DNS Big Data Analytics

DNS-OARC Fall 2015 Workshop

October 4th 2015

Maarten Wullink, SIDN

SIDN labs
Internet Research & Innovation

SIDN

- Domain name registry for .nl ccTLD
- > 5,6 million domain names
- 2,45 million domain names secured with DNSSEC
- SIDN Labs is the R&D team of SIDN

SIDN labs
Internet Research & Innovation

SIDN

DNS Data @SIDN

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

SIDN labs
Internet Research & Innovation

SIDN

ENTRADA

ENhanced Top-Level Domain Resilience through Advanced Data Analysis

- **Goal**: data-driven improved security & stability of .nl
- **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

SIDN labs
Internet Research & Innovation

SIDN

Requirements

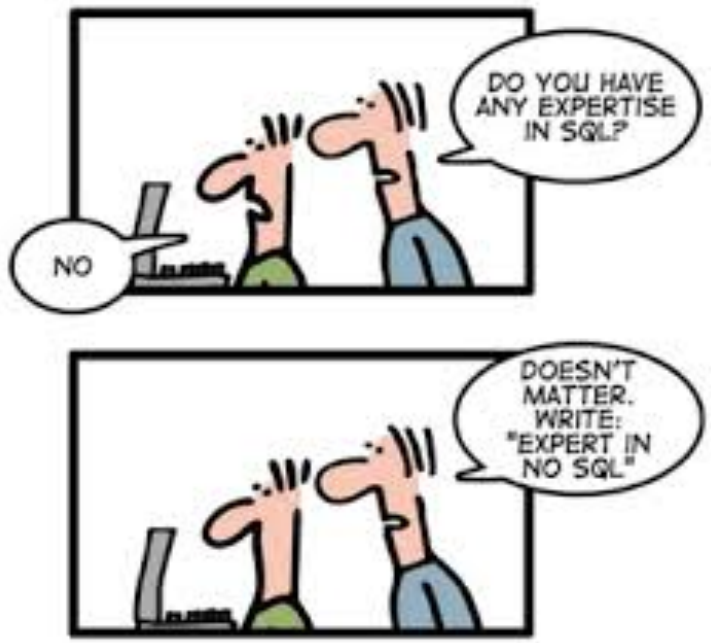
- SQL support
- Scalability
- High performance
- Capacity for >1 year of DNS data
- Extensibility
- Stability
- Don’t spend too much money!

Query Engine Options

Engines galore!

