2022北京网络安全大会
2022 BEIJING CYBER SECURITY CONFERENCE
全球网络安全 倾听北京声音

爬虫对抗实践与思考

吴春来 云盾智慧安全科技有限公司

Content

背景介绍

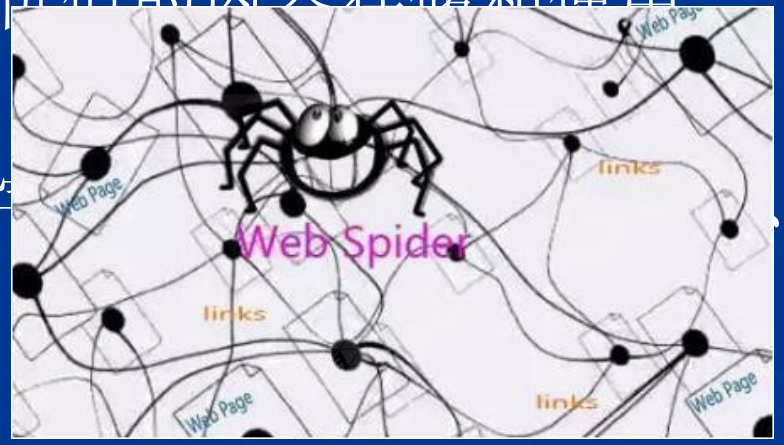
- 爬虫是什么
- 5W1H
- 爬虫带来的问题
- 相关术语定义

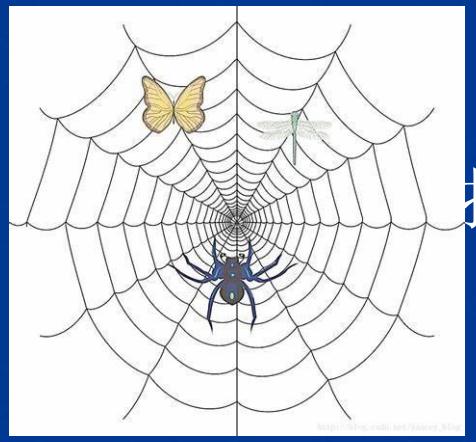
爬虫是什么？

- 出于各种目的，人们往往需要对互联网中的特定数据进行采集和整理，但人的精力是有限的，为了提高效率和准确性，就需要开发使用爬虫。

- 网络爬虫（下文简称爬虫），是一种会模拟人类行为访问网站的程序，它可以按照我们预设的规则自动化地请求网页数据，然后从中提取有价值的内容存储和使用。

- 关键字：[illegible]、自动化、[illegible]据提取





5W1H

WHO

个人&组织
企业

WHY

个人需求
商业需求

WHAT

公开数据
新闻政策
资源抢占
商品信息

WHERE

个人设备
云服务器

WHEN

被动采集
定时采集
持续采集

HOW

自动化脚本
集成式框架
分布式服务

爬虫带来的问题

主要问题

- 服务器负载过高，成本大幅提升
- 恶意的商业竞争，影响市场稳定
- 羊毛党和黄牛党，破坏既定秩序

应对措施

- 内部，业务厂商发力业务层风控，反爬虫反作弊，开放 API 引导爬虫
- 外部，安全厂商提供云防护平台、业务风控平台，云端进行流量清洗
- 政策，立法越来越关注数据安全，相关的黑灰产业链上下游判例不断

