大数据集群安全策略研究

Research on Security Strategy of Big Data Cluster

张　彬[1]，曹京卫[2]，刘　果[1]，李长连[1]（1. 中讯邮电咨询设计院有限公司，北京 100048；2. 中国联通智网创新中心，北京 100033）

Zhang Bin[1]，Cao Jingwei[2]，Liu Guo[1]，Li Changlian[1]（1. China Information Technology Designing & Consulting Institute Co.，Ltd.，Beijing 100048，China；2. China Unicom Intelligent Network Innovation Center，Beijing 100033，China）

摘　要：

数据中心在为应用系统提供数据服务的同时，需要保证其大数据平台的安全性。针对Hadoop大数据集群面临的安全威胁，从防火墙策略、用户认证、权限控制和日志审计等方面，提出了一套完整的大数据平台安全加固的解决方案，提高整个数据中心的安全性。

关键词：

数据安全；防火墙；用户认证；权限；日志审计

doi：10.12045/j.issn.1007-3043.2022.03.011

文章编号：1007-3043（2022）03-0058-06

中图分类号：TN918

文献标识码：A

开放科学（资源服务）标识码（OSID）：

Abstract：

Data center needs to ensure the security of big data platform while providing data service for application system. In view of the security threats faced by Hadoop cluster，it analyzes and studies the firewall policy，user authentication，authority control and log audit，and proposed a complete solution of security reinforcement for big data platform，which improves the security of the data center.

Keywords：

Data security；Firewall；User authentication；Permission；Log audit

引用格式：张彬，曹京卫，刘果，等. 大数据集群安全策略研究[J]. 邮电设计技术，2022（3）：58-63.

0　前言

随着《网络安全法》[1]和《国家网络空间安全战略》[2]的发布和实施，网络安全问题被提升到了国家战略高度。大数据平台作为用户核心业务和机密数据的载体，一方面要保证其平台内部的数据安全，另一方面，也要保障大数据平台与应用提供者之间的接口安全。

大数据平台建设大多采用分布式架构。作为大数据平台的标准解决方案，Hadoop成为目前最流行的大数据分布式系统架构[3]，已广泛应用于各个大数据产品生态体系。Hadoop平台由多个组件搭建而成，所以各个组件的安全共同影响并决定着大数据平台的安全。

目前，Hadoop存在以下几个方面的安全问题：内部网络攻击或越权访问[5]；缺乏必要的安全认证机制；权限管理依赖Linux账户的权限，控制能力较弱并且缺少细粒度的访问控制；缺乏统一的审计日志监控；无法保证数据传输安全。针对以上安全问题，本文从主机防火墙策略、Kerberos用户认证和加密、Apache Ranger细粒度权限控制和统一日志审计3个方面，提出一套完整的大数据平台安全加固的解决方案。

1　安全加固方案

1.1　防火墙安全方案

传统的防火墙已经将外网和内网隔离开[4]，但实

收稿日期：2022-02-18

际上，80% 的攻击和越权访问来自内部网络[5]。大数据集群一般部署在内部网络，并且大多数厂商会租用数据基地的服务器来部署自己的集群。数据基地和外网有专业的防火墙做隔离，但是内网之间，各个厂商所租用的服务器都要接受数据基地综合监控平台的监管，并没有做到完全的资源隔离，不能排除内网员工恶意攻击或者做一些不恰当的操作。此时就需要依赖服务器操作系统的主机防火墙来阻断内网中其他服务器的非法访问。

