doi:10.3969/j.issn.1000-1247.2018.11.014

基于大数据的工业互联网安全初探

毛华阳
北京亚鸿科技发展有限公司

摘要 工业互联网安全问题对工业运行、经济发展、社会稳定带来严重影响，成为国家安全的重要组成部分。从工业互联网大数据安全分析角度出发，从工业互联网数据采集、存储、分析、应用等几个方面深入阐述利用现代信息技术，整合各种信息资源，搭建工业互联网大数据安全分析能力。

关键词 工业互联网 大数据 安全

1 引言

传统工业经过几次工业革命，已经进入瓶颈期，单纯从工业内部发展很难突破。随着互联网和大数据等信息技术的广泛应用，信息技术将会在工业领域产生一次新的工业革命，给工业带来全面的革新。德国主导的“工业4.0”，在信息互联基础上将大数据、云计算等信息技术引入了工厂，推动智能工厂的发展，以提高生产效率，降低能源消耗。美国倡导的“工业互联网”理念，采用先进的互联网技术打造新型的产业生态，旨在用互联网对产业进行全面革新。我国早在2013年9月就发布了《信息化和工业化深度融合专项行动计划（2013—2018年）》，将促进实体经济实质性和持续发展的重点放在“两化深度融合”，提出互联网与工业融合等创新行动。

2 工业互联网数据采集

工业互联网数据采集通常有两种技术，分别是主动探测和被动监测。

(1)主动探测技术：搭建工业互联网爬虫系统，对工业互联网组件进行主动探测和流量采集，完成数据收集，为工业互联网分析提供及时可控的数据。

(2)被动监测技术：在工业互联网网络关键节点、工业互联网平台、工业设备和系统部署基于DPI技术的工业互联网数据采集探针，实现对工业互联网流量的被动监测。

通过主动探测和被动监测两种方式，结合威胁情报资源库、漏洞库、第三方安全特征库，实现基于网络关键节点、工业互联网平台、工业互联网应用设备和系统、工业APP、标识解析系统的企业级和省级网络安全监测技术能力，建立多种恶意行为和安全漏洞的全网监测手段，实现对全网网络安全监测的能力。

3 工业互联网大数据整合和分析

工业互联网大数据存储分析依托HDP等通用大数据组件，通过构建SQL引擎、图引擎、离线计算引擎、内存计算引擎、流式计算引擎、作业调试引擎等核心引擎，打造大数据开发项目管理、数据集成、数据开发、数据管理和大数据挖掘分析5大子系统，并为数据业务层提供用户UI，包括可拖拽式的DAG开发流程图，以及标准化的API服务。

工业互联网大数据存储分析依托HDP等通用大数据组件构造的工业互联网大数据平台，主要功能分为两部分：第一部分是基于开源组织Hontonworks公司开发的开源HADOOP版本组件，构建SQL引擎、图引擎、离线计算引擎、内存计算引擎、流式计算引擎、作业调试引擎等核心引擎；第二部分是提供一站式大数据应用开发和管理平台，打通工业互联网数据加工的任督二脉。其提供数据集成、数据开发、数据管理、数据治理等全方位的产品服务，以及一站式开发管理的界面，专注于数据价值的挖掘和探索。将HADOOP作为核心的计算、存储引擎，提供海量数据的离线加工分析、数据挖掘的能力。

通过工业互联网大数据平台，可对数据进行数据传输、数据转换等相关操作，从不同的数据存储引入数据，对数据进行转化处理，最后将数据提取到其他数据系统。

3.1 大数据采集交换平台

大数据采集交换平台通过文件、消息等方式快速将工

业互联网原始数据集成到大数据平台，通过实时计算引擎、离线计算引擎的业务ETL处理，将原始数据经过数据清洗、过滤、核验，合并汇总后加载到数据仓库中，基于数据仓库进行数据的归类、分层、统计分析、数据挖掘，生成工业互联网平台业务数据专题库，如IP库、域名库、企业库、DNS库、漏洞库。