相关术语定义

- 爬虫

使用技术手段，通过一定规则自动化获取网站数据的一种方式。

- 反爬

使用技术手段，阻止自动化获取网站数据的一系列方法。

- 拦截

对爬虫的访问进行有效限制，使其请求无法到达服务器。

- 误伤

在爬虫对抗过程中，错误的将普通用户识别为爬虫进行限制。

相关术语定义

- 成本

攻守双方为了达成目的，所需要的物理资源与人力资源的总和。

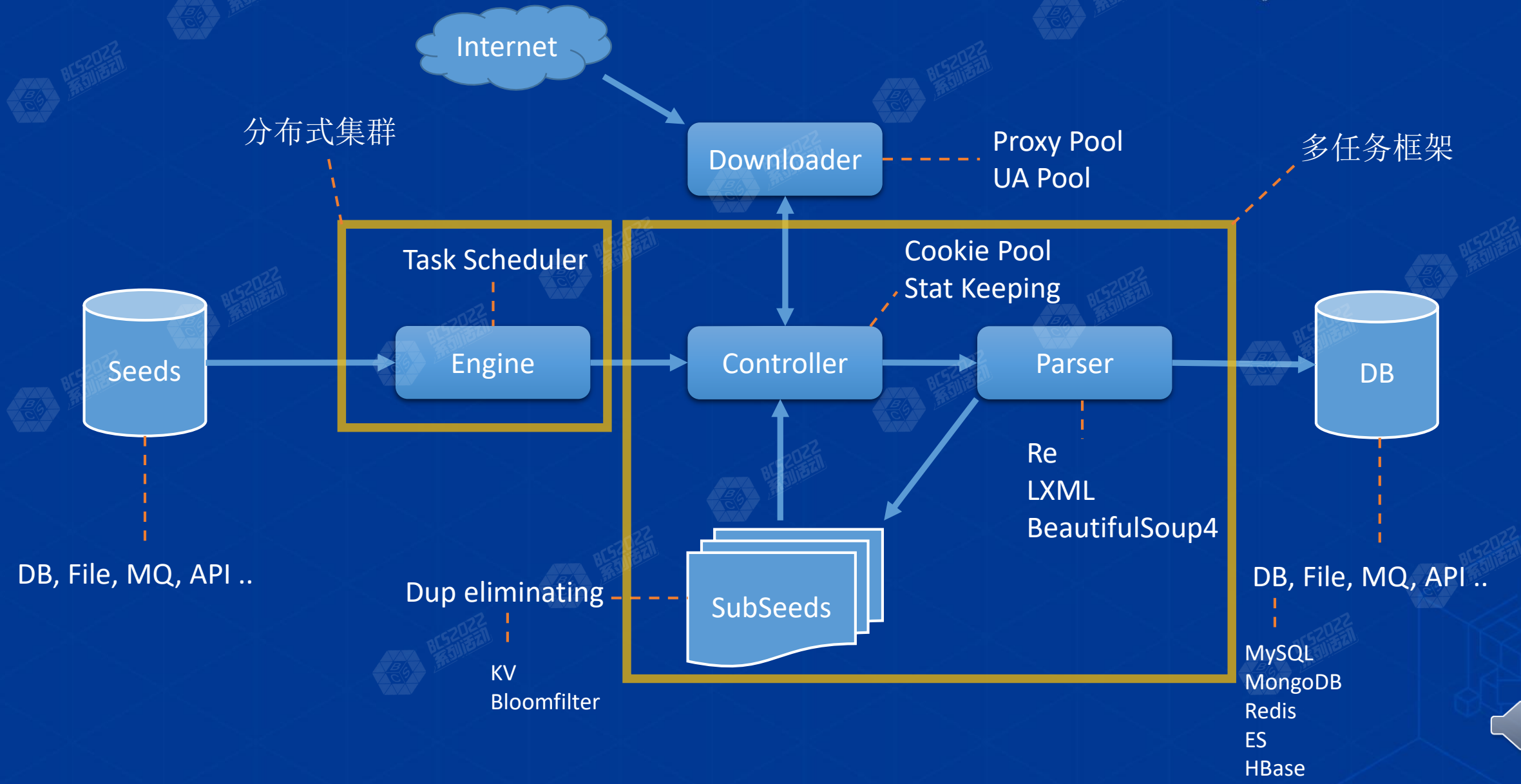
低级的爬虫，性能好，成本低，易被限制。
高级的爬虫，性能低，成本高，难被限制。

在对抗过程中，当成本高到一定程度的时候，往往付出与收益不成正比，无条件死磕是不划算的。

