Towards Putting the “Open” in OpenINTEL

Mattijs jonker

UNIVERSITY OF TWENTE.

SURF

SIDN

NLNETLABS

Outline

1. **What the OpenINTEL project is about**
2. **Data Sets**
3. **Measurement Infrastructure**
4. **Analysis Infrastructure**
5. **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?"
- This resulted in our flagship *forward DNS* measurement (March 2015)
  - Largely seeded from zonefiles (and several Top lists)
  - Over time we expanded by adding gTLD and ccTLD zones
- In 2020 (Feb), the daily *reverse DNS (IPv4)* measurement saw the light of day
- Our goal:
  - To become the long-term memory of the DNS
  - To facilitate research and make DNS measurement data available

Project’s Objectives

- Using zonefiles is accompanied by data sharing challenges
  - For gTLDs we solve this by verifying that requester has zonefile access
  - For closed ccTLD zones (/w strict contracts) it is challenging
- A few years ago, during AIMS-KISMET 2020, in this very auditorium, I wondered**:** *“Why don’t we collect CT, given it’s a rich source of domains that are already public?”*
  - This prompted CT log scraping efforts (more details in backup slides)
- In Jan 2024 we rolled out a CT-sourced fDNS measurement
  - To expand coverage of country-code TLDs

    … and to be able to share more measurement data publicly

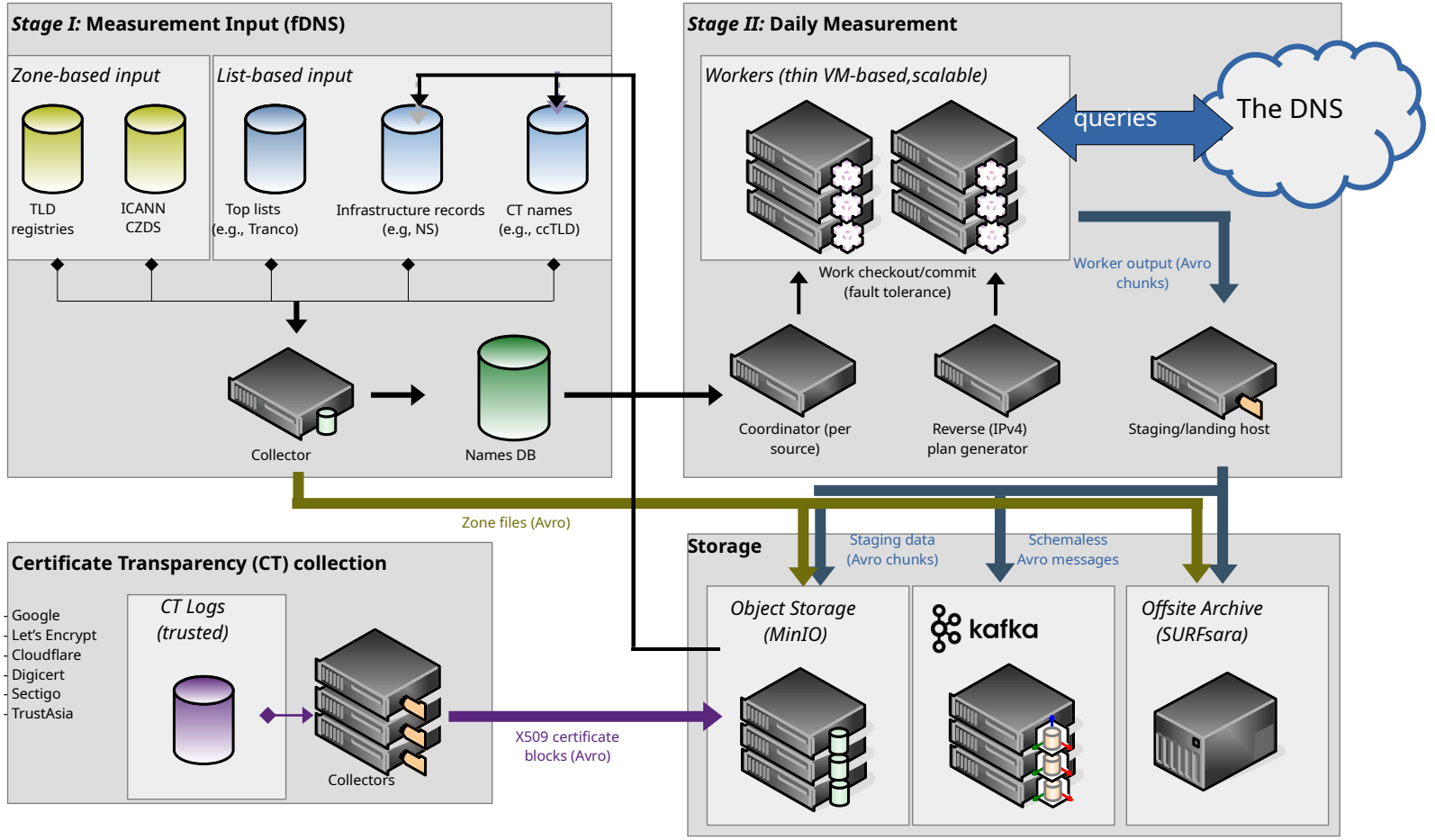
The primary fDNS and rDNS measurements

- OpenINTEL performs active forward and reverse DNS measurements
- We send a fixed set of queries, once every 24 hours
  - Covering many hundreds of millions of domains per day (gTLDs, ccTLDs, Top lists, and CT-sourced) for *fDNS*
  - Covering a sensible part of the IPv4 address space for *rDNS*
  - With a “measurement feedback loop” (e.g., we measure infra records)

