大数据驱动的网络综合监测系统的设计与实现*

曾　彬[1,2],张文沛[2]

(1.湖南友道信息技术有限公司,湖南 长沙 410000; 2.成都卓源网络科技有限公司,四川 成都 610041)

摘　要:当前的网络监测分析平台存在着检测数据来源单一、数据割裂、数据整合能力弱等不足,特别是新型攻击手段对传统基于已知规则库进行安全监测方法带来巨大冲击。充分利用大数据分析及预测技术,实现对 DPI、DFI、NetFlow、主动测量和扫描、SNMP、SLA 等多元化数据的采集,支持流量、性能、安全等结构化及非结构化数据的处理和存储,呈现以态势感知、流量透视、回溯分析、性能监控、安全检测、资产管理功能为核心,融合设备与拓扑管理、专题分析、攻击反制、异常文件识别和还原、主动测量等手段,具备不依赖规则而检测网络威胁的能力,打造一体化网络综合、智能化监测分析解决方案。

关键词:网络流量分析;安全监测;深度数据包检测;大数据;

中图分类号:TP393　　**文献标识码:**A　　DOI: 10.19358/j.issn.2096-5133.2020.02.001

引用格式:曾彬,张文沛.大数据驱动的网络综合监测系统的设计与实现[J].信息技术与网络安全,2020,39(2):1-7.

Design and implementation of network integrated monitoring system driven by big data

Zeng Bin[1], Zhang Wenpei[2]

(1. Hunan YouDao Information Technology Co., Ltd., Changsha 410080, China;
2. Chengdu ZhuoYuan Network Technology Co., Ltd, Chengdu 610041, China)

Abstract: Current network monitoring and analysis platforms have insufficient capabilities, such as single source of detection data, data fragmentation, and insufficient data integration capabilities, in particular, new types of attack methods have brought a huge impact on traditional security monitoring methods based on known rule bases. This paper makes use of big data analysis and prediction technology, to realize the collection of diversified data such as DPI, DFI, NetFlow, active measurement and scanning, syslog, SNMP, SLA, etc., to support the processing and storage of structured and unstructured data such as traffic, performance, security, etc., and to present situation awareness, traffic perspective, backtracking analysis, performance monitoring, security detection, asset management functions as the core, integrating equipment and topology management, thematic analysis, exception file identification and restore, active measurement and other means, with the ability to detect network threats without relying on rules, to create an integrated network comprehensive, intelligent monitoring and analysis solution.

Key words: network traffic analysis; security monitoring; deep packet inspection; big data

0　引言

随着网络攻击越来越复杂和隐蔽,传统以“防护”为主的安全体系面临极大挑战。各种检测技术从不同的角度发现网络中可能存在的安全问题,但无法准确和有效地找出网络中存在的真实威胁。以勒索、挖矿病毒等高级持续性威胁为代表的新型攻击手段,绕过了传统边界防护设备,传统的安全监测方法大都是基于已知规则库进行监测,但对未知威胁则无能为力[1-2]。对正在发生或已造成损失的入侵行为无法做到完整的溯源取证和损失评估。

另一方面,网络节点多、规模大、设备和资源众多,新业务层出不穷,宏观管理层面,缺乏对全局网络性能表现、资产设备、流量分布、用户行为、安全状况等的掌控[3]。微观运维层面,缺乏对网络原始数据的保存与检索分析能力,不能对故障及安全事故进行溯源查证[4]。

*基金项目:“十三五”共用技术和领域基金预研项目(31512080205)

随着网络带宽的持续增加,海量网络数据的实时采集、分析与存储是一个巨大的挑战,如果长期存储,代价巨大,另外解析出的流数据、元数据、包数据、事件数据等,存储模式差异巨大,如何融合分析也面临巨大挑战[5]。现有系统的交互能力都较弱,专业分析人员使用工具与数据进行交互分析,如事件调查、历史回溯、条件组合等的效果都有待提高,才能满足高质量高效的分析需求。