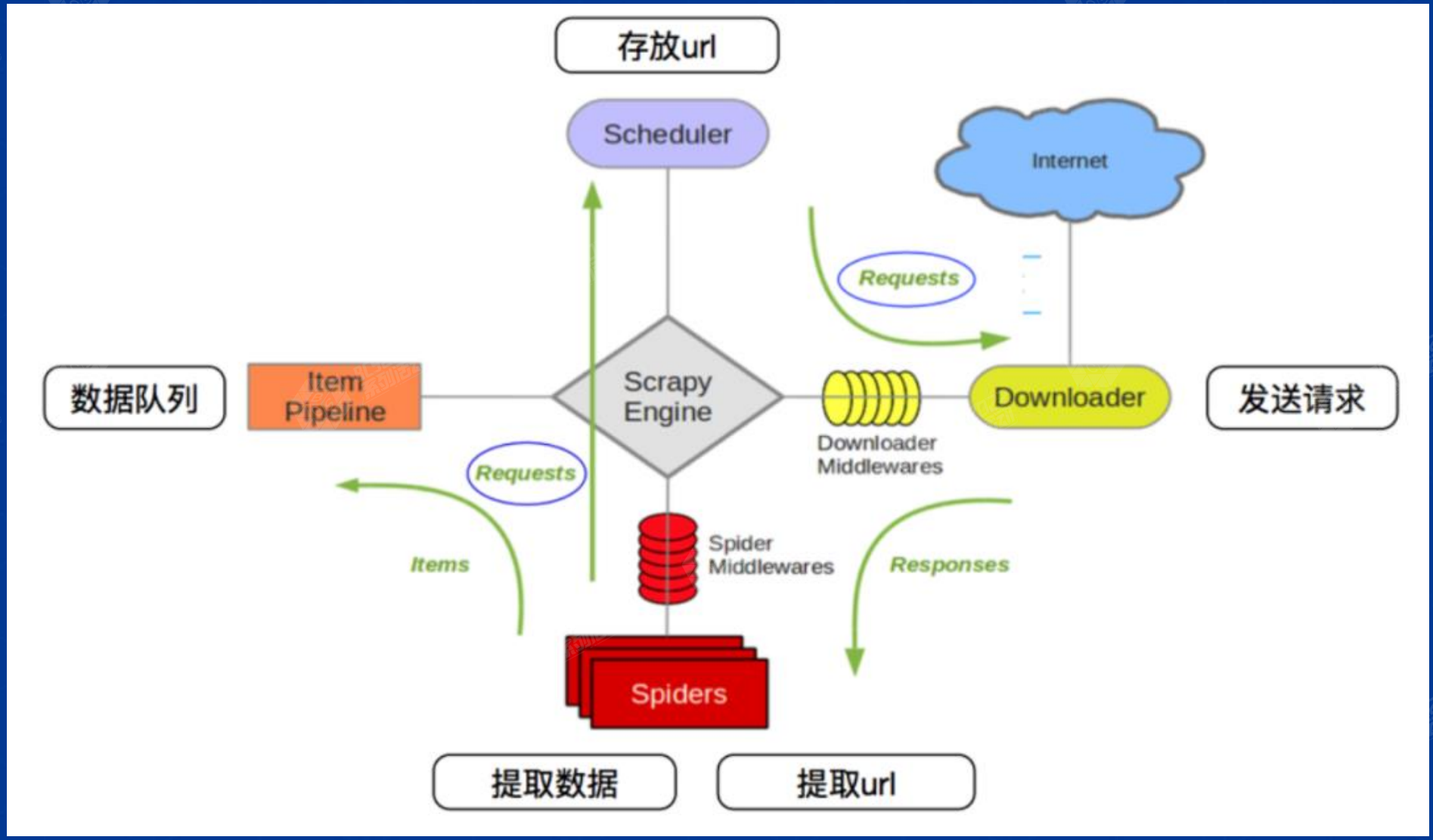
架构概述

- 爬虫服务架构
- 防护系统架构

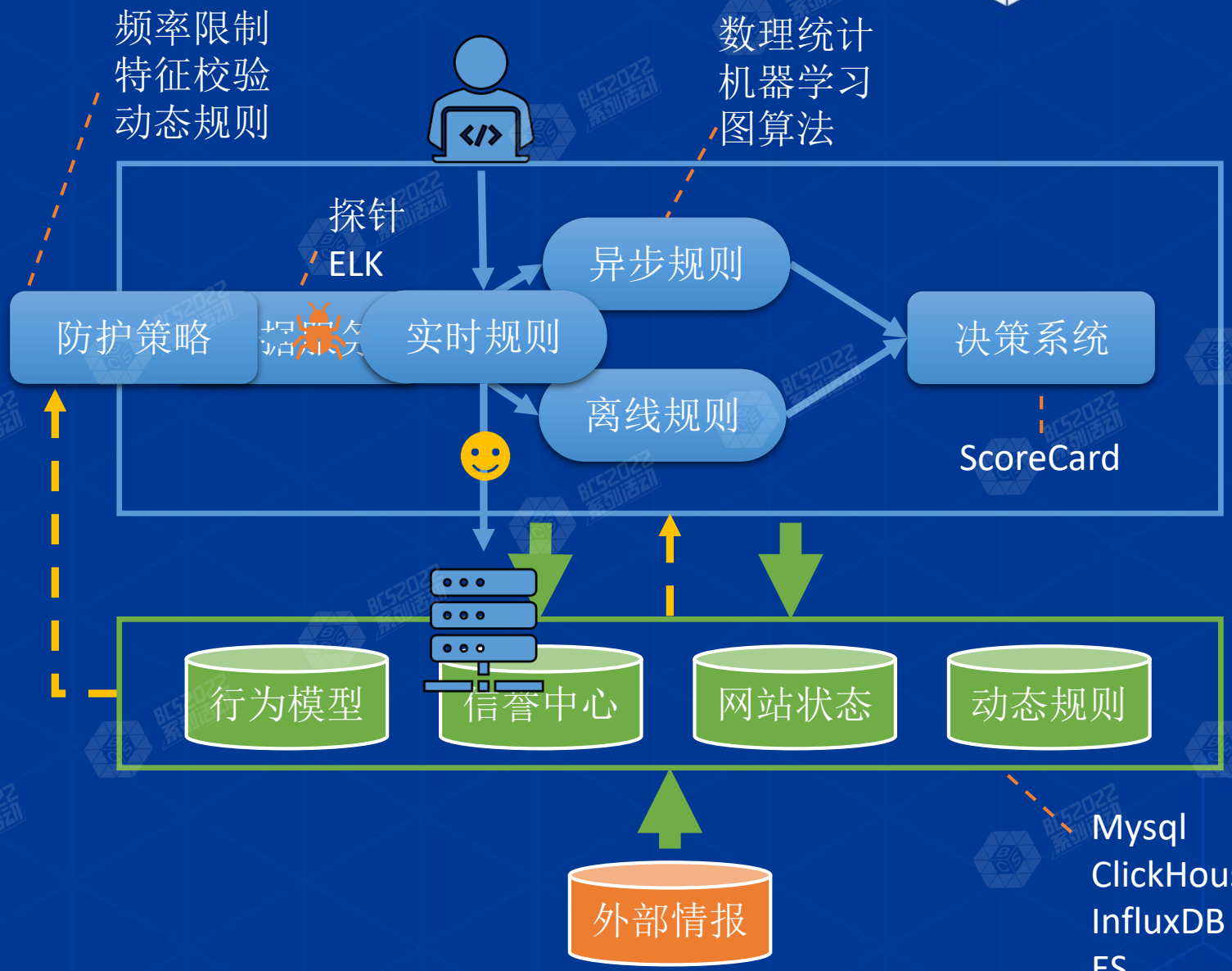
爬虫服务架构


Internet
分布式集群
Downloader
Proxy Pool
UA Pool
多任务框架
Task Scheduler
Cookie Pool
Stat Keeping
Seeds
Engine
Controller
Parser
DB
Re
LXML
BeautifulSoup4
DB, File, MQ, API ..
