基于开源工具集的大数据网络安全态势感知及预警架构

琚安康 郭渊博 朱泰铭

(中国人民解放军信息工程大学 郑州 450001) (数学工程与先进计算国家重点实验室 郑州 450001)

摘 要 对信息系统安全防护而言,大数据是一把双刃剑。信息量的巨增使得数据价值密度更小,给APT等攻击行为提供了更好的藏身环境;但大数据处理技术对海量数据的聚合、挖掘和分析又使得准确检测及预测攻击威胁成为可能。为增强信息系统的威胁感知与攻击预警能力,构建大数据威胁处理平台势在必行。基于最新的开源大数据组件集,构建了集数据收集整理、数据存储、离线分析发现、实时关联检测、威胁预警和态势呈现等功能于一体的、支持全流程安全事件处理过程的、完整的网络安全态势感知及预警架构,与现有同类平台架构相比,其具有高可用、可扩展、易部署等特点,且能较好地支持威胁情报的引入。

关键词 开源工具,大数据,态势感知,威胁预警

中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.023

Framework for Big Data Network Security Situational Awareness and Threat Warning Based on Open Source Toolset

JU An-kang GUO Yuan-bo ZHU Tai-ming

(PLA Information Engineering University, Zhengzhou 450001, China)

(State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China)

Abstract Big data is a double-edged sword for information system security protection. On the one hand, data value density decreased because of the dramatic increase in the amount of information, which provides a better shelter for attacks like APT. On the other hand, its processing technology in aggregation, mining and analysis of huge amounts of data makes it possible to identify security threats accurately. In order to strengthen the perceiving threat ability of information system, it is imperative to build a big data threat analyzing platform. Based on open source big data components, we proposed a situational awareness and threat warning platform for data collection and reduction, data storage, off-line analysis, real-time correlation, threat warning and situation awareness. Compared with existing platforms, this architecture has the advantages of high availability, scalability, and it is easy to deploy and is suitable for introducing threat intelligence.

Keywords Open source tools, Big data, Situational awareness, Threat warning

1 引言

随着互联网的大规模增长,各种入侵和攻击网络工具及手段层出不穷,这使得对入侵行为的检测和预防成为网络系统的关键组成部分。恶意入侵可以定义为威胁网络资源(如用户账号、文件系统、系统内核等)的完整性、机密性或可用性的行为。入侵检测系统和入侵防御系统可对网络流量和系统运行状态进行监测,以发现恶意活动、产生报告并对入侵行为进行报警或阻止。但传统的IDS/IPS均面临着误报、漏报的问题。当前,随着APT等高级攻击手段的出现,以及大数据环境下攻击行动的高度隐匿性,给入侵行为的检测带来诸多新问题:

1)安全事件数量巨大。对于大型企业来说,单一安全数据源(如网络流量数据、日志数据)已经面临着大数据处理的问题,而多来源、多维度的安全数据势必为数据处理带来更大的挑战。

2)安全事件来源多样、种类繁杂。随着数据接入多样性的增加,可供安全人员分析的事件类型也多种多样。目前各类安全数据来源有:系统日志、应用日志、防火墙、IDS/IPS等安全设施的日志、网络数据包、NetFlow数据等,这些数据来源广泛且种类繁杂,若不能有效地加以利用,就无法提升系统的整体安全性,而多源异构数据的引入又给安全分析带来了难以避免的复杂性。

3)网络安全新形势的需要。高级可持续威胁(APT)采用定制化攻击手法,隐蔽性高,常规手段难以及时检测,这就要求分析人员必须对较长时间的安全数据进行掌握,时间跨度

到稿日期:2016-04-27 返修日期:2016-07-22 本文受国家自然科学基金(61501515)资助。

琚安康(1995—),男,硕士生,主要研究方向为信息安全与大数据,E-mail:jusissp@yeah.net;郭渊博(1975—),男,博士,教授,博士生导师,主要研究方向为信息安全;朱泰铭(1991—),男,硕士生,主要研究方向为信息安全。

大势必带来更大的数据量的挑战，而安全检测的及时性对分析处理的速度提出了较高的要求。

4)安全管理者缺乏对整个网络安全态势的全局实时感知能力。传统方法产生的报警数量多且绝大部分是误报，这使得安全管理者不能有效掌握系统运行的安全状况，难以评估网络整体安全态势，对当前是否受到攻击、受到何种攻击、面临何种安全威胁、存在哪些安全风险不能有效感知。

安全事件集中处理最大的弱点是待分析处理的数据量巨大，那些体量庞大、冗余多且独立分散的安全事件显然不能直接作为响应依据，同时网络安全防护也有实时性要求。因此，要解决大数据时代攻击行为的检测与预警问题，最有效的方法还是要依赖于大数据技术本身的能力，上述问题的根本解决途径是大数据技术支持下的网络安全事件关联处理。

针对目前网络安全数据体量巨大、来源多样、增长速度快、价值密度低等特点，借助大数据技术对多源异构安全数据进行分析处理，全面感知系统的安全状况，成为提高系统整体安全能力的关键方法。