网络中现存的大量异构安全设备的告警从不同维度对攻击事件进行展现,虽然数据之间存在着关联,但是传统的分析方法无法将海量数据进行汇总和分析,缺乏对流量的深度分析能力和对未知攻击的检测能力[6-7]。Gartner 在 2013 年提出网络流量分析(NTA)是未来应对高级威胁的主要手段之一,它通过融合深度数据包检测、机器学习等技术,通过网络行为分析检测网络中的可疑行为[8-9]。

人工智能可以对网络流量内外所包含的无数元数据的海量关联进行分析;AI 技术可将包括企业运营日志数据和外部威胁情报服务的多个信息源整合分析,具备处理上百万数据点以及生成预测的能力,获得最准确的网络风险评估[10-11]。基于大数据和人工智能的网络运维减少了人为差错,智能化的监控有利于提高网络的安全防御水平[12]。

本文通过多元化采集、细化监控粒度,实现网络、主机、应用、数据、终端等物理或虚拟资产的全面监控。对网络安全态势进行多维画像、分析和预测,开展异构网络安全事件的综合分析,突破对多源安全威胁事件进行处理、挖掘和分析等关键技术,达到快速展现和解决网络的安全态势,形成相关的算法和系统,从而提升整体网络的安全能力,实现智能、联动、快速响应的“主动”防御,构建关键设备、资产的基础设施安全保障体系。

1 系统设计

系统采用全新的分布式协同测量体系架构,支持虚拟化部署与分权分域部署。总体架构方面主要包括:数据采集层、数据接入层、数据处理层、数据存储层、可视化展示层,如图 1 所示。

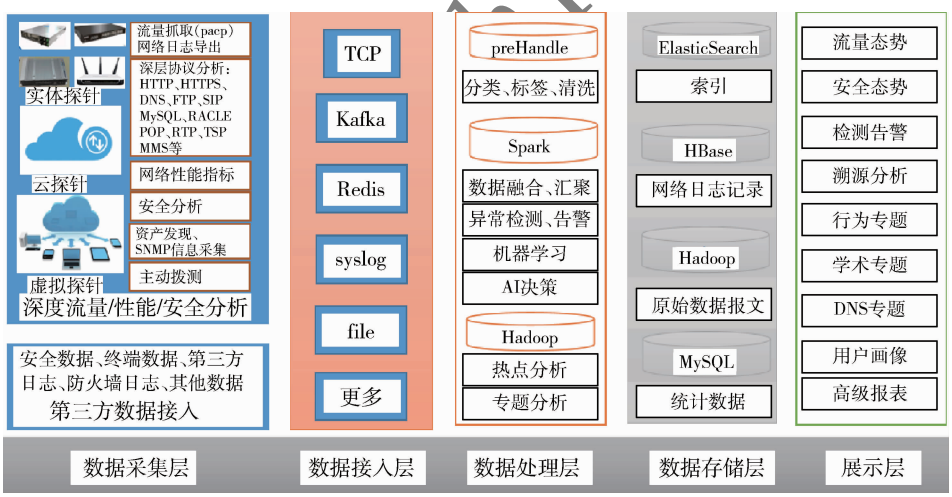

图 1 系统整体架构图

1.1 整体架构

数据采集层结合实时与非实时的数据采集方式,尽可能全方位立体化地收集网络空间多维度的数据源,实现实时海量异构安全数据采集。其中海量异构安全数据主要包括:安全日志、探测数据、网络流量、网络环境数据以及一些固有属性数据等内容。以主动或被动的方式收集漏洞评估数据、关键系统的运行日志及安全告警日志,以上数据包括实时数据和非实时数据,采用开源采集引擎。

数据接入层负责对收集到的数据汇总、预处理及分发。对汇聚平台上传的数据进行负载均衡,分发到数据处理模块,然后对数据进行去重、打标签、脱敏等操作。根据不同接口形式与格式分享数据。

数据处理层主要实现日志的关联分析、数据挖掘、异常发现等,包含实时分析和离线分析两个分析引擎。实时分析引擎拟采用 Storm/Spark Streaming 架构,用于数据实时统计、实时查询、实时抽样;离线分析引擎拟采用 Spark/ELK 架构,用

