基于大数据分析的
内部异常行为检测系统
白皮书

北京创元启安科技有限公司

2015 年 5 月

目　录

第1章 前言

大家都知道，安全是对抗，不可能完全防范，系统无非两种状态**已经被攻破的，或即将被攻破的**。做安全的思路应该从防止安全入侵这种不可能的任务转到了防止损失这一关键任务上，防范措施必不可少，但是基于预警、响应的时间差更关键。

从未来看，企业安全将会发生一个大的转变：即以“信息和人”为中心的安全策略，结合全面的内部监控和安全情报共享。全方位的内部监控和安全情报是保护信息安全的主要手段。实际的安全工作中，很多用户知道要严防死守外部侵袭，但往往忽略了内部威胁对系统造成的破坏，实际上大多数安全威胁都来自内部。内部的异常行为主要来源：

1.1 APT 攻击的内部潜行

APT 攻击者企图隐藏一切，但当攻击发生时，**流量和行为无论如何伪装也会展现出来。**

外部的攻击要真正达到目的必须经过“内部潜行”才能接触到敏感数据乃至盗取数据，从 APT 攻击的攻击链可以看出 APT 攻击需要经过好几个环节：

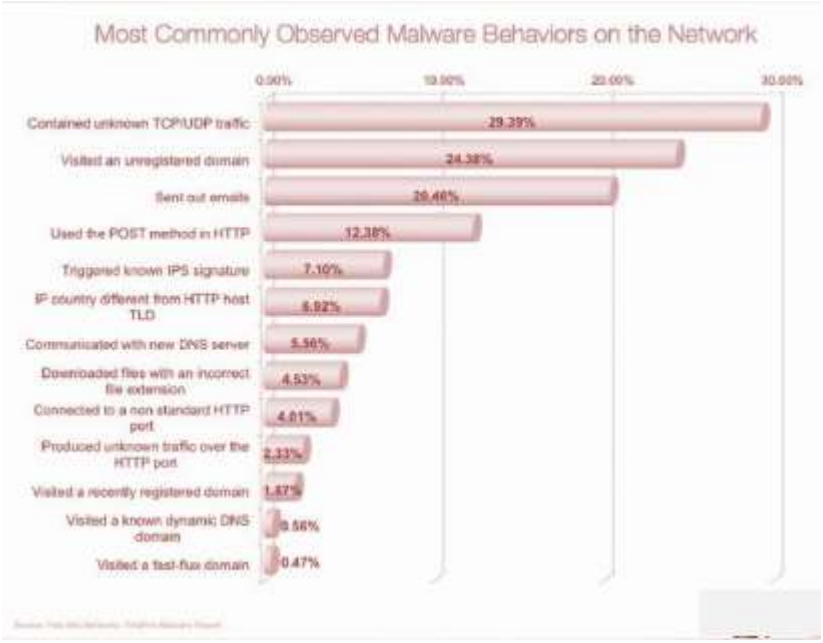
- Reconnaissance（侦查，充分的社会工程学了解目标）；
- Weaponization（定向的攻击工具的制作。常见的工具交付形态是带有恶意代码的 pdf 文件或 office 文件）；
- Delivery 把攻击工具输送到目标系统上，APT 攻击者最常用这三种来传送攻击工具，包括邮件的附件、网站（挂马）、USB 等移动存储；
- Exploitation 攻击代码在目标系统触发，利用目标系统的应用或操作系统漏洞控制目标；
- Installation 远程控制程序（特马）的安装，使得攻击者可以长期潜伏在目标系统中；

- ✧ Command and Control (C2):被攻破的主机一般会与互联网控制器服务器建立一个 C2 信道，即与 C2 服务器建立连接；
- ✧ Actions on Objectives:经过前面六个过程攻击者后面主要的行为包括：

1、偷取目标系统的信息，破坏信息的完整性及可用性等。

2、进一步以控制的机器为跳转攻击其它机器，扩大战果。

外部攻击者发起 APT 攻击，其中的部分环节 Delivery、Exploitation、Installation、Command and Control (C2)、Actions on Objectives 都需要通过“内部潜行”才能接触到敏感数据达到盗取或破坏的目的。



1.2 来自于内部的威胁

企业的安全管理危险份子有时可能就是坐在你身旁的普通同事。在恶意内部人员的情况中，这些员工很愤怒或者不满，他们可能准备离职或者已经被解雇，并仍然可以访问企业系统。这些攻击者非常危险，因为他们已经知道如何访问网络，并能轻松地访问大量信息，而不需要费力。

第二种类型的内部威胁源于偶然的数据上传、未能安全地处理文件，以及具有意想不到后果的复杂的交互。无论它是如何发生，当员工意外暴露数据时，都可能发生内部攻击。

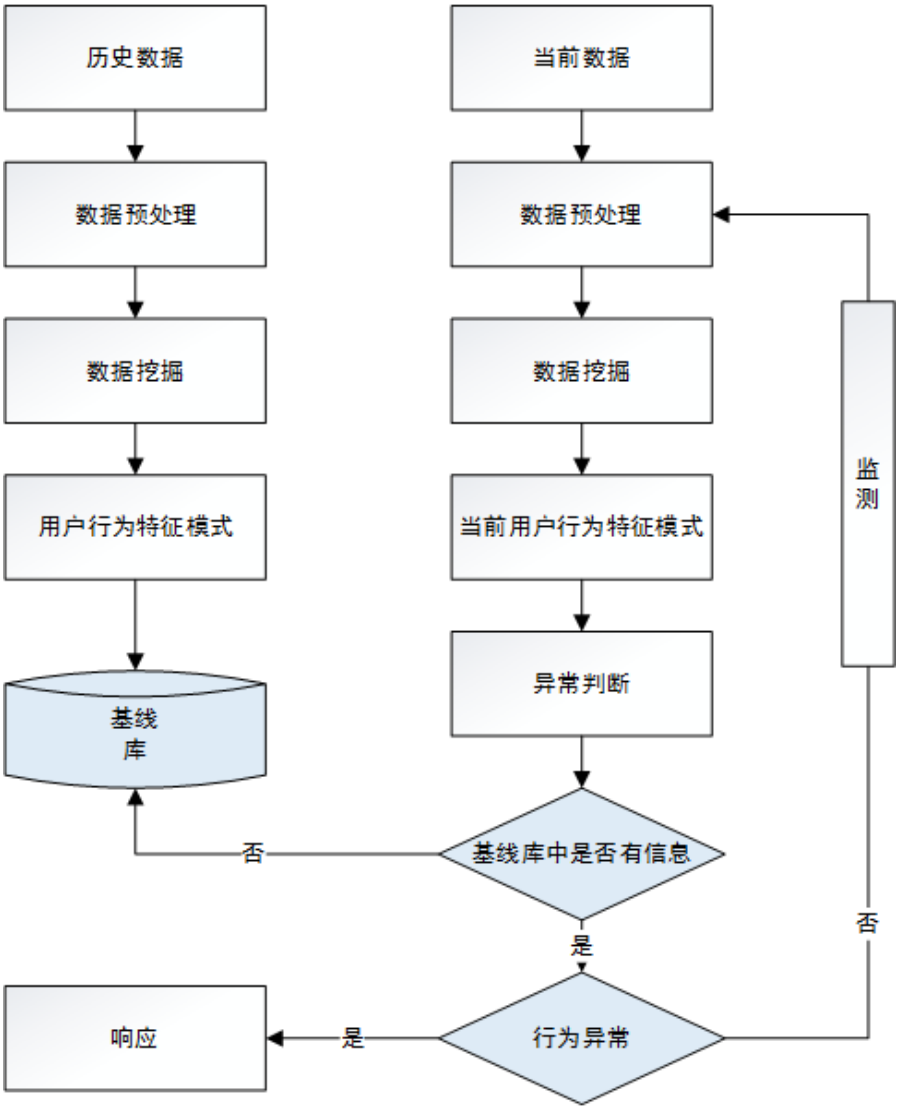
内部攻击也可能通过合作伙伴或者第三方的形式，这些第三方被授予访问权限，并可能意外暴露数据。我们看到过多少因为笔记本、U 盘或文件处理不当(其中可能包含数以千计敏感数据)而造成的数据泄露事故?这些泄露事故并不是恶意的内部人士，而是没有受过教育和轻率的内部人员给企业和企业客户造成的损害。

