Early phishing detection in .nl through the ENTRADA platform

SIDN Relatiedag

26 november 2015

Giovane Moura & Maarten Wullink

SIDN labs
Internet Research & Innovation

Inhoud

- Deel 1: Phishing detectie toepassing
  - (Giovane): Big Data toepassing (in het Engels)

- Deel 2: ENTRADA
  - (Maarten): DNS big data platform

SIDN labs
Internet Research & Innovation

SIDN

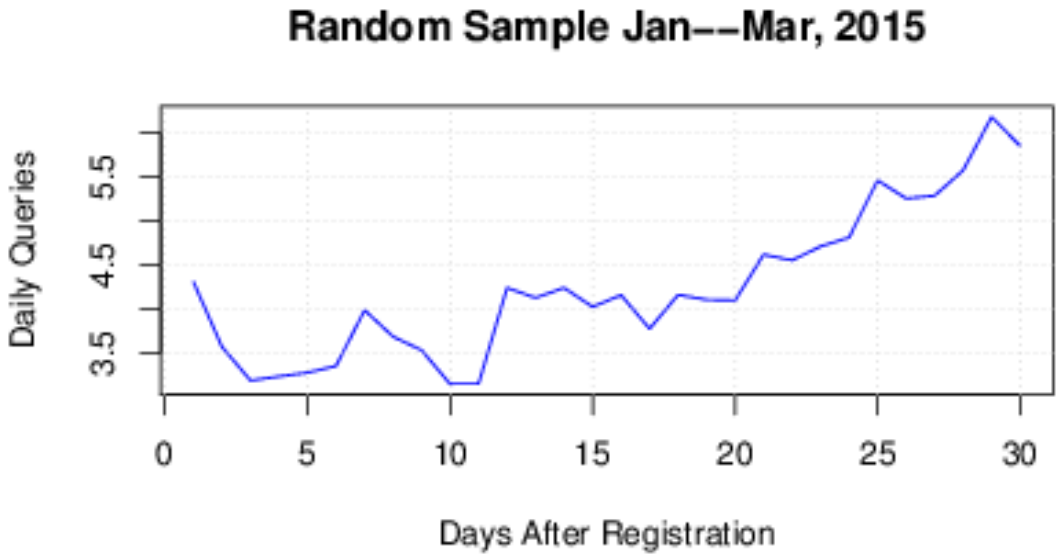
Motivation

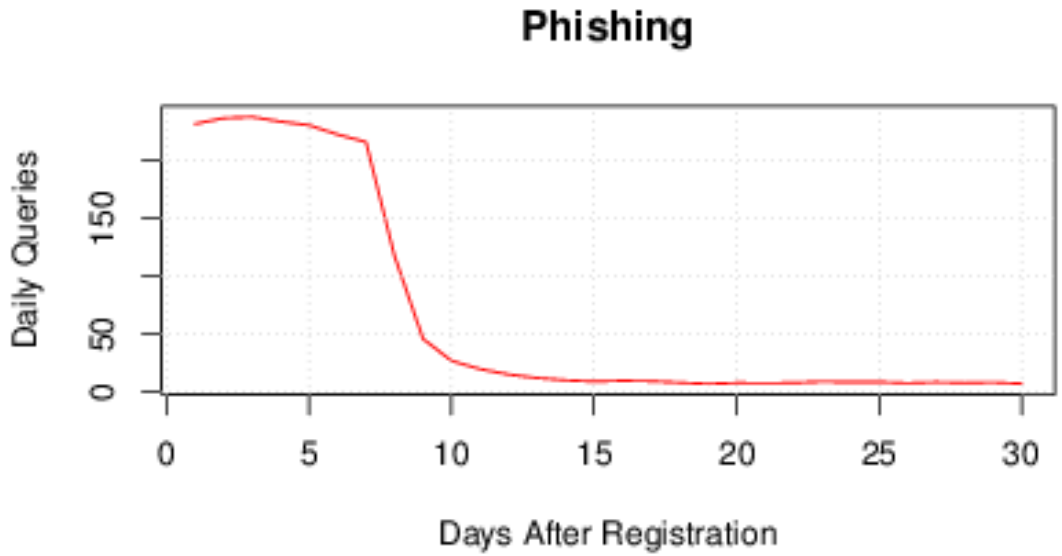
- Our goal is protect the .nl domain:
  - Domains; users, hosting, registrars; registrants
- Many compromised .nl domains are hacked CMS's
- Some (e.g.: phishing) are newly registered domains
  - Helps with credibility
- Potential damage is huge:
  - Blacklisting IPs of hosting providers, etc.
  - Internet users losing money

Introduction (1/2)

- This project: how to detect those newly registered domains?

- Newly registered malicious domains have an abnormal initial DNS lookup [1]
  - @Registrars/Hosting: do you see that in your DNS/web logs?





