Netflow informatie snel doorzoekbaar maken

Remco Poortinga – van Wijnen

28 juni 2024

Agenda

01 SURF network

02 nfdump/nfsen toolset & issue

03 possible solutions

04 processing setup

05 impressive numbers

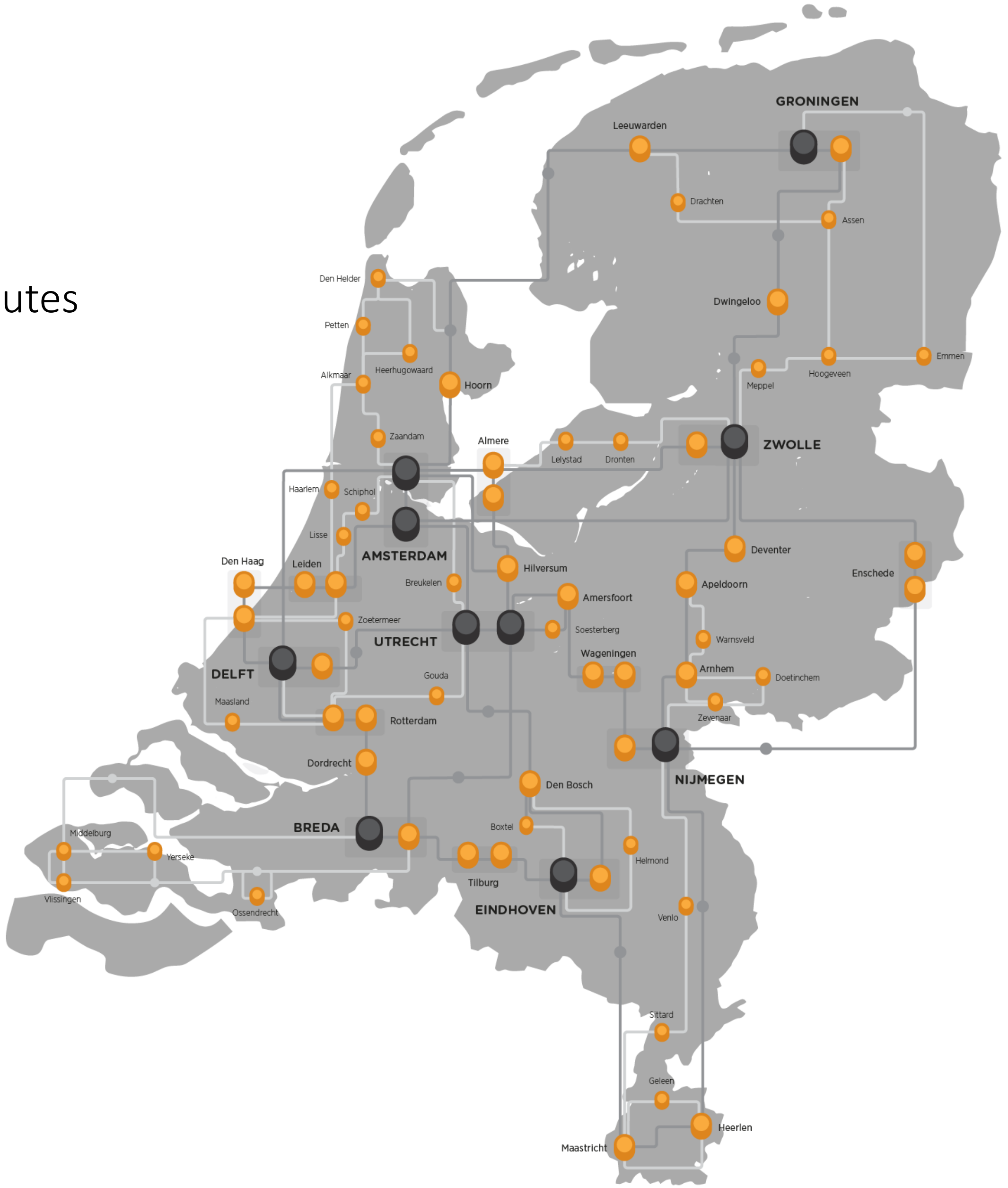
06 Other uses & examples

07 conclusions

The SURF network

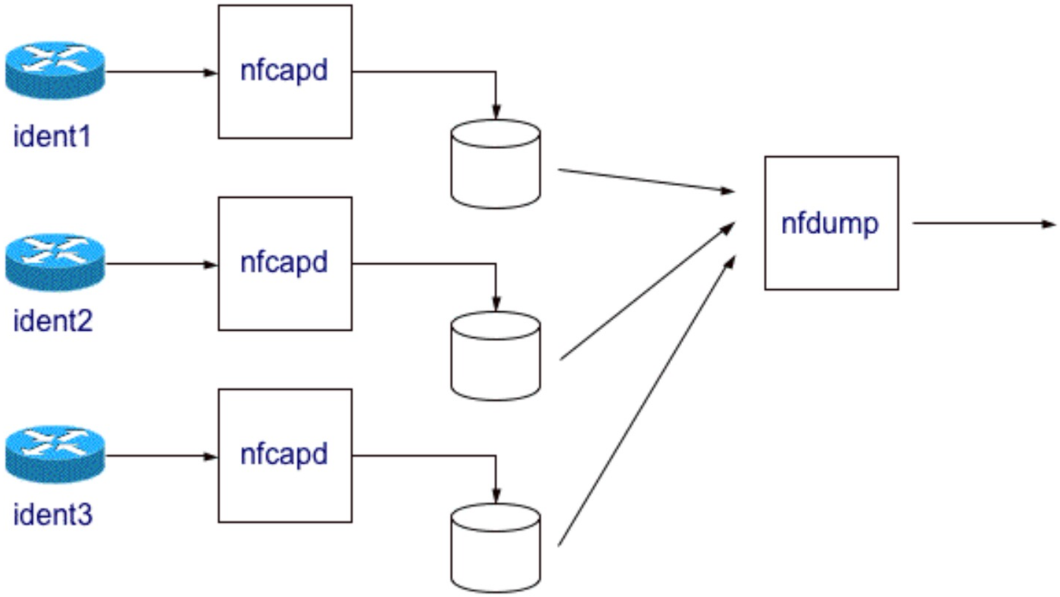
- Points-of-Presence (PoP) in data centers of Universities, Applied Universities, Vocational Colleges, Research Institutes throughout the Netherlands
- ~ 12,000 km fibre cables
- ~ 460 routers
- ~ 1,5 million end users