目前，大数据处理技术发展迅猛，开源技术在大数据领域已经占据了主导地位，基于 Hadoop 生态圈的大数据平台已经被业界广泛使用，部署规模从几十台到几万台，可以存储和分析 PB 级别数据，Hadoop 生态圈也在不断完善，目前已经能够实现并行计算、高速计算、流式计算等计算框架。基于大数据技术的解决方案可以扩展安全视角、提供安全智能分析、应对新型安全威胁，可有效解决传统方案的瓶颈问题，为网络安全态势感知和威胁预警提供新的解决思路。

2 研究现状

近年来，各类安全厂商基于大数据处理技术，结合网络安全分析方法，提出了各自的安全态势预测和安全管理应用系统，如 360 天眼、阿里云盾态势感知系统、安恒网络空间态势感知监测预警通报管理系统等，借助云端强大的计算能力、数据挖掘技术和可视化分析技术的优势，对那些能够引发网络安全态势发生变化的要素进行全面、快速和准确的捕获和分析，帮助安全管理人员从总体上把握网络安全态势，以及开展预警通报、应急处置等工作。

目前也有很多学者基于 hadoop 和其他大数据平台开展网络安全应用研究。Lee 等人[1]基于 Hadoop(如 HDFS，MapReduce和 Hive)开展了 DDoS 攻击检测技术实验，对存储记录的流量数据进行分析，处理效果可以达到 14 Gbps 的吞吐量。Cheon 和 Choe 等人[2]提出了一种基于 Snort 和 Hadoop 的分布式入侵检测系统体系结构，并开展了一系列实验来分析添加额外的 Hadoop 的节点是否可以提高整体处理效率，实验中使用的是记录文件而不是实时数据，结果发现，Hadoop 集群节点数的增加带来了性能效率的提高(处理相同数据集花费更少的时间)，与只有一个基于 Hadoop 分析节点相比，8 个处理节点的工作性能增加了 424%。

内华达大学的 Bingdong Li[4]在研究现有 NetFlow 应用技术的基础上，在其博士论文中提出了基于大数据技术的网络安全监控系统设计方案，该系统包括用于分布式实时数据收集组件 Flume[15]和 Kafka[16]、实时分布式流数据处理组件 Storm[17]、NoSQL 数据存储组件 Cassandra，以及数据处理状态的用户界面。Li 使用决策树和支持向量机模型对用户行为分析建模，对主机角色进行分类，实验结果显示可以达到较高的用户识别精度，系统为使用者提供了交互式查询分析接口，辅助实现网络的实时监测和可视化态势呈现，是一种有效的态势感知和分析系统。

2014 年 9 月在 Cisco 支持下的开源的 OpenSOC[10](Open Security Operations Center)，是安全大数据应用落地的又一重要成果，OpenSOC 是一种新兴的信息安全大数据处理框架，它将 Hadoop，Storm，ElacticSearch 等多种开源大数据分析工具进行有机结合，提出了一种新的解决思路，具有可扩展、部署灵活等特点，但主要针对网络数据包处理场景，且易用性差。

可以看出，现在这些基于大数据处理技术的安全检测研究或解决方案基本上都是利用其在数据处理方面的优势，克服了传统方法的瓶颈问题，实现了较高的数据分析效率，但都是针对特定应用场景提出的，其技术方法和架构具有片面性且不易部署应用；现有研究对如何开展基于大数据的网络安全态势感知缺少设计指导，在现实应用中缺少一种易于部署的支持各种处理模式的大数据安全分析框架。

在前人工作的基础上，本文综合现有大数据技术，集成整合开源大数据处理组件，结合攻击检测预警核心技术，设计了一种覆盖数据处理流程各个阶段的网络安全态势感知预警架构，综合网络入侵及恶意行为检测、规则挖掘、安全态势呈现等功能，实现对网络安全态势的全面感知，以及对安全威胁的及时预警。

3 整体思路

为有效应对海量安全数据带来的诸多挑战，提高对网络安全态势的感知能力，本文从现有大数据处理技术出发，提出一种安全大数据的集成化解决方案，实现全流程的安全数据分析监控。

本文通过将开源大数据产品和安全分析方法进行集成，为企业或机构提供一种具有安全预警和态势感知功能的大数据基础平台，为事前预警、日志分析、事后报警、运维管理等环节提供一个应对新型安全威胁的系统框架。此框架可以快速地从不同类型的系统、设备上采集数据，对整个网络中的安全事件进行存储和关联分析，通过对已有安全信息的合理分析，实时检测网络威胁和生成安全预警，并对这些信息进行有效呈现。

本文目标并不是开发一个全新的态势感知系统，而是综合利用丰富、强大的各种开源大数据工具，结合当前最前沿的大数据存储、分析、管理技术，综合集成一个能够对海量安全日志数据等(包括结构化数据、半结构化数据、非结构化数据)实现存储、实时处理、离线分析等工作，实现网络安全态势感知和威胁预警的系统框架。该框架具有以下特点：

1)整体采用开源大数据组件，易于维护和部署使用；

2)系统自身具有鲁棒性，可应对海量数据处理需求；

3)搭建成本廉价，适用于大规模分析场景；

4)支持离线处理与实时流数据分析；

5）态势呈现交互可视化；

6）支持威胁情报的引入；

7）便于后续扩展应用。

4　框架设计

结合安全分析的主要需求，此框架从安全数据处理的整个流程角度出发，设计一种高可用、可扩展、易于部署的安全大数据处理架构，主要包括：数据收集整理、数据存储、规则挖掘提取、实时关联分析、安全态势呈现等功能模块，各子系统之间通过数据逻辑通路连接，合作实现对网络安全态势的分析预警。

