文章编号:1671-4598(2016)04-0149-03 DOI:10.16526/j.cnki.11—4762/tp.2016.04.044 中图分类号:TP301 文献标识码:A

基于公钥基础设施的 Hadoop 安全机制设计

陈 卓,王有春,平佳伟
(上海航天电子技术研究所,上海 201109)

摘要:为提高私有云平台的安全性,将云平台应用于航天领域,对现有基于 Hadoop 的云平台的安全机制做了深入的研究,分析了 Hadoop 官方团队 apache 推出的 Kerberos 身份认证体系,详细介绍了 Kerberos 安全体系的原理和在 Hadoop 中的工作流程,指出了 Kerberos 体系存在的过度依赖 KDC,采用对称密钥加密体制,客户端与 Hadoop 分布式文件存储系统(HDFS)的网络接口通过明文传输数据等缺陷;设计了一种基于公钥基础设施(PKI)体系的安全认证机制,能有效解决 Kerberos 中存在的上述缺陷,并将这种安全认证机制实际应用到 Hadoop 集群中。

关键词:Hadoop;HDFS;Kerberos;PKI;公钥

Design of Security Mechanism in Hadoop Based on Public Key Infrastructure

Chen Zhuo, Wang Youchun, Ping Jiawei
(Shanghai Aerospace Electronic Technology Institute, Shanghai 201109, China)

Abstract: To improve the security of private cloud platform and apply cloud platform to aerospace area, this paper made deep research on the security mechanism in cloud platform. In this paper, we analysed the Kerberos released by apache, then we introduce the principle of Kerberos and the workflow of Kerberos in Hadoop, we point out that Kerberos have some defects, first, Kerberos is too dependent on KDC, second, Kerberos use symmetric cryptographic communication, third, the communication between client and HDFS uses plaintext. So we designed a new authenticated scheme based on PKI which can solve those problems in Kerberos, and we practical apply it in Hadoop platform .

Keywords: Hadoop; HDFS; Kerberos; PKI; public key

0 引言

随着云计算的飞速发展,强大计算存储能力使得它被广泛地应用,作为未来重要的数据处理存储方式,不久的将来将会应用到各个领域,包括航天等特殊领域,人们对数据安全性的要求会越来越高,对云计算安全的研究是重中之重。Hadoop 是一种被广泛应用的云计算实现方法,IBM、雅虎、阿里、百度都把它作为自己云平台的底层架构,但 Hadoop 在安全性上存在一定的缺陷,因为它最初的设计目的只是建立一个高效的并行计算模型,应用在单一用户数据可控的环境下,但随着 Hadoop 平台的演变和发展,以 Hadoop 为基础的云平台的逐渐普及,它的用户群体和构建 Hadoop 集群的环境复杂而多样,它的安全缺陷逐渐显露出来。为提高 Hadoop 安全性,Apache 将 Kerberos 认证技术引入到 Hadoop 中,但 Kerberos 在较大的集群中存在一些缺陷,本文提出的方案重点解决平台用户身份认证和数据传输加密问题。

1 Hadoop 概述

Hadoop 最核心的两个模块分别是 HDFS 和 MapReduce[1]。其中 HDFS 是 Hadoop 的存储模块,为云计算系统提供分布式存储的底层支持,具有高容错,易扩展等特点;MapReduce 是一种 Hadoop 提供的计算模型,为云计算中的分布式并行任务处理提供支持,被设计用来高效处理大数据问题,能够让用户直接进行分布式程序开发。用户的数据信息都存储在 Hadoop 的分布式存储系统 HDFS 中,Hadoop 中的身份认证和数据保护都要在 HDFS 中进行。

2 Hadoop 现有的认证体系分析

Hadoop 中现有的安全机制主要分为两部分:一部分是在用户层面上对用户实施身份认证,访问控制;另一部分是在数据层面对存放其中的数据进行家加密、备份、恢复,下面对两个部分的原理和存在的缺陷进行分析[2]。

2.1 用户层面安全模块

Hadoop 中用户的访问权限一般分为“-r”、“-w”、“-x”,即只读、写入和执行,由 Kerberos 身份认证体系进行用户的身份认证和访问授权[2]。

在 Kerberos 认证体系中,密钥分配中心(KDC)是整个体系的核心,客户端首先通过 KDC 进行身份认证,随后向 KDC 发送访问节点数据的请求,KDC 通过请求后给客户端颁发票据,客户端利用票据可以通过节点的认证获取访问权限[3],详细流程如图 1 所示。

Kerberos 认证体系的安全性虽然能够基本满足大部分场景的安全需求,但这种认证方式还是存在以下的 3 点缺陷:

KDC 在 Hadoop 集群中过于重要,每一个新的任务执行都要有 KDC 的认证授权,一旦 KDC 出现故障,整个 Hadoop 集群无法运行。

KDC 安全防护过于薄弱。客户端、Hadoop 集群的密钥均存放在 KDC 中,一旦 KDC 被攻破,则整个集群的数据对于攻击者都是透明的,而在 Kerberos 认证体系中,KDC 是建立在

收稿日期:2015-10-07; 修回日期:2015-11-10。

作者简介:陈 卓(1993-),男,河南驻马店人,硕士研究生,主要从事测试与控制方向的研究。
王有春(1974-),女,安徽蚌埠人,硕士研究生导师,主要从事测试与发控方向的研究。

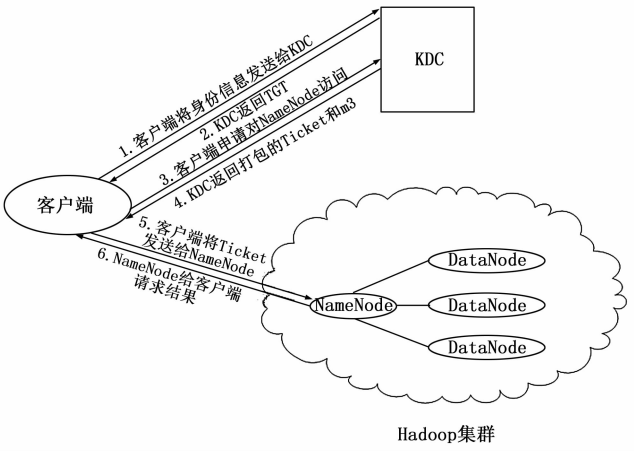

图 1 Kerberos 工作流程

Hadoop 集群的节点中，没有很高的安全保证。

Kerberos 认证协议采用对称密钥加密体制。通信双方所使用的密钥是一样的，如果一方的密钥被攻击者获取，那么攻击者就可以通过密钥访问通信的另一方，这也是对称密钥加密体制的一个缺陷。

2.2 数据层面的安全模块

当用户把数据存入 Hadoop 集群后，Hadoop 集群首先对数据进行分片存储，并对数据进行备份，默认备份系数为 3，分别存放到 Hadoop 的不同节点上，这样一旦主节点检测到数据节点故障，就可以通过启用备份节点保持数据完整性，并补充备份节点数目。这种完善的备份恢复机制使得 Hadoop 中的数据具有很高的可靠性。

上述备份恢复机制虽然让集群中数据很难丢失，但并没有完善的数据加密机制。Hadoop 集群中各个节点之间的通信收到 Kerberos 安全协议保护，但客户端和集群之间的通信使用的是明文传输，信息很容易截获。

3 需求分析

通过对 Hadoop 现有安全机制的分析，可以看出新设计的安全机制要解决原有安全机制的缺憾，需要满足一下 4 个方面的要求：

1）新的安全机制能够在自身出现故障后短期内 Hadoop 可以正常运行；

2）新的安全机制存放密钥的地方应有充分的安全保证；

3）新的安全机制应避免对称密钥加密体制的缺陷；

4）新的安全机制能对客户端与节点之间进行数据传输加密。

4 云计算安全机制设计

目前主流的安全技术有 PKI、SSL 和 VPN，它们实施难度相差不大，其中 PKI 和 SSL 都支持证书的验证和有效性查询，这在 Hadoop 认证体系中是非常重要的，因为 Hadoop 就是根据证书的真实有效性来判断用户身份提供访问权限；此外 PKI 相比于 SSL，不仅支持双向认证，与应用的结合度也很友好，基于 PKI 设计的安全体系不仅能很好地与 Hadoop 相结合，而且能够让 Hadoop 对用户进行验证授权的同事被用户进行双向认证，所以本文用 PKI 技术作为安全体系设计的基础。

在数据层面上，本文提出了 HTTPS 协议作为客户端与 HDFS 服务器之间的传输协议，能够对传输的数据进行加密，有效弥补了原有安全体系在数据层面的缺憾。

通过 PKI 进行用户身份认证，HTTPS 进行数据加密传输，使得新的安全体系有了很大的提升。

4.1 基于 PKI 的用户身份认证体系

4.1.1 PKI 原理

PKI 的英文全称是 Public Key Infrastructure，也就是公钥基础设施，采用非对称密钥加密体制，利用公钥理论和技术为用户提供身份认证和数字签名等服务的公钥基础设施，并提供公钥和证书管理体系。

