安全小课堂第七十八期【大数据安全】

京东安全 京东安全应急响应中心 2017-11-17

大数据时代已经到来，大数据渗透到各个行业领域，逐渐成为一种生产要素发挥着重要作用，成为未来竞争的制高点。然而，大数据掀起新一轮生产率提高和生活方式改变的同时，随之而来的是安全挑战。

JSRC **安全小课堂第七十八期**，邀请到**兜哥**作为讲师就**大数据安全**为大家进行分享，同时感谢白帽子们的精彩讨论。

什么是大数据安全?

京安小妹

兜哥:

我个人理解，大数据安全有三个方向：

一是大数据架构或者说软件自身的安全；二是使用大数据做正向的安全建设，解决目前的安全问题，这个是目前很多大公司以及创业公司重点做的；三是使用大数据作为工具发起攻击，这个目前已经出现很多案例了，比如AI攻击验证码以及一些智能自动化攻击工具。总的来说就是这三个方面。

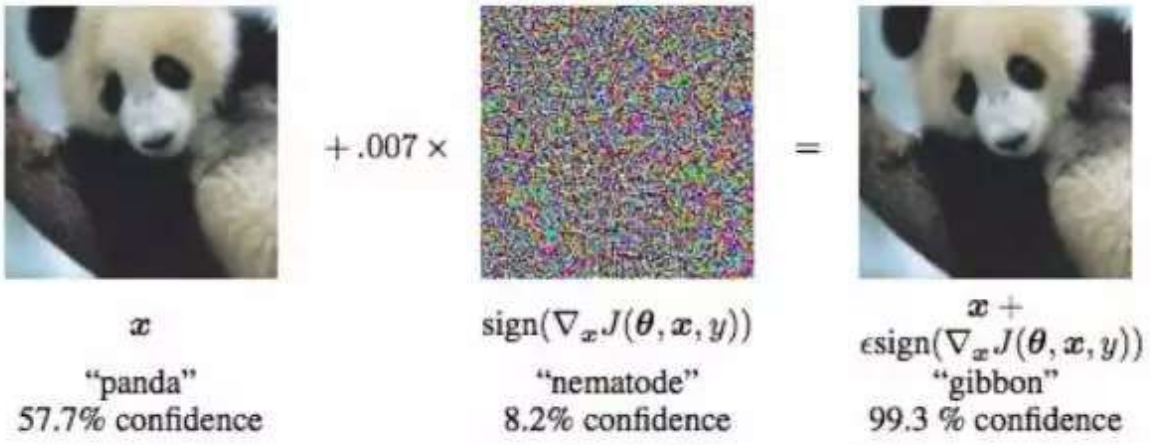
讲师

大数据安全未来会面临哪些挑战？

京安小妹

兜哥：

这是个好问题，也是个非常开放的问题。一是大数据相关软件的自身安全问题，比如ES，Hadoop、Kafka机器学习算法。之前已经出现过多起针对大数据软件的攻击，比如针对ES的删除数据以及加密敲诈、针对公网开放的Kafaka的数据截取等攻击。早期出现的就是污染机器学习样本，误导机器学习。这就是一个典型的针对机器学习算法的攻击：



左边是一个正常的熊猫，通过一定算法把它和噪声混合生成的图像，我们肉眼可以很轻松的辨别是熊猫，但是机器学习算法却识别为长臂猿了，这方面的例子很有意思。还有把停车牌识别为60公里限速的。二是就是大数据环境的用户隐私保护问题，这也是一个研究方向，尤其是新网安法以后，这几乎就是迫在眉睫的问题了。三就是如何使用大数据技术，如何在海量数据中挖掘APT攻击，这个是安全领域火了好几年的一个点，很多创业公司在都弄。四就是大数据导致大量用户行为数据、商业数据高度集中，这导致攻击目标更大，我们应该如何防范黑客针对大数据平台的攻击，这是一个很实际的问题。五是对大数据访问的合规使用也是很大挑战，和传统的数据库审计相比，大数据平台的审计更加麻烦。首先平台是开放的，协议是多样的，大数据平台自身也是快速发展，针对它的审计产品还远远没有跟上。如何发现开发、分析、管理人员窃取大数据平台的数据，这个是个复杂的问题。六是在有大数据之前，大量的商业分析，机器学习的数据集规模较小，尚可以通过人工梳理清洗，但是在大数据环境下如何防范黑客对攻击样本的污染（已经出现针对机器学习样本的攻击）也是个问题，大数据环境下，人工做数据清洗、过滤样本是个非常奢侈的事情。最后一个，也是我觉得很挑战的是：黑产利用搜集的海量的个人数据，分析处理后可以更加有效的进行钓鱼、电信欺诈以及APT攻击，我们更要加强对用户隐私数据的保护。这是我个人的一些理解，大数据安全本身就是个非常宽泛的话题。

