ENTRADA: Enabling DNS Big Data Applications

Maarten Wullink - SIDN | APWG eCrime 2016

June 2nd 2016 - Toronto

What if…

You have many TB’s of network data?

And you want to:

1. Store it efficiently
2. Query it efficiently (SQL with interactive response times)
3. Quickly test a large number of hypotheses on your data
4. Continuously keep adding new data

You could...

1. Convert pcap to text format like csv and use Linux utilities
2. Run 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 stream warehouse (DSW)

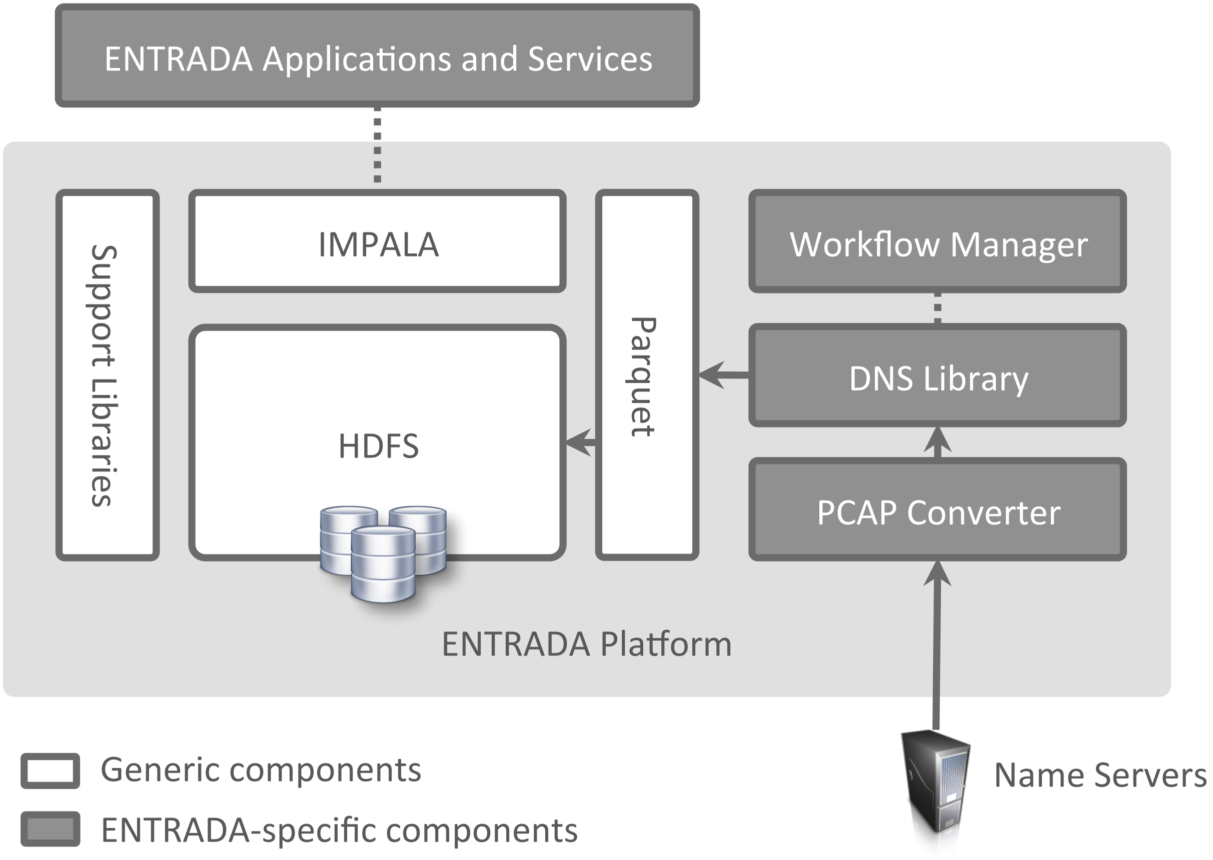
- ENTRADA is our open source Hadoop-based DSW (**entrada.sidnlabs.nl**)