系统整体采用分层设计和模块化设计思想，其整体架构如图 1 所示。

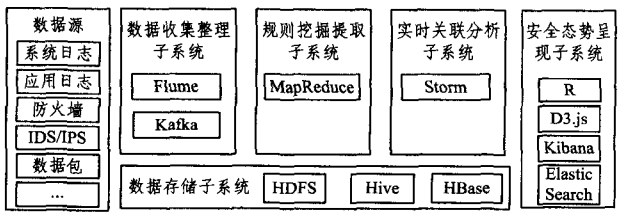

图 1　基于大数据的安全态势预警框架整体设计图

多源异构安全数据主要来源于：网络中的安全设备（如防火墙、IDS/IPS、蜜网等）、网络设备（如路由器、交换机）、系统日志、服务和应用日志、网络流量（大小、成分）、关键网络设备性能（CPU 利用率、端口利用率等）等；数据类别有：日志数据、网络数据包、指标数据等。按照网络体系结构可将数据来源分为 3 类：数据链路层主要是网络设备日志，网络层主要是安全设备数据，应用层数据的主要表现形式是操作系统日志和应用程序日志。

系统以上述多源异构安全数据作为数据处理来源，以 HDFS，Hive，HBase 作为数据存储基础，将 Flume 和 Kafka 作为数据传输和预处理通道，实现数据采集汇聚与系统内的数据高速传输；对数据的处理采取基于 Storm 的实时关联分析和基于 MapReduce 的离线处理模式，态势呈现系统采用 R 语言和 D3. js 等可视呈现技术，并结合 ElasticSearch 作为索引，实现数据和网络态势的交互式可视分析，为用户提供良好的查询接口。

图 2 给出了系统中数据处理的基本流程，客户端采集到的数据交由 Flume 进行汇总后有两条处理路径：1）经 Kafka 缓存后再交由 Storm 进行实时流分析；2）将数据存储至 HDFS，然后再根据需要采用 MapReduce 批处理的方式对离线存储的数据进行挖掘分析，最后可以根据需求对整个流程中产生的安全数据进行检索呈现。

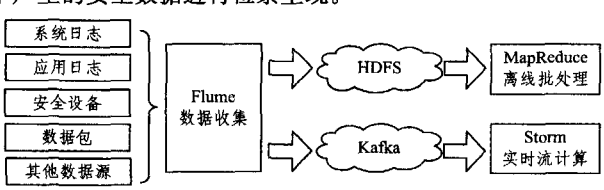

图 2　数据处理流程

4.1　数据收集整理子系统

数据收集整理子系统主要实现各种网络设备产生的不同类型数据的收集工作，初步整理校验收集到的数据后将其送入大数据平台，同时将数据存储至 Hbase 数据库或者 Hive 数据仓库内。数据收集整理子系统实现了数据收集、上报，以及对采集数据初步整理的功能。数据预处理的主要工作包括：数据关联、数据打标、数据清洗、数据集成、数据归约等，完成对采集到的原始数据的初步整理，方便和简化下一步数据处理过程。

数据收集可在目标资产上安装数据探针，或在关键节点部署数据采集设备等，由客户端及各类安全设备将采集到的安全数据上报给大数据处理中心，而后先交由 Flume 对数据进行汇总、存储，Kafka 则作为数据中间件，缓存数据后交由 Storm 进一步关联处理。

Flume[15] 是 Cloudera 提供的一个高可用、高可靠、分布式的海量日志采集、聚合和传输的系统，支持在日志系统中定制各类数据发送方，主要用于数据的收集汇聚；同时，Flume 也提供对数据进行简单处理，并写到各种数据接收方（比如文本、HDFS、Hbase 等）的能力。Flume 具有可靠性强、功能可扩展、便于维护等特点，Flume 采用分层架构，可以适用于有日志搜集和聚合需求的绝大多数分布式处理场景。

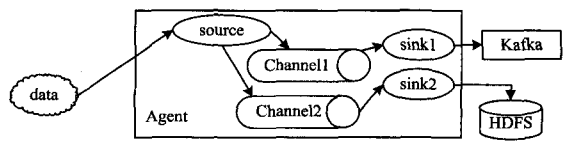

图 3　Flume 结构图

Flume 主要包括以下 3 个部分。

source：处理数据的来源。数据来源可以是 Web Service 中封装的客户端（AVRO 客户端），可以是 NETCAT 服务，也可以是一个不断增长的日志文件。

channel：提供了一层缓冲机制来实现数据的事务性传输，最大限度地保证数据的安全传输。这层缓冲可以在内存中，或者在文件、数据库中，可以由用户自定义实现。

sink：将数据转发到目的宿主，或者继续将数据转发到另外一个 source，实现数据的接力传输，可以通过 AVRO Sink 来实现。

Kafka[16] 是一个分布式的、可划分的、冗余备份的持久性的消息发布-订阅系统，它主要用于处理活跃的流式数据。Kafka 最初由 LinkedIn 公司开发，之后成为 Apache 项目的一部分。Kafka 主要有以下特点：提供高吞吐量的消息发布和订阅；支持持久化操作；分布式系统，易于向外扩展；单个节点宕机能自动平衡负载分配；支持实时和离线两种计算场景。

