ENTRADA:
An *Open Source* Platform for Network Data Analysis

Moritz Müller | 2nd YETI DNS Workshop

2016-11-12 in Seoul, South Korea

SIDN LABS

SIDN

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

DNS Data @SIDN

> 3.1 million distinct resolvers

> 1.3 billion queries 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

- **Problem**: Existing solutions do not work well with large datasets and have limited analytical capabilities.

Requirements

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

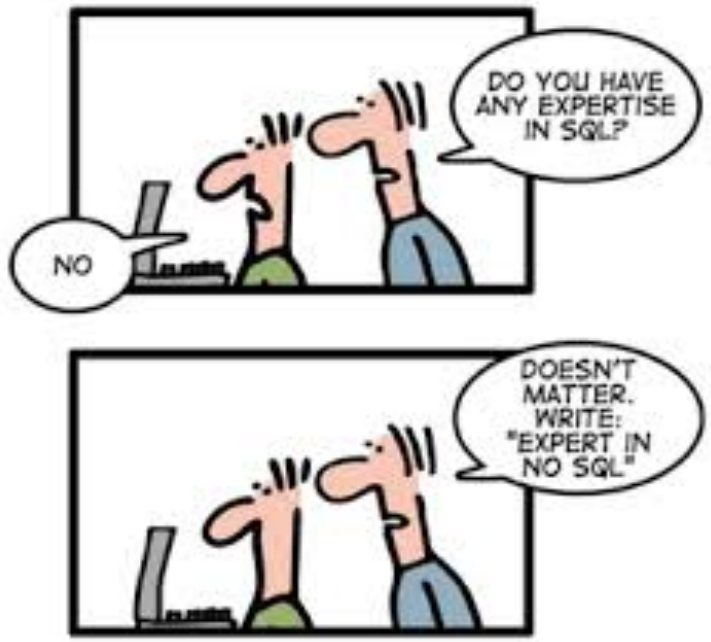
Query Engine Options

Engines galore!

Evaluated SQL and NoSQL solutions

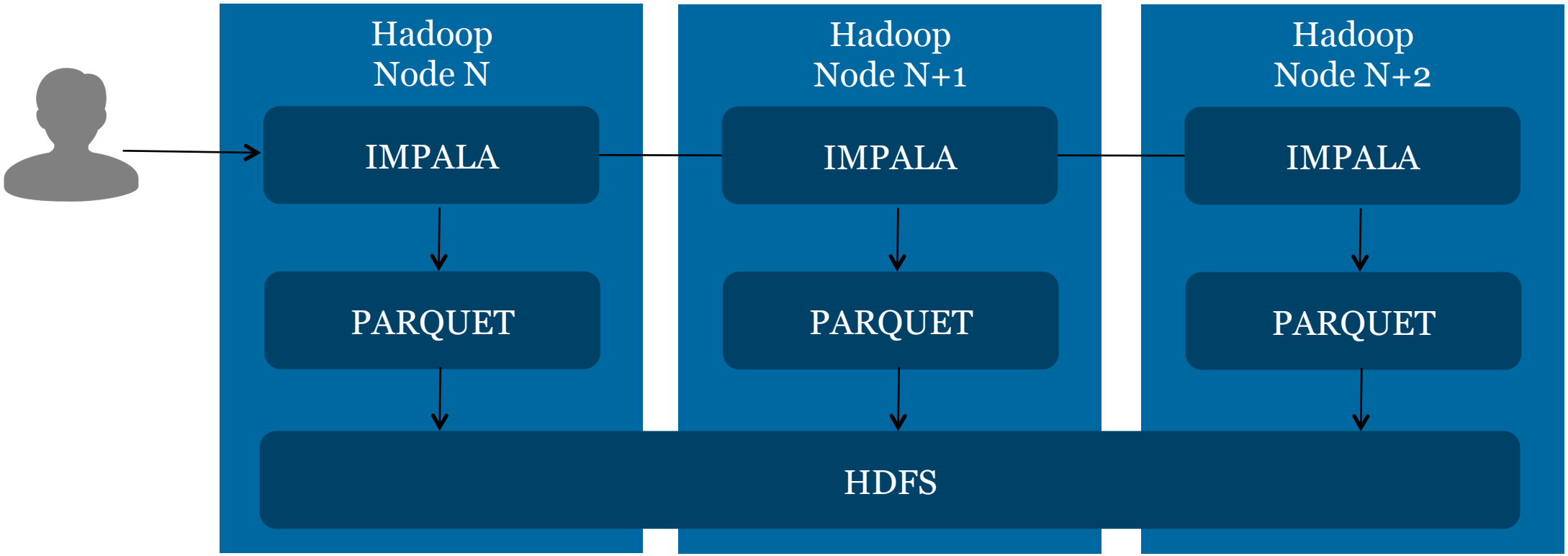
- Relational SQL (PostgreSQL)
- MongoDB
- Cassandra
- Elasticsearch
- Hadoop (HBASE + Apache Phoenix or Hive)

