Towards Putting the “Open” in OpenINTEL

Mattijs jonker

UNIVERSITY OF TWENTE.

SURF

SIDN

NLNETLABS

Outline

1. **What the OpenINTEL project is (primarily) about (very briefly)**
2. **Funding (past and current)**
3. **Data Sets**
4. **Measurement Infrastructure**
5. **Analysis Infrastructure**
6. **Value Add / Opportunities**

Project's Objectives

- More than **ten** years ago, we started with **an idea:**

  "Can we measure (large parts) of the global DNS on a daily basis?"

  … setting a goal to make DNS measurement data available to facilitate research and to become the long-term memory of the DNS

- This resulted in our flagship *forward DNS* measurement (March 2015), which we continue to expand by adding new sources of domain names
- In 2020 (Feb), a *reverse DNS (IPv4)* measurement was added
- In 2024 (Jan), we rolled out the first CT-sourced fDNS measurement (to expand coverage of ccTLDs)

Funding (past and current)

- Project Initiation and first years of planning/execution
  - REDACTED
- PMs
  - Roland (SURF, later NLnet Labs)
  - Mattijs (NWO D3)

Funding (past and current)

- External funding
  - REDACTED
  - Incentive programs
  - Tried RIPE CPF, ICANN, Mozilla DFL
  - *Something promising is brewing (confidential)*
- Research Project Grants that “depend” on OpenINTEL
  - “TIDE” (SIDN Funds) – Olivier*
  - “MADDVIPR” (NWO/DHS) – Raffaele*
  - “MASCOT” (NWO) – Etienne

Data Sets that we collect

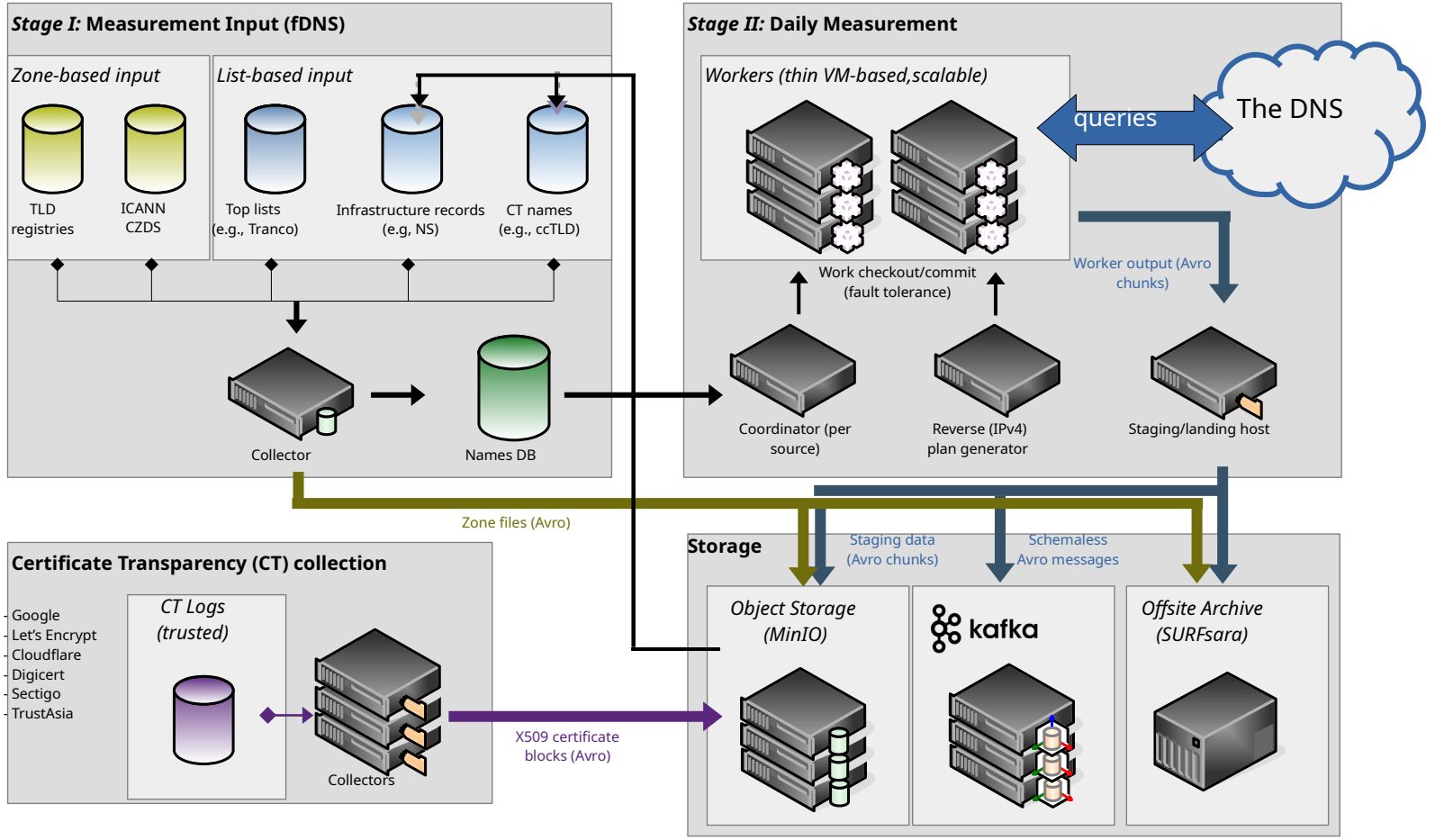
- **OpenINTEL** primarily performs **active** forward and reverse DNS measurements
- We send a fixed set of queries, **once every 24 hours** and **at scale**
  - covering many hundreds of **millions** of domains per day (gTLDs, ccTLDs, other sources) for *fDNS*