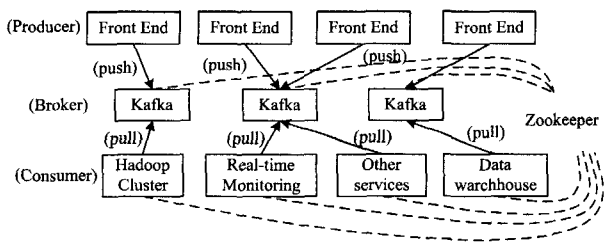

图 4　Kafka 结构图

Kafka 的整体架构非常简单，是显式分布式架构，其结构图如图 4 所示。Producer，Consumer 实现 Kafka 的注册接口，

数据从 Producer 发送到 Broker，Broker 承担一个中间缓存和分发的作用。多个 Broker 协同合作，Broker 将数据分发到系统中的 Consumer，Broker 的作用类似于活跃数据和离线处理系统之间的缓存；其中 Producer，Broker(Kafka)和 Consumer 都可以有多个。

Kafka 集群接收到 Producer 发过来的消息后，将其持久化到硬盘，并保留消息指定时长(支持自定义配置)，但不关注消息是否被消费，Consumer 则从 Kafka 集群 pull 数据，并控制获取消息的 offset。也就是说，Producer 用于收集数据，Broker 用于数据的中间存储，而 Consumer 则用于数据的消费订阅。

Kafka 和 Flume 都是可靠的消息收集系统，通过适当的配置能保证零数据丢失，可以很好地结合起来使用。若设计中需要从 Kafka 到 Hadoop 的数据流，也可以使用 Flume 代理并配置 Kafka 的 Source 读取数据后存储，或者可以直接利用 Flume 与 HDFS 及 HBase 的结合功能。

4.2 数据存储子系统

数据存储子系统是整体框架的存储基础，主要负责对异构数据类型的统一存储，为数据处理和检索查询提供支持，包括操作系统日志、应用日志、防火墙数据、IDS/IPS 日志、网络数据包、Netflow 数据等多源异构数据。数据存储子系统不仅支持由收集整理子系统采集的各种安全数据存储，也要为中间处理结果提供存储保障。

HDFS 结构如图 5 所示。

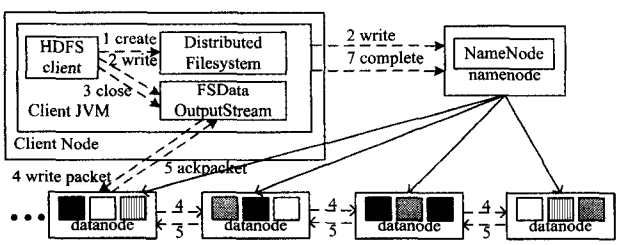

图 5 HDFS 结构图

在此框架设计中选取 HDFS，Hive，Hbase 作为存储组件。其中，HDFS 作为分布式文件系统是整体的存储基础；Hive 是建立在 Hadoop 上的数据仓库，用于存储结构化数据，主要功能为将 HQL(类 SQL)转换为并行计算框架能够识别的程序，使用 HQL 语言查询存放在 HDFS 上的数据；Hbase 是一种分布式 Key/Value 系统，主要以表格的形式存储数据。Hbase 与 Hive 都是将 Hadoop 作为底层存储，两种工具具有不同的特点，对于结构化的数据如日志文件、关系数据类型数据适合于用 Hive 进行存储查询，Hbase 通常用来快速索引数据。

4.3 规则挖掘提取子系统

规则挖掘提取子系统采用 MapReduce 编程模式，主要实现对离线数据的分析计算，用于满足对实时性要求不高的业务需求，可实现预测性分析、数据挖掘等目标功能。采用数据挖掘与机器学习等技术，结合聚类分析、离群点分析、分类算法和统计学方法等，将过去的安全数据进行深度分析和处理，得到数据的深层价值，挖掘出隐藏度较高的攻击方式。

一个 MapReduce 框架程序由 Map 函数和 Reduce 函数组成，可在多个实例下并行处理键值对，有很好的水平扩展性；同时 MapReduce 很好地解决了海量计算速度慢的问题。Mahout 是一个基于 Hadoop 的机器学习和数据挖掘的分布式计算框架，在 MapReduce 模式下封装实现了大量数据挖掘经典算法，为基于 Hadoop 的应用开发提供了算法模型的接口，大大降低了大数据挖掘产品的开发难度。

MapReduce 计算框架如图 6 所示。

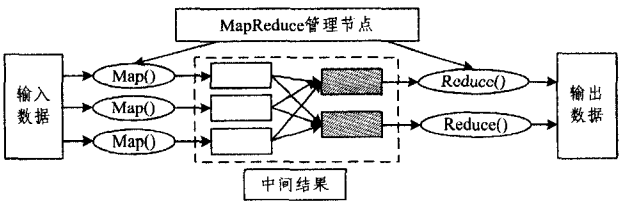

图 6 MapReduce 计算框架

在规则挖掘提取子系统中，关键内容是数据挖掘算法和专家知识的集成，以实现对新规则的挖掘提取。此模块在实际应用中可以扩展集成威胁情报的功能支持，以提升对隐藏较深的攻击方式(如 APT)的发现能力。

