TLD Data Analysis

ICANN Tech Day, Dublin

October 19th 2015

Maarten Wullink, SIDN

SIDN

- Domain name registry for .nl ccTLD
- > 5,6 million domain names
- 2,46 million domain names secured with DNSSEC
- SIDN Labs is the R&D team of SIDN

DNS Data @SIDN

- > 3.1 million distinct resolvers
- > 1.3 billion query's daily
- > 300 GB of PCAP data daily

ENTRADA

ENhanced Top-Level Domain Resilience through Advanced Data Analysis

- **Goal**: data-driven improved security & stability of .nl and the Internet at large
- **Problem**: Existing solutions for analyzing network data do not work well with large datasets and have limited analytical capabilities.
- **Main requirement**: high-performance, near real-time data warehouse
- **Approach**: avoid expensive pcap analysis:
  - Convert pcap data to a performance-optimized format (key)
  - Perform analysis with tools/engines that leverage that

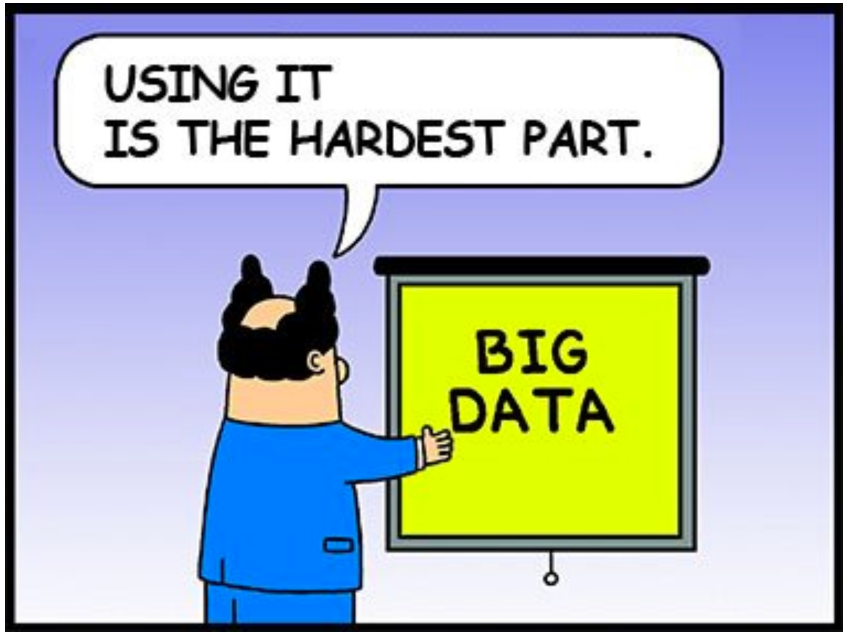
Use Cases

Focussed on increasing the security and stability of .nl

- Visualize DNS patterns (visualize traffic patterns for phishing domain names)
- Detect botnet infections
- Real-time Phishing detection
- Statistics (stats.sidnlabs.nl)
- Scientific research (collaboration with Dutch Universities)
- Operational support for DNS operators

Example Applications

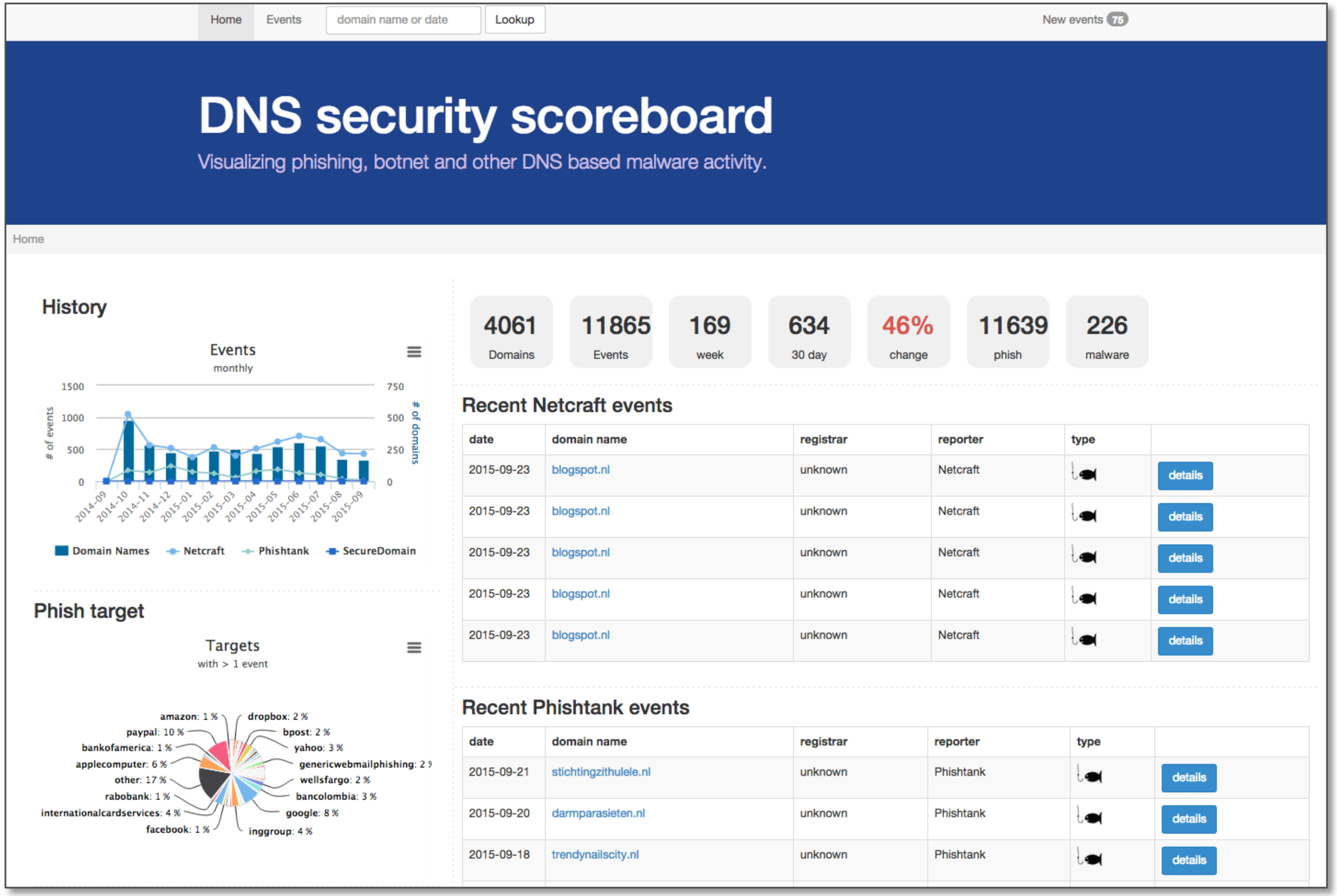
- DNS security scoreboard
- Resolver reputation



