Your world. Our domain.

DNS Big Data

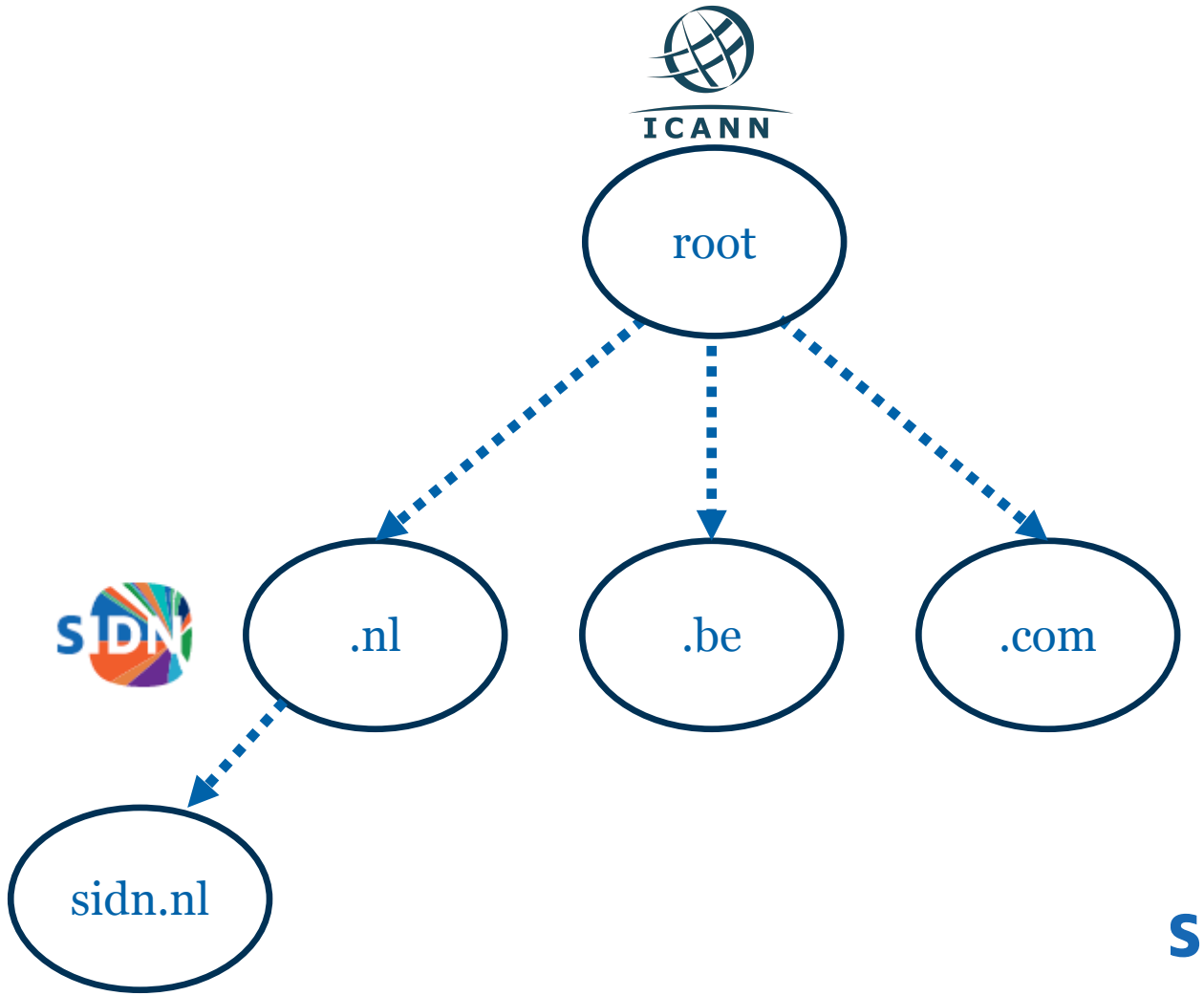
Maarten Wullink | SIDN TechTalk
Arnhem, 18 september 2019

SIDN LABS

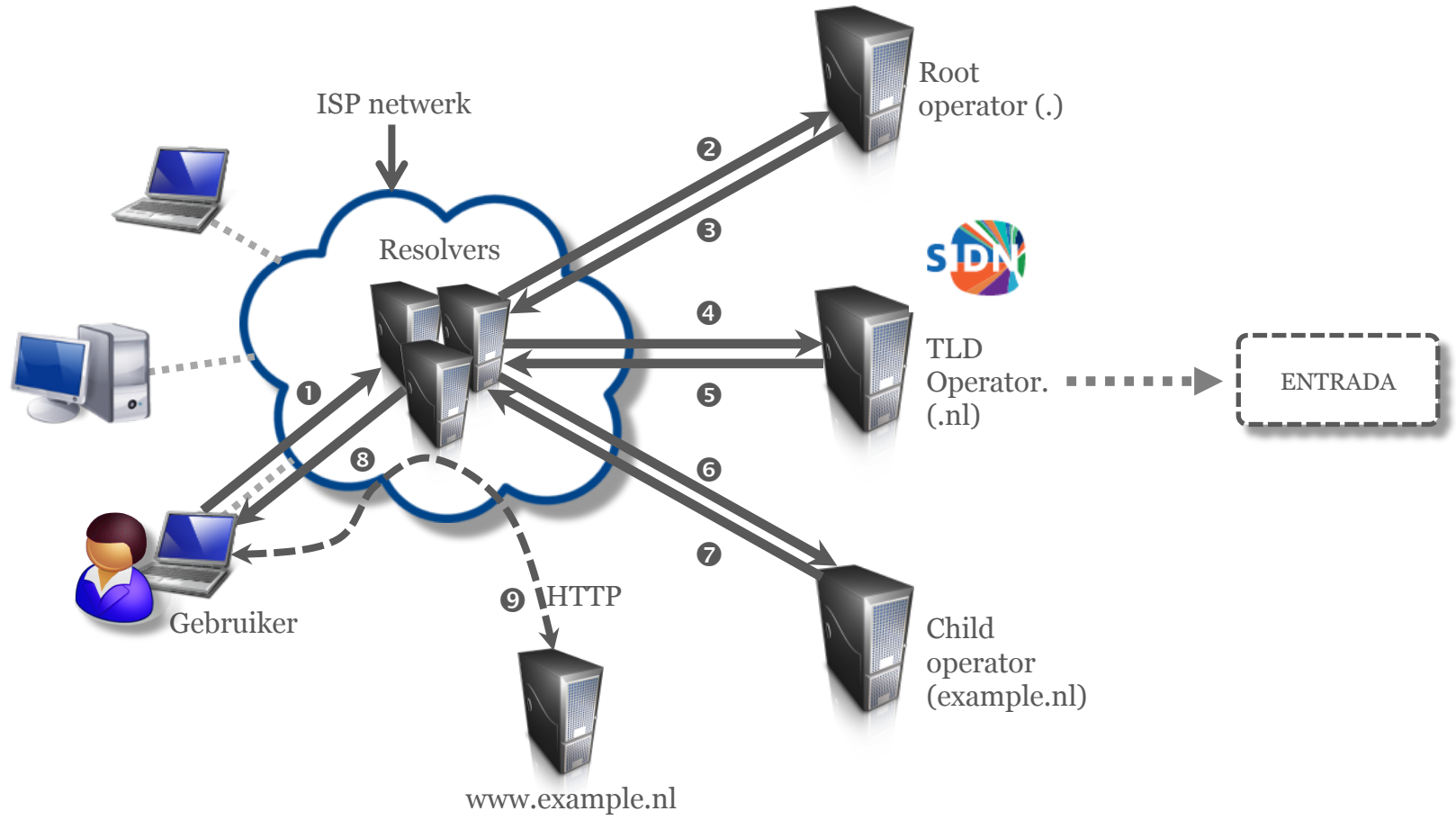
Wat is het Domain Name System (DNS)