3.1.1 数据采集

数据采集是针对离线的大批量数据通过定义数据来源和去向的数据源和数据集，提供一套抽象化的数据抽取插件（称之为 reader）、数据写入插件（称之为 writer），并基于此框架设计一套简化版的中间数据传输格式，从而达到任意结构化、半结构化数据源之间数据传输的目的。作为一套生态系统，每接入一套新数据源，该新加入的数据源即可实现和现有数据源的互通。

利用Flume NG技术，对工业互联网数据进行集成。Flume NG是一个分布式、可靠、可用的系统，能够将不同数据源的海量日志数据（TB、PB级）进行高效收集、聚合、移动，最后存储到一个中心化数据存储系统中。Flume NG工作流程如图1所示。

Event：一个数据单元，带有一个可选的消息头。

Flow：Event从源点到达目的点的迁移的抽象。

Client：操作位于源点处的Event，将其发送到Flume Agent。

Agent：一个独立的Flume进程，包含组件Source、Channel、Sink。

Source：用来消费传递到该组件的Event。

Channel：中转Event的一个临时存储，保存有Source组件传递过来的Event。

Sink：从Channel中读取并移除Event，将Event传递到Flow Pipeline中的下一个Agent（如果有的话）。

Flume的核心是把数据从数据源收集过来，再送到目的地。为了保证输送成功，在送到目的地之前，会先缓存数据，待数据真正到达目的地后，删除自己缓存的数据。

Flume传输数据的基本单位是Event，如果是文本文件，通常是一行记录，这也是事务的基本单位。Event从Source流向Channel，再到Sink，本身为一个byte数组，并可携带headers信息。Event代表着一个数据流的最小完整单元，从外部数据源来，到外部的目的地去。

3.1.2 数据交换

实时数据分发是一个流式数据（Streaming Data）的处理平台，提供对流式数据的发布（Publish）、订阅（Subscribe）和分发功能，可以轻松构建基于流式数据的分析和应用。

实时数据分发服务基于大数据平台，具有高可用、低延迟、高可扩展、高吞吐的特点，与流计算引擎FLINK无缝连接，用户可以轻松使用SQL进行流数据分析；也提供分发流式数据到各种产品系统的功能。

3.2 大数据开发平台

数据开发平台提供可视化开发界面、任务调度运维、快速数据集成、多人协同工作等功能，为用户提供一个高效、安全的数据开发环境，拥有强大的Open API，为数据应用开发者提供良好的再创作生态，是大数据平台中的一个子系统。数据开发平台功能架构如图2所示，数据开发平台技术架构如图3所示。