DNS Security Scoreboard

Goal: Visualize DNS patterns for malicious activity

How: Combine external phishing feeds with DNS data



DNS security scoreboard

Visualizing phishing, botnet and other DNS based malware activity.

History

Phish target

4061	11865	169	634	46%	11639	226
Domains	Events	week	30 day	change	phish	malware

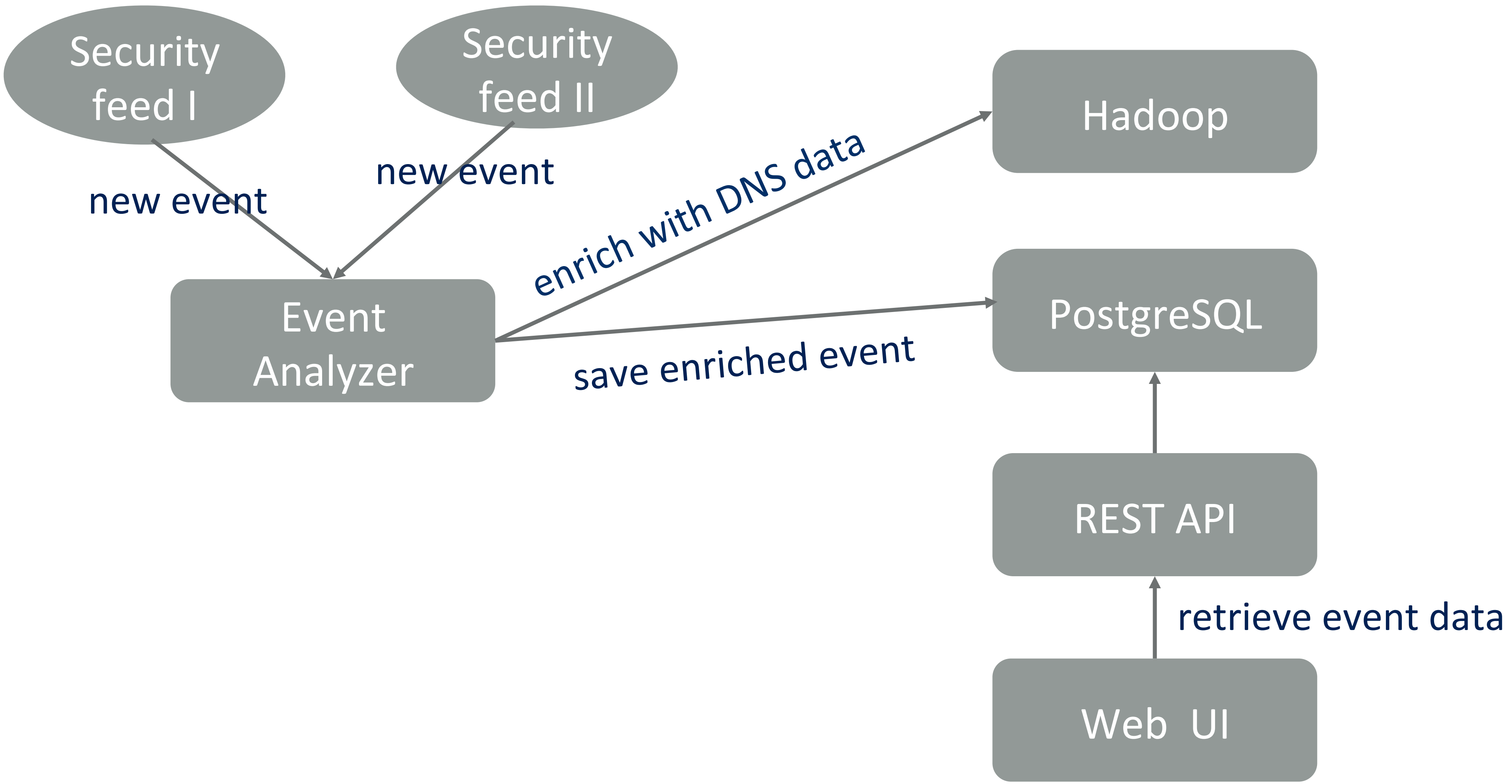
Recent Netcraft events

date	domain name	registrar	reporter	type	
2015-09-23	blogspot.nl	unknown	Netcraft		details
2015-09-23	blogspot.nl	unknown	Netcraft		details
2015-09-23	blogspot.nl	unknown	Netcraft		details
2015-09-23	blogspot.nl	unknown	Netcraft		details
2015-09-23	blogspot.nl	unknown	Netcraft		details