于数据挖掘和机器学习相关的分析。两种分析方式可以联动,利用离线的经验结果来判别实时发现的疑似异常。

数据存储层负责向展示层提供统计指标及原始数据。数据仓库拟采用分布式的存储结构,并采用高效的压缩、索引等算法保证数据的完整性、一致性且能够快速读取。使用 Redis 存储高频使用数据;HDFS 存储 PCAP 报文;HBase 存储实时统计数据;MySQL 存储汇聚数据、用户配置数据等。

展示层主要以 B/S 的方式,主要包括流量分析、性能监控、漏洞监测与管理、运行监测与管理、安全管理、攻击告警等模块。同时,将利用新型的可视化交互技术,将抽象的数据转变为人可感知的信息,将海量异构数据以图形图像的方式表现出来,在海量异构数据可视化层面,采用基于 WebSocket 的多数据并行、实时推送技术,辅以基于结构的方法、基于统计的方法、基于分割的方法、基于图压缩的方法,实现对大规模的数据监控。

经过大数据架构分析后的数据,可以供流量透视系统、性能分析与预测系统、安全态势与反制系统、流量朔源与取证系统、报表等系统实时或者直观可视化地展示数据。

1.2 大数据存储及检索方案设计

大数据平台对各个区域内流量进行实时采集并长期存储,通过全量采集数据并存储分析,结合特征规则检测方式,对网络流量进行安全特征检测及告警,同时可以提供告警日志及流量会话分析日志等信息通过 API 接口输出给第三方统一分析平台。系统自身能够长期采集并存储分析网络通信流量数据,包括 IP 地址、IP 会话、TCP/UDP 会话统计分析,并提供数据追溯分析能力,能够对任意时间段的数据进行追溯取证。通过自定义警报规则,能够对网络全流量进行检测预警,发现网络中安全隐患。对于分析日志及告警日志,系统提供 API 接口,能够通过 API 接口输出给第三方数据分析平台,实现结果统一分析及展示,如图 2 所示。

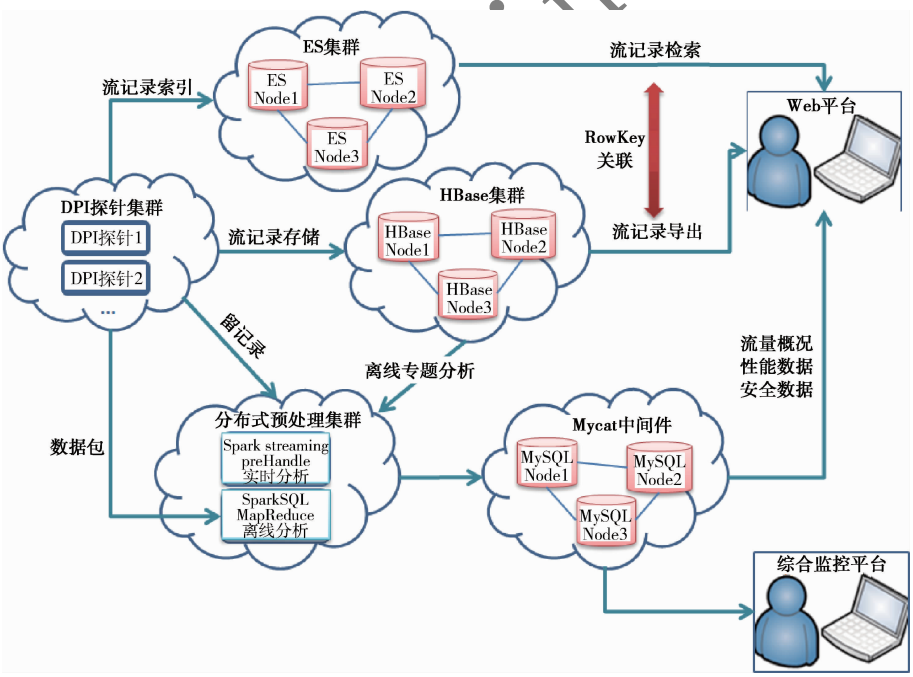

图 2 大数据采集调度设计方案

采用 HBase 作为原始数据的存储主体。采用 key-value 形式存储的非结构化数据库,通过其主键快速查询到原始记录。并且通过分布式扩展、数据压缩、批量写入,既可以保证海量流记录写入速度,又能保证系统的扩展性,在单机性能不足时可增加存储服务器节点解决问题。

