

ENTRADA: Enabling DNS Big Data Applications

Moritz Müller | SDNRG @ IETF97 | Soul, South Korea

2016-11-14



What if...

You have 100 TB or more of pcap data?

You want to:

1. Store it efficiently
2. Query it efficiently (interactive response times)
3. Test a large number of hypothesis on your data
4. Continuously keep adding new data

You could...

1. Convert it to text format like csv
2. Hadoop MapReduce jobs on csv/pcap
3. Store it in a RDBMS
4. ...

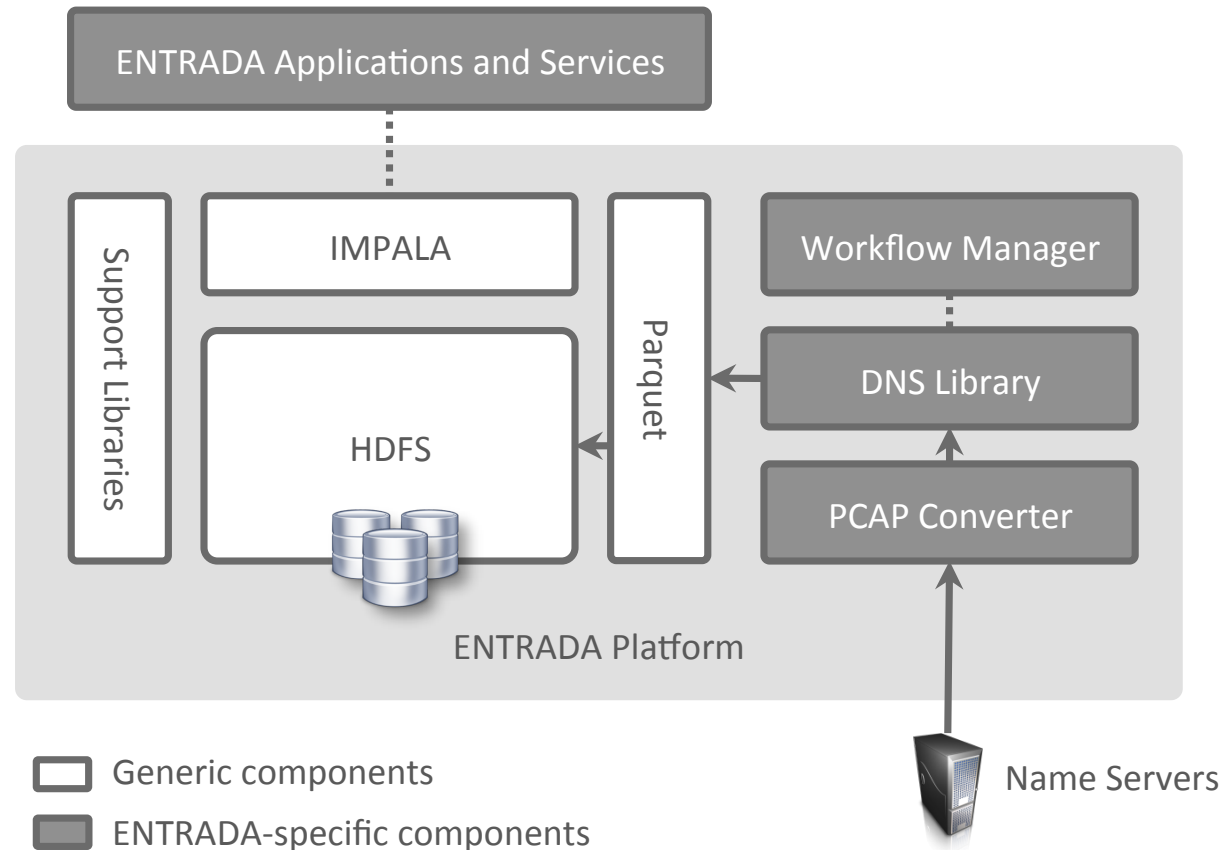
With most options it will be hard to scale and deliver interactive response times

What to do?

- Build your own data streaming warehouse (DSW)
- ENTRADA is our open source DSW (**entrada.sidnlabs.nl**)
- Analyze 50TB pcap data equiv in under 3.5 minutes with a small 4 node cluster
- Our use case: network (DNS, TCP/IP, ICMP) analytics
 - But extensible to other protocols

ENTRADA

ENhanced Top-Level Domain Resilience through Advanced Data Analysis



ENTRADA@SIDN

- We are a TLD registry
- Use it to increase security and stability
- Operational for 2 years
- Capturing data for .nl name servers
- 150 Billion rows (DNS query+response pairs), 21 TB of data

