北京航空航天大学研究生课程考核记录

2018－2019学年 第一学期

学号 ZF1821308 姓名 夏涛 成绩

课程名称：《大数据原理与分析》

论文题目：大数据在工业控制系统信息安全领域的应用

任课教师评语：

任课教师签字： 考核日期： 年 月 日

# 问题的提出

## 工业互联网的发展

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。当前，以新一代信息技术引领未来制造业发展，加快塑造国际竞争新优势已成为主要工业强国的共识，以美国先进制造、德国工业4.0等为代表，虽然各国战略计划各有侧重，但究其本质，均是以工业互联网为驱动的智能制造变革。工业互联网是互联网和新一代信息技术与工业系统全方位深度融合所形成的产业和应用生态，是工业智能化发展的关键综合信息基础设施，其本质是以机器、零部件、控制系统、信息系统、产品以及人之间的网络互联为基础，通过对工业数据的全面深度感知、实时动态传输与高级建模分析，形成智能决策与控制，催生智能化生产、网络化协同、个性化定制、服务化延伸等新型制造模式。在新一轮科技革命和产业变革背景下，工业互联网是实现智能制造并进而抢夺全球制造业竞争优势的关键，并对互联网的未来发展、全球竞争和国际治理产生深远影响。

## 工业控制系统信息安全

工业控制系统信息安全（以下简称工控安全）是实施制造强国和网络强国战略的重要保障。近年来，随着中国制造全面推进，工业数字化、网络化、智能化加快发展，我国工控安全面临严重的威胁。主要表现在以下方面：1.大部分制造企业厂商安全意识不强；2.工控设备数量和状态不明；3.安全漏洞数量不断增多；4.安全威胁加速渗透、攻击手段复杂多样。以上这些问题的解决需要寄希望于以大数据、人工智能等为代表的新兴技术。大数据技术在工业互联网设备数据的采集，存储，管理，集成，处理，分析等方面将发挥重要的作用。本文主要探讨大数据技术在工业联网设备的日志分析方面的应用，主要通过对工控设备的日志进行分析来进行恶意IP统计和异常检测，进而对某一厂商或地区的工控设备安全态势进行判断。

# 数据来源

工业互联网数据的来源是多种多样的，在工业互联网网络关键节点、工业互联网平台、工业设备、工业APP等，以及工业爬虫从部分网站爬取的威胁情报信息、设备漏洞信息、以及蜜罐收集的网络攻击信息等。大部分的数据均属于文本数据，本文主要分析的工业联网设备的安全日志基本都是文本型数据，当前，在大型局域网内部，日志文件的种类众多、格式不一、体量庞大(达到 ＴＢ 、 ＰＢ 级别)，传统的基于关系数据库的日志分析方法性能急剧下降，无法满足海量数据的处理需求。在大数据领域，日志是广泛使用的数据采集方法之一 ，具有４Ｖ 特征：数据体量巨大(volume)、数据类型繁多(variety)、有价值但密度低(value)、处理速度快(velocity)。因此，大数据技术是解决日志分析的更有效手段。