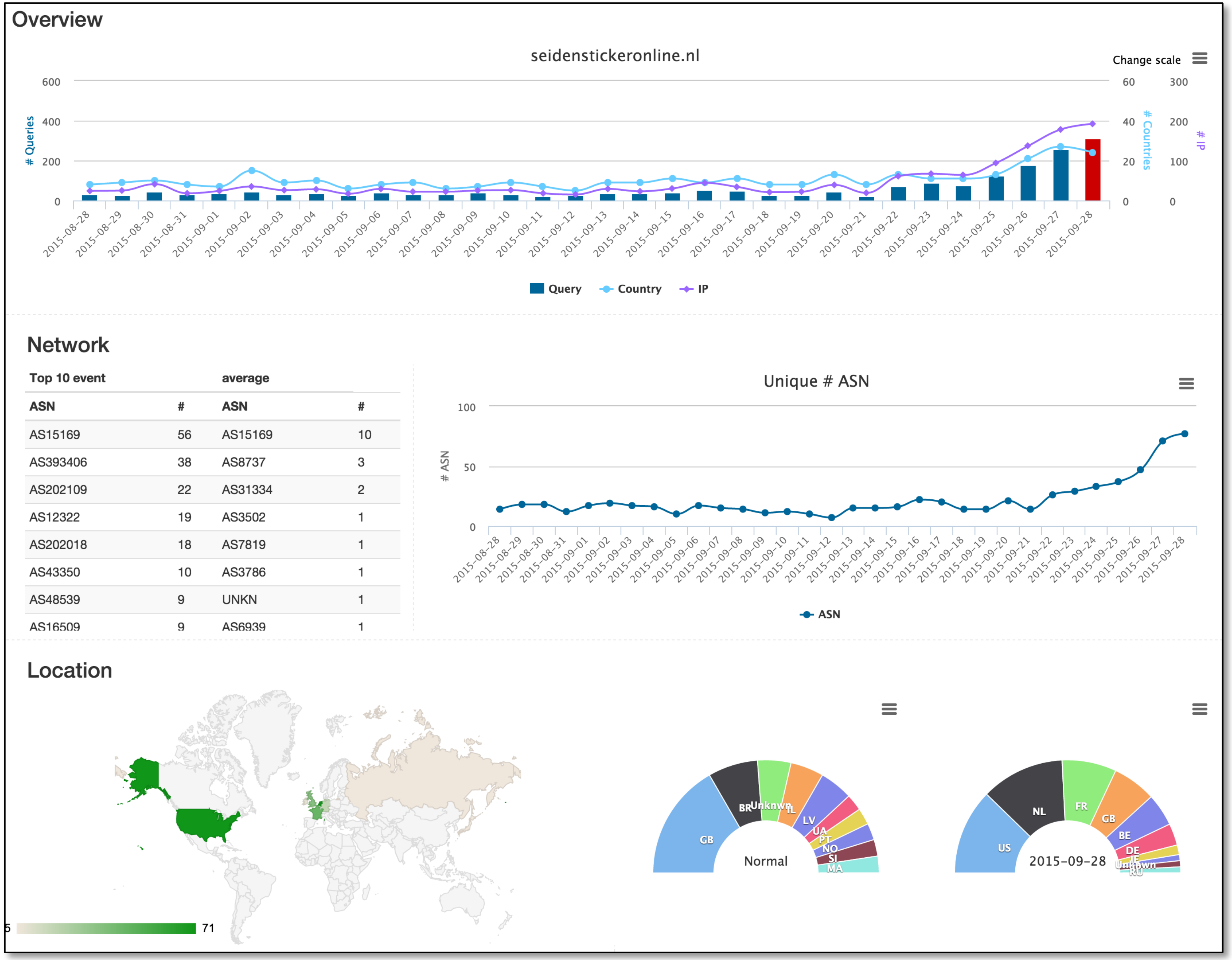
Recent Phishtank events

date	domain name	registrar	reporter	type	
2015-09-21	stichtingzithulele.nl	unknown	Phishtank		details
2015-09-20	darmparasieten.nl	unknown	Phishtank		details
2015-09-18	trendynailscity.nl	unknown	Phishtank		details

Architecture



Traffic Visualization



Network

Top 10 event		average	
ASN	#	ASN	#
AS15169	56	AS15169	10
AS393406	38	AS8737	3
AS202109	22	AS31334	2
AS12322	19	AS3502	1
AS202018	18	AS7819	1
AS43350	10	AS3786	1
AS48539	9	UNKN	1
AS16509	9	AS6939	1