Use Cases

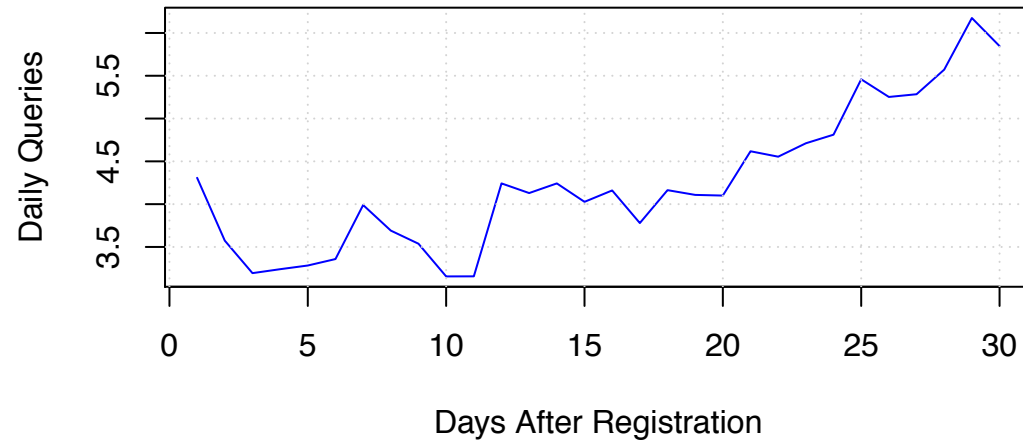
Focus on increasing the security and stability of the DNS

- Statistics (stats.sidnlabs.nl)
- Scientific research
- Support for DNS operators
- **Malicious domain detection**
- **Botnet infection detection**