→ **SQL on Hadoop (Impala + Parquet +HDFS)**



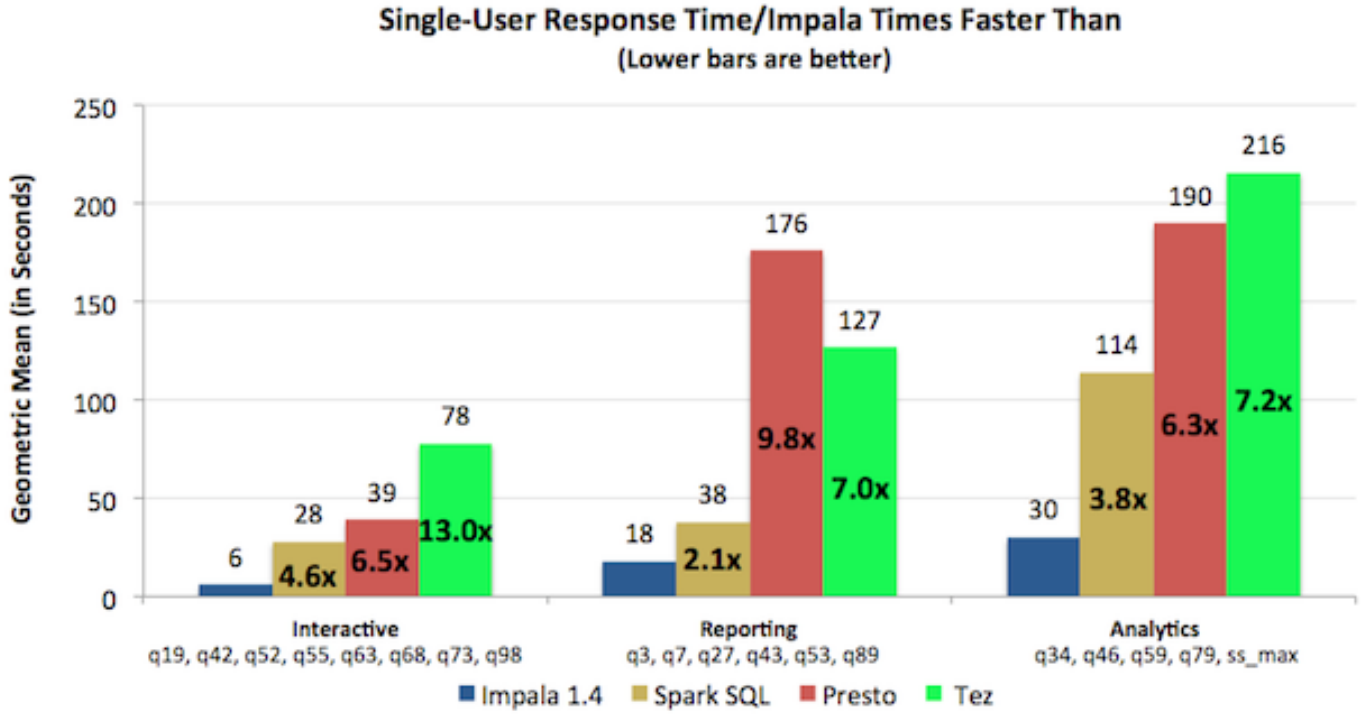
SQL on Hadoop

Best fit for our requirements



Impala Query Engine

- MPP (massively parallel processing)
- Low latency and high concurrency for BI/analytic queries on Hadoop
- Excellent performance compared to other Hadoop based query engines



Impala (2)