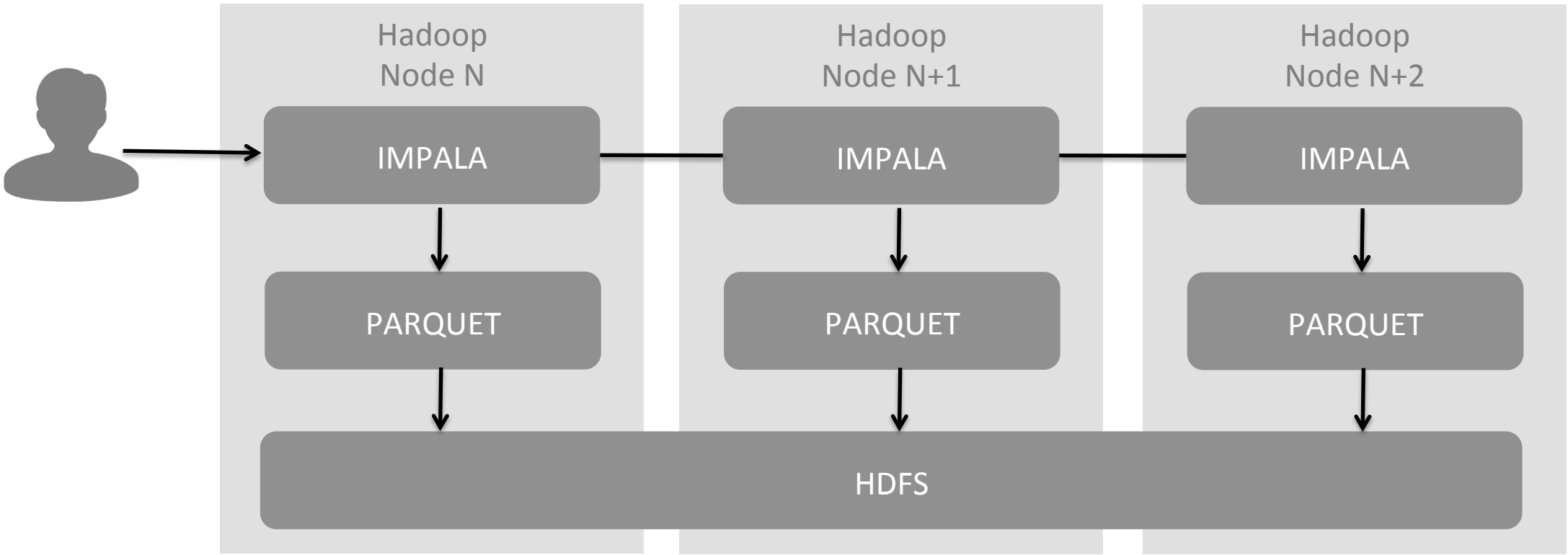
Evaluated SQL and NoSQL solutions

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



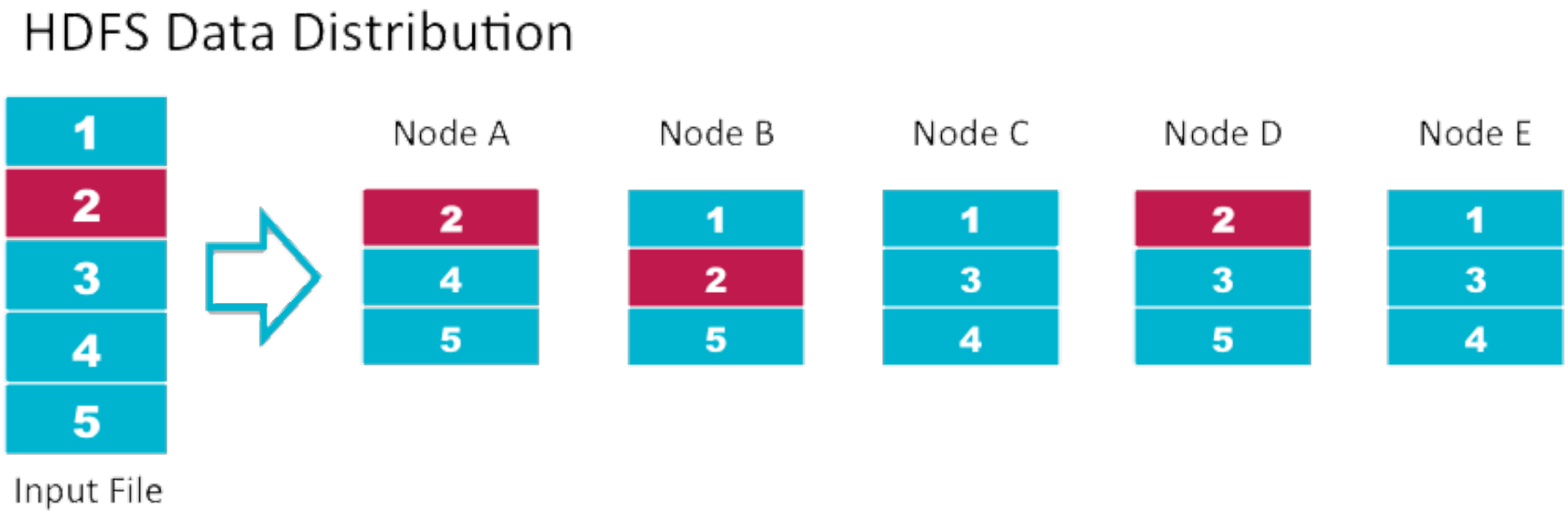
SQL on Hadoop

Best fit for our requirements



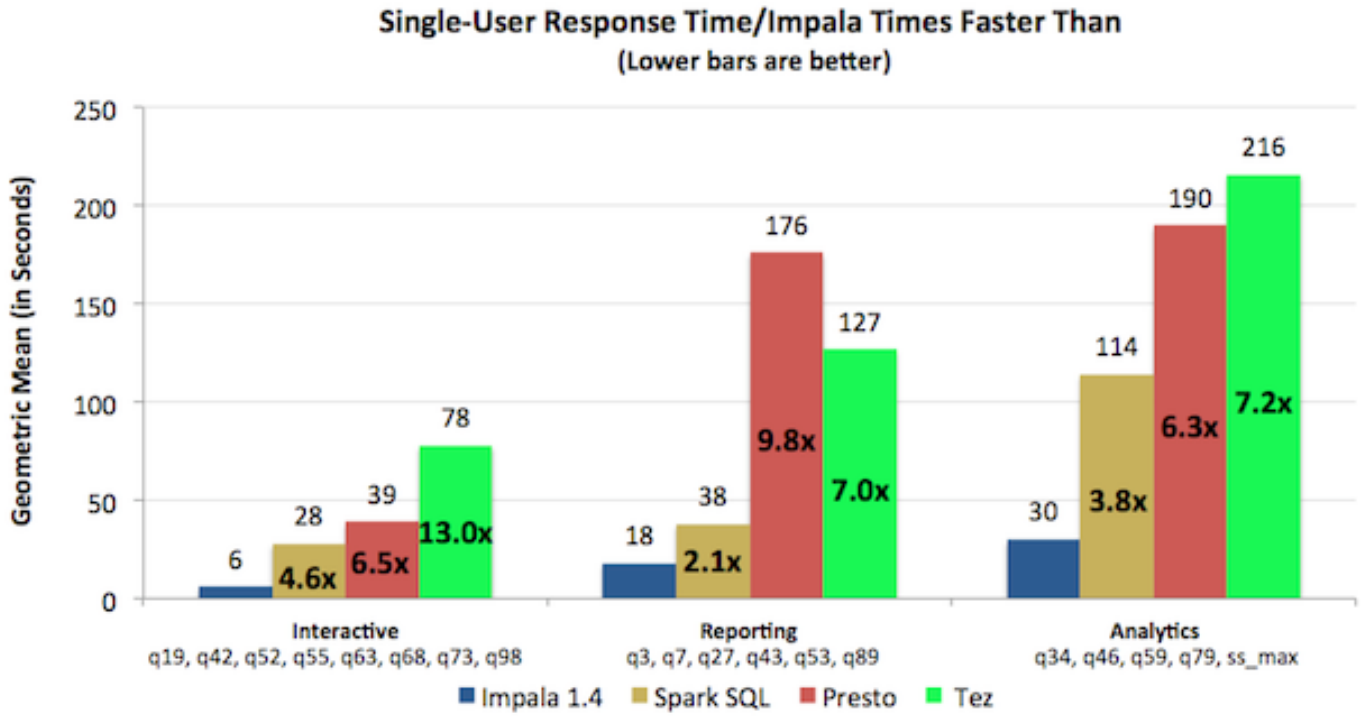
HDFS

- Distributed file system for storing large volumes of data
- High availability through replication of data blocks
- Scalable to hundreds of PB's and thousands of servers



Impala query engine

- **MPP** (massively parallel processing)
- Inspired by Google Dremel paper
- Provides low latency and high concurrency for BI/analytic queries on Hadoop
- Excellent performance when compared to other Hadoop based query engines.



SIDN labs
Internet Research & Innovation

SIDN

Impala (2)

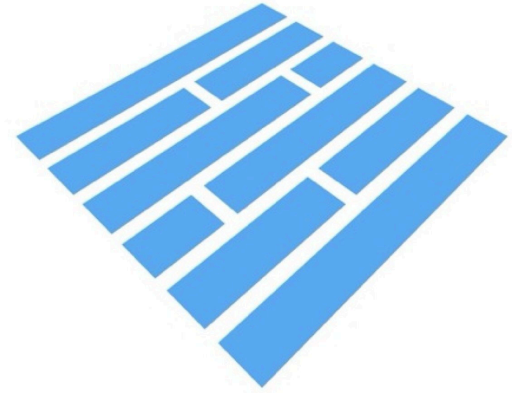
Data formats

- Text
- Hadoop formats
- Apache Avro
- Apache Parquet

Interfaces

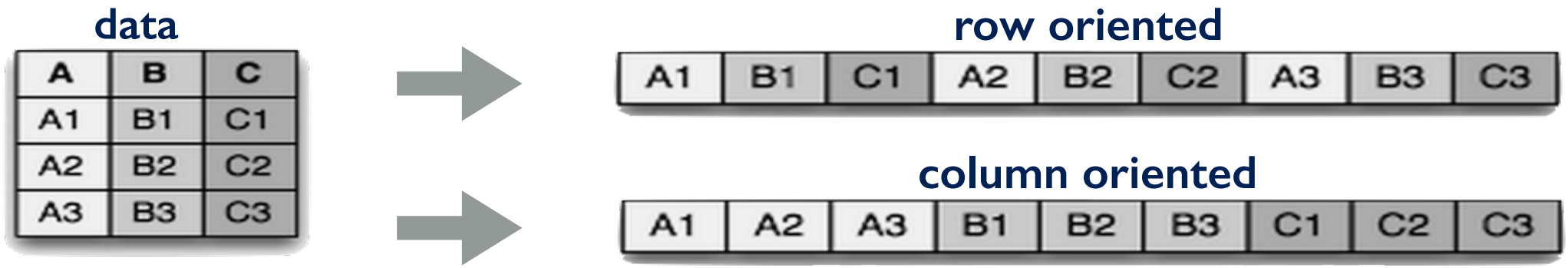
- Web-based GUI
- Command line (impala-shell)
- Python (Impyla)
- JDBC

Apache Parquet



- Why not just use the PCAP files?
  - Reading (compressed) PCAP data is just too slow
  - Analytical engines cannot read PCAP files

