

浅谈大数据检索平台



分享者：dark3r@SAINTSEC



图个里安，有子 → → →

PART 01 为何需要它

通用漏洞、漏洞情报……

PART 02 怎么去创造它

平台框架、数据爬取、索引……

PART 03 后续 · 新思路

网络那么大，我要去看看

1

PART 01

“它”是个什么东西？

一个能够协助我们进行应用指纹提取、通用漏洞挖掘、日常辅助使用等等……



为什么需要它？

随着现在SRC平台的愈加增多，白帽子疯狂的刷榜，这个时候，我们需要为团队成员武装一些挖洞辅助装备。

它能够做什么？

例如：在向CNVD,CNNVD提交漏洞的时候能够提供通用漏洞案例查找，能够对企业资产进行汇聚等等……

对比与现有的大数据检索平台(zoomeye、shodan、fofa)，它的存在有什么意义？

不管是zoomeye、shodan、fofa，这些搜索某些数据时候，其实都是存在一定的限制。而我需要的是一个对我能够完全开放的平台。能够不阻拦我对数据的向往。

造轮子现在是一个很火热的词，具体我们不去深究，而我的个人见解是如果轮子能够造出自己的特色，那也算是一种创新，同一个轮子，别人造的是单车轮，而我造的却是宝马轮。

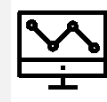


2

PART 02

我要创造它

从设计框架到数据爬取，从爬取到可视化分析。
一步一步往前走。



数据行星

大数据中心



The diagram features a central 'Big Data Center' (大数据中心) surrounded by four data collection points, each represented by a black dot and a line. These points are: 'Global IP Collection' (全球IP收集) at the bottom-left, 'Global Domain Collection' (全球域名收集) at the bottom-right, 'Fingerprint Rule Library' (指纹规则库) at the top-right, and 'Historical Vulnerability Library' (历史漏洞库) at the top-right. The entire system is titled 'Data Planet' (数据行星) at the top. The background consists of several concentric, slightly tilted ellipses.