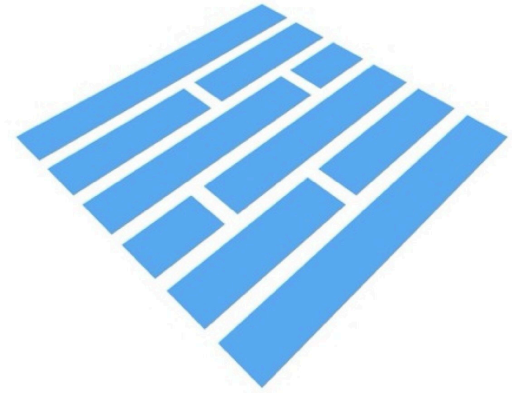
Data formats

- Text
- Hadoop formats
- Apache Avro
- Apache Parquet

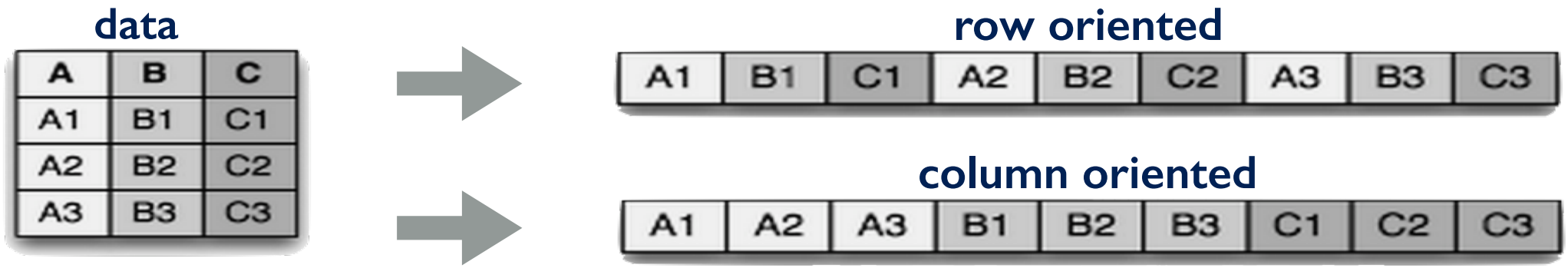
Interfaces

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

Apache Parquet

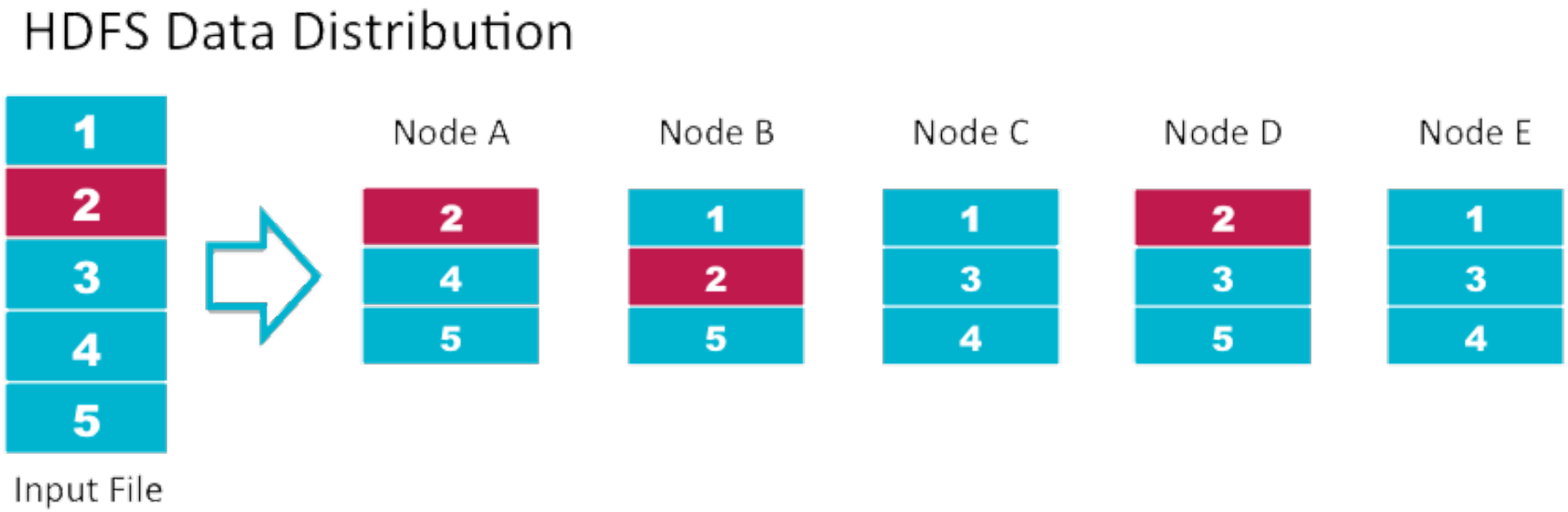


- Why not just use the PCAP files?
  - Reading (compressed) PCAP data is just too slow
  - Analytical engines cannot read PCAP files



HDFS

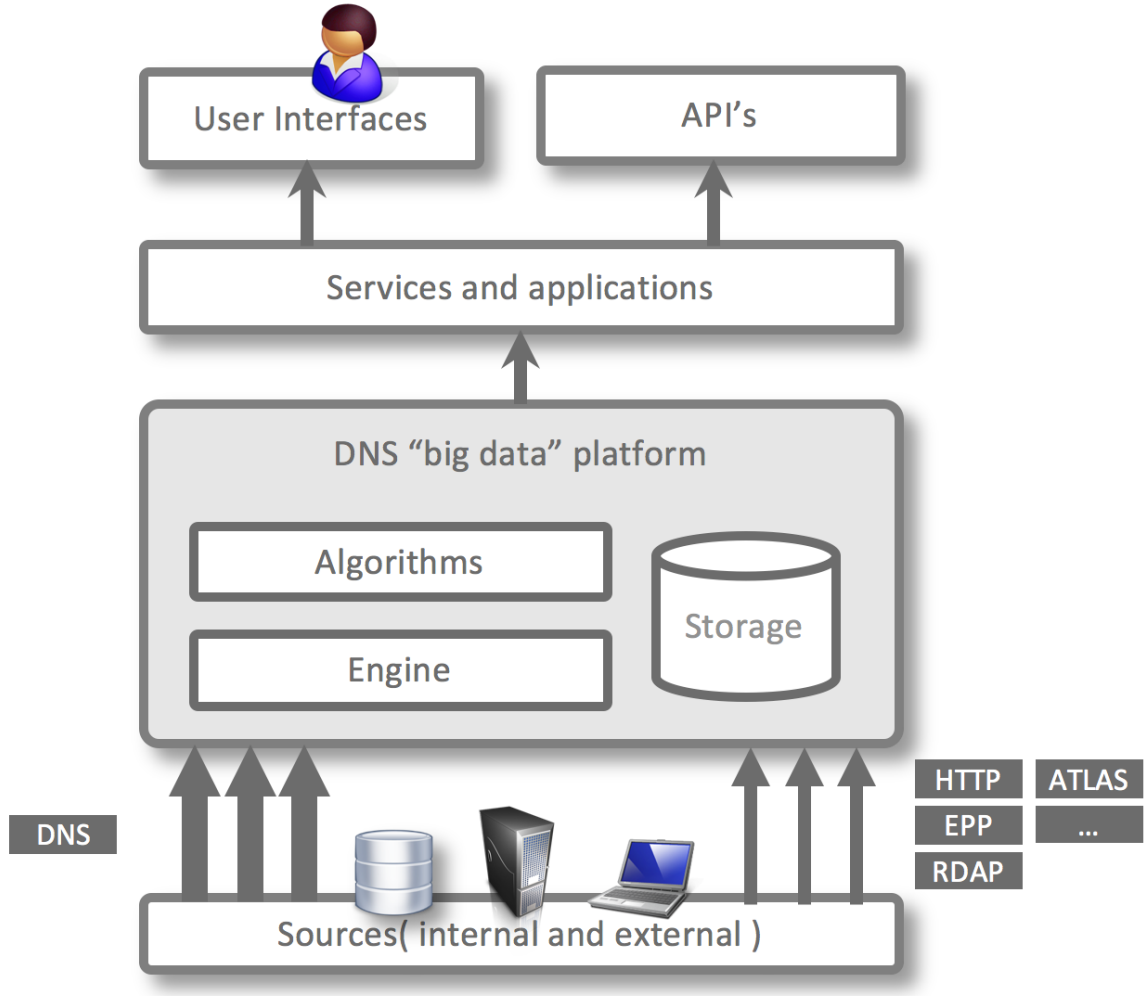
- Distributed file system for storing large volumes of data
- High availability through replication of data blocks
- Scalable to hundreds of PB's and thousands of servers