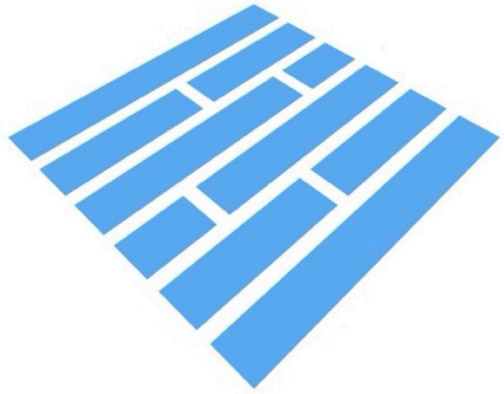
- Columnar storage format



SIDN labs
Internet Research & Innovation

SIDN

Apache Parquet (2)



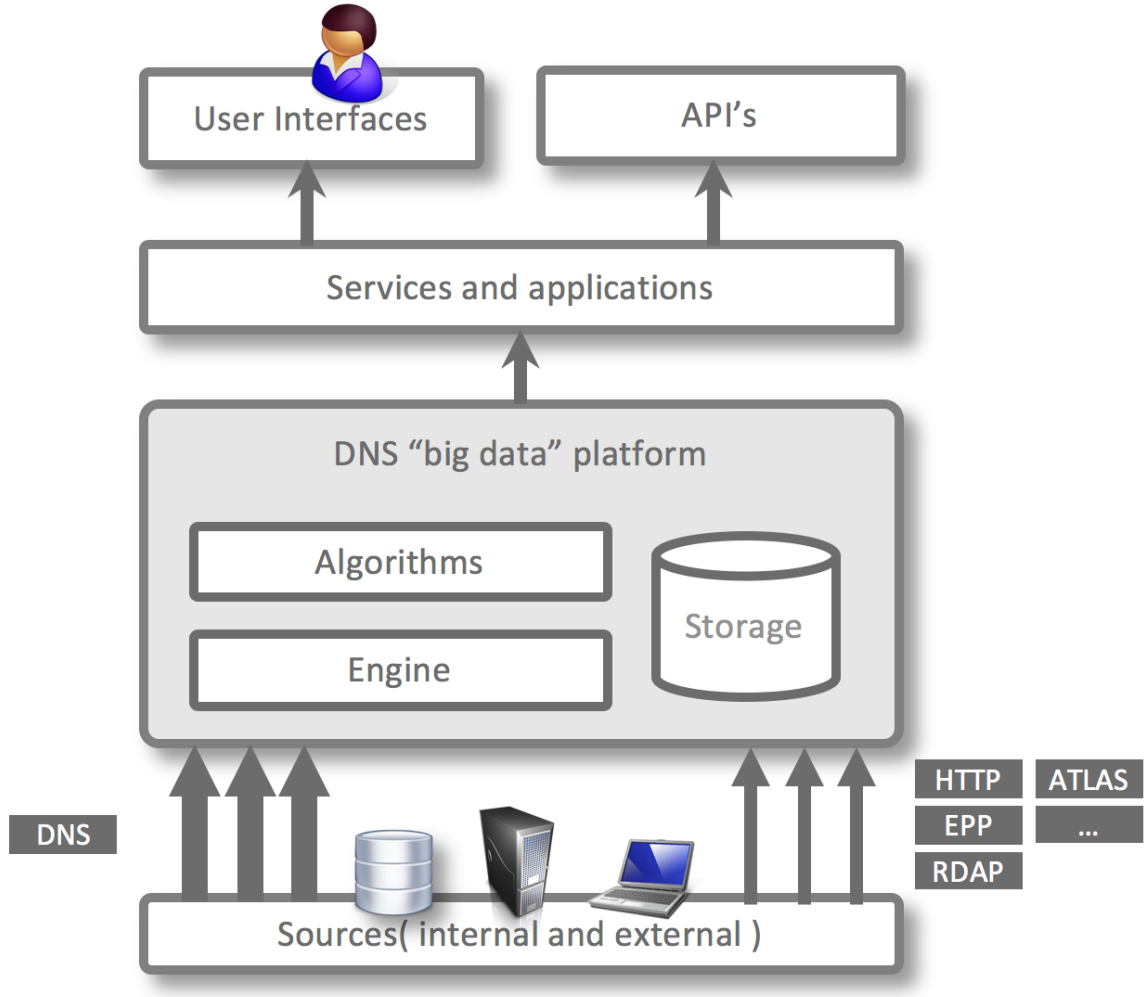
- Columnar storage allows for efficient encoding/compression
  - multiple encoding schemes
  - support for Snappy compression

- Partition data (e.g. by year, month, day and server)
  - Partition pruning allows Impala to skip data we are not interested in

- Other analytical engines such as Apache Spark can use the same Parquet data.

ENTRADA Architecture

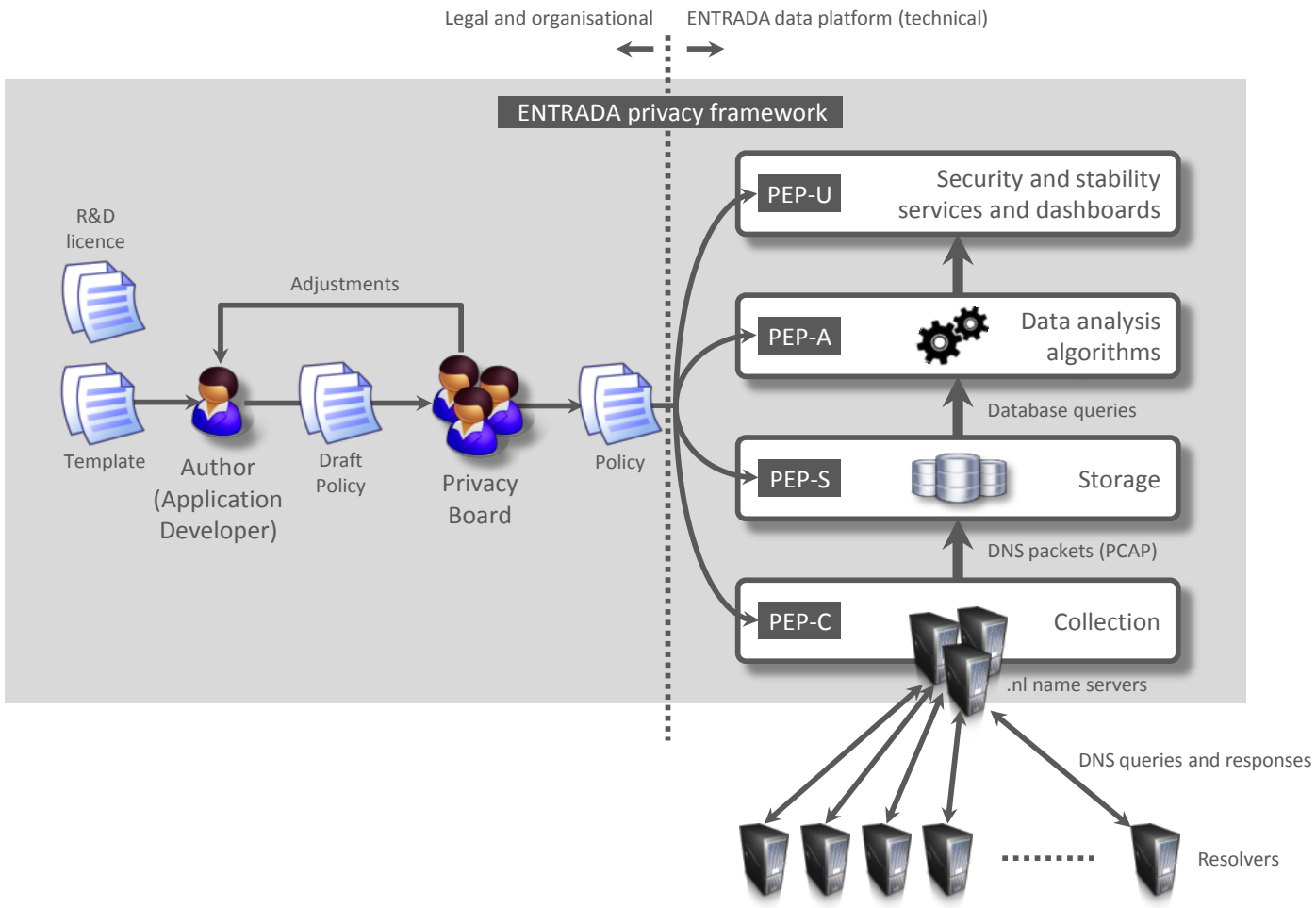
- ‘DNS big data’ system
- Goal: develop applications and services that further enhance the security and stability of .nl, the DNS, and the Internet at large
- ENTRADA main components
  - Applications and services
  - Platform
  - Data sources
  - Privacy framework