Dup eliminating
SubSeeds
DB, File, MQ, API ..
KV
Bloomfilter
MySQL
MongoDB
Redis
ES
HBase

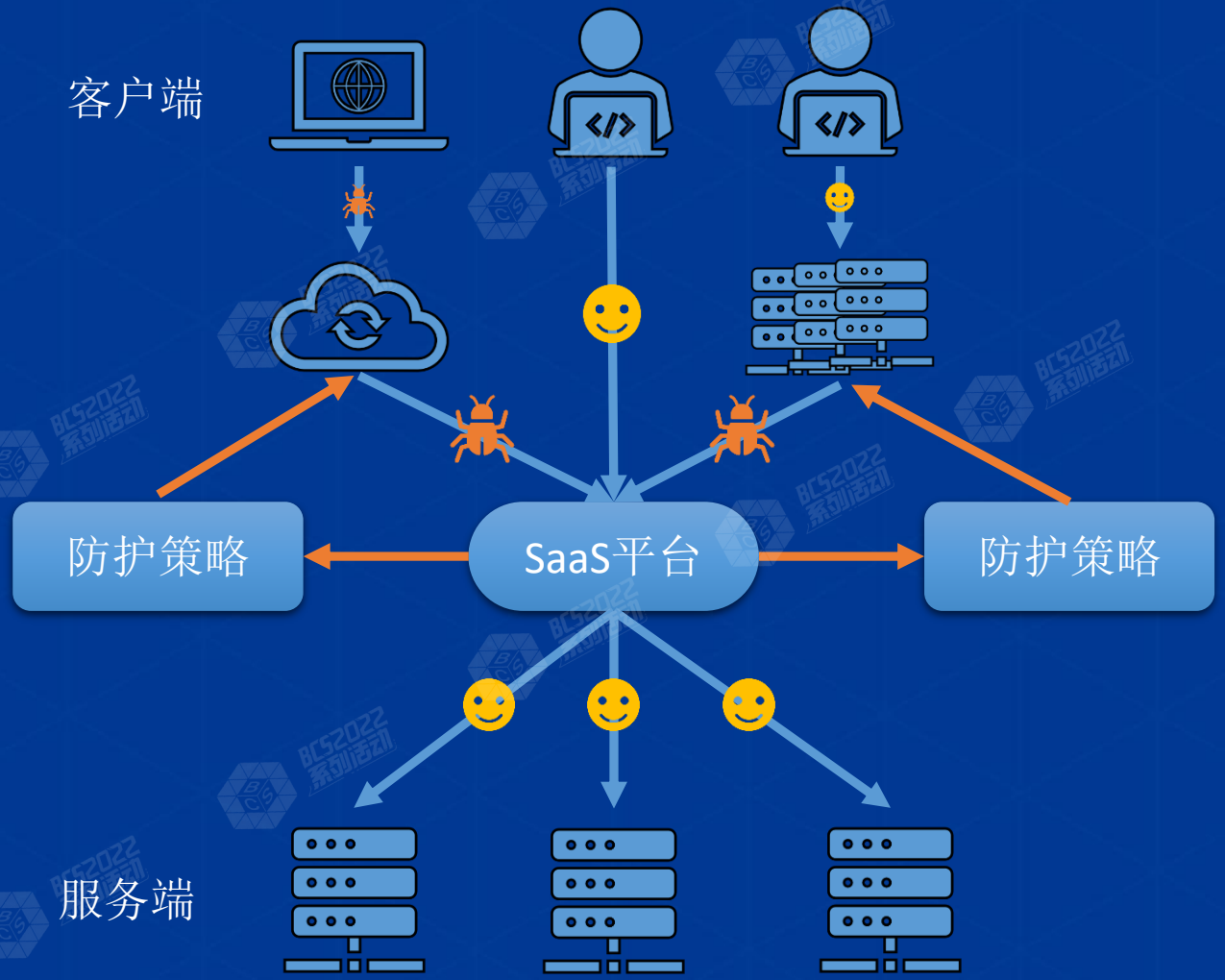
爬虫服务架构 — Scrapy



防护系统架构


频率限制
特征校验
动态规则
数理统计
机器学习
图算法
探针
ELK
异步规则
防护策略
实时规则
决策系统
离线规则
ScoreCard
协议验证
人机识别
图片验证
诱导欺骗
行为模型
信誉中心
网站状态
动态规则
外部情报
Mysql
ClickHouse
InfluxDB
ES

