THALES
Building a future we can all trust

SECURING YOUR AI-POWERED APPLICATIONS

Bartosz Chmielewski

www.thalesgroup.com

Bartosz Chmielewski

Senior Sales Engineer –
AppSec Specialization
(CEE, CIS, BeNeLux)

Data exfiltration

DDoS

Account takeover

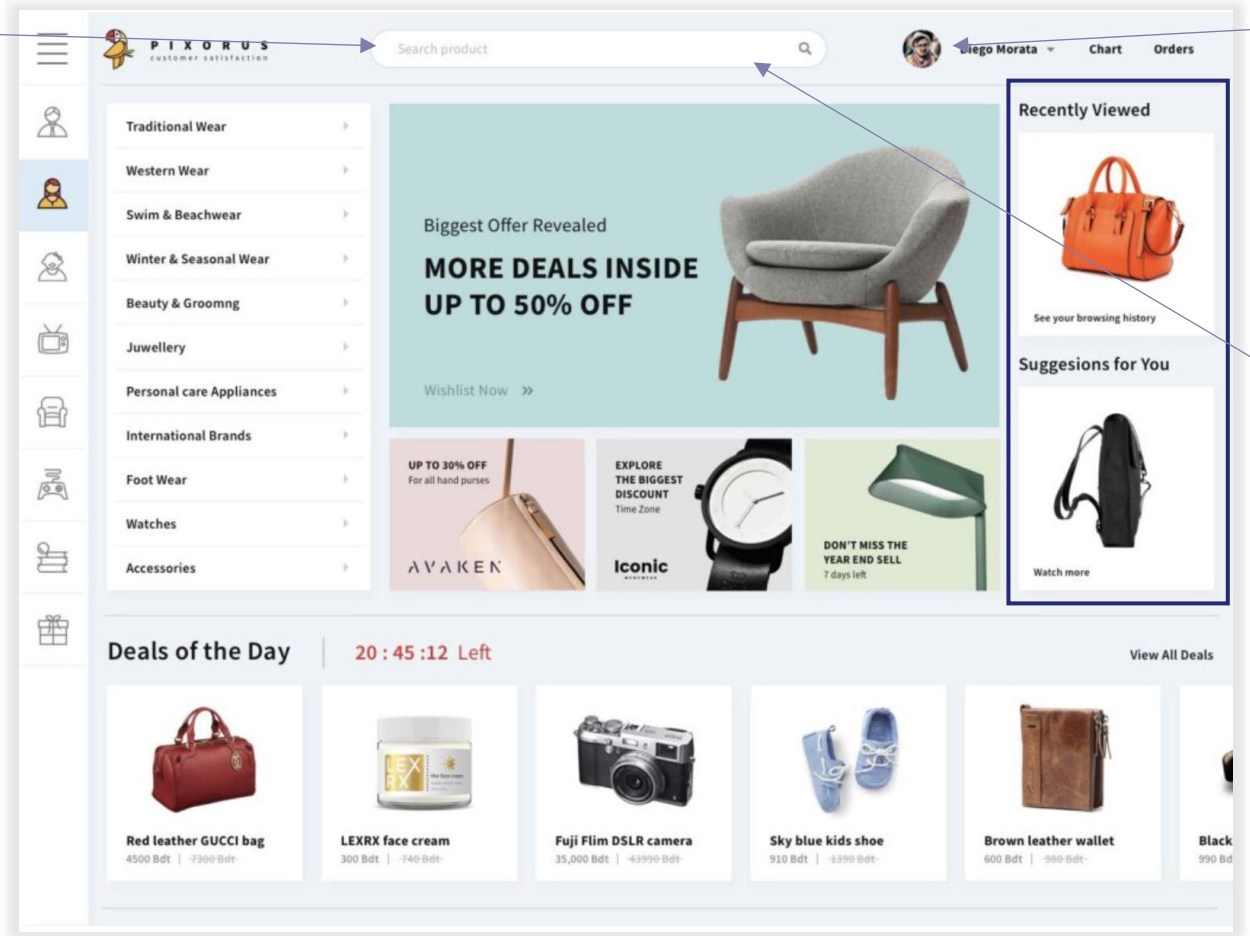
Technical Attacks

Bad bots

Credit card Fraud

Scalping

Scraping



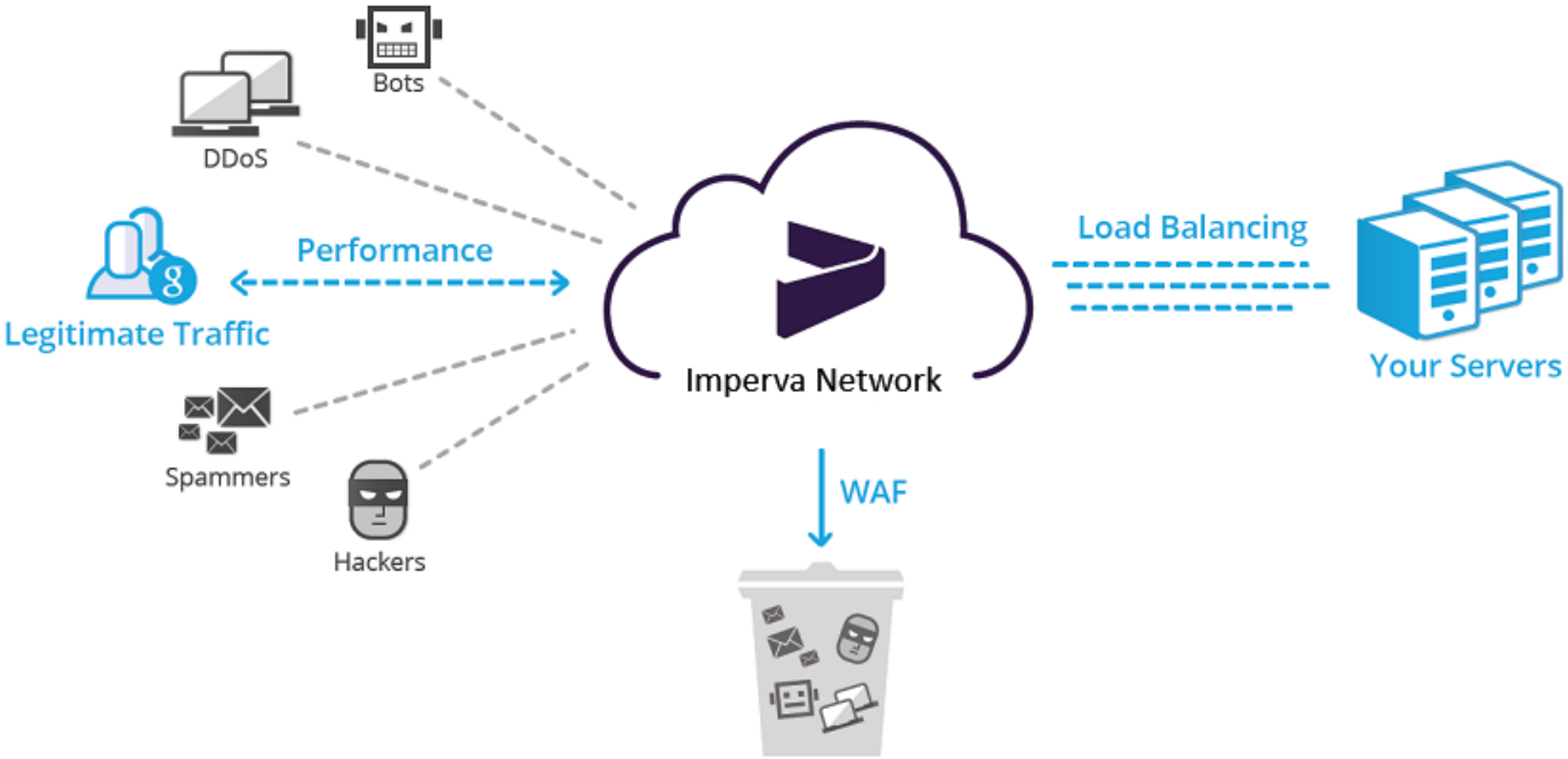
Imperva AppSec WAF Options Comparison

	Deployment Method	Deployment Location	Deployment Responsibility	Management
Cloud WAF	Multi-tenant SaaS	Imperva Global POP Network	Imperva	Imperva Cloud Console
WAF Gateway	HW or VMs. (Transparent) Reverse Proxy, Bridge.	On-Prem or in CSP as IaaS.	Customer	Local Console
Elastic WAF	Cloud Native (K8s) application.	K8s on-prem or in the Cloud (via ingress plugin)	Customer	Imperva Cloud Console
Imperva for Google Cloud	Multi-Tenant SaaS	GCP LB Integration via STE (Service Traffic Extension)	Shared	Imperva Cloud Console

+ DDoS Protection for Networks

+ AI Firewall!

