大数据技术在信息网络威胁情报中的运用研究

孙　辉,罗双春,李余彪
(78111 部队,四川 成都 610031)

摘　要:网络空间安全的形势随网络攻击技术的不断升级日趋严峻,对信息网络安全构成巨大威胁,而基于大数据处理技术的网络安全威胁情报是一种有效应对手段。围绕信息网络面临的安全挑战,从网络防御的角度,介绍了大数据处理主流平台和关键技术,提出了基于大数据的网络威胁情报技术架构,阐述了网络威胁情报生成过程中大数据技术的关键作用,最后讨论了网络安全威胁情报在网络安全事件处置过程中的运用和作用。

关键词:大数据;威胁情报;信息网络
中图分类号:TP391　　**文献标识码**:A　　DOI: 10.19358/j.issn.2096-5133.2020.05.006
引用格式:孙辉,罗双春,李余彪. 大数据技术在信息网络威胁情报中的运用研究[J]. 信息技术与网络安全,2020,39(5):28-32.

Study on application of big data in threat intelligence of information network

Sun Hui, Luo Shuangchun, Li Yubiao
(78111 Troop, Chengdu 610031, China)

Abstract: With the developing of attack technologies, the security situation of cyber space is becoming much worse which causes tremendous threat to information network. The threat intelligence based on big data technology is an effective method to cope with the crisis. From the perspective of security defense, the paper introduces main platforms and critical technologies of big data, proposes the technology architecture of threat intelligence, and elaborates the key point of big data in the production process of threat intelligence. Finally, the paper discusses the effect and application of threat intelligence in handling network security events.

Key words: big data; threat intelligence; information network

0　引言

随着网络空间攻击理论和技术的深入发展,对网络攻击的方式逐渐聚集到使用 APT(高危持续攻击)方式和利用尚未公开的零日漏洞上来,这使得现实网络安全防护面临的挑战和风险日益上升。对于这类新型网络攻击方式,防火墙、入侵检测系统等传统的被动式安全防护手段虽然能起到一定的抵御攻击的作用,但其基于已知攻击行为特征的检测响应模式,面对新型攻击时往往导致防御失效。2013 年提出应对这类新型攻击的方法,即网络威胁情报。网络威胁情报是基于证据的知识,包括机制、环境、推论、指标、可行性建议等,它能够针对潜在的风险或威胁提供建议和决策[1]。随着网络威胁情报概念的出现,业界也开展了相应的研究。魏为民等人在文献[2]中介绍了网络威胁情报各类标准的核心内容,指出企业应根据自身环境定制威胁情报。李超等人在文献[3]中归纳了大数据环境下适用于威胁情报分析的相关技术,提出了威胁情报分析的流程。单琳在文献[4]中从技术和管理角度分析了威胁情报的特点、优势及不足,并结合我国现状提出相关建议。徐锐等人在文献[5]中从网络防御的视角介绍安全威胁情报的作用和工作流程,并探讨了安全威胁情报的应用场景和研究热点。吕宗平等人在文献[6]中提出一种基于攻击链结合网络异常流量检测的威胁情报分析方法。林晨希等人在文献[7]中详细介绍了网络安全威胁情报的生成与共享方法,并分析总结了网络安全威胁情报的发展现状。

网络威胁情报技术得益于大数据、云计算等技术的应用和支撑日趋完善,国家相关安全机构、网络安全厂商纷纷建立了网络安全威胁情报中心,与各类网络安全防护产品一起,为国家网络空间筑起

坚实的安全屏障。

1 大数据技术

从2012年开始,“大数据”已经成为关注度最高的词汇之一并延续至今。大数据从概念上讲,具有被称为“4V”的显著特征,即规模巨量(Volume)、类型多样(Variety)、流转迅速(Velocity)、较低的价值密度(Value),是超出传统处理技术能力的大量数据的集合。要对这类数据进行处理和分析,需要借助新兴的平台和技术,即大数据处理技术。

1.1 大数据处理主流平台

大数据处理平台是一种计算平台,包括硬件系统、操作系统和运行库,在其上可支持各类大数据处理算法运行。当前业界中存在的大数据平台较多,本文主要对较为流行的Hadoop和Spark平台进行介绍。