我们并不能清楚的了解自己的网络。

在实际应用中，网络的流量非常多，且非常大，手上的工具要么只针对 IP 层，要么只针对攻击行为，仅有的几个做应用层分析的产品，又要指定具体要监控哪个应用甚至是哪个链接，没有称手的工具，想让管理人员立即说清楚网络里面有什么，有没有危险存在，多数网管都会觉得这样的问题很难回答。对于安全人员来说，我们需要一个得力的工具，告诉我们网络里面到底有什么！

第2章　实现原理

异常行为分析技术通过对流经设备的流量进行连续、实时监控来分析流量信息，建立行为基线，利用统计分析、关联分析和机器学习等多种技术手段来检测流量和用户/应用行为中的异常模式，以发现异常行为。异常可以与同类对象做比较而得出，也可以与历史数据做比较而得出。



第3章 异常行为检测安全情境

3.1 关键资产的识别

通过分析内部网络流量，采用自动化学习的方式识别组织内的网络资产，同时将组织内的设备或资产显示在一张逻辑图上，很便利的看到设备之间的互联关系；

允许用户使用资产管理的方式对设备进行属性的定义，以便对核心资产生成的流量进行识别。比如可根据资产上存放的数据的重要性（数据的分类分级很重要），考虑资产的使用者比如关键员工和高管的设备等定义资产的属性。

3.2 行为基线的建立

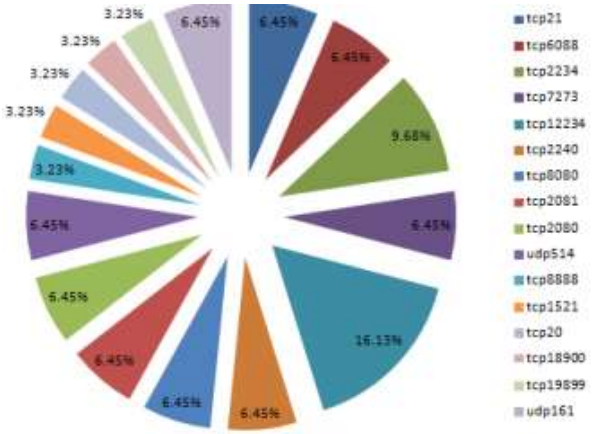
该系统专为用户掌握和管理信息系统运行情况和精细化安全管理而设计。通过旁路抓包获得网络中的流量数据，经过重新组包、匹配、识别，提取信息系统的数据流运行状态，帮助用户自动识别、梳理应用，提取网络中的业务数据流，，建立网络行为基线。

