ENTRADA: Enabling DNS Big Data Applications

Moritz Müller | SDNRG @ IETF97 | Soul, South Korea

2016-11-14
What if...

You have 100 TB or more of pcap data?

You want to:
1. Store it efficiently
2. Query it efficiently (interactive response times)
3. Test a large number of hypothesis on your data
4. Continuously keep adding new data
You could...

1. Convert it to text format like csv
2. Hadoop MapReduce jobs on csv/pcap
3. Store it in a RDBMS
4. ...

With most options it will be hard to scale and deliver interactive response times
What to do?

• Build your own data streaming warehouse (DSW)

• ENTRADA is our open source DSW (entrada.sidnlabs.nl)

• Analyze 50TB pcap data equiv in under 3.5 minutes with a small 4 node cluster

• Our use case: network (DNS, TCP/IP, ICMP) analytics
  • But extensible to other protocols
ENTRADA

ENhanced Top-Level Domain Resilience through Advanced Data Analysis

[Diagram showing ENTRADA Platform with ENTRADA Applications and Services, IMPALA, HDFS, Parquet, Workflow Manager, DNS Library, PCAP Converter, Name Servers, and Support Libraries. The diagram indicates Generic components and ENTRADA-specific components.]
ENTRADA@SIDN

• We are a TLD registry

• Use it to increase security and stability

• Operational for 2 years

• Capturing data for .nl name servers

• 150 Billion rows (DNS query+response pairs), 21 TB of data
Use Cases
Focus on increasing the security and stability of the DNS

• Statistics (stats.sidnlabs.nl)
• Scientific research
• Support for DNS operators
• Malicious domain detection
• Botnet infection detection
**Observation:** Phishing domains have unique query patterns

**nDEWS:** a New Domains Early Warning System for TLDs  G. Moura / M. Müller / M. Wullink / C. Hesselman. IEEE/IFIP International Workshop on Analytics for Network and Service Management (AnNet 2016)
nDEWS Architecture

Every day workflow

Newly Registered Domains

Registry DB

Get Query Characteristics

ENTRADA

Cluster Domains

Legit Domains

Suspicious Domains

Share with Registrar

Σ PReq: popularity
Σ PIPs: resolver diversity
Σ PCC: country diversity
Σ PASes: AS diversity
Detecting botnet infections 1/3

Malicious activity:
- Spam-runs
- Botnets
- DNS-amplification attacks
Detecting botnet infections 2/3

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 afbeelding er niet goed in? Selecteer de foto, klik 'Format' in het lint en selecteer 'Crop'.
4. De afbeelding is nu te verschuiven, door met een linkermuisklik vast te houden op de afbeelding en de muis naar de gewenste richting te bewegen.
Detecting botnet infections 3/3
Summary

• We have shown ENTRADA, a DSW built using open-source big data tools

• It enables quick hypothesis testing and application development using SQL.

• We have shown some example use cases, which can be easily extended

• Download it and contribute to it.
Questions?

Moritz Müller
Research Engineer
moritz.muller@sidn.nl
@dhr_moe
www.sidnlabs.nl

TEKST NIVEAUS

1. Eerste bullets (20 pt.)
2. Tweede bullets (20 pt.)
3. Derde bullets (20 pt.)

Niveau omhoog
Niveau omlaag

Start

Uitgebreide tekst (20 pt.)

Kop 1 (24 pt.)

ENTRADA
An open source platform for network data analytics

Get Started

Performance
Analyze the Parquet data equivalent of about 56 terabytes of pcap data in under 3.5 minutes with a small 4 data-node cluster. Read the performance evaluation in our research paper.

Analytics
Use an efficient columnar data format with a massively parallel SQL query engine for low latency and high concurrency analytic queries.

Query Language
Query your data using the SQL-92 standard and standardized interfaces for Java and Python, which makes it easy for anyone to start analyzing network data.

Data model
Using a data model designed for DNS, TCP, UDP and ICMP network data enabling fast analytics with precomputing, enrichment and pre-joining of request and response packets.

Monitoring
View real-time process and network data metrics with Graphite and Grafana dashboards. Visualizing ENTRADA processes and ingested network data.

Storage
Automatic conversion to a columnar data format with efficient compression and encoding schemes is used to optimize the data volume and query performance.