1.1.1 Hadoop平台

Hadoop是一种分布式的重量级大数据分析处理平台,它构建在谷歌公司开发的大数据分析处理编程模型和MapReduce框架之上。Hadoop对大数据处理的过程是将大型的数据处理业务分割为多个小任务,分配给多台计算机进行处理。它是分布式计算处理计算架构的开山鼻祖,在大数据处理方面应用非常广泛。这种分布式处理的过程具有以下几方面的优点:(1)可靠性高,(2)处理性能高效,(3)成本低廉,(4)扩展性好。从数据底层存储方面来说,该平台通过构建Hadoop分布式文件系统(Hadoop Distributed File System,HDFS)来支撑数据处理任务,它将平台中所有计算机的硬盘资源汇集在一起,进行调用、分割、存储;在资源调度管理层面,通过YARN(Yet Another Resource Negotiator)集群调度和管理软件,对所有资源进行统一调度管理;从数据处理计算方面来说,该平台使用MapReduce实现框架对数据进行分布式分析计算。此外,经多年发展,Hadoop生态圈也日趋成熟,许多功能组件被集成到平台中,如ZooKeeper、HBase、Hive、Pig、Sqoop等。

1.1.2 Spark平台

Spark是一种轻量级的大数据分析处理平台,其系统架构的核心是Spark Core,在其中实现了类Map算子和类Reduce算子,也为用户提供了Filter、Join、GroupbyKey等功能丰富的算子。与Hadoop平台相比,Spark平台的最大优点就是处理速度快。Spark平台提供了内存计算框架来减少输入/输出开销,此外,Spark提供了同时满足流式计算、批处理和交互式查询等不同应用场景数据处理的能力。该平台采用弹性分布式数据集(Resilient Distributed Dataset,RDD)的方式来存储海量数据,这为系统节约了绝大部分磁盘读写所耗费的时间。该平台还支持许多功能丰富的组件,包括:Spark Streaming,用于实时流计算;Spark SQL,用于各类查询;MLlib,用于数据挖掘、机器学习;GraphX,用于图计算等。Spark可以运行在独立集群中,也可以运行在Hadoop、Amazon EC2等云环境中。

1.2 大数据处理关键技术

大数据分析处理技术不是一种单一的技术,而是一整套贯穿大数据采集、存储、分析处理及应用全过程的分析和处理技术的总称。人们使用这些与以往不同的新兴技术处理结构化、非结构化的数据,从中得到分析处理结果[8]。这些技术包括大数据采集、预处理、存储和管理、分析和挖掘、可视化呈现等。

1.2.1 数据采集与预处理技术

根据采集方式的不同,采集技术分为主动采集和被动采集。主动采集通过Telnet、SNMP、HTTP、FTP、SSH协议、ODBC/JDBC接口、代理和插件、漏洞和端口扫描、蜜罐和蜜网等技术进行采集;被动式采集通过有线和无线、集线器和交换机、Syslog、NetFlow/IPFIX/sFlow、Web Service/MQ、DPI/DFI等技术进行采集。数据预处理方法包括数据清洗、集成、规约、变换、融合等,在分析前去除采集的数据中可能存在的无关、重复、错误等噪声,保证数据分析准确性。

1.2.2 大数据存储与管理技术

该技术的运用主要针对结构化、非结构化、半结构化数据的存储、调度、管理相关事务,主要包括分布式数据库、分布式文件系统、分布式协调系统、资源调度管理、非关系型数据库等。

1.2.3 大数据分析与挖掘技术

该技术主要由机器学习技术、神经网络技术、统计分析技术、数据库分析技术等组成。其中机器学习技术包括决策树、贝叶斯网络、逻辑回归、随机森林、高斯混合模型、支持向量机等;神经网络技术包括自组织神经网络、前向神经网络等;统计分析技术包括判别分析、回归分析、聚类分析等;数据库分析技术包括联机分析处理、多维数据分析、面向

属性的归纳法等。

1.2.4 大数据可视化呈现技术

该技术通过生动、直观、形象的方式将数据分析处理结果呈现给用户。该技术包括层次与网络结构数据可视化、时空数据可视化、多变量数据可视化、文本和跨媒体数据可视化等。

2 信息网络威胁情报技术架构

当前各类信息网络应当构建由安全大数据技术支撑的内网威胁情报系统,在传统的纵深防御安全架构之上,形成“安全策略—监测预警—威胁感知—应急处置”的闭环,第一时间感知威胁,精准应急响应,实现从被动防御向积极防御、从单点防御向整体防控的转化。

