大数据平台Hive组件深度审计实现技术研究

Research on Implementation Technology of Hive Component Deep Audit in Big Data Platform

冀　文[1]，田　峰[1]，康　乾[2]，叶荣伟[2]（1. 中国移动信息技术有限公司，北京 100032；2. 中国移动杭州研发中心，浙江 杭州 310000）

Ji Wen[1], Tian Feng[1], Kang Qian[2], Ye Rongwei[2]（1. China Mobile Information Technology Co., Ltd., Beijing 100032, China; 2. China Mobile Hangzhou Research&Development Center, Hangzhou 310000, China）

摘　要：

针对Hive组件的深度审计方法，在现网中结合DPI技术，深度分析Hive组件的交互方式和交互内容，设置多维度的审计分析规则，重塑组件行为、流量和内容侧的交互行为，挖掘大数据组件运行的安全风险。大数据平台Hive组件深度审计方法结合分布式计算能力，能满足现网超大规模流量的分析需求，可适用于数据访问风险监测、数据共享防泄露等场景，实现对Hive组件的深度审计功能。

关键词：

Hive；深度审计；DPI；大数据

doi：10.12045/j.issn.1007-3043.2019.04.007

中图分类号：TN915.08

文献标识码：A

文章编号：1007-3043（2019）04-0030-05

Abstract：

Based on the deep audit method of Hive components, combined with DPI technology in the current network the interactive mode and content of Hive components are analyzed in depth. The multi-dimensional audit analysis rules is set to reshape component behavior, traffic and content side interaction, and dig the risks of big data component operation. Combined with distributed computing capacity, the method can meet the analysis requirement of super-large-scale traffic in the existing network, and can be applied to data access risk monitoring, data sharing and anti-disclosure scenarios, and realize the deep audit function of Hive components.

Keywords：

Hive; Deep audit; DPI; Big data

引用格式：冀文，田峰，康乾，等. 大数据平台Hive组件深度审计实现技术研究[J]. 邮电设计技术，2019(4)：30-34.

1 概述

大数据业已成为产业发展的创新要素，不仅在数据科学与技术层次，而且在商业模式、产业格局、生态价值与教育层面，大数据都能带来新理念和新思维[1-2]。在充分认知大数据产业发展重要性的同时，也要充分意识到大数据安全对大数据应用发展的重要性[3-4]。

大数据平台是实现大数据分析能力的基础，而大数据行为的合规审计则是一种保护平台的有效方法。目前，这类审计能力基本上是基于组件的安全日志、运行日志和审计日志的组合分析，存在一定局限性。

a）缺少原生分析能力。大数据平台的组件基于开源软件，在设计时缺少安全机制，例如Hadoop生态系统[5]，其本身没有审计功能。

b）日志字段粒度太大。可审计日志包括安全日志、运行日志和审计日志，但是其字段不完整，记录的信息模糊，例如HDFS[6]，其日志缺少对操作目录的记录。

c）审计能力不可扩展。新增日志字段需要修改

收稿日期：2019-02-18

源代码，分析能力不可扩展，无法满足分析审计规则更新的需求[7]。

针对上述问题，本文提出了大数据组件深度审计方法。该方法主要通过采集和解析网络流量数据，提取全量组件访问、共享过程安全日志进行分析，实现第三方集中式安全深度审计。

2 大数据组件深度审计实现原理

下面对大数据组件深度审计方法中涉及的相关概念做一个基本介绍。

a）深度包检测技术[8]（DPI——Deep Packet Inspection），是一种基于应用层的流量检测和控制技术，当IP数据包、TCP或UDP数据流通过基于DPI技术的带宽管理系统时，可通过深入读取IP包载荷内容对OSI七层协议中的应用层信息进行重组[9]，从而得到整个应用程序的内容，然后按照系统定义的管理策略对流量进行整形操作。

b）Hive[10]是建立在Hadoop上的数据仓库基础构架。它提供了一系列的工具，可以用来进行数据提取转化加载（ETL），是一种可以操作和分析Hadoop中存储的数据的机制。