一个完整的 PKI 系统由一下几部分组成：认证中心（CA），注册机构（RA），数字证书库（LDAP），密钥备份恢复系统，证书撤销处理系统，PKI 应用接口系统组成[4]。

一个用户申请证书的实际流程如图 2 所示。

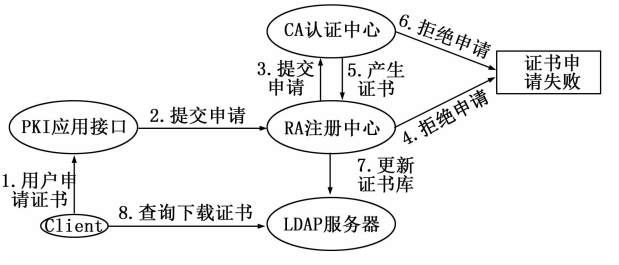

图 2 PKI 原理

4.1.2 公钥密码体制

公钥密码算法的核心在于它的一对公钥和私钥，公钥是所有人都可以获取使用的，私钥是用户单独保存使用的，如果用其中一个进行加密，则只能用另一个密钥进行解密，可以用于数字加密和数字签名。

目前常用的公钥加密算法有 RSA 和椭圆曲线算法。

RSA 算法基于一个十分简单的数论事实：将两个大素数相乘十分容易，但是想要对乘积进行因式分解十分困难，可以将乘积作为公钥，RSA 密钥越长，保密强度越高，十分便于理解和应用，是目前应用最广泛的公钥方案[5]，它的缺点则是 RSA 密钥的位数很长，使得加密的计算量很大。

椭圆曲线加密算法是一种新兴的加密算法，基于椭圆曲线对数问题 ECDLP，给定素数 P 和椭圆曲线 E，对于 $Q=KP$，已知 K 和 P 计算 Q 比较容易，由 Q 和 P 计算 K 则比较困难。相比于 RSA，椭圆曲线算法占用内存和计算量都更少。

本文采用 RSA 作为公钥加密算法，RSA 算法原理更为简单，而且远远比椭圆加密算法成熟，实现起来更为简单，虽然它在密钥的产生和认证过程中运算量上比较大，但密钥的产生和认证只是在 CA 中和客户端的初始认证中发生，并不影响 Hadoop 集群的运行速度，所以本文选用更为成熟稳定的 RSA 算法。

4.1.3 散列函数

散列函数能够把任意长的报文转换成固定长度输出的消息摘要，不同的报文会产生不同的消息摘要，所以可以通过消息摘要验证信息的完整性，应用于 PKI 认证系统中的数字证书[6]。目前广泛应用的散列函数算法有 MD5 和 SHA，本文采用 SHA-256，因为 SHA 算法比 MD5 具有更强的安全性，SHA-256 在 SHA 算法系列中可以很好地保证安全性又不会带来过大的计算量。

4.1.4　身份认证流程

在 PKI 认证体系中，客户端在访问 HDFS 之前，首先应该在 CA 中进行身份认证，获得自己的身份证书，当访问 HDFS 时，提交自己的身份证书，HDFS 通过 CA 验证客户端身份的真实性，通过验证后 HDFS 给客户端相应的权限访问数据，如图 3 所示。

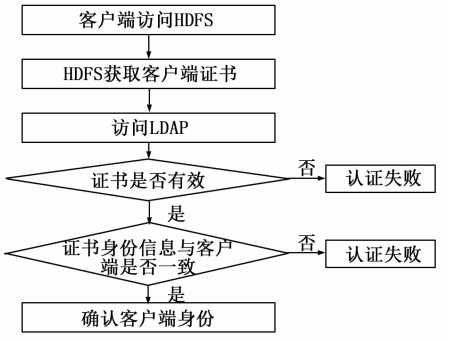

图 3　身份认证流程

4.2　数据加密

在数据层面上，Hadoop 虽然有着完善的备份恢复机制，能够保证数据很难丢失，但它的数据加密机制并不完善，在客户端与集群之间的数据传输并没有相应的安全机制，直接采用 HTTP 协议进行明文传输，数据很容易被攻击获取，本文提出用 HTTPS 作为客户端和 HDFS 之间的数据传输协议。