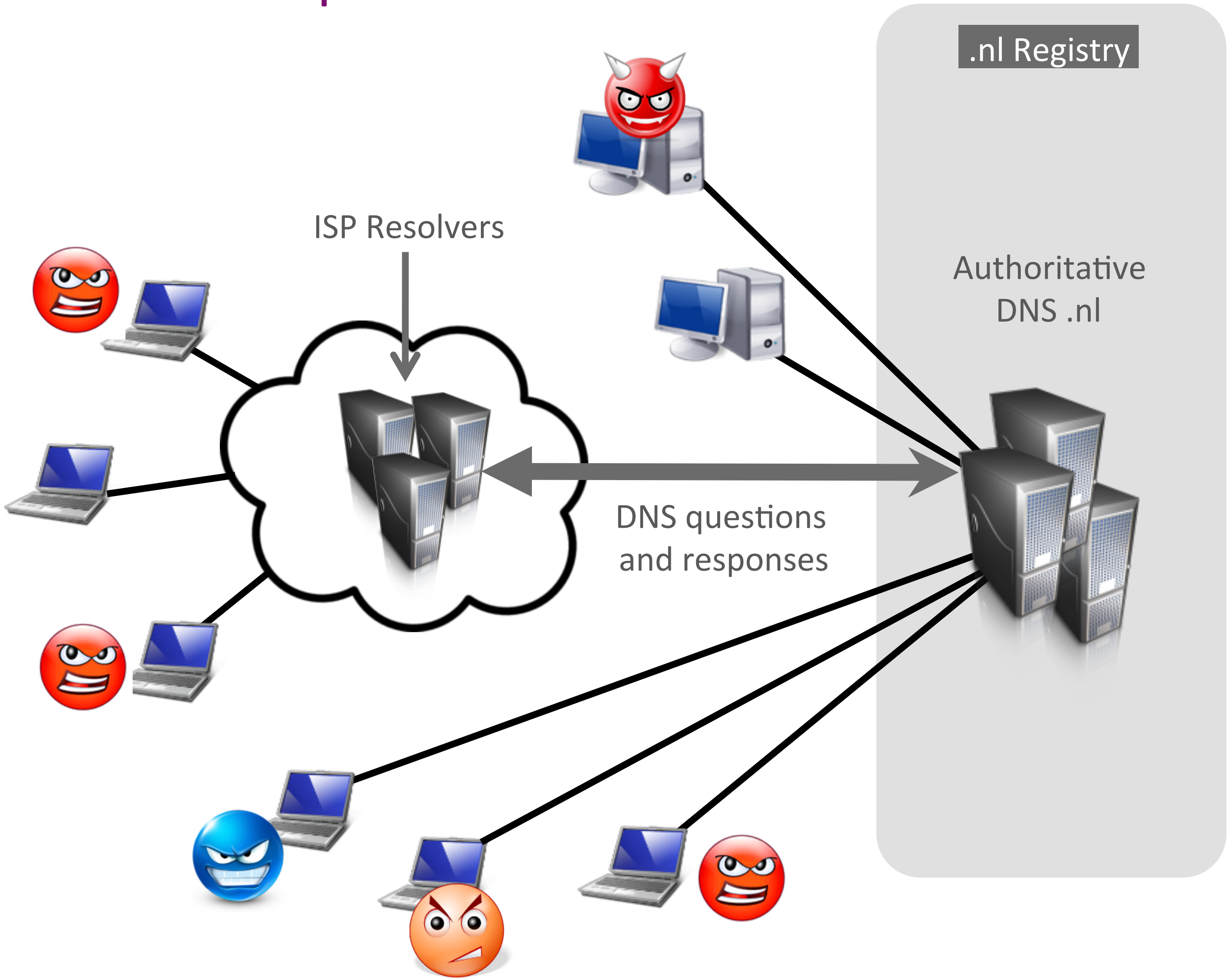
Location

Resolver Reputation (RESREP)

Goal: Try to detect malicious activity by assigning reputation scores to resolvers

How: “fingerprinting” resolver behaviour

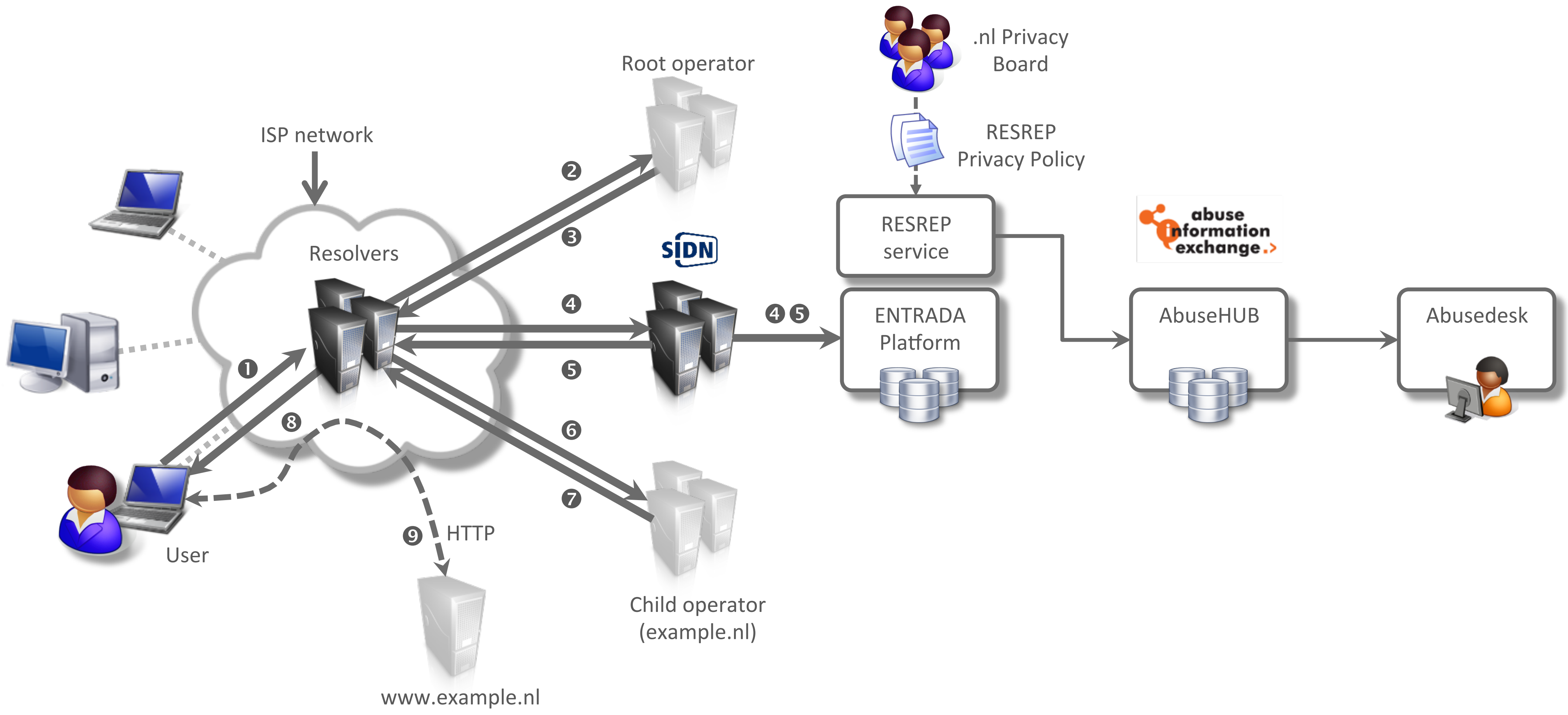
RESREP Concept



Malicious activity:

- Spam-runs
- Botnets like Cutwail
- DNS-amplification attacks

RESREP Architecture



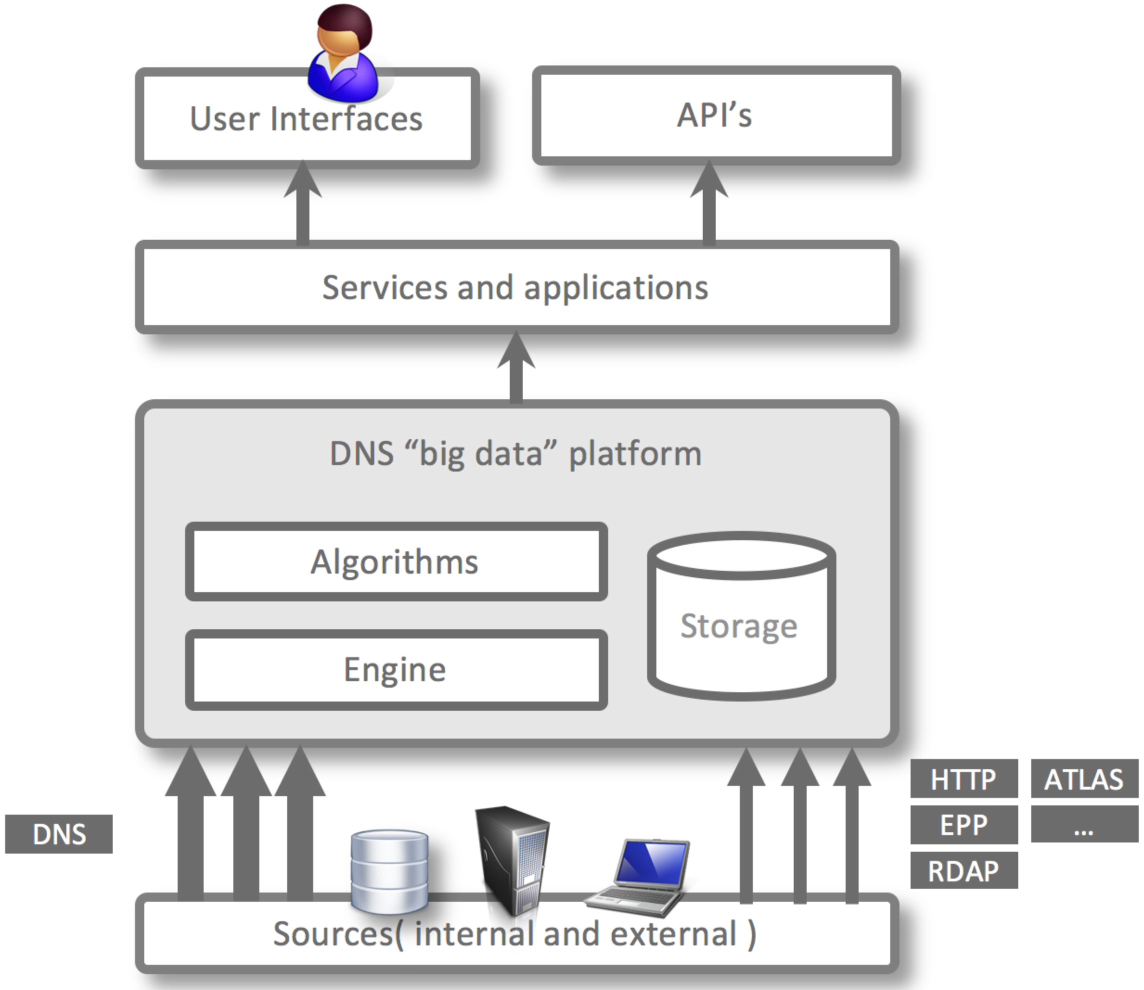
SIDN labs Internet Research & Innovation SIDN

ENTRADA Architecture

- ‘DNS big data’ system

- Goal: develop applications and services that further enhance the security and stability of .nl, the DNS, and the Internet at large

- ENTRADA main components
  - Applications and services
  - Platform and data sources
  - Privacy framework
  - Platform + privacy framework = ENTRADA plumbing



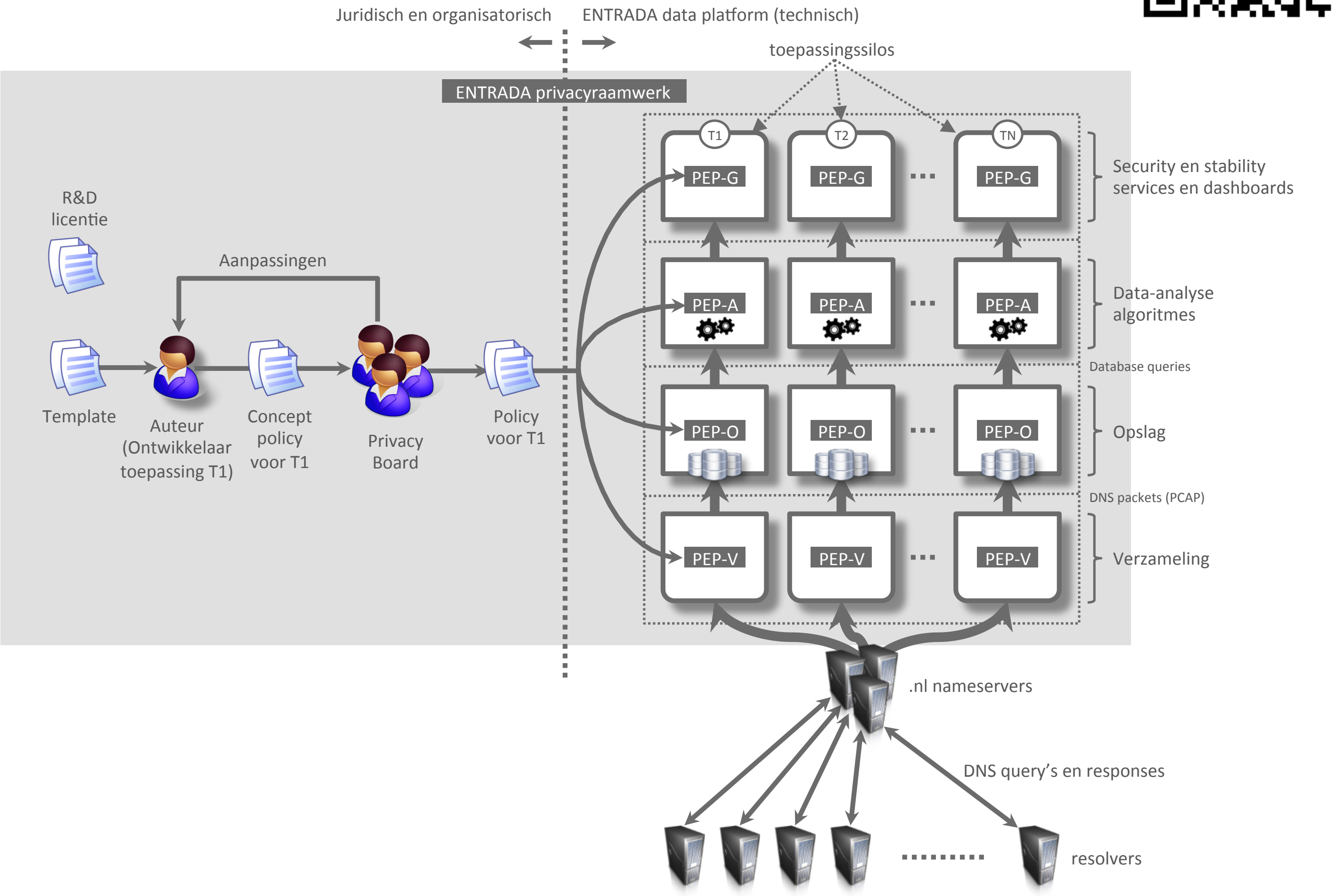
SIDN labs
Internet Research & Innovation
SIDN