ENTRADA Architecture

Main components

- Data sources
- Platform
- Applications and services
- Privacy framework



Cluster Design

nano sized

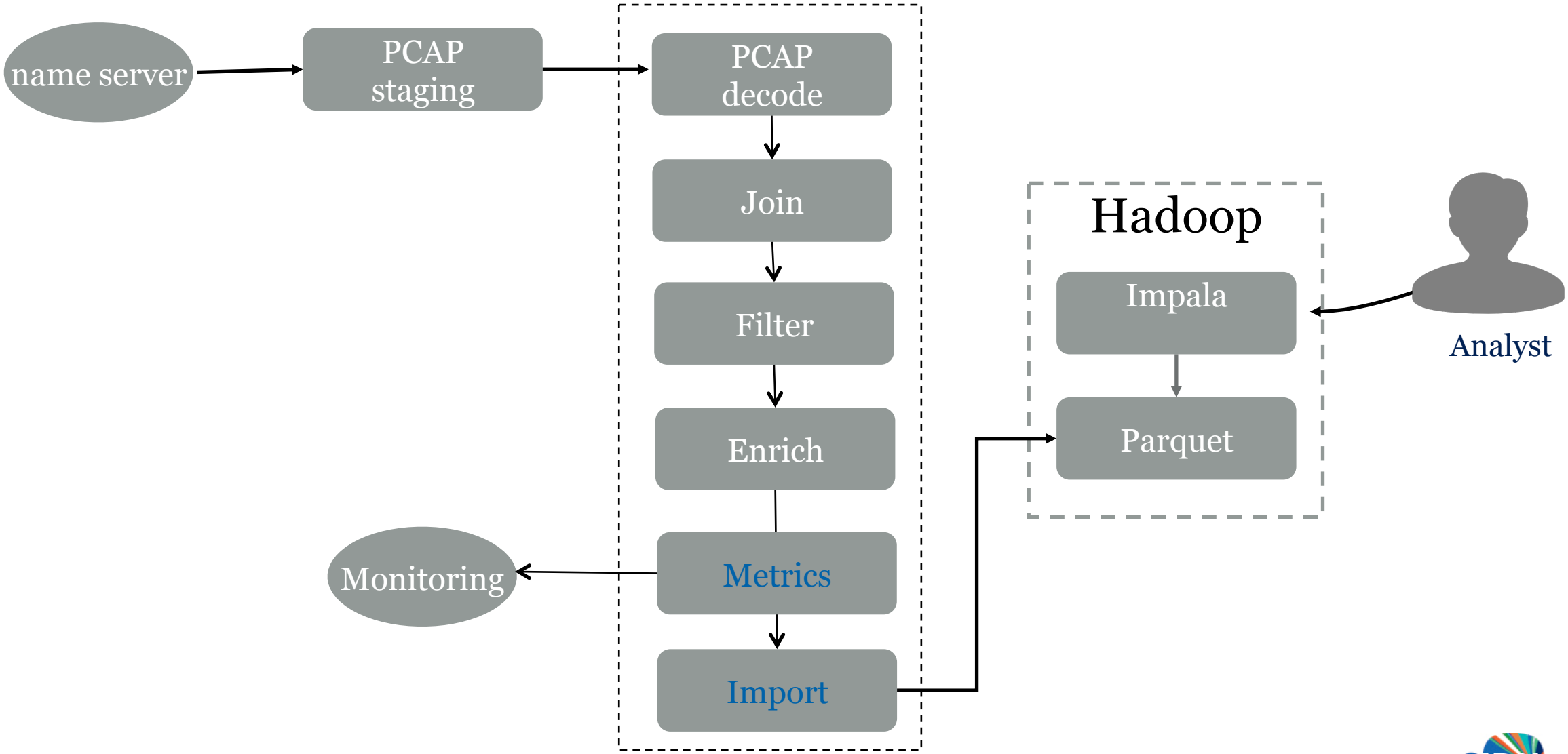
location I
management node

location II
data nodes

location III
data nodes

2Gb/s network

Workflow

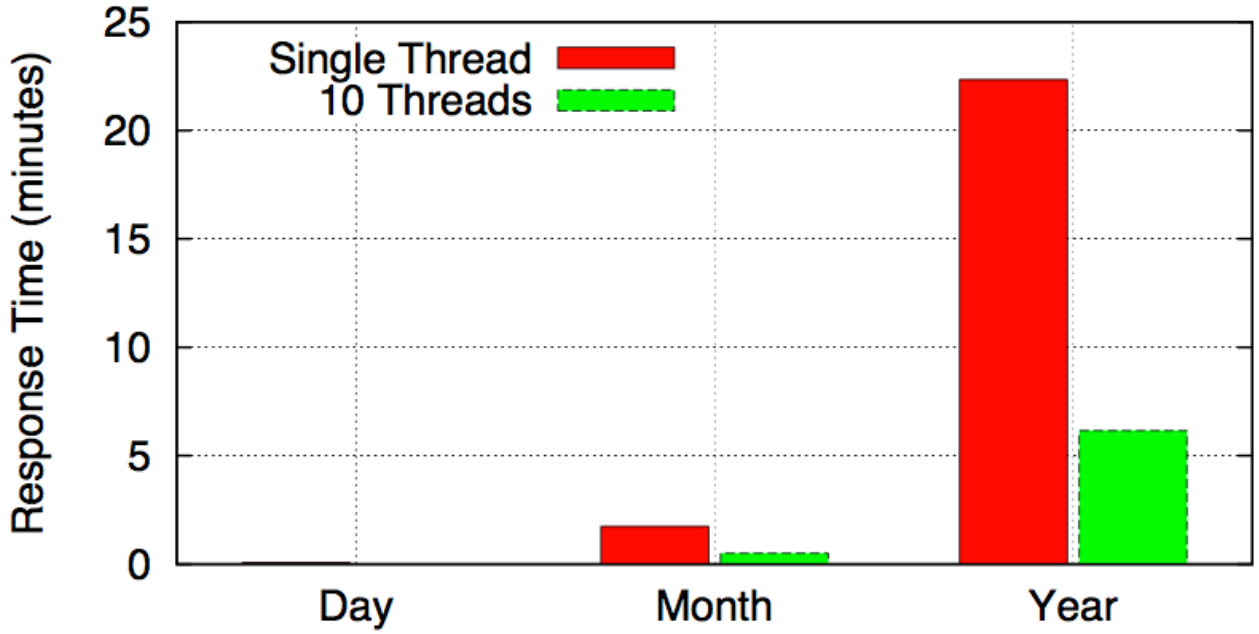


Query data available for analysis within 10 minutes

Performance

Example query, count # ipv4 queries per day.

```
Select day, month, year, count(1)
from dns.queries
where ipv=4
group by day, month, year
```