3.2.1 指纹识别

系统能够支持应用的自动发现、识别、提取等功能，通过对应用的分析，对**业务基本特征指纹**可自动识别，包括：

- 资产识别：资产的 IP 等信息。
- 应用识别：服务端口、应用系统、应用特征

以下显示的应用的端口指纹：

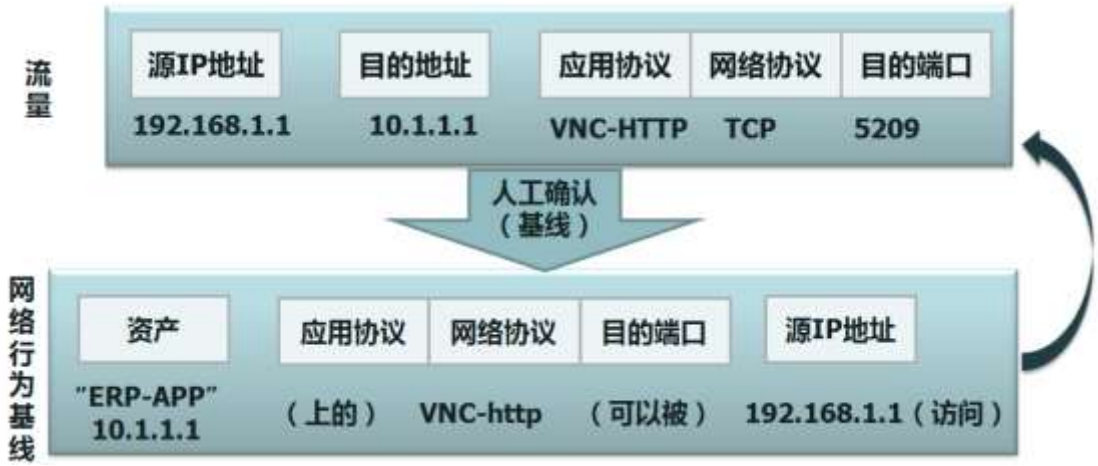


3.2.2 行为基线建立

快速生成行为基线

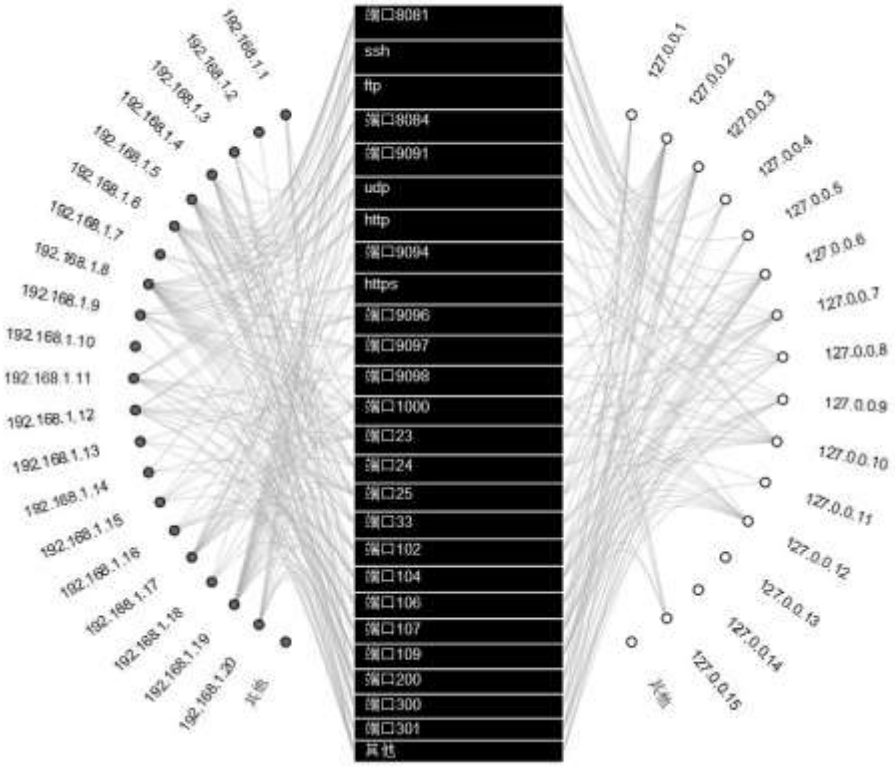
网络行为基线是指从真实存在的应用访问关系列表中，经过自动生成和人工调整建立符合应用逻辑的访问关系表，称之为业务应用网络行为基线，简称行为基线。它是应用系统内部运行、外部互连是否安全的重要依据。网络行为基线由：源地址、目的地址、目的端口、协议等要素组成。

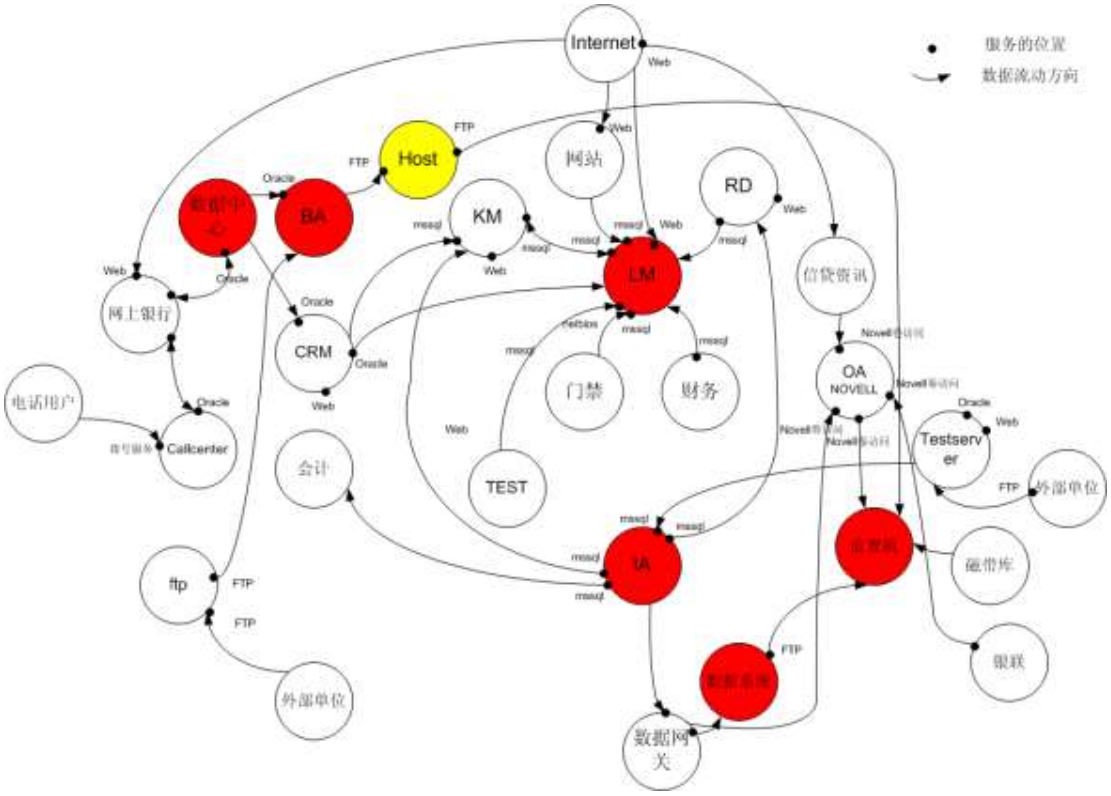
系统可以将初期学习的数据流量，人工确认后自动提取为网络行为基线，成为被认可为业务数据流，网络行为基线具体形成过程大致为几个步骤，第一步由设备解析流量，识别流量的源、目的及应用协议（部分）、网络协议和目的端口，第二步由人工方式确认流量合法性，第三步根据确认的结果形成网络基线，为识别异常行为做准备。这样可以帮助用户快速进入业务流监控管理状态，可以在以后的工作中再逐渐对业务网络行为基线进行动态优化。



业务流可视化

异常行为分析系统分析后生成可视化互连关系图，以“业务流”为全新的维度视角， 针对在运行系统和网络， 建立基于业务数据流的可视化系统模型。它脱离了传统基于物理拓扑和数据特征为基础的安全防护模式， 最大限度地反应了用户业务流和安全管理之间的关系， 帮助用户建立起基于业务流及可信资产的白名单机制。让企业网络安全管理者清晰、明确地看到企业与外部的关系，即“谁在访问我，我在访问谁”；