Protocol voor koppelen en opvragen van informatie gerelateerd aan domeinnamen

- Hiërarchisch
- Gedecentraliseerd



DNS Resolving



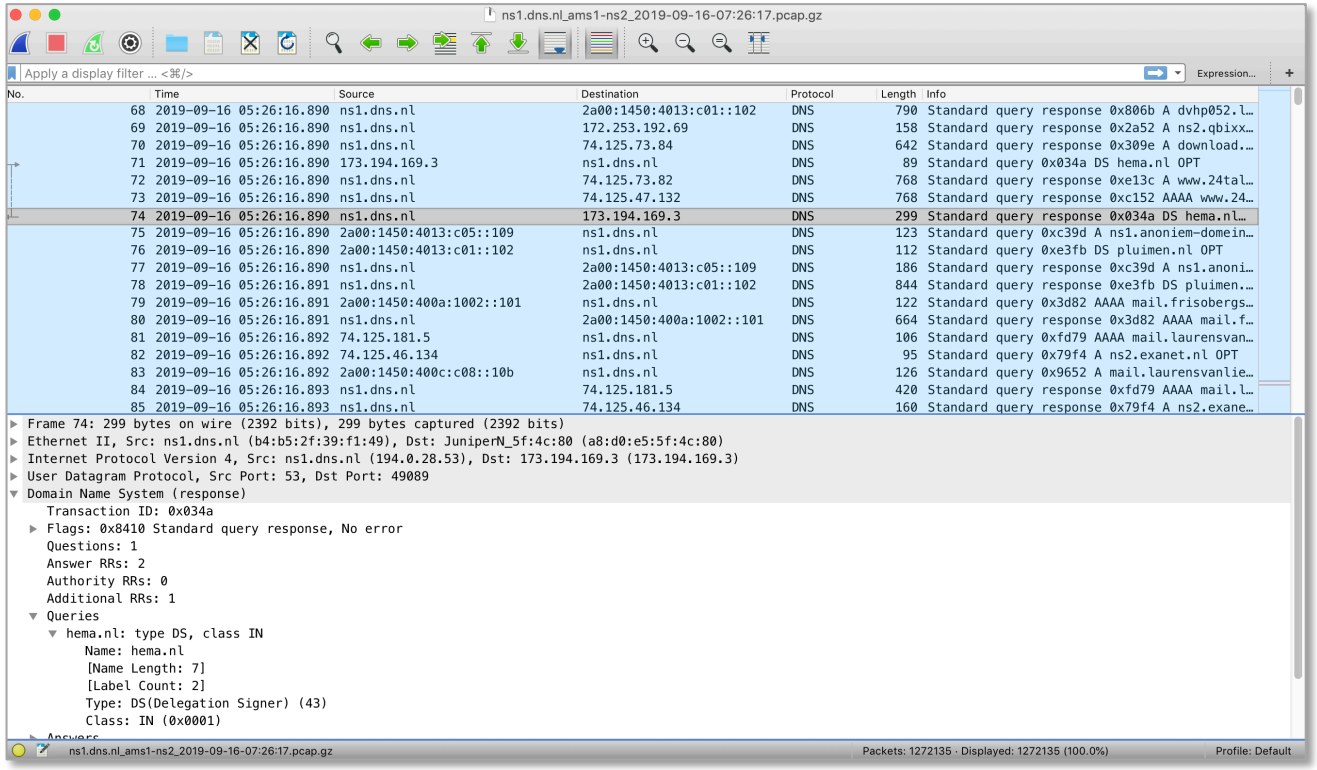
DNS data analyse

TCPDUMP & LIBPCAP

WIRESHARK

1. Capture DNS data met Tcpdump
2. 1 pcap bestand per 5-10 minuten
3. Analyse met Wireshark/Tshark

Klaar?



Wireshark voorbeeld

SIDN LABS

ENTRADA

ENhanced **T**op-Level Domain **R**esilience through **A**dvanced **D**ata **A**nalysis