3.2.1 项目管理

数据开发平台是以项目为单元实现对物理空间和计算资源的管理，项目管理模块实现平台项目的创建，成员的管理，物理、计算资源的配置。

(1)项目配置

可通过配置项目的操作，对当前项目空间的属性进行管理和配置，如发布目标、项目别名等。

(2)项目成员管理

可通过项目成员管理操作，对当前项目空间的成员进行管理和配置，选择用户权限管理模块中的用户添加到当前项目中，并配置部署、开发、访客、运维等角色。

3.2.2 数据开发

通过平台界面实现大数据作业的开发，DAG流程图的拖拽，业务流程以及数据流程的贯穿，提供脚本SQL、

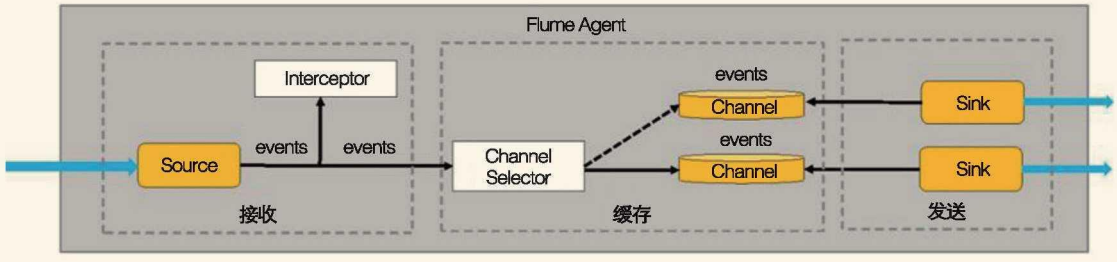

图1 Flume NG工作流程

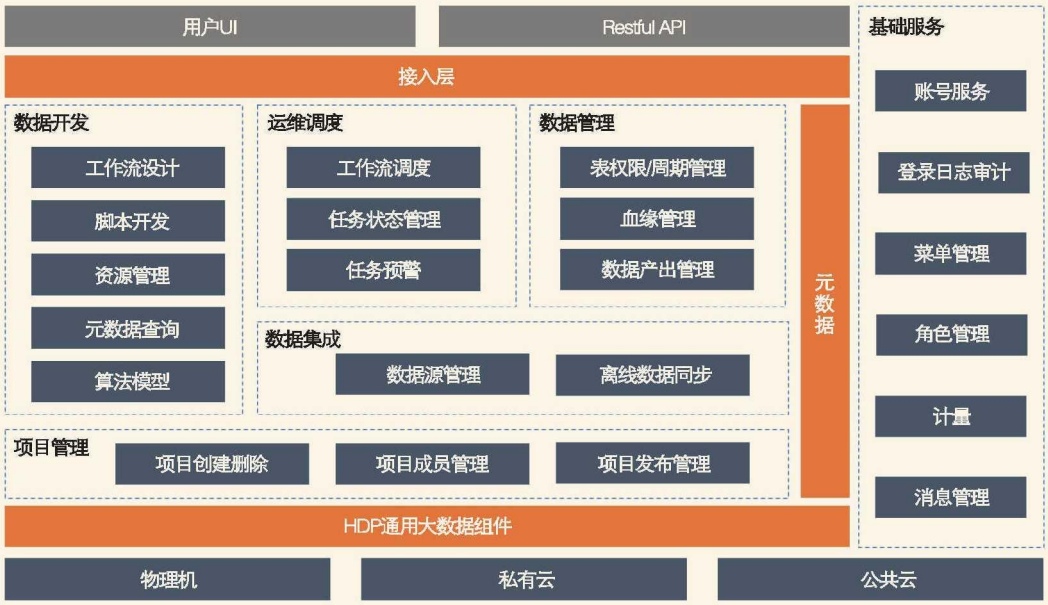

图2　数据开发平台功能架构

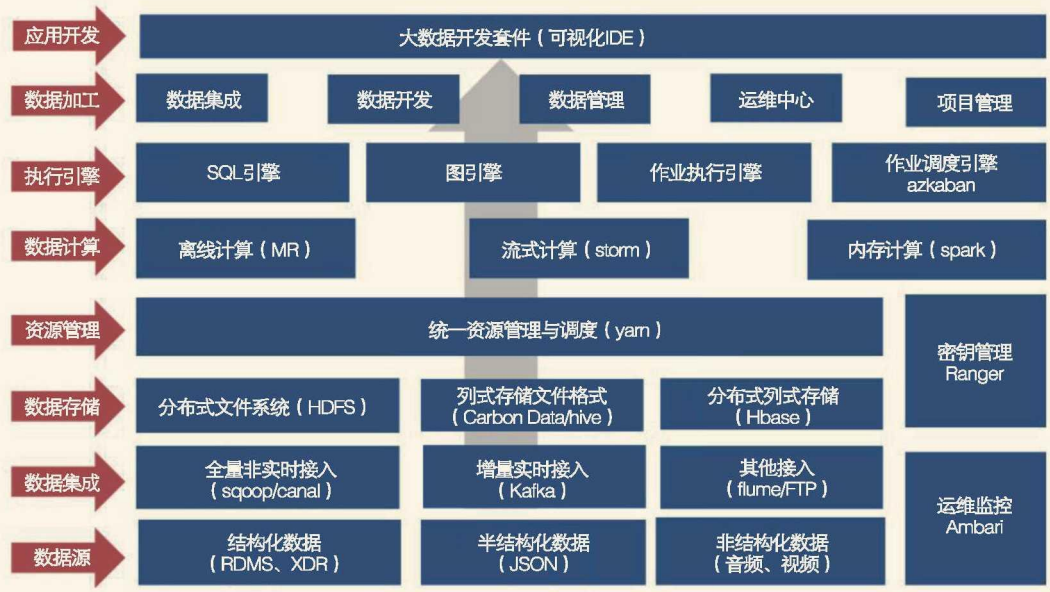

图3　数据开发平台技术架构

shell，任务mapreduce、spark等离线和实时作业的在线开发。

3.2.3 数据管理

数据管理模块中可以看到组织内全局数据视图、分权管理、元数据信息详情、数据生命周期管理、数据表/资源/函数权限管理审批等操作。

(1)数据权限申请

针对平台中的表、函数、资源三种数据类型进行严格的权限申请后才能使用。

(2)新建表

数据开发过程中需要创建表来存储数据同步、数据加工的结果数据，数据管理模块提供可视化建表、语句建表两种方式创建表。

(3)修改生命周期

数据表管理模块对数据表进行了分类，并对各个分类提供不同的表信息以及表操作管理功能，以便广大开发者管理自己的数据表。在数据表管理中，可以对表进行以下操作：

生命周期设置、表管理（包括修改表的类目、描述、字段、分区等）、表隐藏/取消隐藏和表删除。

(4)查看表详情