采用 Mycat 作为 MySQL 的分布式中间件,将数据均匀分布存储至多个个 MySQL 数据库中,做到负载均衡。查询数据时,可通过 Mycat 中间件调用各个 MySQL 数据库进行数据查询,提高查询速度。在每个 MySQL 数据库中可采用分表策略进行数据存储,例如每天分一个表或者每 6 个小时分一个表,每

个表也可以建立相应的分区，将数据均匀分布在各个分区，在数据查询时只需在某些分区扫描数据，可提高查询效率。

在 HBase 的基础上，采用 ElasticSearch 搜索引擎对所有查询条件（IP、端口、协议等）建立二级索引，在需要多条件查询原始流记录时，先通过 ElasticSearch 进行索引搜索，由于 ElasticSearch 内部采用性能优越的 Lucene 搜索算法，可快速找到符合条件的唯一性的 key 值，然后利用 HBase 一级索引的优势，通过 key 值快速取出原始流记录，如图 3 所示。

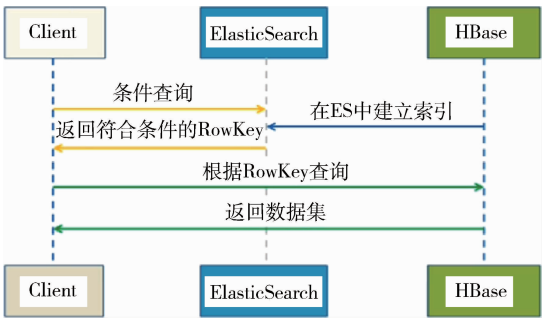

图 3　海量测量数据快速检索机制设计

2　关键技术

2.1　多维度测量数据融合与动态关联

系统支持分布式测试探针获得的指标进行综合分析，数据采集、存储、本地缓存，支持流量、性能、安全等结构化及非结构化数据的存储。对网络数据分析结果进行分层次建模分析，获得网络流量基准特征，并通过实时监测对网络异常行为进行告警，并自动触发深度流量分析功能，获得造成网络异常行为的地址、协议以及端口号等特征，有必要情况下自动触发相关网络设备的信息查询，包括接口状态查询等；并结合网络设备信息自动触发主动网络性能探测，对网络状况进行实时跟踪分析，从而帮助定位网络故障。

异构信息之间的挖掘、关联分析算法和技术：设计信息关联引擎和相关算法，包括事件采集、事件归一化（将各种日志格式标准化以消除异构产品来源间的差异）、聚合（将频繁重复事件如大量的端口扫描告警，聚合为一条事件）、事件关联分析引擎等几个部分，如图 4、图 5 所示。基于上述方法，开展

图 4　多维度测量数据的融合管理

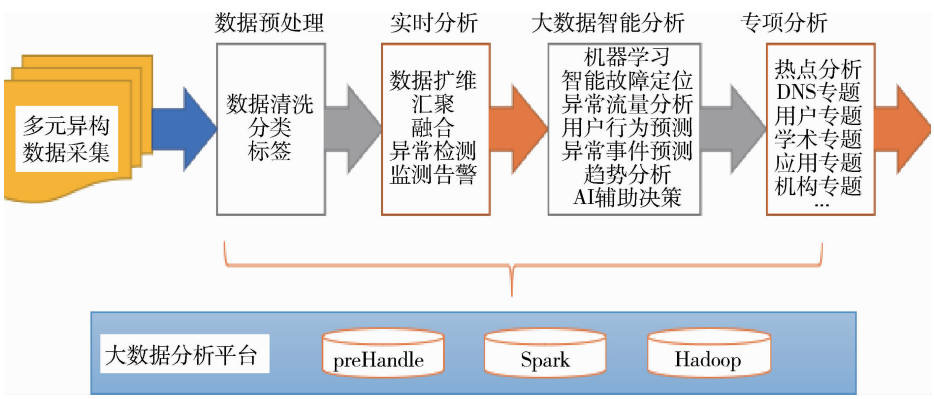

图 5　基于大数据挖掘的网络综合态势与关联分析

异构网络性能的综合分析,对网络异常行为进行预判和快速响应,基于数据挖掘算法,利用网络态势感知及预测技术,突破对多源威胁事件的处理、挖掘和分析等关键技术,达到快速展现和解决网络的安全隐患,形成相关的算法和系统。牵引性指标有:数据包监测性能、流量应用识别比率、支持的应用种类。