威胁情报(Threat Intelligence)是指以证据为基础的知识，包括背景、机制、指标、影响和实践建议等，以及关于资产现有或新兴威胁的知识，并告知相关主体以对这些威胁或危险做出回应。威胁情报是对现有安全机制在广度和深度上的扩展，引入威胁情报功能对提高系统整体安全分析能力有很大帮助，情报驱动的安全策略是日后安全技术发展的重要趋势。

4.4 实时关联分析子系统

实时关联分析子系统可集成多种数据分析方法，是对流数据进行实时分析处理的一种在线数据处理平台。网络安全事件关联分析能将不同来源的报警信息进行去伪存真，从而挖掘出真正的网络攻击事件。流数据可以抽象地看成一组有序到达、连续产生、无限增长并且数据规模不可控的动态数据流。对流数据的实时关联是复杂事件处理(CEP)的典型应用过程，通常需结合知识库和实时在线学习等技术。在网络安全领域，通过对实时产生的海量异构安全事件进行关联分析，可及时发现安全异常和突发敏感事件，对特定事件的发展趋势进行预测，并及时对危急情况进行预警。

在此系统设计中选用 Storm 作为流数据处理工具，对接收到的实时数据进行关联分析，分析完成后将分析结果和告警信息送入存储子系统，以供日后查询分析，或提交给态势呈现子系统直接显示到可视化界面中，展示实时网络安全态势。

Storm[17]是由 BackType 开发并被 Twitter 于 2011 年开源的分布式实时流式数据处理系统，其主要应用场景为实时分析、在线机器学习、持续计算、分布式 RPC、ETL 等。

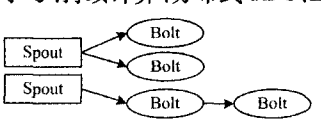

图 7 一个典型的 Storm 拓扑抽象结构

在 Storm 中以 Topology 作为一个计算单位，Topology 本身由一个有向无环图组成，每个任务节点称为一个 Component，节点间以 Tuple 作为最基本的数据传输单位。一个典

型的 Topology 抽象结构如图 7 所示，其中 Spout 是流数据处理的起点，负责为 Topology 从特定的数据源发送数据，而 Bolt 可以接收来自 Spuot 以及其他 Bolt 的数据并进行处理，在完成特定的业务逻辑后，再将数据流发送给下一个 Bolt 或直接消费。

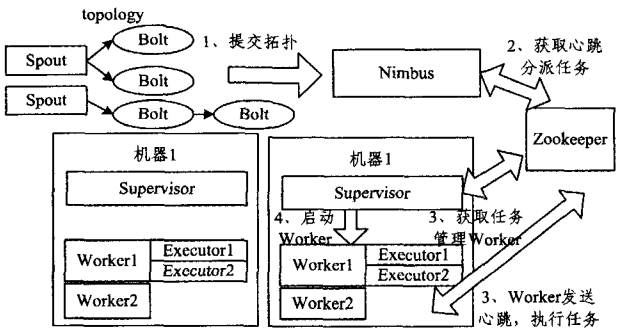

图 8　Storm 结构图

Storm 集群由一个主节点和多个工作节点组成。主节点运行了一个名为“Nimbus”的守护进程，用于分配代码、布置任务及检测故障。每个工作节点都运行了一个名为“Supervisor”的守护进程，用于监听工作，开始并终止工作进程。Nimbus 和 Supervisor 都能快速失败，而且是无状态的，这样一来它们就变得十分健壮，两者的协调工作由 Zookeeper 来完成。

实时关联分析子系统通过对实时流数据进行关联分析，产生新的安全警报，为安全管理人员和系统使用人员进一步分析安全态势提供支持。采用 Storm 进行流数据处理具有以下特点：编程简单，支持多种编程语言；部署和运维便捷，支持水平扩展；容错性强。但是需要针对不同应用场景编写计算拓扑，既要完成对数据的及时处理，也要兼顾效率和准确性。

4.5　安全态势呈现子系统

异构的数据源和持续增长的数据量给分析人员带来了繁重的负担，可视化是当前大数据技术中一项重要的应用，在大数据分析中扮演着越来越重要的角色。安全态势呈现子系统以数据可视化技术为支撑，采用 R 语言、d3. js 等可视化技术将安全分析的结果和实时安全状态进行可视呈现。

Elastic Search[18] 是基于 Lucene 构建的一个大数据搜索引擎，支持分布式多用户全文搜索，可达到实时搜索的能力；Kibana 则为 Elastic Search 的检索结果提供了丰富的表现形式；R 语言是贝尔实验室开发的 S 语言的一种开源实现，它提供了一系列统计和图形显示工具。Elastic Search，Kibana，R 语言和 D3. js 等可视化技术结合可为用户提供较好的呈现效果。

通过将大数据平台中存储的数据和中间运算部分结果按照需求进行可视化呈现，可帮助系统管理员实现对网络整体安全态势的掌握，以应对日趋复杂的网络安全形势。安全态势呈现子系统为用户提供了高度可交互操作，可极大地提高系统的可用性和易用性。

5　架构应用

5.1　威胁模型

在企业网络环境中，安全威胁主要来自于 4 个层面：物理层、网络层、应用层以及系统层，通过对这 4 个层面中相关设备和人员信息数据的采集，可以对网络整体安全态势实现全面感知。