防护系统架构 — SaaS平台

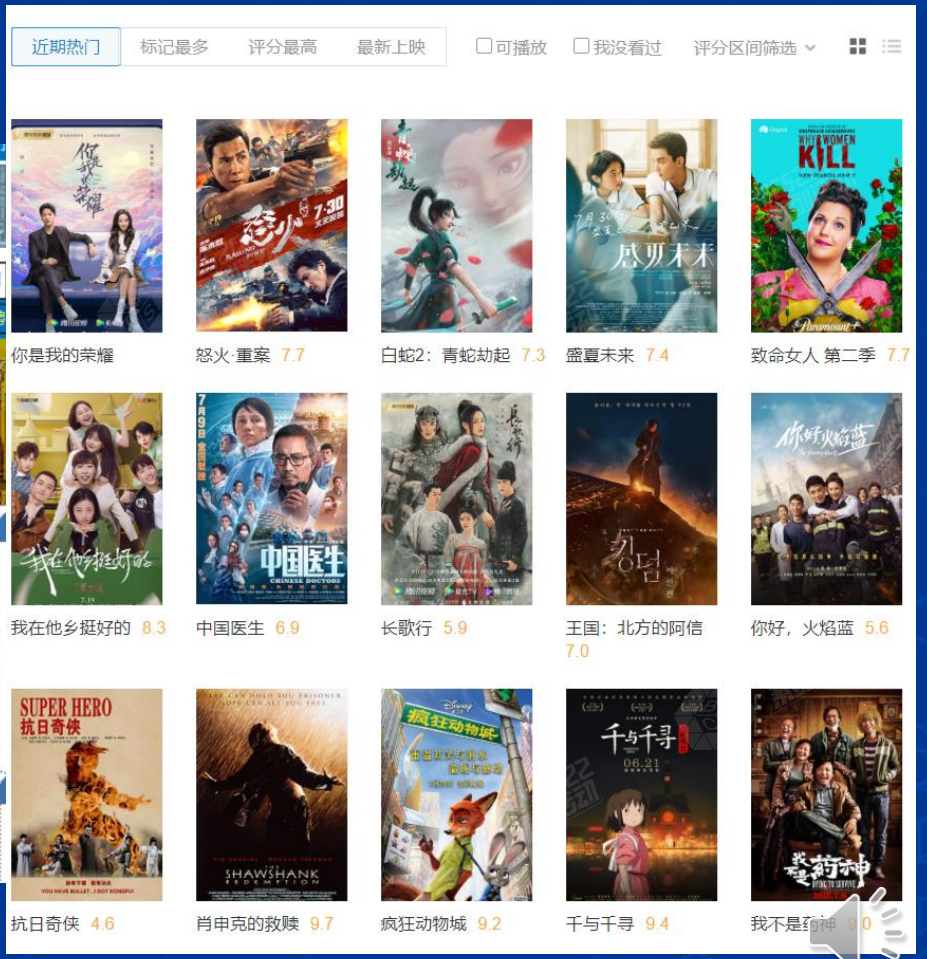


红蓝对抗

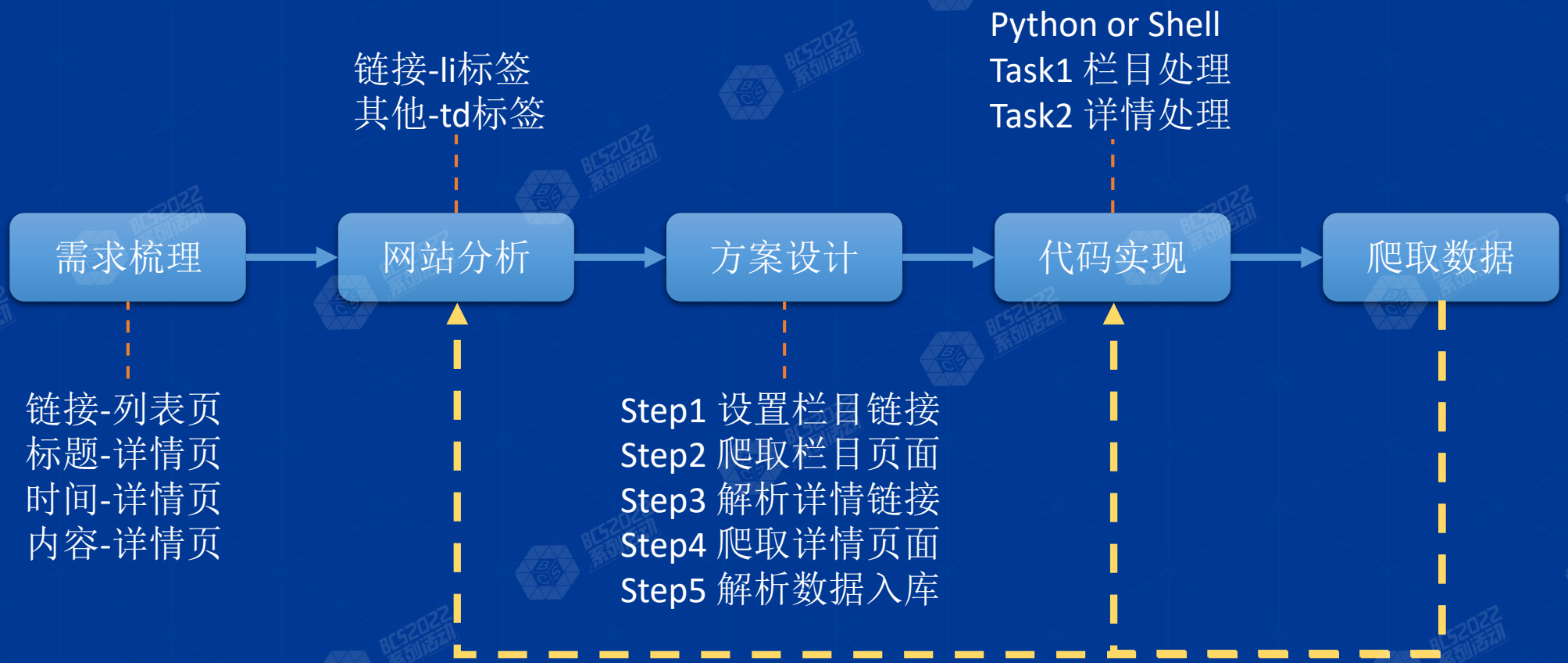
- 对抗初级阶段
- 对抗中级阶段
- 对抗高级阶段

对抗初级阶段 — 网站设计

- Web1.0 时代，依赖于服务端渲染，网站以静态资源和链接为主，只有部分网站使用了 JS 做一些简单逻辑。



对抗初级阶段 — 爬虫思路


链接-li标签
其他-td标签
Python or Shell
Task1 栏目处理
Task2 详情处理
需求梳理
网站分析
方案设计
代码实现
爬取数据
链接-列表页
标题-详情页
时间-详情页
内容-详情页
Step1 设置栏目链接
Step2 爬取栏目页面
Step3 解析详情链接
Step4 爬取详情页面
Step5 解析数据入库