Query response times

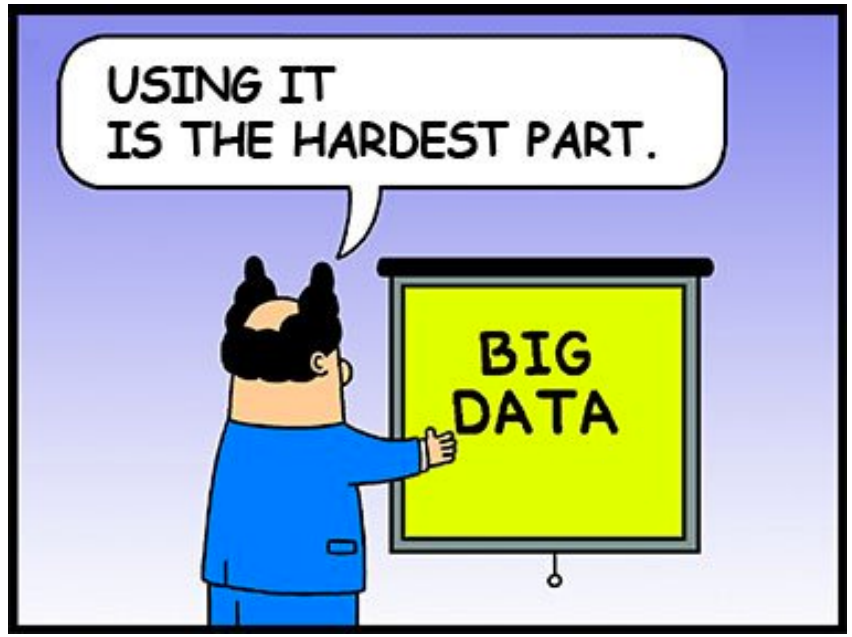
1 Year of data is 2.2TB Parquet ~ 52TB of PCAP

SIDN LABS

Use Cases

Focussed on increasing the security and stability of .nl

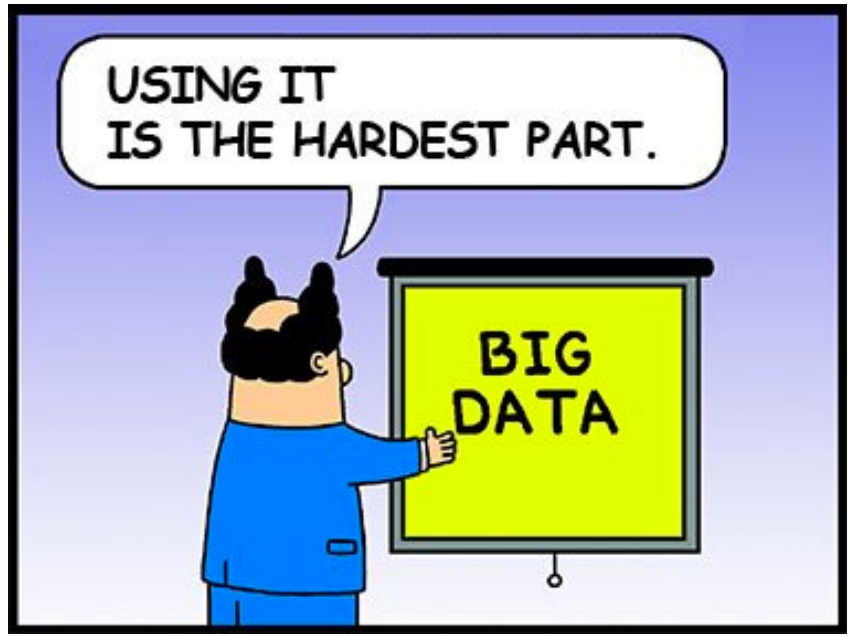
- Visualize DNS patterns
- Statistics (stats.sidnlabs.nl)
- Scientific research
- Support for operators
- Real-time Phishing detection
- Detect botnet infections



Use Cases

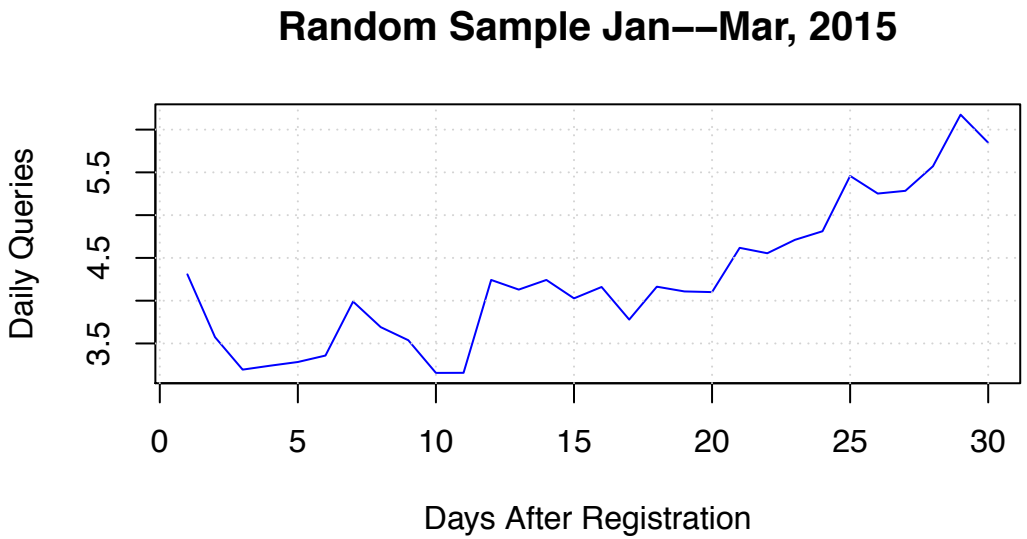
Focussed on increasing the security and stability of .nl