首先将各个层面的所有可能发生事件进行概括和分类，得到事件模式集合 P 以及事件集合 E，定义如下：

事件模式集合 $P=\bigcup_{i=1}^{4} P_i$，其中 $P_i=\{p_i^1,\cdots,p_i^n,\cdots\}$，$p_i^n$ 表示第 i 层的一个事件模式，$p_i^n=(a_1,a_2,\cdots,a_s)$，其中 a_1，$a_2,\cdots,a_s$ 是不同的特征属性，不同的特征属性值表征不同的事件模式。

事件模式是对事件的概括分类，一种事件模式代表了一类事件集合，可以采取机器学习的分类算法完成，并采用实时学习算法将事件进行分类，可事先定义部分事件模式，其余的在系统运行中动态生成。

事件集合 $E=\bigcup_{i=1}^{4} E_i$，其中 $E_i=\{e_i^1,\cdots,e_i^n,\cdots\}$，$e_i^n$ 表示第 i 层发生的一个事件，$e_i^n=(a_1,a_2,\cdots,a_t)$，其中 $a_1,a_2,\cdots,a_t$ 同样表示不同的特征属性，但与 p_i^n 相比，e_i^n 属性值多出了事件实体标识、发生时间、发生区域等特征。若 e_i^n 的一般属性值与 p_i^n 相符合，则称事件 e_i^n 具有事件模式 p_i^n，具有相同事件模式的事件称为同一类型的事件。同类事件可理解为：为达到某种特定效果在不同的时间、区域、实体上的具体实现。

在企业的网络环境中，攻击者要实现一定的目标任务，某些事件的发生是实现后续攻击目的的先决条件，因此引入攻击进程的概念。攻击进程指为完成一定攻击目的的事件模式组合，攻击进程库 $T=\{T_1,T_2,\cdots,T_i,\cdots\}$，其中攻击进程 T_i 是不同层次事件模式的组合，代表实施攻击的各个步骤。

攻击者为达到某种突破目的，隐藏期可能较长，通常可以采用状态机、推理机、贝叶斯网络等方式建模安全威胁场景。通过对不同类型事件之间（即事件模式之间），以及不同时空具体实现下的事件之间的关联，找出真实的攻击进程，从而实现安全监测预警的目的，提高对系统安全态势的感知能力。

5.2　应用过程

下面以企业网络环境下的网络安全监测预警为例，说明此框架在实际部署中的应用过程。首先以敏感数据窃取为例，说明现有安全威胁的主要实施步骤，分析其检测方法。

通过下述活动，攻击者可完成窃取内部网络数据至外部站点的目的：

1）渗透侵入 DMZ 中的 Web 服务器，并以此作为继续侵入内部数据库管理系统的立足点；

2）攻击者通过远程访问服务，在 Web 服务器与内部数据库之间创建网络隧道；

3）利用网络隧道，将关键文件或数据传输至 DMZ 中的 Web 服务器，并通过各个跳板，将敏感数据传输至外部站点；

4）攻击者擦除攻击痕迹，删除活动日志数据。

在每一个活动步骤中，攻击者都会在网络中留下活动踪迹，攻击者虽然有意识在目的达成后擦除攻击痕迹，但管理者只需建立日志服务器，将记录的日志数据实时传输至日志管理系统，即可实现对此行为的破解。

图 9 示出了典型企业网络部署环境。随着 APT 攻击越来越频繁，为应对来自 Internet 的攻击威胁，HIDS，NIDS 等

常用方法很难起作用。传统 SIEM 方法采集多种来源的数据，通过处理分析得到高层次安全场景，一旦网络规模扩大，伴随多源安全数据的引入和时间的推移，数据规模急剧增长，巨大的数据量对系统处理能力提出了较大需求，安全管理系统出现运算瓶颈问题。如上例中给出的威胁场景，单一的检测手段难以检出，辅助检测的日志数据又被海量数据所淹没，因此将大数据处理技术与安全分析技术结合就显得至关重要。

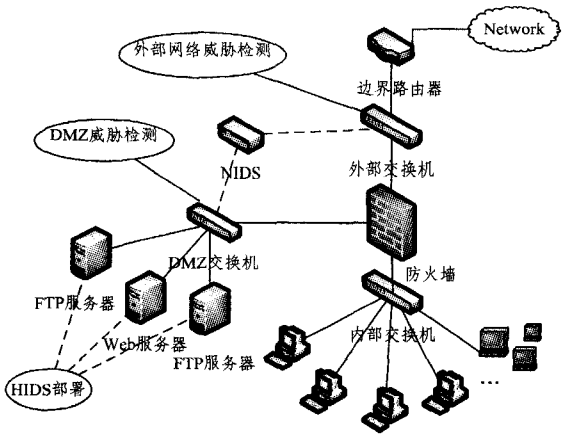

图 9 典型企业网络环境

本文提出的大数据检测框架可有效解决大规模企业网络环境中的网络安全监测预警问题。以图 9 中假设的企业网络环境为例，采集数据可分为 3 类：从边界路由器、内部交换机、DMZ 交换机等网络设备采集的通信数据；从部署的各类安全设备（如 IDS、防火墙）采集的安全异常数据；从内部主机节点及应用程序采集的日志数据。通过部署数据采集探针、代理服务等，在由数据收集整理子系统 Flume 组件接收后，一方面经过 Kafka 的缓冲后，由 Storm 完成实时关联检测，三者的结合可实现对每秒数十万条消息的处理；另一方面交由数据存储子系统，实现采集数据的持久化存储。

