ENTRADA: Enabling DNS Big Data Applications

Giovane C. M. Moura - SIDN | NMRG @ IETF96
Berlin, Germany
What if...

You have many TB’s of network data?

And you want to:
1. Store it efficiently
2. Query it efficiently (SQL with interactive response times)
3. Quickly test a large number of hypotheses on your data
4. Continuously keep adding new data
You could...

1. Convert pcap to text format like csv and use Linux utilities
2. Run 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 stream warehouse (DSW)
- ENTRADA is our open source Hadoop-based DSW ([entrada.sidnlabs.nl](http://entrada.sidnlabs.nl))
- Analyze 50TB of converted pcap data in under 3.5 minutes using a small cluster
- Our main use case: network (DNS, TCP/IP, ICMP) analytics
ENTRADA

ENhanced Top-Level Domain Resilience through Advanced Data Analysis
ENTRADA@SIDN

• We are the TLD registry of the Netherlands (.nl)

• Use ENTRADA to further increase security and stability

• Operational for over 2 years

• Capturing data from .nl name servers

• 160 billion rows (DNS query+response tuple), 21 TB of data
More ENTRADA details

For design choices and a performance evaluation, see our 2016 NOMS paper:


See: https://www.sidnlabs.nl/publications
Example Use Cases

• Statistics (stats.sidnlabs.nl)
• Scientific research
• Insight for DNS operators
• Malicious domain detection
• Botnet client detection
• Measuring uptake of email security
Malicious Domain Detection (1/2)

**Observation:** New phishing domains have distinct query patterns

---

Malicious Domain Detection (2/2)

*Every day workflow*

- **Newly Registered Domains**
- **Domain Characteristics**
- **Cluster Domains**

**Registry DB**

**ENTRADA**

- $\Sigma$ PReq: popularity
- $\Sigma$ PIPs: resolver diversity
- $\Sigma$ PCC: country diversity
- $\Sigma$ PASes: AS diversity

**Flowchart**

- **Normal**
- **Suspicious**

**Notify Registrar**
Botnet Client Detection (1/2)
Botnet Client Detection (2/2)

- **Graph:**
  - Title: DNS queries from fingerprinted botnet
  - X-axis: Date (June 2015 to January 2016)
  - Y-axis: Weekly Queries
  - Data points show spikes in DNS queries in October and December 2015.

- **Map:**
  - Title: Botnet activity by country
  - Color scale indicates percentage of botnet activity, ranging from 0 to 10.
  - Dark areas represent higher botnet activity in regions such as Africa and Asia.

---

*Source: SDN Labs*
Uptake of DKIM/DMARC (1/3)

- Email security standards DKIM (RFC 6376) and DMARC (RFC 7489)

- Approach: count standardized labels

```sql
WHERE is DKIM/DMARC used most?

select country, count(1) as total
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)
group by country
```

Use standard SQL for analysis
Uptake of DKIM/DMARC (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><strong>42.60</strong></td>
</tr>
<tr>
<td>IE</td>
<td>84,515,235</td>
<td><strong>17.26</strong></td>
</tr>
<tr>
<td>NL</td>
<td>79,052,717</td>
<td><strong>16.15</strong></td>
</tr>
<tr>
<td>BE</td>
<td>67,963,161</td>
<td><strong>13.88</strong></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.
### Uptake of DKIM/DMARC (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>11,831,129</td>
<td>2.30</td>
</tr>
<tr>
<td>Yahoo</td>
<td>AS26101</td>
<td>10,248,57</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 4 large e-mail providers.
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 real world example use cases

• ENTRADA can be extended to other use cases

• Download and contribute!
Future Work

• More DNS research in collaboration with research partners

• Develop data-driven applications and services based on ENTRADA

• Facilitate ENTRADA user community