- **Doel**: verhogen van de veiligheid en stabiliteit van de .nl-zone

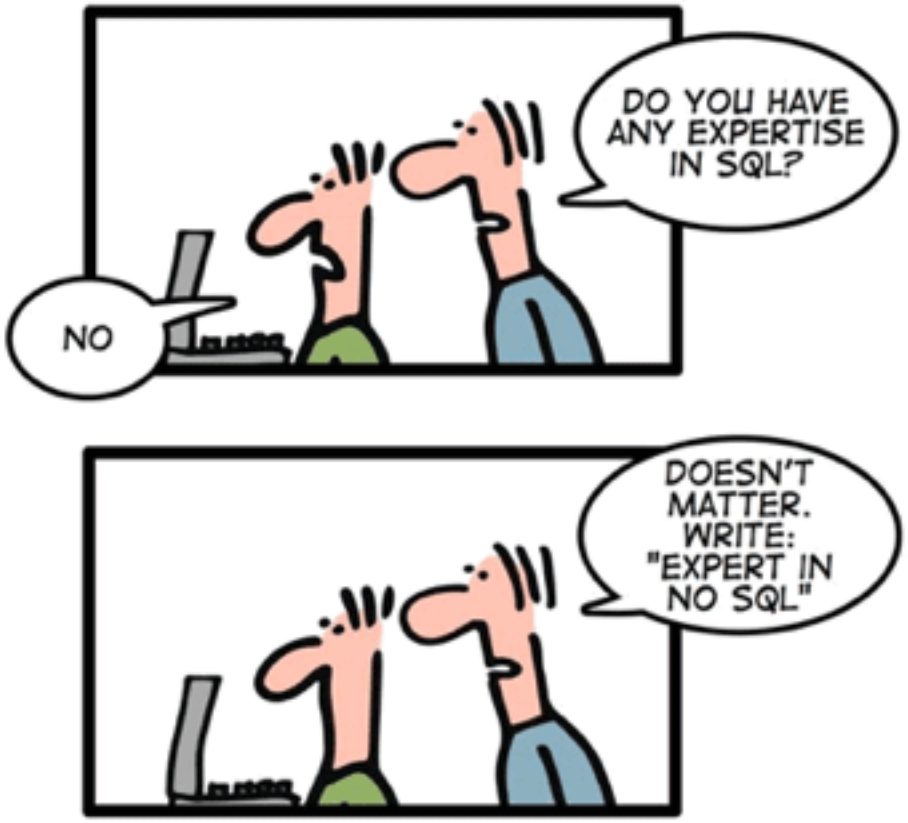
- **Requirements**:
  - Goede performance
  - Hoge beschikbaarheid
  - Semi real-time data warehouse
  - **SQL support**

SIDN LABS

Technologie selectie

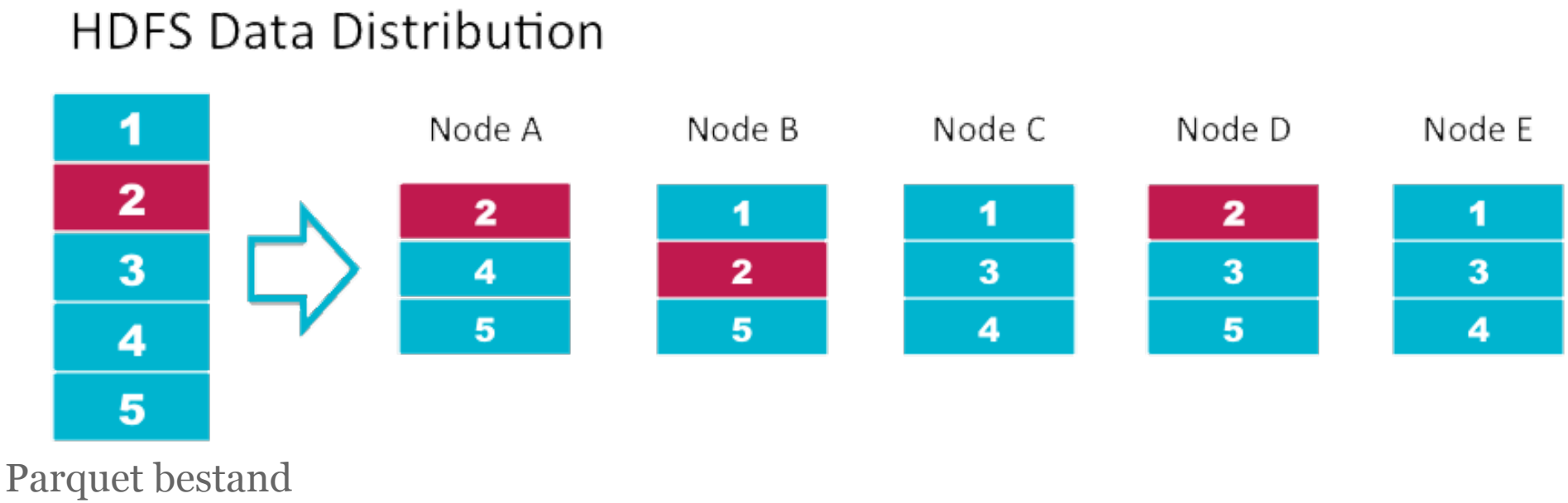
SQL and NoSQL oplossingen geëvalueerd (2015)

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



HDFS

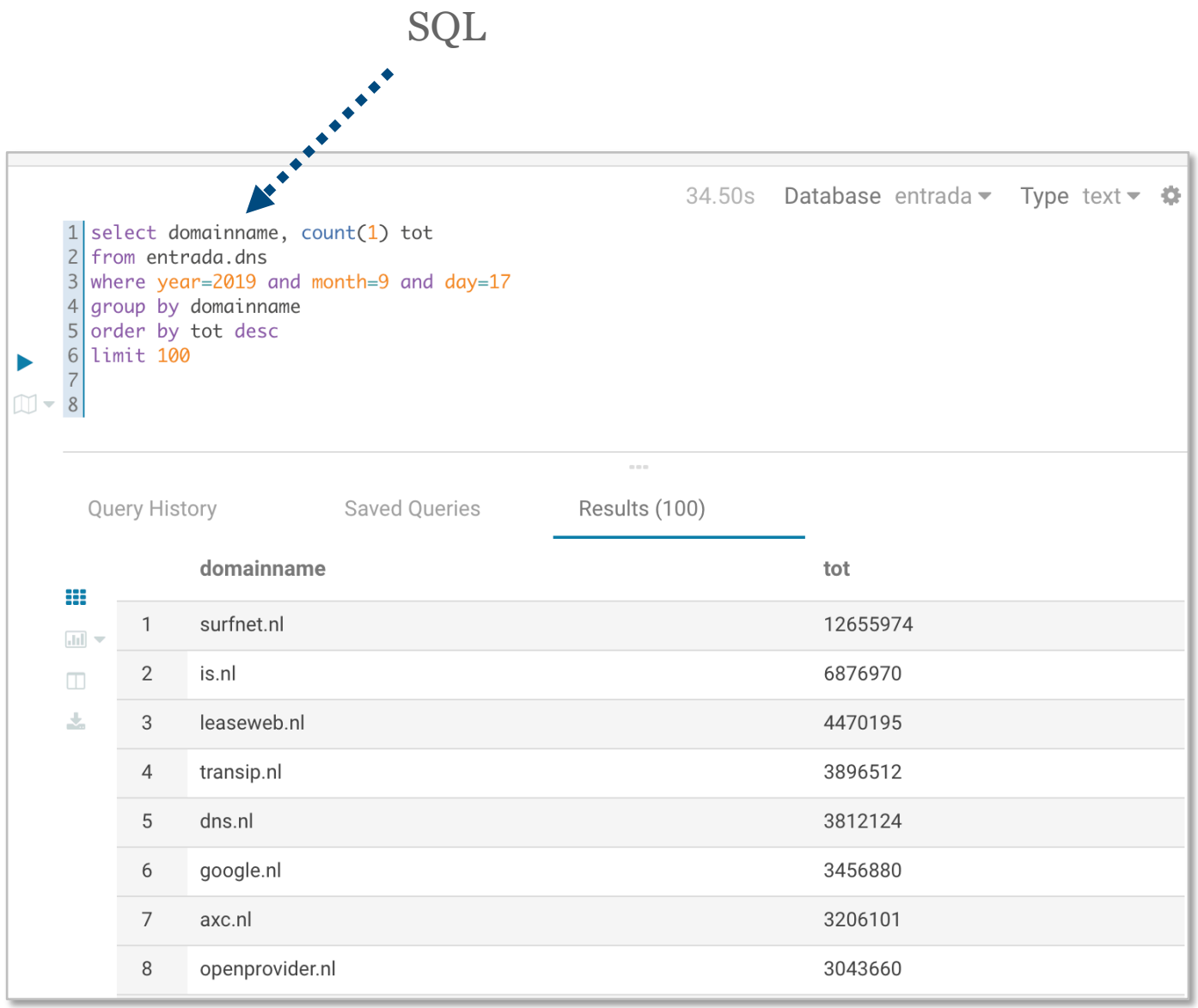
- Hadoop gedistribueerd bestandssysteem
- Hoge beschikbaarheid doormiddel van data replicatie
- Schaalbaar tot >100 PB en >1K servers