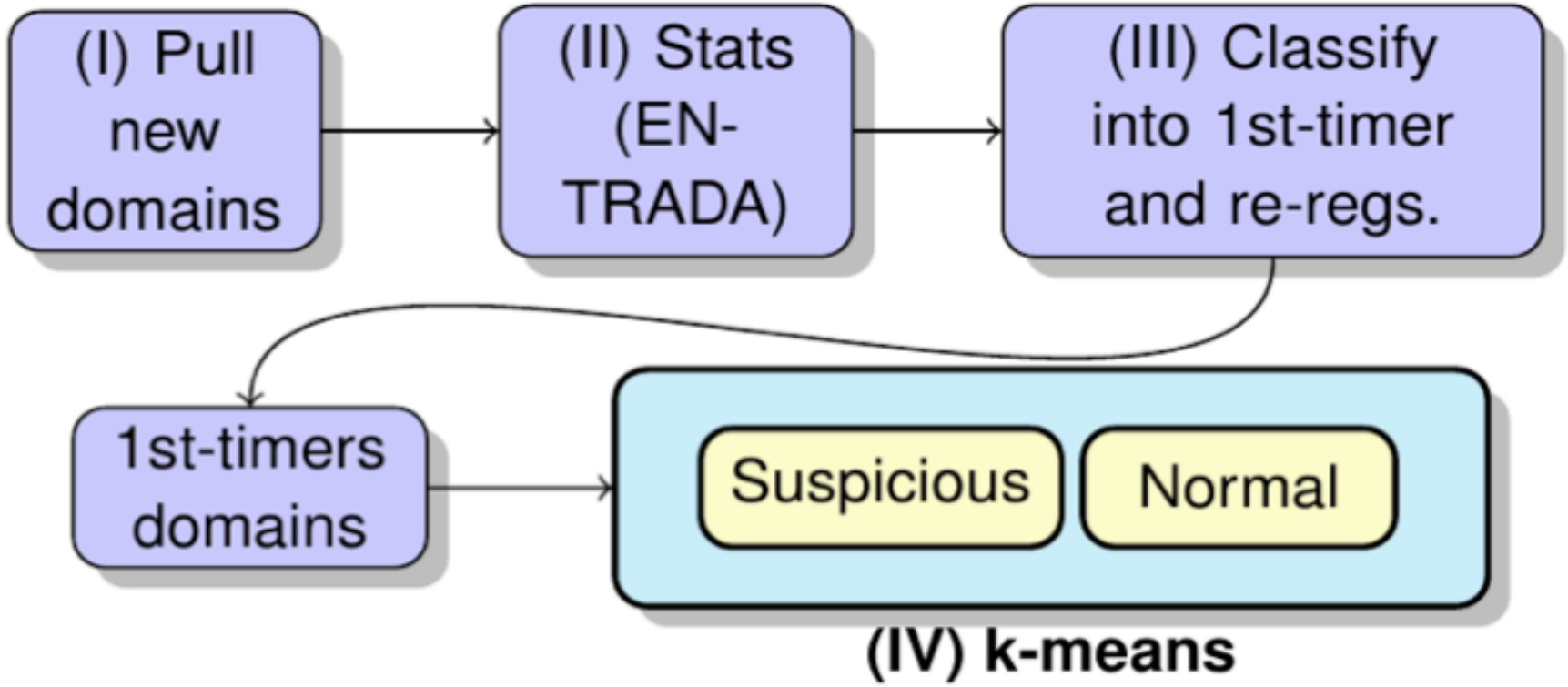
Internet Research & Innovation

Introduction (2/2)

- Why is that?
  - Assumption: spam-based business model
  - Automated
  - Maximize profit before being taken down

- **Question: can we use this to improve security in the.nl zone?**

New Domains Early Warning System



Evaluation (1/2)

Key	Value
Interval	Jan 1st, 2015 to Aug 30th 2015
Average .nl zone size	$\sim$ 5,500,000
$\sum$ new domains	586,201
New domains - first timers	476,040(81.2%)
New domains - re-registered	110,161 (18.8%)
Total DNS Requests	32,864,402,270
DNS request new domains (24h)	826,740
DNS request new domains - first-timers (24h)	420,362

Evaluation (2/2)

Cluster	Size	$\sum$ *Req*	$\sum$ *IPs*	$\sum$ *CC*	$\sum$ *ASes*
Normal	132,425	4.31	3.06	1.64	1.43
Suspicious	2,956	55.03	27.87	4.99	7.43

Validation with Historical Data (1/3)

- Were those “suspicious” domains really malicious?

- Very hard to verify on historical data: if they had pages; they might be gone or diff by now

- Results with historical data:
  - Content analysis: 148 “shoe stores” , 17 adult/malware
  - 19 phishing domains (out of 49 reported by Netcraft on the same period)
  - VirusTotal: 25 domains matched

Validation with Historical Data (2/3)

- Why so many (5–10) new shoes stores per day?
- Most counterfeit product = 40% of US Border seizures
- Shoes are a smart play: high demand, and low penalties
- @Registrars/Hosting: do you see this too?