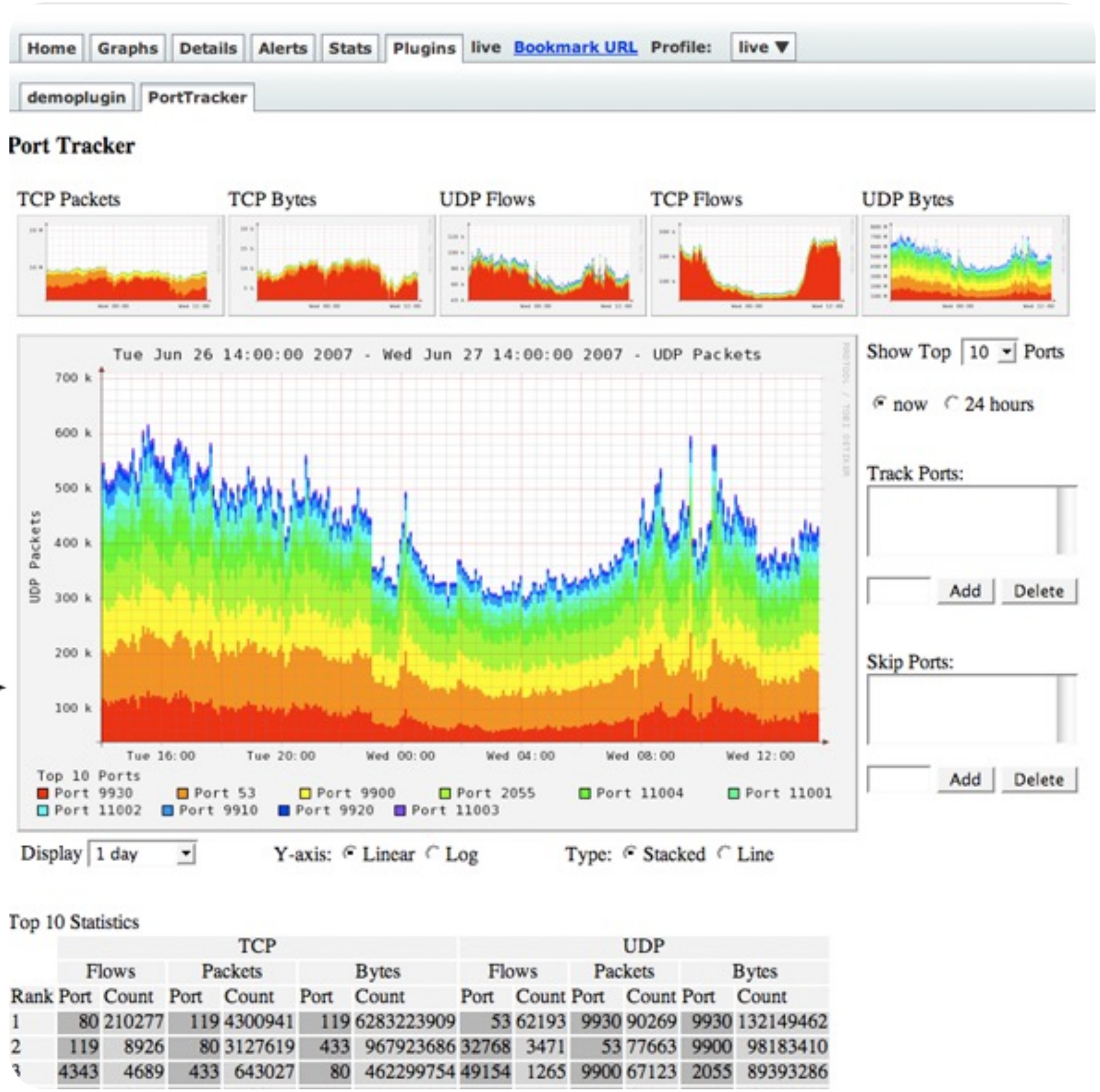
- Netflow 1/100 sampling
  - 2..4 billion flows/day
  - > 80.000 flows/s peak
    - Mondays 12:00 – 12:05



nfdump/nfsen toolset

- SURFcert tool of choice for analysing network traffic
- Works really well for analysis
  - If you know where & (especially) when to look
  - Determining that can be really slow and cumbersome





Port Tracker

Show Top 10 Ports

● now ○ 24 hours

Track Ports:

Skip Ports:

Display 1 day

Y-axis: ● Linear ○ Log

Type: ● Stacked ○ Line

Top 10 Statistics

	TCP						UDP					
	Flows		Packets		Bytes		Flows		Packets		Bytes	
Rank	Port	Count	Port	Count	Port	Count	Port	Count	Port	Count	Port	Count
1	80	210277	119	4300941	119	6283223909	53	62193	9930	90269	9930	132149462
2	119	8926	80	3127619	433	967923686	32768	3471	53	77663	9900	98183410
3	4343	4689	433	643027	80	462299754	49154	1265	9900	67123	2055	89393286

Possible solutions

MySQL (2009)

Transactional database not developed for large data analytics. Much slower.

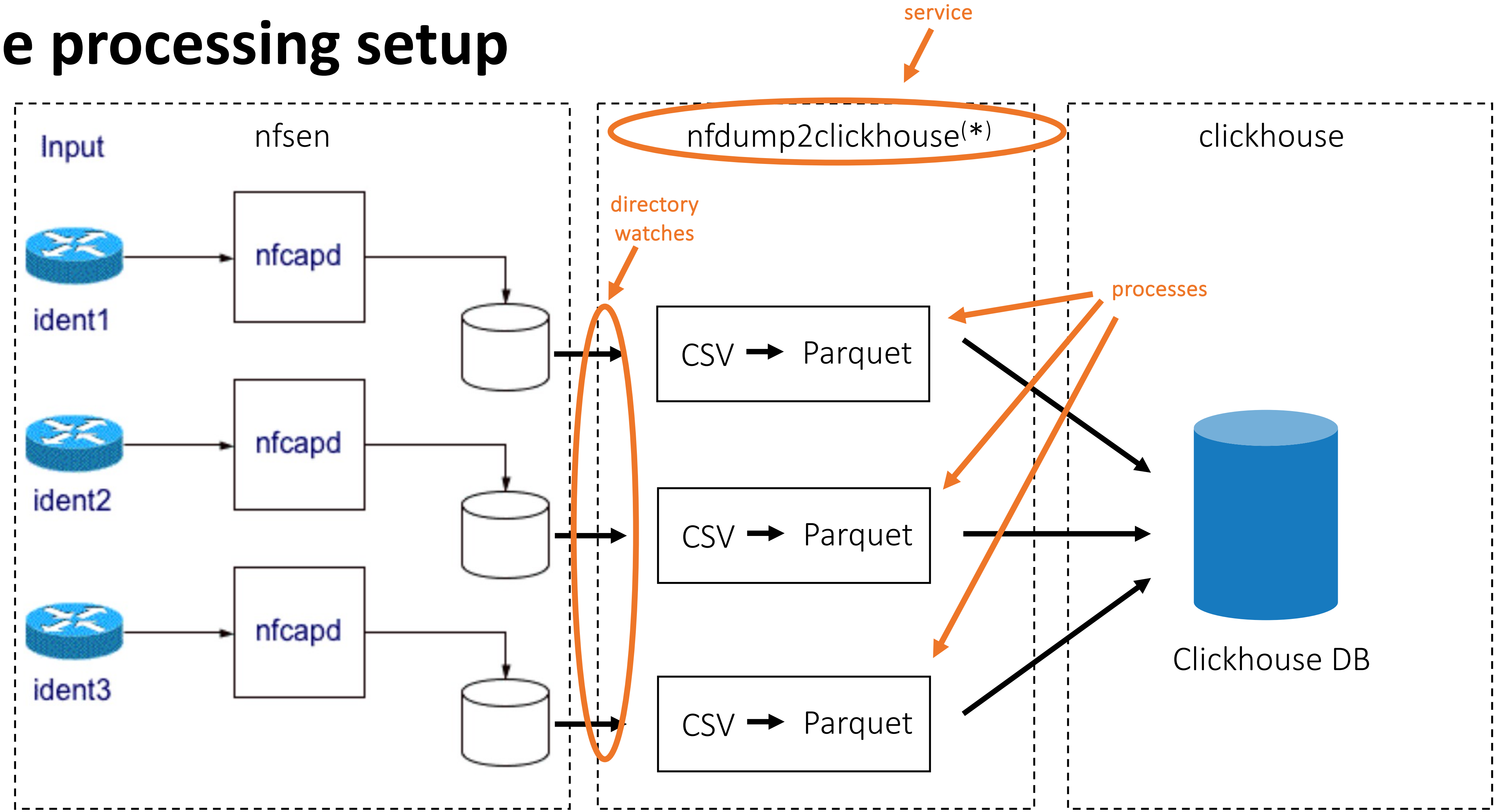
Hadoop+spark (2018)

Faster for timeframes over 3.5 hours, slower for shorter ones. Complex/specialized.

ClickHouse

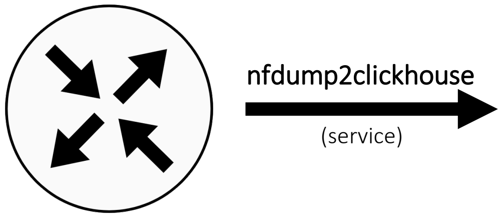
Column oriented analytical database. Goldilocks ‘biggish data’ solution.