Hadoop 集群部署时，各个节点需要完全互信，并且集群运行时，各个服务占用端口比较多，集群在高负荷运行时，开启防火墙会影响集群的整体性能，所以官方文档建议所有主机关闭防火墙。但是，数据安全是提供任何服务的前提，遵循“默认拒绝，最小放开”的原则，在不影响服务通信的前提下，集群内所有服务器开启主机防火墙。对 Hadoop 集群内开放所有 IP 和端口访问权限；对集群外的主机，只开放某些 IP 下的具体服务端口访问权限。以 HDFS 和 Kafka 为例。

a）HDFS 客户端访问 HDFS 服务：对客户端放开用于获取文件系统 Metadata 信息的 8020 端口（core-site.xml 配置文件中参数 fs.default.name namenode）；对客户端放开用于数据传输的所有 DateNode 节点的 50010 端口（hdfs-site.xml 配置文件中参数 dfs.datanode.address）。

b）Kafka 生产者或者消费者访问 Kafka 集群：对客户端放开 Kafka 集群所有节点之间通信的 RPC 端口 9092（server.properties 配置文件中参数 listeners 或者 advertised.listeners）；对客户端放开连接 Zookeeper 的端口 2181（server.properties 配置文件中参数 zookeeper.connect）。

主机防火墙作为网络侧细粒度的访问控制机制，能够有效阻止主机外任何外网或者内网节点的非法访问。

1.2 用户认证和加密方案

用户身份认证是对访问者身份进行识别并确认的过程，是数据访问控制的基础，同时也是实现大数据安全架构的基础。Hadoop 集群默认采用基于操作系统账号的 Simple 认证，没有安全性保证，用户只需在客户端的操作系统上建立一个同名账号，即可伪装成任何用户访问集群[6]，一旦拥有 Hadoop 用户权限，便能随意查看、复制 HDFS 上的内容，或者进行其他操作[7]。Kerberos 作为一种可信的第三方认证服务[8]，无需主机操作系统认证，在非安全网络通信中，可使用 Kerberos 通过共享密钥的方式向另一个实体表明其身份信息[9]。

1.2.1 Kerberos 认证机制

使用 Kerberos 时，一个客户端需要通过以下 4 个步骤来获取服务。

a）认证：客户端向认证服务器（AS）发送服务请求。

b）授权：AS 生成 2 个票据（Ticket）返回给客户端，一个用客户端密钥加密，另一个用服务端密钥加密。

c）服务请求：客户端向服务器出示服务 Ticket，证实自己的合法性。

d）验证服务器：服务器给客户端回复一条消息，客户端比较服务器返回的时间戳是否一致，来验证服务器的真实性。

Kerberos 支持单点登录，即当客户端通过了 Kerberos 的认证后，便可以访问多个服务实体。

1.2.2 Kerberos 服务高可用方案

在实际生产环境中，集群规模都比较大，部署单个节点的 Kerberos 认证服务往往会出现单点故障问题。对于一个启用了 Kerberos 服务的 Hadoop 集群来说，KDC 的高可用是必须要考虑的。Kerberos 服务支持一主多备的模式，通过 kprop 服务将主节点的数据同步到各个备节点。Kerberos 服务高可用认证系统架构如图 1 所示。

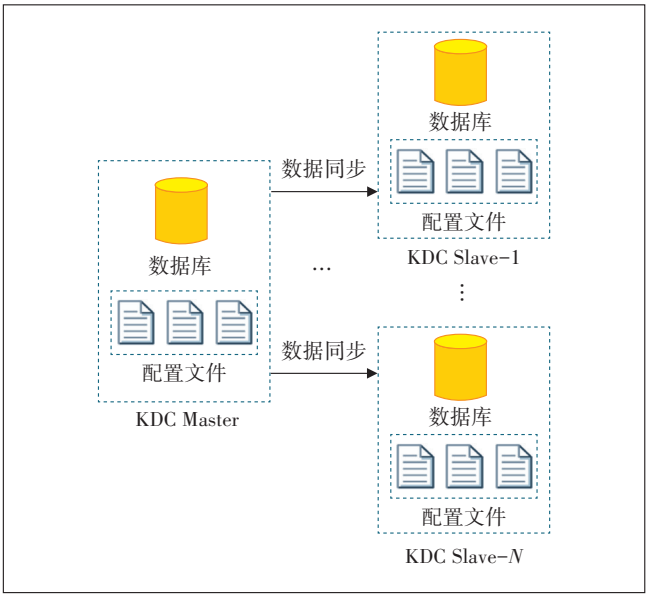

图 1　Kerberos 高可用架构

Kerberos 中每个 KDC 都包含数据库的副本。主

KDC包含域（Realm）数据库的可写副本，它以固定的时间间隔复制到从KDC中。所有数据库更改（例如密码更改）都在主KDC上进行，当主KDC不可用时，从KDC提供Kerberos票据给服务授权，但不提供数据库管理。