Validation with Current Data

- “Shoes” sites dominate it, depending on the day
- Adult and malware is also detected; we now download screenshots and content as we classify
- False positives: rapidly popular political websites and others
- Labs results on: act on the data to improve security on .nl
- Start a pilot: we would like to share this data with hosting/registrars

ENTRADA

DNS data @SIDN

- > 3.1 miljoen unieke resolvers per maand
- > 1.3 miljard query's per dag
- > 300 GB PCAP data per dag

ENTRADA

- *ENhanced Top-Level Domain Resilience through Advanced Data Analysis*
- Doel: data-driven verhogen van de veiligheid en stabiliteit van .nl en het Internet in de breedte
- Probleem: Wat doe je als je > 50TB aan compressed PCAP data snel wil doorzoeken?
- Belangrijkste requirement: high-performance, bijna real-time data warehouse
- Aanpak
  - Transformeer data naar een geoptimaliseerd opslag formaat
  - Analyseer data met een parallelle query engine

SIDN labs
Internet Research & Innovation

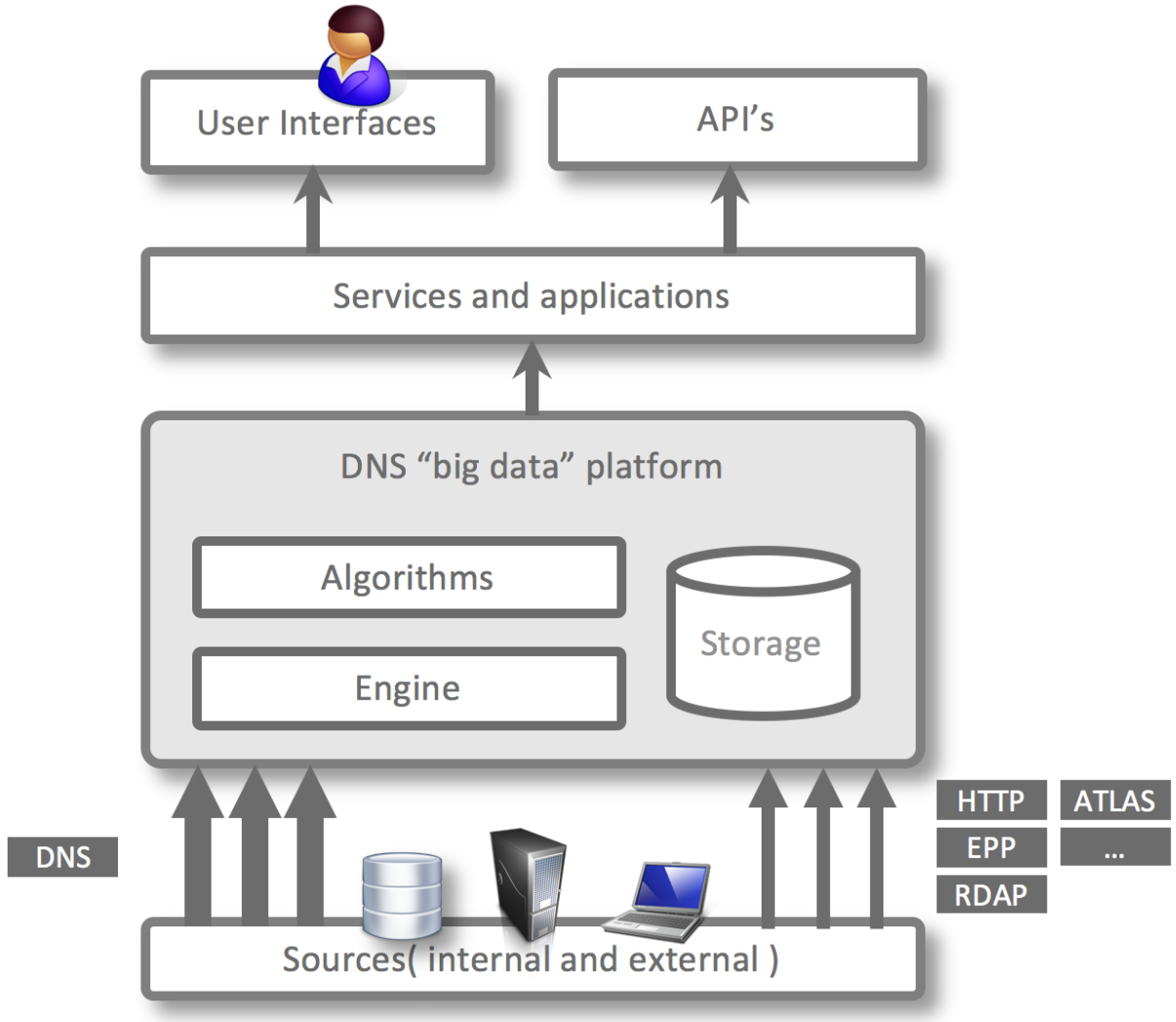
SIDN

Use cases

- Visualisatie van DNS patronen (patronen van phishing domeinnamen)
- Botnet infecties detecteren
- Real-time Phishing detectie
- Statistieken (stats.sidnlabs.nl)
- Wetenschappelijk onderzoek (in samenwerking met Nederlandse universiteiten)
- Operationele ondersteuning van DNS operators