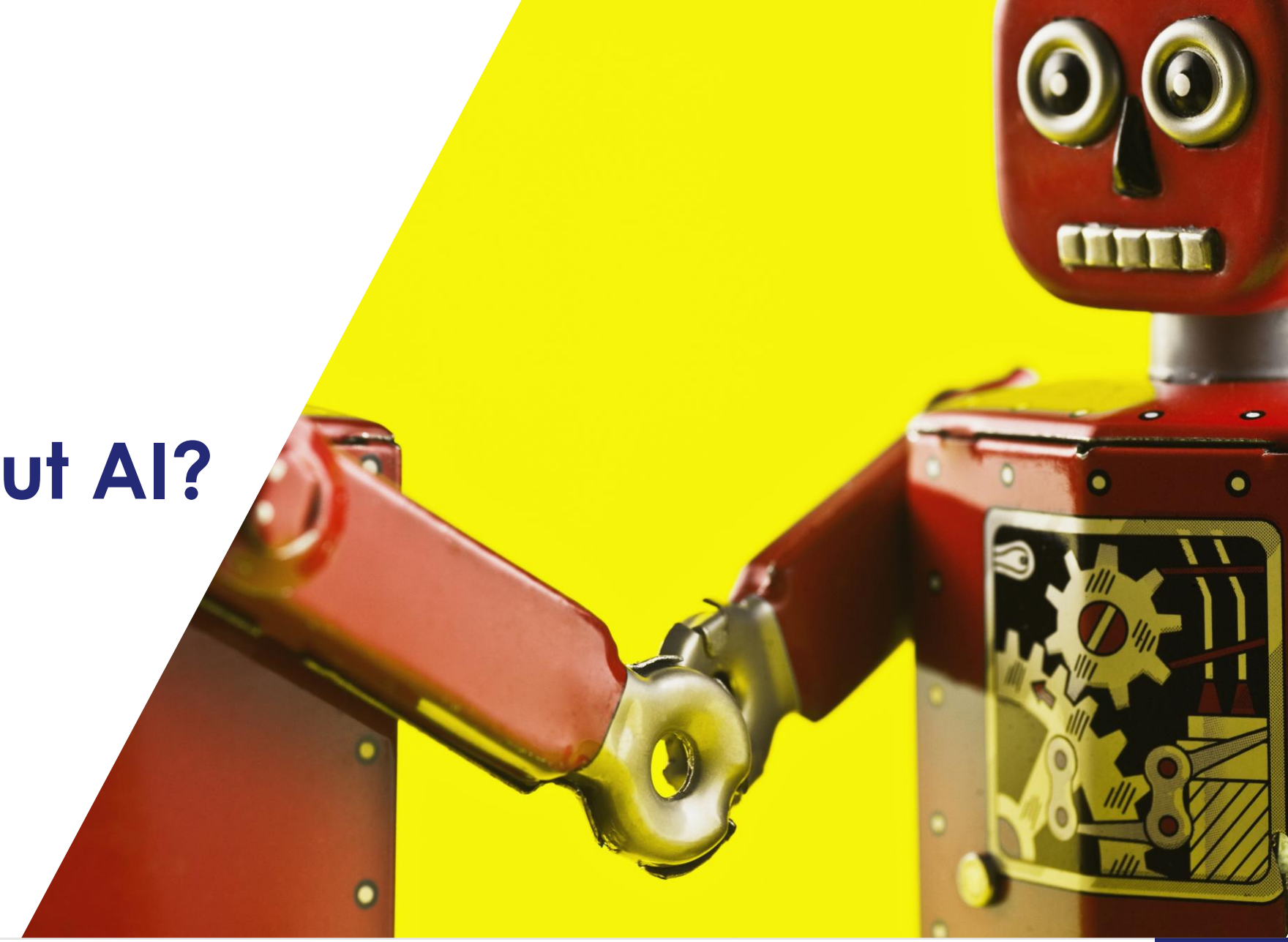
Imperva Cloud WAF



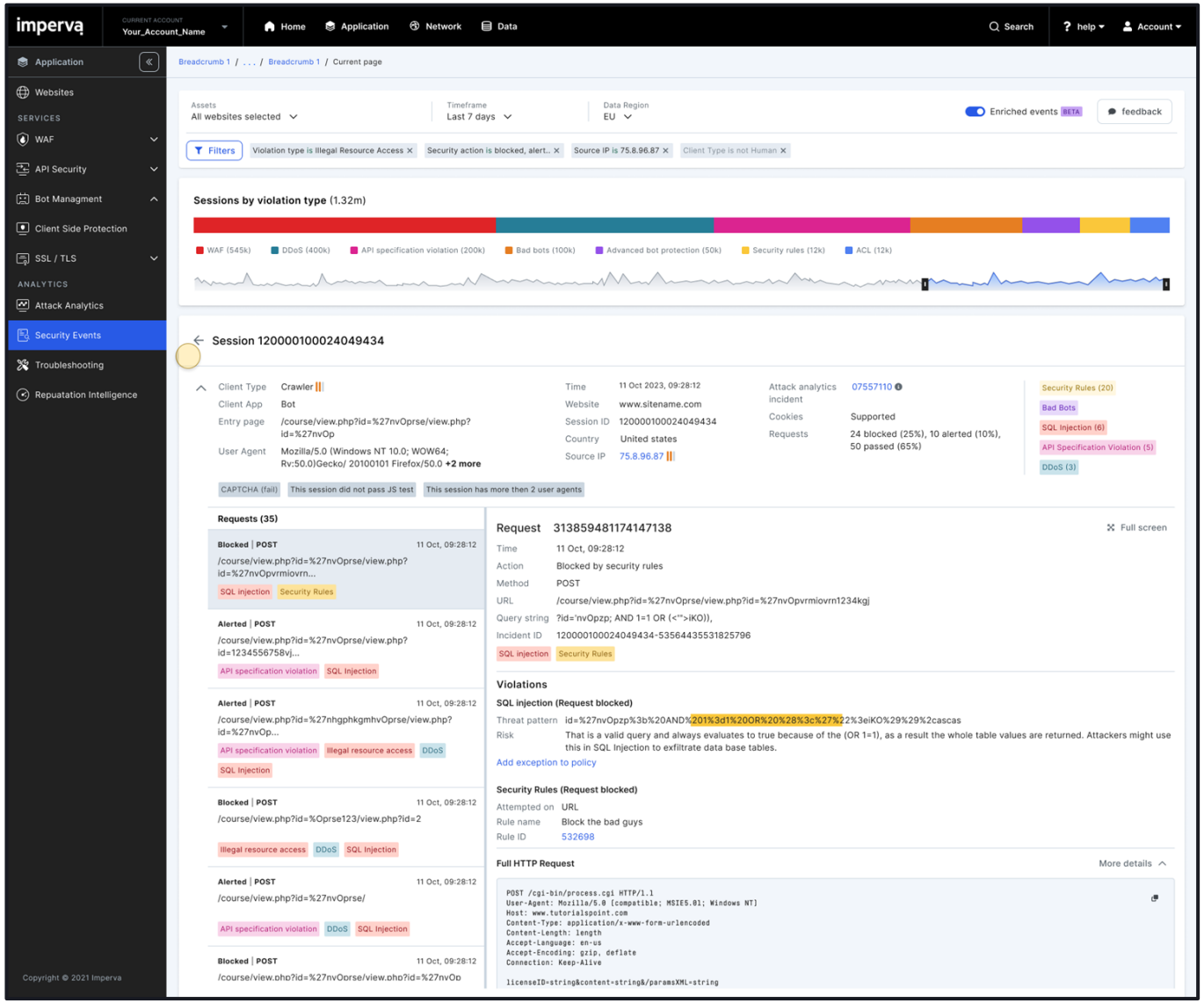
Imperva Global Network

- Global Mesh
- CDN
- Anycast Network