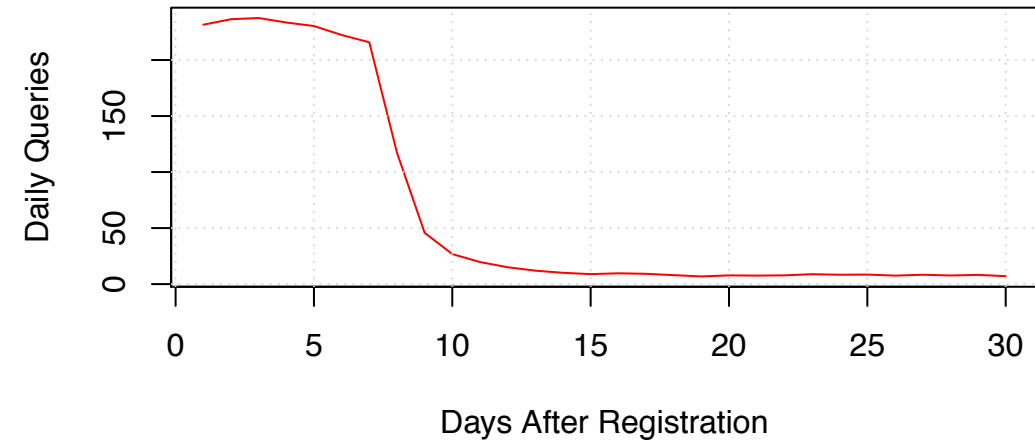
Malicious Domain Detection 1/2

Observation: Phishing domains have unique query patterns

Random Sample Jan--Mar, 2015



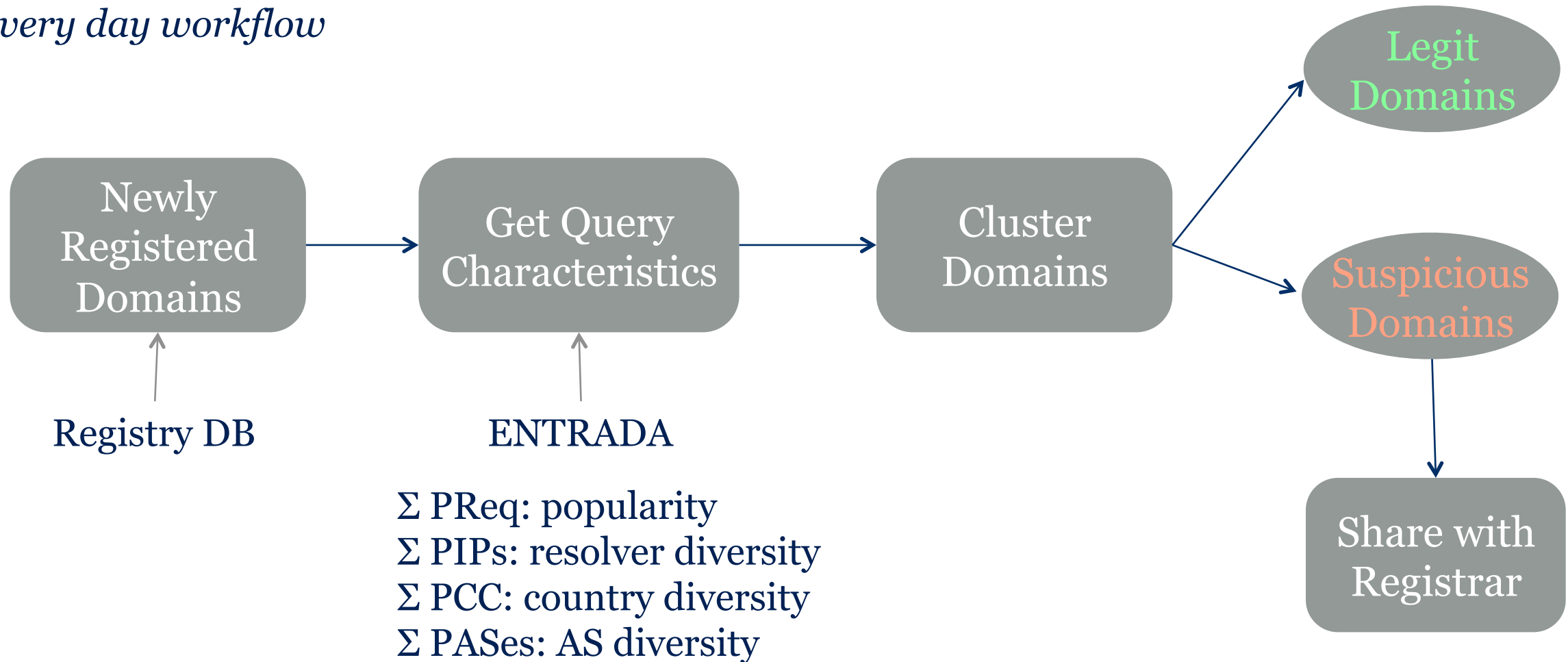
Phishing



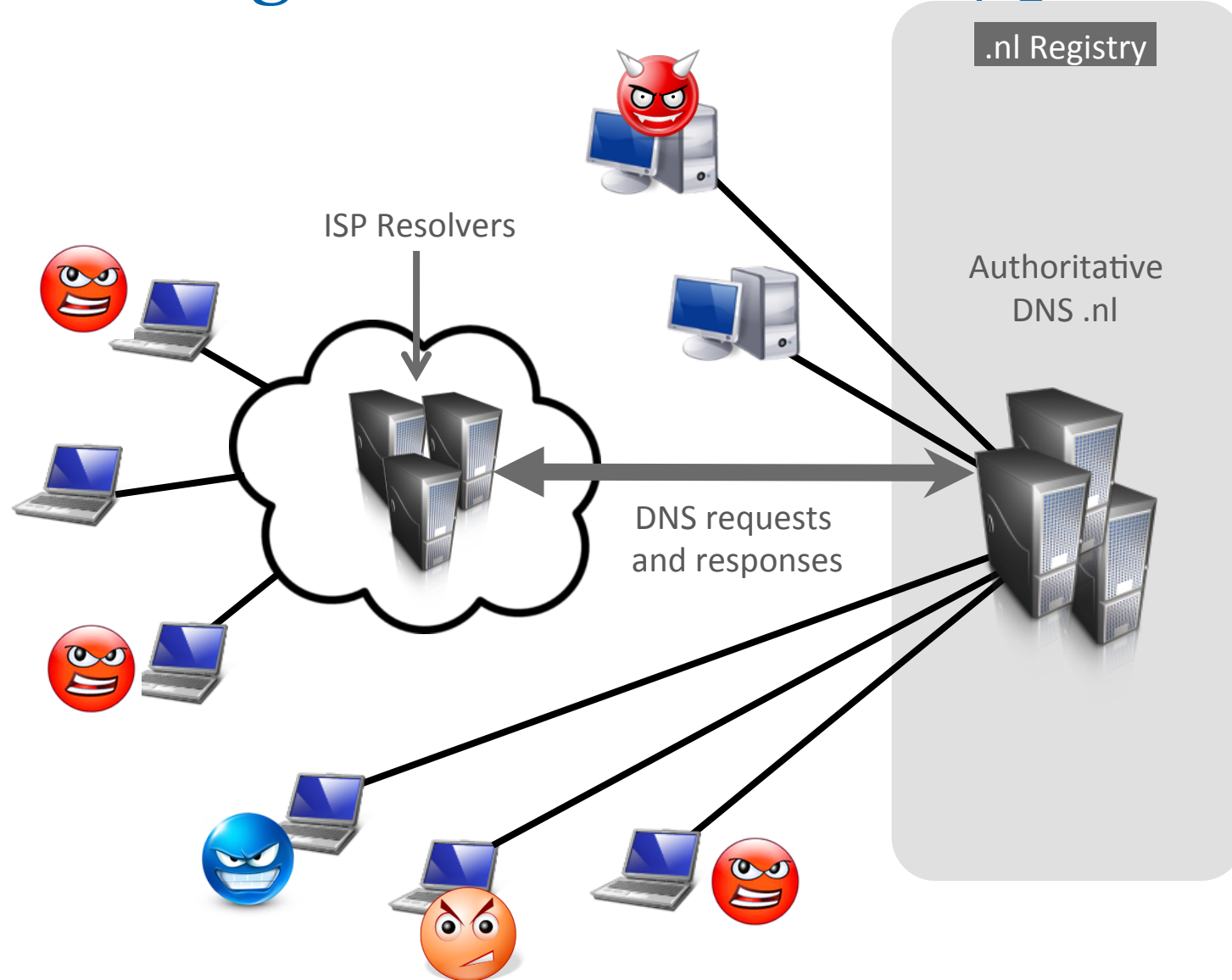
nDEWS: a New Domains Early Warning System for TLDs G. Moura / M. Müller / M. Wullink / C. Hesselman. IEEE/IFIP International Workshop on Analytics for Network and Service Management (AnNet 2016)