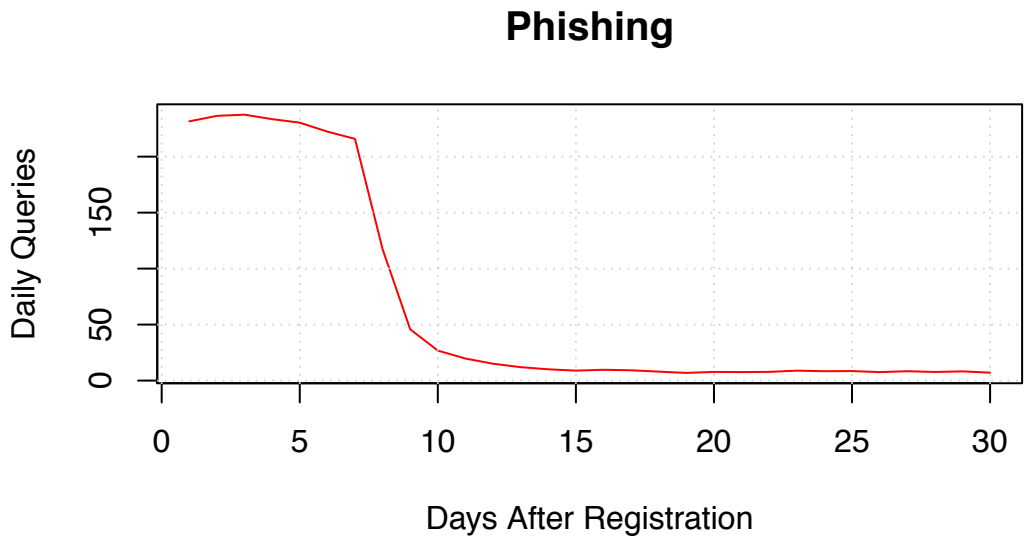
- Visualize DNS patterns
- Statistics (stats.sidnlabs.nl)
- Scientific research
- Support for operators
- **Real-time Phishing detection**
- **Detect botnet infections**



Malicious Domain Detection with nDEWS

Observation: Phishing domains have unique query patterns

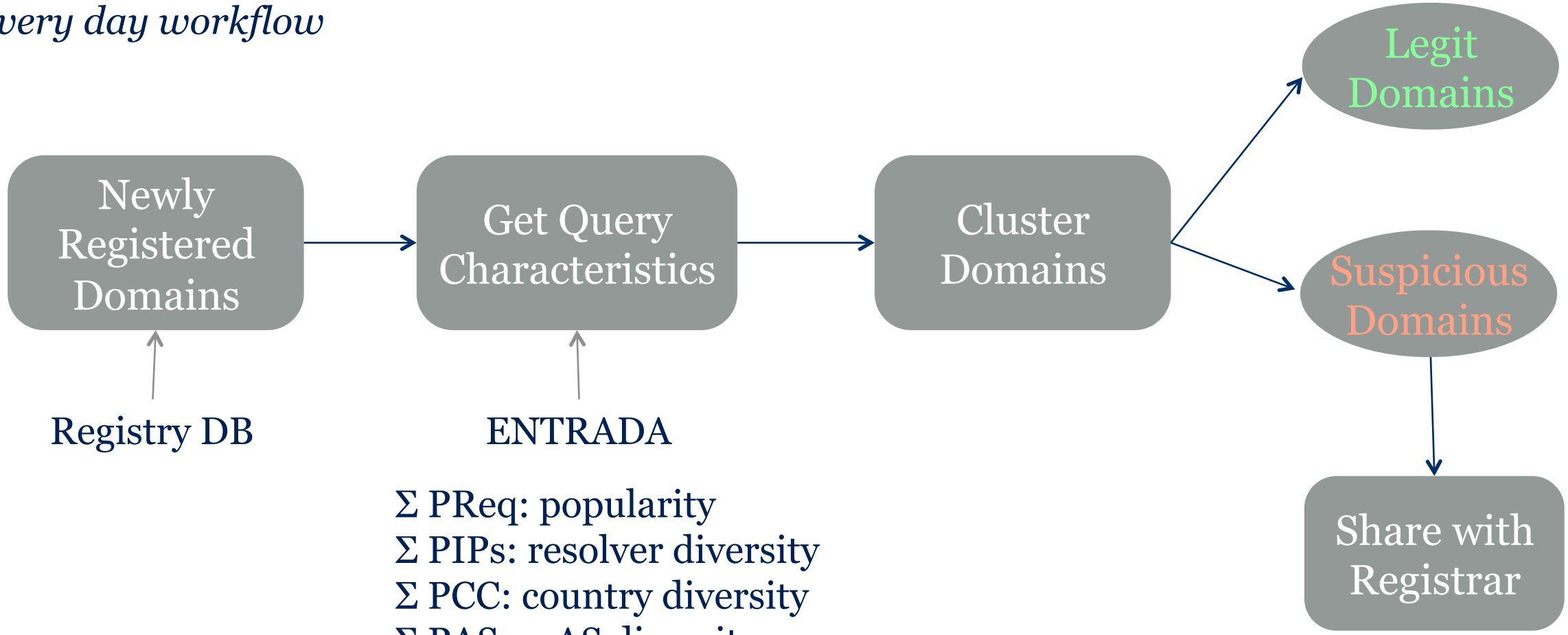




Ref: Moura, G.C. M., Muller, M., Wullink, M, Hesselman, C.: nDEWS:
a new domains early warning system for TLDs . In: IEEE/IFIP AnNet 2016