- Analyze 50TB of converted pcap data in under 3.5 minutes using a small cluster

- Our main use case: network (DNS, TCP/IP, ICMP) analytics

ENTRADA

ENhanced Top-Level Domain Resilience through Advanced Data Analysis



ENTRADA Applications and Services

Support Libraries

IMPALA

HDFS

Parquet

Workflow Manager

DNS Library

PCAP Converter

ENTRADA Platform

Name Servers

□ Generic components

■ ENTRADA-specific components

ENTRADA@SIDN

- We are the TLD registry of the Netherlands (.nl)

- Use ENTRADA to further increase security and stability

- Operational for over 2 years

- Capturing data from .nl name servers

- 160 billion rows (DNS query+response tuple), 21 TB of data

SIDN LABS

More ENTRADA details

For design choices and a performance evaluation, see our 2016 NOMS paper:

"ENTRADA: a High-Performance Network Traffic Data Streaming Warehouse", IEEE/IFIP Network Operations and Management Symposium 2016 (NOMS 2016), Instanbul, Turkey

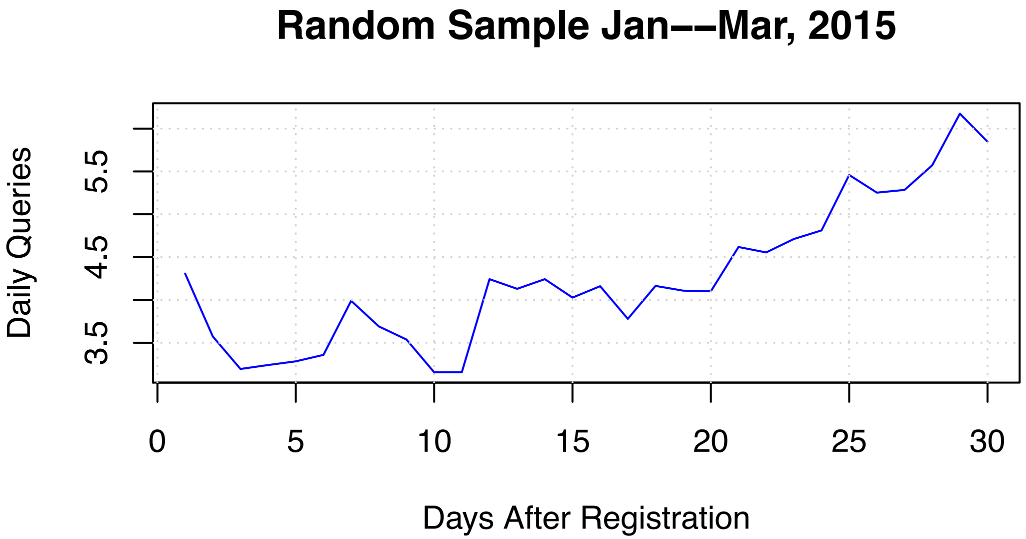
See: https://www.sidnlabs.nl/publicaties

