

ENTRADA:
An open source platform
for network data analysis

Moritz Müller | Netnod Spring Meeting 2016-04-17



SIDN

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

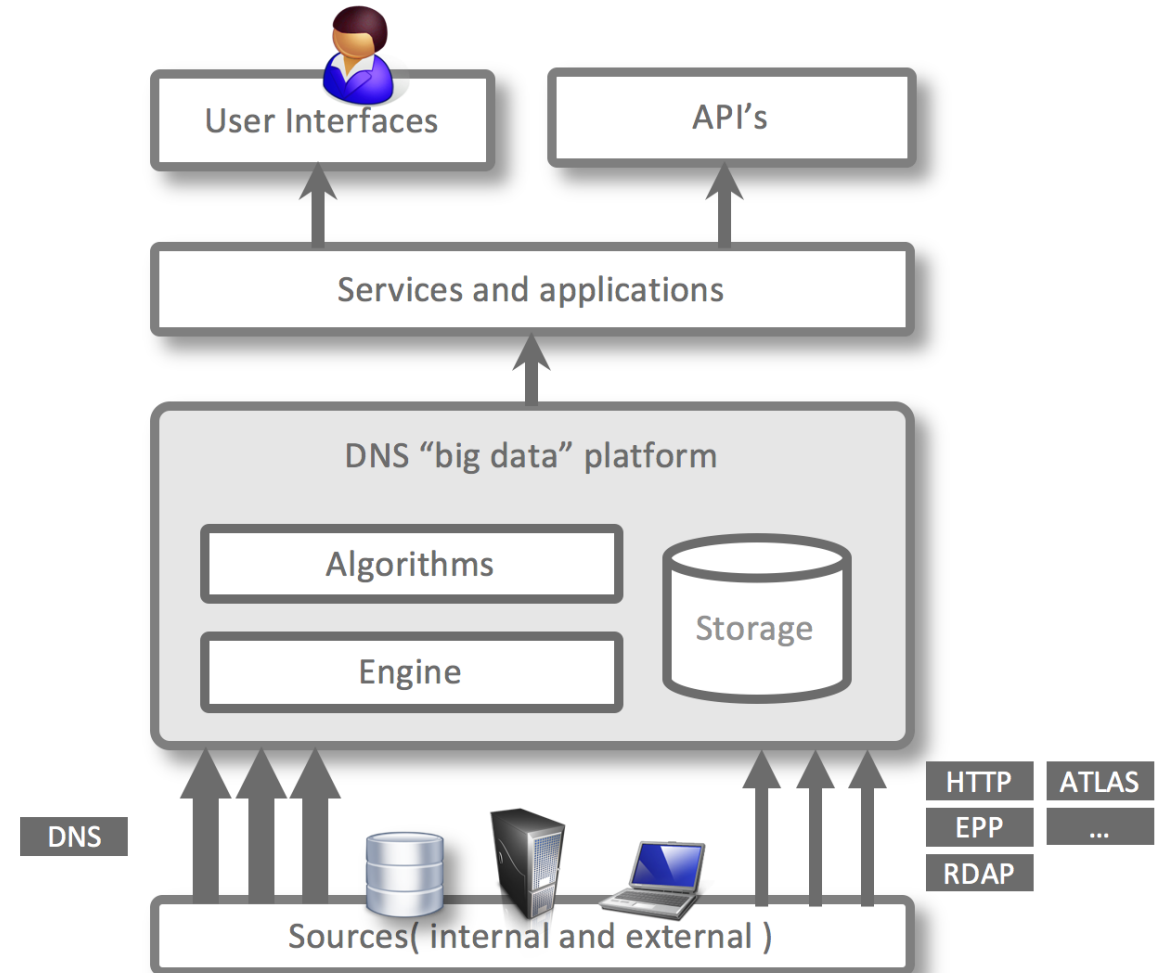
Requirements

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

ENTRADA Architecture

Main components

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

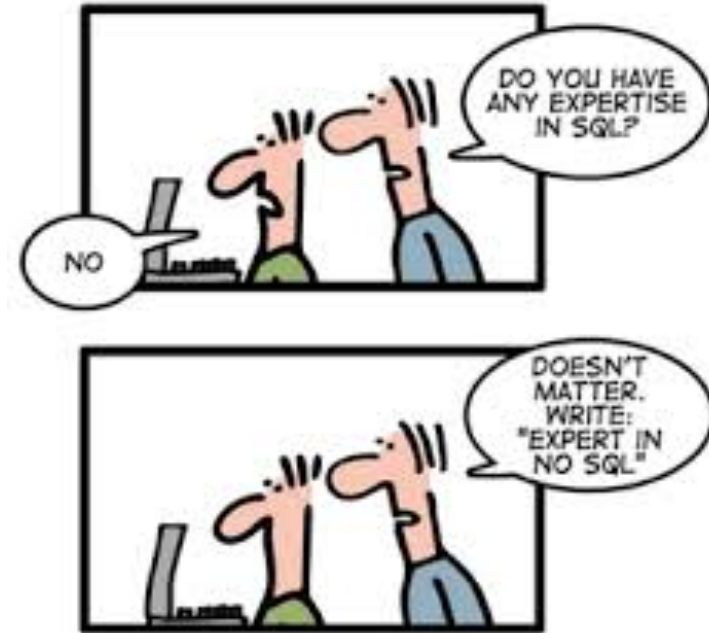


Query Engine Options

Engines galore!