The processing setup



(*) https://github.com/poorting/nfdump2clickhouse

Clickhouse flows table



nfdump2clickhouse

(service)

netflow

flows		
ts	Start time of the flow	DateTime
te	End time of the flow	DateTime
sa	Source IP address	String
da	Destination IP address	String
sp	Source port	UInt16
dp	Destination port	UInt16
pr	Protocol (e.g. ‘TCP’ or ‘UDP’)	String
flg	Flags (if pr is ‘TCP’)	String
ipkt	Number of packets in this flow	UInt64
ibyt	Number of bytes in this flow	UInt64
smk	Source mask	UInt8
dmk	Destination mask	UInt8
ra	Router (IP) Address	String
in	Input interface number	UInt16
out	Output interface number	UInt16
sas	Source AS number	UInt16
das	Destination AS number	UInt16
exid	Exported id	UInt16
flowsrc	Additional label for this source	String

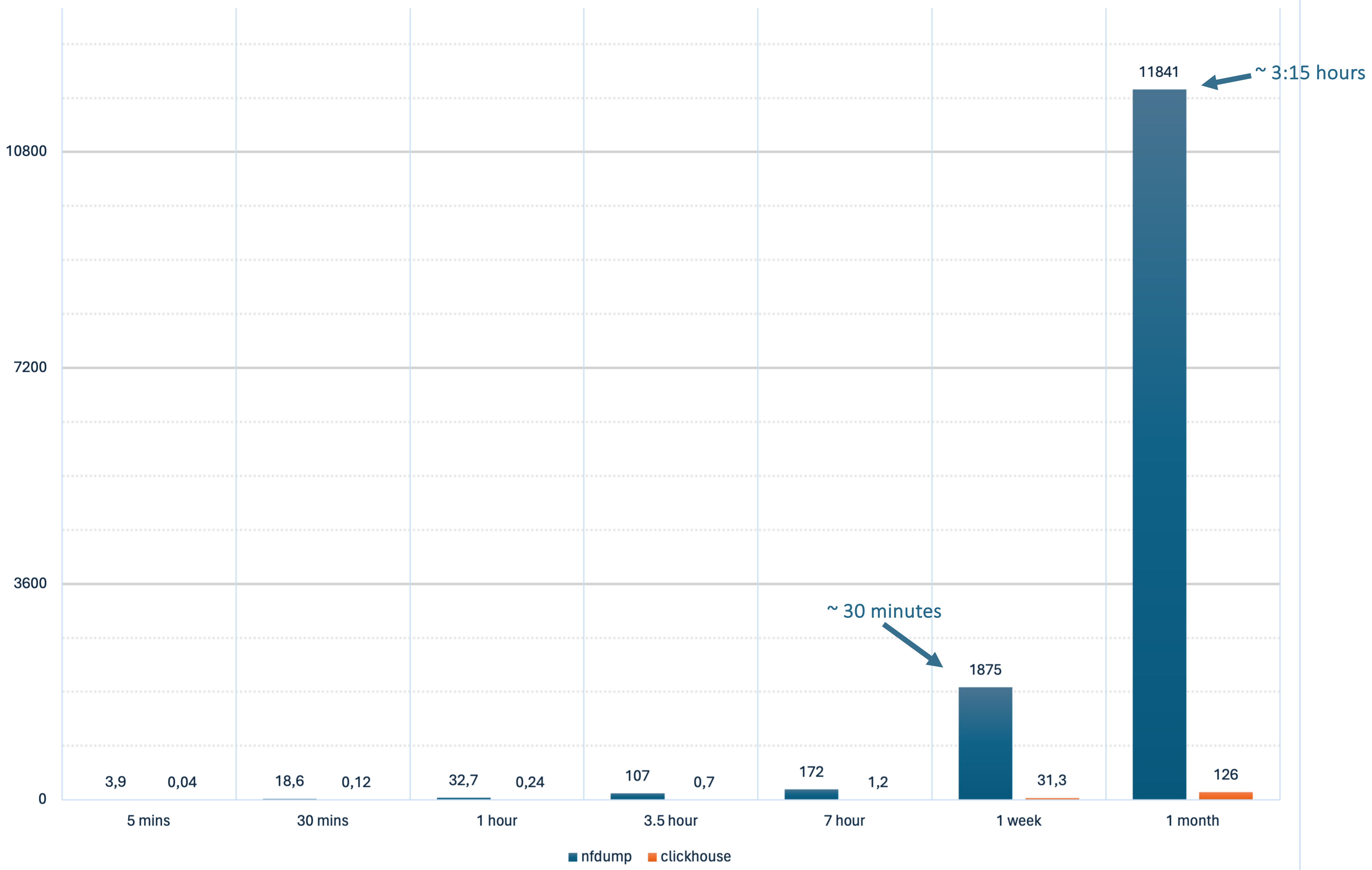
So how fast is this thing anyway?

Comparing clickhouse to nfdump

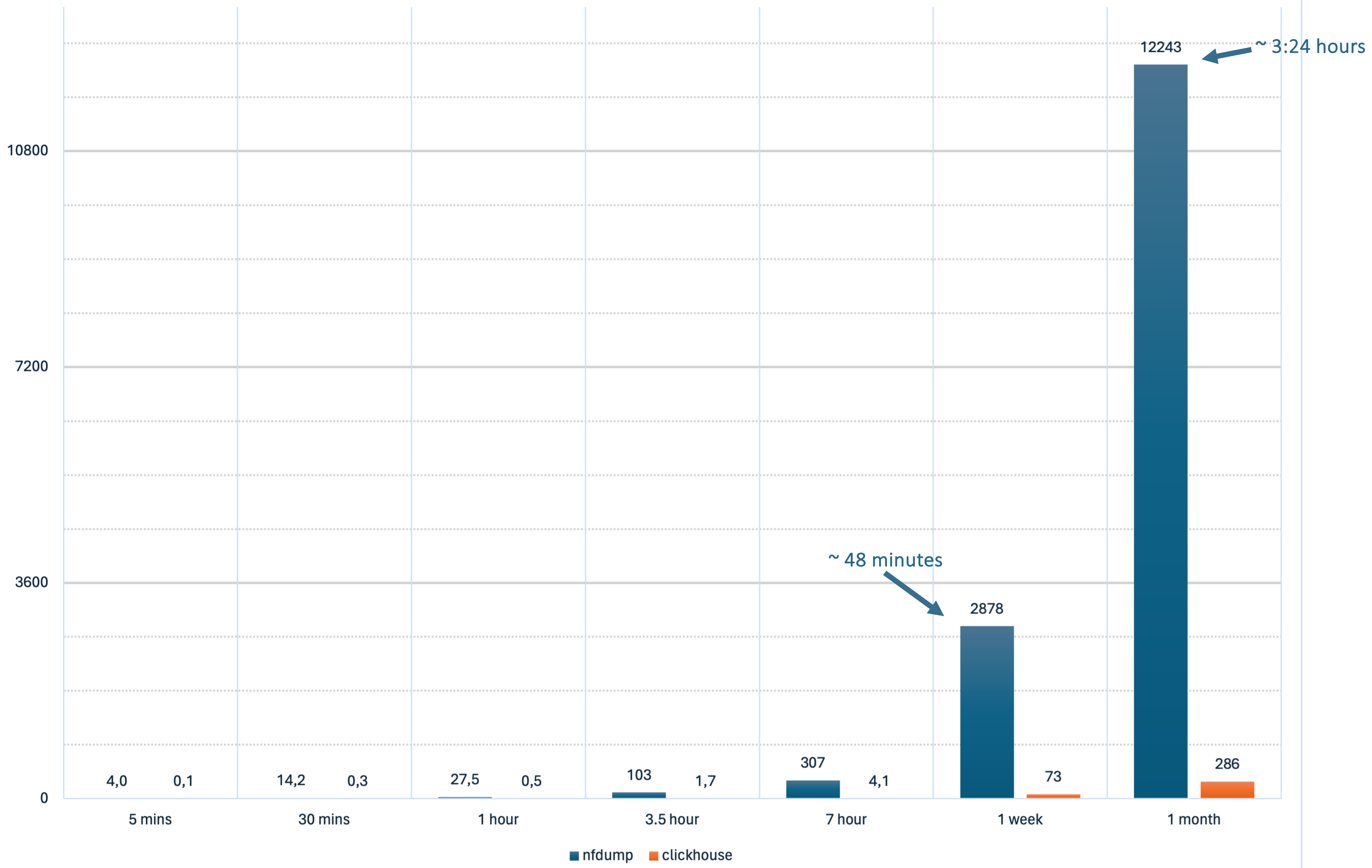
- Searches:
  - specific destination IP address
  - specific destination IP address + port
  - specific destination IP range + port excluding IP range from results
- Searches executed on the same machine
  - 24/48 core AMD EPYC 7443P Processor
  - 128 GB of memory
  - 2x7 TB of storage

timeframe	# of flows
5 minutes	23 million
30 minutes	136 million
1 hour	268 million
3.5 hours	997 million
7 hours	1.16 billion
1 week	23 billion
1 month	95 billion