Kerberos的同步机制只复制主数据库的内容，相关配置文件必须手动复制到每个Slave中，Kerberos服务配置高可用的主要步骤如下。

a）主备节点安装Kerberos服务。

b）主节点Kerberos配置文件修改。

c）主节点配置文件拷贝至备节点。

d）主节点数据同步至备节点并配置同步任务。

Kerberos高可用服务为Hadoop提供了较强的认证及授权保护，所有节点必须通过认证确认身份后才能访问集群资源，有效避免了针对Hadoop集群的恶意使用或篡改。

1.2.3 Kerberos数据加密

客户端和服务端相互认证阶段使用的是长期密钥，而认证结束后，客户端和服务端都会获取到认证服务器随机生成的临时会话密钥，此密钥将作为临时通信密钥来加密需要传输的内容，一段时间后会话密钥会过期，需要重新申请，这样保证了数据传输的安全性。

1.3 Ranger权限管理和日志审计方案

大数据集群中各个组件的访问权限控制都依赖于HDFS分布式文件系统的权限控制。HDFS的权限设计基于POSIX模型，权限划分为用户、用户组以及其他用户的读写执行权限。每个用户使用不同的Linux账户便能访问到相应文件。若用户已获得HDFS NameNode地址和端口号，在HDFS客户端安装完成后，使用与NameNode相同的用户名，即可获得任意文件的访问权限，这就无法保证各个组件的数据安全性。

Apache Ranger作为一个集中式安全管理框架，能为Hadoop生态组件提供操作、监控、管理复杂数据权限的能力[10]，并且可以对用户的行为日志进行统一的审计管理[11]。

1.3.1 Ranger权限管理

细粒度权限管理在数据级别没有共性，不同的组件对应的业务资源是不一样的，细粒度权限控制如表1所示。

Ranger基于策略来抽象出用户、资源以及权限之间的关系，从而进一步延伸并形成自己的权限模型。

表1　细粒度权限控制表

业务组件	资源	权限项
HDFS	FilePath	Read Write Execute
HBASE	Table Column-family Column	Read Write Create Admin
HIVE	Database Table Column	Select Create Update Drop Alter Index Lock Read Write All
YARN	Queue	submit-app admin-queue

用户：用用户（User）或者用户组（Group）来表示。

资源：用（服务，策略）二元组来表示；一条策略对应一个服务，并且是唯一对应，但一个服务可以对应多个策略。

权限：用（AllowACL，DenyACL）二元组来表示，AllowACL代表允许执行，DenyACL中则代表拒绝执行。

用户访问决策树如图2所示。

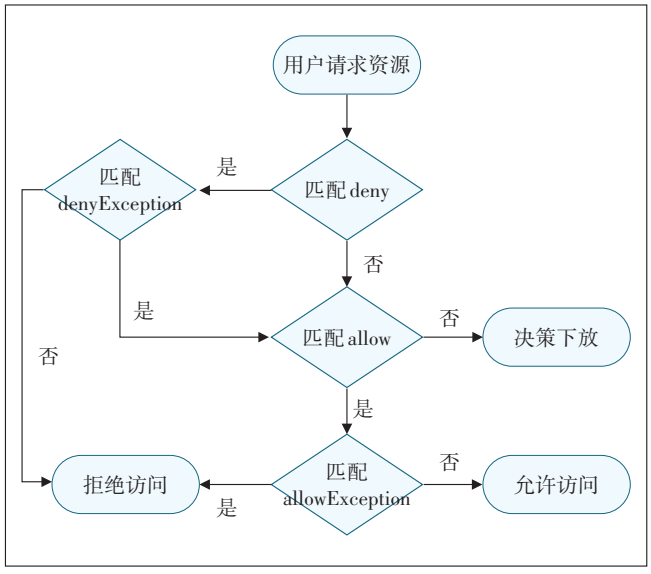

图2　访问决策树

一条策略有4组决策项：allow、allowException、deny、denyException，优先级由高到低依次是：denyException、deny、allowException、allow。如果没有策略能决策访问，一般情况是认为没有权限拒绝访问，此时Ranger会将决策下放给系统自身的访问控制层，如HDFS分布式文件系统的权限控制层。