SIDN labs
Internet Research & Innovation
SIDN

ENTRADA architectuur



SIDN labs
Internet Research & Innovation

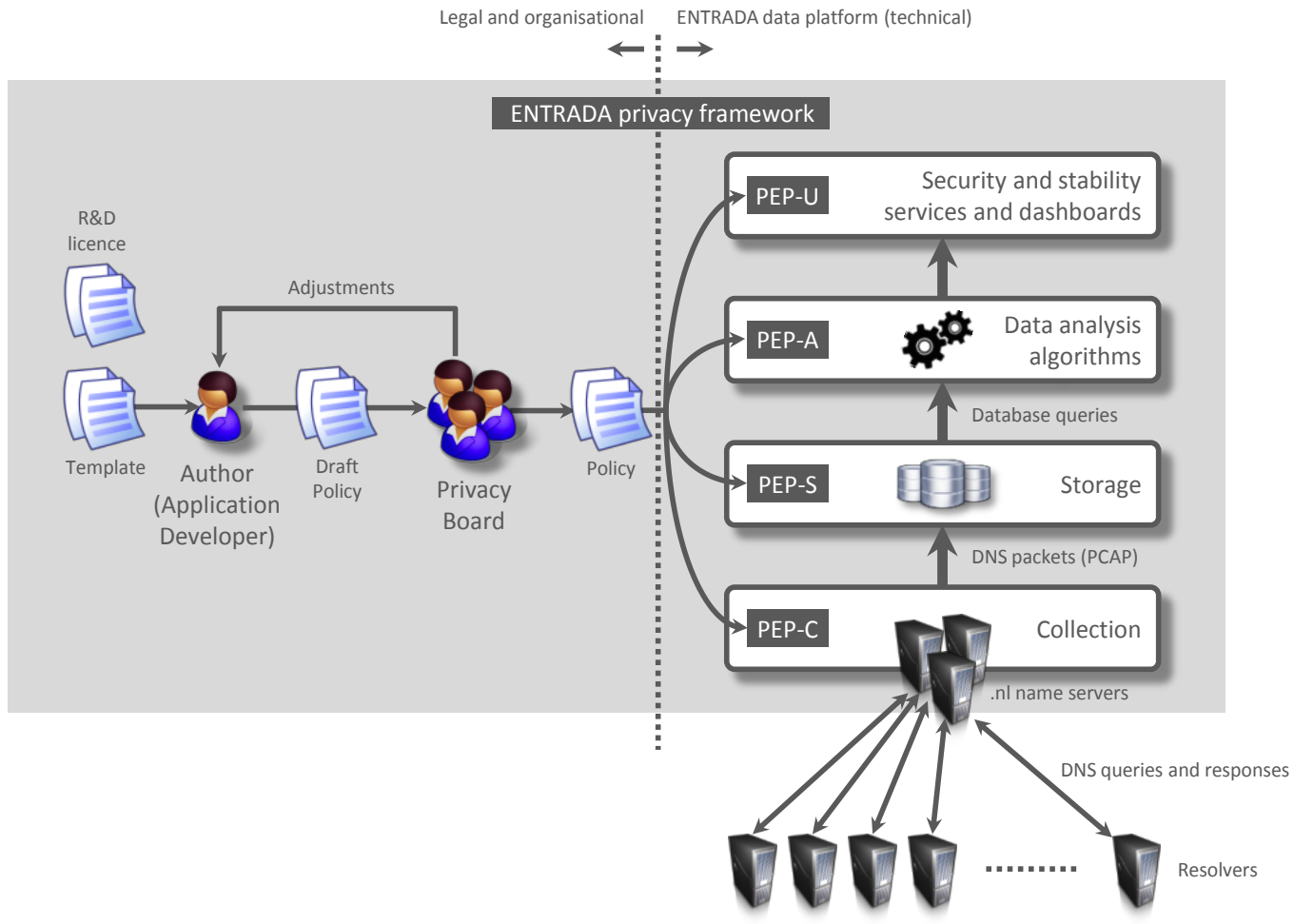
SIDN

ENTRADA privacyraamwerk

- Onderdeel van de “ENTRADA basis”

- Belangrijkste concepten
  - Applicatie-specifieke privacy policy
  - Privacy Board
  - Policy Enforcement Points