Find destination IP address



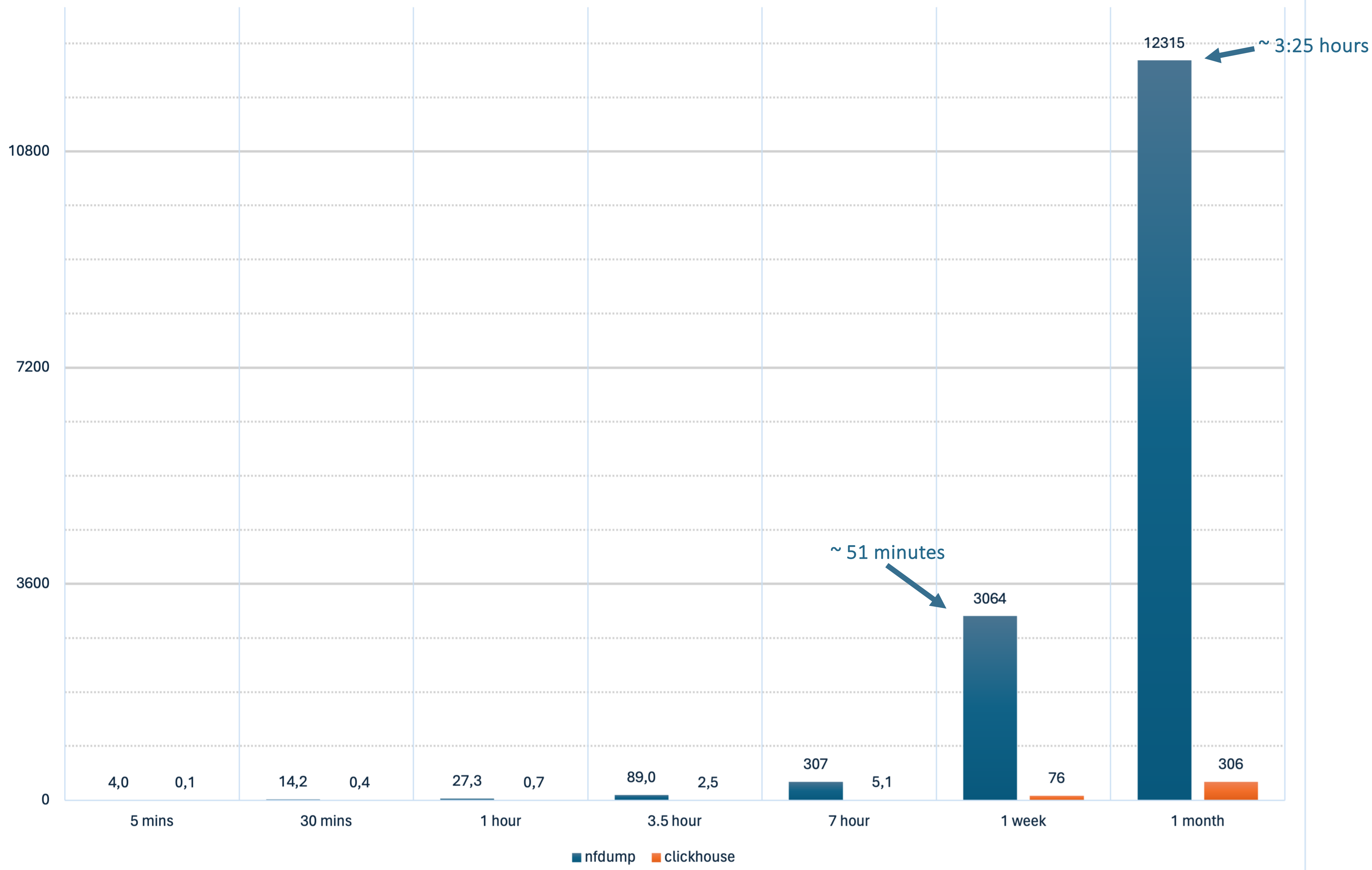
Find destination IP address + Port



~ 3:24 hours

~ 48 minutes

Find destination IP range + Port, except source IP range



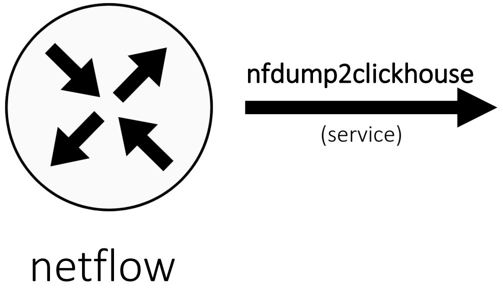
~ 3:25 hours

~ 51 minutes

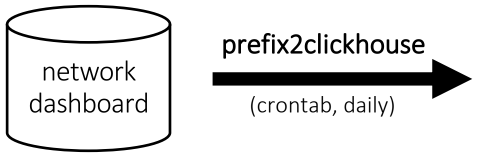
But wait…

- Finding IP addresses is all well and good,
  - but we deal with institutions
- It would be nice to link an IP address to an institution directly

- Just create a table with all prefixes of our network
  - Luckily there's an API for that! ☺



flows		
ts	Start time of the flow	DateTime
te	End time of the flow	DateTime
sa	Source IP address	String
da	Destination IP address	String
sp	Source port	UInt16
dp	Destination port	UInt16
pr	Protocol (e.g. 'TCP' or 'UDP')	String
flg	Flags (if pr is 'TCP')	String
ipkt	Number of packets in this flow	UInt64
ibyt	Number of bytes in this flow	UInt64
smk	Source mask	UInt8
dmk	Destination mask	UInt8
ra	Router (IP) Address	String
in	Input interface number	UInt16
out	Output interface number	UInt16
sas	Source AS number	UInt16
das	Destination AS number	UInt16
exid	Exported id	UInt16
flowsrc	Additional label for this source	String



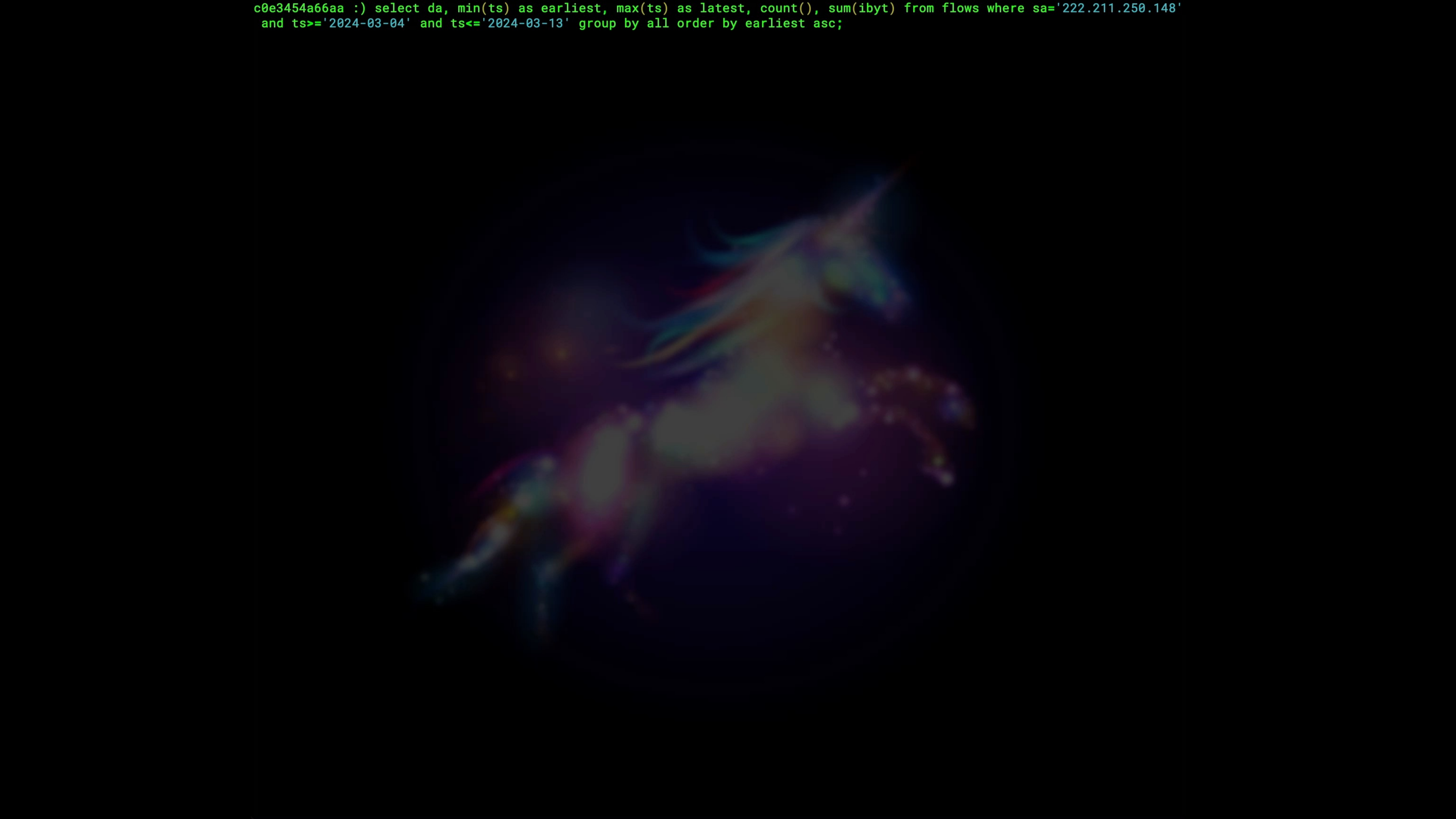
prefixes		
prefix	Network prefix	String
customer_id	Customer id	String
abbreviation	Abbreviation of the org	String
name	Name of the organisation	String