Parquet bestand

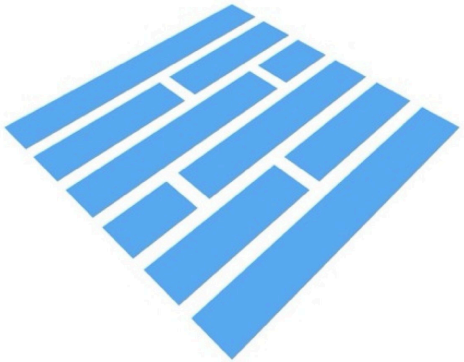
Apache Impala

- MPP Query engine
- SQL support
- JDBC, WEB, CMD



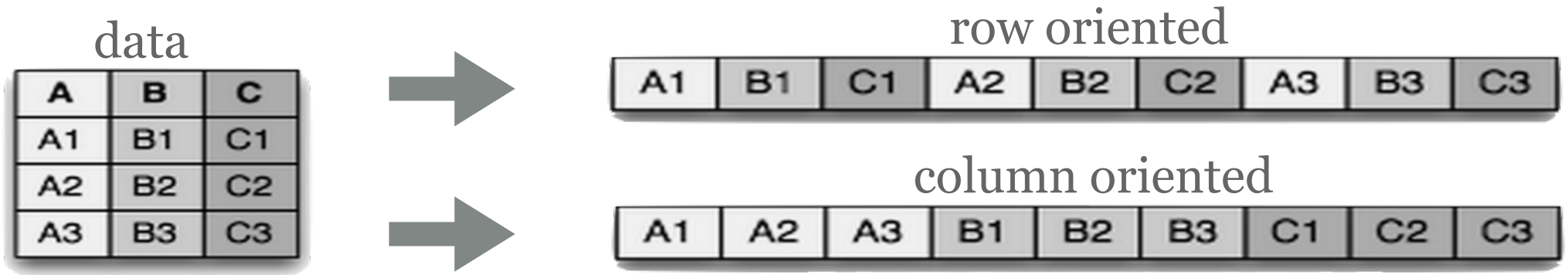
	domainname	tot
1	surfnet.nl	12655974
2	is.nl	6876970
3	leaseweb.nl	4470195
4	transip.nl	3896512
5	dns.nl	3812124
6	google.nl	3456880
7	axc.nl	3206101
8	openprovider.nl	3043660

Apache Parquet

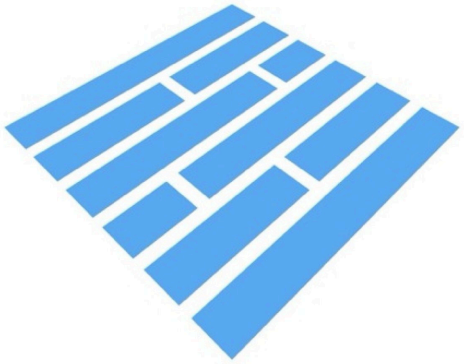


- Columnar storage format
  - Ontwikkeld door Twitter & Cloudera

- Voordelen Parquet
  - Bestanden zijn kleiner en efficiënter te lezen (minder disk IO)



Apache Parquet (2)