离线挖掘子系统对存储的各类数据进行离线分析，找出海量数据中的异常事件，再结合人工分析判断，挖掘出攻击进程 T_i，而威胁情报的引入则大大减轻了人工分析的负担，从而保持攻击进程库的不断更新。实时关联分析子系统则根据攻击进程库实时检测攻击行为，分别匹配对应于攻击进程各个阶段的事件模式，生成多步攻击场景；同时将检测出的攻击报警上报，交由安全管理人员后续处理，实现企业网络安全监测预警功能。

针对企业网络安全监测预警应用场景，此框架不仅可以应对由网络规模和多源数据带来的海量数据处理问题，而且可以结合威胁情报的引入有效应对新型安全威胁，提高监测预警的效率和准确度，同样可将此框架应用于网络流量分类、主机识别等大数据处理场景，实现对网络的全面感知。

6 架构特点

上述从数据收集整理、数据存储、规则挖掘提取、实时关联分析以及安全态势呈现 5 个方面对基于大数据的网络安全态势感知预警框架进行了介绍，分析了各部分功能和具体实现方式，涵盖了完整的安全事件处理周期，并以企业网络环境下的多步攻击检测场景为例，说明了此框架在实际部署中的应用过程。

本文针对网络安全数据源种类多、数据量大、实时处理与批量数据并存的特点，提出并设计了大数据网络安全态势感知预警架构，主要具有以下 5 个关键特性：

1）基于开源大数据工具集。目前开源大数据工具的发展已相对成熟，具有案例丰富、社区活跃、文档完备、使用风险较小等优点，基于开源大数据工具部署业务系统是大数据技术发展的主流趋势。

2）海量事件处理能力。处理海量安全数据是现有安全大数据处理框架所需具备的最基本能力，此框架在开源大数据工具基础上进行设计，具备高效采集和分析处理海量的多源异构安全事件的能力，实现了对安全事件和告警信息的及时存储，并能够进行关联匹配和安全态势的可视化展示，可有效应对海量安全数据带来的挑战。

3）离线分析与实时关联结合。离线挖掘可以从历史数据中找出隐藏度高的异常事件，而对安全数据的实时关联可以及时发现和处理网络异常，有效缩短应急响应时间，二者的结合对提高系统的安全能力起着至关重要的作用。

4）全流程安全事件处理。框架设计从整体处理流程出发，分别从收集整理、存储、规则挖掘、实时关联、可视呈现等方面，给出了一种易于部署应用的大数据安全分析平台，为各个处理模块的设计应用提供指导。

5）支持引入威胁情报。威胁情报共享和应用对消除技术孤岛起着关键作用，单个安全专家的分析能力毕竟有限，威胁情报的共享可有效减弱这个因素带来的影响，也是安全管理平台后续发展、能力提升的重要手段。

与 Li 提出的处理系统和 OpenSOC 相比，本文所提架构设计在扩展性和部署应用方面有显著优势，且支持威胁情报的后续引入，三者的特性对比如表 1 所列。

表 1 系统支持特性对比

支持特性	基于开源组件	可扩展	实时/离线	易于部署	支持威胁情报
Li	√	×	√/×	√	×
OpenSOC	√	√	√/×	×	×
本文	√	√	√/√	√	√

综上，本文从开源大数据软件出发，结合安全分析技术的基本思路，提出并设计了一种较为理想的安全大数据分析框架，覆盖全流程安全事件处理流程，支持威胁情报的引入，此框架可有效提升安全管理的自动化、精确化、实时化水平，为基于开源大数据工具的安全分析系统设计提供思路。

结束语 针对如何在大数据环境下开展网络安全应用的问题，本文从数据收集、存储、分析处理、可视呈现的整个流程角度，结合实时和离线两种应用场景，采用 Hadoop，Flume，Kafka，Storm，Hive，Hbase，Elastic Search 等一系列开源大数据工具，提出并设计了一种基于大数据的网络安全态势感知预警框架，为基于开源大数据工具开展安全态势感知提供思路，此框架将现有大数据处理技术与安全事件管理需求相结合，建立基于开源组件的大数据处理平台体系，涵盖安全数据处理流转的整个流程。通过对已有安全数据的合理分析，对

网络威胁事件产生预警，实时检测网络攻击行为，并采用可视化的方法对结果进行呈现。可根据安全需求将其部署于信息系统安全防护，也可用于 APT 攻击检测等应用场景。

本文更多关注框架设计，给出了各个部分的主要功能和具体实现方式，未涉及过多具体分析方法，下一步的工作重点是结合此框架开展应用，并在此框架下充实完善各种算法细节。

参 考 文 献

[1] LEE Y. Toward scalable internet traffic measurement and analysis with Hadoop[J]. Acm Sigcomm Computer Communication Review, 2013, 43(1): 5-13.

[2] CHEON J J, CHOE T Y. Distributed Processing of Snort Alert Log using Hadoop[J]. International Journal of Engineering & Technology, 2013, 5(3): 2685-2690.