nDEWS Architecture

Every day workflow



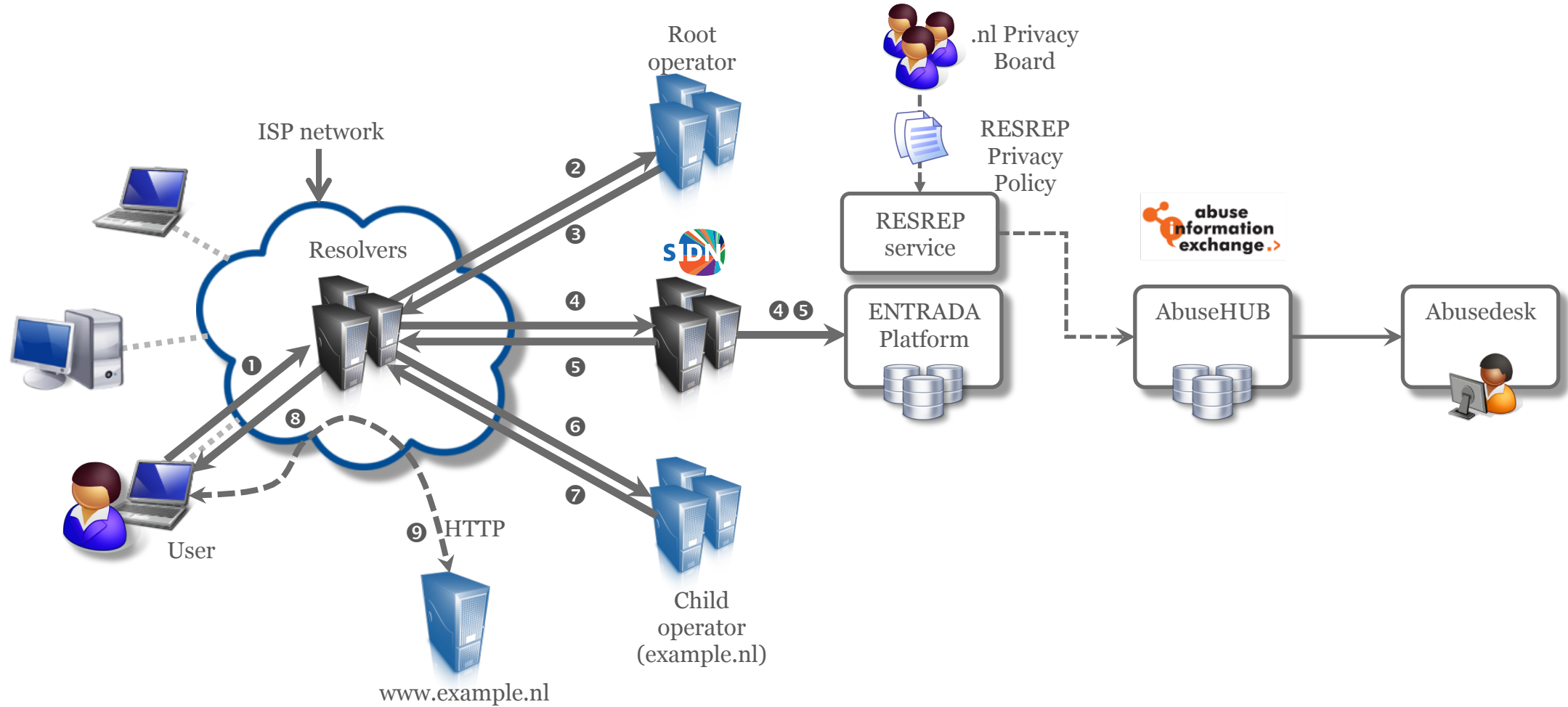
Detecting botnet infections 1/3



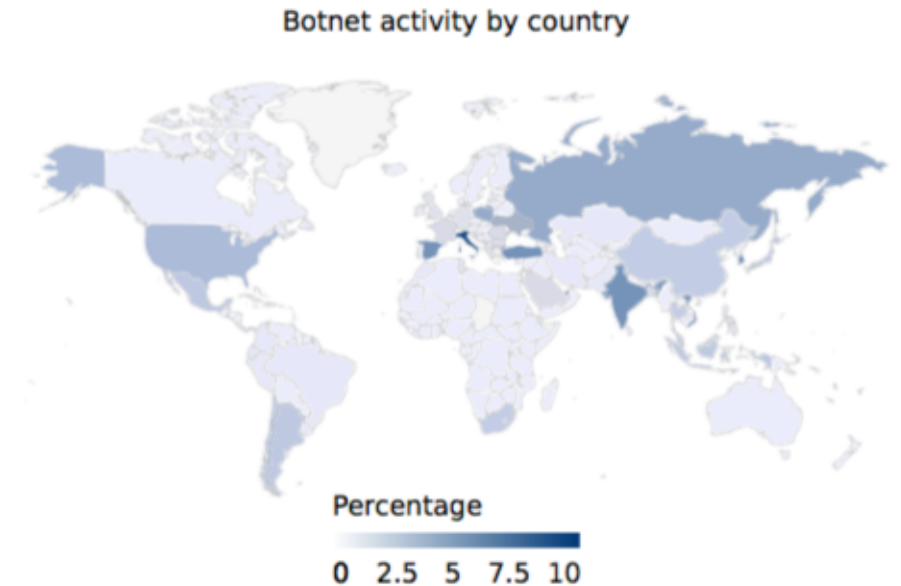
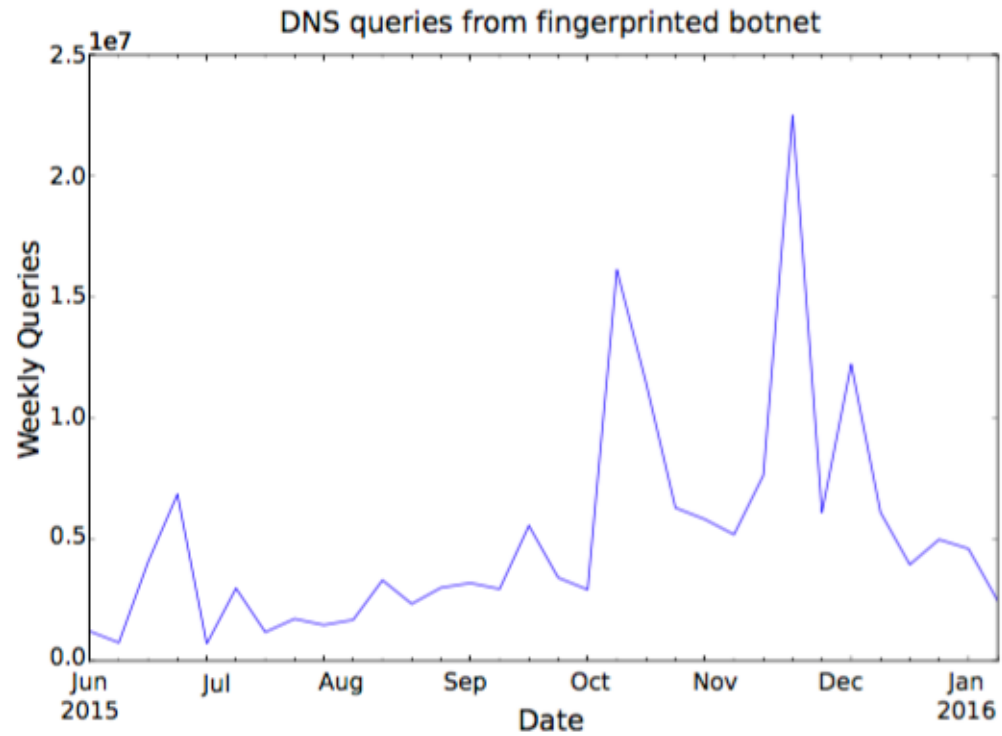
Malicious activity:

- Spam-runs
- Botnets
- DNS-amplification attacks

Detecting botnet infections 2/3



Detecting botnet infections 3/3