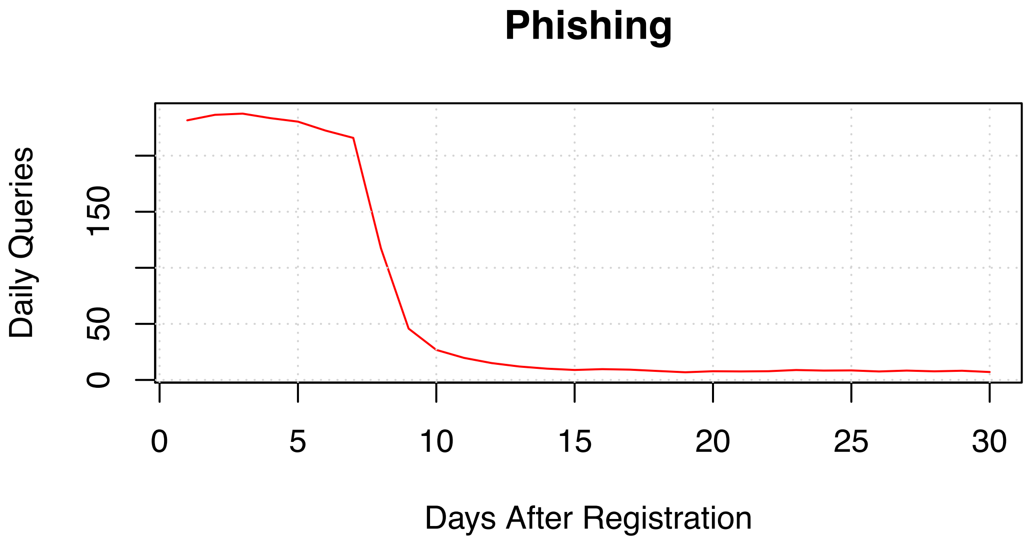
Example Use Cases

- Statistics (**stats.sidnlabs.nl**)
- Scientific research
- Insight for DNS operators
- **Malicious domain detection**
- **Botnet client detection**
- **Measuring uptake of email security**

Malicious Domain Detection (1/2)