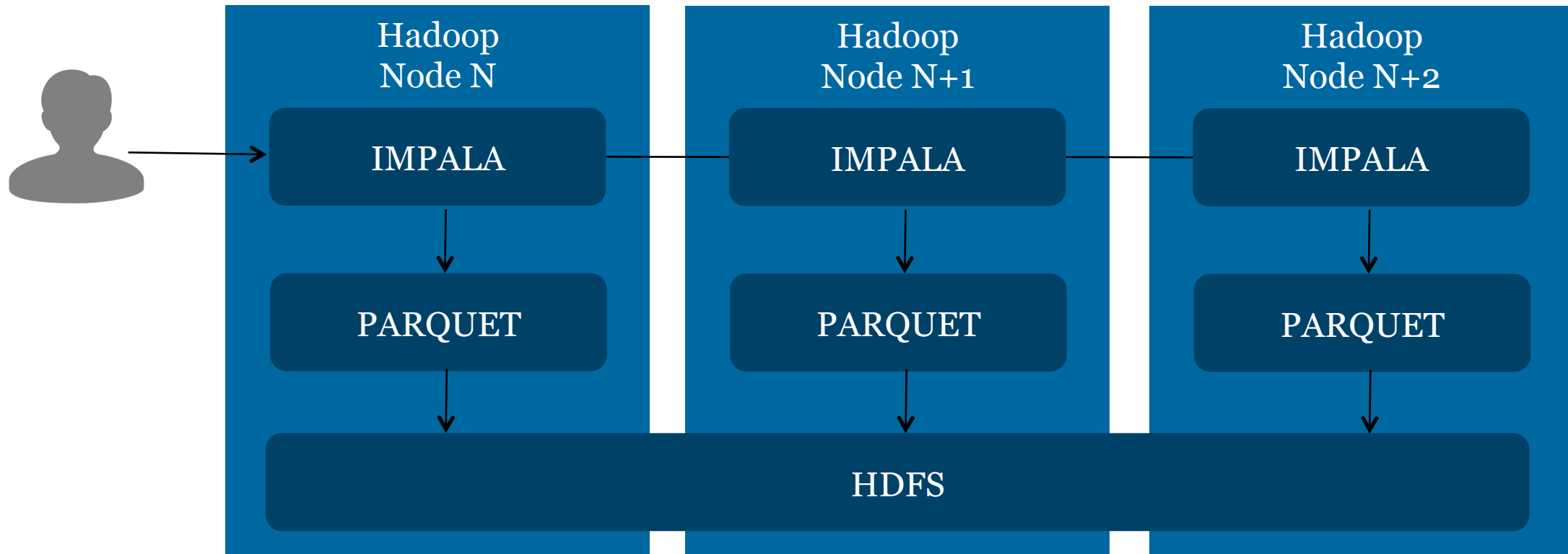
Evaluated SQL and NoSQL solutions

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



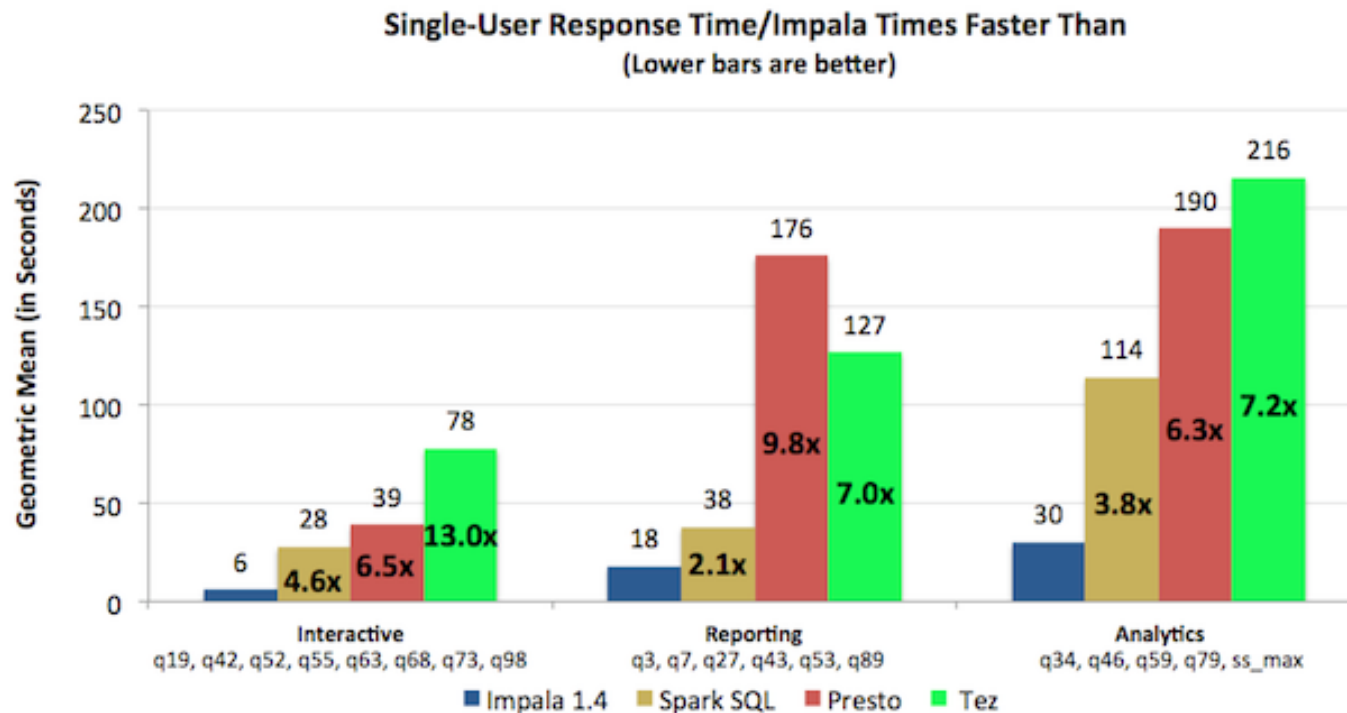
SQL on Hadoop

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

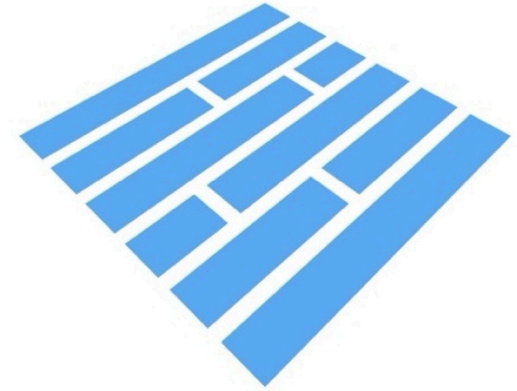
- Text
- Hadoop formats
- Apache Avro
- Apache Parquet