Summary

- We have shown ENTRADA, a DSW built using open-source big data tools
- It enables quick hypothesis testing and application development using SQL.
- We have shown some example use cases, which can be easily extended
- Download it and contribute to it.

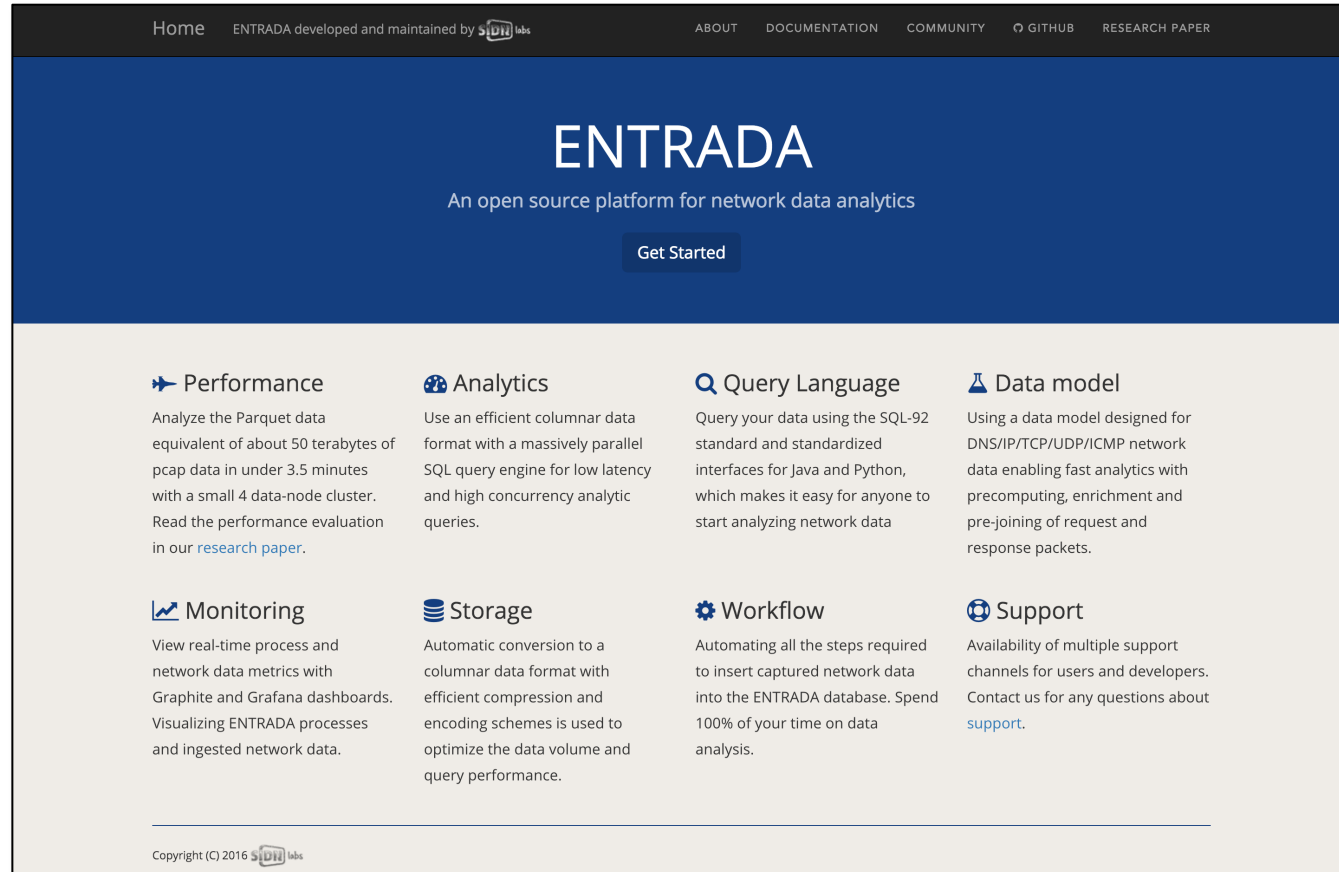
Questions?