全球IP收集

争对全球IP及IP所开放的端口和
协议版本。

全球域名收集

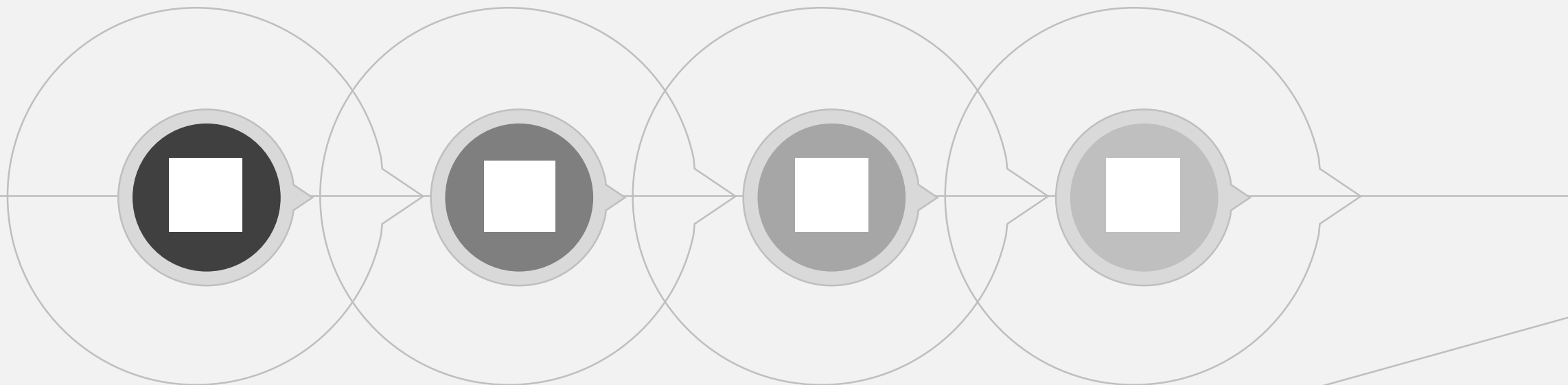
争对全球约3.4亿域名进行爬取

指纹规则库

对数据进行指纹规则建立。

历史漏洞库

争对已知的漏洞进行匹配，提供
思路学习。



准备阶段

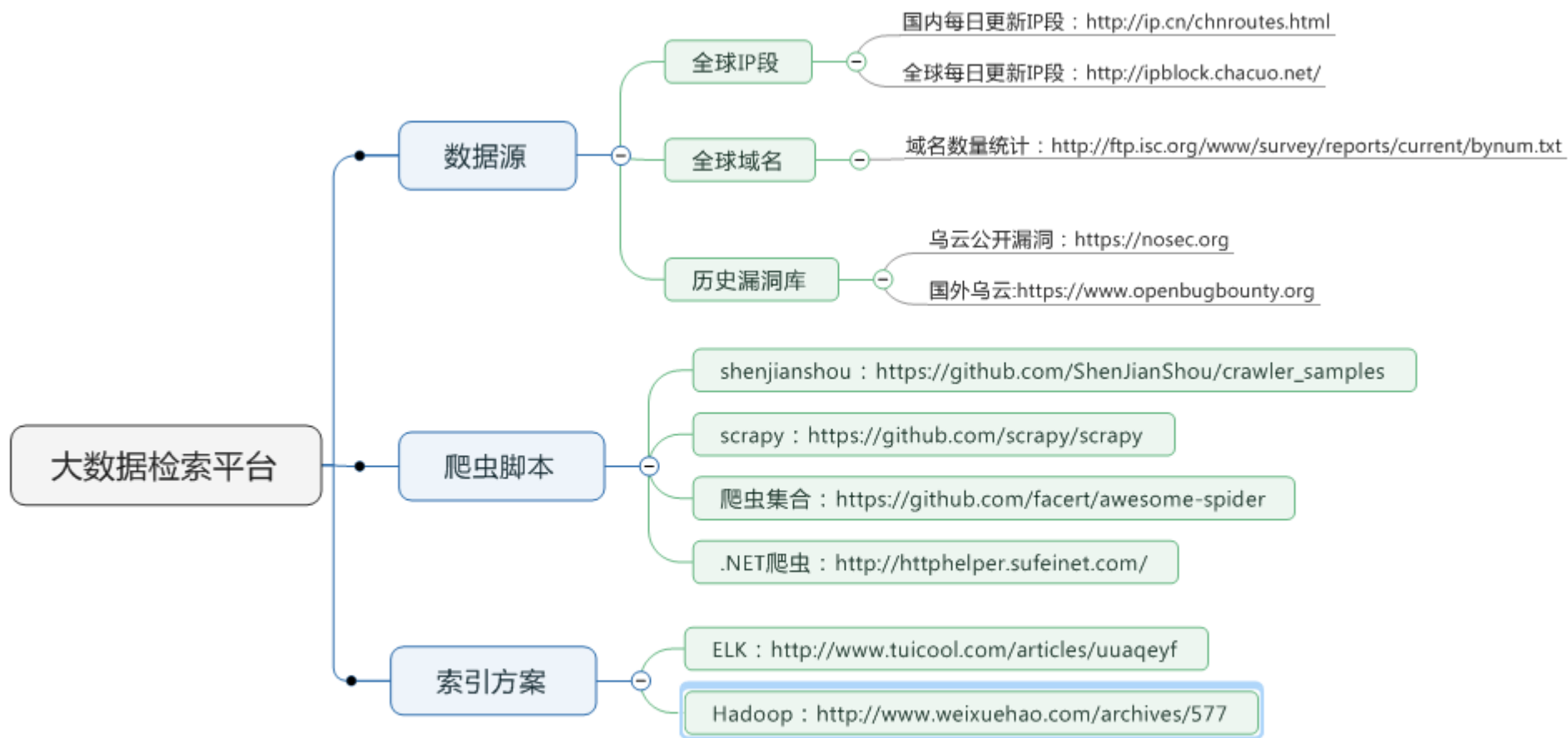
设计数据库结构，收集数据源，可以从Shodan,历史DNS记录等等进行下手。

工作阶段

编写爬虫工具，对数据进行爬取入库。
接着对端口扫描工具优化扫描规则。

后续阶段

对数据进行生成索引，争对方案进行
选择后续阶段的数据变化持续监控。



名	类型
id	bigint
title	text
host	text
ip	text
protocol	text
port	int
domain	text
os	text
spider_time	date
app_server	text
header	text
body	text
geo_country_code	text
geo_country_name	text
geo_region	text
geo_city	text
geo_postal_code	text
geo_lat	text
geo_lon	text
geo_continent_code	text