[3] CHARISHMA P, VENKATESH K. Big Data Security Analytic Solution using Splunk[J]. International Journal of Engineering Research & Applications, 2015, 5(4): 50-53.

[4] LI B. Network Security Monitoring and Analysis Based On Big Data Technologies[D]. Dissertations & Theses, 2013.

[5] MARCHAL S, JIANG X, STATE R, et al. A Big Data Architecture for Large Scale Security Monitoring[C]// Proceedings of the 2014 IEEE International Congress on Big Data. IEEE Computer Society, 2014: 56-63.

[6] SAURABH R. Big Data Analytics and Challenges: Network Security and Instruction Detection [J]. International Research Journal of Computers and Electronics and Engineering, 2015, 3(1): 290-295.

[7] MA Z, SMITH P. Determining Risks from Advanced Multi-step Attacks to Critical Information Infrastructures[M]// Critical Information Infrastructures Security. Sprin-ger International Publishing, 2013: 142-154.

[8] ALSERHANI F M. Knowledge-Based Model to Represent Security Information and Reason About Multi-stage Attacks[M]// Advanced Information Systems Engineering Workshops. Springer International Publishing, 2015: 482-494.

[9] LIN S, LI Y, DU X. Study and research of APT detection technology based on big data processing architecture[C]// International Conference on Electronics Information and Emergency Communication. IEEE, 2015.

[10] Opensoc[OL]. http://opensoc.github.io/

[11] XU H. Research on the Tecom Fundamental Network Information Security Awareness Based on Big Data Analyzation[J]. Journal of Information Security Research, 2015(3): 253-260. (in Chinese)
徐浩. 基于大数据分析的电信基础网安全态势研究[J]. 信息安全研究, 2015(3): 253-260.

[12] LI M G, XIAO Y, CHEN J F, et al. Big Data-based Framework for Security Event Mining[J]. Communications Technology, 2015, 48(3): 346-350. (in Chinese)
李明桂，肖毅，陈剑锋，等. 基于大数据的安全事件挖掘框架[J]. 通信技术, 2015, 48(3): 346-350.

[13] FU Y, LI H C, WU X P, et al. Detecting APT attacks: a survey from theperspective of big data analysis[J]. Journal on Communications, 2015, 36(11): 1-14. (in Chinese)
付钰，李洪成，吴晓平，等. 基于大数据分析的 APT 攻击检测研究综述[J]. 通信学报, 2015, 36(11): 1-14.

[14] SUN D W, ZHANG G Y, ZHENG W M. Big data stream computing: Technologies and instances[J]. Journal of Software, 2014, 25(4): 839-862. (in Chinese)
孙大为，张广艳，郑纬民. 大数据流式计算：关键技术及系统实例[J]. 软件学报, 2014, 25(4): 839-862.

[15] Flume[OL]. http://flume.apache.org.

[16] Kafka[OL]. http://kafka.apache.org.

[17] Storm[OL]. http://storm.apache.org.

[18] Elastic Search[OL]. https://www.elastic.co/products/elasticsearch.

(上接第 104 页)

[4] WU D H, YANG W, LONG K. Security Protection Architecture and Critical Technology for Cyberspace[J]. Information Security and Communications Privacy, 2014(7): 79-80. (in Chinese)
吴东海，杨文，龙恺. 网络空间安全防护体系及关键技术研究[J]，信息安全与通信保密, 2014(7): 79-80.

[5] ZHAO T, GAO K L, ZHENG X J, et al. Research on technical framework and cyber security protection system of IOT in smart grid[J]. Electric Power, 2012, 45(5): 87-90. (in Chinese)
赵婷，高昆仑，郑晓崑，等. 智能电网物联网技术架构及信息安全防护体系研究[J]. 中国电力, 2012, 45(5): 87-90.

[6] GAO K L, XIN Y Z, LI Z, et al. Development and Process of Cybersecurity Protection Architecture for Smart Grid Dispatching and Control Systems[J]. Automation of Electric Power Systems, 2015, 39(1): 48-52. (in Chinese)
高昆仑，辛耀中，李钊，等. 智能电网调度控制系统安全防护技术及发展[J]. 电力系统自动化, 2015, 39(1): 48-52.

[7] ZHANG S P, LI J Z, ZHANG F Q, et al. Research and implementation of data center security system based on cloud computing[J]. Computer Engineering and Design, 2011, 32(12): 3965-3979. (in Chinese)
张水平，李纪真，张凤琴，等. 基于云计算的数据中心安全体系研究与实现[J]. 计算机工程与设计, 2011, 32(12): 3965-3979.

[8] JIANG C Z, YU Y, LIN W M. Research on Electric Information Network Security Situation Awareness Model Based on Intelligent Agent[J]. Computer Science, 2012, 39(12): 98-101. (in Chinese)
蒋诚智，余勇，林为民. 基于智能 Agent 的电力信息网络安全态势感知模型研究[J]. 计算机科学, 2012, 39(12): 98-101.

[9] DING X H, ZHAO W D, JU Y, et al. On Demand Security Framework for Cloud Computing[J]. Computer Science, 2014, 41(Z11): 284-287. (in Chinese)
丁鲜花，赵卫栋，俱莹，等. 云计算的按需防护安全框架[J]. 计算机科学, 2014, 41(Z11): 284-287.