Interfaces

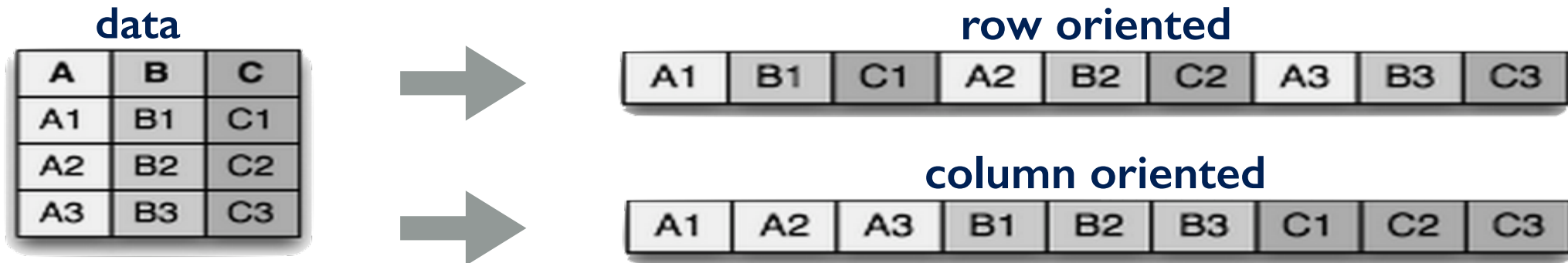
- Web-based GUI
- Command line (impala-shell)
- Python (Impyla)
- JDBC



Apache Parquet



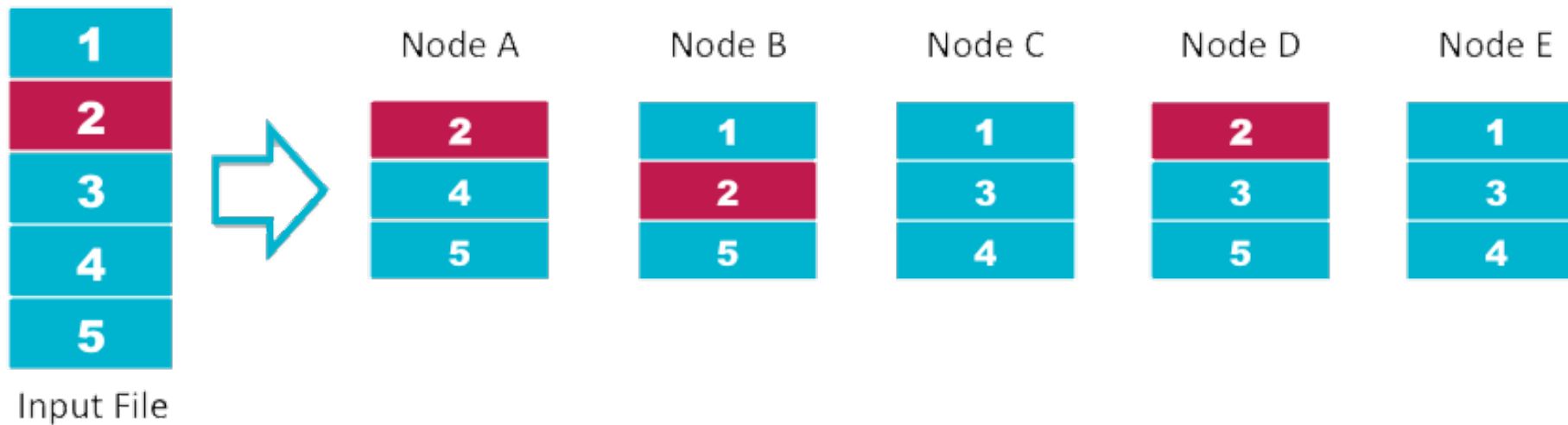
- Why not just use the PCAP files?
 - Reading (compressed) PCAP data is just too slow
 - Analytical engines cannot read PCAP files