2.2 面向海量网络数据的存储与快速检索

应对海量数据的挑战,基于交叉表聚合的技术,处理数据存储、检索、共享与分发方法,系统实现了多时间粒度、多侧面立体式数据存储以及分布式存储。系统还支持通过外置磁盘阵列进行存储,保证结果可长期保存、可回溯,同时提出灵活方便的高效数据共享和分发方案,实现数据的高效共享,系统支持将通过高速探针采集得到的流量以多种形式导出共享,包括原始数据包、流(如 NetFlow V5/V9、sFlow 和 cFlow,以及自定义的 iFlow 格式,等等)、流量统计信息和抽象得出的流量事件(如流量构成、分布趋势和异常告警等系统专有信息等),如图 6、图 7 所示。

提供高效网络流量的索引方法,包括网络流量的索引特征,索引存储空间的消耗,索引记录插入的速率,索引的查询响应时间。采用 Redis 缓存上报日志流,ES 存储进行全量计算的数据处理框架,分开处理离线数据和实时数据,分别使用 Hive/Spark 来处理离线数据以及实时数据流的分析。

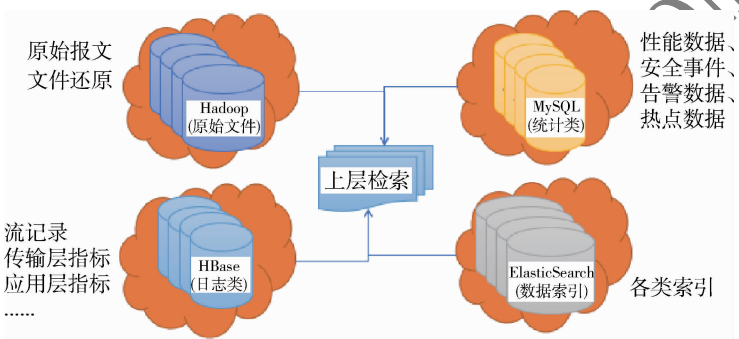

图 6 海量数据分布式存储、快速检索能力示意图

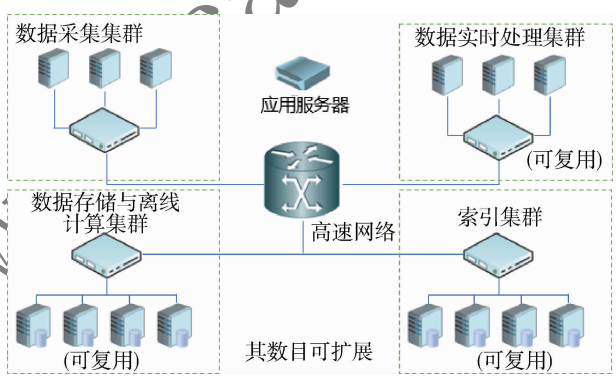

图 7 海量数据的分布式集群部署示意图

2.3 支持软件定义与虚拟化的新型测试架构

平台支持虚拟化与云端部署,实现功能模块的动态加载与销毁,可以扩展流量控制、攻击溯源反制、威胁文件还原等功能,具有分布式的海量网络测量管理数据共享平台,提供开放、通用、透明的数据查询分析接口。

研究实现测试探针与虚拟化、软件定义和边缘计算技术的结合。现有方案中多是独立探针形态,且位置不能更靠近用户侧。平台中既有独立探针,又有与软件定义智能网关结合的虚拟化探针,和独立的虚拟化探针(云化部署)实现完整的调度、配置、结果采集和数据分析框架,如图 8 所示。