ENTRADA Privacy Framework

- Part of the “ENTRADA plumbing”

- Key concepts
  - Application-specific privacy policy
  - Privacy Board
  - Enforcement Points

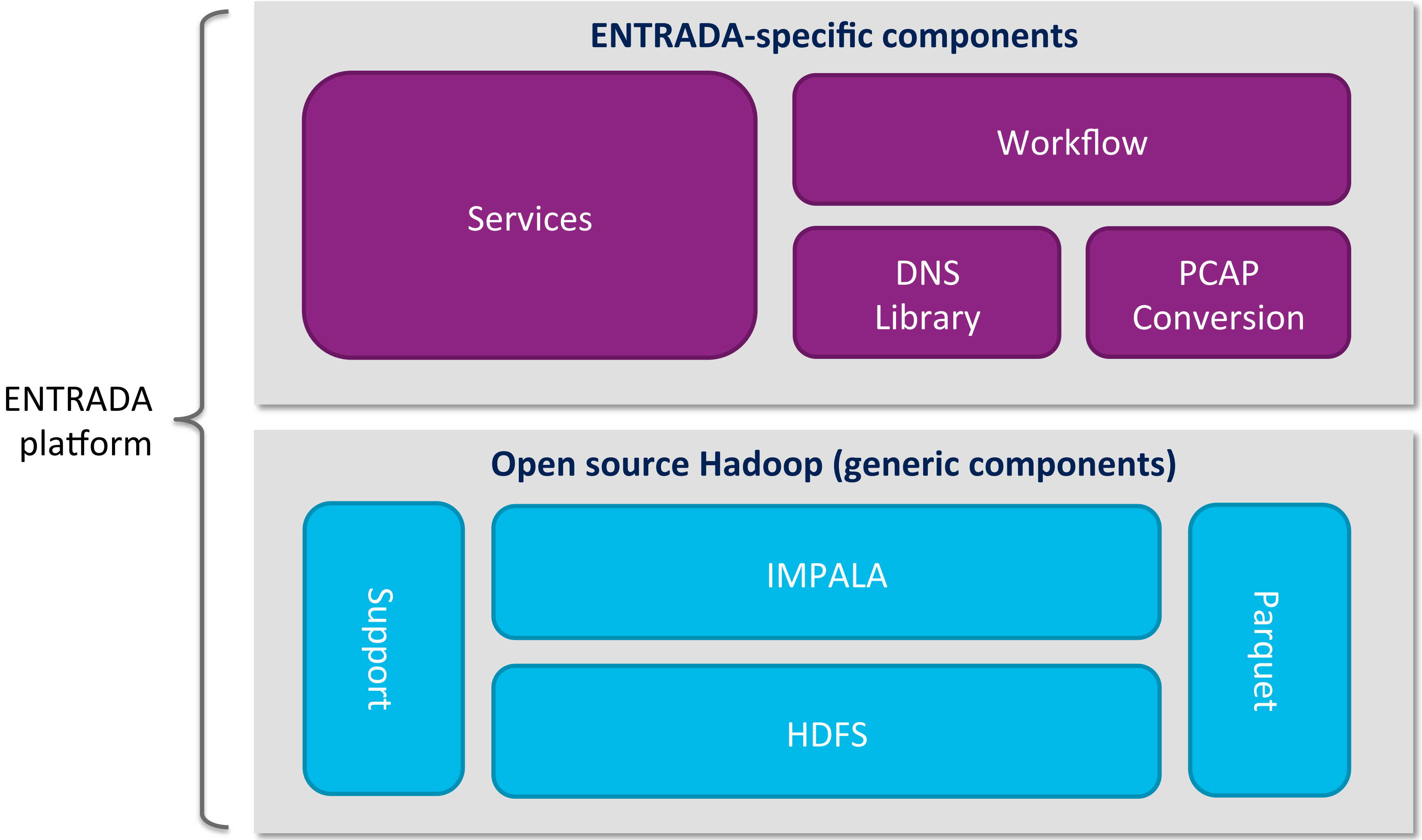
- Policy elements include
  - Purpose
  - Data used
  - Filters
  - Retention period
  - Type of application (R&D vs. production)