A real-life use case

- Wageningen University & Research (WUR) noticed a high number of failed authentications
- 200.000 attempts in 24 hours
  - 15 attempts per account, then on to the next
  - Appeared to be from a very old account list
  - None succeeded
- Coming from an IP address in China
- Shared this info with SCIRT
  - SURF Community of Incident Response Teams
- Does this IP address show up in combination with other institutions?

WAGENINGEN
UNIVERSITY & RESEARCH

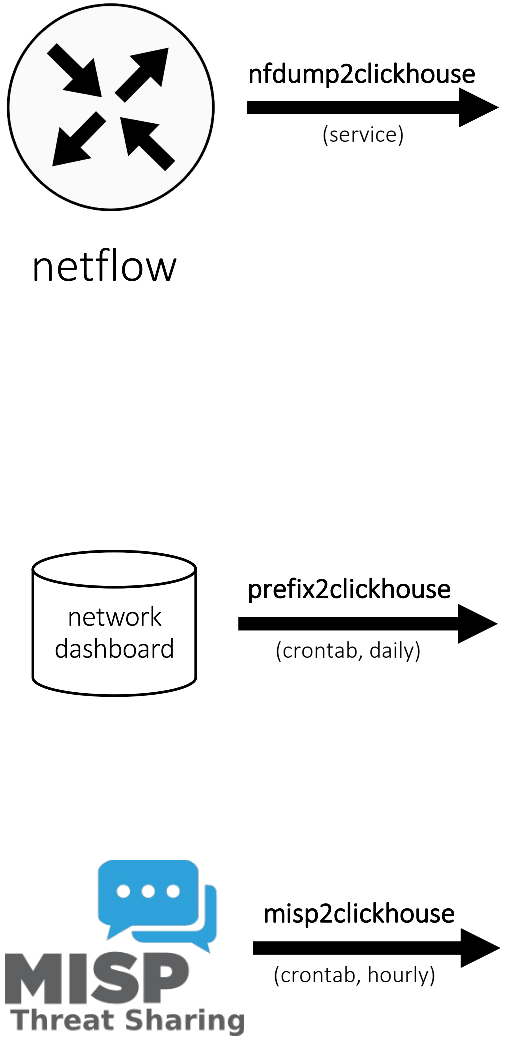
SURF



```
c0e3454a66aa :) select da, min(ts) as earliest, max(ts) as latest, count(), sum(ibyt) from flows where sa='222.211.250.148'
 and ts>='2024-03-04' and ts<='2024-03-13' group by all order by earliest asc;
```

But wait…

- If we can find IPs (+ports) quickly
  - Why not use this to find IoCs?
- Just create a table with known IoCs (IP+port)
  - Luckily there’s an API for that! ☺
- Search for IoCs by comparing flows with IoCs
- Or use materialized views!
  - store results of a query in a table
  - updated automatically when underlying data changes



flows

ts	Start time of the flow	DateTime
te	End time of the flow	DateTime
sa	Source IP address	String
da	Destination IP address	String
sp	Source port	UInt16
dp	Destination port	UInt16
pr	Protocol (e.g. ‘TCP’ or ‘UDP’)	String
flg	Flags (if pr is ‘TCP’)	String
ipkt	Number of packets in this flow	UInt64
ibyt	Number of bytes in this flow	UInt64
smk	Source mask	UInt8
dmk	Destination mask	UInt8
ra	Router (IP) Address	String
in	Input interface number	UInt16
out	Output interface number	UInt16
sas	Source AS number	UInt16
das	Destination AS number	UInt16
exid	Exported id	UInt16
flowsrc	Additional label for this source	String

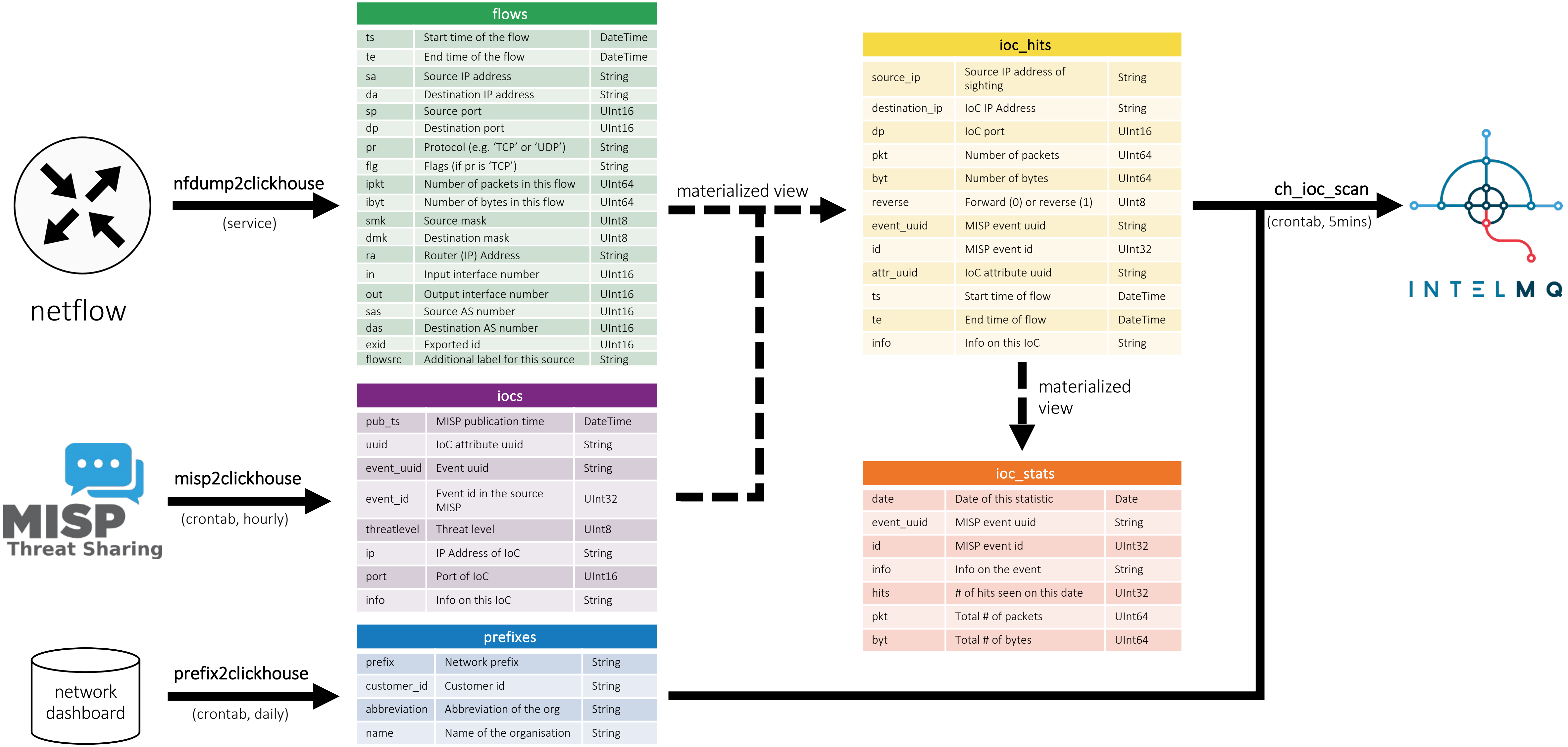
prefixes

prefix	Network prefix	String
customer_id	Customer id	String
abbreviation	Abbreviation of the org	String
name	Name of the organisation	String

iocs

pub_ts	MISP publication time	DateTime
uuid	IoC attribute uuid	String
event_uuid	Event uuid	String
event_id	Event id in the source MISP	UInt32
threatlevel	Threat level	UInt8
ip	IP Address of IoC	String
port	Port of IoC	UInt16
info	Info on this IoC	String