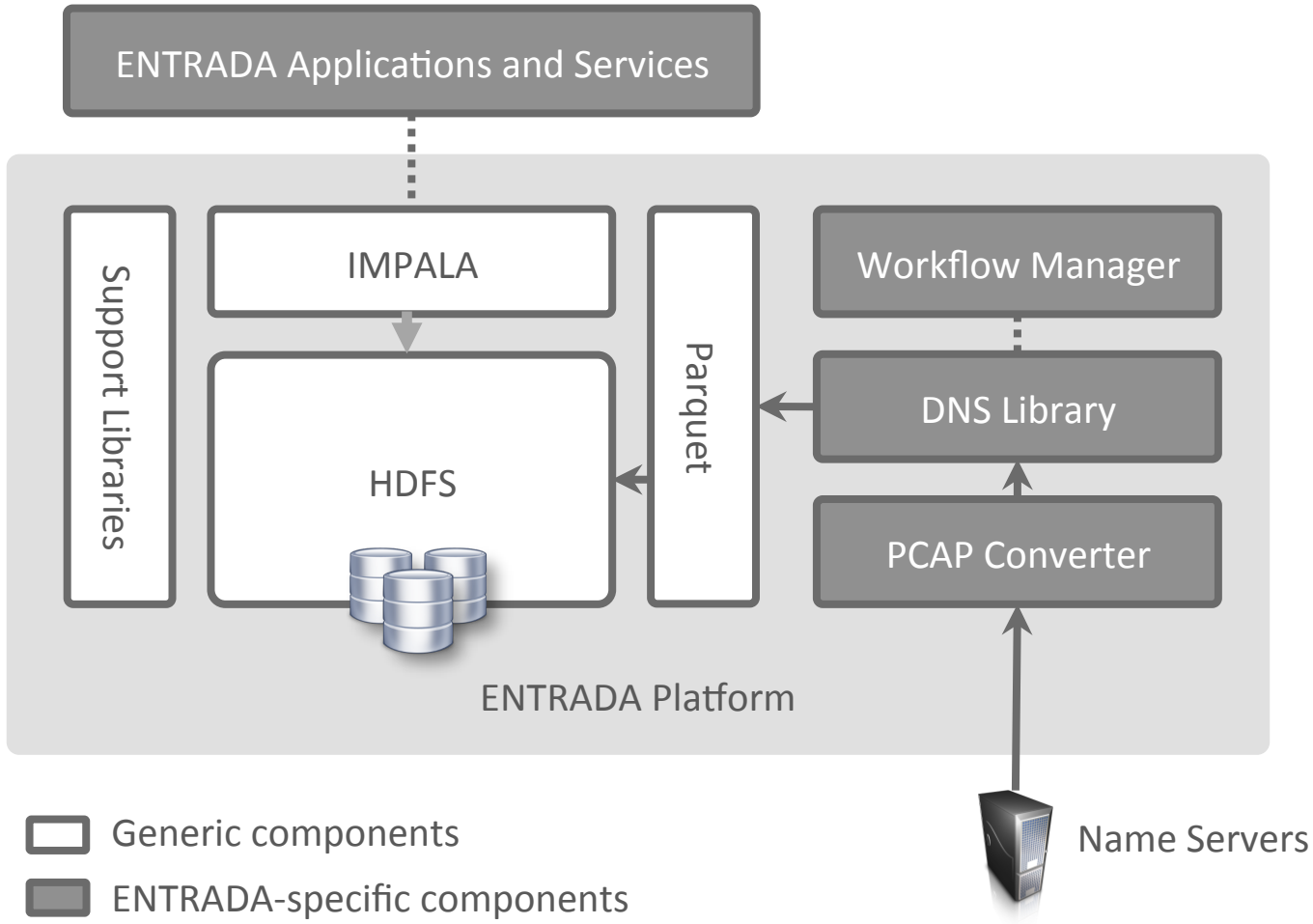
- Efficiente encoding/compressie
  - Verschillende encoding schemes (RLE, DICT ...)
  - Snappy compressie support

- Partitionering van data (jaar, maand, dag, server)
  - Partition pruning, negeer data waar we niet in geïnteresseerd zijn

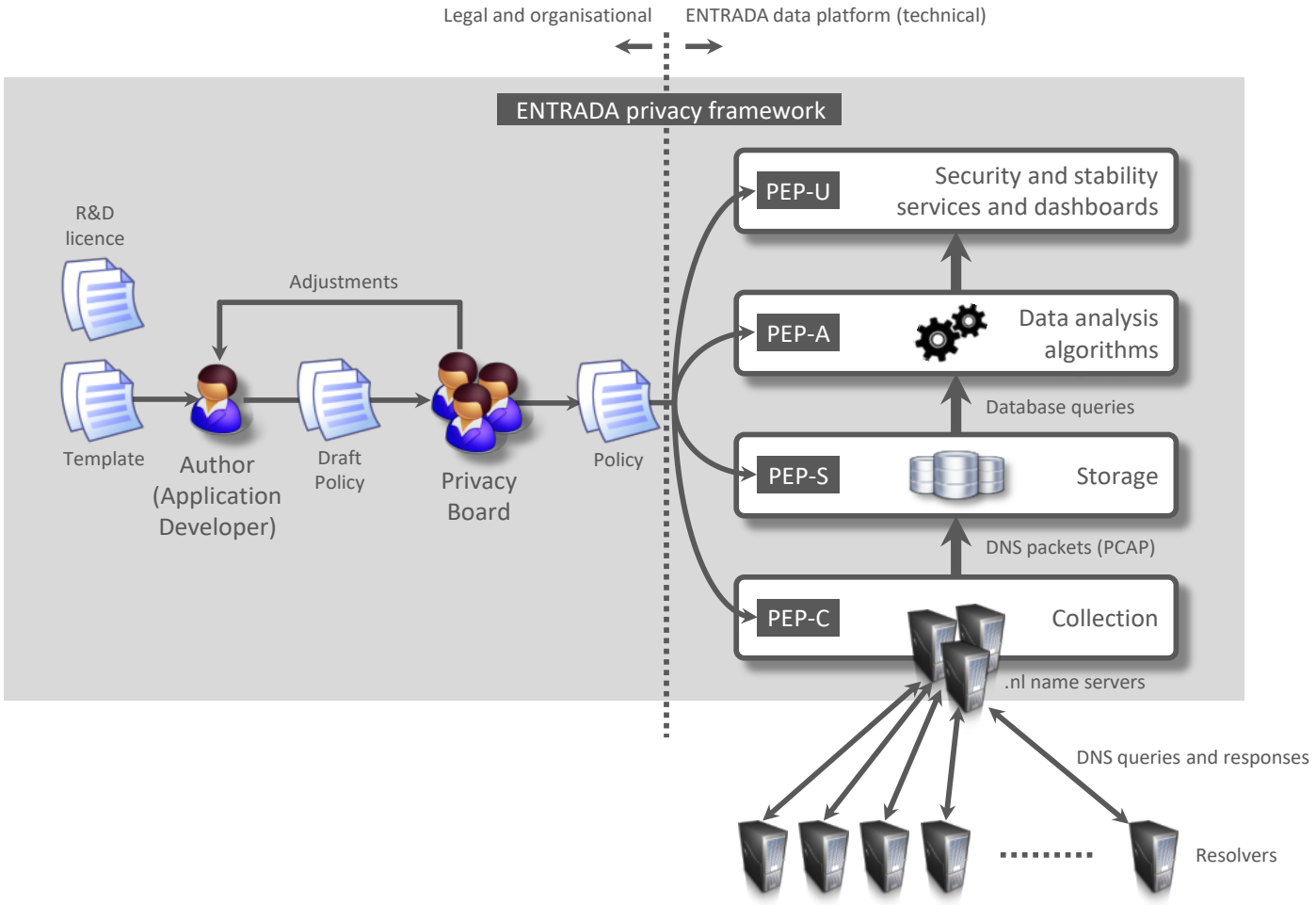
- Query engine Parquet support (Impala, Spark, AWS Athena)