Moritz Müller
Research Engineer


moritz.muller@sidn.nl

 @dhr_moe

www.sidnlabs.nl











The screenshot shows the homepage of the ENTRADA project. The header is dark blue with the project name 'ENTRADA' in white. Below the header, there is a navigation bar with links for Home, ABOUT, DOCUMENTATION, COMMUNITY, GITHUB, and RESEARCH PAPER. The main content area is light blue and features a large 'ENTRADA' title and a subtitle 'An open source platform for network data analytics'. A 'Get Started' button is prominently displayed. The page is organized into a grid of feature cards, each with an icon and a brief description. The footer contains the copyright information for 2016 SIDN Labs.


Home ENTRADA developed and maintained by  SIDN Labs ABOUT DOCUMENTATION COMMUNITY GITHUB RESEARCH PAPER

ENTRADA

An open source platform for network data analytics

[Get Started](#)

-  **Performance**
Analyze the Parquet data equivalent of about 50 terabytes of pcap data in under 3.5 minutes with a small 4 data-node cluster. Read the performance evaluation in our [research paper](#).
-  **Analytics**
Use an efficient columnar data format with a massively parallel SQL query engine for low latency and high concurrency analytic queries.
-  **Query Language**
Query your data using the SQL-92 standard and standardized interfaces for Java and Python, which makes it easy for anyone to start analyzing network data
-  **Data model**
Using a data model designed for DNS/IP/TCP/UDP/ICMP network data enabling fast analytics with precomputing, enrichment and pre-joining of request and response packets.
-  **Monitoring**
View real-time process and network data metrics with Graphite and Grafana dashboards. Visualizing ENTRADA processes and ingested network data.
-  **Storage**
Automatic conversion to a columnar data format with efficient compression and encoding schemes is used to optimize the data volume and query performance.
-  **Workflow**
Automating all the steps required to insert captured network data into the ENTRADA database. Spend 100% of your time on data analysis.
-  **Support**
Availability of multiple support channels for users and developers. Contact us for any questions about [support](#).