1.3.2 Ranger Audit日志审计

大数据集群各个组件都可以配置审计日志，这些审计日志会以文件的方式存储在各个节点上，查询审计日志比较繁琐，不利于追踪、定位用户发起的任务。

Ranger在控制访问权限的同时会记录所有用户访问集群资源的行为并写入日志，这些审计日志统一存

储在HDFS文件系统中。Ranger提供一个统一的审计管理平台来展示和统计这些审计日志，审计管理包含以下6个部分。

a）Access：Access页为管理员提供所有已开启审计策略的服务活动数据，记录用户试图访问资源的行为和决策结果（允许或者拒绝）。

b）Admin：Admin选项卡记录管理员操作Ranger Web UI的所有事件，包括创建、删除、修改策略等操作。

c）Login Sessions：Login Sessions页面记录每个用户登录Ranger Web UI的会话信息。

d）Plugins：插件选项卡显示安全代理的上载历史，用于检查组件是否成功地与Ranger通信。

e）Plugin Status：插件状态选项卡显示每个插件的有效策略，包括相关的主机信息以及插件下载和开始执行策略的时间。

f）User Sync：用户同步页面记录从Unix或LDAP同步的用户和组信息，默认1 min同步1次。

通过Ranger审计管理界面可以看到详细的用户操作日志信息，方便管理员监控、查询用户的历史操作，做到有证可查。

2 安全加固方案部署

2.1 防火墙策略部署

本节以访问HDFS文件系统为例对防火墙策略部署进行说明，具体如图3所示。

假设数据采集区的采集服务器从公网采集过来的数据要写入核心数据区的HDFS文件系统，首先要开放kerberos服务的kdc端口88和admin_server端口749，用作kerberos用户认证，其次需要核心数据区的NameNode节点对数据采集区开放用于获取文件系统Metadata信息和接收Client端RPC连接的端口8020，另外还要对数据采集区的节点开放所有DateNode的50010端口，此端口是DateNode服务端口，用于数据传输。

以“默认拒绝，最小放开”的原则，既能保证用户的访问，又最大限度地保证网络安全。

2.2 Kerberos认证部署

为了确保Hadoop集群的安全，集群内每个节点都必须设置Kerberos认证。首先，选择安全且独立的节点安装并配置Kerberos服务；然后，将集群中所有服务、节点和用户标识都保存至Kerberos数据库中；最