SIDN LABS

ENTRADA


ENTRADA Applications and Services
Support Libraries
IMPALA
HDFS
Parquet
Workflow Manager
DNS Library
PCAP Converter
ENTRADA Platform
Generic components
ENTRADA-specific components
Name Servers

Privacy raamwerk



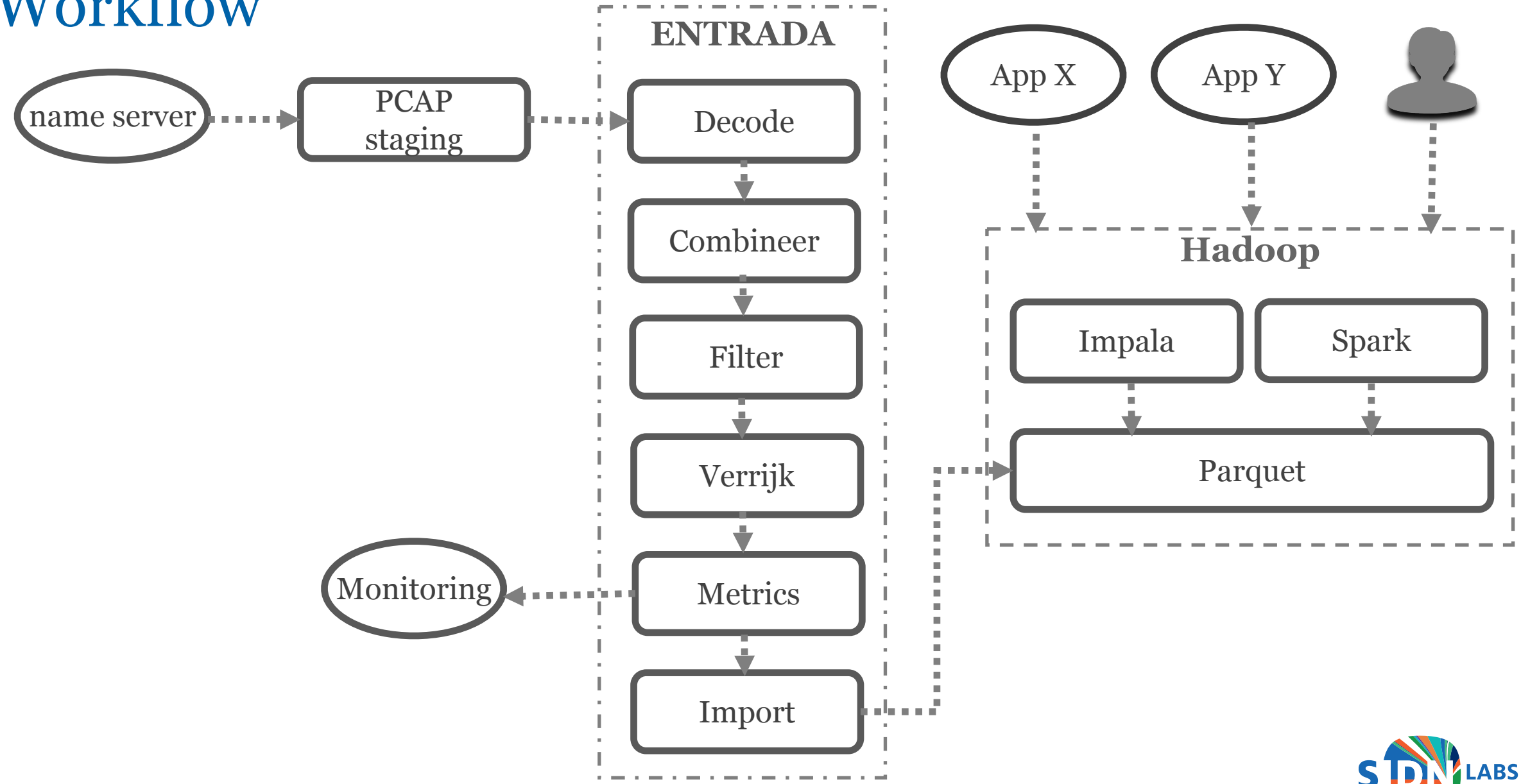
Belangrijke concepten

- Applicatie-specifiek privacy policy
- Privacy Board
- Enforcement Points

Policy elementen

- Doel
- Gebruikte data
- Filters
- Retentie periode
- Type applicatie (R&D vs. production)

Workflow



DNS data beschikbaar voor analyse in ~10-15 minuten

ENTRADA@SIDN Labs

- 4 jaar operationeel
- 2 anycast .nl name servers (28 sites)
- 1,2 biljoen (1,2 x $10^{12)}$ records (DNS query+response)
- 65 TB data (netto)

Hardware (mini-cluster)

1x Management node

Virtual machine (VMWARE)

9x Data node

HP ProLiant DL380
Xeon 2.10 GHz 8 core CPU
128GB RAM
25-30 TB storage (SAS + SATA)
2x1Gb netwerk