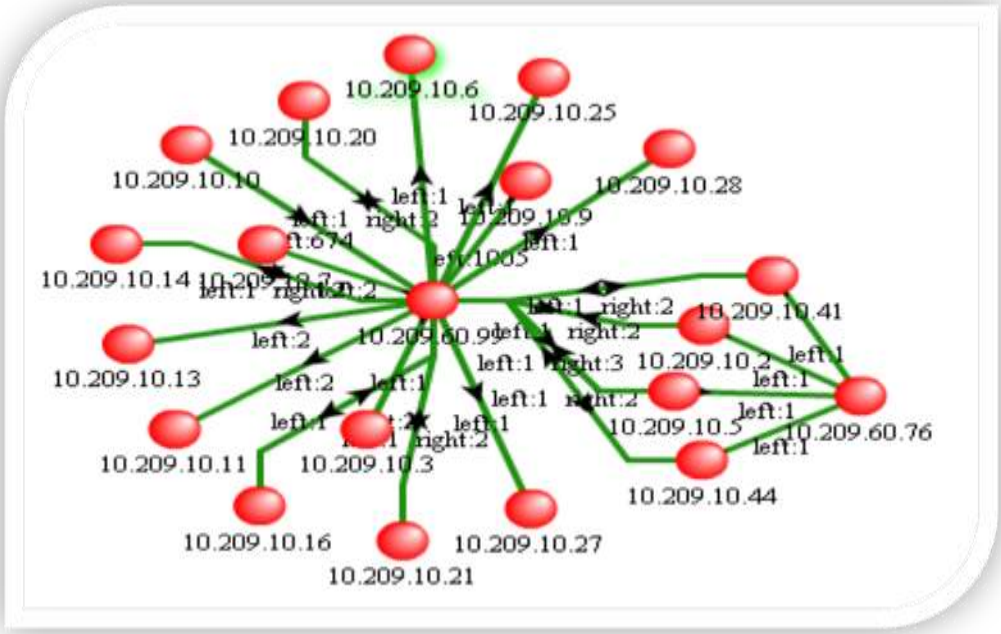




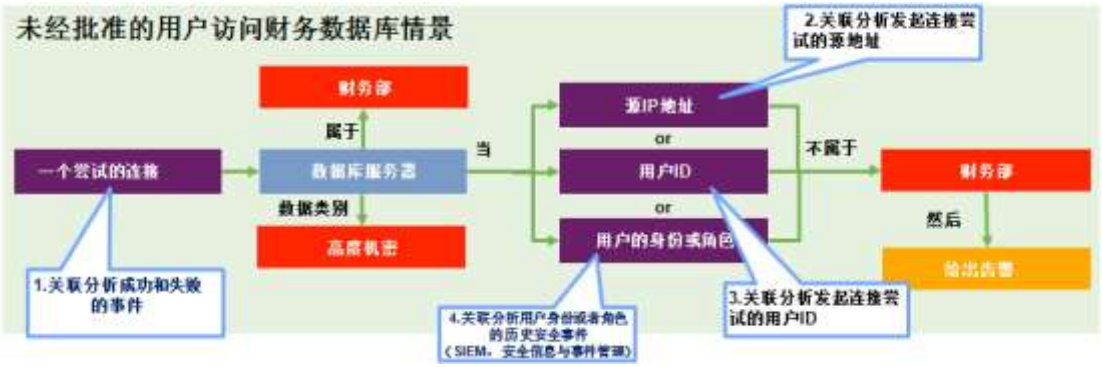
3.2.3 敏感数据的访问基线

敏感数据窃取是用户面临的最严重安全问题，本系统可针对梳理敏感数据库服务器进行专门的访问基线梳理，通过网络行为的检测发现数据盗用等行为。敏感数据的访问基线包含的信息主要有：

- ✧ 用户对敏感数据服务器的的访问行为频次、时间、访问源等：如 A 部门用户每天平均访问 220 次数据服务器。
- ✧ 外发敏感数据行为：如外发用户、设备、时间、频率和目地等
- ✧ 内部业务系统和服务器敏感数据访问历史等
- ✧ 用户或设备频繁外发加密文件数据



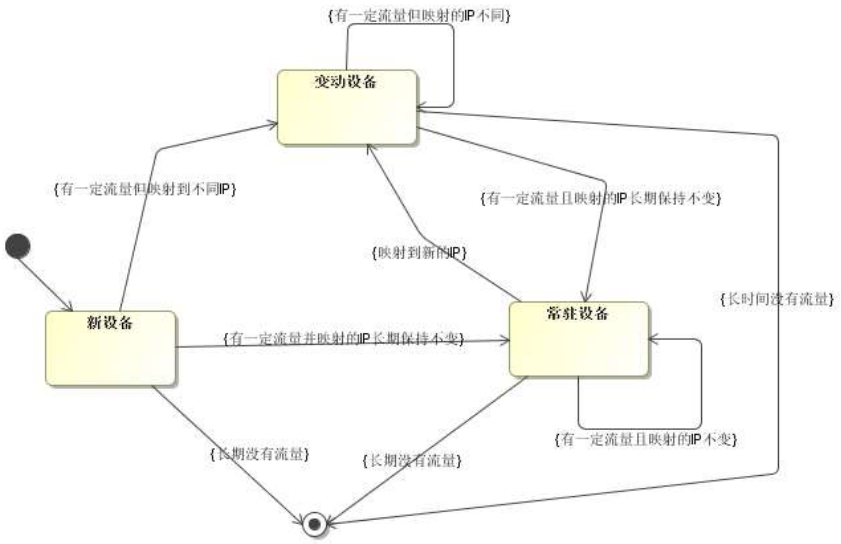
未授权的敏感数据访问逻辑模型如下：



3.3 异常访问行为的可视化

3.3.1 异常情境 1：未知资产的发现

系统通过对网络互联关系的全面监控， 捕捉每一条互联关系并生成纪录，对任何非法在线的入网设备哪怕只要在线进行一次连接动作， 就一定能够发现， 并对非法设备的连接行为进行取证。同时对系统中有网络信息变动的设备也可以第一时间发现。



3.3.2 异常情境 2：非法或长期未使用的应用发现

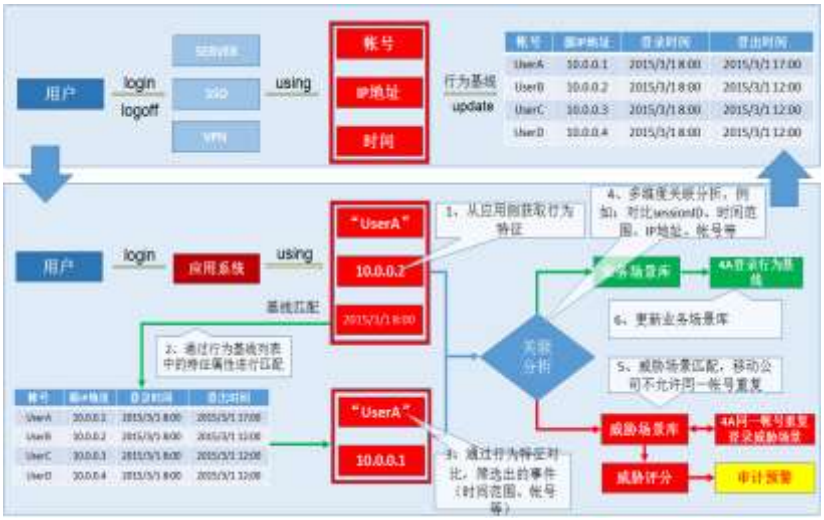
建立应用可用性基线（流量、RTT、连接数），监控应用可用性，对内网用户和应用的行为模式进行匹配分析，通过该系统可以：

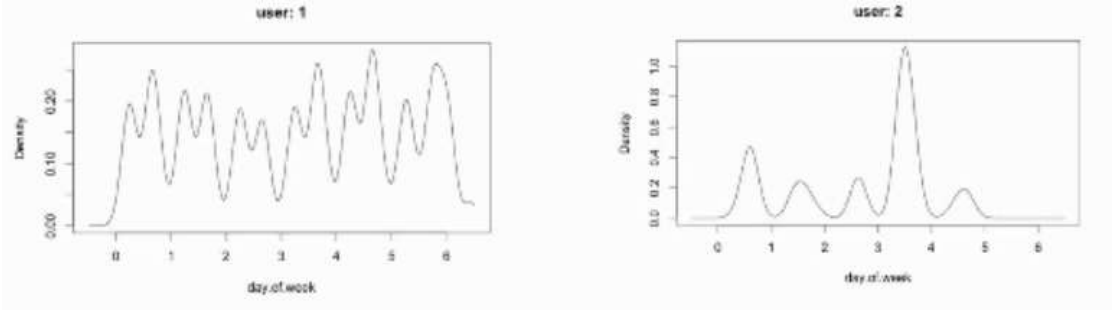
- 发现未备案应用上线
- 发现长时间未使用应用
- 发现长时间未使用边界访问控制策略
- 发现未经审批新上线应用
- 发现长时间未使用的应用