包括表的基本信息、存储信息、字段信息、分区信息、产出信息、变更历史、血缘信息和数据预览。

(5)数据质量

数据质量主要对分区表数据的准确性和数据质量进行校验，按照通用和自定义的规则进行校验和检查，并通过可视化的工具对问题数据和任务进行记录和展示。

3.2.4 运维中心

运维中心是对任务和实例展示/操作的地方。可以在任务列表中看到全部任务，可以对展示的任务进行测试、补数据、添加报警、修改责任人、冻结/解冻等操作。在任务运维中，可以看到所有任务的实例，可以对展示的实例进行终止、重跑、解冻等操作。

(1)运维概览

运维概览主要是对任务运行情况的报表展示。

(2)任务列表

任务列表主要展示了提交到调度系统的全部任务，主要有离线周期任务、实时任务和手动任务这三个子分类。

(3)任务运维

任务运维主要展示了任务提交到调度系统后，经过调度系统/手动触发运行后生产实例的展示列表，主要有周期实例、手动实例、测试实例、补数据实例这4个子分类。

(4)报警

报警主要是对任务运行情况的监控，若被监控的任务未运行或运行失败，将会收到提示信息，主要有监控记录和监控设置这两个子分类。

3.3 大数据应用平台

大数据应用平台是以标签中心为基础，建立在跨多个计算资源之上的统一逻辑模型，可以在“标签”这种逻辑模型视图上结合画像分析、规则预警、文本挖掘、个性化推荐、关系网络等多个业务场景的数据服务模块，通过接口的方式进行快速的应用搭建。这种方式的好处，一在于屏蔽应用开发人员对于下层多个计算存储资源的深入理解与复杂的系统对接工作；二在于通过数据服务的形式透出，有助于IT部门对数据使用的管理，避免资源的重复和冗余。简单来说，因为大数据计算能力的增强，只需要把需要使用的数据在模型当中进行管理后，即可通过API方式进行相应的计算对接到产品界面端上，或通过提供的界面配置功能直接生成可以独立部署的代码，快速搭建相应的大数据应用产品。