Resources

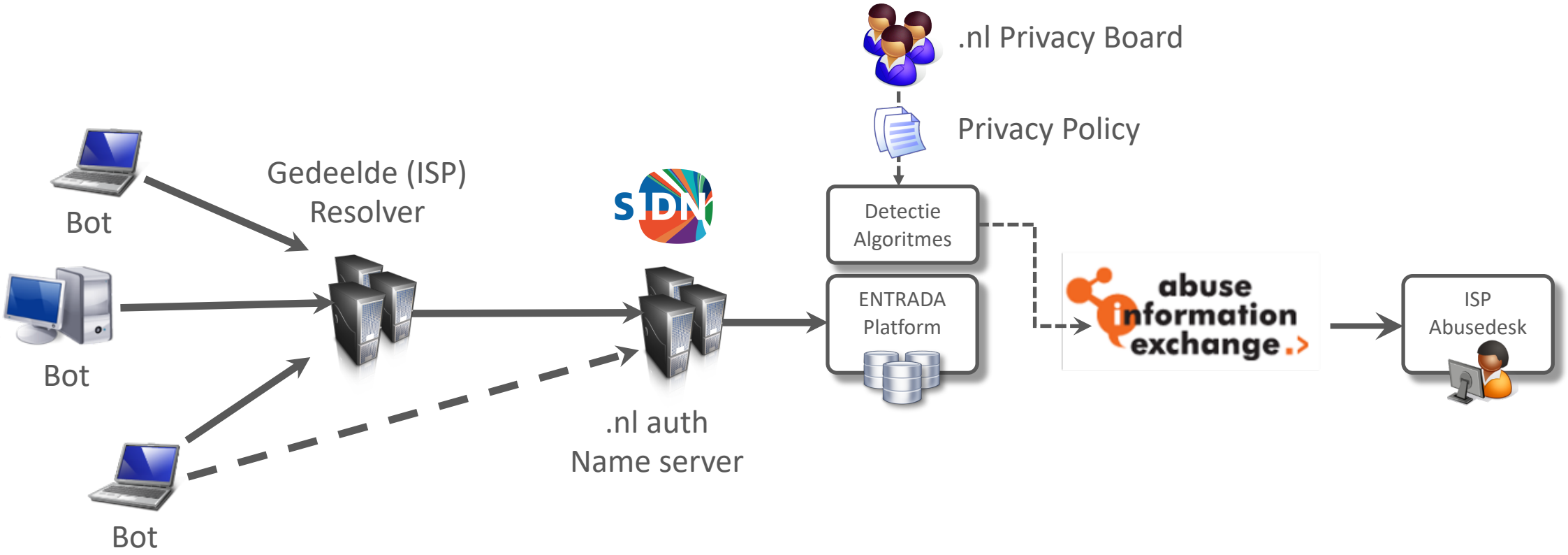
142 CPU cores
~250 TB storage
~1 TB RAM

Voorbeeld Use Cases

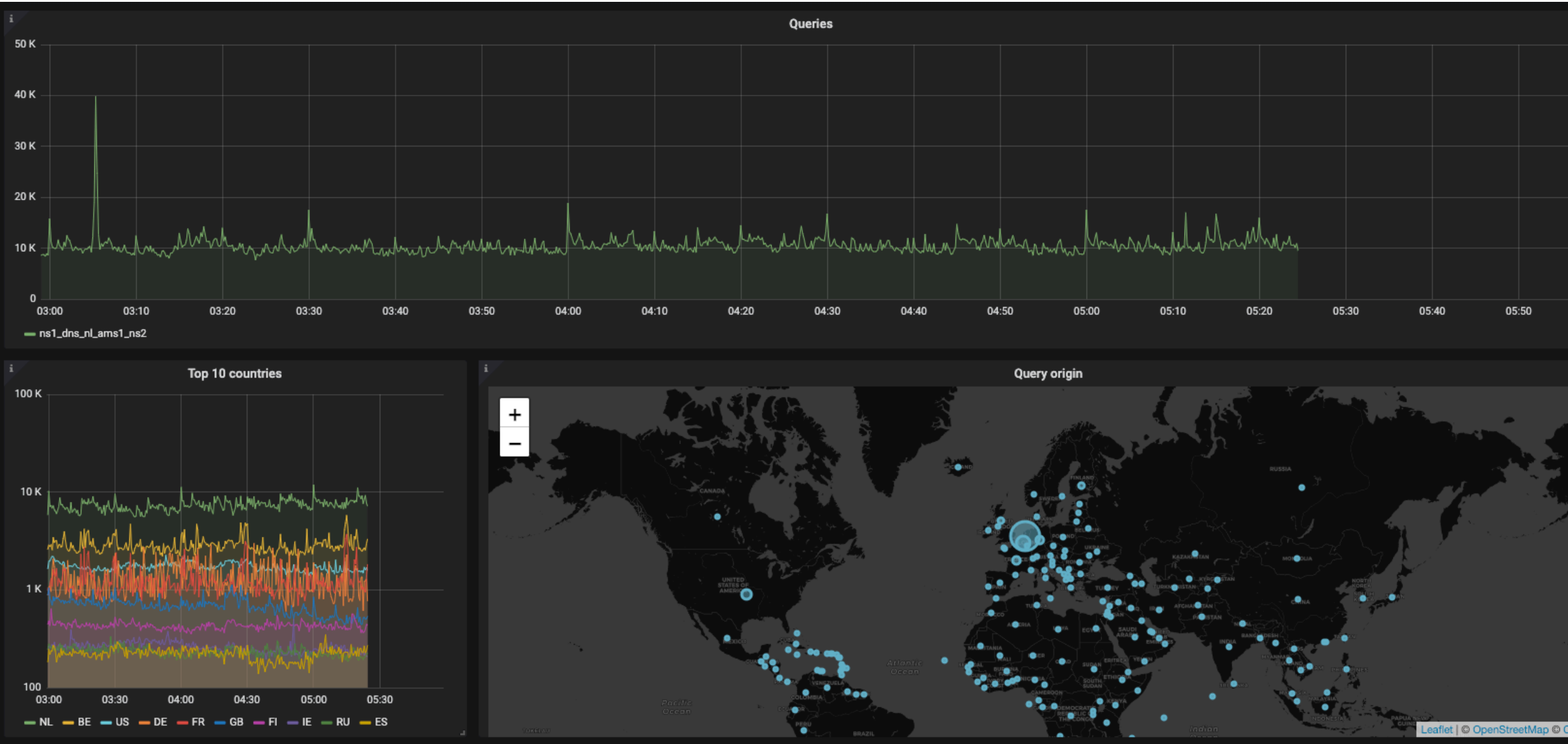
- Statistieken (**stats.sidnlabs.nl**)
- Onderzoek
- Inzicht voor DNS beheerders
  - Monitoring
- Detectie van misbruik
  - Botnets
  - Fake webshops