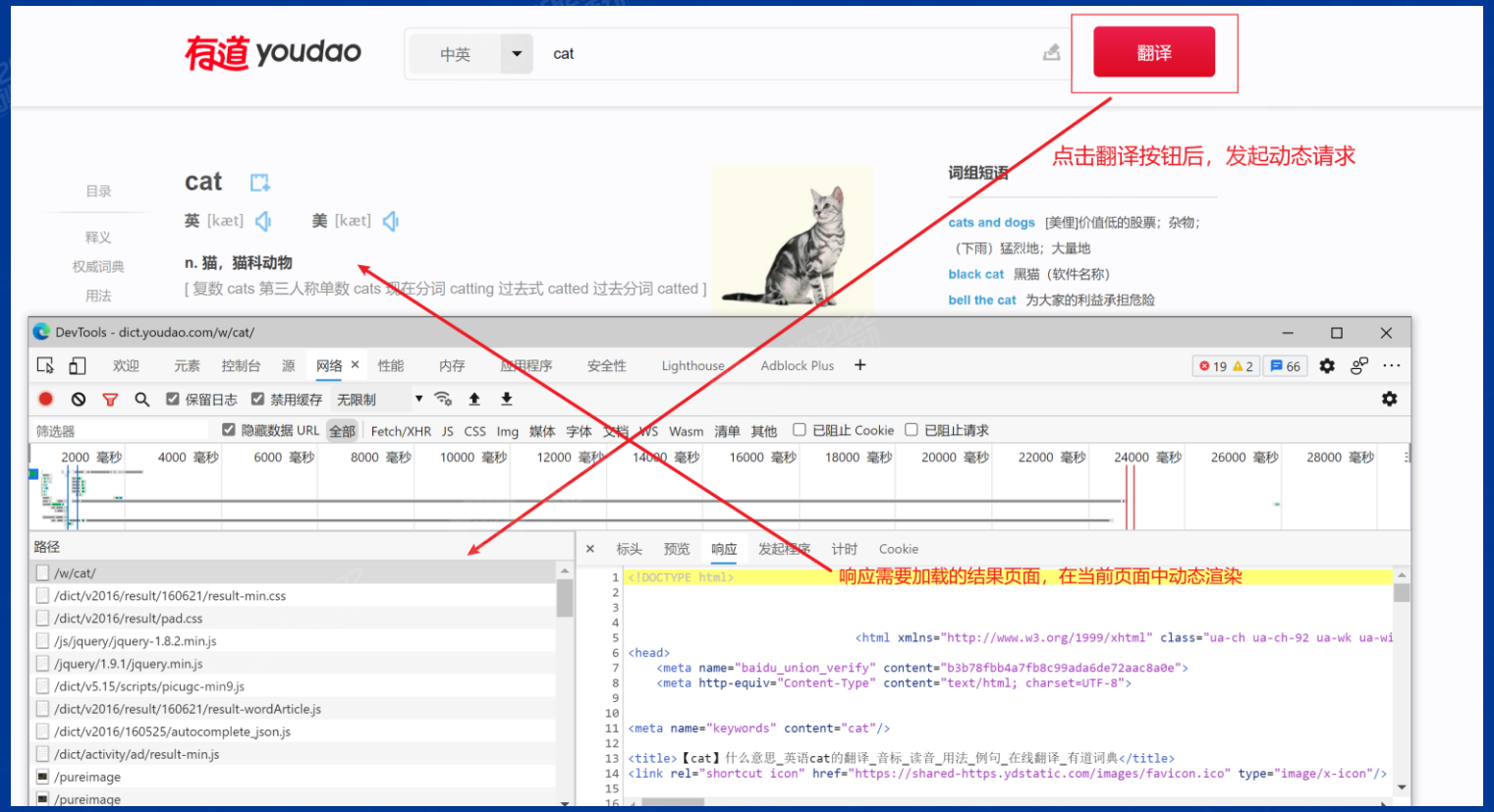
对抗初级阶段 — 对抗视角

- 无视 robots 协议，直接爬数据
- 设置 User-Agent 和 Referer 头部
- 运营商 NAT 接入或使用动态拨号
- 每次爬取数据都清空 Session
- 维护 UA 池，时不时切换一下

- robots.txt 协议
- Headers 检查
- IP 频率控制
- Session 频率控制
- UA 频率控制
- 注册登录机制

对抗中级阶段 — 网站设计

- Web2.0 时代，随着 jQuery、HTML5 和 V8 引擎的出现，前端技术开始快速发展。越来越多的工作通过 JavaScript 进行处理，数据也开始通过 API 传输，网站开始朝着动态化的方向发展。



对抗中级阶段 — 对抗视角1

控制访问频率

创建代理池，通过代理爬取

WebRTC 局域网地址
Session/UA 变换频率
IP 访问频率

蜜链引诱

JS 能力验证

对抗策略 – JS 验证

发起 HTTP 请求

/page

判断 Token 有效性

ok

nil or invalid

计算 JS 中的问题

1+1=
?