HTTPS 协议是 HTTP 和 SSL 协议的结合，能够在 HTTP 的基础上对传输的数据进行 SSL 加密，有效地保证了传输数据的安全。

当客户端对 HDFS 服务器进行访问时，HDFS 服务器和客户端都要互相提供数字证书给对方，通过验证后，随机产生一对密钥，通过这对密钥加密的密文信道传输数据。

5　实验结果与分析

5.1　平台构建及证书申请

在硬件方面课题采用四台计算机来组建一个简易的 Hadoop 集群，这 4 台计算机配置如表 1 所示。

表 1　Hadoop 集群硬件信息

cz01	192.168.1.1	Pentium-Dual-Core 2.5G	150GB	Ubuntu 14.04
cz02	192.168.1.2	Pentium-Dual-Core 2.5G	150GB	Ubuntu 14.04
cz03	192.168.1.3	Pentium-Dual-Core 2.5G	150GB	Ubuntu 14.04
cz04	192.168.1.4	Pentium-Dual-Core 2.5G	150GB	Ubuntu 14.04

实验平台采用 cz01 作为 master 节点，其它 3 台计算机作为 slave 节点建立 Hadoop 集群。首先为各个节点之间配置 SSH 协议来实现节点之间加密免登陆访问，随后修改节点的系统环境变量添加 Hadoop 信息，配置 Hadoop 的端口、备份节点、分片大小等信息。

构建 PKI 认证体系，将一台安全可靠的第三方计算机作为 PKI 体系的 CA 认证中心，将 RA 和 LDAP 服务器放在 Hadoop 的 NameNode 节点上，组成了 PKI 基本的硬件架构；在软件方面采用 Openssl1.0.1 作为密钥加密的基础，apache2 作为 PKI 进行证书注册申请的服务器基础，mysql 存放 LDAP 服务器中的用户证书信息，Openca 为用户提供整个 PKI 服务。

在安装 Openca 时需要将 apache2，mysql，Openssl 信息集成到其中，如以下代码所示：

./configure－－prefix＝/usr/local/openca－－with－openca－tools－prefix＝/usr/local/openca－tools－－with－httpd－user＝daemon－－with－httpd－group＝daemon－－with－httpd－fs－prefix＝/usr/local/apache2－－with－htdocs－fs－prefix＝/usr/local/apache2/htdocs/pki－－with－db－name＝openca－－with－db－type＝mysql－－with－service－mail－account＝" webmaster@cz01. cn"

当客户端访问一个 Hadoop 集群时，首先需要向 CA 申请身份认证。通过 PKI 应用接口向 RA 注册机构提出证书申请，将客户端的身份信息提交给 RA，如图 4。RA 通过审核将信息传递给 CA，随后 CA 生成证书。

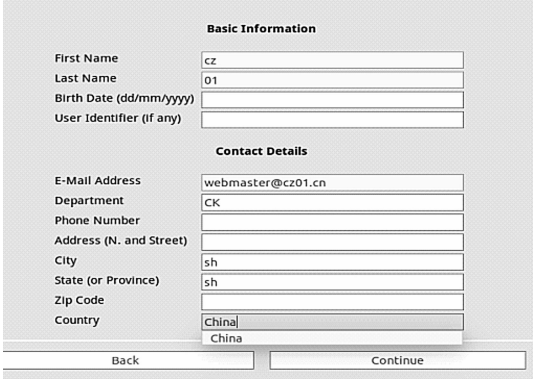

图 4　客户端提交证书申请到 RA

客户端访问 HDFS，将自己的身份证书发送给 HDFS，如图 5，HDFS 对客户端证书通过 CA 验证，验证通过则客户端获得访问权限。

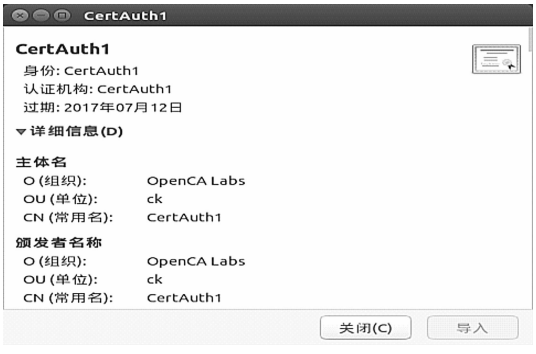

图 5　用户的身份证书

客户端通过身份验证后，通过 HTTPS 协议与 HDFS 进行数据传输时，首先会要求双方提供数字证书，如图 6 所示，双方各自交换身份证书并进行验证，通过验证后会随机产生一对密钥，通信双方就可以通过这对密钥对数据进行加密通信。

5.2　实验分析

将 PKI 技术与 Hadoop 集群结合起来，使用新的身份认证方式和传输协议，Hadoop 集群的安全性在多个方面比原有 Kerberos 安全体系有了提高，

首先在 Kerberos 安全认证体系中，Hadoop 集群的运行过度依赖 KDC，每一个新的任务的发起都需要有 KDC 的参与，而

（下转第 166 页）