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

HDFS Data Distribution



Cluster Design

nano sized

location I
management node



location II
data nodes



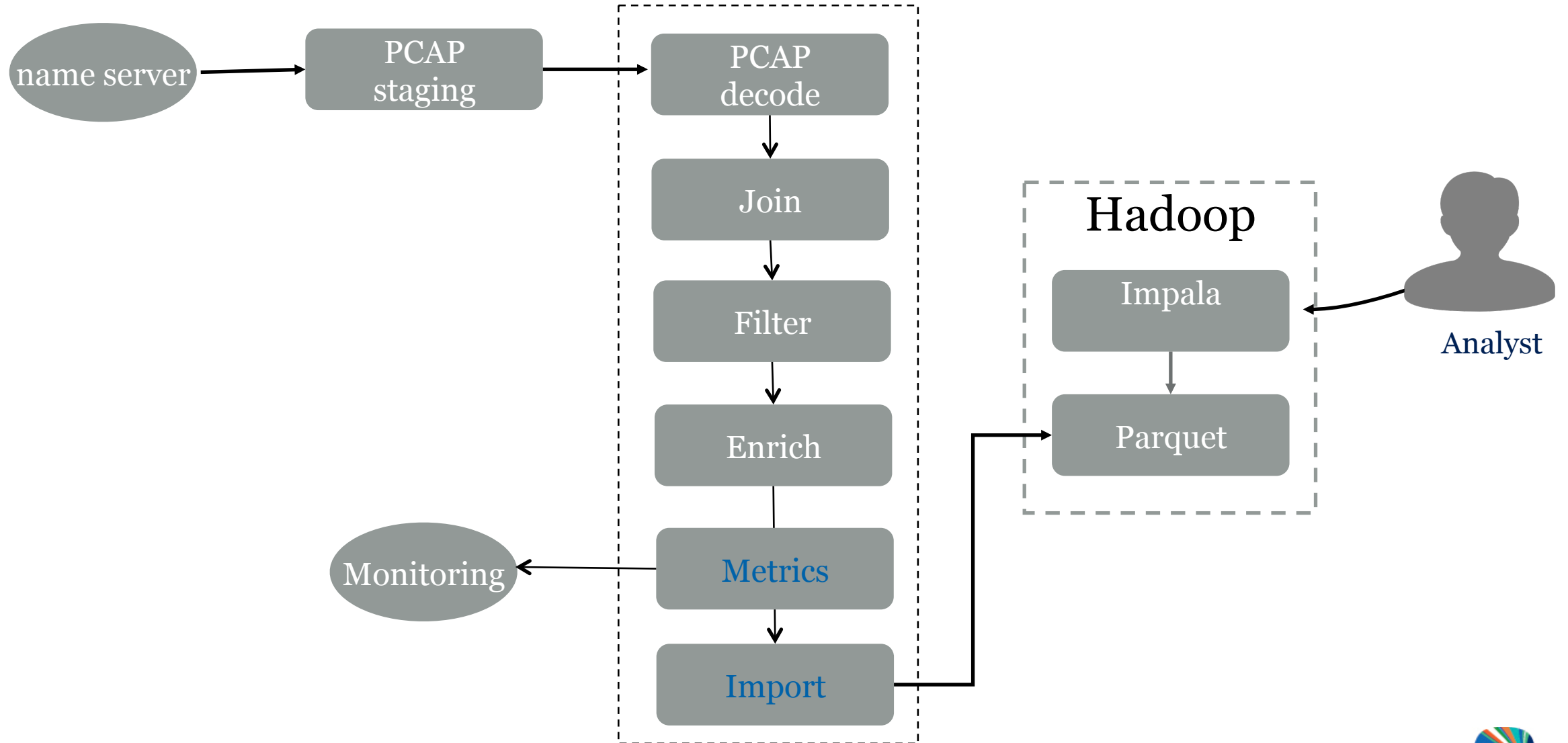
location III
data nodes



2Gb/s network



Workflow

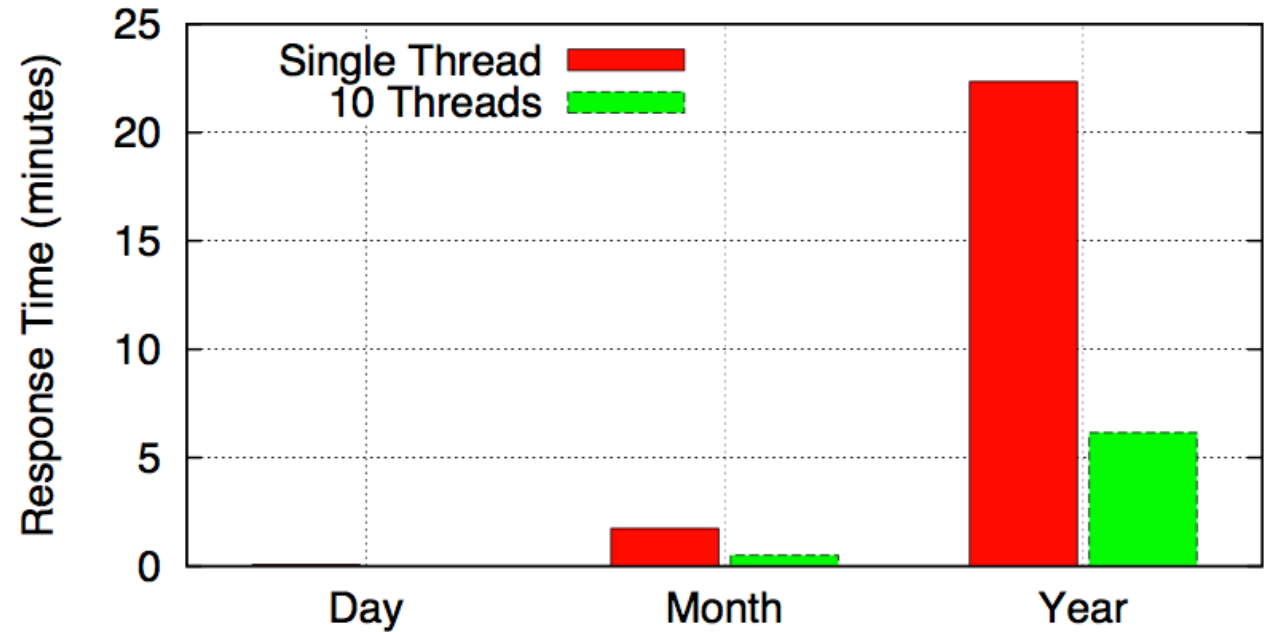


Query data available for analysis within 10 minutes

Performance

Example query, count # ipv4 queries per day.

