Open INTEL

Creating a long-term “memory” for the global DNS

Mattijs Jonker

UNIVERSITY OF TWENTE.

SURF NET

SIDN

NLNETLABS

Introduction

- Almost **five years** ago, we started with **an idea:**

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

- In this talk, I will discuss:
  - The data we gather (nowadays)
  - How do we perform our measurements
  - Which data do we share
  - And planned improvements / additional data

How we perform our measurements

- **OpenINTEL** performs an **active** measurement, sending a fixed set of queries for all covered names, **once every 24 hours**
- We do this **at scale**, covering over **227 million** domains per day:
  - **gTLDs**:

    .com, .net, .org, .info, .mobi, .aero, .asia, .name, .biz, .gov

    + almost 1200 "new" gTLDs (.xxx, .amsterdam, .berlin, ...)
  - **ccTLDs:**

    .nl, .se, .nu, .ca, .fi, .at, .dk, .ru, .рф, .us, .na, .gt, .co
  - Various **other** sources:

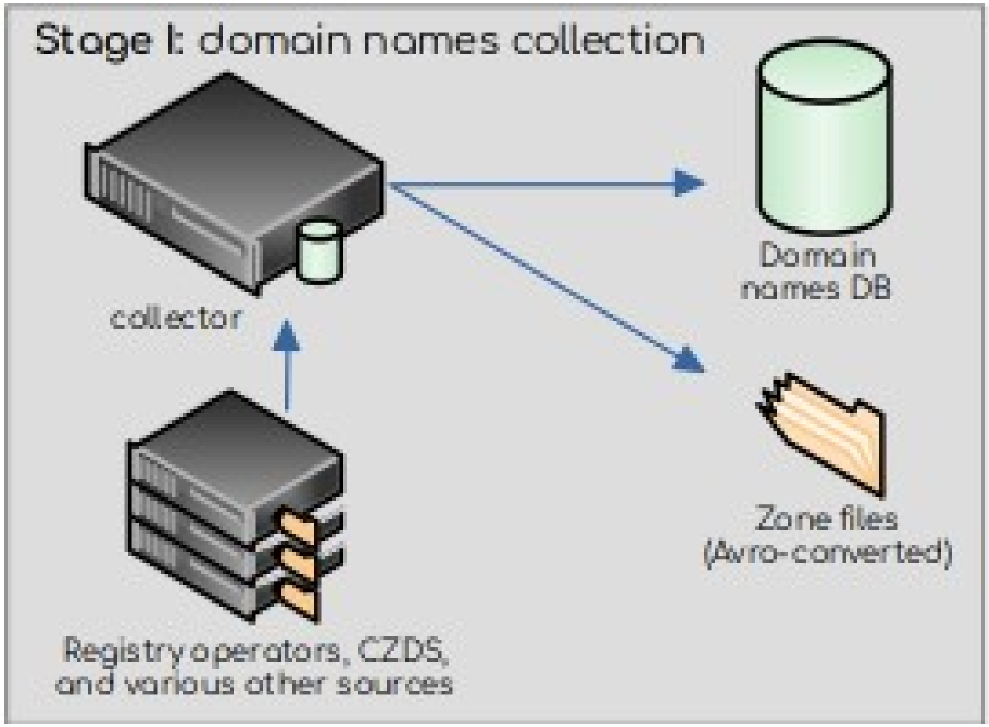
    Alexa top 1M, Cisco Umbrella, diverse blacklists

How we perform our measurements

- The measurement process involves three stages
  1. Extraction of **names**
  2. Active measurement
  3. Streaming and persisting data

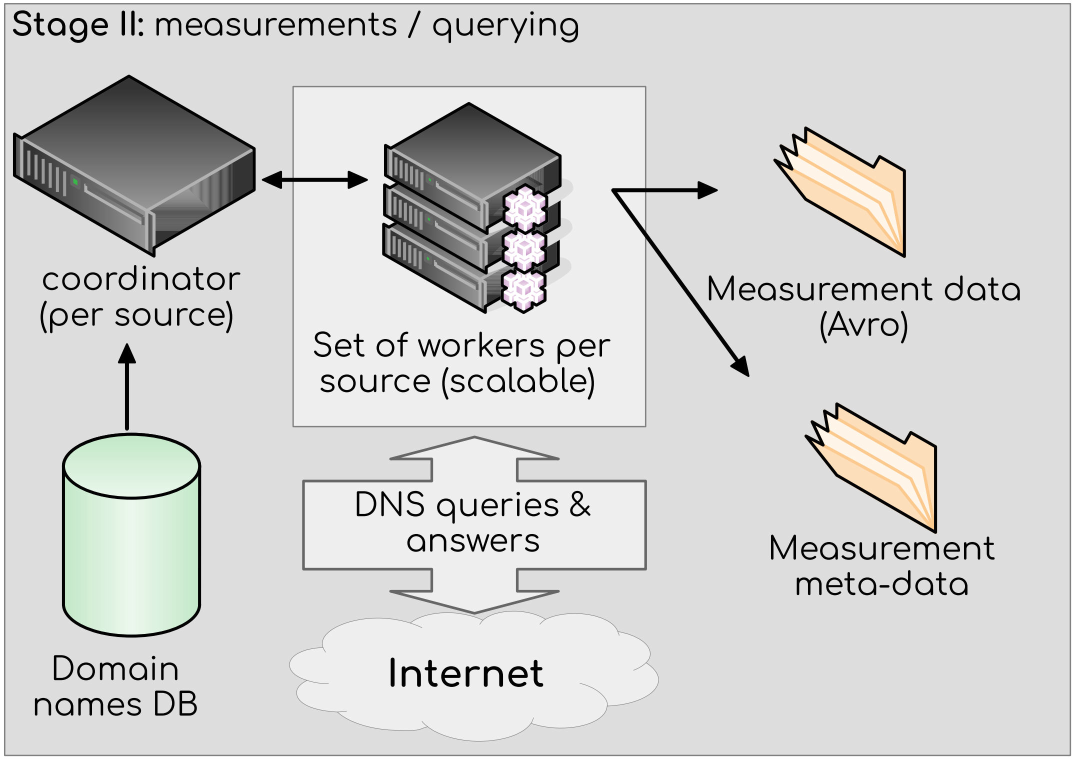
Stage I: collecting names

- Extraction of **names** from zone files and other sources (at least once daily)
- Store state of covered namespace in "names to measure" DB
- Convert zone files to Avro