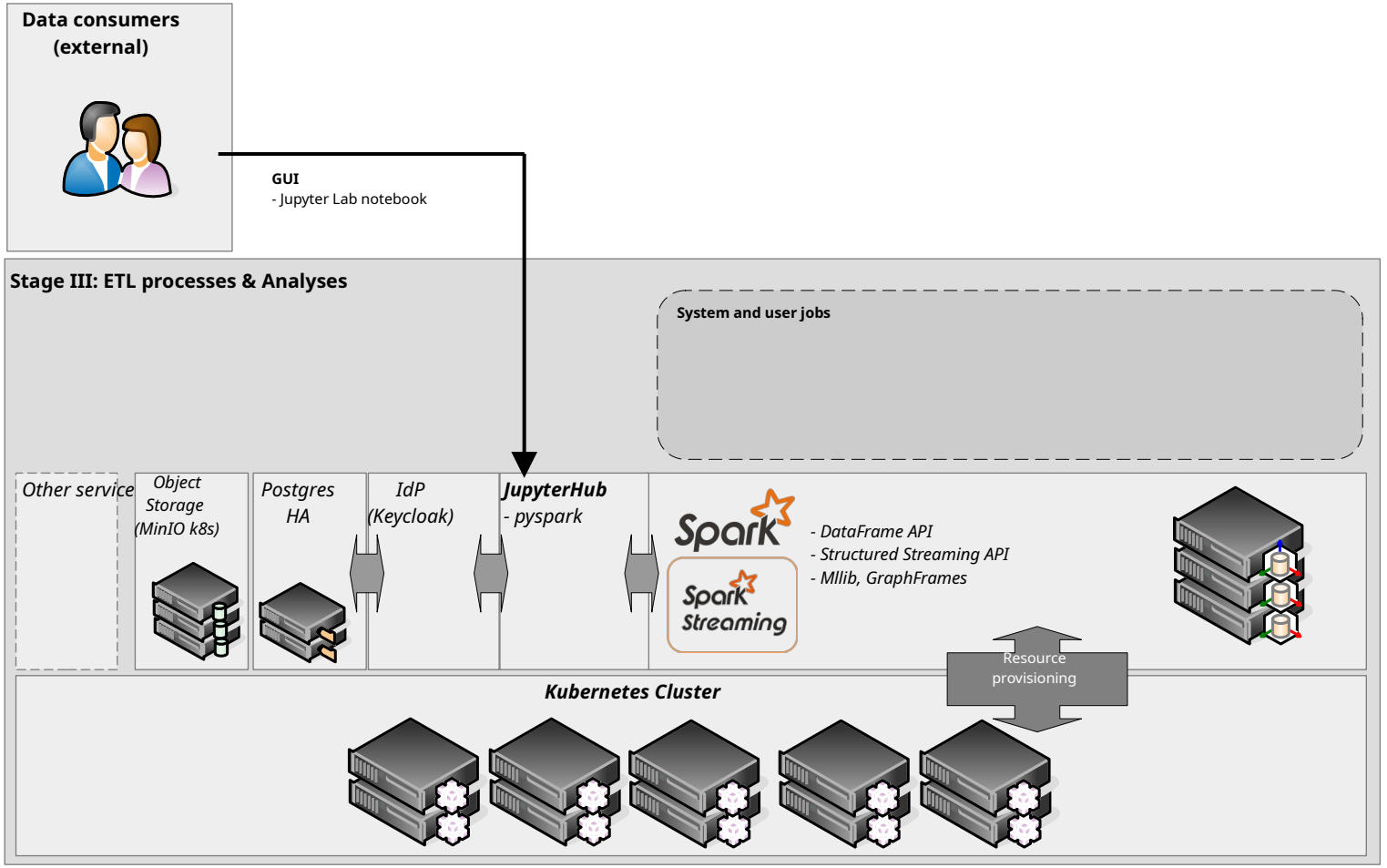
Datasets and sharing

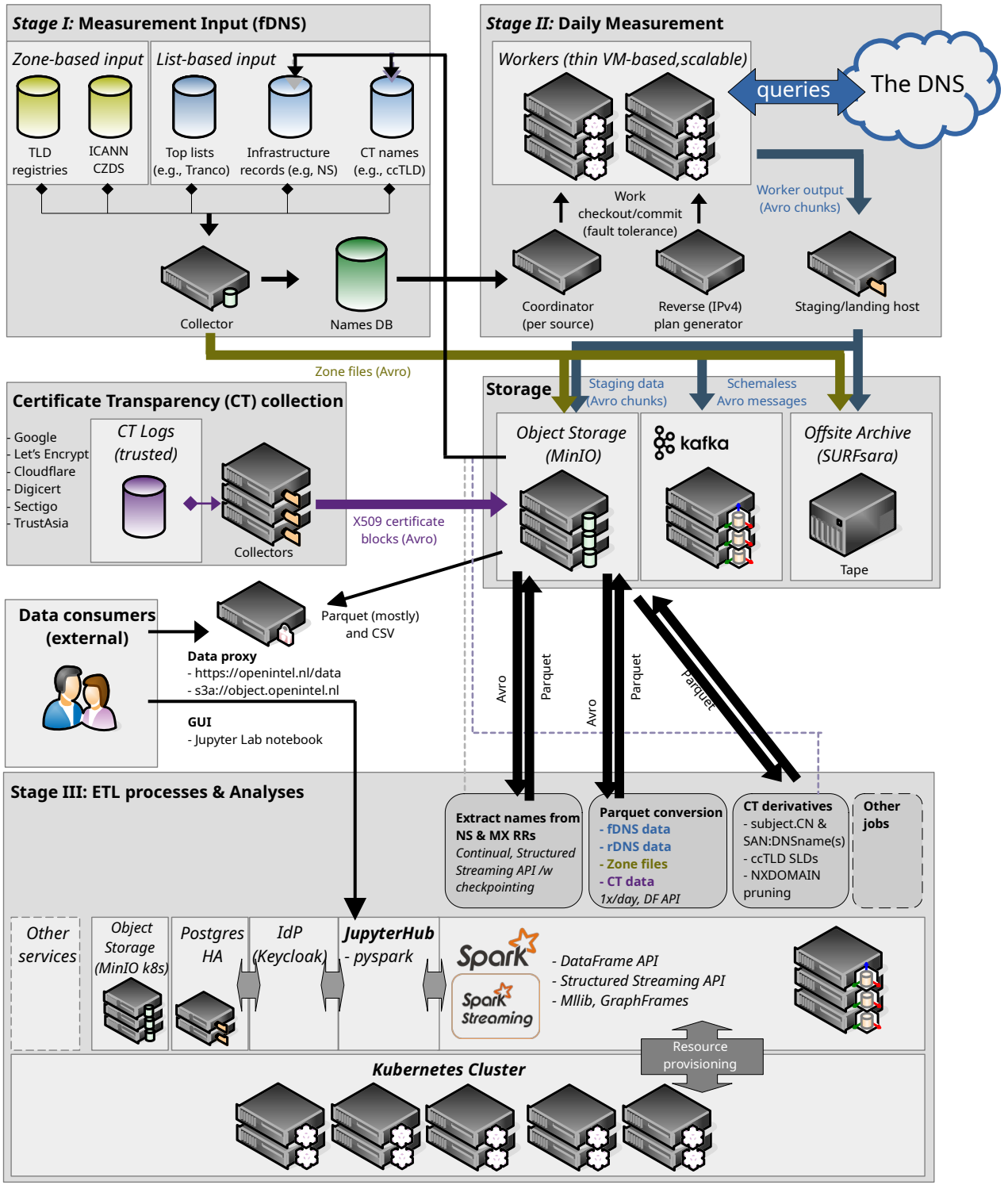
- Zone-based measurement data (fDNS)
  - gTLDs – *Show us your contract*
  - ccTLDs (open zone) – *Public data*
  - ccTLDs (closed) – *Defer to registry, if deemed promising*
- Reverse DNS measurement data – *Closed, case-by-case*
- List-based measurements (fDNS)
  - Top lists – *Public data*
  - Infrastructure (e.g., MX records)– *Closed, case-by-case*
  - CT-based ccTLD – *Pending publication*
- Domain lists
  - ccTLD names∈CT – 307 ccTLDs, weekly, validity-based
- Zonestream (Transparency presentation)
  - NRDs – *Public*
  - ZoneDiff – *Public*
- RIR-level rDNS zones and route obj. delegation (in collab /w 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