Copyright (C) 2016  SIDN Labs

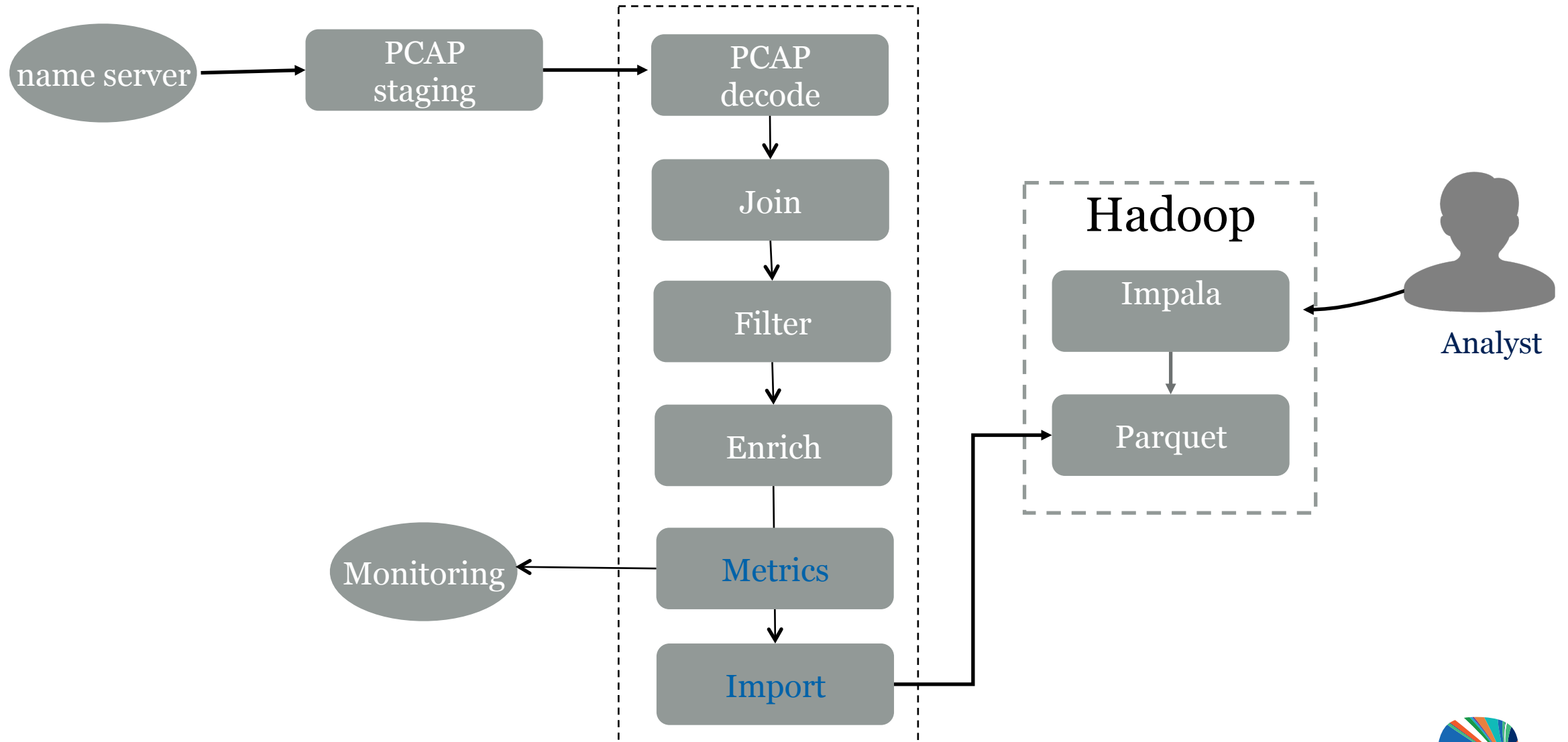
entrada@sidn.nl
entrada.sidnlabs.nl



Future Work

- More DNS research in collaboration with research partners
- Develop more data-driven applications and services based on ENTRADA
- Build an active ENTRADA users community

Workflow

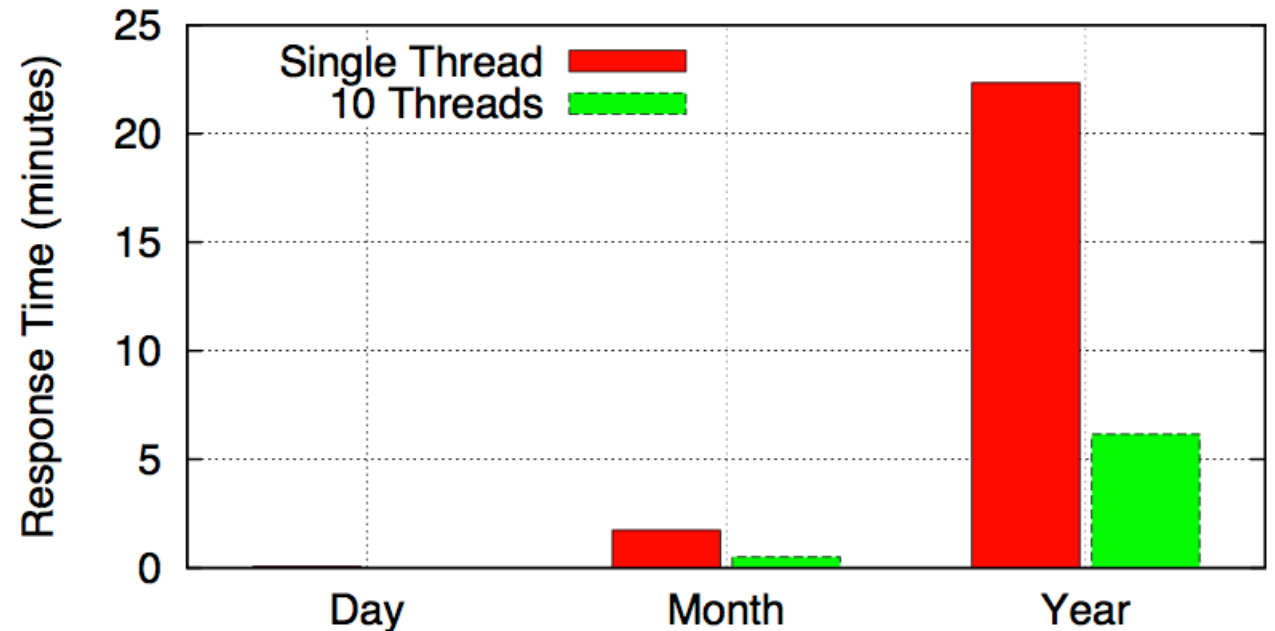


Query data available for analysis within 10 minutes

Performance

Example: count # daily ipv4 queries.