3 系统实现与应用

系统通过多元化采集、大数据存储检索与 AI 智能分析,在 2 至 7 层用户通信数据报文还原分析,呈现态势感知、流量透视、回溯分析、性能监控、安全检测、资产管理六大主体功能,并结合各类设备与拓扑管理、专题分析、机器学习建模分析、攻击反制,

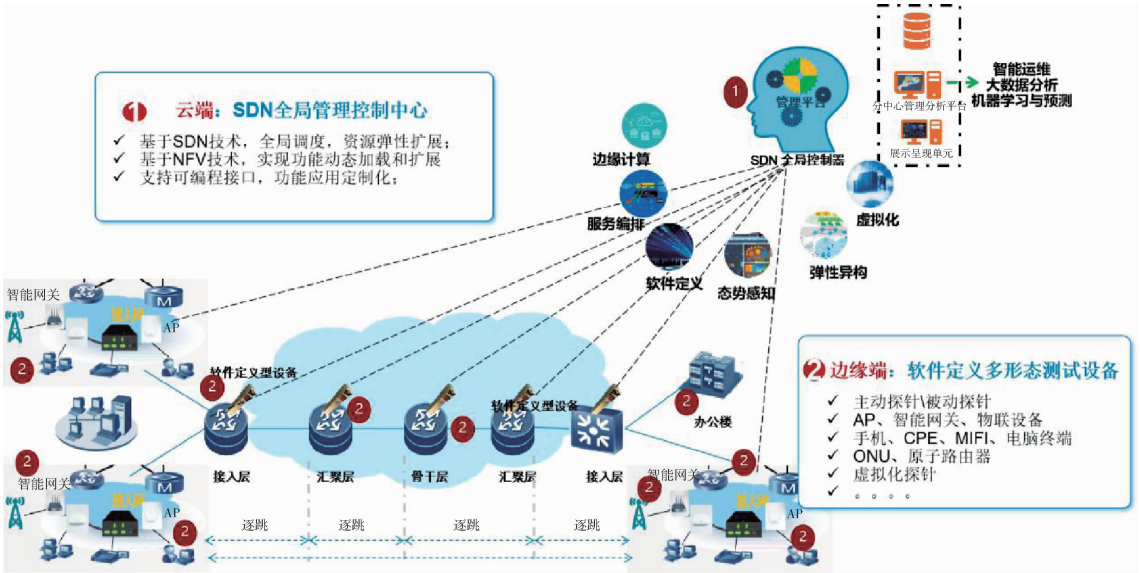

图 8　支持软件定义与虚拟化的新型测试架构

无损探测、异常文件识别和还原、主动测量、工控网络监控等功能，打造一体化网络综合、智能化监测分析解决方案。

（1）态势感知。全局展示网络走势、流量组成/分布情况、性能/告警提示，虚拟现实可视化分攻击方与被攻击方视角查看安全事件轨迹，展示攻击类型、严重等级、严重地理区域、攻击以及被攻击部门、机构、个人最严重 TOP 等信息。如图 9 所示。

（2）流量透视。支持实时监控流量，显示时间粒度可支持到 1 s，并可实时查看当前流量的应用、用户、外部地址的成分。系统支持以虚链路为基本对象，网络流量的应用、应用组、用户、用户组，国家、城市、外部地址、虚链路、学术平台之间的数据应支持相互关联，支持多维度的 Drill-Down 分析。

（3）性能监控。系统可以对 Web 网页、DNS、邮件、数据库、语音等主流业务访问服务的综合评价分析，包括建立连接时间、响应时间、连接成功率等关键性能指标；支持对业务的综合对比分析，按地区、按服务器角度对比分析，评价业务状况，定位性能故障和瓶颈。

图 9　网络态势感知

(4)回溯分析。保存会话记录数据包含记录时间、端口、MAC 地址、持续时间、总字节数、总包个数、速率等 25 个以上关键指标。安全事件能自动触发原始数据报文保存,留存攻击、入侵、病毒等数据报文,同时支持数据报文在线解码、在线分析、下载等功能。通过基准线告警能触发业务性能、网络性能、网络会话、安全事件数据的相互关联,在业务性能中查看关联会话记录,在会话记录中可以查看关联网络性能指标,在安全事件中查看关联原始数据包,在安全事件中查看关联会话记录。