    … and with some “measurement feedback” (e.g., we measure infrastructure records)
  - covering a sensible part of the IPv4 address space for *rDNS*
- Mere domain lists (recent development)
  - Registered domains under ccTLD, extracted from CT (pending publication)

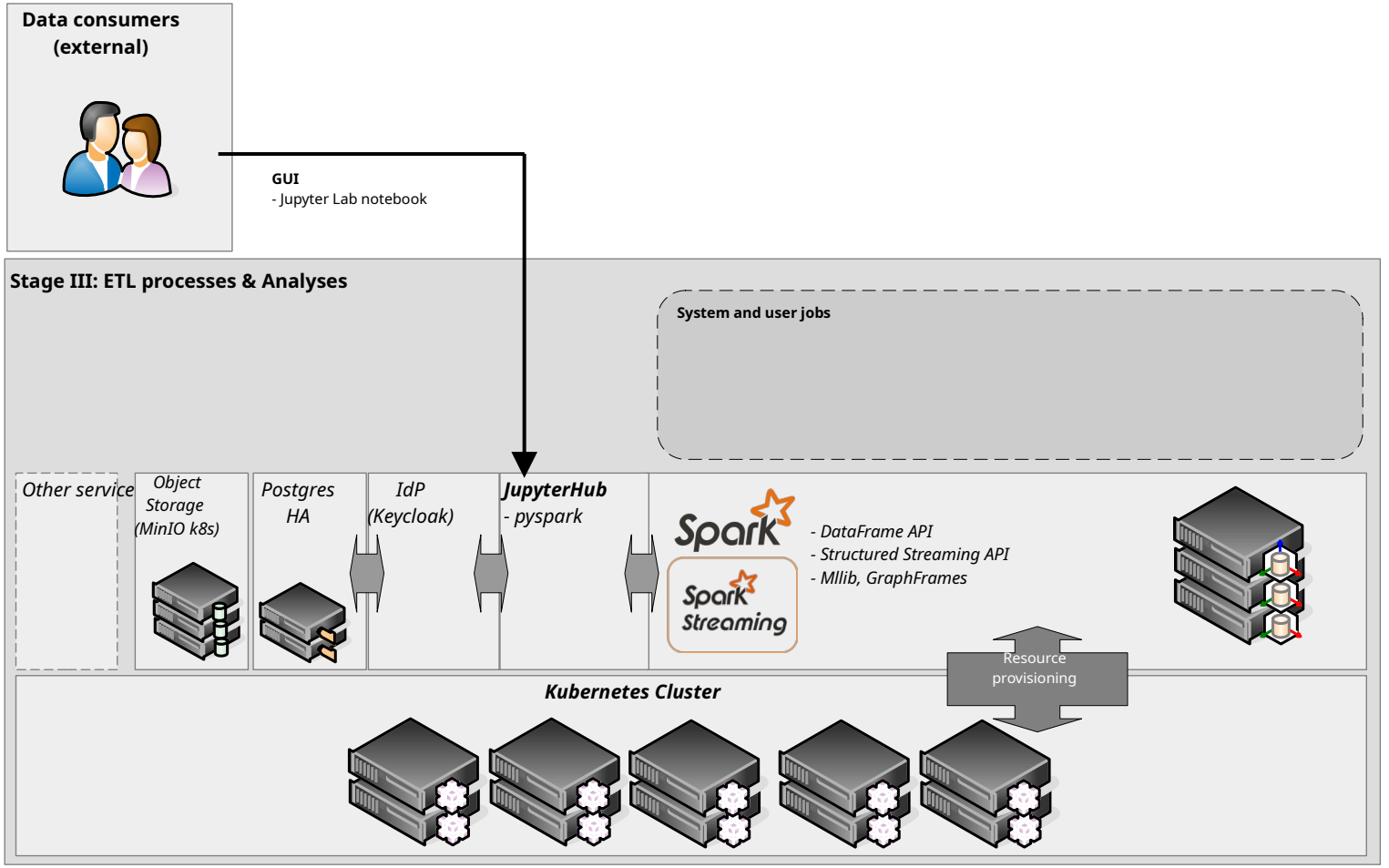
Data Sets and Sharing

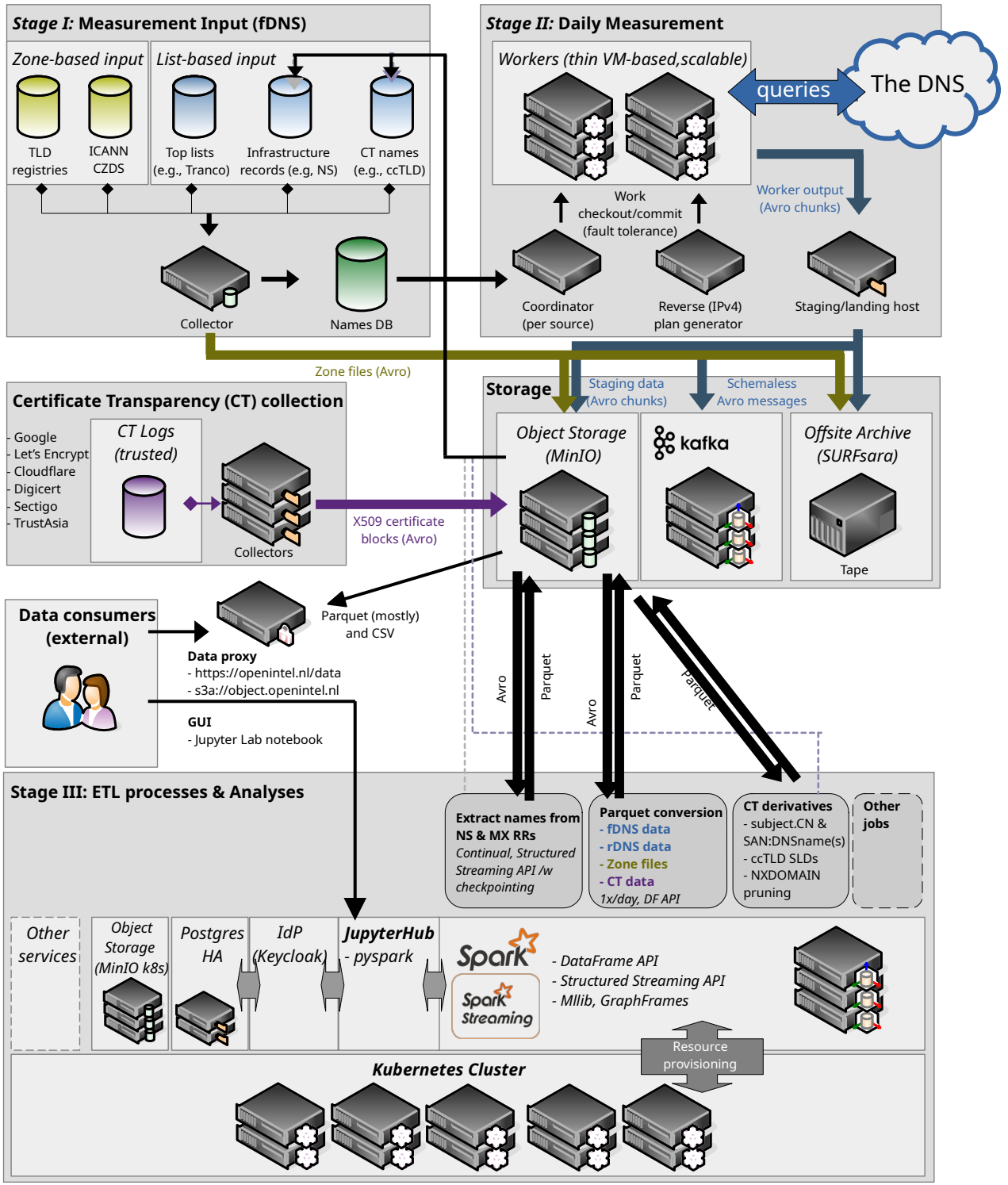
- Zone-based measurement (fDNS)
  - gTLDs – *Show us your contract*
  - ccTLDs (open zone) – *Have at it, hoss*
  - ccTLDs (closed) – *Defer to registry, if deemed worth the effort*
- Reverse DNS – Closed for now due to PII in PTRDNAMEs
- List-based measurements (fDNS)
  - Top lists – *Have at it*
  - Infrastructure – *on request*
  - CT ccTLD (2024-01-26 → current) – *alongside paper publication*
- Domain lists
  - CT ccTLD (#307) weekly, validity-based (2015-01-05 → current) – *alongside paper*
- Zonestream
  - NRDs – *public*
  - ZoneDiff – *public*
- RIR-level rDNS and route obj. delegation data (daughter proj., in collab /w Ioana @ SimulaMet) – *public, http://rir-data.org*