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



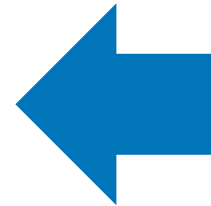
Query response time

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

E-mail security 1/3

- What is the usage of DMARC/DKIM?
 - Count standardized labels, see RFC 6376 and RFC 7489

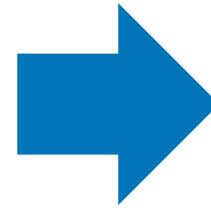
```
Select country,count(1)
from dns.queries
where qtype =16
and (qname like '%_domainkey.%'
or qname like '_dmarc .%')
and rcode=0
and ((year=2014 and month>6) or
year=2015)
and server='ns1.dns.nl'
group by country
```



Use standard SQL for analysis

E-mail security 2/3

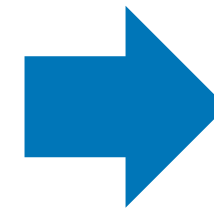
Country	# Queries	Percentage
US	208,533,790	42.60
IE	84,515,235	17.26
NL	79,052,717	16.15
BE	67,963,161	13.88
FI	9,112,053	1.86
RU	7,306,873	1.49
DE	7,119,556	1.45
GB	5,897,734	1.20
CN	5,446,895	1.11
DK	2,958,891	0.60



89.9% of queries
originate from top 4 countries

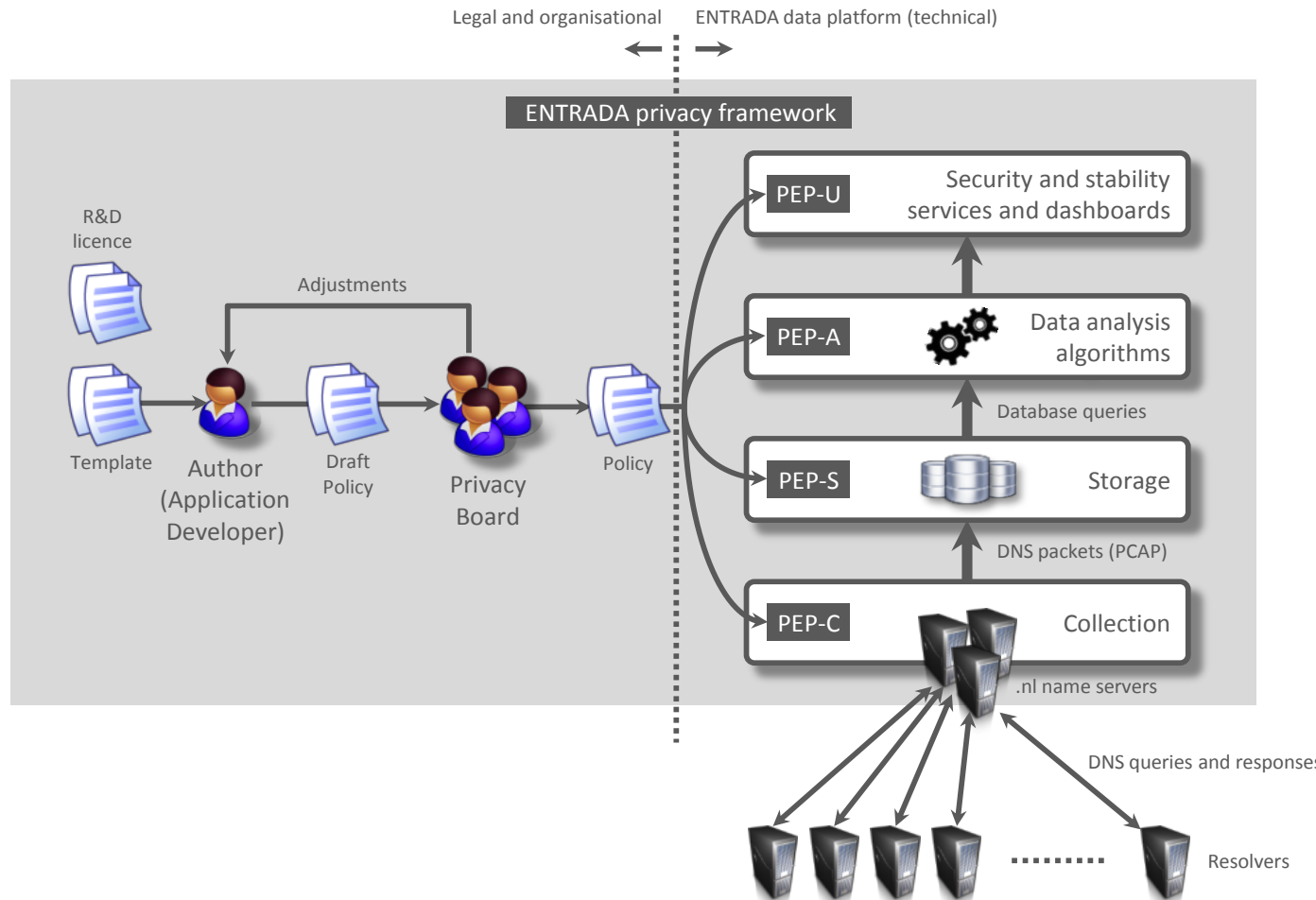
E-mail security 3/3

Provider	ASN	# Queries	Percentage
Google	AS15169	302,465,578	61.79
Microsoft	AS8075	51,556,416	10.53
Unknown	UNKN	15,788,699	3.22
AOL	AS1668	12,971,456	2.65
Yahoo	AS36647	11,83,129	2.30
Yahoo	AS26101	10,24,857	2.07
Yahoo	AS36646	9,150,523	1.87
Yahoo	AS34010	4,522,388	0.92
IDC China Tel	AS23724	4,520,819	0.92
Mail.ru	AS47764	3,659,097	0.75



82.13% of queries originate from large e-mail providers

Privacy Framework



Policy elements:

- Purpose
- Data that is used
- Filters on the data
- Retention period
- Access to the data
- Type of application (Research vs. Production)