nDEWS Architecture

Every day workflow

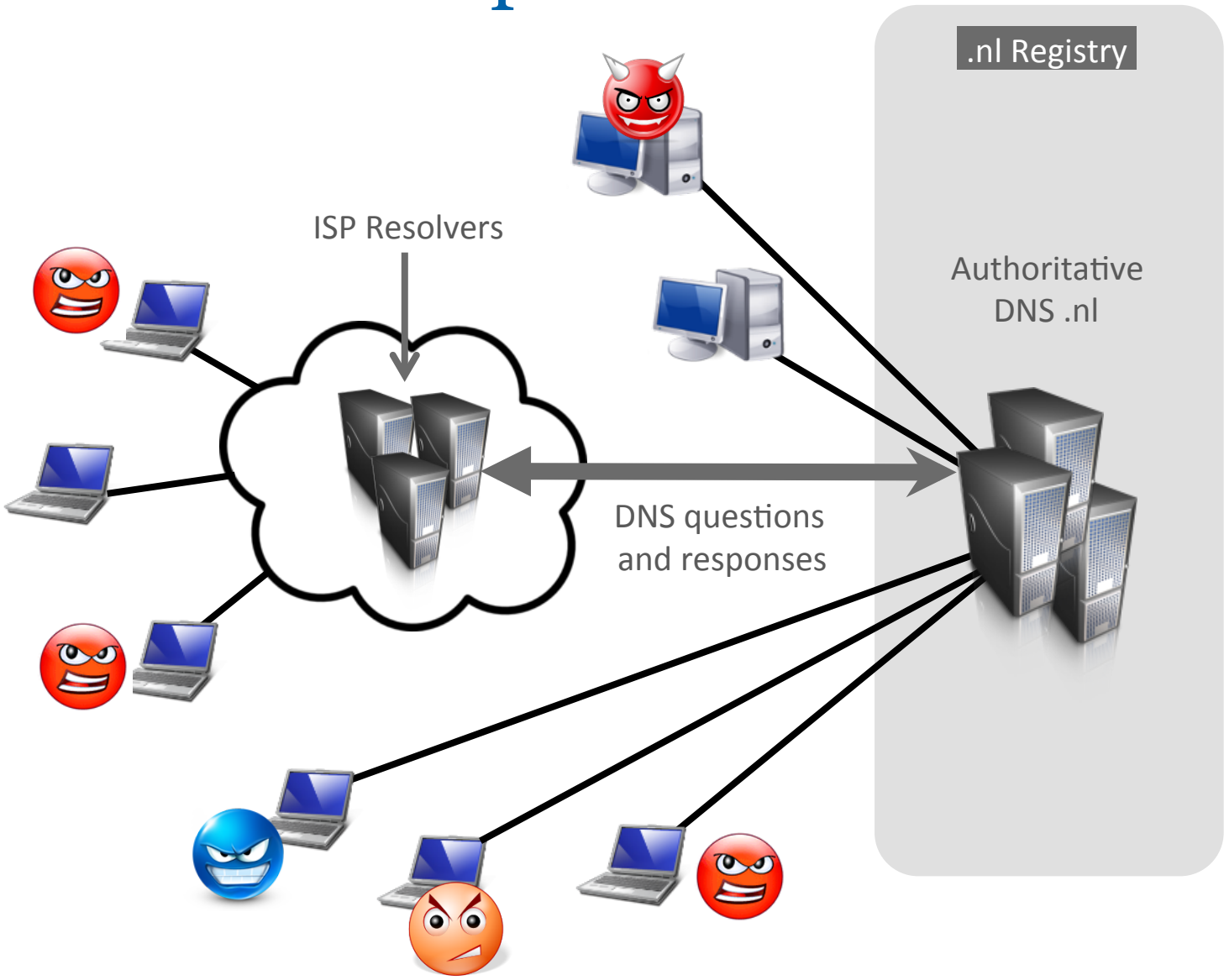


Resolver Reputation (RESREP)