讲师

大数据可能涉及到哪些具体的安全技术问题?

京安小妹

兜哥:

一是大数据相关软件的自身安全问题，刚才举了一些例子了；二是大数据中的数据挖掘技术，比如机器学习，这方面我写的相关的书和文章，已经比较多了，这块是目前安全领域比较热的；三是隐私保护相关的加密技术以及在加密环境下的数据分析检索问题，这是个老话题，加密环境下的数据处理技术在大数据环境下更加重要，尤其是检索和分析技术。：

讲师

安全行业能够利用大数据做些什么?

京安小妹

兜哥：

利益驱动的黑产在技术上始终保持着领先性，传统的基于规则、黑白名单的安全防范技术已经难以跟上黑产进步的速度。大数据或者说机器学习，一方面可以通过学习正常行为模型区分异常行为，发现深入的攻击行为；另外一方面可以训练机器掌握部分人工渗透的技术，提高我们主动发现安全漏洞的能力，就是传说中的智能渗透，目前已经有实现自动学习语法、自动生成攻击载荷的自动化渗透。

讲师

如何建立完整可用的安全大数据平台?

京安小妹

兜哥：

大数据平台自身的各类日志要另外保存与分析审计；
大数据平台的安全配置与软件版本管理。
这个其实内容很多，简单讲就这两条

讲师

有哪些好的安全数据来源或者安全大数据平台?

京安小妹

兜哥:

建议参考OpenSOC的架构，实时分析基于storm 离线、基于spark 数据持久化以及hdfs 实时检索、基于es数据高速公路、基于kafka底层计算，依赖cpu和gpu配合数据采集，依赖logtash flume bro等。

这里讲一下常用的数据来源：

内部数据：

安全设备产生的日志

网络流量

数据库日志

业务日志

服务器系统日志

应用服务器日志

外部数据：

接入威胁情报

特别强调一点全流量很重要，比accesslog丰富几条街。上了https以后，如果你证书部署在负载上，需要把流量镜像部署在负载后面。

讲师

白帽子提问:

1. 如果是全球idc，有好的方式收集日志吗?

   兜哥：跨境数据传输有要求，要参考当地法律了，有国家是禁止原始数据跨境的。技术方案就是本地搜集，统一搜集或者各个州分析完的数据汇总。

2. 学习样本如何清洗，因为样本来源本身就不准确（来自于waf或者正则匹配的结果）

   兜哥：深度学习之前特征工程本身就包括数据清洗和特征提取，工作量占60%以上，这个没啥捷径。

3. 用tfidf取特征，做学习样本是否可行？
4. 只取key1=abc&key2=456中的value值做特征，是否可行？

   兜哥：问题3和4要结合场景，一般基于文本的分类问题，tkidf可以配合词袋这些模型使用，比如垃圾邮件，恶意评论啥的识别。至于问题4，基于get参数的攻击检测可以这么做。

本期JSRC 安全小课堂到此结束。更多内容请期待下期安全小课堂如果还有你希望出现在安全小课堂内容暂时未出现，也欢迎留言告诉我们。

安全小课堂的往期内容开通了自助查询，回复“安全小课堂”或者点击阅读原文进行查看。

最后，广告时间，京东安全招人，安全开发、运营、风控、安全研究等多个职位虚位以待，招聘内容具体信息请扫描二维码了解。

简历请发送：cv-security@jd.com

微信公众号：jsrc_team

新浪官方微博：京东安全应急响应中心