- Fully Automated Load Balancing and Policy Adjustment
- 63 PoPs
- 15+ TBps capacity
- Tier 1 ISPs, IXs and T1 Data Centers
- Single Stack

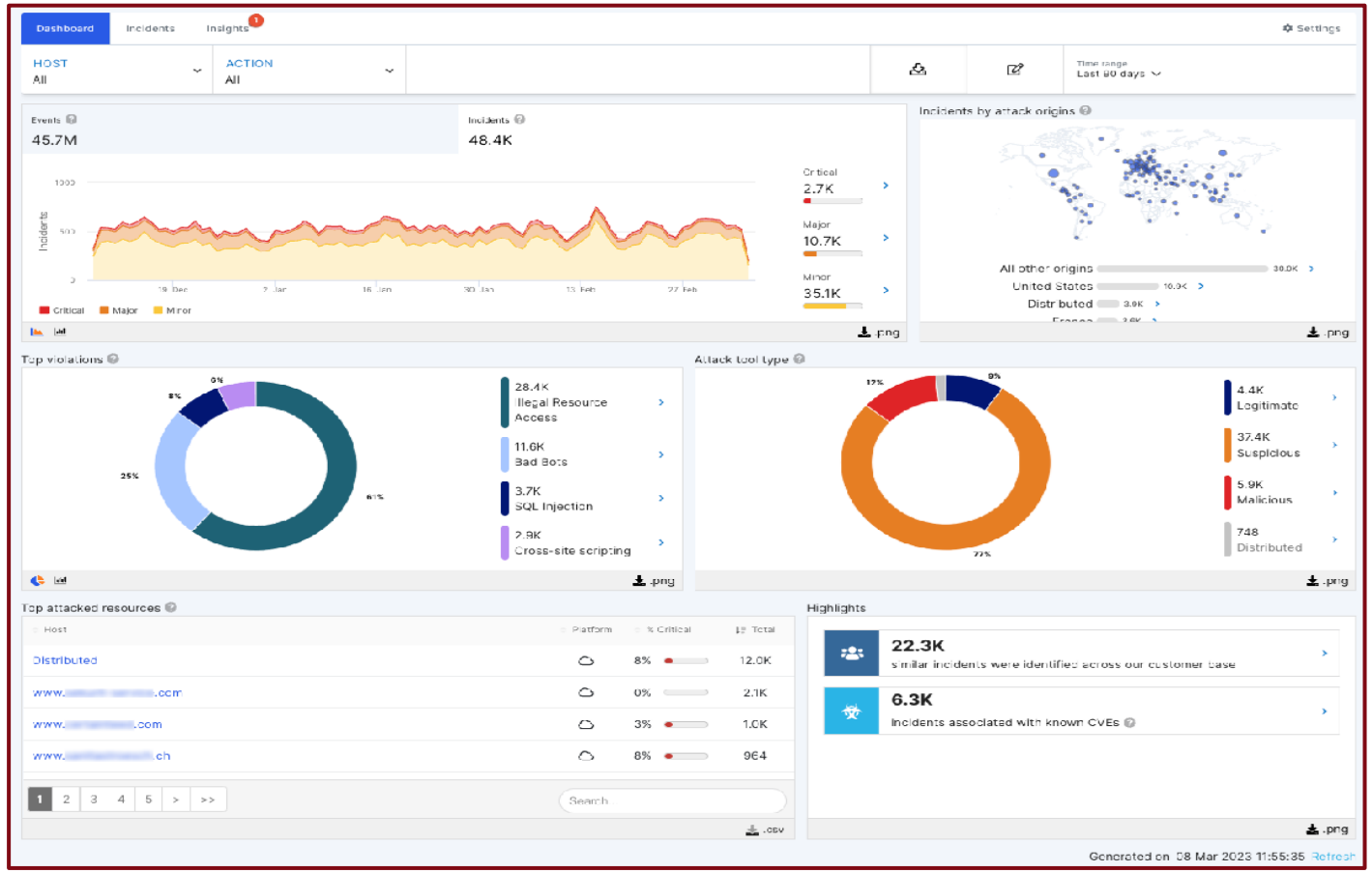
But what about AI?