Clickhouse tables



netflow

nfdump2clickhouse

(service)

MISP Threat Sharing

misp2clickhouse

(crontab, hourly)

network dashboard

prefix2clickhouse

(crontab, daily)

flows

ts	Start time of the flow	DateTime
te	End time of the flow	DateTime
sa	Source IP address	String
da	Destination IP address	String
sp	Source port	UInt16
dp	Destination port	UInt16
pr	Protocol (e.g. 'TCP' or 'UDP')	String
flg	Flags (if pr is 'TCP')	String
ipkt	Number of packets in this flow	UInt64
ibyt	Number of bytes in this flow	UInt64
smk	Source mask	UInt8
dmk	Destination mask	UInt8
ra	Router (IP) Address	String
in	Input interface number	UInt16
out	Output interface number	UInt16
sas	Source AS number	UInt16
das	Destination AS number	UInt16
exid	Exported id	UInt16
flowsrc	Additional label for this source	String

iocs

pub_ts	MISP publication time	DateTime
uuid	IoC attribute uuid	String
event_uuid	Event uuid	String
event_id	Event id in the source MISP	UInt32
threatlevel	Threat level	UInt8
ip	IP Address of IoC	String
port	Port of IoC	UInt16
info	Info on this IoC	String

prefixes

prefix	Network prefix	String
customer_id	Customer id	String
abbreviation	Abbreviation of the org	String
name	Name of the organisation	String

materialized view

ioc_hits

source_ip	Source IP address of sighting	String
destination_ip	IoC IP Address	String
dp	IoC port	UInt16
pkt	Number of packets	UInt64
byt	Number of bytes	UInt64
reverse	Forward (0) or reverse (1)	UInt8
event_uuid	MISP event uuid	String
id	MISP event id	UInt32
attr_uuid	IoC attribute uuid	String
ts	Start time of flow	DateTime
te	End time of flow	DateTime
info	Info on this IoC	String

materialized view

ioc_stats

date	Date of this statistic	Date
event_uuid	MISP event uuid	String
id	MISP event id	UInt32
info	Info on the event	String
hits	# of hits seen on this date	UInt32
pkt	Total # of packets	UInt64
byt	Total # of bytes	UInt64

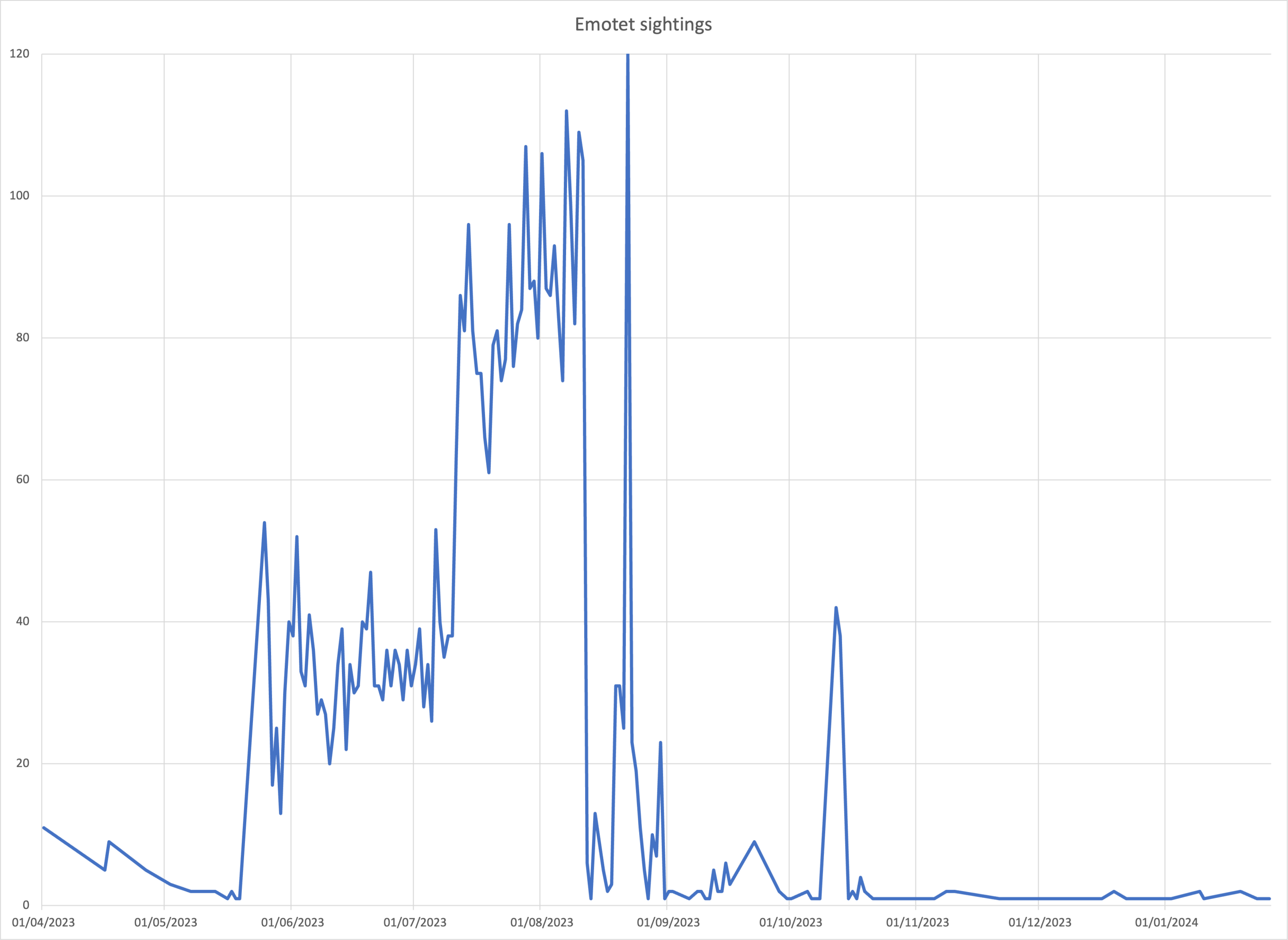
ch_ioc_scan

(crontab, 5mins)