Observation: New phishing domains have distinct query patterns

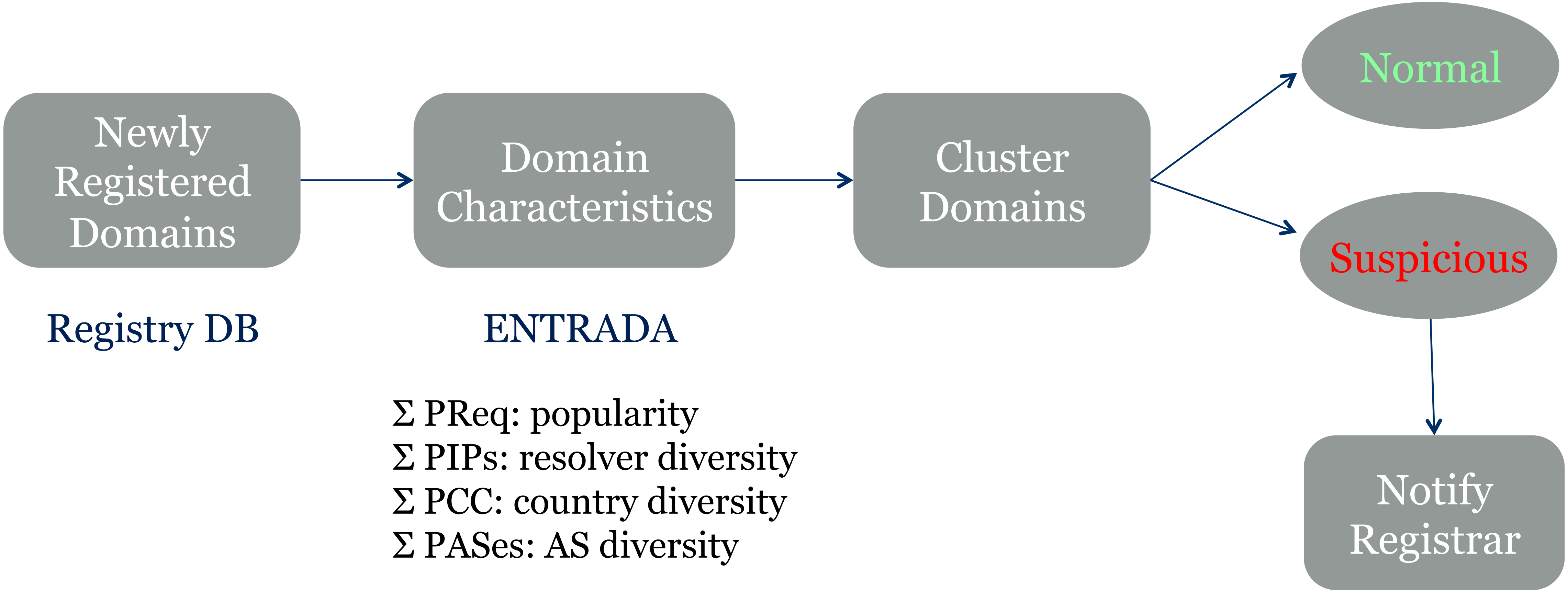

Random Sample Jan––Mar, 2015


Phishing

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

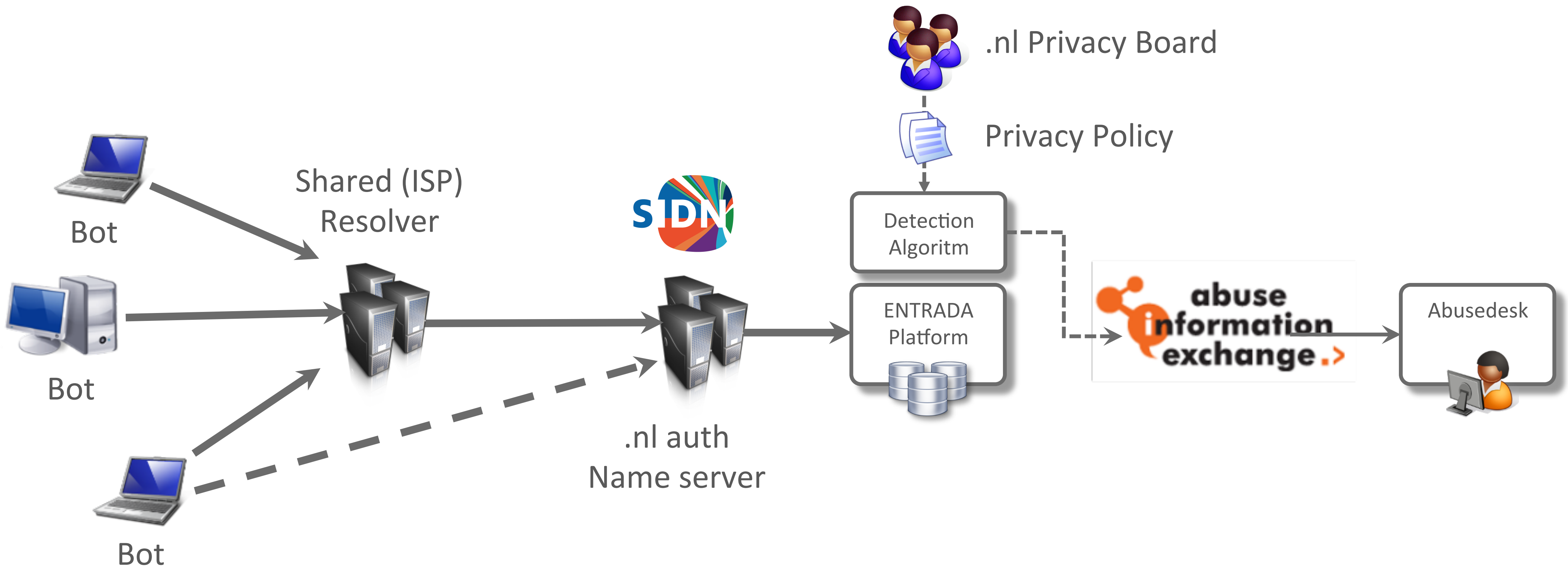
Malicious Domain Detection (2/2)