响应 JS 页面

响应资源数据

a=1+1

n!=2

携带答案进行 XHR 请求

/page?a=n

判断问题与答案一致性

n==2

加载页面数据

html with token

对抗中级阶段 — 对抗视角1

控制访问频率

创建代理池，通过代理爬取

解析实现验证算法，自动化绕过

JS 能力工具集成

NodeJS
PyExecJS
Selenium
PhantomJS

WebRTC 局域网地址
Session/UA 变换频率
IP 访问频率

蜜链引诱

JS 能力验证

JS 混淆技术

webpack
uglifyJS
obfuscator

图片验证码

对抗中级阶段 — 对抗视角2

爬虫携带登录获取的 Token

创建账号池，批量获取 Token

使用黑灰产接码平台注册

API 接入黑灰产账号平台

提供大量账号
提供登录 Token
提供账号+验证码

注册登录机制

基于账号频率控制

账号注册手机绑定

实名注册，登录验证

常用地关联风险评估

实名注册
验证码登录
登录环境绑定

对抗高级阶段 — 网站设计

Web3.0 时代，用户服务应用开始分布在 Web、Android、IOS、HarmonyOS、公众号、小程序等多种平台上。

各种各样的前端框架开始出现，其中作为代表的就是 Vue、React，还有 Flutter 这样跨移动平台框架，前后端数据传输对 API 产生了更强的依赖。

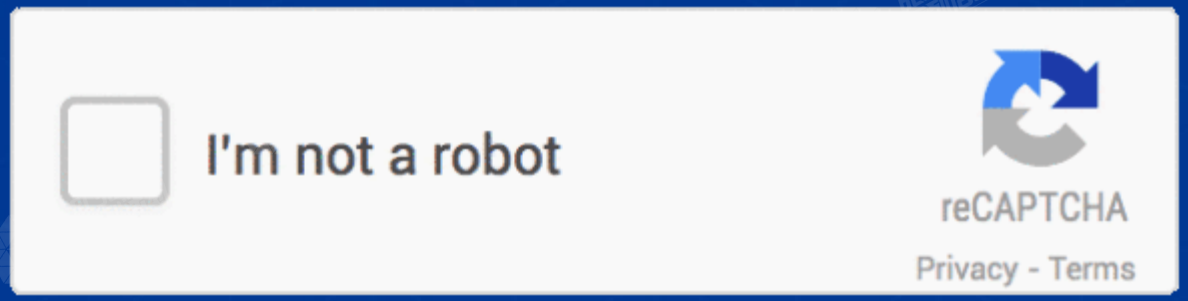
API 安全成为关键命题。

对抗高级阶段 — 对抗视角

Hook 技术	网络劫持
环境定制	逆向破解
第三方缺陷利用	机器学习

人机检测	动态验证
投毒欺骗	数据加密
机器学习	动态协议

对抗策略 – 环境验证


I'm not a robot
reCAPTCHA
Privacy - Terms


虎扑 JR
忘记密码?
没有帐号，立即注册 >
点击按钮开始智能验证
登录
微信登录
QQ登录
手机短信快捷登录

对抗策略 – 文本验证


文本验证
验证内容
ASCII
Unicode
验证方式
图片
语音
短信

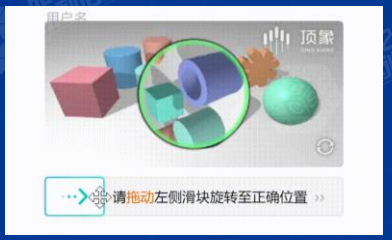

请填写验证码
看不清
换一张
请输验证码