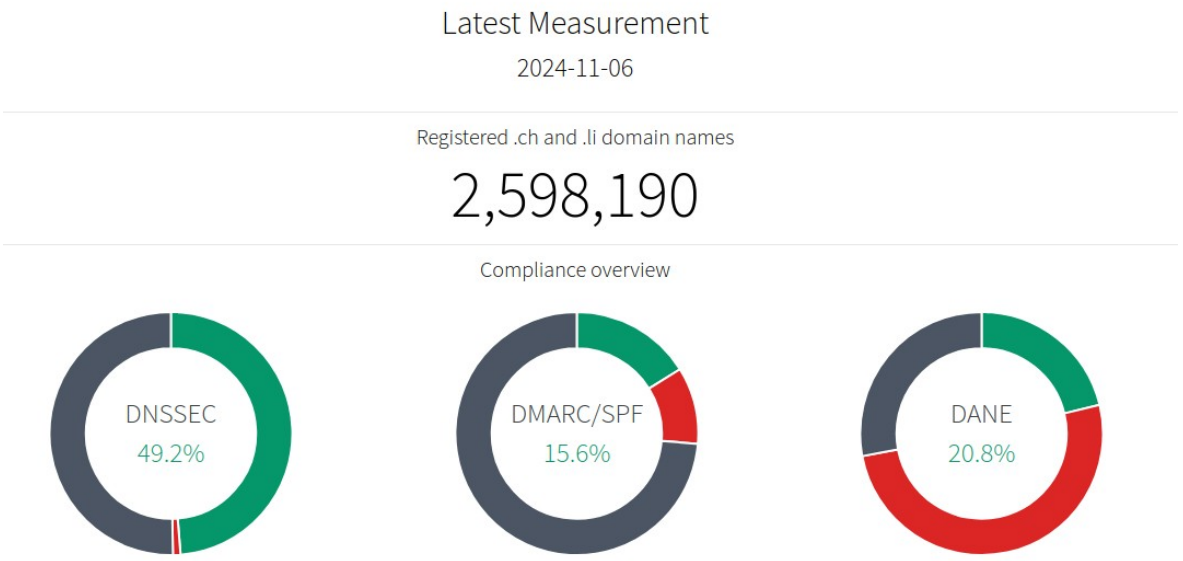
- Academic insights
  - Resulted in quite a few papers – diverse topics incl. security, systems and infrastructure resilience, centralization, (geopolitical) provisioning decisions, operational aspects
  - Data synergy – IYP (/w IIJ), DNSAttackStream (/w CAIDA)
- Use data and compute infra in UT courses: *Internet Measurements, Cloud Networking*
- 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*