图 1 防火墙日志样本

# 解决方案设计

大数据工控设备日志分析系统的功能模型如图 1所示，此方案主要包括日志收集，日志预处理，日志挖掘和日志存储模块。

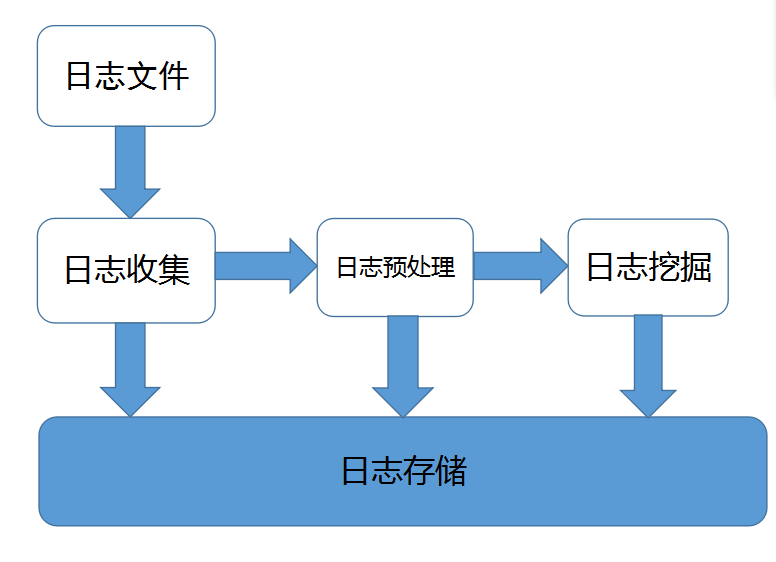


图 2 功能模块

## 日志数据的收集

日志数据的收集是对其进行分析的第一步，而Flume NG 是当前主流的日志收集框架，它是一个分布式、可靠、可用的系统，能够将不同的数据源的海量日志（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。其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，删除已缓存的数据。Flume 工作流程如图 2所示。

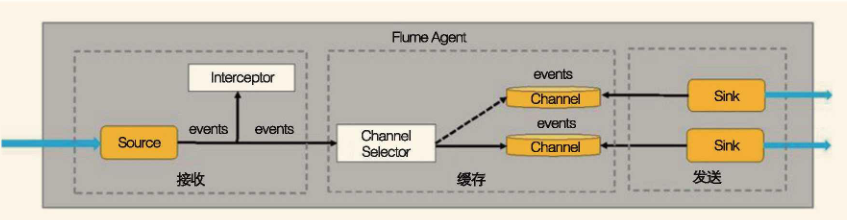


图 3 Flume 工作流程图

## 数据预处理模块

数据预处理作为该系统的一个重要模块，是数据挖掘的基础。本系统则将数据直接存入到 HDFS文件系统中，默认切分为每块为64 MB大小的文件，

保存在各个DataNode上。如下图 3所示。

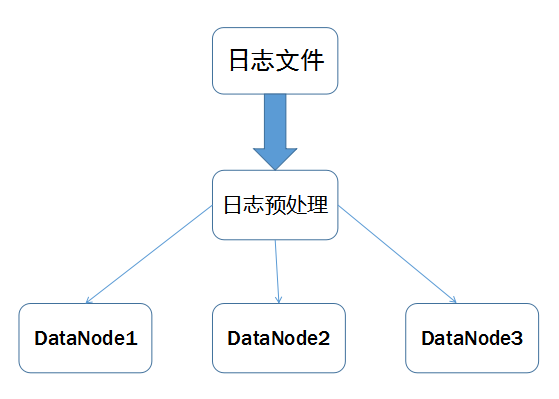


图 4 数据预处理

## 日志挖掘模块

### 3.3.1源IP统计算法

无论是CC攻击，DOS攻击还是DDOS攻击其基本原理都是通过频繁访问服务器，使服务器始终处于高负荷运转，服务器资源被耗尽，最终导致服务器瘫痪。

因此对服务器日志中的IP地址的请求次数进行统计，找出高频访问的IP地址，是检测攻击源，和防御攻击的有效方法。为了提高统计的效率我们可以借助MapReduce并行计算的特点，将传统的单机计算的方法改进为并行算法。Map函数的输入为从HDFS中读取的每一条日志记录，输出为Key/value形式，Key为SourceIP，故输出为SourceIP/1;Reduce函数用于将相同的SourceIP合并，其输入为相同的Key（SourceIP）然后将value相加，其输出为Key/value,SourceIP/n。表示同一个SourceIP对服务器访问的次数。如下图 4所示。

