



业务安全之反爬虫实践

董俊杰 2017.11



目

录

1

爬虫带来的安全风险

2

爬虫的识别与防御

3

反爬虫系统架构

4

思考与Q&A

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



1

爬虫带来的安全风险

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



1 爬虫带来的安全风险

网络爬虫（网络蜘蛛），通过不断访问互联网的各种资源，根据一定的规则收集特定信息的自动化程序或脚本。



Robots协议（robots.txt），告诉网络爬虫，哪些页面可以爬取，哪些页面不可以爬取，达到保护隐私的目的。

黑产

搜索引擎

爬虫技术

灰色地带
(ATS)

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



2

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

爬虫识别与防御

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE





频率限制

通过对运营数据进行统计与分析，限制单位时间内客户端向服务端发起的请求总数。

-运营数据



行为(环境)分析

通过客户端埋点，收集客户端设备信息或用户的操作动作。

-设备指纹
-浏览器插件
-鼠标动作
-屏幕动作



威胁情报

通过蜜罐或其它威胁情报提供方获取各种维度的威胁情报。

-蜜罐信息
-IP黑名单库
-手机号黑名单库



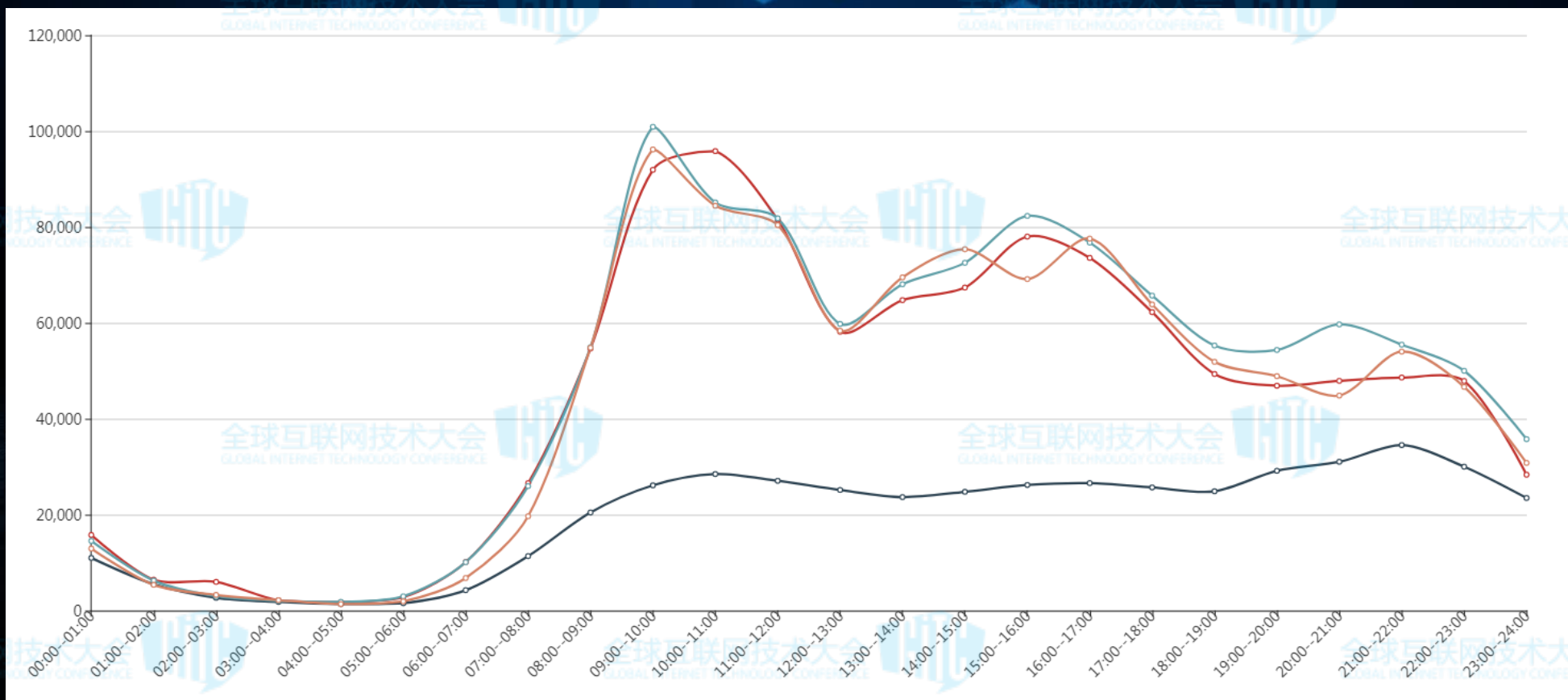
用户画像

通过用户注册信息结合用户历史行为，对用户进行用户画像。

-用户注册信息
-历史行为

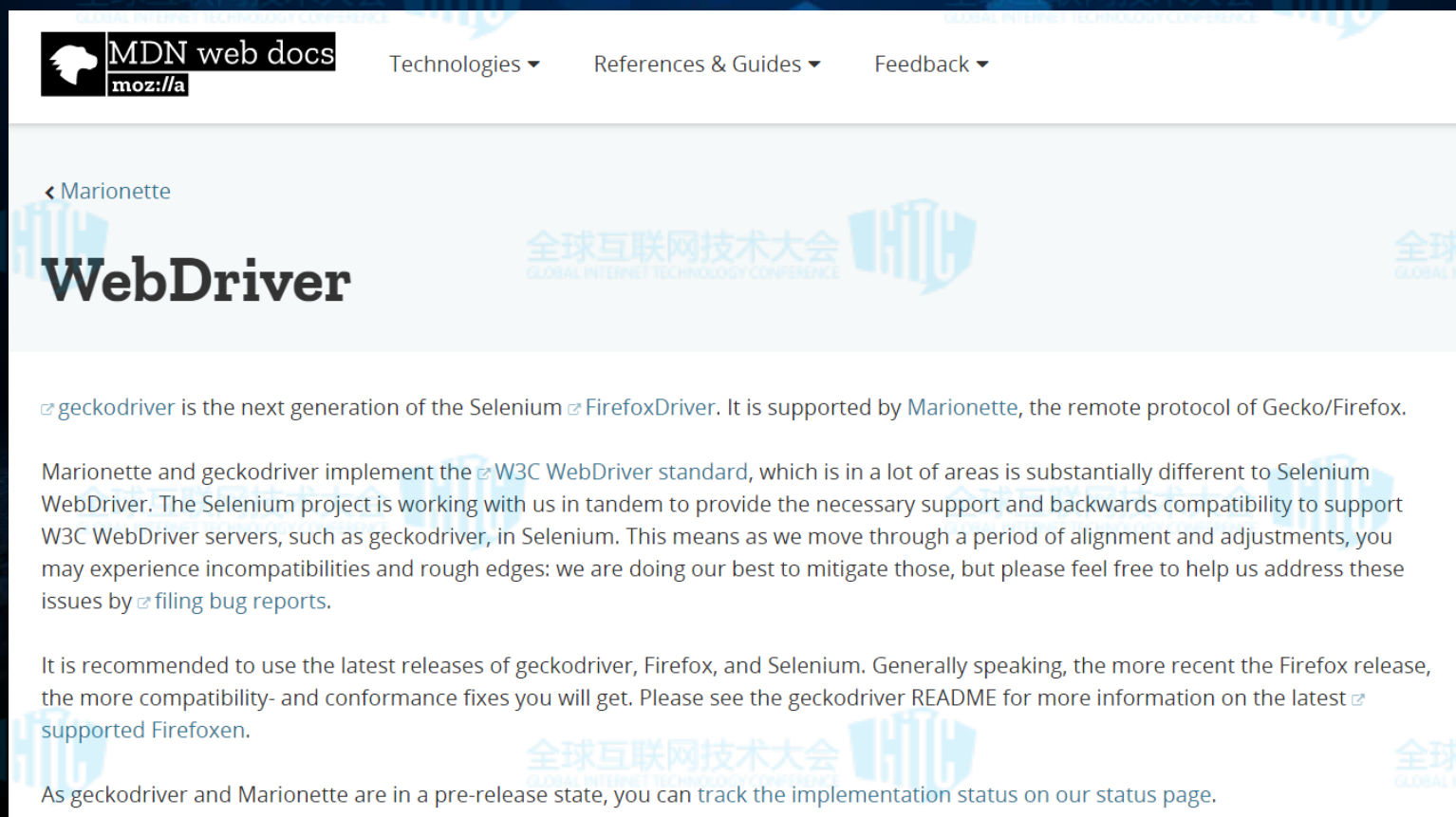
频率限制：

非常考验阈值的设置，需要考虑业务的周期性与非周期性增长，稍有不慎可能影响业务。



行为(环境)分析：

非常依赖前端技术，要关注客户端代码保护与收集信息的验证，基于浏览器驱动的用户行为模拟。



The screenshot shows the MDN web docs page for WebDriver. The page header includes the MDN logo, the text "MDN web docs", and the URL "moz://a". Navigation links for "Technologies", "References & Guides", and "Feedback" are visible. The main content area features a breadcrumb "Marionette" and the title "WebDriver". The text explains that GeckoDriver is the next generation of Selenium WebDriver, supported by Marionette. It notes that Marionette and GeckoDriver implement the W3C WebDriver standard, which differs from Selenium WebDriver. The page also mentions that Selenium is working with MDN to support W3C WebDriver servers like GeckoDriver. Recommendations include using the latest releases of GeckoDriver, Firefox, and Selenium. A link to the GeckoDriver README is provided for more information. Finally, it states that as GeckoDriver and Marionette are in a pre-release state, users can track implementation status on the status page.