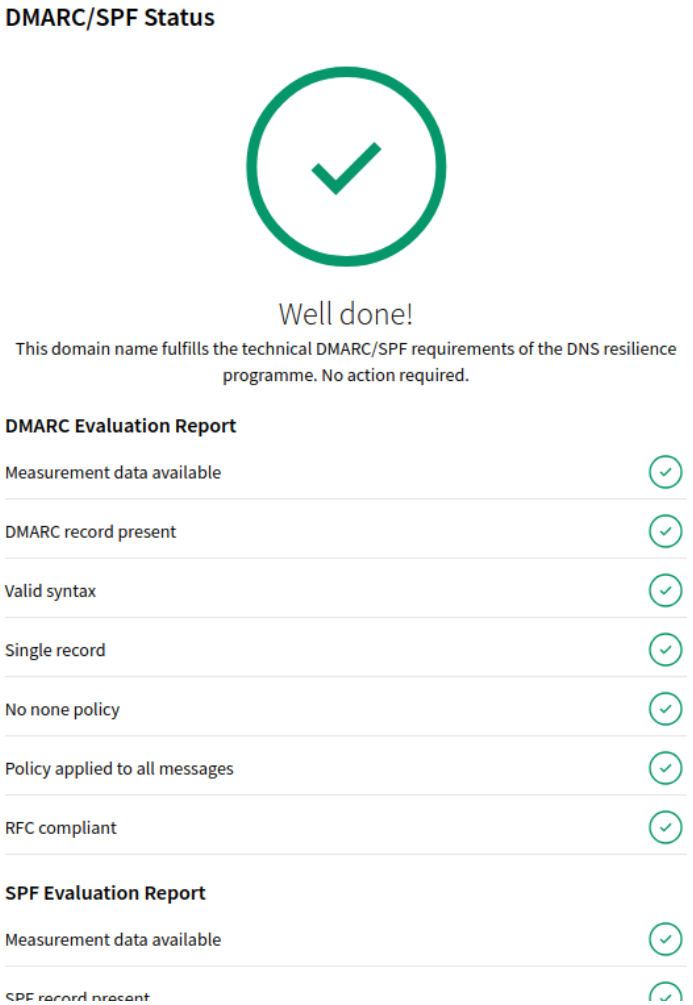
Value Add 2/2

- Operator decision-making
- Incentive programs

  SWITCH.ch – https://dns-resilience.openintel.nl/

  IIS.se (no visual)





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” tool (Separate presentation)
- Do more with the CT data (we store the full certs) ~105 logs, 43B entries, 8B(?) unique certs
- Share Graphs (e.g., SPF trust, delegation structure) (long in the planning)
- Start 1x/week Web crawling (ccTLD & other names of interest)
  - Abuse detection based on content & infra
  - Misinformation / (geo)political content

Questions?

BACKUP SLIDES

Certificate Transparency (CT)101

- Certificate Transparency involves public, append-only logs
  - To audit and monitor issuance by certificate authorities
- Uses CT Internet standard (v2.0 in RFC 9162)
- Browser vendors (e.g., Google) drove adoption: certificates need to be included in logs to be accepted by browser
- Policies have changed over years (e.g., Chrome’s 1 Google-operated log SCT requirement)

Integrating CT in OpenINTEL

- A few years ago, during WIE-KISMET 2019, in this very auditorium, I wondered**:**

  “Why don’t we collect CT, given it’s a rich source of domains that are already public?”
- This prompted scraping efforts
- I found a blog article: *“Retrieving, Storing and Querying 250M+ Certificates Like a Boss”*
  - Offers a tool, *Axeman*, to scrape CT (by “CaliDog” – how befitting)
  - Modified it to output Avro data (more changes necessited over time)
  - Built pipeline for continual scraping, data warehousing
- With various logs exceeding 1-2 billion certs, and ~43 billion CT entries retrieved, stored and queried, 250M doesn’t seem like much anymore
  - Logs are nowadays typically *temporally sharded* (e.g., Nimbus 2024, Oak 2024H1)
- Logs are retired once considered no longer needed

Funding (past and current)

- Project Initiation and first years of planning/execution
  - Initial measurement cluster (2014) and extension (2015) – SURFnet
  - Initial analysis cluster (2015) – SURFnet & SIDN
  - Hadoop nodes extension (2017) – UTwente
  - Measurement cluster replacement (2018) – UTwente
  - Hadoop storage upgrade (2019) – UTwente
  - Streaming server / Kafka broker (2020) – Utwente
- PMs
  - Roland (SURF, later NLnet Labs)
  - Mattijs (NWO D3)

Funding (past and current)

- External funding
  - IIS.se (incentive program, ended)
  - SWITCH.ch (DNS resilience program, 2021-06 → 2026-12)
    - → New compute cluster (2022) and extension (2023)
    - → Research engineer
  - 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