3.3.3 异常情境 3：用户的异常登录

重点分析用户认证登录异常行为：如异常时间、异常 IP、多 IP 登录、非个人用户帐号登录、频繁登录失败等。登录异常行为同时也包括共享账户行为，比如一个账号短时间更换 IP 登陆，一个 IP 登陆了多个账号等。

斯诺登就是一典型的 insider threats 案例，按照安全设计理念，他是能被发现的，比如斯诺登经常要同事的帐号访问系统，比如斯诺登可能比一般员工更多的访问了核心服务器，比如斯诺登可能短时间内打包了很多的敏感数据等，这些行为都可以通过异常行为分析来发现。

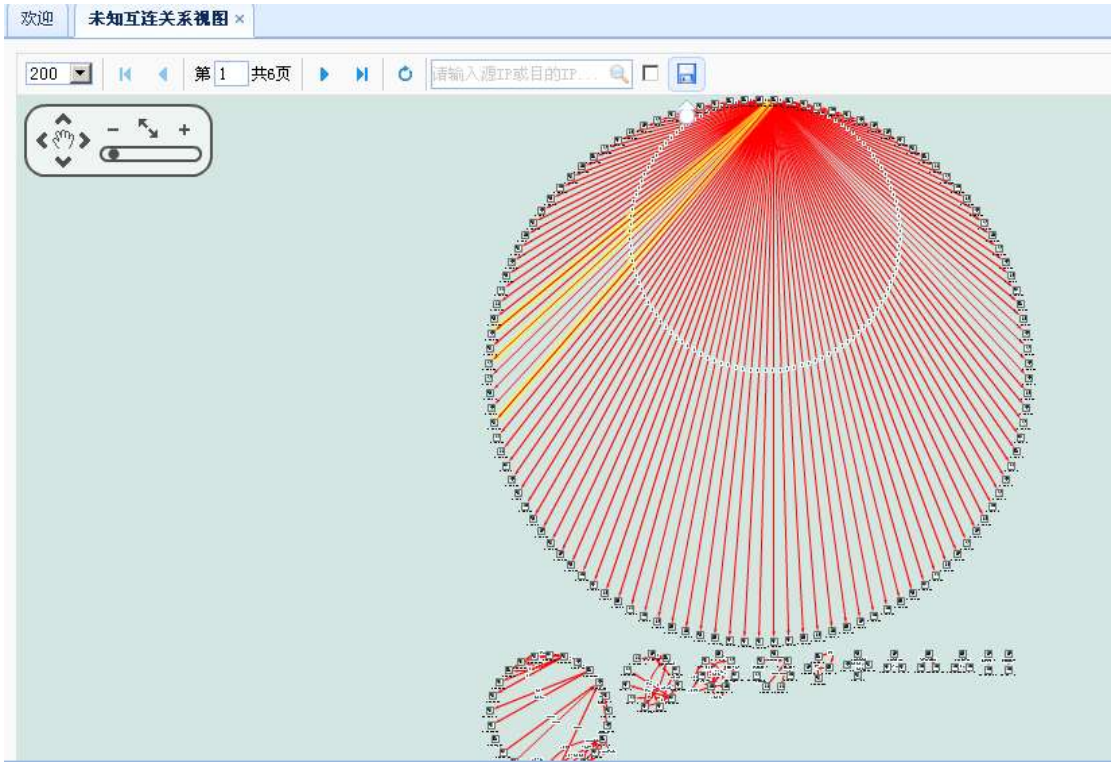




场景类型	场景名称	场景说明	审计内容	预警级别	审计周期	备注
绕行4A	常规绕行4A	此场景审计通过合规手段，绕行4A认证功能的行为	异常登录审计用户直接通过核心域外IP，访问核心域内资源。	高级	实时	无
绕行4A	非常规绕行4A行为	此场景审计直接绕过4A堡垒机，访问核心域的行为	用户也通过正常的方式登录4A平台，通过已授权的主机资源（如windows）访问其他已授权或未授权资源。	高级	实时	非常规绕行行为无法溯源
绕行4A	常规工程类绕行4A系统	审计通过进入机房等行为，绕行系统	因为工程割接、系统故障需进入机房，直接服务器console、ILO、交换机等访问行为。	高级	实时	无
绕行4A	绕行后高频访问	此场景审计正常登陆之后，对核心域其他资产进行高频访问	通过已授权的主机资源（如windows）尝试访问未授权的主机资源。行为疑似扫描、暴力破解密码等攻击行为。	高级	日	无
异常帐号登录	同一帐号同时登录4A	帐号共享异常行为	自然人通过相同主帐号，登录4A平台。	中	日	无
异常帐号登录	同一帐号同时登录4A(wifi)	终端地址、主帐号共享异常行为	自然人通过wifi共享，并共享同一主帐号，登录4A平台。	中	日	无
异常帐号登录	异常帐号登录(主帐号)	主帐号登录异常行为	单一主帐号多IP登录、非工作时间段登录、异地登录、主场景状态异常、均为疑似异常行为	低	日	无

3.3.4 异常情境 4：异常的互联关系

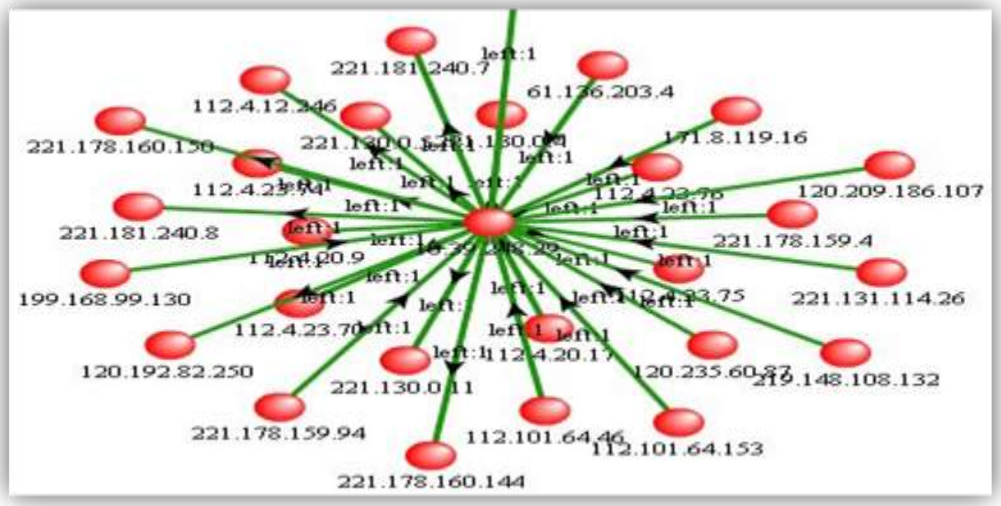
对于不符合白名单和灰名单的互连关系和流量，系统会自动生成未知数据流，并对未知数据流进行识别、匹配，从中提取可供判断的信息，如：单点对多点的快速连接，系统会识别为网络扫描，非正常时段的数据连接，系统会视为异常行为，多点对单点的大流量连接，系统会识别为非法应用等。用户对未知数据流识别、确认、处理后，可将未知数据流自动生产报警或提取到白名单。