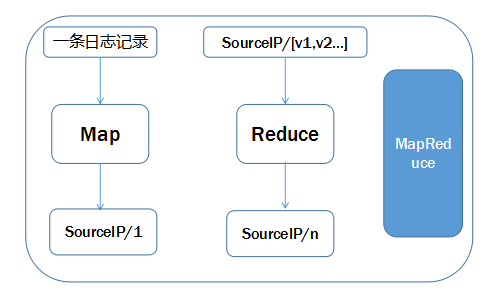


图 5 MapReduce函数

### 3.3.2异常检测算法

网络中对的工控设备的流量日志记录了该设备的行为，当设备遭遇攻击或病毒、木马时，设备会表现出与被攻击前不同的行为状况，我们通过分析设备的流量日志，可以分析出异常流量，发现设备被攻击的行为。异常检测是发现那些行为不符合期望的数据对象，它们与大部分数据不相似、不一致，这种对象被称为异常、噪声或离群点。计算机入侵检测中的异常可能意味着入侵行为的发生。

我们在进行异常检测时，应该以时间轴为参照对同一设备的历史流量数据进行纵向的考量。每一条日志文件应当包含如下信息：设备IP，时间(time)，单位时间内设备产生的流量(traffic)，单位时间内与该设备交互的IP的个数(interactionIP)。我们将traffic和interactionIP投影到二位维平面上，traffic表示横轴，interactionIP表示纵轴，这样我们就成功的将一条流量日志转换成了平面上的一点，该点表示某个IP的设备在某个时间点的流量信息。接下来我们就可以利用基于密度的异常检测算法来判断在该点处设备的流量状态是否正常。基于密度的异常检测算法的思想是：对整个数据空间的密度分布进行计算，存在于低密度区域的数据会被认为是异常数据。具体思路如下：

首先定义,表示，则至少存在k个点使得;定义表示距离点的距离不超过的所有点的集合；定义为可达距离;定义为局部可达密度；定义为局部偏离因子。

但是在实际运用中，由于工控设备的日志数量极为庞大，传统的LOF算法不能有效的处理如此海量的日志数据，因此我们需要借助大数据技术对传统的算LOF法进行改进。我们引入MapReduce并行计算，在集群中完成异常检测。Map函数的输入为流量日志文件，一行为一条记录：IP,time,traffic,interactionIP;输出的Key/Value是：IP/time,traffic,interactionIP。Reduce函数用于汇聚相同的IP的数据输入时IP/list[time,traffic,interactionIP]，输出Key/Value为：IP/time/lof，表示在time时刻IP所标识的设备的局部偏离因子。

# 应用成效

在大数据时代许多传统算法的性能难以满足大数据量的要求，为此我们将LOF算法在Hadoop平台下基于MapReduce模型对其进行改进，重新实现了源IP统计算法和基于密度的异常检测算法，可以应对未来工业互联网时代大批量工控设备日志分析的要求。通过对工控设备流量日志的分析，有效的发现异常流量设备，实现工控设备的安全防御，避免遭受攻击。

# 项目总结和个人思考

在信息安全领域，网络日志的分析与挖掘技术有着广泛的应用。随着工业制造2025规划的实施，工业互联网的建设不断推进，联网工控设备的信息安全正面临越来越严重的威胁，因此，通过分析和挖掘工控设备的日志隐含的信息来对工控设备的安全状况进行评估，发现非法入侵，保护厂商的工控安全具有重要的意义。本文中我们基于大数据技术，在Hadoop平台下重现了源IP统计算法和基于密度的异常检测算法，将传统的算法与大数据技术进行融合，此方案能够有效解决工控设备日志数量庞大的问题，高效的对源IP数量进行统计，和实现异常检测，为防御非法入侵，保障工控设备信息安全提供了一个良好的思路。

当然本文只是针对于大数据时代工控设备日志分析的一个初步尝试，并未提出一套完整的基于工控设备日志分析的检测恶意攻击的方案，因此有待于进一步的优化和设计。