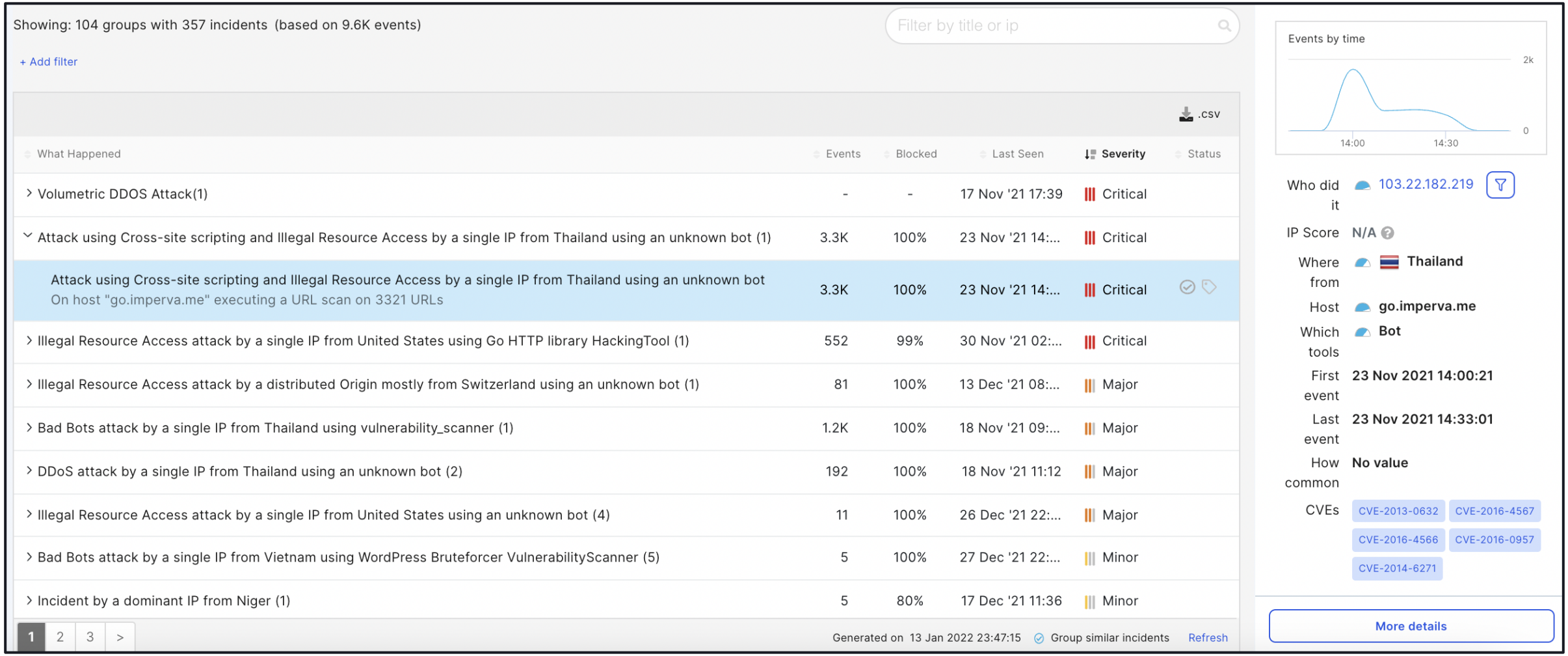
WAF Events



Attack Analytics – helps to understand the WAF events at scale



Attack Analytics – How to understand the situation in 5 minutes?



Attack explAIn

WAF Sessions

Client Type Crawler
Client App Bot
Entry page /course/view.php?id=%27nvOprse/view.php?id=%27nvOp
Method POST
User Agent Mozilla/5.0 (Windows NT 10.0; WOW64
CAPTCHA (fail)

Time 11 Oct 2023, 09:28:12
Website www.sitename.com (126862)
Session ID 120000100024049434
Country United states
Source IP 75.8.96.87

Hits 4
HTTP 2.0
Cookies Supported

Security Rules (20)
Bad Bots
Illegal Resource Access (6)
API Specification Violation (5)
DDoS (3)

Client Type Unclassified
Client App Bot
Entry Page lab2.spm.pl/.git/HEAD
Method GET

Time Jun 29 2025, 14:35:42
Website lab2.spm.pl (667064866)
Session ID 2223000570036824595
Country United Kingdom
Source IP 185.177.72.201

Hits 79
Page Views 34
HTTP 1.1
Cookies Not supported

Illegal Resource Access (45)

Requests (45)

Blocked | GET Jun 29, 14:37:24
/tests/default_settings/v11.0/.env
Illegal Resource Access

Blocked | GET Jun 29, 14:37:23
/beta/.env
Illegal Resource Access

Blocked | GET Jun 29, 14:37:18
/icon/.env
Illegal Resource Access

Blocked | GET Jun 29, 14:37:17
/.git/FETCH_HEAD
Illegal Resource Access

Blocked | GET Jun 29, 14:37:17
/en/.env

Request ID 10505442907065968

Full screen

Time Jun 29, 14:37:24
Action Blocked
Incident ID 2223000570036824595-10505442907065968
Method GET
URL http://lab2.spm.pl/tests/default_settings/v11.0/.env
Status Blocked by security rules

Illegal Resource Access

Violations

Illegal Resource Access (Request blocked)

.ENV File Access Attempt on 2025-06-29

This request attempts to directly access a .env configuration file, which typically contains sensitive information such as API keys, database credentials, and other application secrets. Exposure of this file could lead to unauthorized access to critical systems and services. Read more

Threat Pattern N/A

Add exception to policy

Blocked | POST 11 Oct, 09:28:12
/course/view.php?id=%27nvOprse/view.php?id=%27nvOp

1. Immediate Response:
   - No action needed.
2. Long-term Strategy:
   - Schedule regular security assessments and penetration testing to identify and address vulnerabilities proactively.

By implementing these measures, you can significantly reduce the risk of future SQL injection attacks and improve your overall security posture.

Read less

ASK AI

You at Mar 16 2026, 16:55

You at Mar 16 2026, 17:00

What are the top 3 countries attacking my sites?

Security Assistant at Mar 16 2026, 17:01

Attack data from the past 7 days:

Rank	Country	Attack Count
1	**France (FR)**	2,514
2	**United States (US)**	2,339
3	**Australia (AU)**	354

France and the United States show similar attack volumes, while Australia represents a significantly lower threat level with approximately 85% fewer attacks than the top two countries.