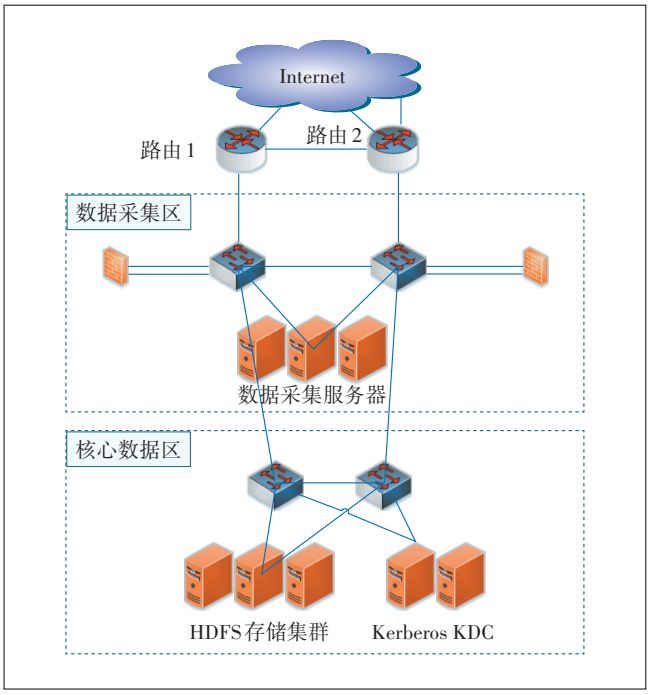

图3　数据采集拓扑图

后，修改所有节点的Hadoop配置，配置文件中加入Kerberos认证，方可启动Hadoop集群。详细配置流程如下。

a）安装KDC，选择2个独立节点进行安装，要保证绝对安全；在安装成功后，需要修改3个配置文件：krb5.conf、kdc.conf和kadm5.acl，包括配置KDC所在的位置、管理服务器、主机名与Kerberos领域名之间的映射等。最后，在所有节点上安装Kerberos客户端。

b）配置Hadoop服务标识。在KDC主节点中分别创建HDFS、Mapred、Yarn三者的对应标识，以确保Kerberos认证Hadoop的守护进程，然后创建HTTP服务标识。

c）为Hadoop服务创建Keytab文件。Keytab文件中包含1个键值对：Kerberos标识、基于Kerberos密码生成的加密密钥。该文件的作用是当服务在后台运行时，在没有人机交互的情况下进行认证。

d）启用KDC主节点kprop服务将主节点的数据同步到KDC备节点。

e）向集群节点分发Keytab文件。每个节点都需要创建相应的Keytab文件，在创建完成后需将Keytab文件移动到节点的/etc/hadoop/conf目录。

f）修改Keytab文件权限，确保Keytab文件的所有者才可以进行查看。

g）设置Hadoop配置文件，Kerberos生效。在更新

配置文件前，要先关闭集群，并重新设置core-site.xml和hdfs-site.xml配置文件。

在Hadoop开源社区提供的大数据平台框架上部署Kerberos认证服务较为复杂，如果使用Ambari搭建的HDP集群，部署Kerberos认证会简便很多。Ambari服务会自动为Hadoop集群中的服务创建用户密钥和Keytab文件，并协助用户修改Hadoop配置文件。

2.3 Ranger权限管理和日志审计部署

Apache Ranger是Hadoop生态中的安全管理框架，主要由Hortonworks开源和维护，和Hortonworks HDP结合得比较好，通过Ambari服务，用户可以直接安装Ranger服务和相关插件。使用原始的方式部署Ranger，需要从官网下载源码手动编译，并修改大量的配置文件。简单介绍一下使用Ambari集成Ranger服务的注意事项。

a）Ranger服务需要依赖第三方的数据库如MySQL、Oracle、Postgres等，建议使用集群依赖的存储元数据的数据库。

b）Ranger服务节点需要安装数据库客户端。

c）建议在数据库中提前创建好Ranger管理用户、密码和database。

d）如果使用MySQL作为数据库，请修改MySQL配置文件my.cnf，添加skip_ssl来关闭SSL验证，修改后重启MySQL服务

e）执行ambari-server setup指定配置Ambari-Server的JDBC驱动信息。

f）Ranger deny策略默认不会启用，如果需要启用，需要在Ranger部署完后在Ambari中Ranger配置下添加：ranger.servicedef.enableDenyAndExceptionsInPolicies=true。