Workflow
Automating all the steps required to insert captured network data into the ENTRADA database. Spend 100% of your time on data analysis.

Support
Availability of multiple support channels for users and developers. Contact us for any questions about support.

Copyright © 2019 ENTRADA

entrada@sidn.nl
entrada.sidnlabs.nl

Moritz Müller
Research Engineer
moritz.muller@sidn.nl
@dhr_moe
www.sidnlabs.nl
Future Work

• More DNS research in collaboration with research partners

• Develop more data-driven applications and services based on ENTRADA

• Build an active ENTRADA users community
Workflow

Query data available for analysis within 10 minutes

name server → PCAP staging → PCAP decode → Join → Filter → Enrich → Metrics → Import

PCAP staging

Monitoring

Hadoop

Impala

Parquet

Analyst

Workflow

PCAP staging

PCAP decode

Join

Filter

Enrich

Metrics

Import
Performance

Example: count # daily ipv4 queries.

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

Query response time

1 Year of data is 2.2TB Parquet ~ 52TB of PCAP
E-mail security 1/3

- What is the usage of DMARC/DKIM?
  - Count standardized labels, see RFC 6376 and RFC 7489

Select country,count(1)
from dns.queries
where qtype =16
and (qname like '%_domainkey.%'
or qname like '_dmarc .%')
and rcode=0
and ((year=2014 and month>6) or
year=2015)
and server='ns1.dns.nl'
group by country

Use standard SQL for analysis
E-mail security 2/3

<table>
<thead>
<tr>
<th>Country</th>
<th># Queries</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>208,533,790</td>
<td>42.60</td>
</tr>
<tr>
<td>IE</td>
<td>84,515,235</td>
<td>17.26</td>
</tr>
<tr>
<td>NL</td>
<td>79,052,717</td>
<td>16.15</td>
</tr>
<tr>
<td>BE</td>
<td>67,963,161</td>
<td>13.88</td>
</tr>
<tr>
<td>FI</td>
<td>9,112,053</td>
<td>1.86</td>
</tr>
<tr>
<td>RU</td>
<td>7,306,873</td>
<td>1.49</td>
</tr>
<tr>
<td>DE</td>
<td>7,119,556</td>
<td>1.45</td>
</tr>
<tr>
<td>GB</td>
<td>5,897,734</td>
<td>1.20</td>
</tr>
<tr>
<td>CN</td>
<td>5,446,895</td>
<td>1.11</td>
</tr>
<tr>
<td>DK</td>
<td>2,958,891</td>
<td>0.60</td>
</tr>
</tbody>
</table>

89.9% of queries originate from top 4 countries.
### E-mail security 3/3

<table>
<thead>
<tr>
<th>Provider</th>
<th>ASN</th>
<th># Queries</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>AS15169</td>
<td>302,465,578</td>
<td>61.79</td>
</tr>
<tr>
<td>Microsoft</td>
<td>AS8075</td>
<td>51,556,416</td>
<td>10.53</td>
</tr>
<tr>
<td>Unknown</td>
<td>UNKN</td>
<td>15,788,699</td>
<td>3.22</td>
</tr>
<tr>
<td>AOL</td>
<td>AS1668</td>
<td>12,971,456</td>
<td>2.65</td>
</tr>
<tr>
<td>Yahoo</td>
<td>AS36647</td>
<td>11,83,129</td>
<td>2.30</td>
</tr>
<tr>
<td>Yahoo</td>
<td>AS26101</td>
<td>10,24,857</td>
<td>2.07</td>
</tr>
<tr>
<td>Yahoo</td>
<td>AS36646</td>
<td>9,150,523</td>
<td>1.87</td>
</tr>
<tr>
<td>Yahoo</td>
<td>AS34010</td>
<td>4,522,388</td>
<td>0.92</td>
</tr>
<tr>
<td>IDC China Tel</td>
<td>AS23724</td>
<td>4,520,819</td>
<td>0.92</td>
</tr>
<tr>
<td>Mail.ru</td>
<td>AS47764</td>
<td>3,659,097</td>
<td>0.75</td>
</tr>
</tbody>
</table>

82.13% of queries originate from large e-mail providers.
Privacy Framework

Policy elements:
- Purpose
- Data that is used
- Filters on the data
- Retention period
- Access to the data
- Type of application (Research vs. Production)