Stage I: domain names collection
collector
Domain names DB
Zone files (Avro-converted)
Registry operators, CZDS, and various other sources

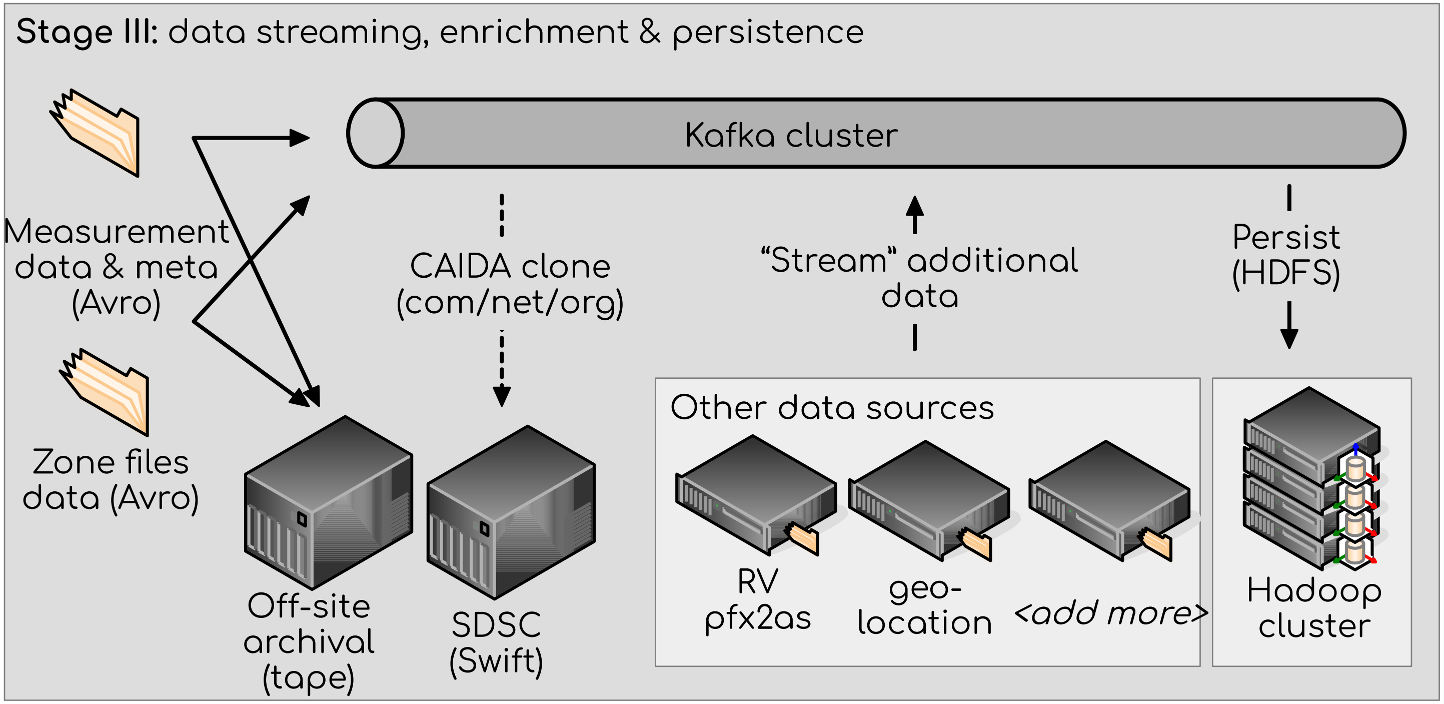
Stage II: main measurement

- Actively sending queries for all collected **names** (daily)
- Workers write results to files, chunked per 100k names
- Also track measurement performance (meta-data)



Stage III: storage and persistence

- We stream the data (measurement, meta, zone files) to a Kafka cluster
  - Allows near real-time stream-based analysis (WIP)
- Data is persisted in HDFS
  - allowing batch-based, longit. analyses (many successes)
- Clone data off-site (archive on tape & CAIDA clone)
- We are adding additional data to our streaming system (e.g., CTLs, RPKI data, ...)



What do we have, in simple numbers

- Started measuring February, 2015
- We collect over $2.4 \cdot 10^9$ DNS records each day
- So far, we collected over $3.6 \cdot 10^{12}$ results (3.6 trillion)

Which data do we share

- We share open data publicly
  - Open sources (e.g., .se, .nu, Alexa)
  - As Avro files on openintel.nl, /w “light“ docs
- We share closed data with other researchers
  - Typically require them to have registry operator contracts
- We share closed data with the respective registry operators

Ongoing and planned improvements

- More and improved data sharing
  - Aggregate datasets
  - Public Kafka broker
  - Rolling stats & insights (openintel.nl)
  - Jupyter containers (Dockerfile) /w example analyses

    (also for education purposes)
- Fusing more data in streaming system
  - e.g.: certificate transparency logs, BGP events, outages, DoS attacks, ...
- Reverse address space measurements (in-addr.arpa)
- Targeting additional authoritative(?) name servers
- Support distributed (multi-VP) measurement

Questions ?