MDN web docs
moz://a

Technologies ▾ References & Guides ▾ Feedback ▾

← Marionette

WebDriver

[GeckoDriver](#) is the next generation of the Selenium [WebDriver](#). It is supported by Marionette, the remote protocol of Gecko/Firefox.

Marionette and [GeckoDriver](#) implement the [W3C WebDriver](#) standard, which is in a lot of areas is substantially different to Selenium WebDriver. The Selenium project is working with us in tandem to provide the necessary support and backwards compatibility to support W3C WebDriver servers, such as [GeckoDriver](#), in Selenium. This means as we move through a period of alignment and adjustments, you may experience incompatibilities and rough edges: we are doing our best to mitigate those, but please feel free to help us address these issues by [filing bug reports](#).

It is recommended to use the latest releases of [GeckoDriver](#), Firefox, and Selenium. Generally speaking, the more recent the Firefox release, the more compatibility- and conformance fixes you will get. Please see the [GeckoDriver README](#) for more information on the latest supported Firefoxen.

As [GeckoDriver](#) and Marionette are in a pre-release state, you can track the implementation status on our [status page](#).

威胁情报：
如何巧妙的设置蜜罐（everything is honey pot）；
威胁情报平台的数据量与准确性。

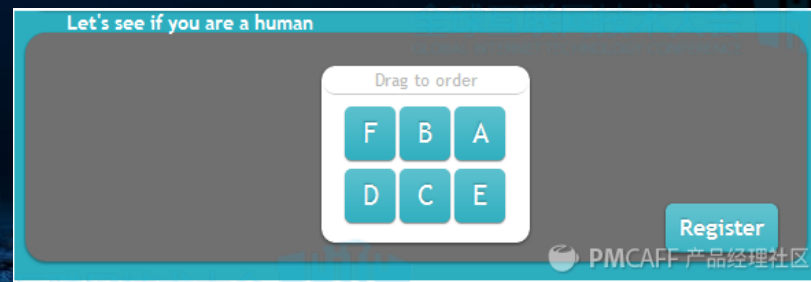
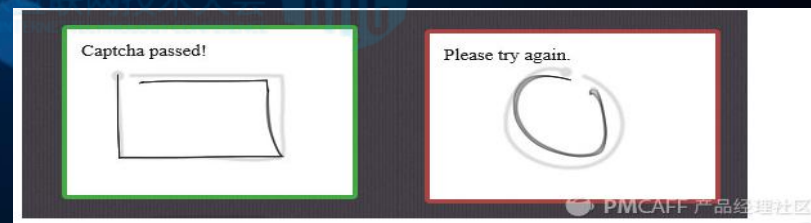
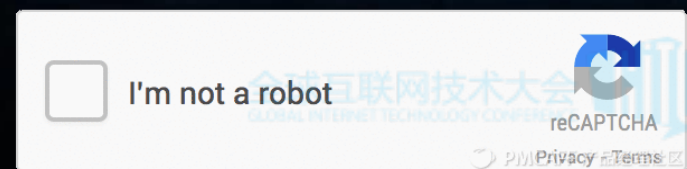
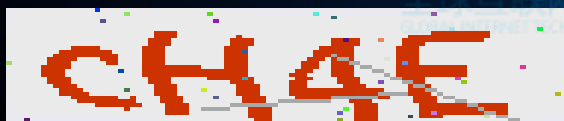


用户画像：
结合具体的业务场景，区分正面样本与负面样本。





图形（点选）验证码：
用户体验相对较好，防御效果一般



短信（邮件）验证码：
用户体验较差，防御效果优于图形验证码

请选择验证身份方式：

昵称：

已验证手机： 若该手机号已无法使用请联系客服

请填写手机校验码：

一封邮件已经寄到了您的暴雪游戏电子邮箱。
请输入您收到的验证码。 [重新发送验证码](#)

输入验证码.....