信息网络威胁情报体系以大数据处理技术为基础,以安全数据接入、安全数据存储、安全数据分析、共享基础框架为支撑,结合威胁情报防御组织体系和标准体系,共同为威胁情报运用共享提供支撑,其技术架构如图 1 所示。

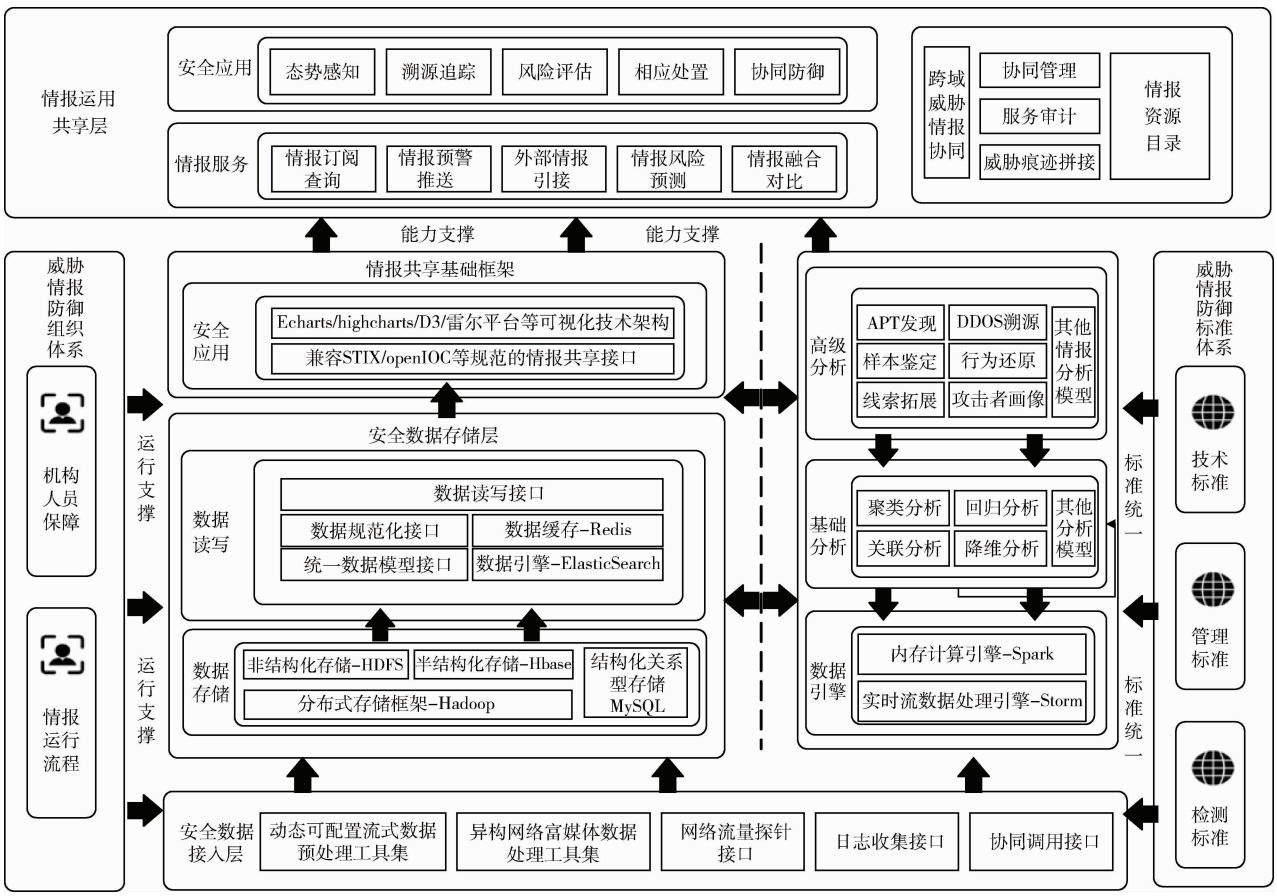

图 1 威胁情报技术体系架构

2.1 安全数据接入层

采集各类安全数据并对数据进行预处理,由各类结构化和非结构化的数据采集接口、数据处理工具集组成。

2.2 安全数据存储层

将接入层采集并预处理后的安全数据存储到 HDFS、MySQL 等结构化、非结构化数据库中,通过数据读写接口为安全数据分析层和情报共享基础框架提供数据调用和管理工作。