图1是Hive组件构成和交互方式示意图。Hive组件由HiveServer、MetaStore、WebHCat等构成[11]，支持多种Hive Client（如Beeline、Thrift、JDBC、ODBC驱动的客户端）与组件进行数据交互。

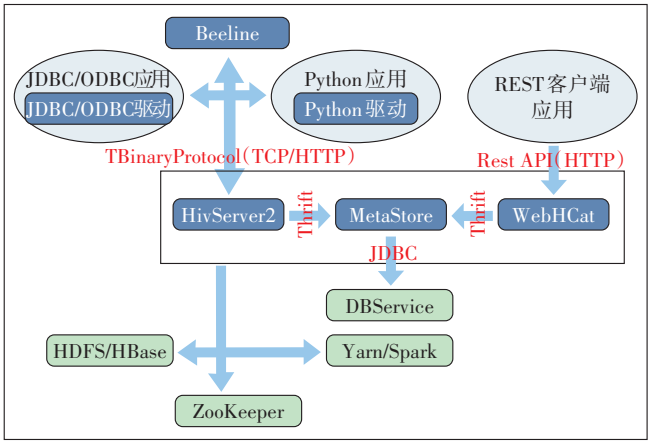

图1　Hive组件交互示意

大数据平台Hive组件深度审计方法重点研究Hive组件和CLI、SDK应用、Web客户端的交互机制，形成可操作协议解析功能模块。而DPI技术具有的支撑全栈协议解析能力和能够提供通用的可插拔应用层协议解析的能力，是此方法实现的基础[12]。

3 大数据组件深度审计方法

3.1 研究目标

大数据组件深度审计技术的目标是及时发现大数据组件运行的安全风险，实现对Hive组件的深度审计。具体来说，该技术通过Hive组件原生能力，实现组件交互行为的重构；通过DPI技术替代常规的日志分析，可以控制分析的字段和粒度，保证审计能力的可扩展性，提供多维度的交互分析。

3.2 深度审计方法

深度审计方法重点阐述Hive组件和客户端的交互信息，基于DPI能力进行数据包特征提取和协议识别，剥离非应用层数据，获取核心交互行为，并结合多维度的分析方法，审计Hive组件交互行为。

面向Hive平台的深度审计方法的流程如图2所示，主要包括以下步骤。

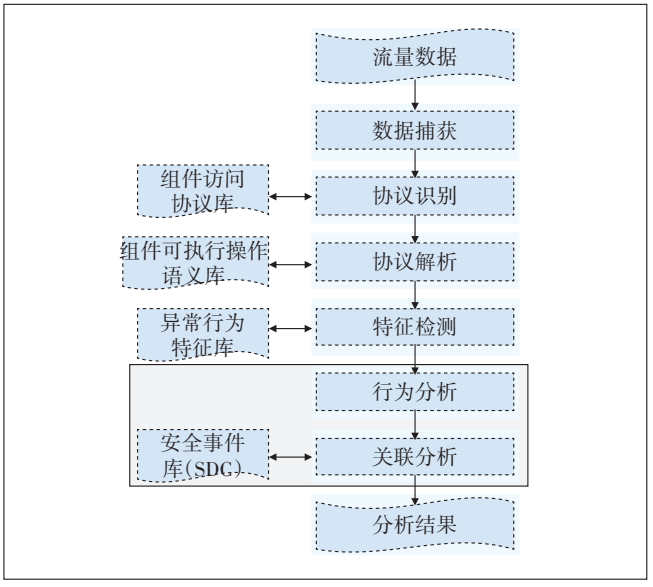

图2　深度审计方法流程示意图

a）审计系统通过采集模块，捕获网络流量数据包。

b）识别网络流量数据包访问协议，过滤非审计范围的数据包。

c）解析应用测协议内容，提取协议中包含的Hive操作相关信息，如操作方法、操作对象、操作参数等。

d）识别访问操作行为是否符合异常行为特征，异常与正常行为分类存储。

e）访问操作行为语义分析，构建操作行为上下

文。

f) 操作行为关联分析,发现深度安全风险。

g) 输出审计分析结果。

深度审计方法的实现主要基于以下2个关键技术。

a) 基于DPI的Hive组件协议解析技术:基于DPI分析技术,获取全量的Hive组件流量,结合组件访问协议库和可执行语意库,匹配流量负载的数据字段,完成协议判断和协议解析。表1是部分交互数据包的结构。