INTELMQ

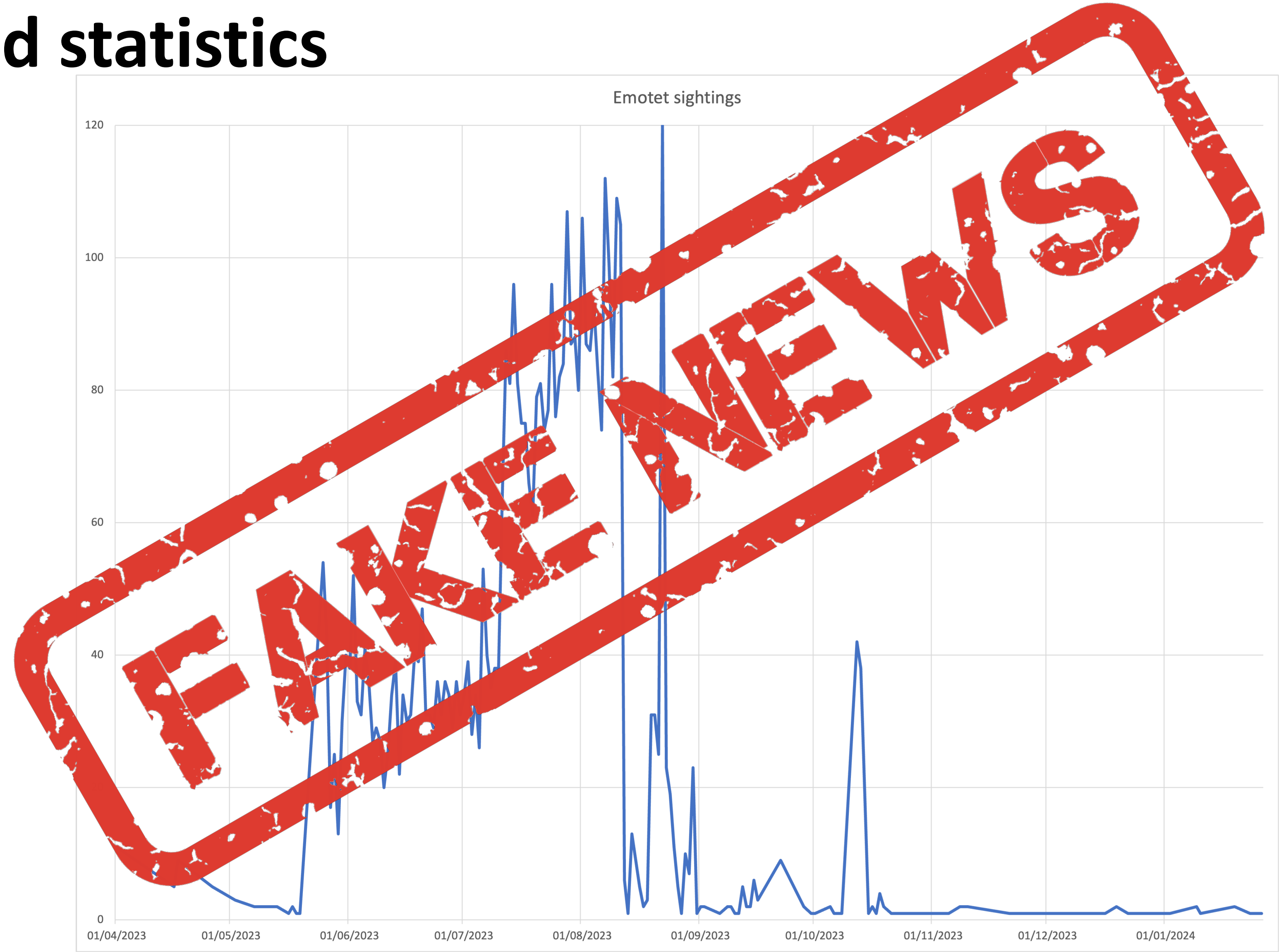
Statistics

- IoC hits of the same event aggregated by date
- Such as Emotet sightings on the SURF network
- Looks like a lot, but…
  - Allmost all are TOR exit nodes,
  - or machines that are scanning
  - This specific IoC is 3 years old…



Lies, damned lies, and statistics

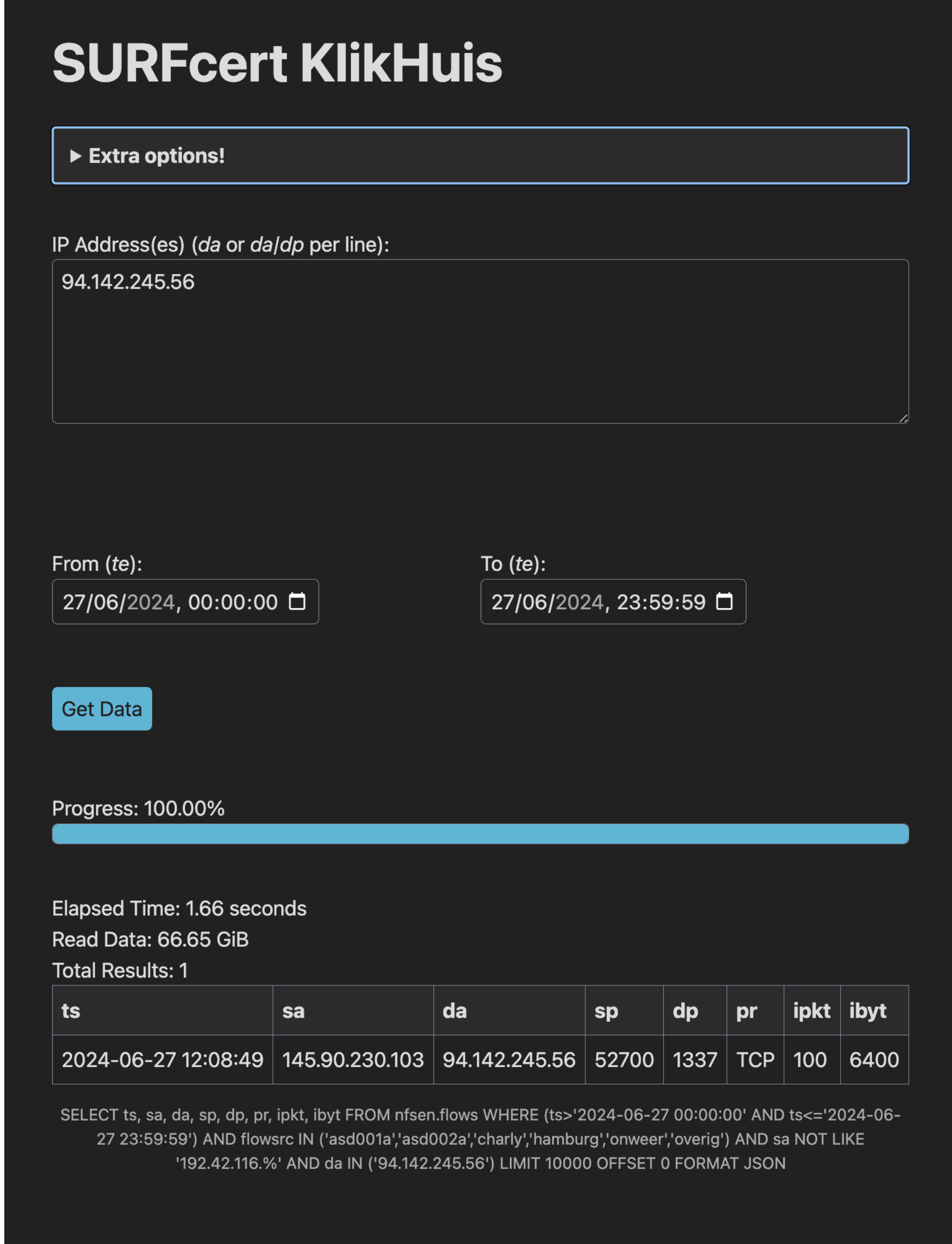
- IoC hits of the same event aggregated by date
- Such as Emotet sightings on the SURF network
- Looks like a lot, but…
  - Allmost all are TOR exit nodes,
  - or machines that are scanning
  - This specific IoC is 3 years old…



SURF

For the quick & easy search

- Easy to use web front-end
- Standard/most used type of queries



SURFcert KlikHuis

IP Address(es) (*da* or *da*/*dp* per line):
94.142.245.56

From (*te*): 27/06/2024, 00:00:00

To (*te*): 27/06/2024, 23:59:59

Progress: 100.00%

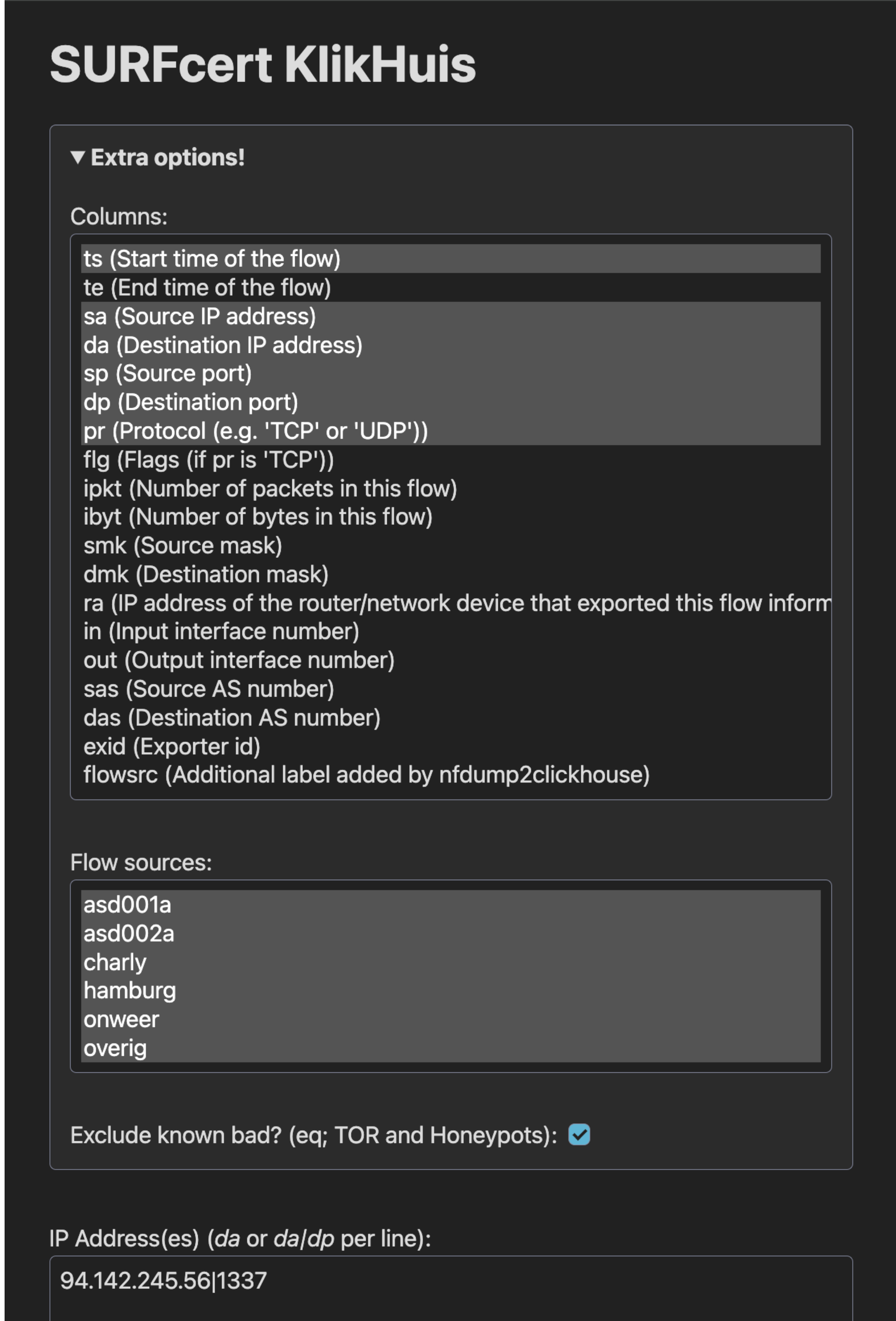
Elapsed Time: 1.66 seconds
Read Data: 66.65 GiB
Total Results: 1