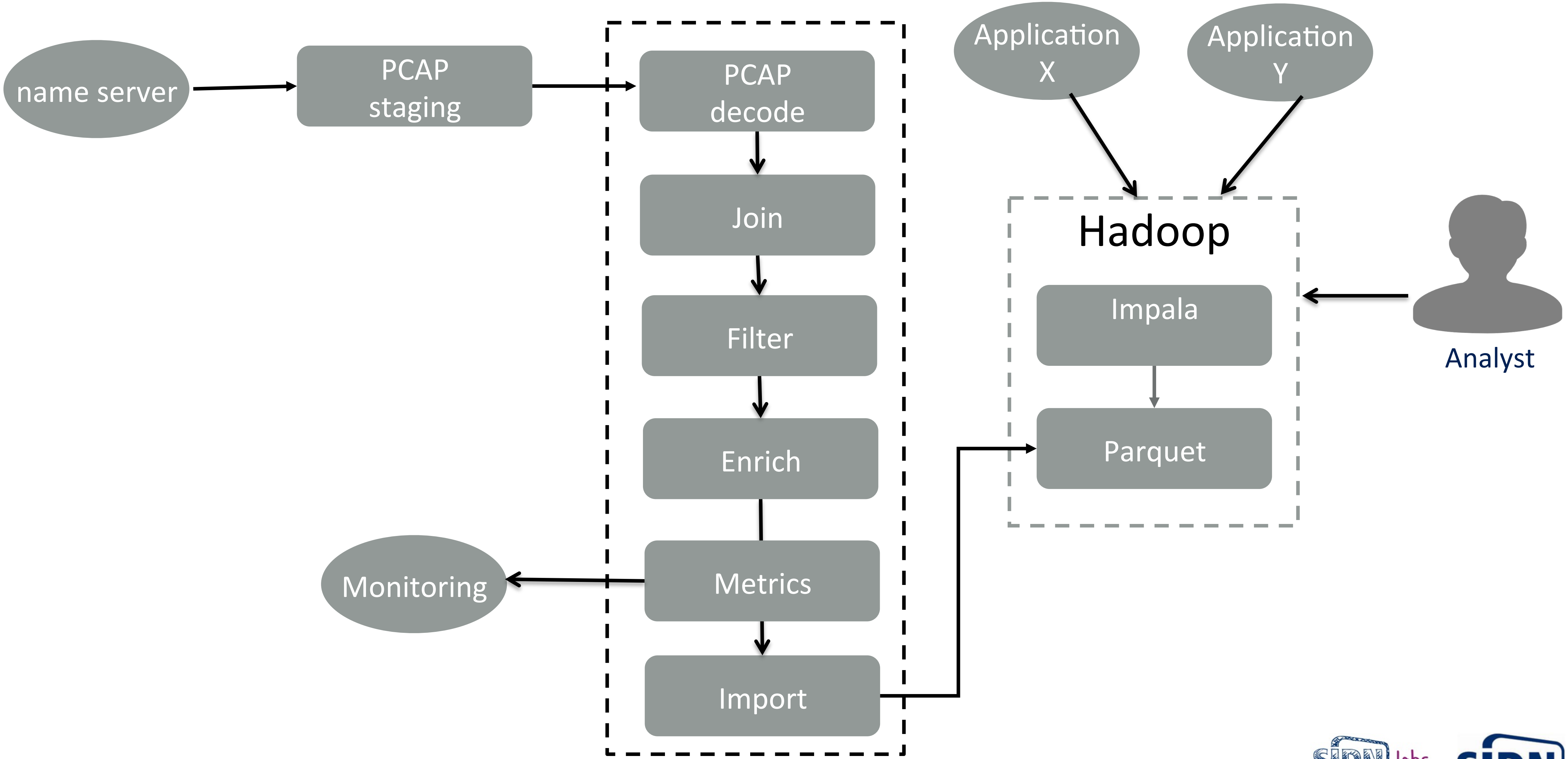
SIDN labs Internet Research & Innovation

SIDN

ENTRADA Technical Architecture



Workflow

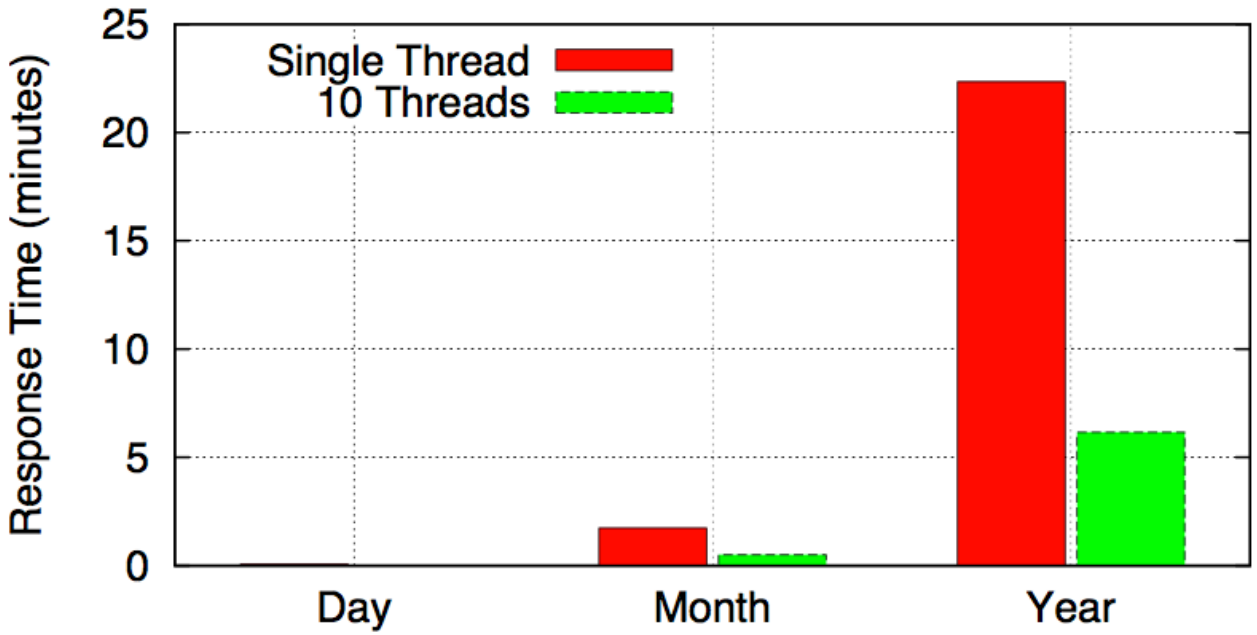


Query data available for analysis within 10 minutes

Performance

Example query, count # ipv4 queries per day.

```
select
concat_ws(’-’,day,month,year),
count(1)
from dns.queries
where ipv=4
group by
concat_ws(’-’,day,month,year)
```