SIDN labs
Internet Research & Innovation

SIDN

ENTRADA Privacy Framework



Download paper:
http://goo.gl/GvsfzQ

Policy elements:

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

SIDN labs
Internet Research & Innovation

SIDN

Cluster Design

nano sized

location I
management node

location II
data nodes

location III
data nodes

2Gb/s network

SIDN labs
Internet Research & Innovation

SIDN

Hardware

Management node

HP ProLiant DL380

Xeon 1.9 GHz 12 core CPU

64GB RAM

3 TB storage

Data node

HP ProLiant DL380

Xeon 1.9 GHz 12 core CPU

64GB RAM

6 TB storage

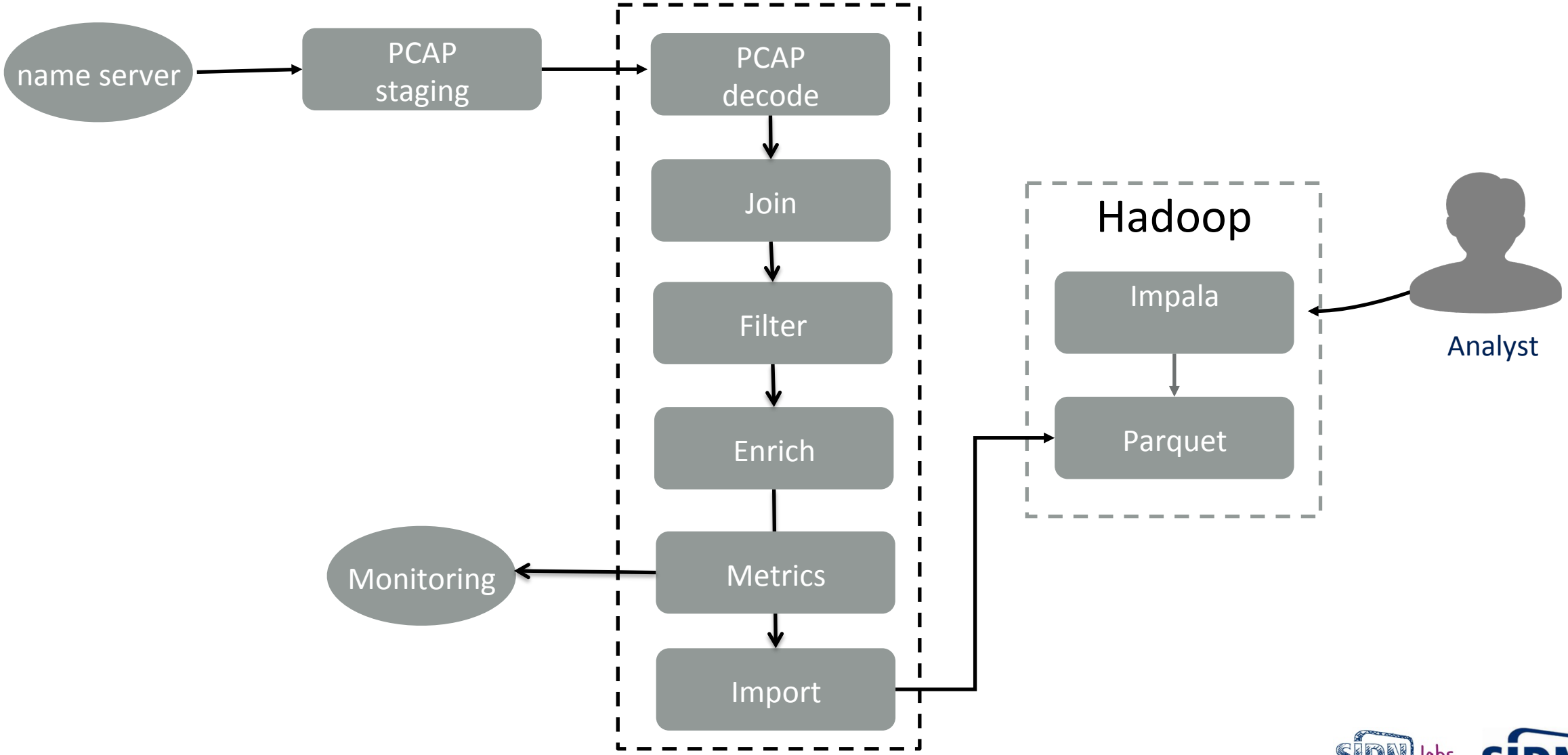
Scaling

- Vertical by adding more resources
- Horizontal by adding more data nodes

SIDN labs
Internet Research & Innovation