3.3.5 异常情境 5：非法外联或跳板攻击

对于网络设备被控制、web 服务器被控制作为跳板攻击，都会存在对外的异常连接，比如对外部发起扫描攻击，或者直接访问 c&c 服务器，如果出现这些行为都代表内部用户确实出了问题，该访问都会通过颜色标识进行告警。

同时系统可以有效监控网络内的主机是否存在自动外联网络的行为，可实现监控系统及应用软件是否存在内置软件对外连接机制和行为，用户定位网络中哪些主机产生较多流量，并频繁与外进行互连，杜绝企业信息涉露的可能行。



3.3.6 异常情境 6：数据盗取行为

大多数用户的日常行为是可预测的，每天的日常活动都差不多。恶意的内部人员在偷盗数据或搞破坏前一定有异常的行为。对于可疑的员工连接关键资产一定要引起足够重视。这种异常通常未必是一个确定的违规行为，但它可以作为重要的调查信息；

通过检测一个主机的流入和流出数据流可以发现内部人员恶意的数据偷取行为。数据科学是非常适合这种类型的分析，该系统采用数据科学技术自动化的检测网络内部用户的敏感数据的访问、偷取等行为，机器学习能主动发现在堆积如山的数据中恶意的内部人员。大部分内部人员可能通过物理介质比如 USB 设备偷取数据，另外常见的是通过网络来来偷取数据，后一种可以通过分析一个主机的流入和流出数据流，可以通过检测网络流量中是否存在隐蔽通道行为、加密通道行为、TOR 活动、内部频繁潜行（staged hop）等，如果存在这些情况基本可以判定发生了内部的数据盗取威胁，通过关联分析，可以发现内部人员数据偷取的整体行为视图。

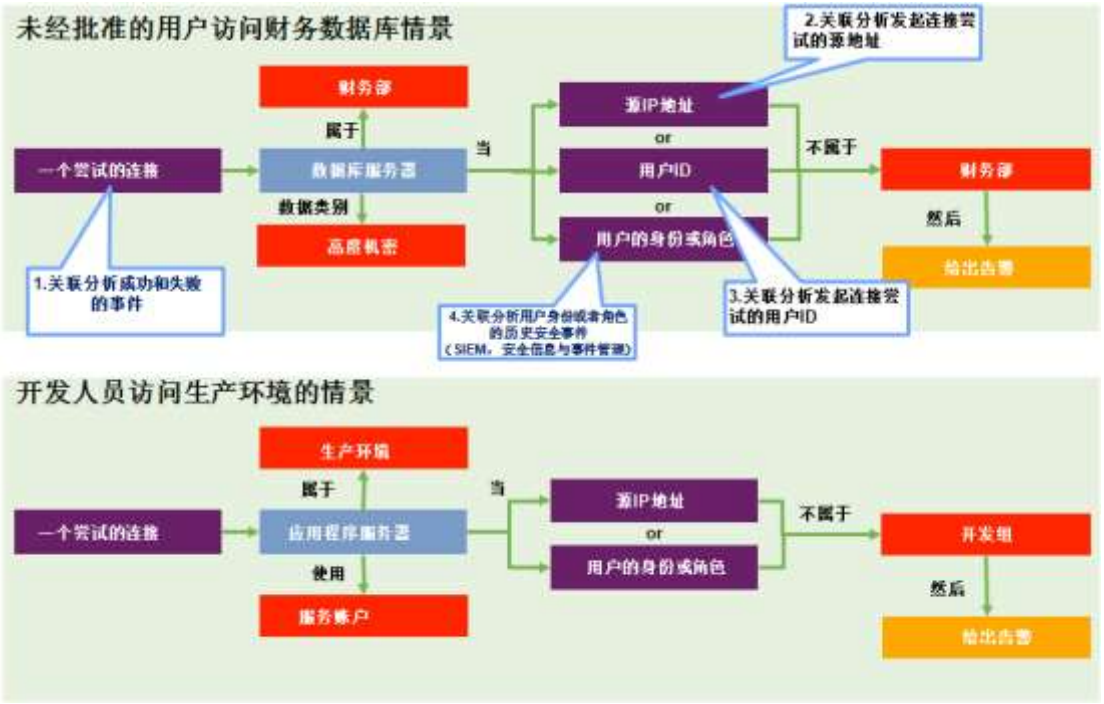
3.3.7 异常情境 7：针对企业内部的安全攻击

内部用户之间的攻击对传统防御系统很难发现，同时 APT 攻击进入内网后，攻击者一定会进行内部的横向潜行，去攻击内部的相关服务器，系统也可以识别内部用户执行的端口扫描、暗网扫描、暴力破解等恶意行为。同时可以将异常的行为关联到组织的关键资产，提

供一个内部威胁的量化指标。系统针对不同的恶意行为通过不同的告警颜色进行标识。

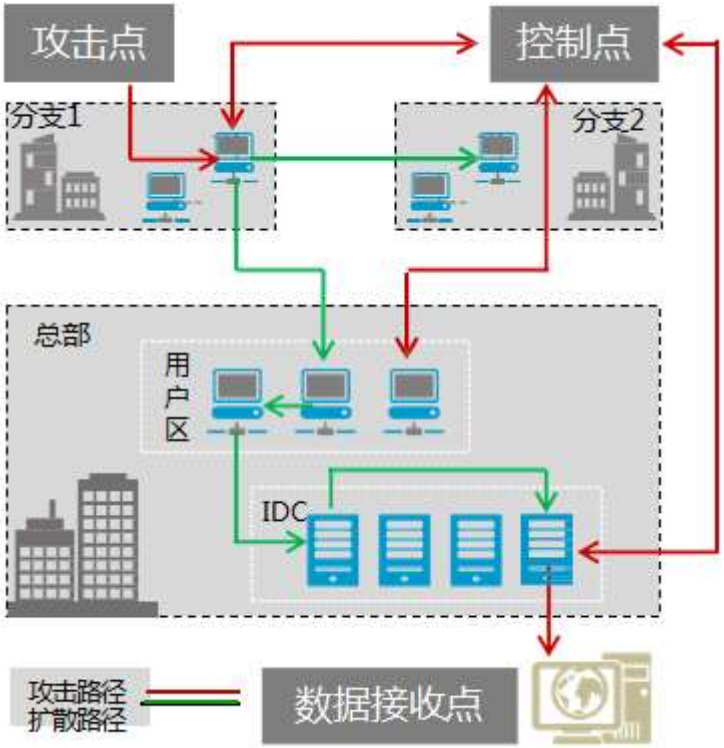
3.3.8 异常情境 8：关键资产异常访问

对关键资产的访问需要进行严格的监控，比如该系统通过对数据库的敏感表访问过程中，对其访问路径实现上下文分析，是否存在非经过堡垒访问数据库敏感表的情况，此人是否存在越权访问情况。



