TLD Data Analysis

ICANN Tech Day, Dublin
October 19th 2015
Maarten Wullink, SIDN
SIDN

- Domain name registry for .nl ccTLD
- > 5,6 million domain names
- 2,46 million domain names secured with DNSSEC
- SIDN Labs is the R&D team of SIDN
DNS Data @SIDN

• > 3.1 million distinct resolvers

• > 1.3 billion query's daily

• > 300 GB of PCAP data daily
ENTRADA
ENhanced Top-Level Domain Resilience through Advanced Data Analysis

• **Goal**: data-driven improved security & stability of .nl and the Internet at large
• **Problem**: Existing solutions for analyzing network data do not work well with large datasets and have limited analytical capabilities.
• **Main requirement**: high-performance, near real-time data warehouse
• **Approach**: avoid expensive pcap analysis:
  • Convert pcap data to a performance-optimized format (key)
  • Perform analysis with tools/engines that leverage that
Use Cases

Focussed on increasing the security and stability of .nl

- Visualize DNS patterns (visualize traffic patterns for phishing domain names)
- Detect botnet infections
- Real-time Phishing detection
- Statistics (stats.sidnlabs.nl)
- Scientific research (collaboration with Dutch Universities)
- Operational support for DNS operators
Example Applications

• DNS security scoreboard

• Resolver reputation
DNS Security Scoreboard

**Goal:** Visualize DNS patterns for malicious activity

**How:** Combine external phishing feeds with DNS data
Traffic Visualization

Overview

Network

<table>
<thead>
<tr>
<th>ASN</th>
<th>#</th>
<th>ASN</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS15169</td>
<td>66</td>
<td>AS15169</td>
<td>10</td>
</tr>
<tr>
<td>AS38406</td>
<td>38</td>
<td>AS8737</td>
<td>3</td>
</tr>
<tr>
<td>AS202109</td>
<td>22</td>
<td>AS31334</td>
<td>2</td>
</tr>
<tr>
<td>AS1232</td>
<td>19</td>
<td>AS3502</td>
<td>1</td>
</tr>
<tr>
<td>AS20218</td>
<td>18</td>
<td>AS7819</td>
<td>1</td>
</tr>
<tr>
<td>AS4350</td>
<td>10</td>
<td>AS3768</td>
<td>1</td>
</tr>
<tr>
<td>AS48539</td>
<td>9</td>
<td>UNKN</td>
<td>1</td>
</tr>
<tr>
<td>AS14039</td>
<td>6</td>
<td>AS4919</td>
<td>1</td>
</tr>
</tbody>
</table>

Unique # ASN

Location
Resolver Reputation (RESREP)

**Goal:** Try to detect malicious activity by assigning reputation scores to resolvers

**How:** “fingerprinting” resolver behaviour
Malicious activity:
- Spam-runs
- Botnets like Cutwail
- DNS-amplification attacks
RESREP Architecture

1. Verwijder de bestaande foto en klik op het icoon, om een foto in te voegen:
2. Zoek de gewenste foto en dubbelklik hierop.
3. Staat de aleelding er niet goed in? Selecteer de foto, klik 'Format' in het lint en selecteer 'Crop'.
4. De aleelding is nu te verschuiven, door met een linkermuisklik vast te houden op de aleelding en de muis naar de gewenste richting te bewegen.

ISP network

Resolvers

Root operator

.RESREP

service

ENTRADA

Platform

AbuseHUB

Abusedesk

User

Child operator (example.nl)

www.example.nl
ENTRADA Architecture

• ‘DNS big data’ system

• Goal: develop applications and services that further enhance the security and stability of .nl, the DNS, and the Internet at large

• ENTRADA main components
  • Applications and services
  • Platform and data sources
  • Privacy framework
  • Platform + privacy framework = ENTRADA plumbing
ENTRADA Privacy Framework

- Part of the “ENTRADA plumbing”

- Key concepts
  - Application-specific privacy policy
  - Privacy Board
  - Enforcement Points

- Policy elements include
  - Purpose
  - Data used
  - Filters
  - Retention period
  - Type of application (R&D vs. production)
ENTRADA Technical Architecture

ENTRADA-specific components

- Services
- Workflow
- DNS Library
- PCAP Conversion

Open source Hadoop (generic components)

- Support
- IMPALA
- HDFS
- Parquet
Query data available for analysis within 10 minutes
Performance

Example query, count # ipv4 queries per day.

```sql
select
  concat_ws('-', day, month, year),
  count(1)
from dns.queries
where ipv=4
group by
  concat_ws('-', day, month, year)
```

1 Year of data is 2.2TB Parquet ~ 52TB of PCAP
<table>
<thead>
<tr>
<th>ENTRADA Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name server feeds</td>
</tr>
<tr>
<td>Queries per day</td>
</tr>
<tr>
<td>Daily PCAP volume (gzipped)</td>
</tr>
<tr>
<td>Daily Parquet volume</td>
</tr>
<tr>
<td>Months operational</td>
</tr>
<tr>
<td>Total # queries stored</td>
</tr>
<tr>
<td>Total Parquet volume</td>
</tr>
<tr>
<td>HDFS (3x replication)</td>
</tr>
<tr>
<td>Cluster capacity</td>
</tr>
</tbody>
</table>
Conclusions

Technical:
• Hadoop HDFS + Parquet + Impala is a winning combination!

Contributions:
• Research by SIDN Labs and universities
• Identified malicious domain names and botnets
• External data feed to the Abuse Information Exchange
• Insight into DNS query data
Future Work

• Combine data from .nl authoritative name server with scans of the complete .nl zone and ISP data.

• Get data from more name servers and resolvers

• Expand Open Data program
Questions and Feedback

Maarten Wullink
Senior Research Engineer
maarten.wullink@sidn.nl
@wulliak

www.sidnlabs.nl

https://stats.sidnlabs.nl