SIDN

Workflow



Query data available for analysis within 10 minutes

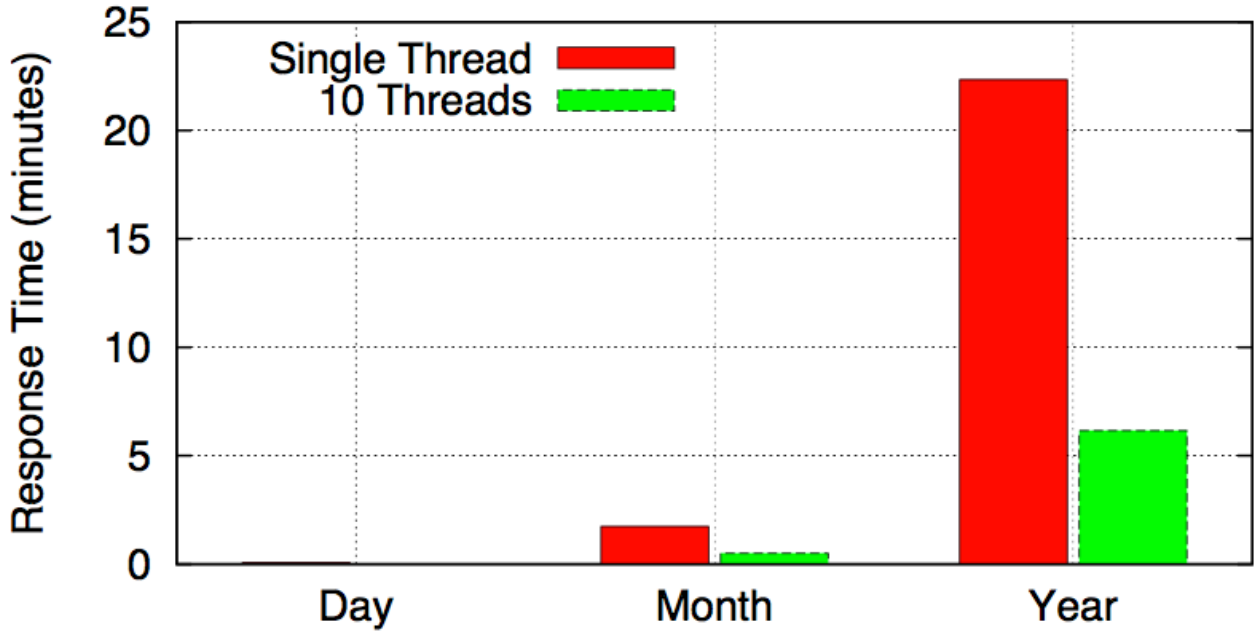
SIDN labs
Internet Research & Innovation

SIDN

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

SIDN labs
Internet Research & Innovation

SIDN

ENTRADA Status

Name server feeds	1
Queries per day	~150M
Daily PCAP volume(gzipped)	~33GB
Daily Parquet volume	~6GB
Months operational	18
Total # queries stored	> 71B
Total Parquet volume	> 3TB
HDFS (3x replication)	> 9TB
Cluster capacity	~150B-200B tuples

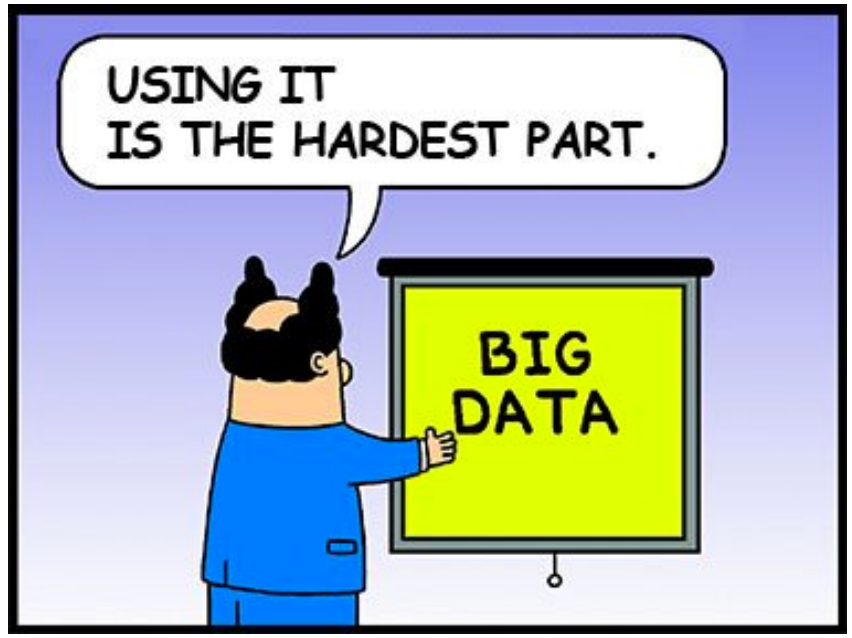
Use Cases

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

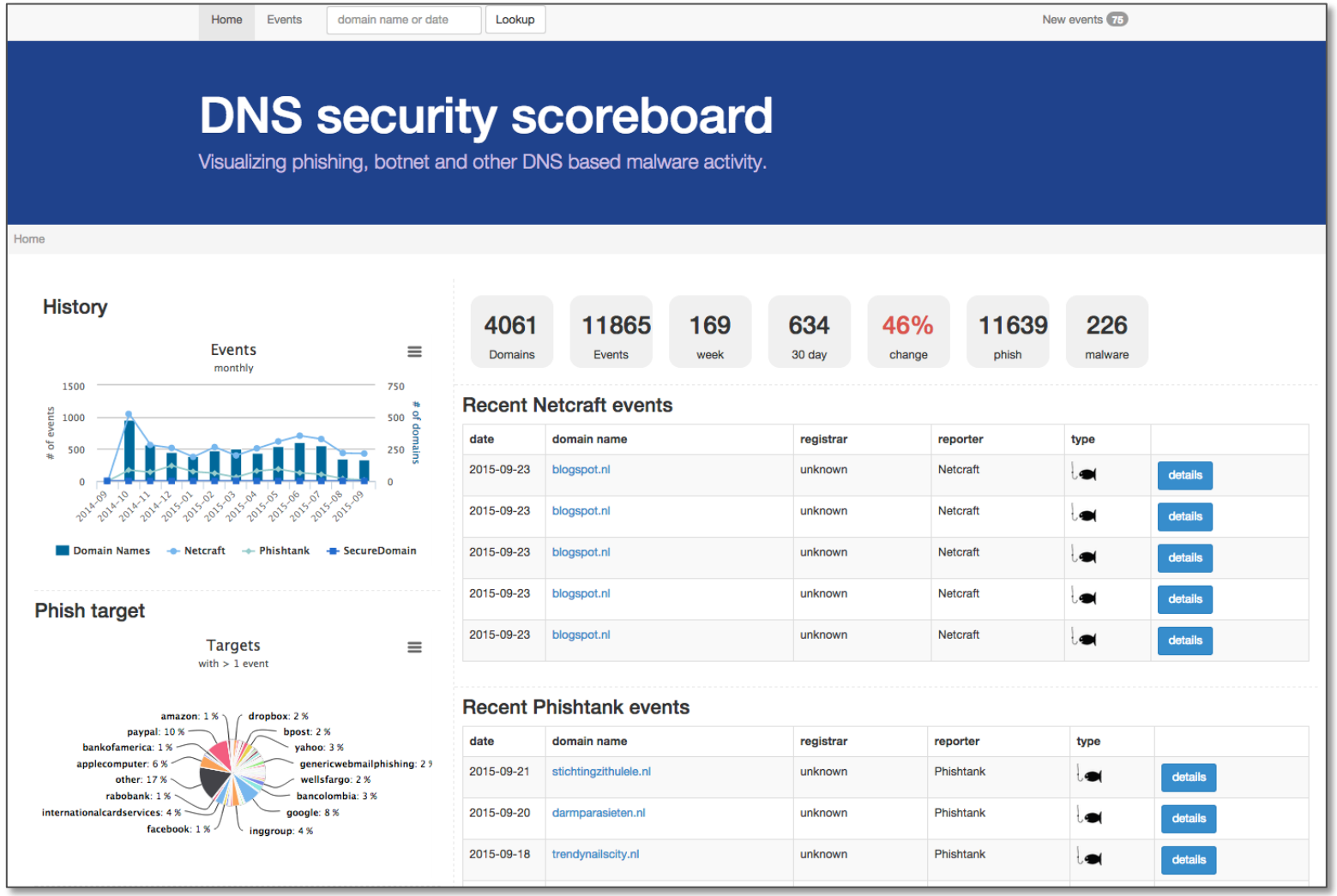
- DNS security scoreboard
- Resolver reputation