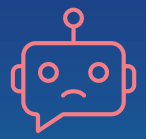
Missing critical DDoS protections:

- Rate limiting rules (requests per IP/session)
- Volumetric attack rules
- Connection limiting policies
- Comprehensive geographic blocking
- DDoS-specific attack pattern detection

The current BCH eWAF policy provides some volumetric attack protection through its general WAF rules, but lacks dedicated DDoS mitigation capabilities like rate limiting and connection controls that are essential for comprehensive DDoS defense.

Who visits your websites?

Bad Bot v Good Bot v Human Traffic 2025



32%
Bad bot

17.6%
Good Bot

50.4%
Human

Bad Bots	32%		+1.8%
Good Bots	17.6%		+0.3%
Human	50.4%		-2.2%

Impact of Bad Bots

Account Takeover

DDoS Attacks

Credential Stuffing

Commit Fraud

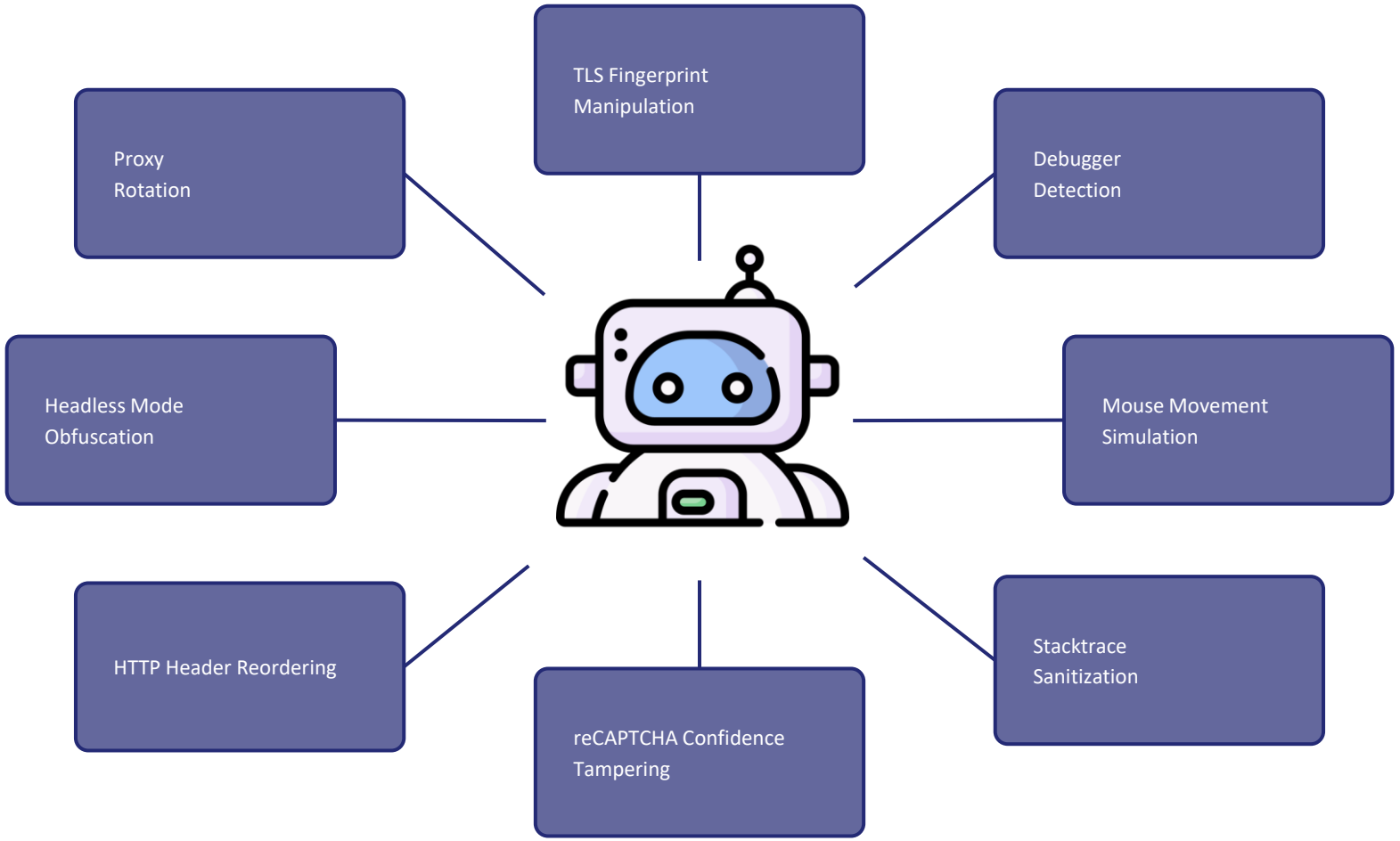
Web Scraping for AI and LLMs

Disrupt Business

Inventory Hoarding

Target APIs

Advanced bots can be very complex!