- Policy elementen zijn o.a.:
  - Doel
  - Data die gebruikt wordt
  - Filters
  - Opslag periode

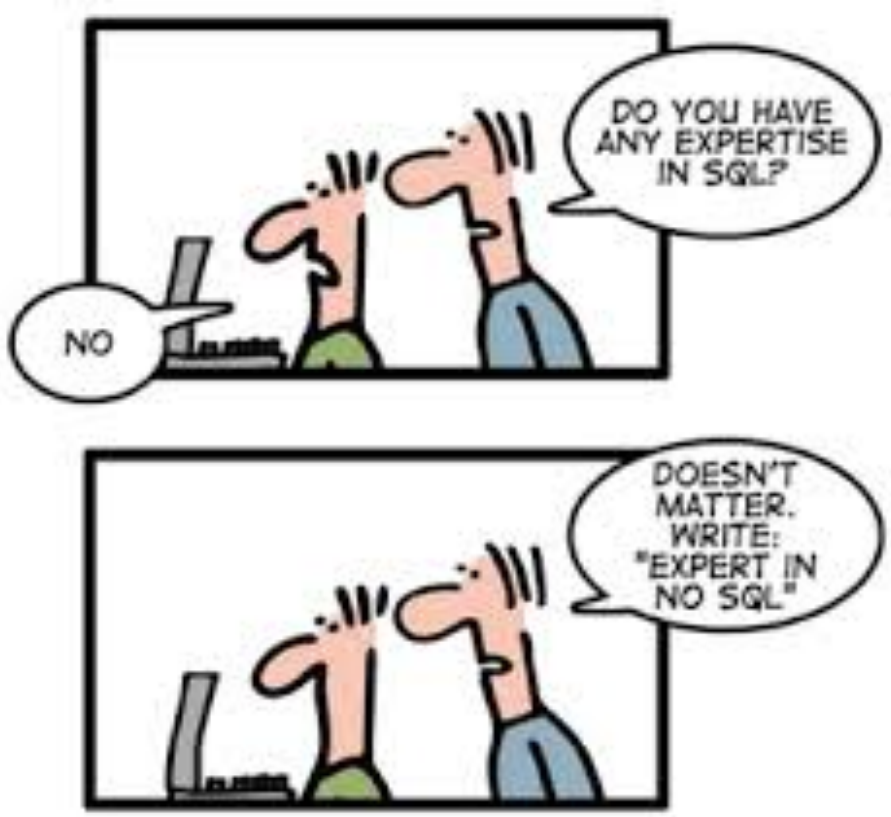


Query engine opties

- Keuze uit veel verschillende opties!

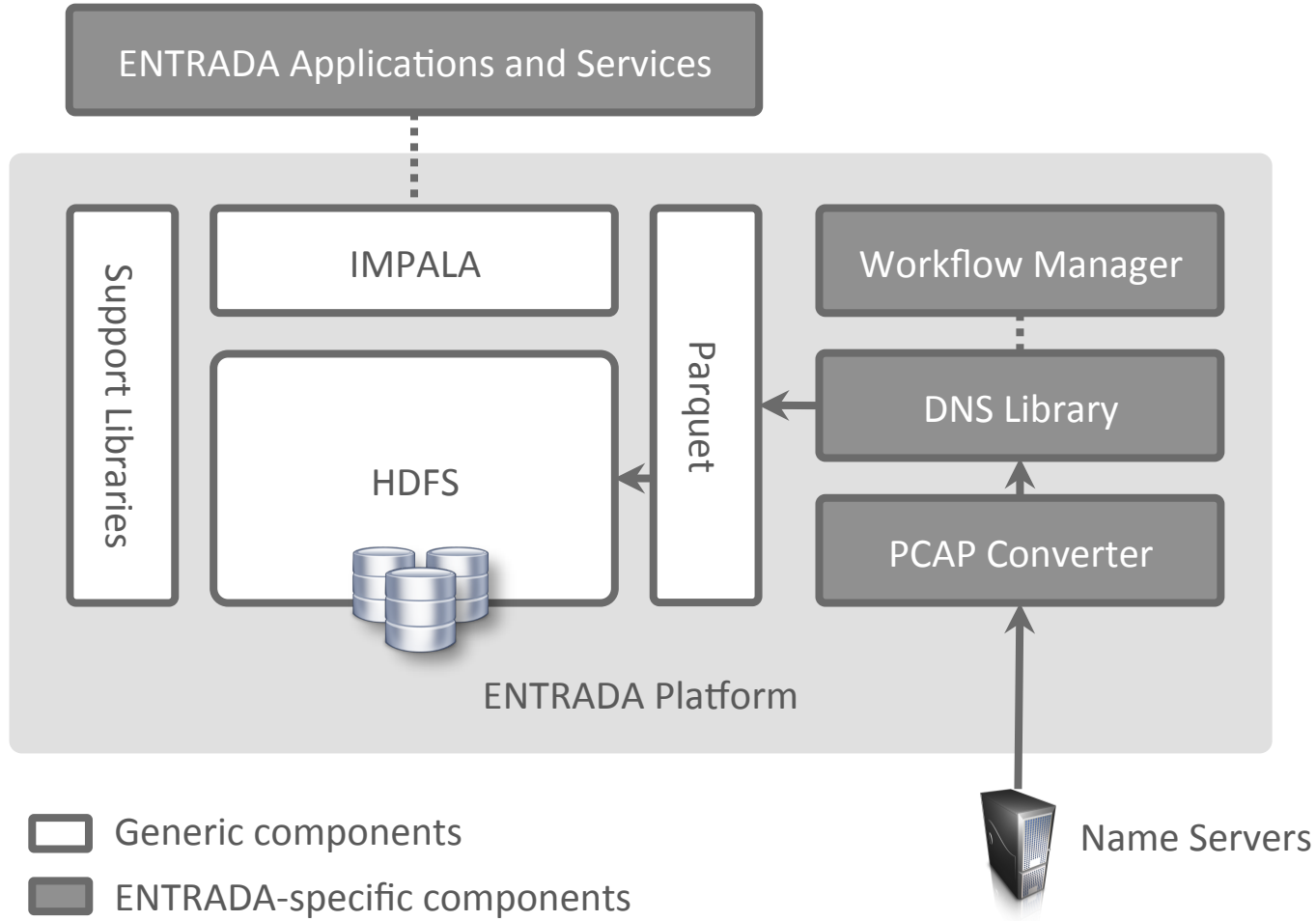
- **SQL and NoSQL oplossingen geëvalueerd**