2.3 安全数据分析层

完成对数据的分析和重要信息提取工作,数据处理引擎采用 Spark、Storm 等计算框架,基础分析包括聚类分析、关联分析、回归分析、降维分析等,高级分析包括 APT 发现、DDoS 溯源、样本鉴定、行为还原、攻击者画像等。

2.4 情报共享基础框架

通过共享接口和 Echart 等可视化组件为情报运用共享层提供服务支持。

威胁情报防御组织体系为情报生成提供运行支撑,威胁情报防御标准体系为情报生成全过程提供统一规范的标准。

2.5 情报运用共享层

实施对威胁情报的组织运用,包括情报服务、安全应用和跨域情报协同等。

3 信息网络威胁情报生成

信息网络威胁情报生成过程主要包括安全数据采集、安全数据预处理、安全数据存储、安全数据分析、威胁情报服务五个步骤。通过生成的情报可形成内网威胁感知地图、支持安全事件溯源分析、促进全网系统协同响应等。

3.1 安全数据采集

一是网络中传统的关系数据库、其他结构化和半结构化数据库中数据的传递和采集,比如使用 Sqoop 工具来完成 Hadoop 的 HDFS 系统与关系数据库中的数据进行转移和抽取;二是采集网络中的实时流量数据,包括流量探针数据、网络交换机的端口镜像数据、安全防护、安全监控装备的系统日志(如入侵检测系统、网络防火墙)及报警日志信息等;三是采集外部网络数据,包括互联网数据或第三方数据,可以使用探测扫描或网络爬虫的方式,采集互联网中的资产基础数据,包括 IP 指纹、Web 指纹、DNS 解析记录、ICP 备案、whois、地理位置、CVE/CNCVE/CNVD/CNNVD 等漏洞信息,也可以使用直接购买、共享交换等形式获取第三方收集积累的情报数据。

3.2 安全数据预处理

将采集的数据存入安全大数据存储区域之前,需要使用各种数据预处理工具,对数据进行清洗、转换、抽取、归并、脱敏等工作。数据抽取过程将收集到的异构多源数据转化为单一或便于处理的构型,对不同来源的情报进行标准化数据表示,以满足快速分析处理的要求;数据清洗针对采集数据中存在的重复、错误、缺失等噪声进行过滤,保证数据的有效性,提升数据的置信度;数据归并通过扫描合并同类数据,保证输入数据的一致性和高质量。

3.3 安全数据存储

支持多种存储方式,包括数据仓库存储、NoSQL 存储、列式云存储等;支持灵活切换不同存储方式,以适应业务变化和快速配置要求,还支持对传统数据集市、数据仓库的数据集成。通过使用 Hadoop 集群技术能够实现多种计算、存储模式的云资源调度。通过建构一体化的数据处理模块,支持对多粒度和多模态数据的融合集成,形成标准的数据集提供给计算服务层分析处理。

3.4 安全数据分析

通过构建合适的数据挖掘、机器学习模型,如聚类模型、预测模型、分类模型等,使用以文本、图像、视频和传感器数据等为代表的多模态计算方法,或是以内存处理、实时流处理和批处理等为代表的多模式计算方法,对获取的多方面数据进行关联分析,挖掘出不同数据之间的隐藏联系,梳理网络整体的安全态势并进行可视化展示,从而提高网络整体在安全管理方面的低延时和可预测能力。

3.5 威胁情报服务

情报服务是威胁情报价值的最终体现,情报推送预警服务将可机读情报转化理解为安全措施、安全规则、安全策略,直接传送到相应的安全设备,其安全防护能力水平可得到有效提升;情报查询服务提供战术或战略情报等人读情报的搜索和汇集功能,主要面向安全运营和安全分析人员;威胁风险预测服务基于威胁情报形成威胁感知理解,评估系统面临的安全风险,预测可能遭受的攻击行为。