表1 Hive组件通信协议字段表

字段名	字段类型	字段长度/bit	字段描述
版本	Int	32	该字段表示TBinaryProtocol协议版本(0x80010000)|函数调用类型(①响应;②异常;③无返回值的请求)
消息名称长度	Int	32	调用消息(函数)名称长度,以字节为单位
消息(函数)名称	String	-	服务端提供的远程调用方法名称,包括操作名称的长度和字段信息
流水号	Int	32	报文序号,相当于mysql协议报文当中的number of fields,即当前报文在本次TCP会话中的序号
消息负载	String	-	指定消息(函数)的参数
结束标记	Byte	8	消息结束标记0x00

b) 多维度分析方法:基于DPI分析技术,获取全量的Hive组件平台通信行为数据后,采用多维度分析方法进行平台数据交互行为的风险审计。该分析方法包括以下3个维度。

(a) 行为审计:通过对大数据组件的网络操作行为,包括用户通过客户端、管理平台对组件的连接、访问以及对数据的增删改查等操作进行监控,根据评估模型识别异常行为并按安全策略实时告警并记录。

(b) 内容审计:提供深入的内容审计功能,可对用户请求的操作及节点间的通信内容进行深入的包探测和分析,提供完整的内容检测,以数据为对象对大数据组件中的数据访问操作进行灵活的细粒度(数据粒度)审计。

(c) 流量审计:提供基于访问协议的流量识别分析能力,识别规范协议外的其他非安全协议的数据传输,提示安全风险。

基于上述关键技术,本文提出的大数据安全审计系统包含了组件行为重构、多维度行为分析等能力,能在相关业务场景中发挥审计作用。

3.3 大数据安全审计系统

本文提出的大数据安全审计系统基于大数据组件深度审计方法实现。通过大数据平台集群侧的全流量镜像能力,系统采集网络原始数据包,解析全网Hive组件的访问流量,并识别大数据组件访问场景中操作行为的关键要素日志,帮助大数据平台运维人员发现异常行为,实现多维度、可视化的安全审计。

3.3.1 功能架构

大数据安全审计系统包括关系模块、存储层、展示层、分析层、采集层和知识库,具体如图3所示。

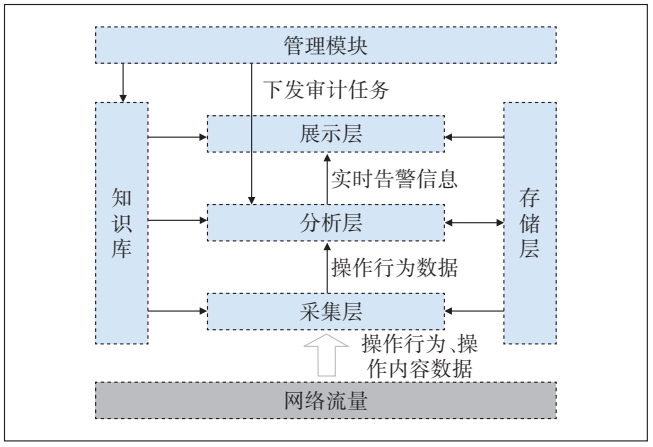

图3 大数据安全审计系统功能架构图

a) 管理模块:主要包括用户认证、权限管理、日志管理、审计策略、审计任务,并提供配置管理和规则库升级等功能。

b) 存储层:审计系统的数据存储位置,可存储数据包括采集到的操作日志和数据内容、审计分析产生的审计结果、审计报告、知识库内容。

c) 展示层:展示审计结果,如监控告警展示和审计报表展示等,包括实时监控告警、审计报表展示、综合统计展示、违规内容数量展示等。

d) 分析层:根据审计需求和审计条件,筛选存储层中操作日志,并进行审计分析,包括根据审计条件快速筛选操作日志,分析操作行为的时效性,根据审计需求全面分析操作行为和操作内容、构建操作会话,实现用户访问的完整审计。