DNS Security Dcoreboard

Goal: Visualize DNS patterns for malicious activity

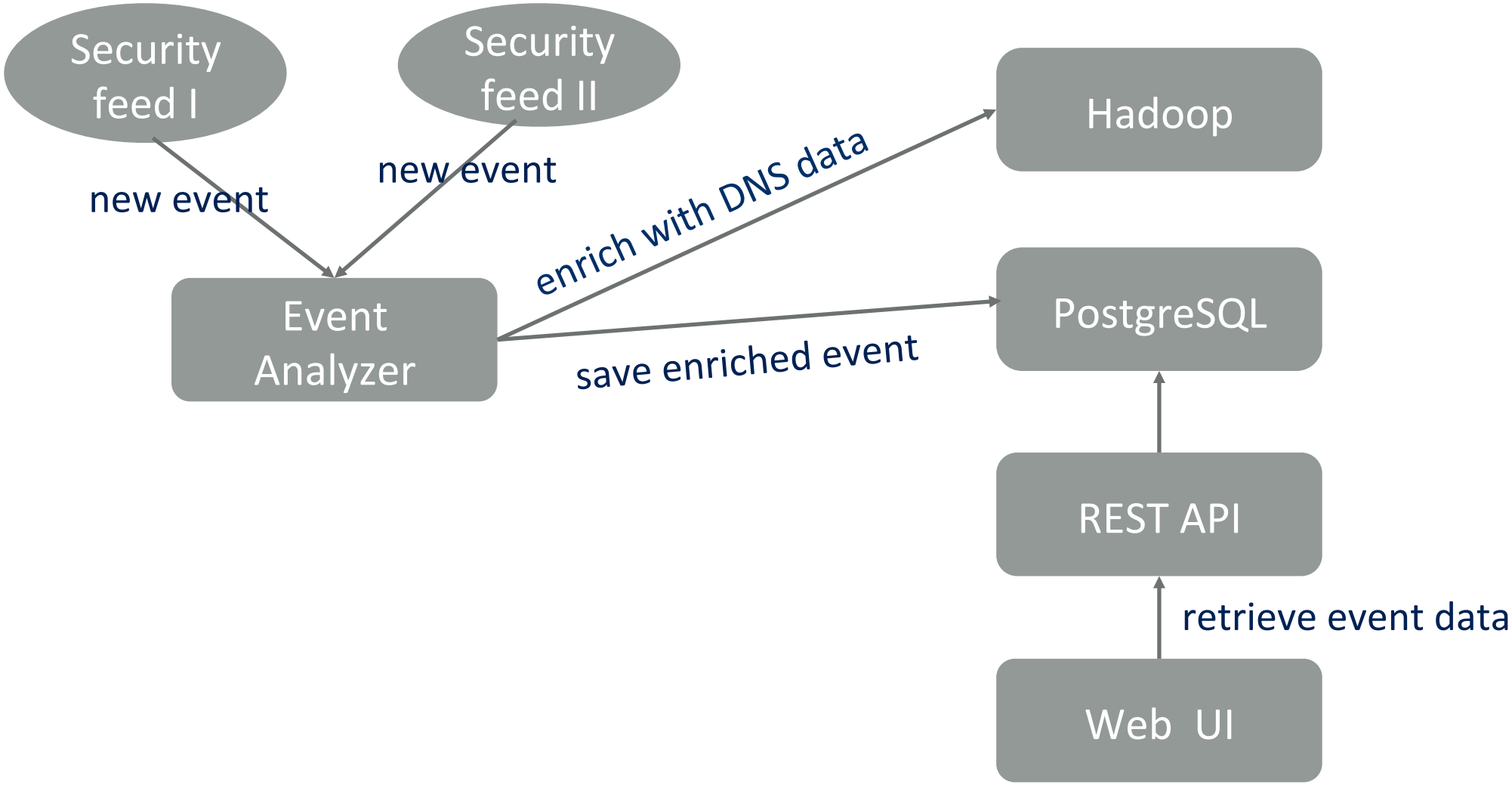
How: Combine external phishing feeds with DNS data