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

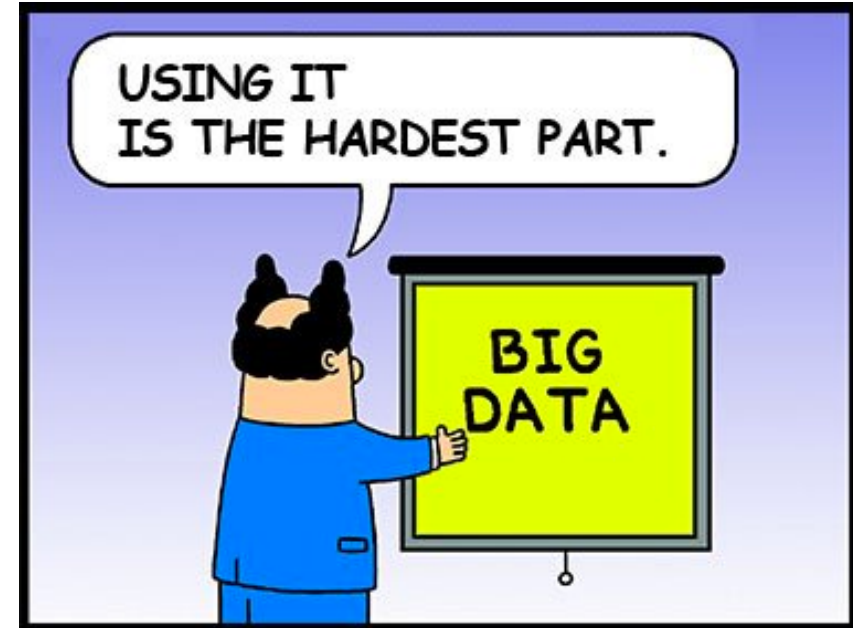


1 Year of data is 2.2TB Parquet ~ 52TB of PCAP

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

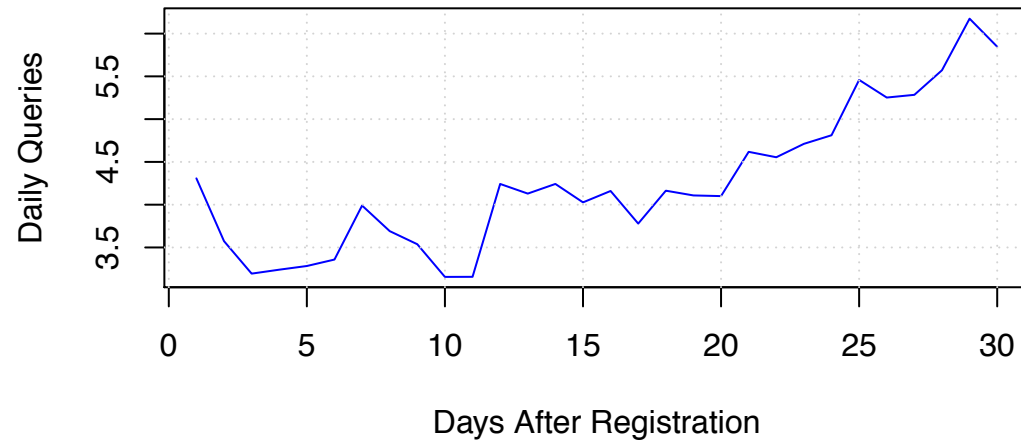
- Visualize DNS patterns
- Statistics (stats.sidnlabs.nl)
- Scientific research
- Support for operators
- **Real-time Phishing detection**
- **Detect botnet infections**