3.3.9 异常情境 9：外部攻击的横向运动

外部的攻击要真正达到目的必须经过“内部潜行”才能接触到敏感数据乃至盗取数据，正常的用户不会在内网不会进行横向的频繁访问，但是攻击者进入到内部后，需要找到目标，其行为明显和正常的用户不同，内部的横向访问其实是发现安全攻击的最好的线索之一。



3.3.10 异常情境 10：运维违规行为

该系统支持分析行为模式并形成报警，包括：行为模式类型，发生此行为模式的五元组条件（源 IP、源端口、目的 IP、目的端口、协议），发生时间等信息。

不同行为模式有不同的含义，例如：

- 违规运维指不允许使用的某些服务，例如：使用3389,23端口的TCP连接；
- 疑似新增应用指应用服务域中不能自动识别的服务；
- 长时间未使用访问策略指在一定时间段内未被数据驱动的访问策略；
- 长时间未使用应用指在在一定时间段内未被数据驱动的应用；
- 垃圾流量指只有去包没有回包的互访关系；

3.4 异常行为的溯源与取证

异常行为分析系统能够完整、准确、快速的记录海量 IT 数据。对于可疑的或明确确定的异常行为，系统会记录原始的网络数据包，通可以快速的完成事件审计、原始数据取证等工作。可以通过告警前后时间段内业务系统访问的全量数据，通过序列分析、聚合统计、关

联分析等多种分析手段，直接判定疑似异常行为是否合法，或分析出用户的可疑行为，找出新的违规行为特征。

以下为系统依据时间序列回溯的一起安全违规事件过程。

3.5 结合威胁情报的自动化检测

通过威胁情报自动化的共享，系统可结合外部的威胁情报，进行异常行为的检测，比如依靠外部的 c&c 恶意域名列表，对所有的用户对外访问的 DNS 进行监测，一般内部用户不会

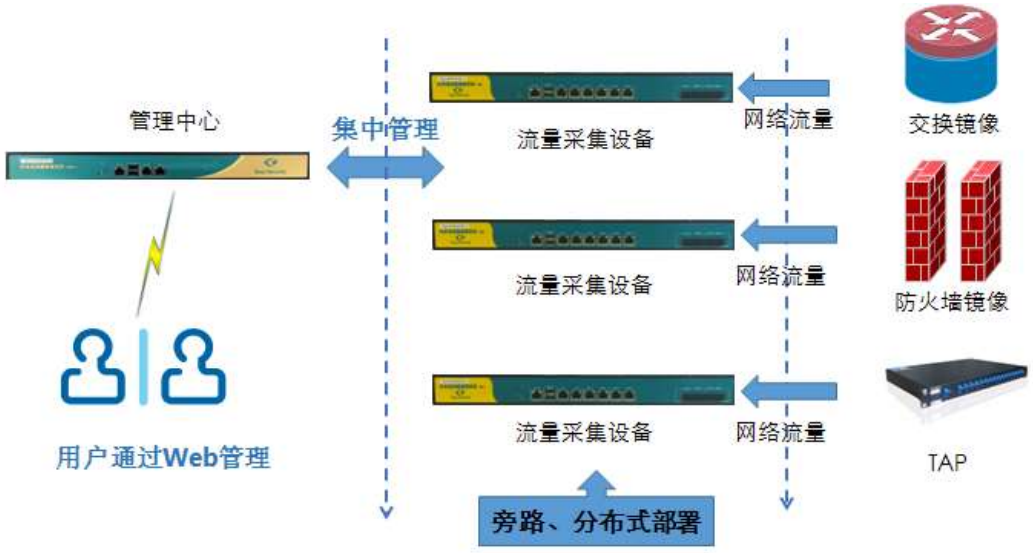
直接访问 c&c 服务器，如果出现这些行为都代表内部用户确实出了问题，该访问都会通过颜色标识进行告警。

第4章　异常行为检测分析平台

4.1　基于流量的采集分析系统

基于流量的采集分析系统主要是对网络实际流量进行采集、处理、存储及分析，实现安全可视化。该系统为“流量采集设备”和“管理中心”两级架构。流量采集设备用于采集网络中的流量，并对流量进行一定程度的分析及处理后传送给管理中心，由管理中心形成用户使用界面。用户可通过浏览器登录管理界面，方便地进行配置与监控。

异常行为分析系统通过白名单技术，定义设备之间的互连关系是否合法。从而实现网络中是否存在非法互连的安全检测目标。



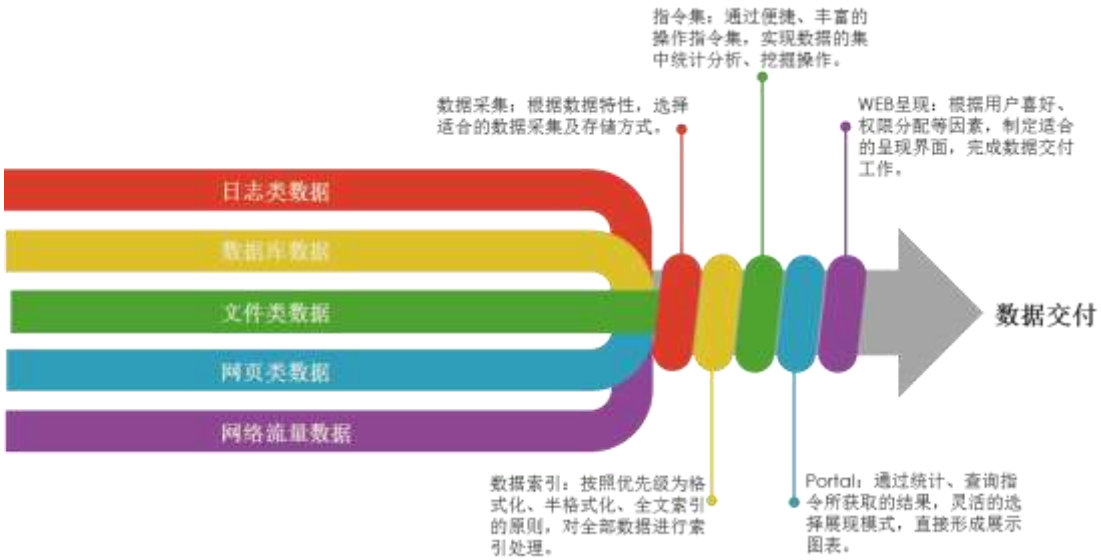
异常行为分析系统为“流量采集设备”（也称为探针）与“管理中心”两级架构；

流量采集设备采用旁路、分布式部署；

管理中心采用集中式管理多台流量采集设备；

用户通过浏览器对管理中心进行配置与管理。

4.2 基于大数据的检索分析平台



图：功能导览

基于大数据的检索分析平台由数据采集模块、索引模块、指令集、可视化模块、WEB 交付模块五部分组成。

数据采集模块的主要功能是根据数据不同特性，选择适合的采集探针及存储方式。目前支持的数据采集种类包括日志类数据、数据库类数据、文件类数据、网页类数据、网络流量类数据。可支持 100PB 规模的海量数据存储。

数据索引技术是基于大数据的检索分析平台的核心技术，在数据采集的同时按照优先级为格式化、半格式化、全文索引的原则，对全部数据进行索引处理，同时将原始数据进行压缩存储，每秒数亿级别的检索速度为后期数据处理功能奠定了坚实的基础。