Measurement Infrastructure



Analysis Infrastructure



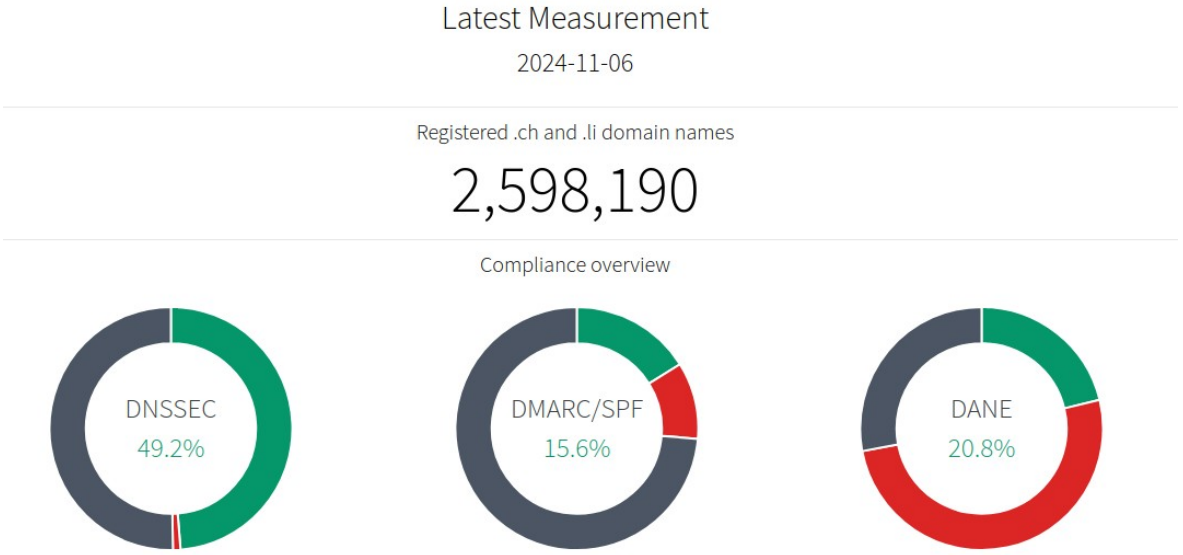

Stage I: Measurement Input (fDNS)
Zone-based input
List-based input
TLD registries
ICANN CZDS
Top lists (e.g., Tranco)
Infrastructure records (e.g, NS)
CT names (e.g., ccTLD)
Collector
Names DB
Stage II: Daily Measurement
Workers (thin VM-based,scalable)
queries
The DNS
Work checkout/commit (fault tolerance)
Worker output (Avro chunks)
Coordinator (per source)
Reverse (IPv4) plan generator
Staging/landing host
Zone files (Avro)
Certificate Transparency (CT) collection
- Google
- Let's Encrypt
- Cloudflare
- Digicert
- Sectigo
- TrustAsia
CT Logs (trusted)
Collectors
X509 certificate blocks (Avro)
Storage
Staging data (Avro chunks)
Schemaless Avro messages
Object Storage (MinIO)
kafka
Offsite Archive (SURFsara)
Tape
Data consumers (external)
Parquet (mostly) and CSV
Data proxy
- https://openintel.nl/data
- s3a://object.openintel.nl
GUI
- Jupyter Lab notebook
Avro
Parquet
Avro
Parquet
Parquet
Stage III: ETL processes & Analyses
Extract names from NS & MX RRs
Continual, Structured Streaming API /w checkpointing
Parquet conversion
- fDNS data
- rDNS data
- Zone files
- CT data
1x/day, DF API
CT derivatives
- subject.CN & SAN:DNSname(s)
- ccTLD SLDs
- NXDOMAIN pruning
Other jobs
Other services
Object Storage (MinIO k8s)
Postgres HA
IdP (Keycloak)
JupyterHub - pyspark
Spark
Spark Streaming
- DataFrame API
- Structured Streaming API
- Mllib, GraphFrames
Resource provisioning
Kubernetes Cluster