.nl-domeinnamen per postcode

Botnet Client Detectie

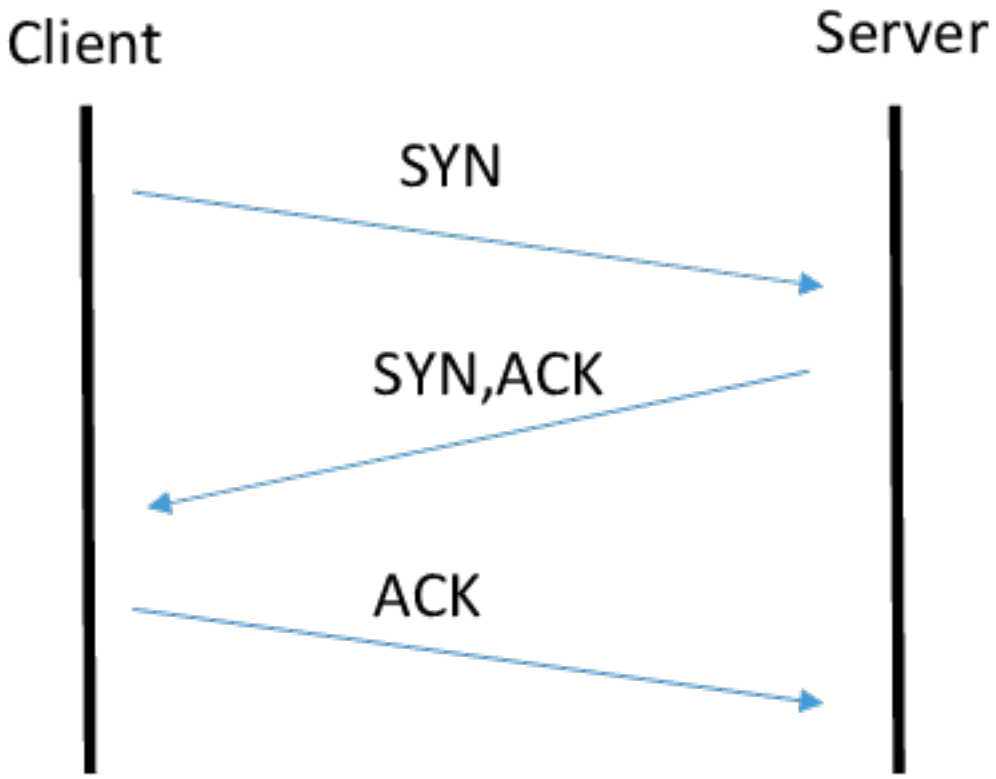


Monitor verkeer

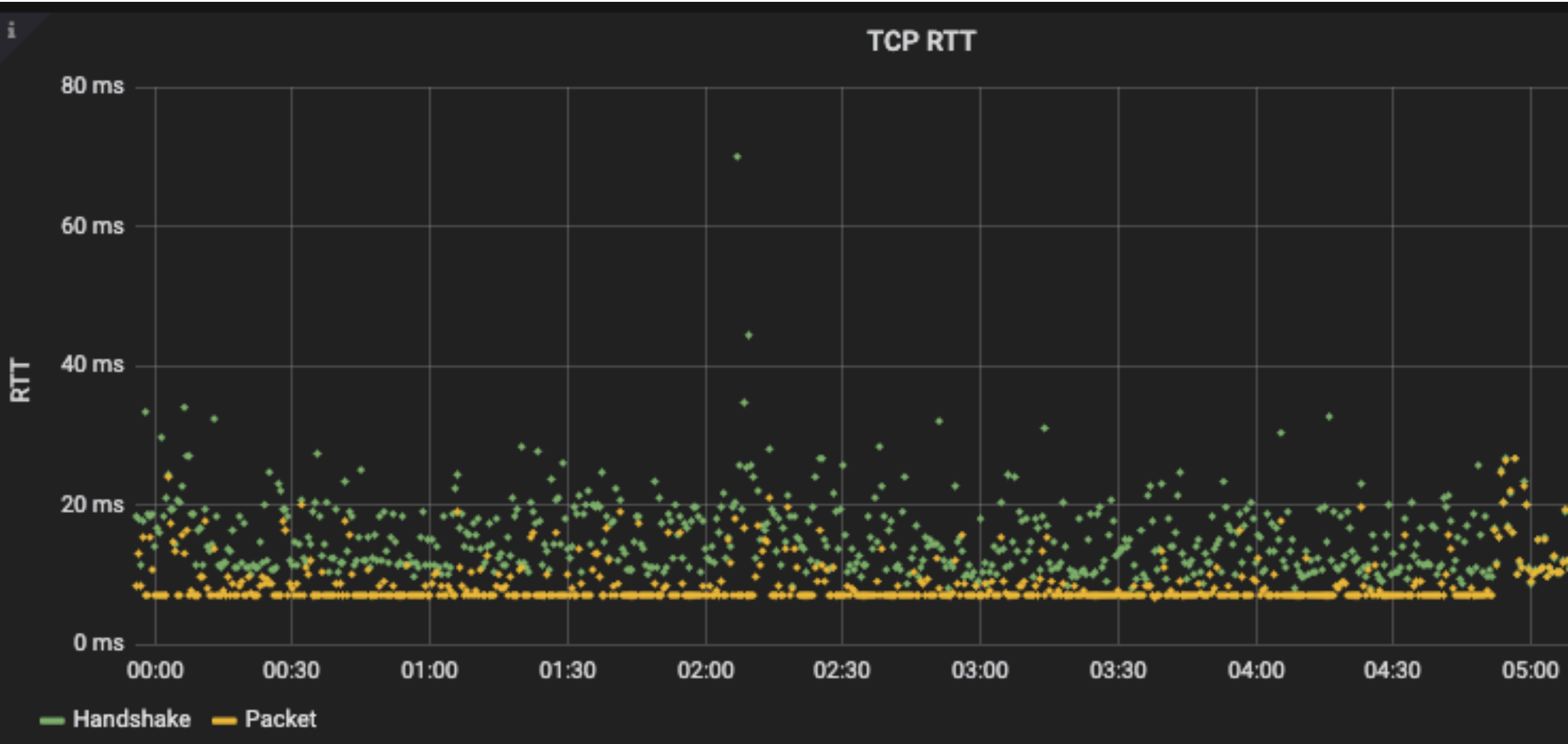


Quality of service monitor

TCP Round Trip Time

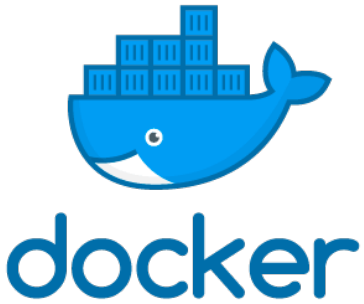


Quality of service monitor



Deployment

- Deploy met Docker
- Hadoop, inhouse of in de cloud
- Amazon S3 + Athena
- Config met docker-compose yaml



https://entrada.sidnlabs.nl

SIDN LABS


SIDN.nl


@SIDN


SIDN

Q&A

www.sidnlabs.nl | stats.sidnlabs.nl

Bedankt!

SIDN LABS