使用Ambari管理界面配置Ranger服务较为简单，这里不再介绍。

3 方案实验验证

3.1 实验目的

通过Hadoop客户端上传数据文件到HDFS文件系统模拟数据采集流程，验证本文大数据平台加固方案的安全性，用户访问控制需满足如图4所示的流程。

3.2 实验环境

使用3台虚拟机组成的Ambari HDP集群作为验证环境，另外1台虚拟机模拟数据采集服务器访问Hadoop集群，实验环境和实验拓扑如表2和图5所示。

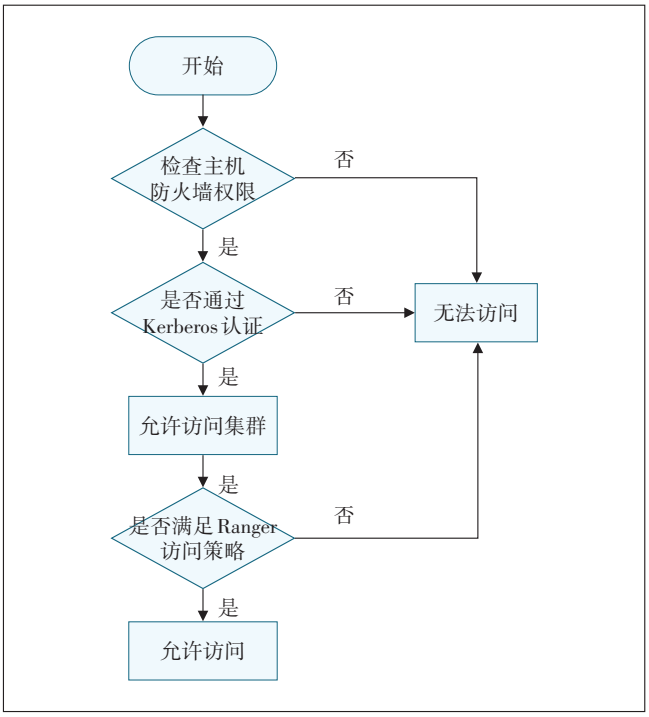

图4　用户访问控制流程图

表2　实验环境

配置名称	配置信息
操作系统	CentOS 7.4
Ambari版本	2.6.2.2
HDP版本	2.6.5.0-292
Hadoop版本	2.7.3
Ranger版本	0.7.0
Kerberos版本	1.15.1-8
防火墙软件	firewalld-0.4.4.4-6.el7
KDC节点	hdp1、hdp2
NameNode节点	hdp1
DataNode节点	hdp1、hdp2、hdp3
数据采集节点	data1

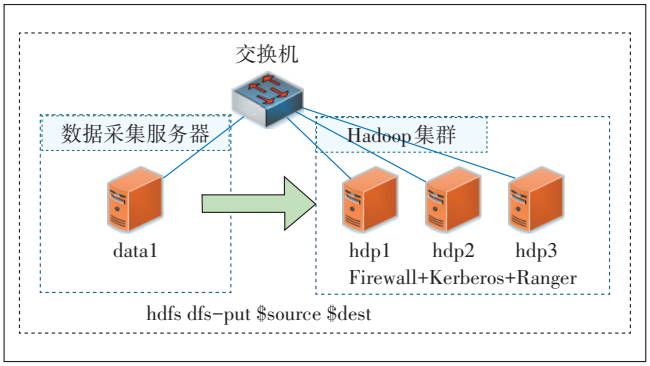

图5　实验拓扑图

3.3 实验操作步骤

实验操作步骤如下。

a）集群内的3台服务器全部开启主机防火墙，默认拒绝所有节点访问。增加访问策略如下：集群内的3台服务器可互访；允许采集服务器data1访问hdp1和hdp2中的Kerberos认证端口88和749；允许采集服务器data1访问hdp1的8020和50010端口；允许采集服务器data1访问hdp2和hdp3的50010端口。

b）为采集服务器data1部署hadoop客户端，保证其能正常访问执行HDFS命令。

c）为采集服务器data1新增Kerberos认证用户hdfstest，并生成Keytab文件；同时将Keytab文件拷贝到data1节点。

d）使用hdfs用户登录hdp1节点，创建HDFS目录/hdfstest，并修改权限为700。

e）登录Ranger管理界面，新增Ranger内部用户hdfstest并添加至用户组hadoop。

f）新增hdfstest用户访问策略，允许访问/hdfstest目录的读写权限。

3.4 验证方式及实验结果

在主机防火墙启用但未开放端口策略的前提下，执行HDFS上传文件命令到/hdfstest目录失败，无法访问目标主机；采集服务器data1节点在没有使用Kerberos认证的情况下，执行HDFS上传文件命令到/hdfstest目录，访问失败，没有有效的证书；禁用Ranger内部用户hdfstest的访问策略，使用Kerberos分配的证书认证后，能够识别用户身份，但用户没有权限访问/hdfstest目录；启用Ranger内部用户hdfstest的访问策略，使用Kerberos分配的证书认证后，上传文件到/hdfstest目录成功；停掉主KDC的进程，使用Kerberos分配的证书认证后，上传文件到/hdfstest目录成功。