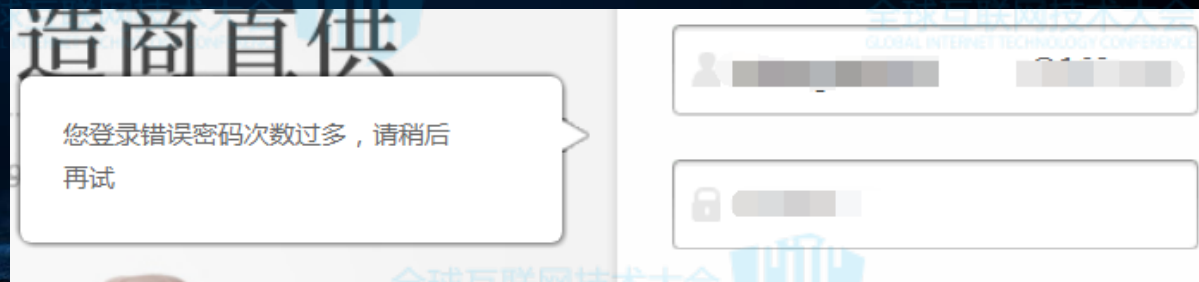
继续

安全验证

请编辑短信“短信验证”发送至1069 8163 0163 331
完成验证，然后输入手机号并点击“已发送”

输入手机号码

限制账号、IP、设备访问：
对用户影响极大，仅适用准确率较高的策略



策略 = 检测KEY + 检测规则 + 动作 + 有效时长

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



3

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

反爬虫系统架构

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

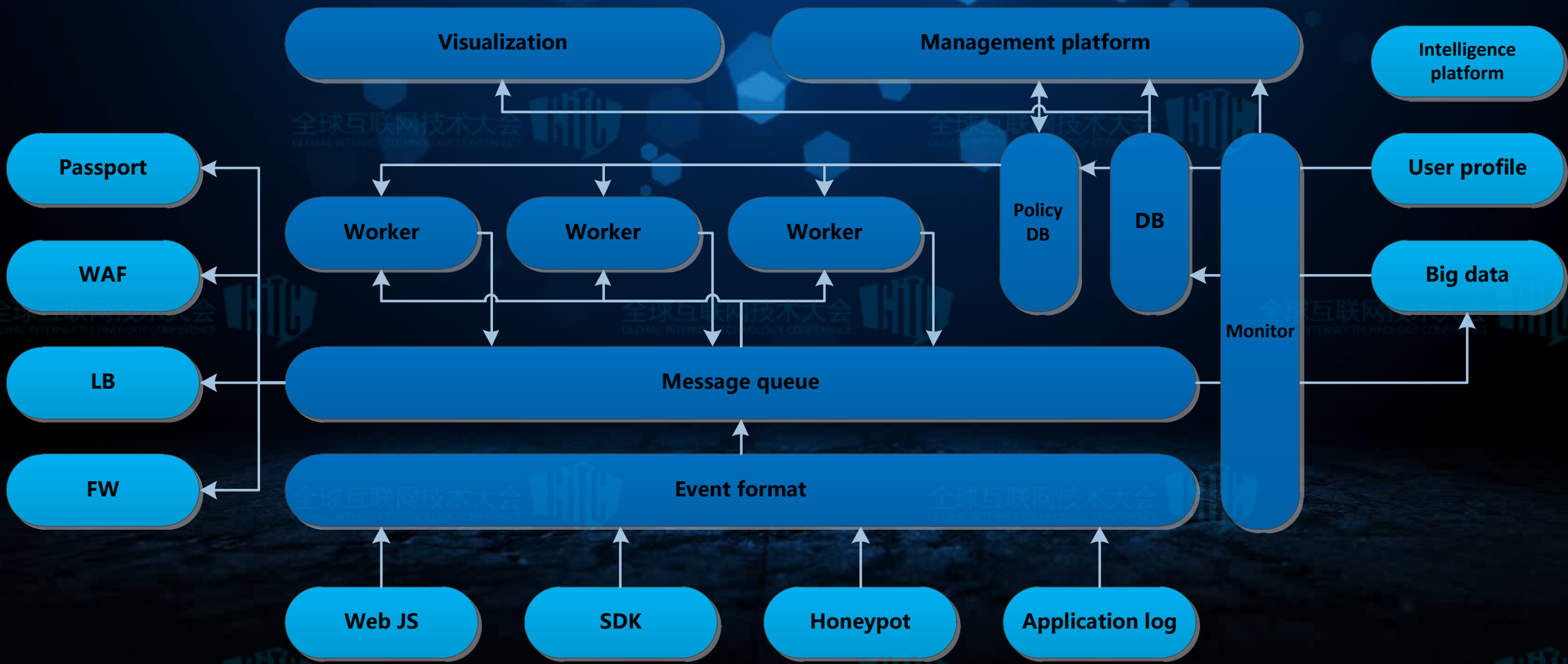


全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE





全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



4

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

思考与Q&A



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



业务安全风险本质上是“人”的不确定性，其行为并不都是非黑即白的，中间还有着一片巨大的灰色地带；安全团队要做的就是尽可能划定黑与白的界线，同时探索这片未知，从而不断缩小灰色地带的范围。



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



Q&A

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



THANKS

THANKS

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