指令集是基于大数据的检索分析平台数据便捷化处理的具体表现，我们定义了一套强大的数据查询、统计语句，用户可以通过这些语句实现各种复杂的查询、统计、分析操作，并将这些操作所获取的指令集直接交付给可视化呈现系统形成图表。这样做的优势是我们可以随时修改图表的内容、形式、维度，而不用估计这些变动对数据模型带来的影响，这将极大的提高用户的使用感知并大幅度降低定制化开发的周期。

可视化系统分为组件与 dashboard 两部分，通过定义组件我们可以将通过指令获取的数据以图表的形式进行展现，再通过 dashboard 管理将多个组件拼叠成用户感兴趣的页面。

WEB 交付模块主要完成权限分配、dashboard 布局等功能。通过一系列灵活的配置，可以快速的完成用户交付工作。

4.3 蜜罐采集分析系统

蜜罐技术本质上是一种对攻击者进行欺骗的技术，通过布置一些作为诱饵的主机、网络服务以及信息诱使攻击者对他们进行攻击，减少对实际系统所造成的安全威胁，更重要的是蜜罐技术可以对攻击行为进行监控和分析，了解攻击者所使用的攻击工具和攻击方法，推测攻击者的意图和动机，从而能够让防御者清晰地了解他们所面对的安全威胁。

蜜罐好比是情报收集系统。蜜罐好像是故意让人攻击的目标，引诱黑客前来攻击。所以攻击者入侵后，你就可以知道他是如何得逞的，随时了解针对服务器发动的最新的攻击和漏洞。还可以通过窃听黑客之间的联系，收集黑客所用的种种工具，并且掌握他们的社交网络。

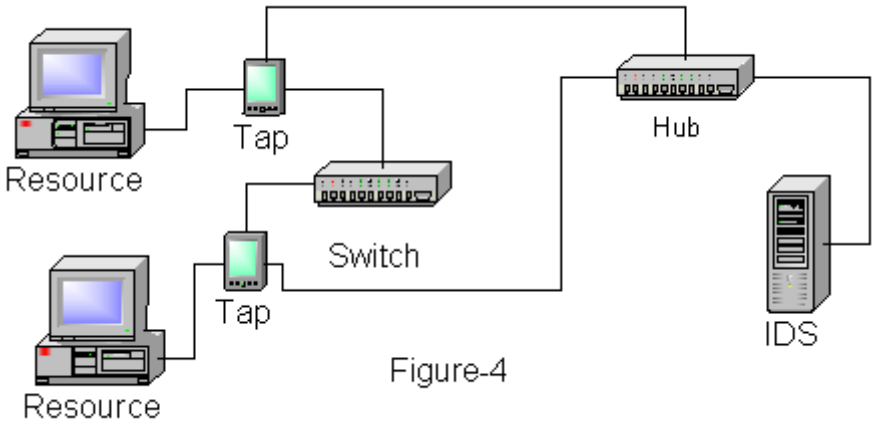
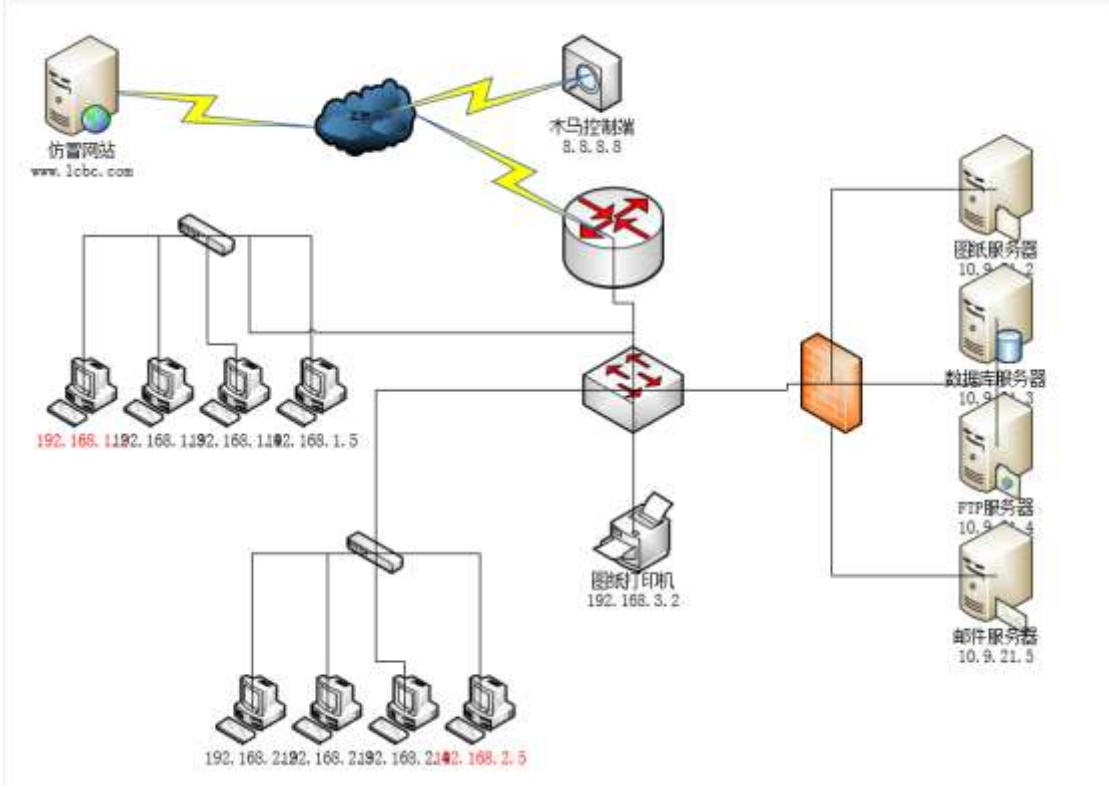


Figure-4

第5章　案例成果



攻击过程：

1、 PC 192.168.1.2 收到钓鱼邮件进而访问仿冒工商银行网站 www.1cbc.com，被利用 IE 浏览器漏洞植入木马程序

2、 PC 192.168.1.2 与木马控制端 8.8.8.8 通过 DNS 协议通讯

3、 PC 192.168.1.2 通过口令批量暴力破解，成功登录业务区的工作站 192.168.2.5

4、 工作站 192.168.2.5 攻击图纸打印机 192.168.3.2，并在打印机中植入后门程序

5、 通过图纸打印机 192.168.3.2 利用缓冲区溢出攻击获得图纸服务器 10.9.21.2 的控制权限

6、 图纸最终通过以上被打开的通道盗走

发现过程：

1、 部署分析平台后，通过一个月的自动学习和人工确认，建立了内部系统的资产白名单、网络流向、网络流量基线

2、（异常触发点）图纸打印机 192.168.3.2 主动连接了图纸服务器 10.9.21.2 的 445 端口，该连接不在基线中

3、 通过平台关联查询，还原了以上涉及资产的连接关系全景图，确认了从图纸服务器 10.9.21.2 已经有通向互联网的通道。