数据应用主要包含三大模块，分别是标签管理、引擎管理、服务管理。各模块的功能如图4所示。

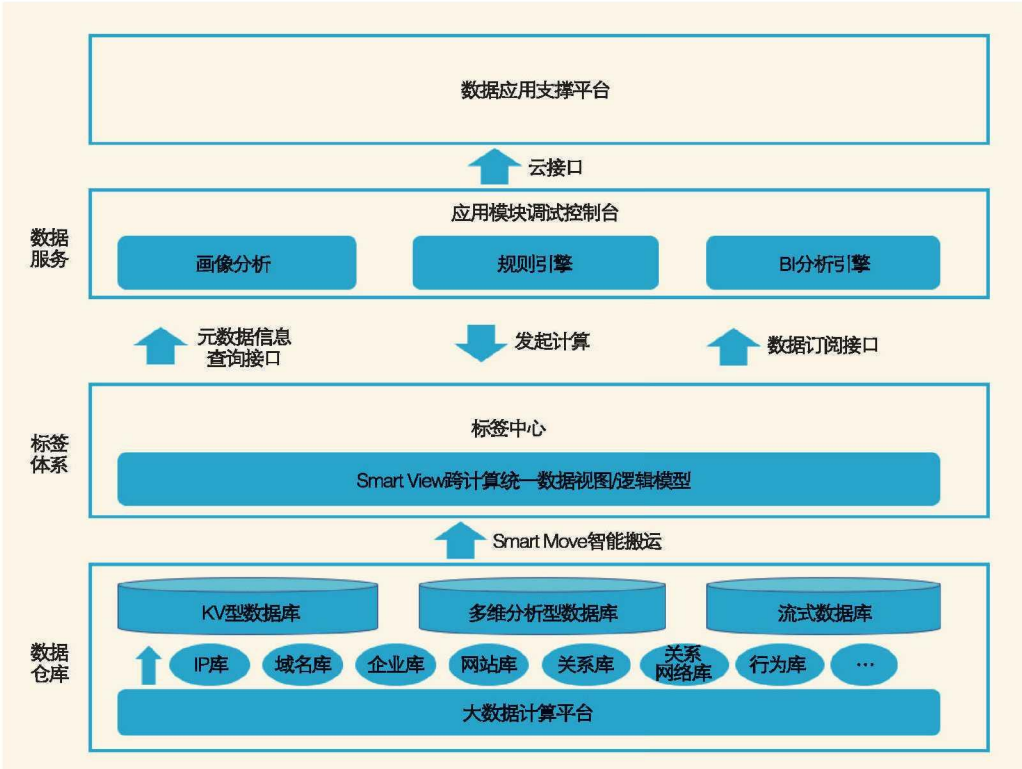

图4　数据应用支撑架构

3.3.1 标签管理

标签类目以树形结构对数据资产进行分类管理、展示和检索，通过类目的形式组织标签可以实现对数据的规范化及标准化管理。让用户通过可理解的分类方式快速查找数据、标签。

标签中心沉淀行业典型场景标签体系，助力客户快速建设自有数据资产。支持标签类目结构、状态管理，支持标签字段管理，支持对标签使用属性进行监控统计，可阅读可理解的标签建模理念，协同顶层业务和底层技术视角。

3.3.2 引擎管理

将标签作用于实际

场景的发动机，如标签通过分析引擎能实现各种聚合的报表，通过查询引擎可得到查询标签的其他标签信息。通过引擎管理来实现分析引擎、数据查询引擎、挖掘算法引擎等各类引擎的上下架管理及引擎生命周期的管理。

(1)机器学习算法引擎

数据挖掘安全分析基于大数据平台的机器学习库MLIB，通过对底层的分布式算法封装，提供拖拉拽的可视化操作环境，让数据挖掘的创建过程像搭积木一样简单。缩短用户与数据的距离，真正实现数据的触手可及。同时提供脚本式算法开发调用，方便将算法嵌入到工业互联网的数据挖掘工程中。

根据业务功能需求，对监测工业互联网流量数据中暴露的工业互联网云平台、工业互联网应用设备和系统、工业APP、固件等重点防护目标的安全情况，以及扫描源组织等威胁源进行跟踪分析。安全分析的结果通过统一存储引擎写入到数据存储模块中，供其他业务系统调用。