e) 采集层:对接网络镜像端口,捕获网络流量,解析协议语义,提取操作行为和操作数据,采集组件操作日志,完成数据清洗和操作日志解析。同时,在该层还需丢弃无关流量并将采集到的数据写入存储层。

f) 知识库:包括审计策略、配置和规则等内容,如

采集流量协议范围、操作行为审计规则、操作内容审计规则、审计策略、系统配置内容（独立模块可包含过滤器配置）、审计报告模板、综合统计模板、操作行为、操作内容黑白名单。

3.3.2　整体实施方案

大数据审计系统整体实施方案如图4所示，通过获取大数据集群的二层网络流量，前置服务器形成流量包pcap文件，交由统一协议解析平台，统一协议解析平台通过协议插件管理器，按配置动态加载大数据组件协议解析插件进行轮询解析，输出解析后的原始日志数据流，通过日志融合规则关联、融合原始日志，形成标准化日志记录数据，并由交互界面呈现审计分析结果[13-14]。

3.3.3　系统核心实施方案

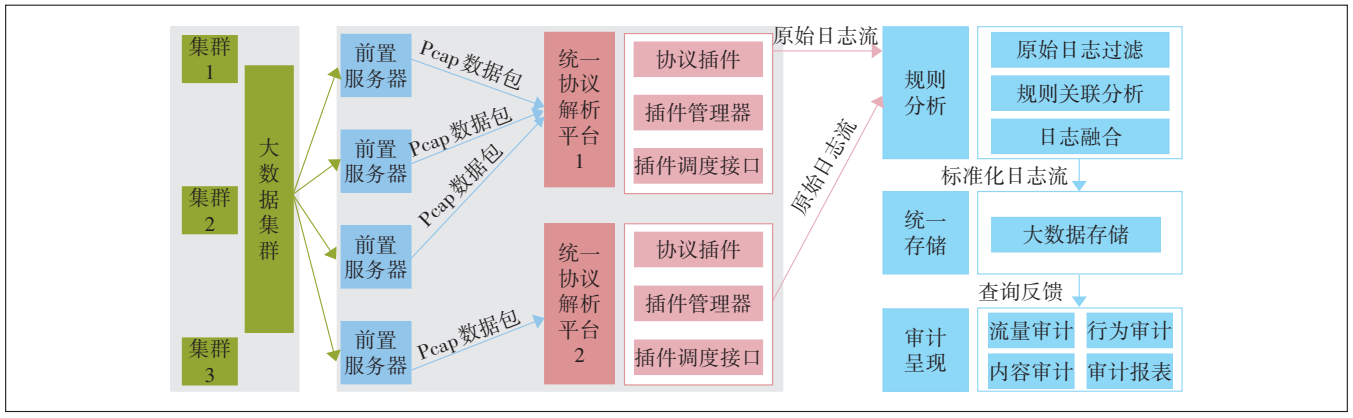

图4　大数据审计系统实施方案图

系统实施方案的核心是采集层和分析层。采集层由采集模块构成，可支持离线文件分析和在线分析的处理模式，其中在线分析需要路由设备提供镜像端口。采集模块通过知识库将操作行为格式化，以供分析器调用分析；格式化数据包以流的形式导入分析器，先进行协议识别分类，然后结合特征知识库，分析平台中执行的操作行为，对异常行为可通过上下文信息构建完整的操作会话，并结合已有安全事件，整体分析操作行为的安全风险。

