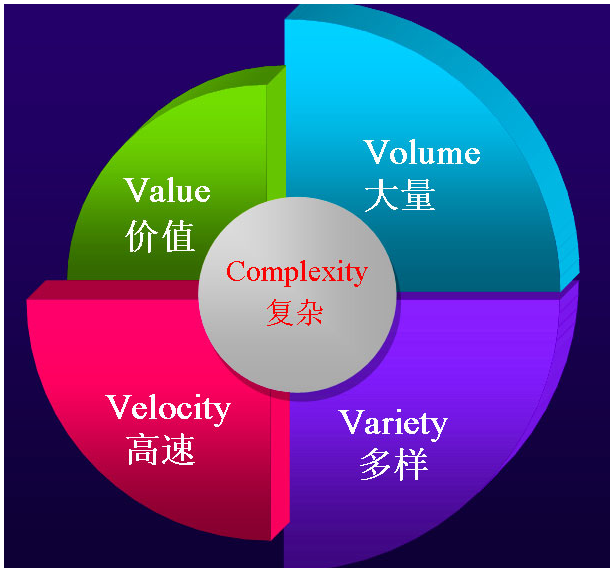

图1 大数据4V1C特征

作者简介：谭正云（1969-），男，湖南张家界人，工程师。
收稿日期：2017-01-17
*本文仅代表作者个人观点，不代表作者所在单位意见。

二、基于大数据的安全技术

（一）基于Hadoop的大数据安全架构

Hadoop是一种分布式的数据和计算框架，已成为全球范围内大数据应用最为广泛的技术架构，在金融业也被大量采用。

当前，金融领域和学术领域都大量采用Hadoop平台开展云计算大数据的应用研究。在不破坏大数据集群的基本功能及大数据本身必要特点的前提下，笔者先就此架构的安全问题及隐患进行分析，并给出相应的安全解决建议。

在分布式数据中，验证异构平台之间的安全和一致性是非常困难的，各个数据结点的安全性、结点之间的整体性和一致性是大数据分布式计算的一个痛点。而与传统集中式数据安全模型不同，存储的数据在集群内部流动，一个数据可能存在多个拷贝，它们在多个节点间移动来确保冗余，这种机制导致数据很难及时、准确地定位存储位置，也无法获知数据备份个数，加大了副本安全保护机制设计上的难度。在访问数据时，一般提供的访问控制最细粒度为schema级别，虽然在大数据环境中还有安全标签和其他高级属性可以借鉴，但这需要在应用和数据存储的设计建立时就考虑进去。Hadoop和大部分的组件没有建立起安全的通信机制，因为它们使用基于TCP/IP（传输控制协议）的RPC（远程过程调用协议），并没有嵌入TSL（安全传输层协议）和SSL等安全机制。另外，客户端可以直接与资源管理者及节点进行交互，增加了恶意代码或链接发送的概率，也难以保证客户端免受数据节点的攻击。这些都是基于HDFS架构的大数据环境的安全隐忧。HDFS架构如图2所示。

基于Hadoop的大数据架构，其安全机制可以通过以下4种方法和技术得以保证。

一是使用Kerberos进行节点验证。Kerberos是最有效的安全控制措施之一，可以集成到Hadoop基础设施

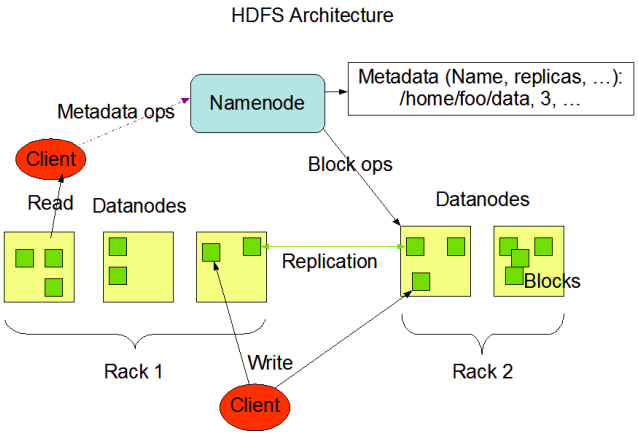

图2 HDFS架构

中。对于集群中的恶意节点和应用程序，它可以验证判别并及时阻断，从而保护管理通道不被攻击。

二是对于恶意客户端发起的获取文件请求，可以通过使用文件层加密对数据进行保护。因为文件是不可读的磁盘映像，不容易被恶意访问，同时，文件层加密还可以提供一致的安全保护，目前市面上已经有产品能做到内存加密保护，进一步提高了文件安全性。