Every day workflow

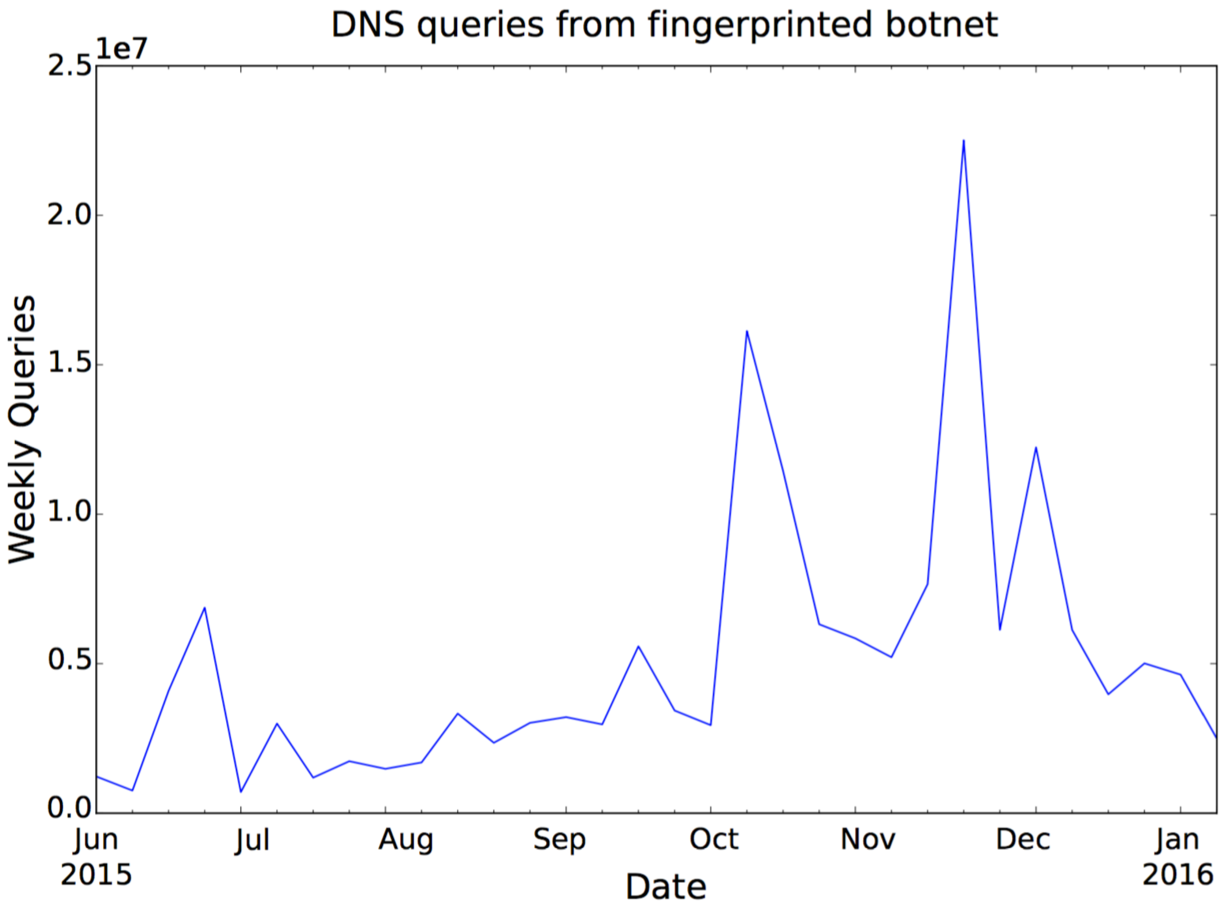


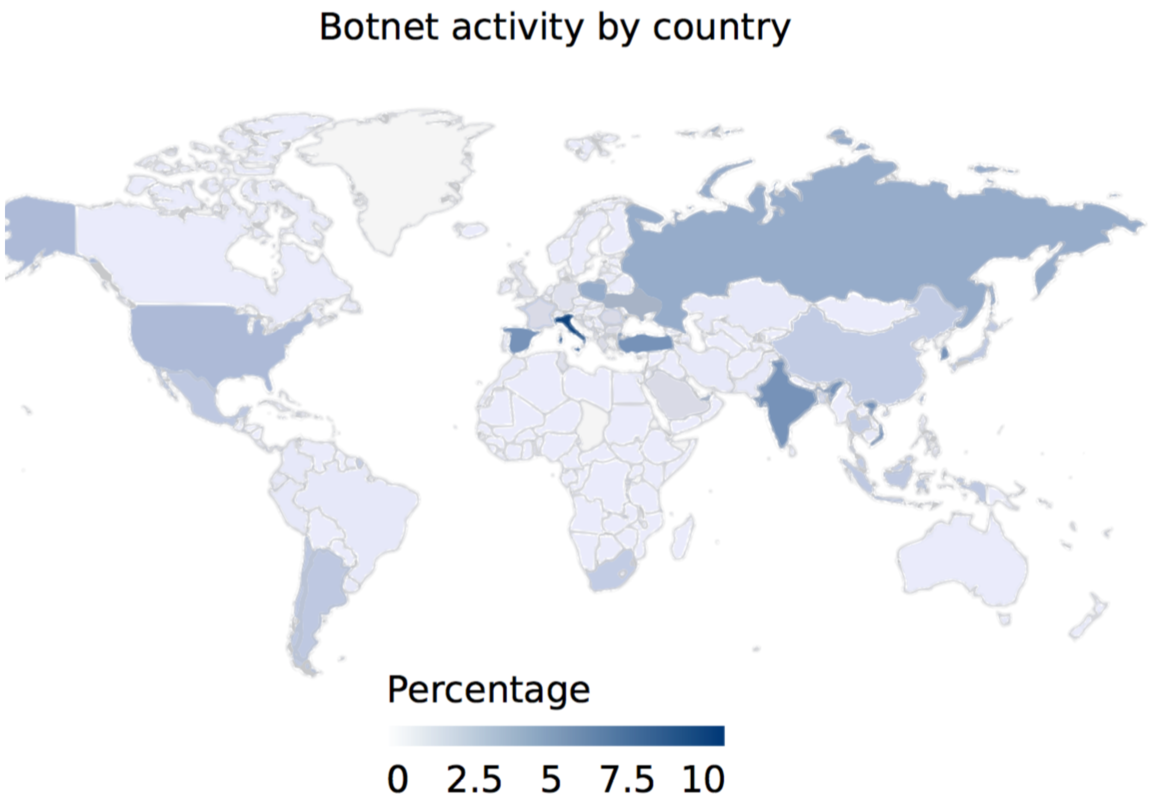
Σ PReq: popularity
Σ PIPs: resolver diversity
Σ PCC: country diversity
Σ PASes: AS diversity

Botnet Client Detection (1/2)



Botnet Client Detection (2/2)





Uptake of DKIM/DMARC (1/3)

- Email security standards DKIM (RFC 6376) and DMARC (RFC 7489)

- Approach: count standardized labels

Where is DKIM/DMARC used most?

```
select country,count(1) as total
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)
group by country
```

Use standard SQL for analysis

Uptake of DKIM/DMARC (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

Uptake of DKIM/DMARC (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 4 large e-mail providers

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 real world example use cases
- ENTRADA can be extended to other use cases
- Download and contribute!