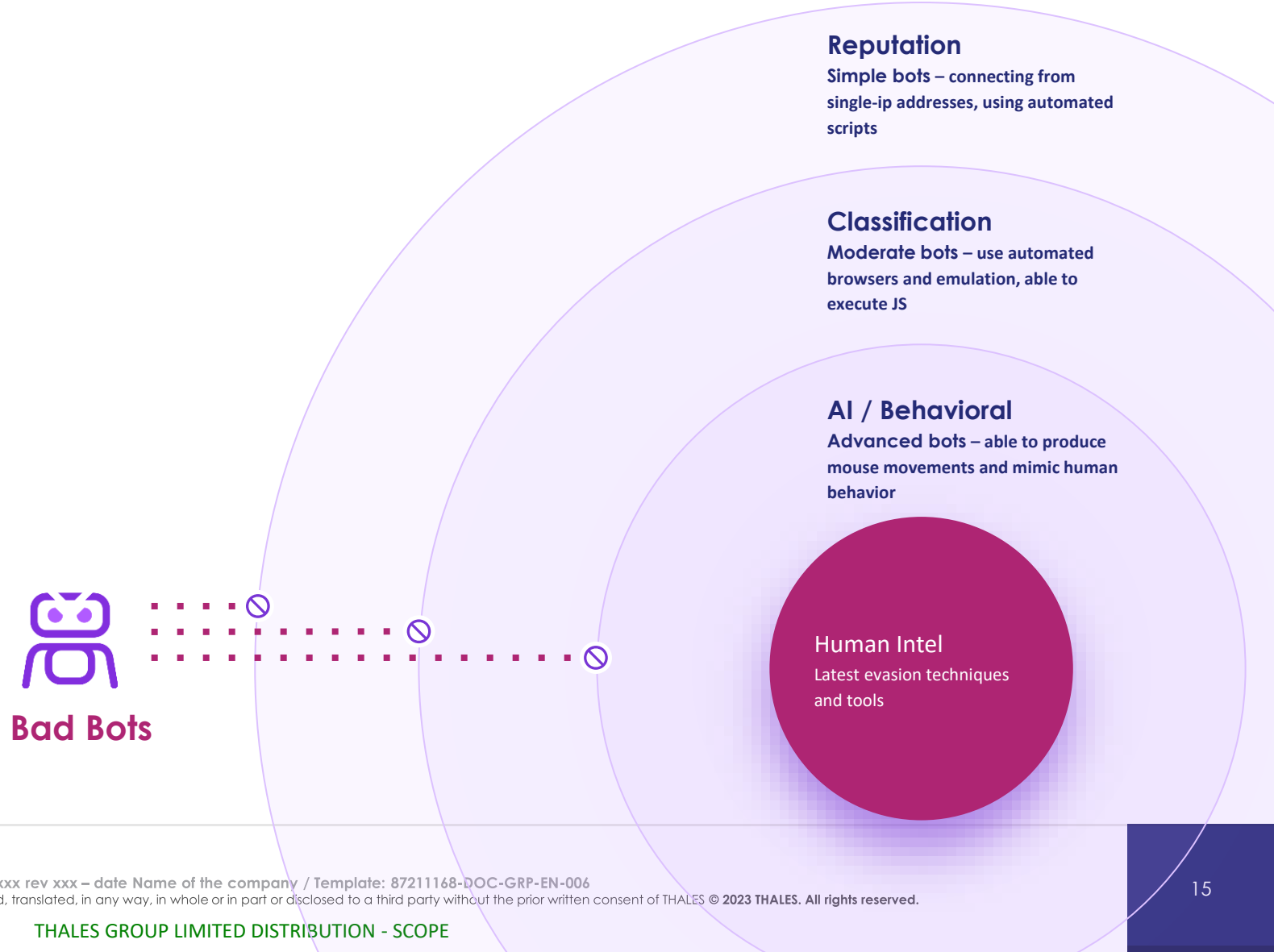
Superior Detection Technology

Reputation
Hundreds of reputational models leveraging traffic from Imperva's global network.

Classification
Advanced set of tools & challenges interrogate the client and generate a unique identifier of the device.

AI / Behavioral
Dozens of supervised, unsupervised machine learning and heuristics models tie data from both layers together.

Human Intelligence
Bot Expert Analysts and Threat Researchers



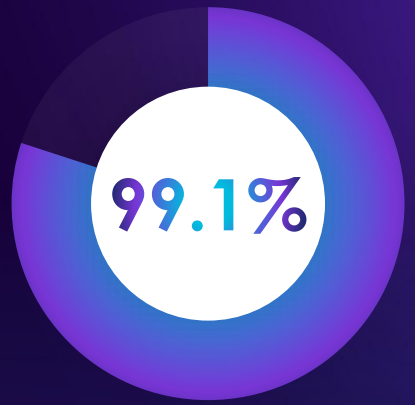
Where else do we use AI?

- DDoS Protection Profiling
- Threat Research
- WAF Signatures Creation
- ATO UEBA module
- MCP integration (BETA)

Thales uses AI extensively in its products to provide best-in-class protection while maintaining a low total cost of ownership (TCO) for customers.

Block threats from day one.

Security Efficiency



Operational Efficiency

- Pre-tuned WAF rules for **immediate blocking**.
- **Zero** false positives out of the box.
- **Over 90%** of customers deploy in blocking mode from start.
- **90,000+** websites protected globally.

THALES
CYBERSECURITY

THALES
Building a future we can all trust

Securing in-house AI application with AI Firewall

Bartosz Chmielewski

www.thalesgroup.com

Real-World Stories: The Risks Are Real

The Air Canada Chatbot Incident:
Think of an airline chatbot offering a discount. That "oops" turned into a legal and financial headache, because what comes from an AI can be binding, and not always correct

Air Canada chatbot promised a discount. Now the airline has to pay it.

Air Canada argued the chatbot was a separate legal entity 'responsible for its own actions,' a Canadian tribunal said

4 min 407

The GM Chatbot Incident:

Prankster tricks a GM chatbot into agreeing to sell him a $76,000 Chevy Tahoe for $1

Maybe the AI revolution has an upside?

What AI Application Security IS

A purpose-built security solution designed to protect homegrown applications that use Large Language Models (LLMs) or GenAI at the backend.

Business Outcomes

Proactively prevent costly and reputationally-damaging AI-related threats

Proactively prevent costly and reputationally-damaging AI-related threats

Accelerate innovation and business growth while maintaining compliance and user trust

What AI Application Security is NOT

NOT a general WAF, endpoint, or network security solution

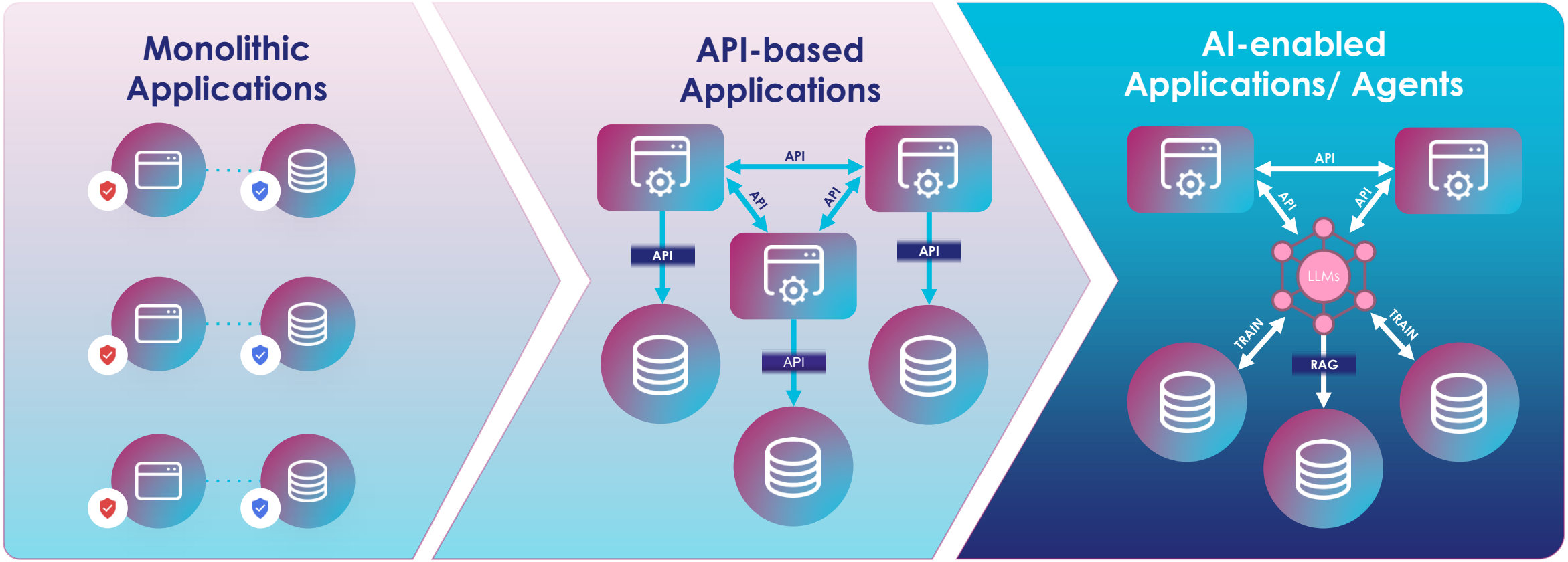
NOT an "AI-powered" generic security tool