（2）敏捷BI分析引擎

BI分析引擎支持Oracle/MySQL/Greenplum/Hive/PostgreSQL/Vertica等28种类型的数据源的快速接入及管理。

统一提供敏捷的商业智能分析工具来快速地完成数据的分析、建模、展示生成API等。

3.3.3 服务管理

数据服务是架设在标签视图之上的业务功能模块，可以通过界面化的配置或API的操作，以标签为粒度对跨计算资源的数据进行统一的业务计算操作。透过数据服务+逻辑建模的组合，既节省工作量又有很好的扩展性。特别是对于工业互联网大数据环境需要整合多个系统数据的前提下，很难一次把所有数据需求全部规划完整，那么这种动态逻辑建模的方式就有非常好的扩展性。从应用的角度来说，由标签模型视图层隔开明细数据复杂的数据结构，能够在相对扁平的标签体系之上进行明细数据的计算和查询，对于数据开发与应用开发的分工流程来说更为合理。

4 工业互联网大数据安全分析应用

通过对工业互联网流量进行深度分析，利用大数据挖掘技术、机器学习和敏捷BI分析能力，可以从多个维度关联发现工业互联网安全事件和隐患，部分安全分析应用如下。

4.1 工业互联网资产管理

(1)资产活跃度分析

通过系统收集活跃的工业资产IP，实时发现现网中的工业资产访问情况，分析工业资产访问活跃度，对用户访问行为进行态势感知分析，同时通过多种方式对数据进行展现。

(2)工业互联网发展分析

围绕工业互联网涉及互联网平台、工控设备、终端等进行发展分析，通过多种图表样式进行展现。

4.2 工业互联网安全监测

(1)工业互联网僵尸事件监测

结合僵尸威胁数据特征和系统数据关联出攻击目标的情况，从而分析出攻击对象、攻击规模、攻击时长等信息，实现对工业互联网僵尸事件的监测分析。

(2)工业互联网木马事件监测

通过流量分析木马指纹特征及周期通信特征，结合系统数据关联出感染的平台情况，分析影响范围等情况。

(3)工业互联网后门事件监测

通过流量分析后门特征及后门控制信息、后门类型、后门使用IP、占用端口、后门特征信息等，结合系统数据关联出感染的平台情况，分析影响范围等情况。

(4)工业互联网蠕虫事件监测

通过工业互联网数据，分析蠕虫特征及蠕虫控制信息、病毒名称、病毒类型等，关联出感染的平台情况，并分析影响范围等情况。

(5)工业互联网恶意APP事件监测

监控工业互联网APP使用情况，对获取的工业互联网APP进行检测分析，并结合系统数据分析恶意APP的使用、下载情况以及相关威胁情况。

(6)工业互联网平台篡改事件监测

监控工业互联网平台站点的安全情况，对获取的平台数据进行篡改分析，并结合系统数据分析被篡改页面的访问情况。

4.3 工业互联网安全分析

(1)攻击画像分析

关联攻击事件、异常流量、漏洞信息、情报信息等，形成目的IP、目的端口、源IP、源端口、漏洞级别、应急情况等多维度的攻击者画像。

(2)访问行为溯源分析

针对攻击事件的IP、域名等索引信息从访问日志数据中分析出攻击者的访问数据，进行行为溯源分析。

4.4 工业互联网安全预警

(1)网络嗅探事件预警

对特定端口、IP、URL及协议的访问情况进行分析，通过配置阈值进行预警。

（下转58页）

还有平台的位置与性能、管理范围、数据库所在地、地面传输通道路径、采用技术、节点使用设备、终端性能参数等对上述指标的优化均有一定关联度。

平台建立后加强近海船舶从业人员的宣传教育仍十分必要。

成本估算目前尚无法估计，主要是没有掌握安装北斗定位器船舶数量（大连沿海的渔船据统计在4千多艘），长期来看北斗导航芯片价格产业规模化后会是一个大幅下降的趋势。

5 系统平台的其他应用

平台可以兼容陆地光缆的维护，由于陆地通信方便得多，因此只要有手机信号的地方，传感器就可以建立连接，传感器可以对声、光、湿度、红外线、震动、坐标位置发生偏移（偷盗或地质灾害发生时）、电源容量下降等发出告警信息和告警位置坐标及终端ID号码。