实验结果满足图4中用户访问控制流程。在停掉主节点KDC进程后，客户端仍然能够通过Kerberos认证并上传数据文件，验证了KDC的高可用性；登录Ranger管理界面后，能够看到详细用户访问审计日志，方便管理员监控、查询用户的历史操作。

实验结果符合预期目标，验证了方案的可行性和有效性，并且该方案已应用到实际生产环境中，能够满足大数据平台的安全需求。

4 结束语

随着Hadoop生态系统的广泛应用，安全漏洞日益增多，本文从主机防火墙策略、Kerberos用户认证和加密、Apache Ranger细粒度权限控制和统一日志审计3个方面，提出一套完整的大数据平台安全加固的解决方案，并通过实验验证该方案的可行性与有效性。下一步将研究针对大数据平台的系统风险评估和安全预警，并提出相应的安全加固方案。

参考文献：

[1] 探讨《网络安全法》出台的重大意义[EB/OL]. [2021-12-24]. http://www.cac.gov.cn/2016-11/07/c_1119866702.htm.
[2] 《国家网络空间安全战略》全文[EB/OL]. [2021-12-24]. http://www.cac.gov.cn/2016-12/27/c_1120195926.htm.
[3] 刘明辉，陈湉，王竹欣. 大数据平台安全威胁与防护技术研究[J]. 信息技术与网络安全，2018，37(1)：65-69.
[4] 郑晓娟. 安全网络构建中防火墙技术的研究与应用[J]. 网络安全技术与应用，2016(3)：25-25.
[5] 曹宇. 用分布式防火墙构建网络屏障[J]. 网络安全技术与应用，2004(2)：18-21.
[6] 中国版大数据“哨兵”观数科技亮相BDTC2016[EB/OL]. [2021-12-24]. https://www.sohu.com/a/123508670_464072.
[7] 王玉龙，曾梦岐. 面向Hadoop架构的大数据安全研究[J]. 信息安全与通信保密，2014(7)：83-86.
[8] 齐忠厚. Kerberos协议原理及应用[J]. 计算机工程与科学，2000，22(5)：11-13.
[9] What is Kerberos [EB/OL]. [2021-12-24]. https://Web. mit. edu/kerberos/.
[10] Apache Ranger[EB/OL]. [2021-12-24]. http://ranger.apache.org/.
[11] 王文杰，胡柏青，刘驰. 开源大数据治理与安全软件综述[J]. 信息网络安全，2017(5)：28-36.
[12] 张立强，何凡，叶卫军，等. 一种基于Kerberos扩展的Web服务安全框架[J]. 武汉大学学报(理学版)，2017，63(2)：95-101.
[13] 毕淏，程晓荣. Kerberos认证协议分析与研究[J]. 电脑知识与技术，2017，13(27)：37-38，59.
[14] 王嘉龙，台宪青，马治杰. 大数据环境下基于用户属性的细粒度访问控制[J]. 计算机工程与设计，2020，41(7)：1801-1808.
[15] 吴晓琴，黄文培. Hadoop安全及攻击检测方法[J]. 计算机应用，2020，40(z1)：118-123.
[16] 陈丽，黄晋，王锐. Hadoop大数据平台安全问题和解决方案的综述[J]. 计算机系统应用，2018，27(1)：1-9.
[17] 张晶，李洪洋，张智钧，等. 大数据时代数据安全治理的网络安全策略[J]. 网络安全技术与应用，2021(1)：67-68.

作者简介：
张彬，毕业于昆明理工大学，工程师，硕士，主要从事网络安全技术及大数据方向的研究工作；曹京卫，毕业于北京科技大学，高级工程师，学士，主要从事互联网优化、网络安全技术的研究工作；刘果，毕业于武汉理工大学，工程师，学士，主要从事网络安全技术方向的研究工作；李长连，毕业于西北工业大学，高级工程师，硕士，主要从事网络安全技术方向的研究工作。