Query response times

1 Year of data is 2.2TB Parquet ~ 52TB of PCAP

ENTRADA Status

Name server feeds	2
Queries per day	~320M
Daily PCAP volume(gzipped)	~70GB
Daily Parquet volume	~14GB
Months operational	18
Total # queries stored	> 74B
Total Parquet volume	> 3TB
HDFS (3x replication)	> 9TB
Cluster capacity	~150B-200B tuples

Conclusions

Technical:

- Hadoop HDFS + Parquet + Impala is a winning combination!

Contributions:

- Research by SIDN Labs and universities
- Identified malicious domain names and botnets
- External data feed to the Abuse Information Exchange
- Insight into DNS query data

Future Work

- Combine data from .nl authoritative name server with scans of the complete .nl zone and ISP data.

- Get data from more name servers and resolvers

- Expand Open Data program

Questions and Feedback

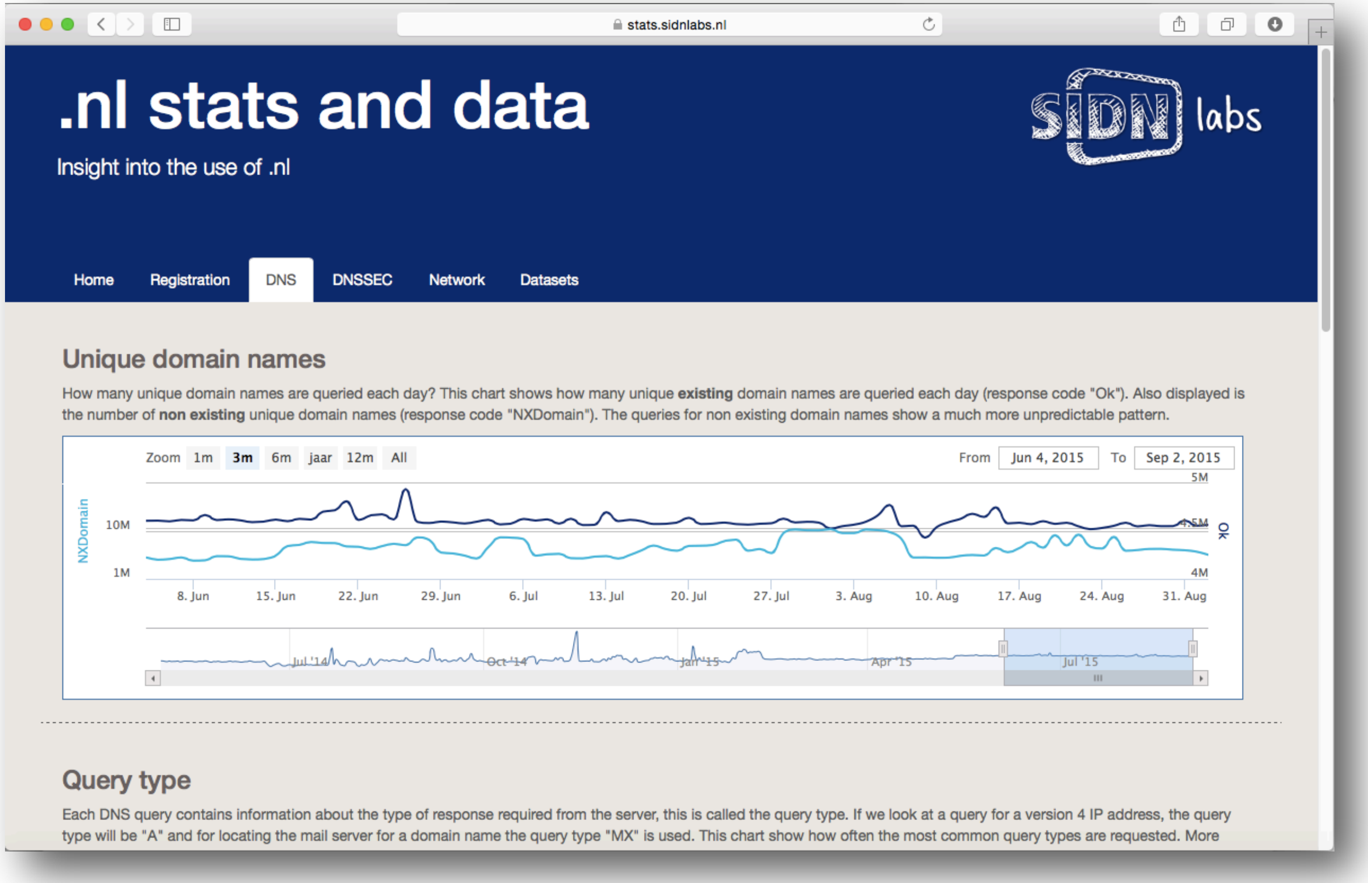
Maarten Wullink

Senior Research Engineer

maarten.wullink@sidn.nl

@wulliak

www.sidnlabs.nl



.nl stats and data

Insight into the use of .nl

Unique domain names

How many unique domain names are queried each day? This chart shows how many unique **existing** domain names are queried each day (response code "Ok"). Also displayed is the number of **non existing** unique domain names (response code "NXDomain"). The queries for non existing domain names show a much more unpredictable pattern.

Query type

Each DNS query contains information about the type of response required from the server, this is called the query type. If we look at a query for a version 4 IP address, the query type will be "A" and for locating the mail server for a domain name the query type "MX" is used. This chart show how often the most common query types are requested. More

https://stats.sidnlabs.nl

SIDN labs
Internet Research & Innovation

SIDN