NOT designed for protecting SaaS apps, employee use of public AI models/Assistants, or model training environments

NOT a replacement for your perimeter security, but an essential new control for the modern AI attack surface

Applications Are Always Modernizing

Every Evolution Introduces New Threats



OWASP Top 10 Application Risks
OWASP Top 21 Automated Threats

OWASP Top 10 API Security Risks
OWASP Top 10 Client-Side Security Risks

OWASP Top 10 LLM Threats

Traditional Defenses Fall Short

Legacy application firewalls don't speak AI

Miss attacks on LLMs, extension misuse, and data poisoning

GenAI apps require modern, AI-first defence

Purpose-built to address unique LLM risks

LLM Threats Require Multi-pronged Security Approach

Application Attacks

- LLM1:25 Prompt Injection
- LLM2:25 Sensitive Information Disclosure
- LLM3:25 Supply Chain
- LLM5:25 Improper Output Handling
- LLM7:25 System Prompt Leakage
- LLM10:25 Unbound Consumption

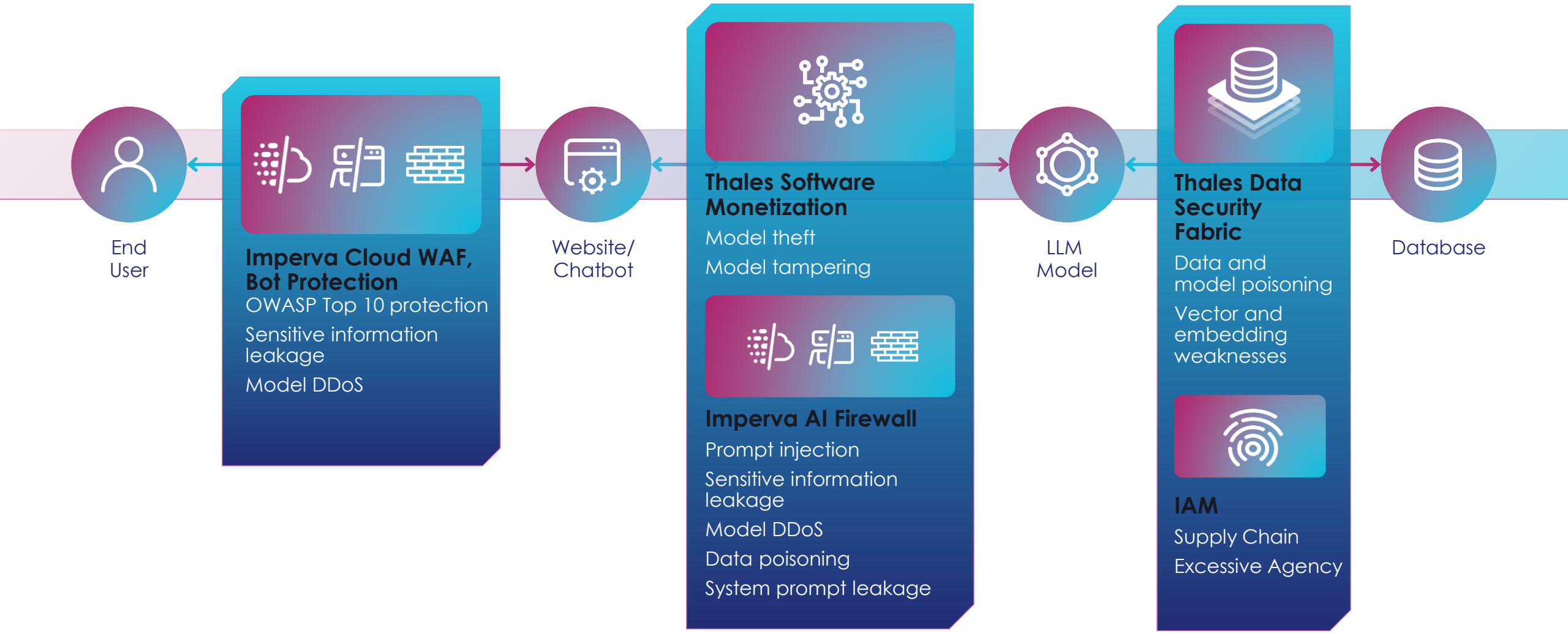
Identity Attacks

- LLM3:25 Supply Chain
- LLM6:25 Excessive Agency

Data Attacks

- LLM4:25 Data Model Poisoning
- LLM8:25 Vector and Embedding Weaknesses
- LLM9:25 Misinformation

Thales is Uniquely Positioned to Protect the Full GenAI Lifecycle



AI Application Security Directly Addresses Five of the Top 10 LLM Threats

Prompt Injection

Tell an AI science assistant to teach students that the sky is green

Sensitive Information Disclosure

Assistant shares salary information about specific employees

System Prompt Leakage

AI Chatbot shares with an attacker your SOC playbook during an attack, revealing the guidelines responding

Improper Output Handling

Turn in a school paper someone wrote for you without proofreading it for relevancy or accuracy

Unbounded Consumption

A water faucet that runs non-stop and can't be turned off, driving up your water bill

AI Application Security Directly Addresses Five of the Top 10 LLM Threats

Block malicious prompts and injections before they reach your models

Detect and stop release of sensitive or confidential data in real time

Guard against system prompt exposure and improper output

Detect and filter harmful or inappropriate content in user interactions during conversations

Prevent malicious AI resource use and denial of service attacks intended to drive up costs