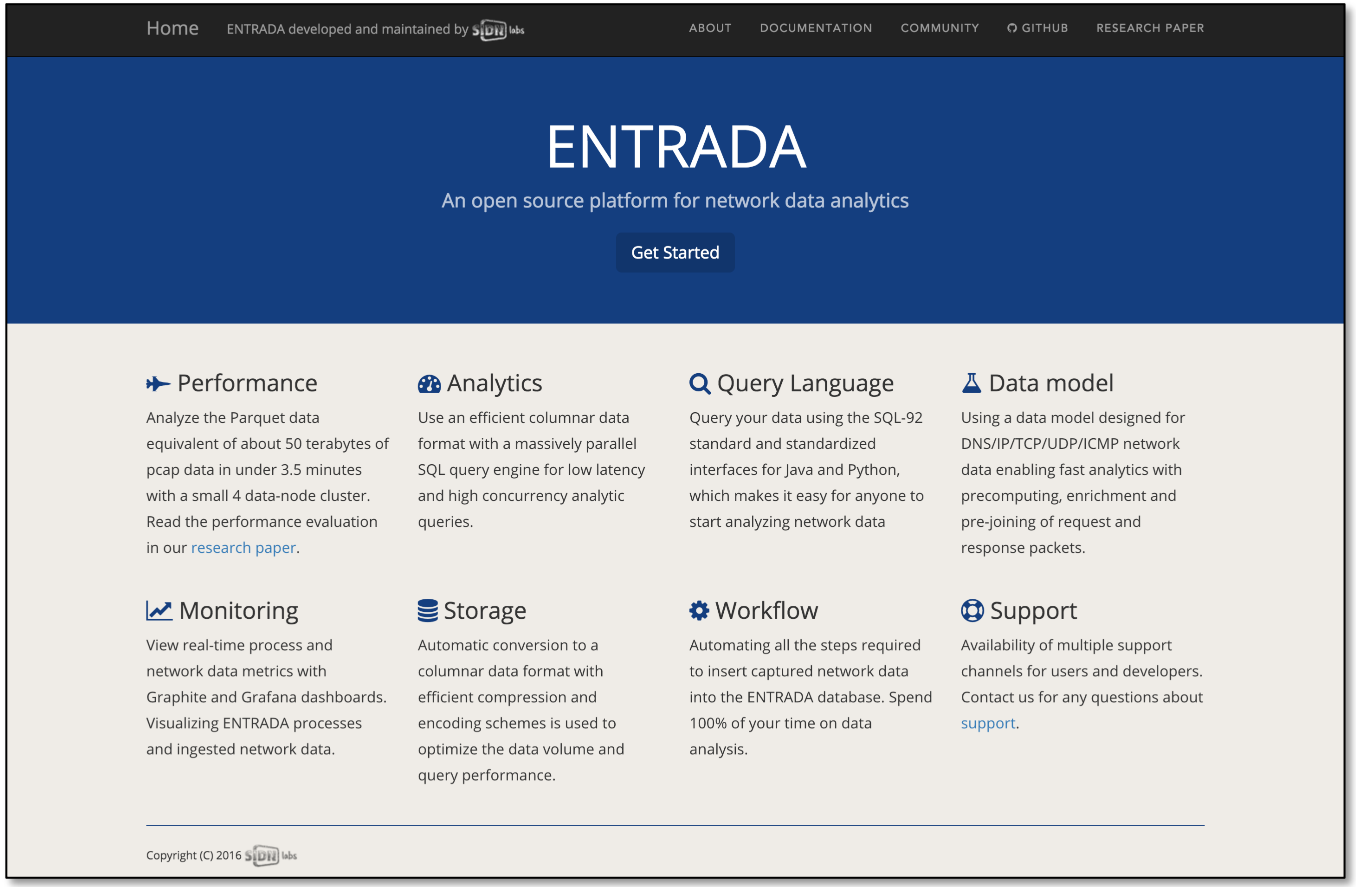
Future Work

- More DNS research in collaboration with research partners
- Develop data-driven applications and services based on ENTRADA
- Facilitate ENTRADA user community

Questions?

Maarten Wullink
Sr. Research Engineer

maarten.wullink@sidn.nl
@wulliak
www.sidnlabs.nl



ENTRADA

An open source platform for network data analytics

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.sidnlabs.nl