- Relationeel SQL (PostgreSQL)
- MongoDB
- Cassandra
- Elasticsearch
- Hadoop (HBASE + Apache Phoenix of Hive)
- SQL on Hadoop (HDFS + Impala + Parquet)



SIDN labs
Internet Research & Innovation
SIDN

ENTRADA componenten



Cluster ontwerp (nano-sized)

locatie I
management node

locatie II
data nodes

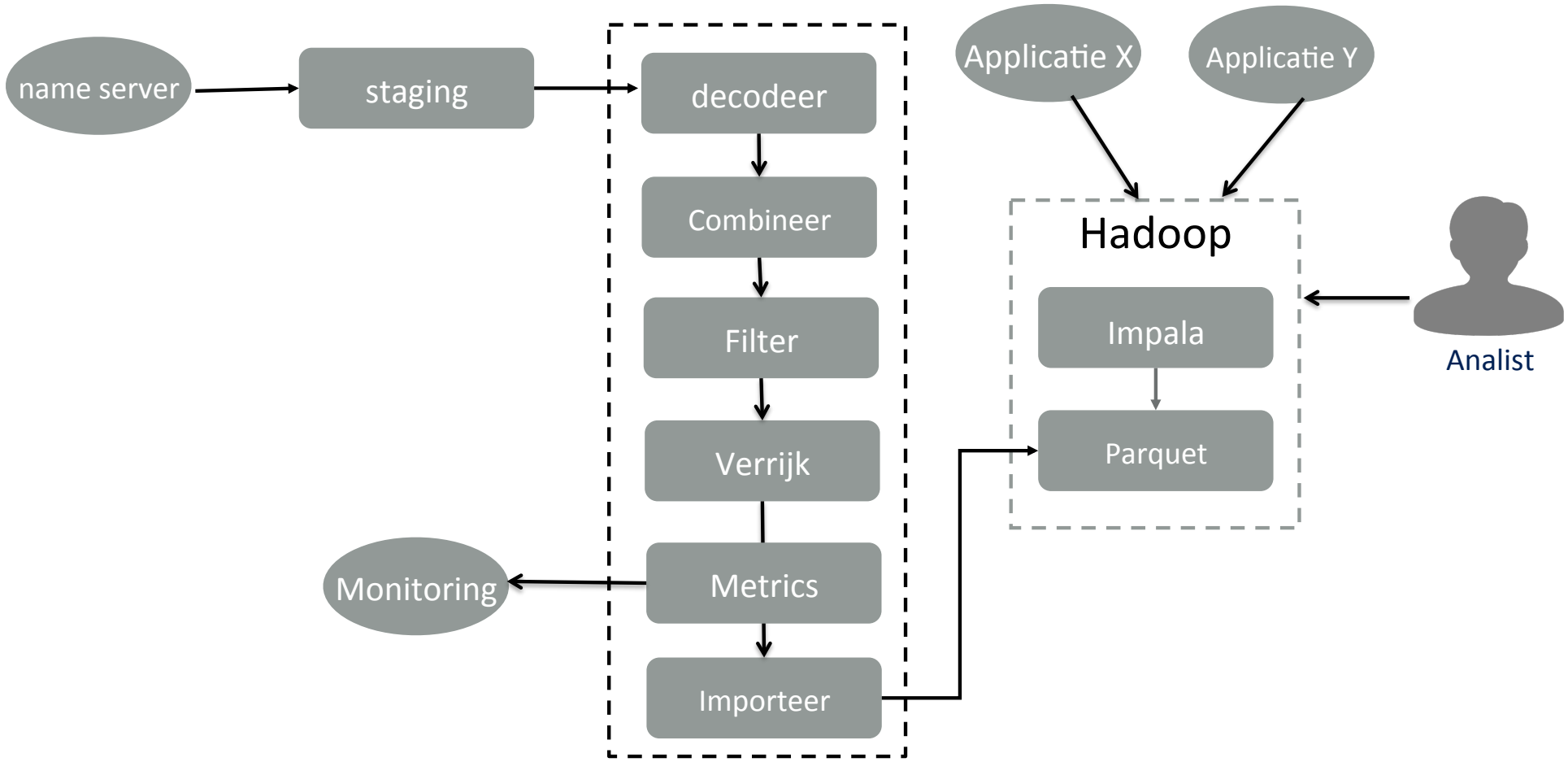
locatie III
data nodes

2Gb/s netwerk

SIDN labs
Internet Research & Innovation

SIDN

Workflow



Query data beschikbaar voor analyse binnen 10 minuten

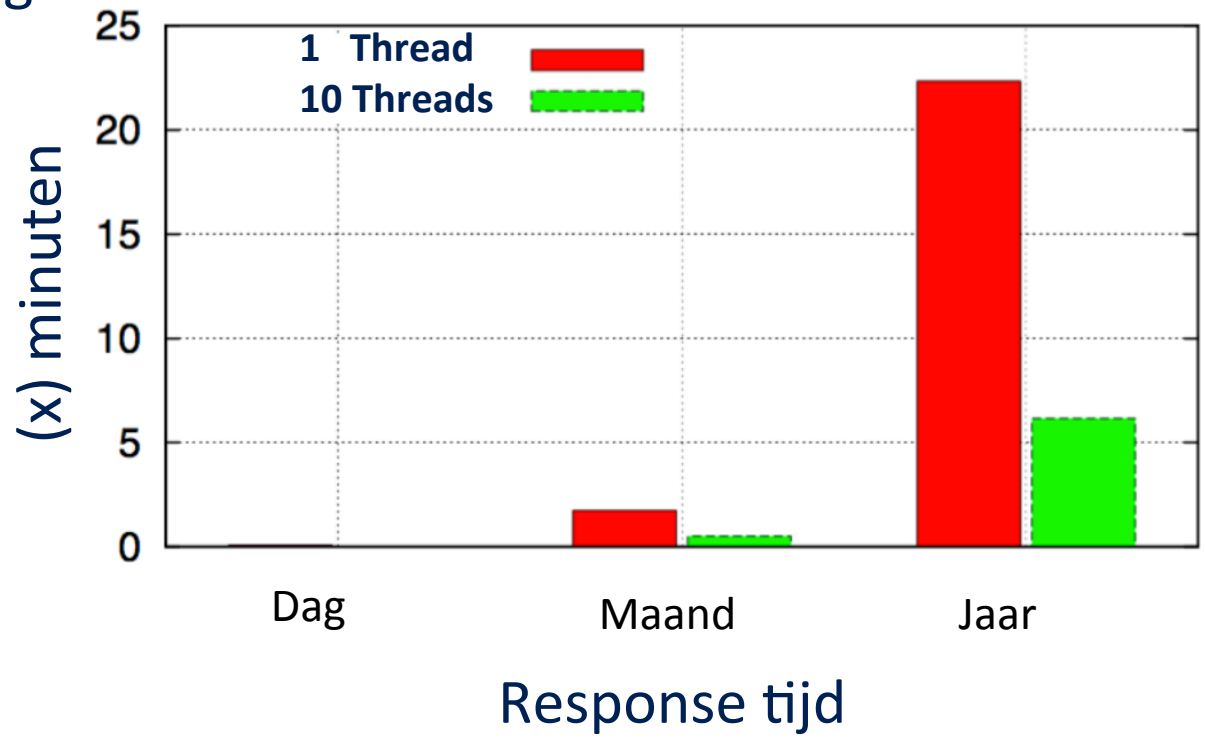
SIDN labs
Internet Research & Innovation

SIDN

Performance

Voorbeeld query: # ipv4 DNS queries per dag.

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



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

SIDN labs
Internet Research & Innovation
SIDN

ENTRADA status

Name servers	2
Queries per dag	~320M
PCAP volume(gzipped) per dag	~70GB
Parquet volume per dag	~14GB
Aantal maanden data	19
# queries opgeslagen	> 86 miljard
Totaal Parquet volume	> 3,6TB
HDFS (3x replicatie)	~ 11TB
Cluster capaciteit	~150 miljard tuples

Conclusies en resultaten

- Technisch: Hadoop HDFS + Parquet + Impala is een perfecte combinatie!