Goal: Detect malicious activity by assigning reputation scores to resolvers

How: “fingerprinting” resolver behaviour

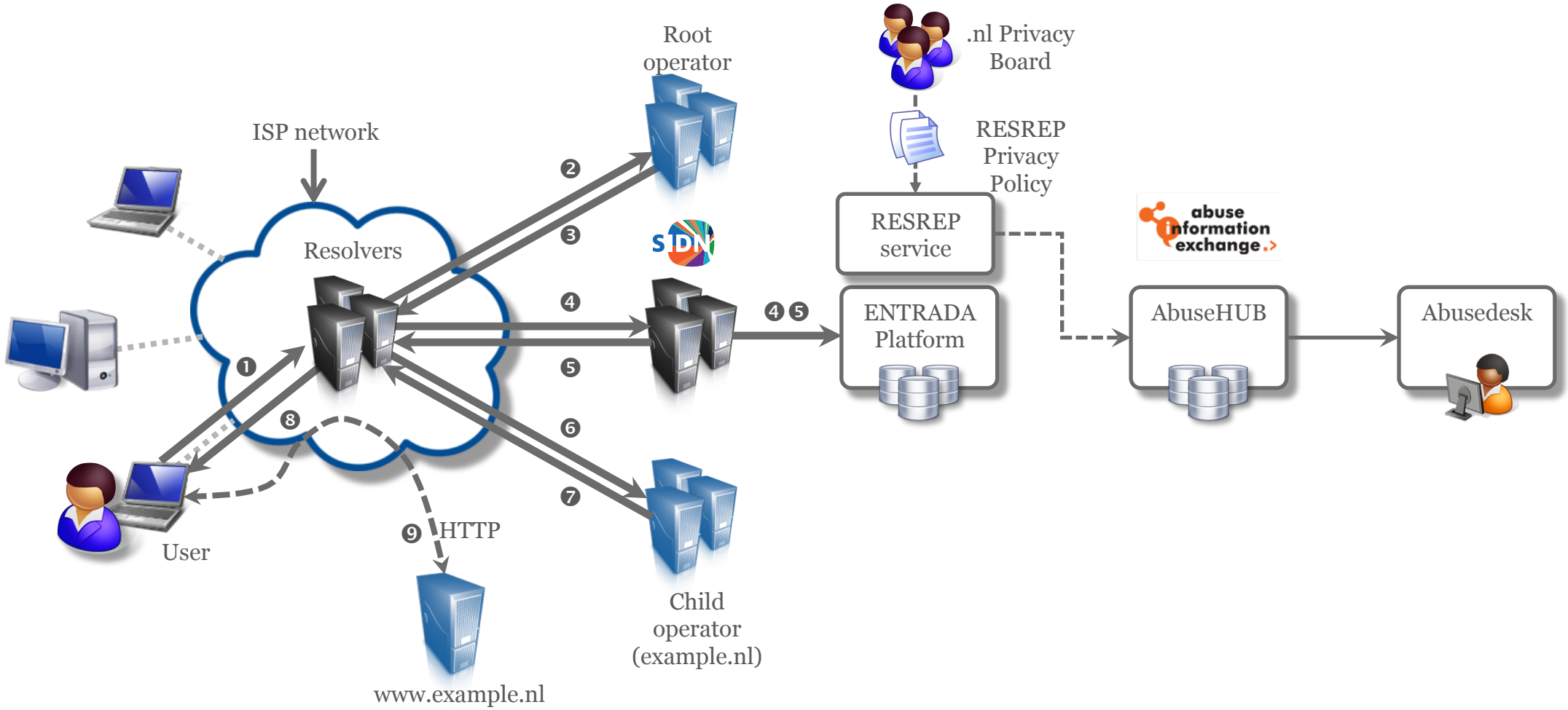
RESREP Concept



Malicious activity:

- Spam-runs
- Botnets
- DNS-amplification attacks

RESREP Architecture



Conclusions and Future Work

- Hadoop HDFS + Parquet + Impala is a winning combination!
- Running for over 2 years
- > 150 billion queries stored

- Develop more use cases
  - DNS Abuse
  - Operational support
- Increase the number of ENTRADA users

It’s open source!

- Since January 2016

- Project site:

entrada.sidnlabs.nl

- GitHub:

github.com/SIDN/ENTRADA/

- 6 registries are using it

Questions? Feedback?

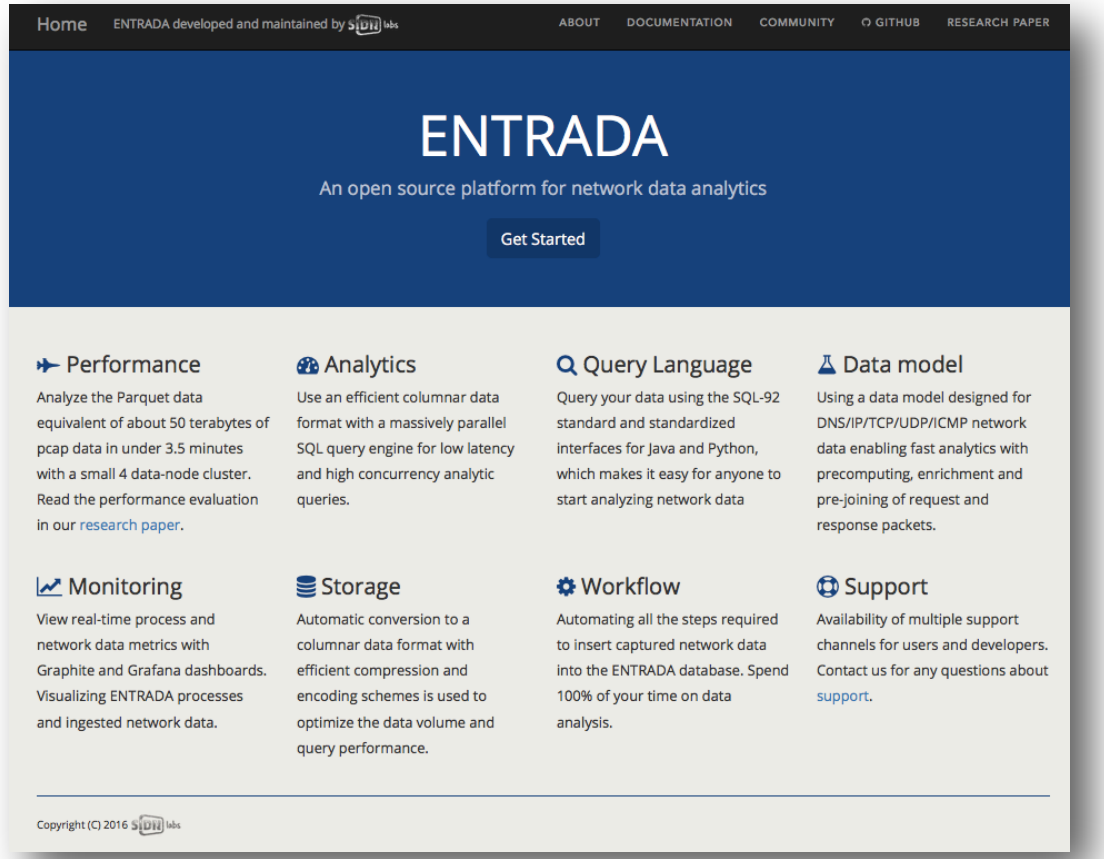
Moritz Müller

Research Engineer

moritz.muller@sidn.nl

@dhr_moe

www.sidnlabs.nl



entrada.sidnlabs.nl