不区分大小写
确认


编辑短信 046336
发送至106981630163999
网易易盾
等待短信验证，剩余 299s

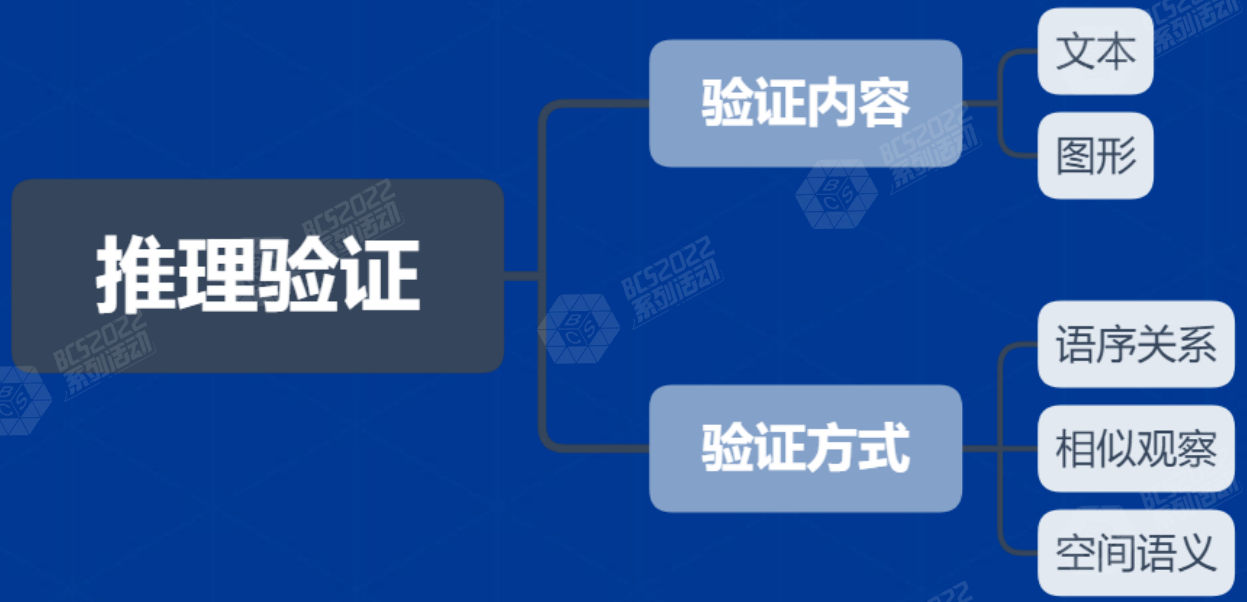

播放语音验证码
请输入语音验证码
刷新验证码
返回
验证

对抗策略 – 滑块验证

- 滑动验证
  - 滑动方式
    - 轨道
    - 自由
  - 表现形式
    - 缺块拼图
    - 旋转拼图
    - 推理拖动



对抗策略 – 推理验证



对抗高级阶段 — 同盟发展

传统代理池
动态 VPS 拨号
账号池
定制工具
爬虫接口平台
人工打码平台

威胁情报
IP 画像
机器学习
安全技术
认证 SDK
风控产品

对抗总结

业务安全检测的思路来自于人的不确定性，而随着技术不断发展机器变得越来越像人，导致“人”的行为并不都是非黑即白的，在黑白之间存在着非常巨大灰色地带。

红蓝对抗的焦点是信息。红方通过技术手段消耗蓝方资源、提升信息获取成本，达到保护信息安全的目的；而蓝方同样通过技术手段提升信息获取效率，争取利益最大化。这将是场旷世持久的战争。

欺诈者总有办法破解你，不过是成本高低罢了，如果你的反欺诈措施足够强大，那欺诈者就不会花很大精力来找你麻烦，转而会去寻找那些防护薄弱的目标。

相比其他工种来说，爬虫开发、风控开发还是比较有意思的，毕竟与人斗其乐无穷。

团队介绍

公司介绍

云盾智慧
iCloudShield

云盾智慧安全科技有限公司

致力于网络安全产品、服务和解决方案的创新与研发，为政府、企事业单位等各行各业机构提供企业级网络安全产品、服务、解决方案的网络安全高科技公司。

云盾智慧安全科技有限公司，于2019年由中国联通与奇安信共同出资成立，公司专注于网络安全产品、服务和解决方案的创新与研发，为各行各业机构提供企业级网络安全产品、服务、解决方案。

公司主要产品有云防护、云监测、风险管理系统、云安全管理平台、态势感知等。

https://www.icloudshield.com/

个人介绍

吴春来，后端程序猿，跟随团队从360到奇安信再到云盾智慧，见证了团队从0到1的发展。

喜欢从0到1的思考和学习过程，热衷于模块化开发和系统设计，目前专注于反爬虫方向的技术研究。

欢迎大家多多交流和指教，一起进步~

THANKS