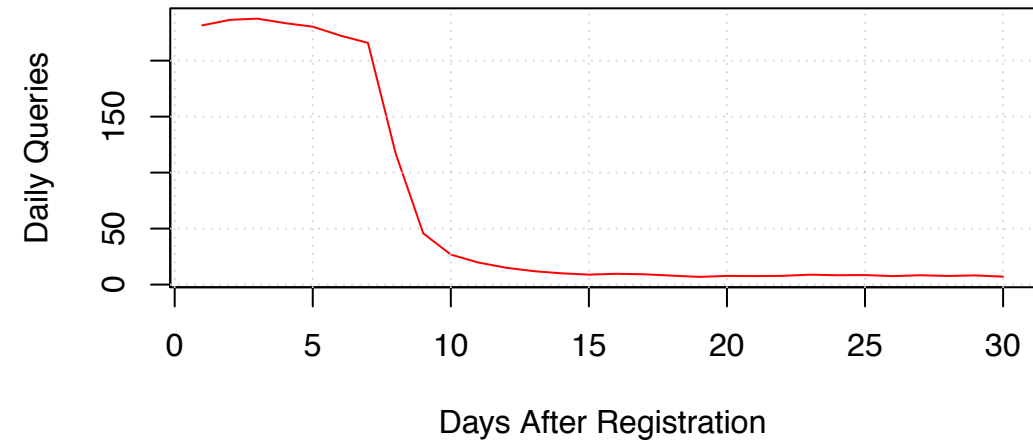
Malicious Domain Detection with nDEWS

Observation: Phishing domains have unique query patterns

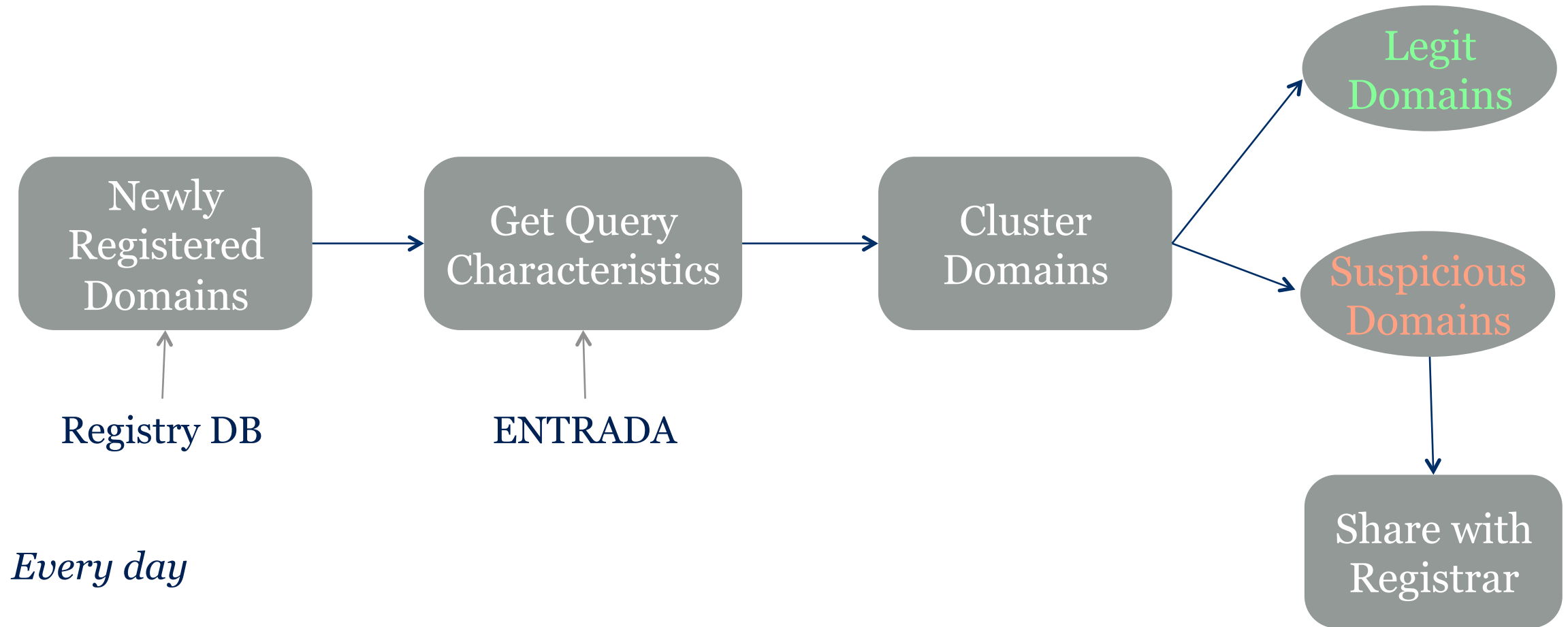
Random Sample Jan--Mar, 2015



Phishing



nDEWS Architecture



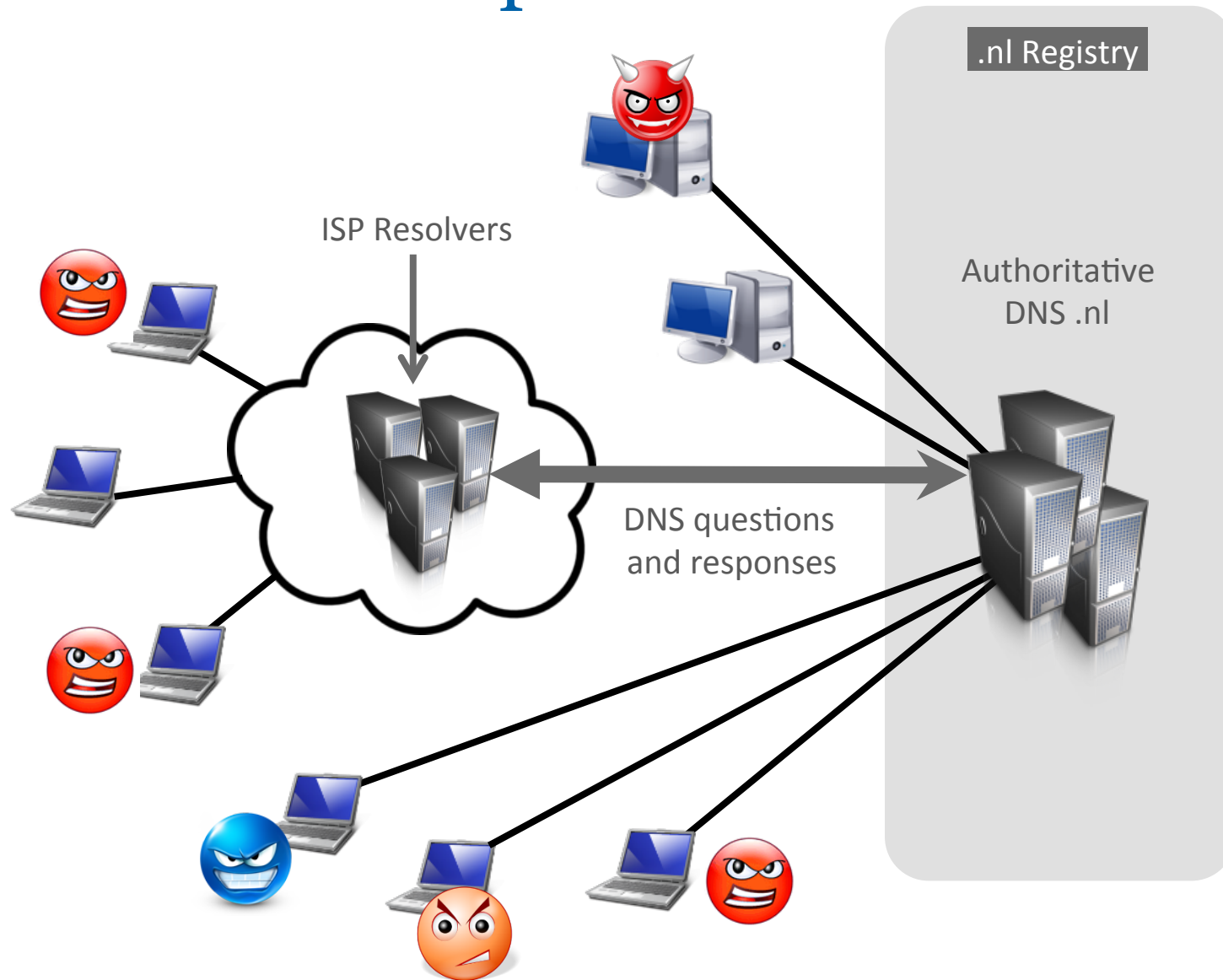
Resolver Reputation (RESREP)