3.3.3.1　基于DPI技术的采集模块设计

采集模块的数据入口为网络路由设备的镜像端口，该模块结构如图5所示，主要包括以下内容。

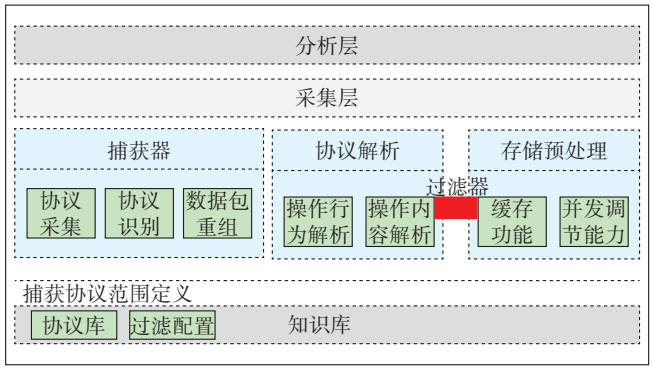

图5　采集模块示意图

a）捕获器：根据审计协议需求捕获有效数据并按协议重组数据包；对捕获的数据包进行基础的协议解析，解析层级在应用层以下。

b）协议解析：通过分析应用层的数据内容，匹配数据负载字段，获取操作行为和操作内容（后续可扩展至操作协议）等数据。

c）存储预处理：将过滤后内容和协议分析结果写入存储层，并将协议解析结果中的操作行为实时发送至分析层。

经过预处理，网络数据流量包被分解为用户访问大数据组件的操作行为和操作数据，该数据一部分将在本地持久化存储，另一部分将继续流转供上层分析模块挖掘访问行为。

3.3.3.2　多维度行为分析模块设计

分析模块将对采集层解析出的操作内容和操作行为进行二次分析。模块对Hive组件的操作数据进行风险等级评估操作，挖掘数据交互过程中的异常日志，结合行为、会话、内容3个维度，深度审计Hive组件数据流转行为（见图6）。

分析模块涉及到以下几个关键知识库。

a）白名单：综合行为特征和数据特征，通过用户自定义操作源、操作对象、操作时间、操作内容、操作环境等内容，指定必须审计的行为和数据。其匹配优

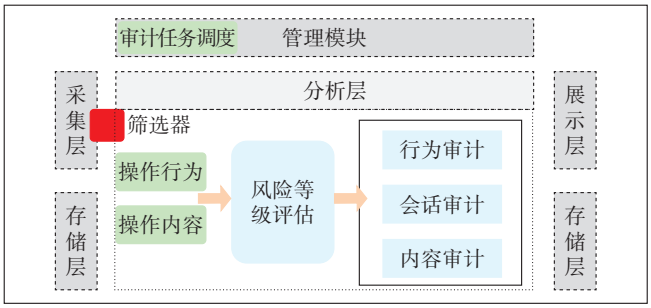

图6　分析模块示意图

先级仅次于黑名单。

b）行为特征：规则库，指明异常行为特征。

c）数据特征：规则库，重要数据泄露风险画像。

根据审计策略，系统加载对应的黑白名单，协助模块获取待分析的操作行为和操作内容数据，这些数据将分别导入各审计分析模块，并结合上述行为特征、数据特征、协议库信息等进行深度用户行为分析。

4 结束语

本文以DPI技术为基础，进行Hive组件的深度审计技术研究，在理论研究的基础上，实现了大数据安全审计系统。该系统基于Hive组件原生能力，利用DPI技术还原负载内容和操作行为，保证对大数据组件审计能力的可扩展性，实现重构组件交互行为，并提供多维度的交互分析能力。大数据安全审计系统能够在数据安全审计与稽核、数据访问风险监测、数据共享防泄露等场景下，提供直观的审计结果，及时发现大数据组件交互过程中的安全风险，保障数据安全。