三是运用密钥管理服务来管理大数据密钥和证书，通过该服务，为不同的应用程序和用户组分发不同密钥，确保文件被有效加密。

四是在节点之间、节点与应用程序之间使用SSL或TLS组件实现安全通信，设计、集成有效的安全通信机制和现成组件。

（二）基于大数据的威胁发现技术

借助大数据挖掘分析技术，金融机构可以主动发现威胁，从而超越传统的PDDR（保护、检测、响应、恢复）模式：比如众所周知的棱镜计划，如果换一个角度来理解，它就是运用大数据的挖掘分析主动发现威胁的成功案例：事先收集全球各地的海量数据，并整合、挖掘、分析，从而发现可能对当局造成威胁的因素，并在这些威胁尚未浮出水面时及时处理和解决。

大数据分析技术也为对抗APT（高级持续性威胁）攻击提供了新的解决手段。APT具有隐蔽能力强、针对性强、攻击手段多、攻击范围广和防范难度高等特点，技术高级，威胁性也大，如图3所示。

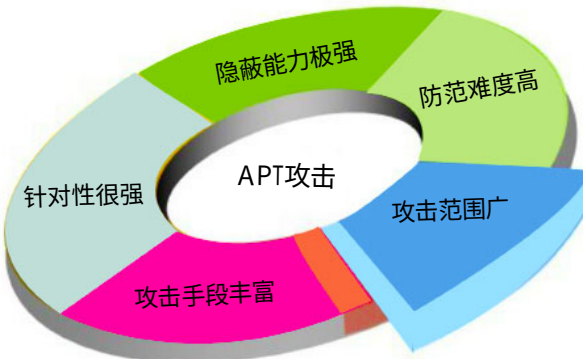

图3 APT特征

为了应对APT攻击，目前已经有沙箱方案、异常检测、全流量审计、深层协议解析异常识别、攻击溯源等方案。APT的潜伏期可能很长，因此，金融机构可以设定一定的时间，并对这个时期的数据进行挖掘分析，从而发现蛛丝马迹，找到攻击源并化解。

（三）基于大数据真实性分析技术

目前，学术界和金融界普遍认为，引入大数据技术的真实性分析是最为有效的方法，基于大量数据综合分析能有效提升真假信息甄别水平。例如，对于用户的银行卡消费行为，可以通过用户画像，来分析客户特征，为鉴别其各种行为的真实性提供参考和依据，如图4所示。

另一方面，引入人工智能的机器学习技术，建立和优化模型，可以进一步提升真假信息的鉴别能力，并随着机器学习和算法模型的进化而不断优化，甚至有可能超过人工鉴别能力。大数据时代的到来，定然会有更多更新、更丰富的安全技术应运而生。金融机构不可能置身事外，但金融机构的数据涉及保密问题，安全措施不能完全依赖外界，必须结合自己的技术特点，依托金融自身收集的大量数据，开展数据分析、建模，来提高信息甄别能力和安全管理水平。按照目前的趋势，将来大数据服务作为底层的技术基础，可帮助各金融机构搭建或定制自己专属的信息安全服务体系，提升金融领域的信息安全水平。

（四）基于大数据的身份认证技术

身份认证是金融业不可回避的问题，不论是对金融客户的身份识别，还是金融机构内部授权管理，都会涉及该问题。金融业传统的身份认证技术主要通过口令和数字证书等硬件来实现，但这个看似严密的身份认证体系其实也面临着安全问题。一是对于用户