- Bijdrages:
  - Onderzoek door SIDN Labs and universiteiten
  - Kwaadaardige domeinnamen and botnets C&C's gevonden
  - Externe data feed naar Abuse Information Exchange
  - Inzicht in DNS query data

Toekomstig werk

- Combineren van data van .nl authoritatieve name server met scans van de hele .nl zone en ISP data
- Data van meer name servers en resolvers
- Open Data programma uitbreiden
- Cluster upgrade

Vragen?

Giovane Moura

Data Scientist

giovane.moura@sidn.nl

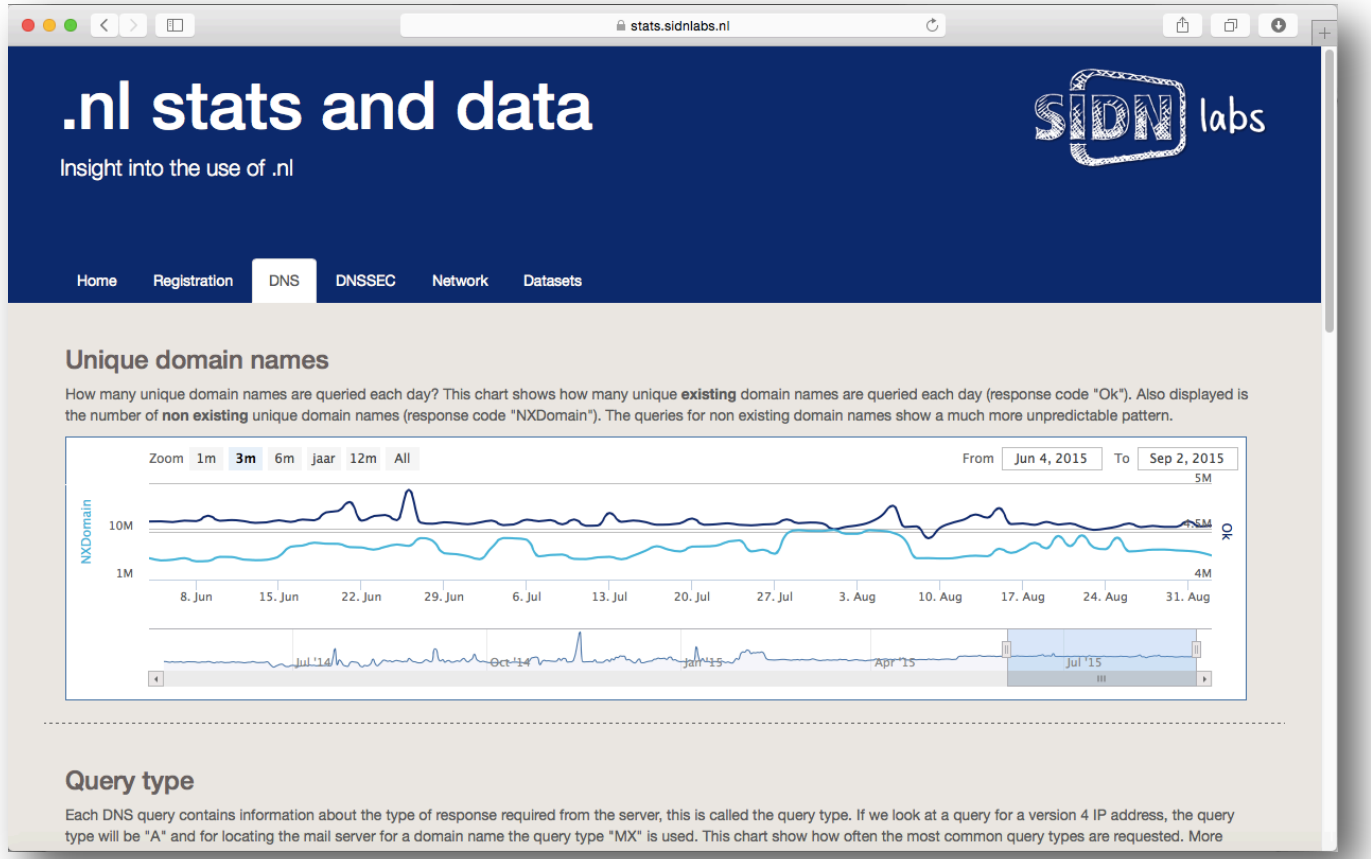
Maarten Wullink

Senior Research Engineer

maarten.wullink@sidn.nl

@wulliak

www.sidnlabs.nl



https://stats.sidnlabs.nl

SIDN labs
Internet Research & Innovation

SIDN