参考文献：

[1] 冯登国，张敏，李昊. 大数据安全与隐私保护[J]. 计算机学报，2014，37(1)：246-258.
[2] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014(9)：1889-1908.
[3] 王元卓，靳小龙，程学旗. 网络大数据：现状与展望[J]. 计算机学报，2013，36(6)：1125-1138.
[4] 顾君忠. 大数据与大数据分析[J]. 软件产业与工程，2013(4)：17-21.
[5] 孟小峰，杜治娟. 大数据融合研究：问题与挑战[J]. 计算机研究与发展，2016，53(2)：231-246.
[6] 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，27(1)：1-9.
[7] 潘富斌. 基于Hadoop的安全云存储系统研究与实现[D]. 成都：电子科技大学，2013.
[8] DERI L，MARTINELLI M，BUJLOW T，et al. nDPI：Open-source high-speed deep packet inspection[C]// Wireless Communications & Mobile Computing Conference. 2014.
[9] nDPI[EB/OL].[2019-01-28]. https://www.ntop.org/products/deep-packet-inspection/ndpi/.
[10] MINAR N，GRAY M，ROUP O，et al. Hive：Distributed Agents for Networking Things[C]// International Symposium on Agent Systems & Applications Third International Symposium on Mobile Agents. 1999.
[11] Hive组件架构[EB/OL].[2019-01-28]. https://hive.apache.org/.
[12] 冉萌，韩玉辉. DPI技术在移动大数据中的应用[J]. 邮电设计技术，2016(8)：33-36.
[13] 丁文超，冷冰，许杰，等. 大数据环境下的安全审计系统框架[J]. 通信技术，2016，49(7)：909-914.
[14] 刘国城. 基于大数据的互联网安全审计过程建模研究[J]. 兰州学刊，2018(3)：92-103.
[15] 詹鹏伟，谢小姣. 大数据系统及关键技术与工具综述[J]. 网络安全技术与应用，2018(8)：50-52.
[16] 曹珍富，董晓蕾，周俊，等. 大数据安全与隐私保护研究进展[J]. 计算机研究与发展，2016，53(10)：2137-2151.
[17] 罗颖. 大数据安全与隐私保护研究[J]. 信息通信，2016(1)：162-163.
[18] 李树栋，贾焰，吴晓波，等. 从全生命周期管理角度看大数据安全技术研究[J]. 大数据，2017(5)：6-22.
[19] 孟金龙. 大数据分析系统概述[J]. 现代电视技术，2018，207(9)：123-124.
[20] 张引，陈敏，廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展，2013，50(S2)：216-233.
[21] 何美斌，胡精英. 基于Hadoop的电信大数据平台安全研究[J]. 信息安全与技术，2015，6(10).
[22] 杨刚，杨凯. 大数据关键处理技术综述[J]. 计算机与数字工程，2016，44(4)：694-699.
[23] 张华. 基于Hadoop的电信大数据平台应用探究[J]. 长春大学学报，2018，28(10)：43-46.
[24] 陈娇，朱焱，丁国富. 大数据环境下Hive访问控制技术研究[J]. 软件导刊，2018，17(12)：187-190，196.
[25] 谷红勋，张霖. DPI：运营商大数据安全运营的基石[J]. 网络空间安全，2016，7(7)：22-26.
[26] 侯慧芳，潘洁. 大数据背景下运营商建设统一DPI系统的思考[J]. 电信科学，2017，33(4)：191-197.
[27] 梁桃红，何丽，张洪革. 一种面向云端存储的大数据安全审计框架[J]. 电子技术与软件工程，2018，140(18)：189-190.

作者简介：
冀文，毕业于中国科学院研究生院，项目经理，博士，主要从事信息安全和网络安全技术研究和管理工作；田峰，毕业于山东工业大学，高级项目经理，学士，主要从事信息安全和网络安全管理和技术研究工作；康乾，毕业于杭州师范大学，安全工程师，硕士，主要从事大数据安全研究工作；叶荣伟，毕业于上海交通大学，高级产品经理，硕士，主要从事数据安全和客户隐私信息保护研究工作。