平台可以兼容所有管道型的线状、点状、块状资产进行维护，与物联网做法相同。其他行业可借助这个平台为己所用，形成共享效应后可以进一步降低单一客户的使用成本，如此可以达到少花钱多办事的效果。并通过注册专业用户的方式来隔离、管理和监控各自的资产。

满足各类公共信息发布：洋流、海况、气象、风暴在所管理资源周边的情况等。

信息平台可以了解运行中的船舶、车辆信息，并以此为基础开发更多的应用，如物流运输管理、运力调度、国防动员、救灾抢险、事故救助、租借车辆管理等，出售给第三方市场的话，只要想得到和用得上的均可开发应用。

6 结束语

近海渔船作业是近年来海底光缆中断的主因，提前发出预警是避免海底光缆中断后带来巨额损失的必要方式，发出能够追溯的有效信息和语音警告是有效的管理方法，北斗导航与短信平台以及VHF通信是可行的工具。平台建设必须对时效性和反应时间予以重视，其与平台的信号处理与传输时延和“禁锚区”范围划定密切相关，平台建设兼顾其他行业应用是降低成本的有效途径。

世界海事新闻网9月7日文章《世界一流海事的国家》刊登了总部位于德国的DNV·GL海事集团研究报告，报告从航运、金融及法律、海事技术和港口及物流等核心指标对全球30个处于领先地位的海事国家进行评估，30个国家按照4大海事核心指标及分级指标的规模和程度进行了先后排名。中国4大指标均排第一，比排名第二之后的国家具有压倒性优势（第二美国，第三日本，第四德国、韩国和挪威并列第五、第六希腊），但从光缆频频阻断这一点看，我们的管理仍存在待改进的地方，如果能够被改进这些问题，国家的海事管理将迈上更高的台阶。

参考文献

[1] 林恒易.电力公司海底光缆运行管理系统的设计与实现[D].福建：厦门大学,2017

[2] 李冬航.卫星导航标准化研究：系统篇[M].北京：电子工业出版社,2016

[3] 刘建.北斗卫星导航系统在交通运输行业的应用与发展[M].北京：人民交通出版社,2017

如对本文内容有任何观点或评论，请发E-mail至ttm@bjxintong.com.cn。

（上接53页）

(2)暴力破解事件预警

通过特定工业互联网平台、URL及协议的多次尝试登录情况，发现登录行为已经超出人工登录频次并多次出现错误，通过配置阈值进行预警。

(3)工业互联网漏洞预警

通过收集到的CNVD漏洞信息、威胁情报（0day漏洞）关联工业资产库分析存在相应漏洞的资产，能够快速排查最新威胁，通过配置阈值进行预警。

(4)工业互联网异常流量预警

通过分析收集到的系统数据，并对监控的数据按照五元组归类，可依据目的IP、目的端口、源IP、源端口的方式统计数据流量和流向信息，通过配置阈值进行预警。

(5)工业互联网仿冒平台预警

分析收集到的系统数据，分析出可能被仿冒的平台URL，交由网络爬虫进行获取，对获取的数据进行仿冒分析。

(6)工业互联网仿冒APP预警

分析收集到的系统数据，找出可能被仿冒的平台URL，交由网络爬虫进行获取，对获取的数据进行仿冒分析。

5 结束语

文中系统地阐述了基于大数据采集、大数据存储、大数据挖掘分析及大数据应用的相关技术，构建工业互联网大数据安全分析平台，实现基于大数据技术的工业互联网安全整体趋势分析、研判分析、预警通报。形成上下贯通、政企协同、多方联动的安全保障体系，推进工业互联网安全技术手段建设，保障工业互联网安全，强化企业主体责任，支撑政府安全分析管理。

如对本文内容有任何观点或评论，请发E-mail至ttm@bjxintong.com.cn。