nsone.net

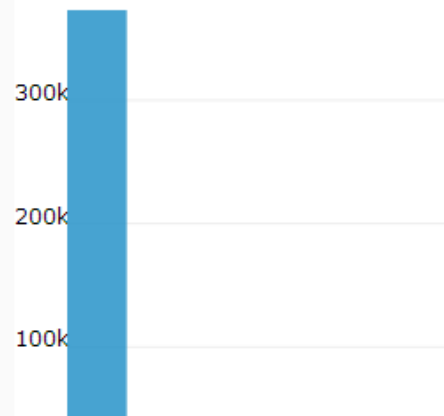
Overview

Graph	Data
-------	------

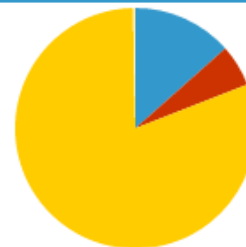
NSONE.NET

Name server History

Graph Data



Monday



New Domains
■ (1,282)

Transferred In
■ (7,680)

Transferred Out
■ (31)

Deleted Domains
■ (535)

Newly Registered Domain Names on NSONE.NET

Currently displaying 50 of 1,282 domain names registered on December 10, 2016 and hosted at the nameserver **nsone.net**.

Domain Name

250years.org

4808beloitavenue.com

8955carsonstreet.com

900northproductions.com

a2smarthomes.net

aacont-accs-cgi-265.com


abacus-inventory.com

abovethelines.com

13.95.216.111		2016-12-03	22	ssh	SSH-2.0-OpenSSH_7.2p2 Ubuntu-4ubuntu2.1
13.95.216.78		2016-12-04	22	ssh	SSH-2.0-OpenSSH_6.6.1p1 Ubuntu-2ubuntu2

A large, bold, black stylized number '3' that serves as a background element for the title.

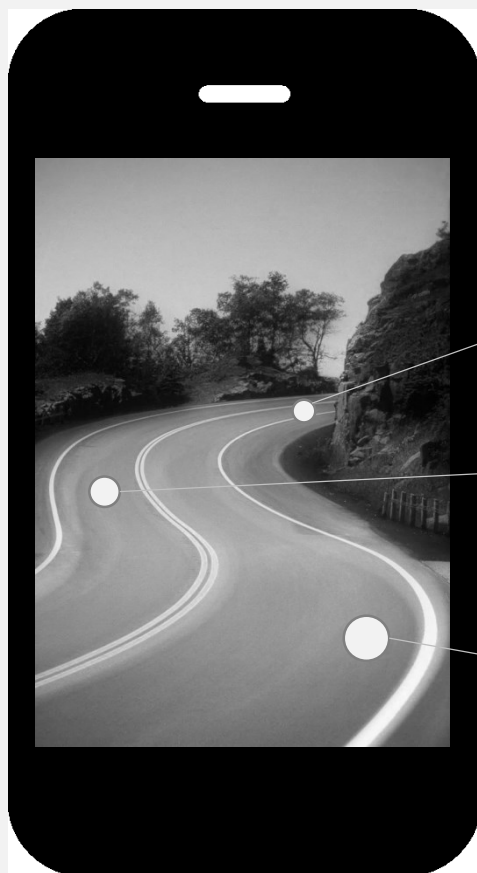
PART 03

A large, light gray circle that frames the subtitle and the quote.

后续·可持续发展

网络那么大，我想去看一看。

未完待续



威胁情报

从数据量足够大的时候
可以依托此平台进行优
化发展成情报收集平台。

不同角度

除了从域名及端口搜集
之外，我们还可以从不
同维度进行收集，如
QQ群，论坛等等...

更多....

还有太多太多的思路，需要老司
机们开车去找啦。。。

THANKS

如有忽漏错误之处，请各位大牛拍砖指教。