Goal: Detect malicious activity by assigning reputation scores to resolvers

How: “fingerprinting” resolver behaviour



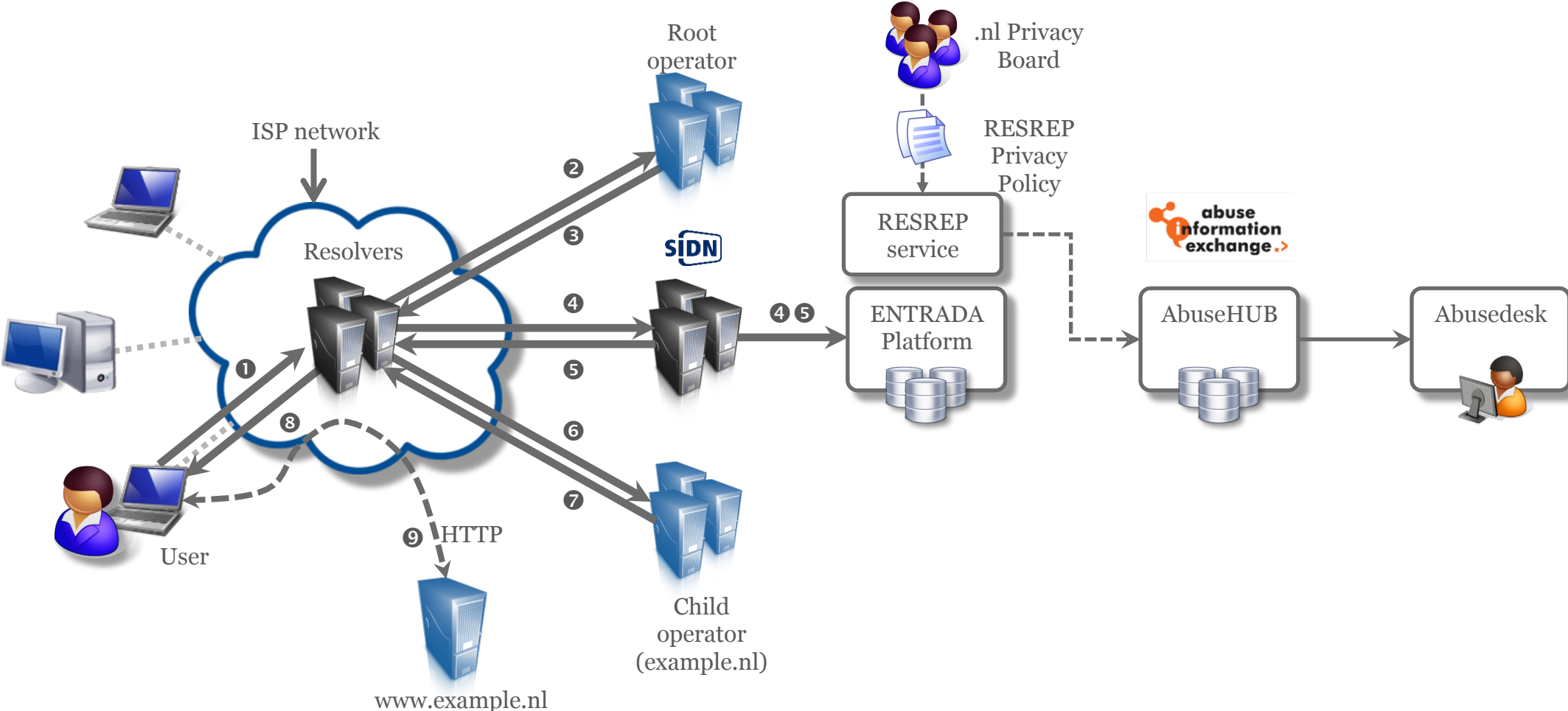
RESREP Concept



Malicious activity:

- Spam-runs
- Botnets
- DNS-amplification attacks

RESREP Architecture



Conclusions

Technical:

- Hadoop HDFS + Parquet + Impala is a winning combination!
- Running since almost 2 years
- > 115 billion queries stored

Contributions:

- Research by SIDN Labs and universities
- Identified malicious domain names and botnets
- Insight into DNS queries, shared at **stats.sidnlabs.nl**



It's open source!