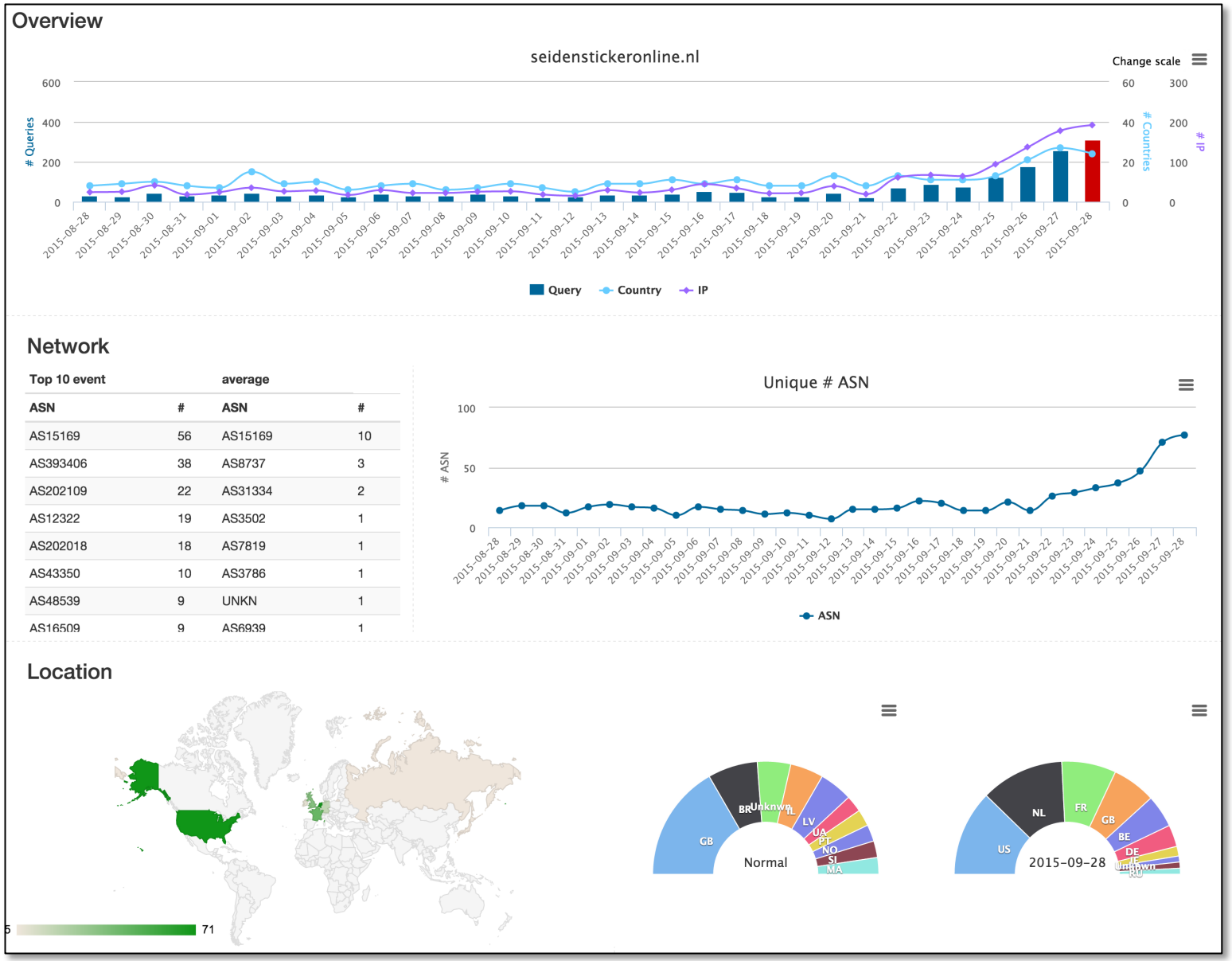
SIDN labs
Internet Research & Innovation

SIDN

Architecture



Traffic Visualization



SIDN labs
Internet Research & Innovation

SIDN

Resolver Reputation (RESREP)

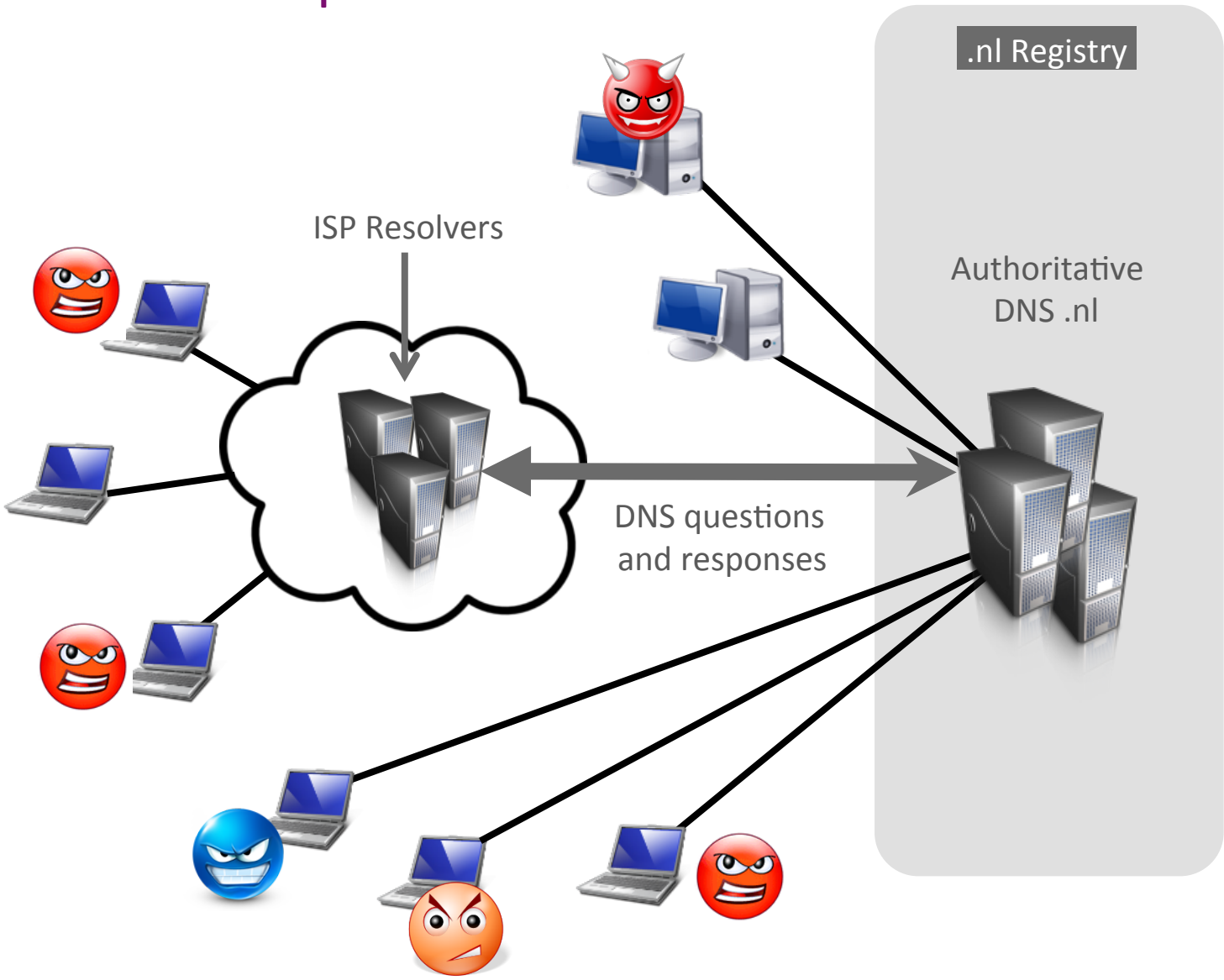
Goal: Try to detect malicious activity by assigning reputation scores to resolvers