ts	sa	da	sp	dp	pr	ipkt	ibyt
2024-06-27 12:08:49	145.90.230.103	94.142.245.56	52700	1337	TCP	100	6400

```
SELECT ts, sa, da, sp, dp, pr, ipkt, ibyt FROM nfsen.flows WHERE (ts>'2024-06-27 00:00:00' AND ts<='2024-06-27 23:59:59') AND flowsrc IN ('asd001a','asd002a','charly','hamburg','onweer','overig') AND sa NOT LIKE '192.42.116.%' AND da IN ('94.142.245.56') LIMIT 10000 OFFSET 0 FORMAT JSON
```

For the quick & easy search

- Easy to use web front-end
- Standard/most used type of queries
- Additional options
  - columns to include
  - flow sources
  - Excluding noisy areas of the network



SURFcert KlikHuis

▼ Extra options!

Columns:

ts (Start time of the flow)
te (End time of the flow)
sa (Source IP address)
da (Destination IP address)
sp (Source port)
dp (Destination port)
pr (Protocol (e.g. 'TCP' or 'UDP'))
flg (Flags (if pr is 'TCP'))
ipkt (Number of packets in this flow)
ibyt (Number of bytes in this flow)
smk (Source mask)
dmk (Destination mask)
ra (IP address of the router/network device that exported this flow inform
in (Input interface number)
out (Output interface number)
sas (Source AS number)
das (Destination AS number)
exid (Exporter id)
flowsrc (Additional label added by nfdump2clickhouse)

Flow sources:

asd001a
asd002a
charly
hamburg
onweer
overig

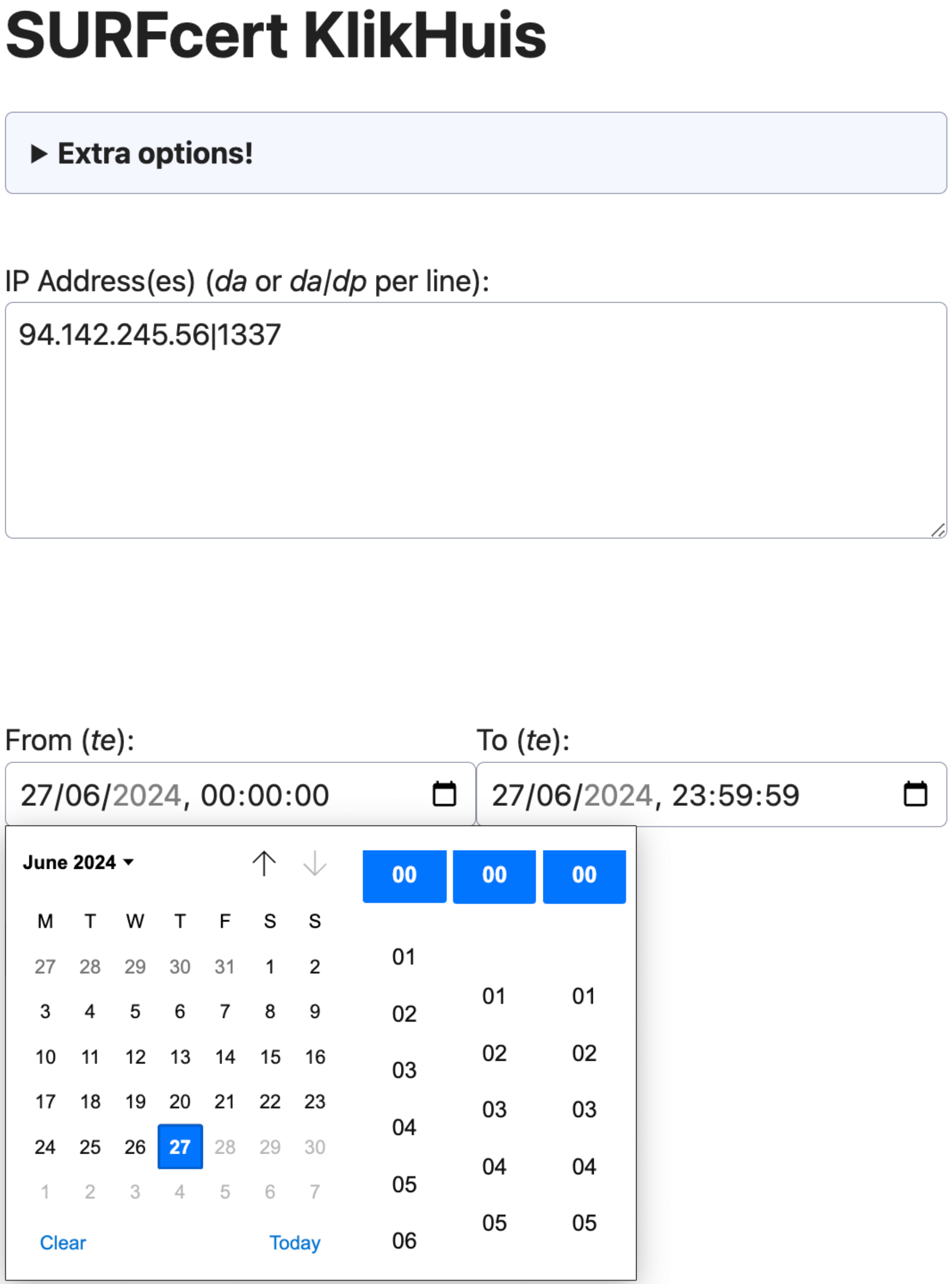
Exclude known bad? (eq; TOR and Honeypots): ☑

IP Address(es) (*da* or *da*/*dp* per line):

94.142.245.56|1337

For the quick & easy search

- Easy to use web front-end
- Standard/most used type of queries
- Additional options
  - columns to include
  - flow sources
  - Excluding noisy areas of the network
- Now also in light mode!



SURFcert KlikHuis

▶ Extra options!

IP Address(es) (*da* or *da|dp* per line):

94.142.245.56|1337

From (*te*): 27/06/2024, 00:00:00

To (*te*): 27/06/2024, 23:59:59

June 2024

M	T	W	T	F	S	S
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
1	2	3	4	5	6	7

Conclusions

- Clickhouse solves main downside of nfsen/nfdump
  - Determining where & when to look, ***fast!***
- Speed enables interactive data analysis
  - Also useful for those ‘*hmm… that’s odd?*’ questions
- Automated watches with materialized views
  - for (curated!) IoC feeds
  - other queries and statistics
- If you have a new hammer, you keep discovering nails
- Easy to explain and (fun to) use(*)

Questions ?

(*) *for the somewhat technically inclined person*