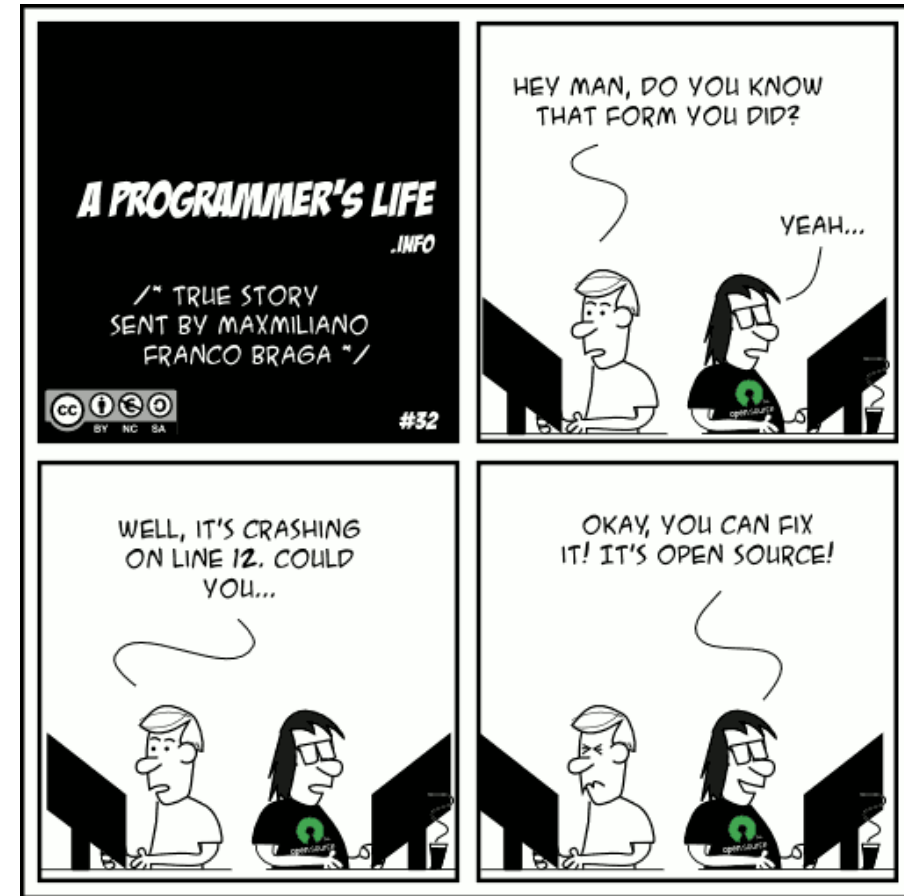
- Since January 2016

- Project site:

[*entrada.sidnlabs.nl*](http://entrada.sidnlabs.nl)

- GitHub:

[*github.com/SIDN/ENTRADA/*](https://github.com/SIDN/ENTRADA/)



Questions? Feedback?

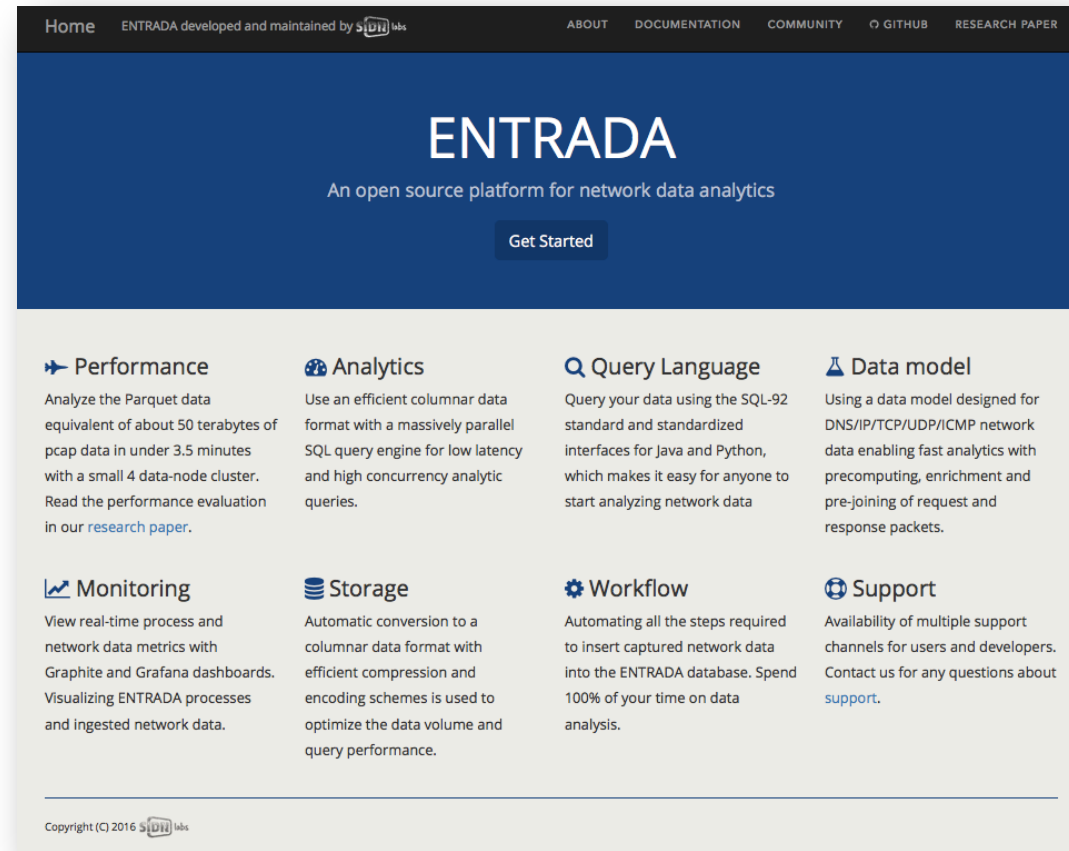
Moritz Müller

Research Engineer

moritz.muller@sidn.nl

 @dhr_moe

www.sidnlabs.nl



The screenshot shows the homepage of the ENTRADA project. The header includes navigation links: Home, ENTRADA developed and maintained by sidnlabs, ABOUT, DOCUMENTATION, COMMUNITY, GITHUB, and RESEARCH PAPER. The main heading is "ENTRADA" with the tagline "An open source platform for network data analytics" and a "Get Started" button. The content is organized into eight feature cards:

- 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 sidnlabs

entrada.sidnlabs.nl