(5)检测网络中存在的各种攻击行为,其中包括端口扫描类攻击、代码溢出类攻击、木马病毒类攻击、蠕虫类攻击、SQL 注入攻击、DOS 攻击、ActiveX 控件攻击、邮件类攻击、不良 IP 检测、个人上网检测、远程过程调用攻击、系统漏洞攻击等 35 个大类 50 000 余种攻击检测行为。支持用户自定义攻击行为:用户通过配置源/目的 IP、端口、特征字符在 payload 中出现的位置、数据包标志位、数据包频率定义安全事件。按任意维度(如告警类型、告警信息、目标国家、目标城市、协议、源地址、目的地址、嫌疑犯用户、嫌疑犯用户组、严重等级)多层次关联分析,多维度的 Drill-Down 分析,并且可以查看告警事件趋势、事件数目、流量状态等详情。

(6)资产管理。监测特定范围内的整体资产/虚拟资产概况,并进行概要统计分类,展示选定资产的总体状态及其安全告警情况,支持类型筛选、核心设定与合法设备配置,支持资产定时扫描和手动立即扫描,资产业务种类超过 2 300 种。支持资产的全生命周期管理,资产的安全状态评分,提示高危资产,对失陷资产的检测和发现,并给出安全加固建议。

4 结论

本文通过异构的网络数据采集边缘终端或软件,产生和获取多元的信息数据,流量和告警的数据采集;通过事件分析,异构安全信息之间的聚合关联分析引擎;大数据平台的支撑和保障海量数据的存储和高速处理。在单个平台内集成“态势感知、流量透视、性能监控、安全检测、回溯分析、主动拨测、资产发现、网络管理、数据挖掘”功能,通过多元化数据源采集与智能关联,实现大数据驱动的智能一体化运维安全监控分析。

参考文献

[1] BOTELHO J. When it comes to troubleshooting and threat detection, NetFlow AND packet capture trump all[J]. Network World (Online), 2013.

[2] Wang Jingyu, Jing Yuhan, Qi Qi, et al. ALSR: an adaptive label screening and relearning approach for interval-oriented anomaly detection[J]. Expert Systems With Applications, 2019, 136: 94-104.

[3] 李艳,王纯子,黄光球,等. 网络安全态势感知分析框架与实现方法比较[J]. 电子学报,2019,47(4):927-945.

[4] 王帅,汪来富,金华敏,等. 网络安全分析中的大数据技术应用[J]. 电信科学,2015,31(7):145-150.

[5] 琚安康,郭渊博,朱泰铭. 基于开源工具集的大数据网络安全态势感知及预警架构[J]. 计算机科学,2017,44(5):125-131.

[6] Lai Yingxu, Chen Yinong, Liu Zenghui, et al. On monitoring and predicting mobile network traffic abnormality[J]. Simulation Modelling Practice and Theory, 2015, 50: 176-188.

[7] 黄伟. 基于机器学习的 AIOPS 技术研究[D]. 北京:北京交通大学,2019.

[8] Gartnet. Gartner Market Guide for NTA 2019[EB/OL]. (2019-12-31). https://www.vectra.ai/download/gartner-market-guide-for-nta-2019.

[9] MOUSAVI S H, KHANSARI M, RAHMANI R. A fully scalable big data framework for Botnet detection based on network traffic analysis[J]. Information Sciences, 2019, 512: 629-640.

[10] 尤肖虎,张川,谈晓思,等. 基于 AI 的 5G 技术:研究方向与范例[J]. 中国科学:信息科学,2018,48(12):1589-1602.

[11] 邬贺铨. 对网络体系变革的思考[J]. 中兴通讯技术,2019,25(1):2-4.

[12] 龙成. 以网络全流量分析为基础的 AIOps 演进[J]. 网络安全和信息化,2018(4):18-20.

(收稿日期:2019-12-31)

作者简介:

曾彬(1979-),男,博士,技术总监,主要研究方向:网络测试、网络安全、大数据分析。

张文沛(1986-),男,硕士,主要研究方向:网络安全、网络测试、安全芯片。

版权声明