How: “fingerprinting” resolver behaviour

SIDN labs
Internet Research & Innovation

SIDN

RESREP Concept



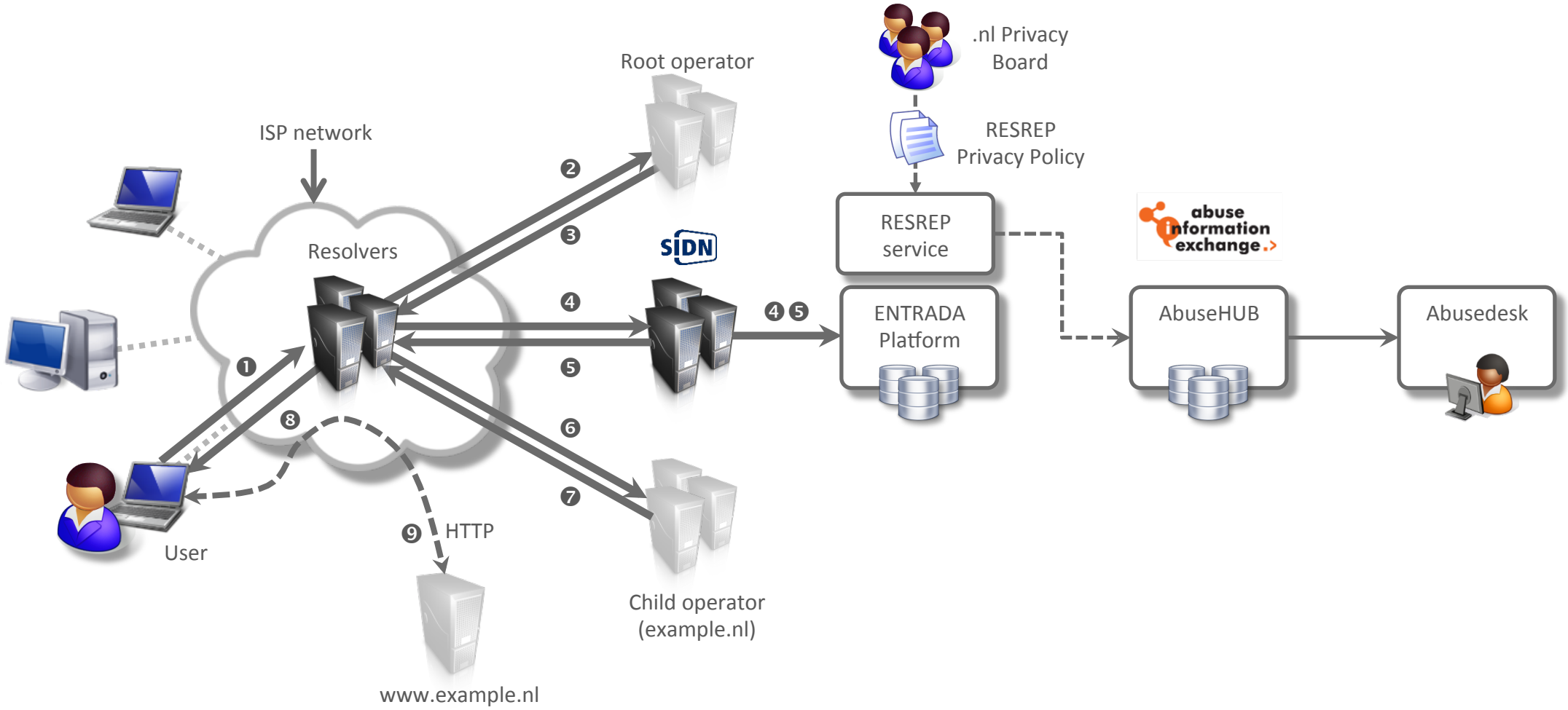
Malicious activity:

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

SIDN labs
Internet Research & Innovation

SIDN

RESREP Architecture



SIDN labs
Internet Research & Innovation

SIDN

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

SIDN labs
Internet Research & Innovation

SIDN

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

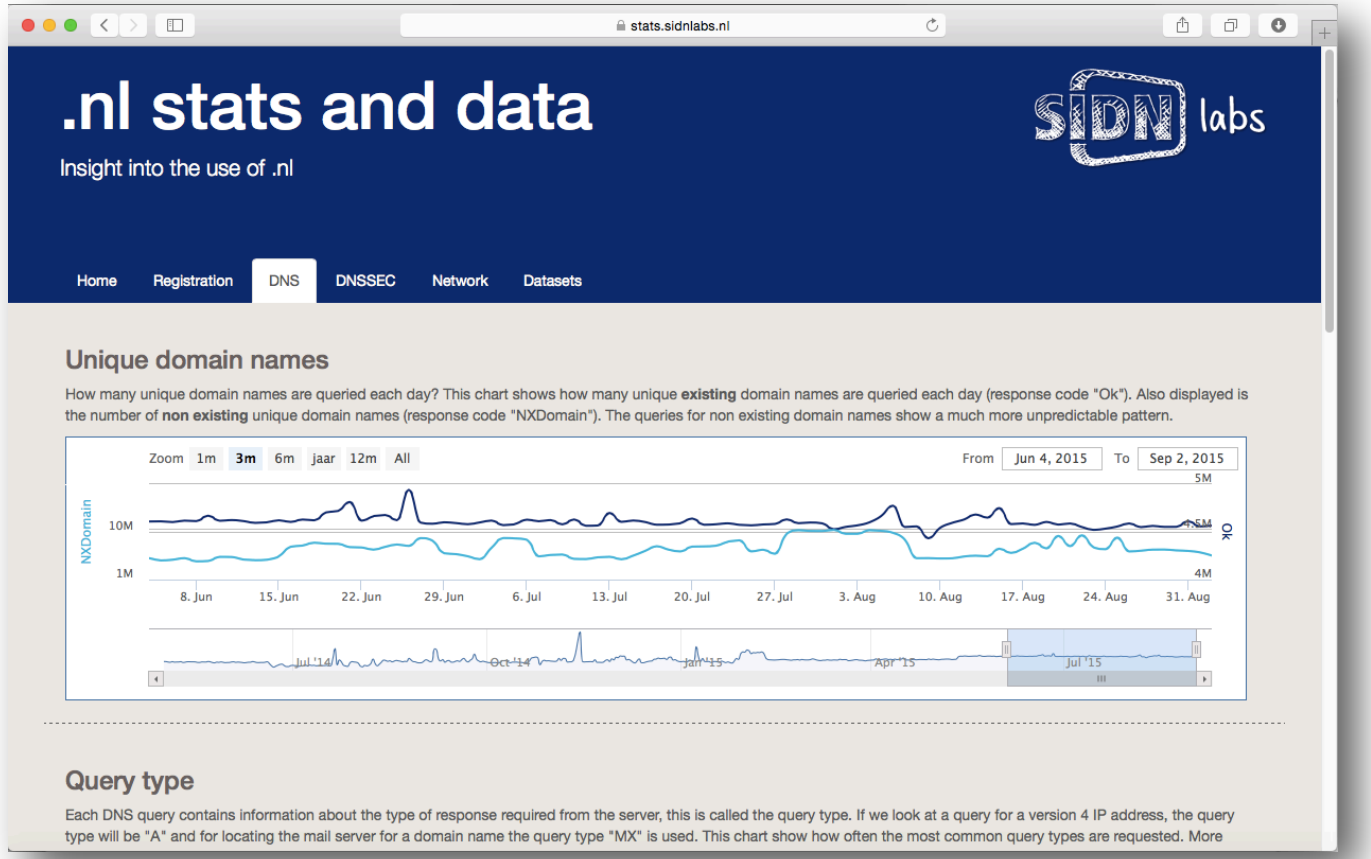
Maarten Wullink

Senior Research Engineer

maarten.wullink@sidn.nl

@wulliak

www.sidnlabs.nl



https://stats.sidnlabs.nl

SIDN labs
Internet Research & Innovation

SIDN