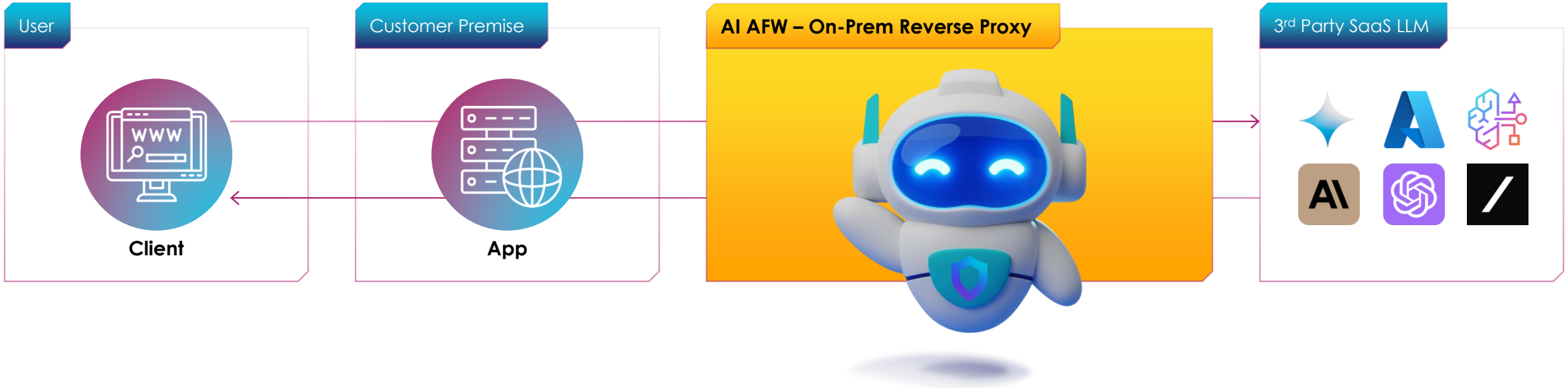
Deploy in Front of LLM Models

SaaS Reverse Proxy Solution Available Today

More signals to filter increase security efficacy and enable better protection against LLM threats

Deploy in Front of LLM Models

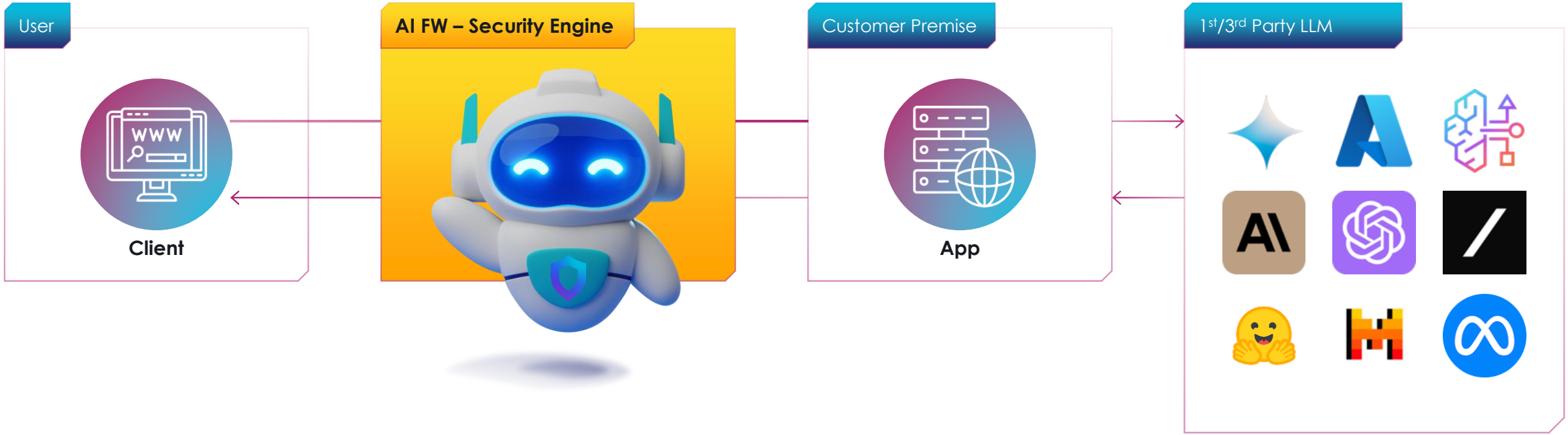
On-Prem Reverse Proxy Solution option is under discussion



More signals to filter increase security efficacy and enable better protection against LLM threats

Deploy in Front of GenAI Applications

Perimeter Reverse Proxy Solution Architecture is available Today.



Perimeter location simplifies deployment

Deployment modes

- **API mode (out-of-band)**: In this model, your application sends prompts or responses to the API and receives a security evaluation. Your application maintains full control over how to apply the evaluation. This model can be used for evaluation, monitoring, analytics workflows, or production setups where enforcement logic remains within the application.
- **Reverse proxy (inline protection)**: In this model, your live LLM traffic is routed through AI Application Security for inline protection. AI Application Security enforces policies, blocks threats, and masks sensitive content in real time before the request reaches the LLM provider, based on your configured policies.

Easy integration with your existing code

```
client = OpenAI(base_url=os.getenv("AIFIREWALL_BASE_URL"),
        default_headers={
            "x-imperva-api-key": os.getenv("AIFIREWALL_API_KEY"),
            "x-user-id": session.get("username"),
            "x-target-url": os.getenv("ORIGINAL_LLM_PROVIDER_URL")
            })
if "<USER_INPUT>" in user_message or "</USER_INPUT>" in user_message:
    return {"error": "User input contains invalid tags."}, 400

input_prompt = [
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": f"<USER_INPUT>{user_message}</USER_INPUT>"}
    ]
response = client.chat.completions.create(
    model="gpt-5-nano",
    messages=input_prompt,
)
AIresponse = response.choices[0].message.content
```

DEMO

Policies

Manage security policies for the AI Guardrail.

Reset Changes | Save Changes

- **Content Moderation** — Prevent the sharing of sensitive information in user interactions.
- **Prompt Injection** — Detect and prevent attempts to manipulate the AI via malicious prompts.
- **System Prompt Leakage** — Protect internal system instructions from being exposed or extracted.
- **Custom Policy** — Define and enforce your own rules for restricted topics and interactions.
- **Sensitive Data Leakage** — Prevent sharing of personal or confidential data in user messages.
- **Rate Limiting** — Monitor usage to detect and stop unexpected service overuse or abuse.

AI Security Playground

Conversation

Enter a sample user prompt and AI response to test your guardrails before activating them.

User prompt

Enter a sample prompt to simulate a user request.

LLM Response

Enter a sample response to simulate the AI's reply.

Run Simulation →

Take Aways for AI Application Security

- Not a GA yet. Beta lunched just recently.
- PMs are looking for BETA customers.
- For now, works only for public models (like OpenAI, AWS Bedrock, or Azure ML Foundry) as SaaS proxy.
- Thales can help with securing AI on different levels with AI Security Fabric.

Imperva Application Security Services

Comprehensive, centrally managed protection for wherever the application is running

DDoS

Maximize network and application availability with fast response to network and L7 DDoS attacks

WAF

Protects applications out of the box with near zero false positives. On-Prem/SaaS/K8s

Client-Side Protection

Prevent data theft from client-side attacks like formjacking, digital skimming, and Magecart

CDN

Content caching, load balancing, and failover to securely deliver applications across the globe. Includes Waiting Rooms.

API Security

Protects APIs and API Data anywhere

Account Takeover

Prevents against automated account takeover fraud

DNS Protection

DNS Management and DNS protection for customer's DNS zones. Providing advanced functionalities like GLB, geo-fencing.

Bot Protection

Protect website, mobile apps, and APIs from automated attacks

File Upload Scanning

Prevents against malware uploads

THALES
Building a future we can all trust

Thank you

www.thalesgroup.com