
**OARC 43**

**26 Oct 2024**

# Project: Crunchy DITL

Jerry Lundström
Software Engineer
jerry@dns-oarc.net

# Analyzing DITL in a different way

- Build a prototype platform that has capabilities to analyze DITL Data
  - Using common scalable open source data tools such as Hadoop, Hive, Parquet, Clickhouse and/or ENTRADA
  - Using high level query languages such as Structure Query Language (SQL)
  - Primarily focused on the DNS message, but include some aspects of the transport
  - Be able to adapt to changes in the DNS, data processing and analysis tools

❤ Verisign

DNS-OARC

Domain Name System Operations Analysis and Research Center

# Picking a “Data Lake”

- Read a lot of articles on Data Lakes
- Asked DITL researchers how they query/process DITL today
- Talked to community members that run Data Lakes today
- Evaluated ClickHouse and Apache Iceberg/Spark
  - Many issues getting Iceberg/Spark to work
  - ClickHouse just worked

DNS-OARC

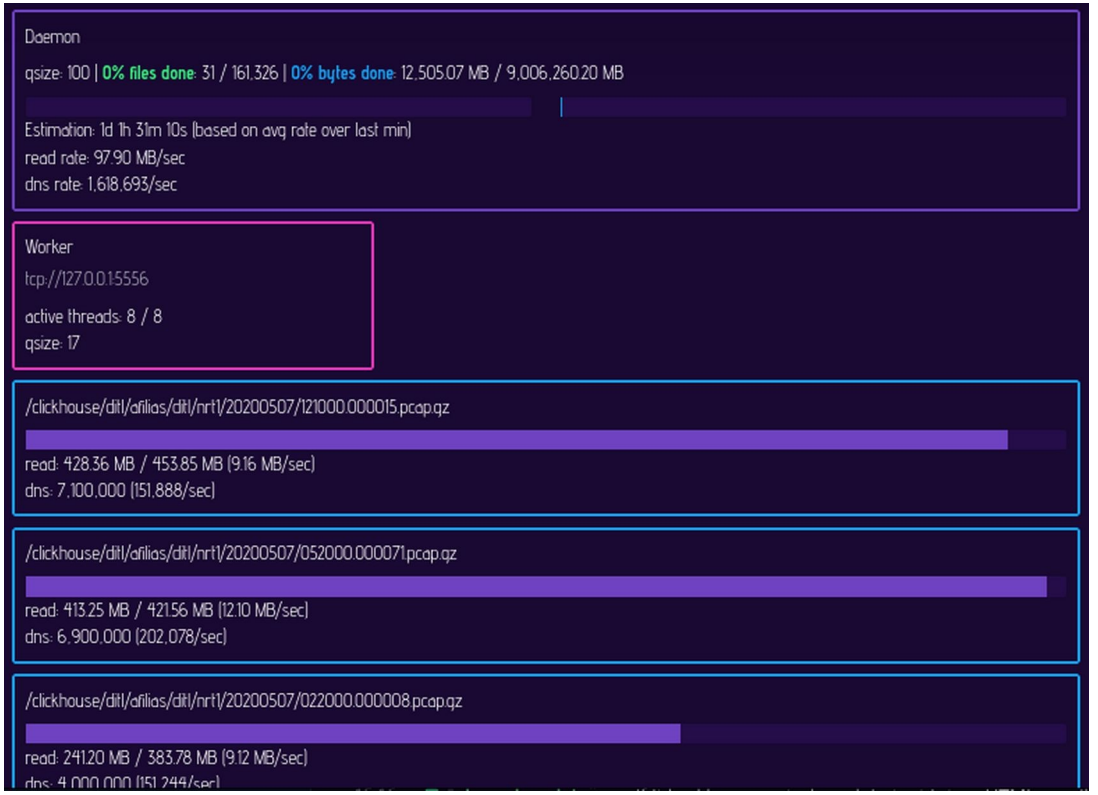
Domain Name System Operations Analysis and Research Center

# ClickHouse

- Picked ClickHouse because:
  - Great performance on common hardware
  - Hardware specification very similar to our Ceph nodes, reusable if project fails
- Two servers, 64GB mem, 5x12TB hdd lvm2 raid5
- Ceph had access/network issues, so copied DITL data locally before processing

# Processing compressed PCAPs

- crunchy-munchy
  - dnsjit, input.zmmpcap, lib.clickhouse
- crunchy-control
  - Flask/socket.io
  - ZeroMQ
- crunchy-explorer →



# Import results

- DITL 2020 RAW compressed PCAPs, 16.3 TB
- 297.93 B (297,925,409,197) DNS records imported
- ~11.5 TB compressed, ~50.7 TB uncompressed in ClickHouse (per schema variant)
- 5 schema variants: 1) DNS hdr/flags as a bitfield or 2) as booleans. 3) QNAME reversed and 4) as an array. 5) site, server, source referrenced externally.

DNS-OARC

Domain Name System Operations Analysis and Research Center

# QNAME reversed in an array

```
>>> "www.example.com".split(".")[::-1]
['com', 'example', 'www']
```

- Opens interesting ways to query data
  - `WHERE qname[1] = ‘com’` A specific TLD
  - `WHERE empty(qname)` Root

DNS-OARC

Domain Name System Operations Analysis and Research Center

# Ready… set… SELECT!

```
SELECT count(*)
FROM crunchy.ditl_bools
WHERE (qr = false) AND (do = true)

    ┌─────count()─┐
1.  │ 207592008622 │ -- 207.59 billion
    └──────────────┘

1 row in set. Elapsed: 174.231 sec. Processed 297.93 billion rows, 480.97 GB
(1.71 billion rows/s., 2.76 GB/s.)
Peak memory usage: 118.27 MiB.
```

DNS-OARC

Domain Name System Operations Analysis and Research Center

# Inventorizing EDNS params from priming queries

```
SELECT
    (edns, edns_flags, edns_bufsize),
    count()
FROM crunchy.ditl_rqn
WHERE empty(qname) AND (qtype = 2)
GROUP BY (edns, edns_flags, edns_bufsize)
ORDER BY count() DESC
INTO OUTFILE 'priming query edns parameters.txt'
FORMAT csv

↖ Progress: 192.93 billion rows, 6.14 TB (25.74 million rows/s., 819.55 MB/s.)
936 rows in set. Elapsed: 9942.741 sec. Processed 297.93 billion rows, 9.51 TB (29.96 million rows/s., 956.93 MB/s.)
Peak memory usage: 218.77 MiB.
```

2h 45min compared to custom C code taking several days

DNS-OARC

Domain Name System Operations Analysis and Research Center

Q&A / Live demo in hallway during breaks!

jerry@dns-oarc.net

@jelu on Mattermost

#OARC Software

DNS-OARC

Domain Name System Operations Analysis and Research Center

# Software Projects & Funding

https://www.dns-oarc.net/oarc/software

- Overview of software developed and maintained by OARC
  - dsc, dsc-datatool, dnscap, dnsperf, dnsjit, drool, packetq, tinyframe, dnswire and more
- Information about funding development, licensing policy, links to GitHub project pages and mailing lists

DNS-OARC

Domain Name System Operations Analysis and Research Center