4 信息网络威胁情报运用

当前恶意组织和个人利用零日漏洞和 APT 攻击呈高发态势,其攻击范围、时间、技术方式均较难预测,对信息网络造成极大威胁。建好用好信息网络威胁情报体系,对尽快识别攻击行为、锁定攻击范围、发出攻击预警、协同应急响应、提升防御效能具有极大的促进作用。

4.1 安全威胁事前预测

基于对攻防对抗的海量安全事件理解,对网络风险进行辨识、分析和控制,在达到发生安全事件之前及时有效地控制攻击危害。基于对海量攻防数据的分析,梳理出实施攻击行动的行为模式,整合为威胁情报,则可越过解析系统安全告警这一耗时步骤,从而使精准检测安全态势和高效理解安全威胁成为现实,若再加上攻击推理树的配合使用,可有效提升安全威胁预测精准度。

4.2 攻击活动事中响应

通过威胁情报高位视角,获取最新的安全威胁信息,特别是 IP 信誉、失陷主机列表、恶意 URL 等入侵指标,将这些指标按预定准则与业务网络中的

实时监控事件流进行关联比对,能够提供与基于特征码完全不同的安全威胁检测能力;通过恶意信息、受害者信息,映射到网络攻击模型,利用带有场景特性的精准情报,分析鉴别内部网络里可能隐藏的已被渗透的攻击跳板,将安全威胁的发现点尽量控制在攻击链源端,根据情况及时发布预警信息,为安全防御展开争取更多的时间和空间,将损失降到最低。

4.3 安全事件事后处置

威胁情报可对定位攻击源提供可靠的数据支持,也可对回溯复现已发生的攻击行为提供可信的数据分析支撑。利用入侵指标,快速判定攻击影响面,利用攻击者攻击战术特点,进行攻击回溯分析。

在预告警分流中,威胁情报可用于区分不同种类的攻击行为,辨识出其中隐藏的 APT 攻击,确保应对措施及时有效;在确定攻击范围大小、溯源分析方面,可使用类型预测等相关指标,预判已出现的攻击痕迹之前或之后可能发生的恶意行为。

5 结论

信息网络面临的安全威胁不仅在数量上快速增长,在技术高水平化和多样化上也呈上升趋势。大数据技术的出现并成功应用到网络威胁情报生产中,为有效应对网络空间新威胁提供了强有力的武器。当前,信息网络安全威胁情报体系发展还处于起步阶段,大数据技术在情报研究工作中的应用还不够深入,对有效信息的筛选还不够精准,这些不足将在今后的研究中逐步解决。此外,在今后的工作中还将继续研究人工智能、机器学习、区块链等技术在信息网络安全威胁情报中的应用,有效促进信息网络威胁情报体系持续高速向前发展。

参考文献

[1] Gartner Research. Definition:Threat Intelligence[R/OL]. (2013-05-16)[2019-12-12]. https://www.gartner.com/en/documents/2487216.

[2] 魏为民,孔志伟,杨朔,等. 基于大数据的安全威胁情报研究[J]. 上海电力学院学报,34(1):53-58.

[3] 李超,周瑛. 大数据环境下的威胁情报分析[J]. 情报杂志,2017(9):28-34.

[4] 单琳. 网络威胁情报发展现状综述[J]. 保密科学技术,2016(8):28-33.

[5] 徐锐,陈剑锋,刘方. 网络空间安全威胁情报及应用研究[J]. 通信技术,2016,49(6):758-763.

[6] 吕宗平,钟友兵,顾兆军. 基于攻击链和网络流量检测的威胁情报分析研究[J]. 计算机应用研究,2017,34(6):1794-1797,1804.

[7] 林晨希,薛丽敏,韩松. 浅析网络安全威胁情报的发展与应用[J]. 网络安全技术与应用,2016(6):12-13.

[8] 朴嘉薇,周颖,郭荣华,等. 网络安全态势感知:提取、理解和预测[M]. 北京:机械工业出版社,2018.

(收稿日期:2020-03-17)

作者简介:

孙辉(1983 -),男,硕士研究生,工程师,主要研究方向:网络安全。

罗双春(1990 -),男,本科,助理工程师,主要研究方向:网络安全。

李余彪(1989 -),通信作者,男,本科,工程师,主要研究方向:网络安全。E-mail:405073973@qq.com。

版权声明