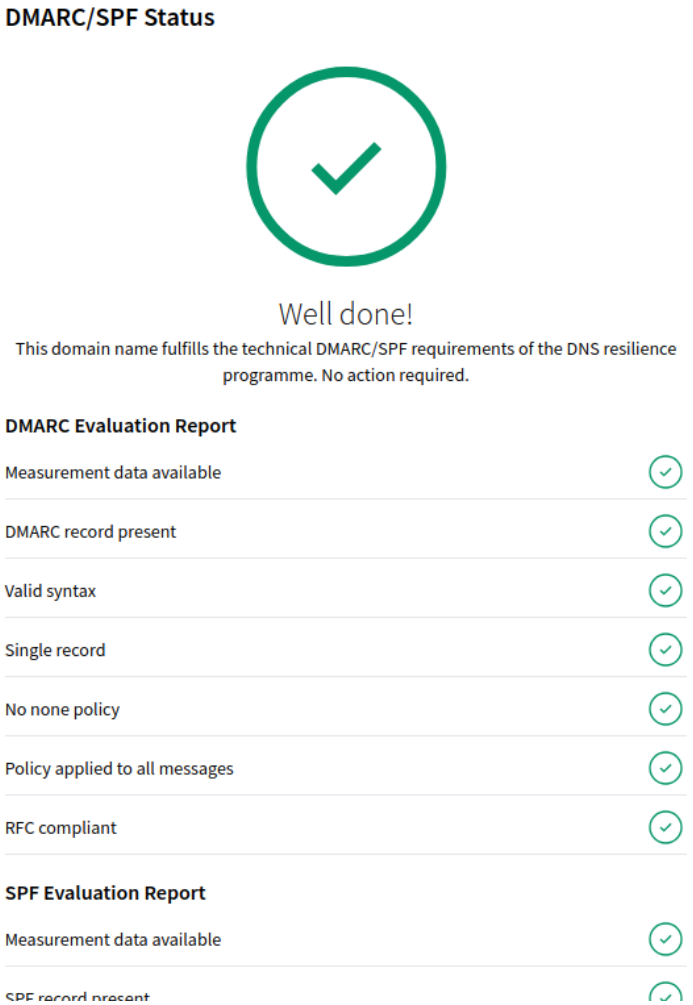
Value Add 1/2

- Academic insights
  - Racked up quite a few papers – diverse topics incl. security, systems and infrastructure resilience, centralization, (geopolitical) provisioning decisions, operational aspects
  - Data synergy – IYP (/w IIJ), DNSAttackStream (/w CAIDA)
- Use in educatoin (UT IM course)
- Policymaking and governance
  - Risk Report Cybersecurity & Economics (2018) – *CPB for Ministry of Justice and Security*
  - Strategic Advisory Report on e-Gov DNS Infrastructure (2022) – *NCSC-NL*

Value Add 2/2

- Operator decision-making
- Incentive programs
- REDACTED





Opportunities / Ideas / Plans

- Share more DNSnames learned from CT logs /w community

  e.g., FQDNs of interest (vpn, citrix, secure)
- Share rDNS extracts (in light of PTRDNAME PII concerns)

  e.g., address delegation structure
- “DNSstream” (BGPstream-like tool & BGP2GO-like idx) (/w Alfred, Ioana @ SimulaMet)
- Do more with the CT data (we store the full certs) ~103 logs, 41B leaf entries, 8B(?) unique
- Share Graphs (e.g., REDACTED, delegation structure) (long in the planning)
- Start 1x/week Web crawling (ccTLD & other names of interest)
  - Abuse detection based on content & infra (exploring collab)
  - Misinformation / (geo)political content
- Provide ccTLD names to Common Crawl so they can expand (exploring)

Questions?