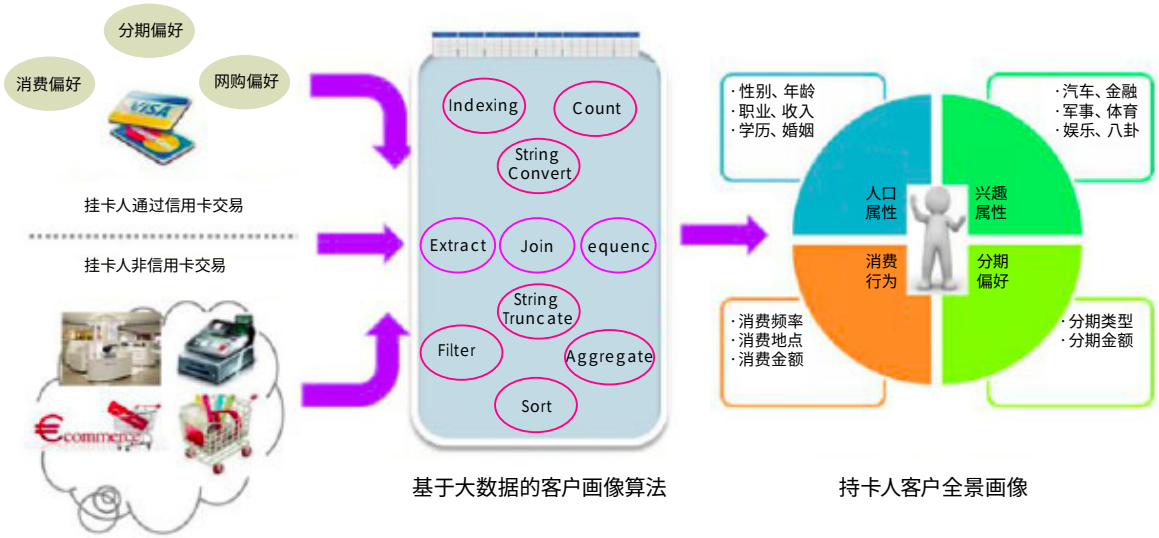

图4 基于大数据的持卡人画像

而言，攻击者总是能够找到方法来骗取本只有用户才知道的信息，比如密码口令和手机动态验证码，或者通过木马等方式直接截取用户的隐私信息，从而通过认证，展开攻击。

二是对于硬件安全而言，虽然增加了安全性，但也加重了用户负担（例如携带硬件USB Key），甚至当用户忘记携带相关硬件时，自身都无法通过验证，降低了便利性。即使是近年兴起的生物认证技术也存在部分缺陷，如生物信息（如指纹、掌纹等）被盗取后，客户无法修改自身信息，面临后续威胁。且生物识别的准确性也存在问题，如人脸识别随着年龄的增长而变化，指纹识别因手指受伤或划痕而无法通过验证，声音识别因咽喉嘶哑而不被系统认可等，而大数据可以提供多维度的身份识别，将用户的多种生物特征进行比对，同时结合用户的行为特征，提高身份识别准确性。

面对种种问题，金融机构只有在身份认证中引入大数据综合分析才能够有效地解决。综合分析用户行为特征、生物特征和设备特征来判定用户身份，如此一来，对攻击者而言，需要掌握用户方方面面的信息才能展开有效攻击，大大增加了攻击难度；对合法用户而言，则大大降低了认证失败的概率。这在提升身份认证的安全性，同时又保证用户身份认证的便利性。

（五）基于大数据的安全规则挖掘技术

在互联网中，为保证网络安全，金融机构会引入防火墙技术和入侵检测技术等。这些技术通常是通过建立一套安全规则或过滤规则达到其安全目标的，而建立这些规则的传统方法是通过专家知识系统，引入大数据技术后，安全规则可以通过数据挖掘等技术来探索、求证、使用。

在众多的挖掘算法中，聚类分析是一项应用较为广泛的技术，该算法把数据按照一定规则来实施分组。聚类算法的好坏判定标准主要在于组内相似性要高，组间差异性要大。在聚类算法中，K-means算法在金融等行业被广泛采用。但是该算法并非十全十美，依然存在一些缺陷，仍在逐步改进中。

一是K-means算法所涉及的两个关键要素：聚类数K和初始质心集，都需要人为选取，而这两个要素选取的标准，对该算法得出结果的质量有很大影响。对此，Stephen C.H.等人提出了改进算法，他们摈弃了人为选取初始条件的做法，改用基于密度的自动聚类，从而提升了K-means算法结果的质量，降低了K-means算法对初始条件和人为选取的过度依赖。

二是K-means算法仅适用于数据项全是数字的情况。这在很大程度上限制了K-means算法的应用范围。针对该问题，在借鉴K-means算法框架的基础上，GENGeng J.K.。提出了一种新的密度聚类算法，采用预抽样的方法将算法时间复杂度控制为线性，同时通过引入次质心的概念，解决聚类失效问题。分析表明，该算法能很好地克服K-means算法的初始条件敏感性和一般密度聚类算法的聚类失效问题，实现较为理想的聚类结果。

三、结束语

信息安全，三分靠技术，七分靠管理。随着大数据的快速发展和在金融行业的规模应用，新形势下的信息安全也面临诸多新挑战，在大数据产业链的各个环节，安全问题无处不在。而大数据本身就能提供新的安全手段，只有在正确完整的安全体系指导下，信息安全建设所需的人财物才能有机整合，提升金融业信息安全水平，也为大数据的发展和应用保驾护航。FTT

参考文献：

[1]林晓轩. 大数据时代下的信息安全治理[J]. 中国信息安全，2015(5):51-52.

[2]周洪.“大数据”时代背景下计算机信息处理技术的分析[J]. 信息与电脑，2015（12）:48-49.