ENTRADA: Enabling DNS Big Data Applications

Maarten Wullink - SIDN | APWG eCrime 2016

June 2nd 2016 - Toronto
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)

• 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
ENHANCED TOP-LEVEL DOMAIN RESILIENCE THROUGH ADVANCED DATA ANALYSIS

ENTRADA Applications and Services

Support Libraries

IMPALA

HDFS

Parquet

Workflow Manager

DNS Library

PCAP Converter

ENTRADA Platform

Name Servers

Generics components

ENTRADA-specific components
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/publicaties
Example Use Cases

- Statistics ([stats.sidnlabs.nl](http://stats.sidnlabs.nl))
- Scientific research
- Insight for DNS operators
- **Malicious domain detection**
- **Botnet client detection**
- Measuring uptake of email security
Observation: New phishing domains have distinct query patterns

Malicious Domain Detection (2/2)

Every day workflow

Newly Registered Domains

Registry DB

Domain Characteristics

ENTRADA

Cluster Domains

Normal

Suspicious

Notify Registrar

Σ PReq: popularity
Σ PIPs: resolver diversity
Σ PCC: country diversity
Σ PASes: AS diversity
Botnet Client Detection (1/2)

1. Verwijder de bestaande foto en klik op het icoon, om een foto in te voegen:
2. Zoek de gewenste foto en dubbelklik hierop.
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.
Botnet Client Detection (2/2)

DNS queries from fingerprinted botnet

Botnet activity by country

Weekly Queries

Date

Percentage
0, 2.5, 5, 7.5, 10

10^7
Uptake of DKIM/DMARC (1/3)

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

• Approach: count standardized labels

<table>
<thead>
<tr>
<th>Where is DKIM/DMARC used most?</th>
</tr>
</thead>
<tbody>
<tr>
<td>select country,count(1) as total</td>
</tr>
<tr>
<td>from dns.queries</td>
</tr>
<tr>
<td>where qtype=16</td>
</tr>
<tr>
<td>and (qname like &quot;%_domainkey.%&quot;</td>
</tr>
<tr>
<td>or qname like &quot;_dmarc.%&quot;)</td>
</tr>
<tr>
<td>and rcode=0</td>
</tr>
<tr>
<td>and ((year=2014 and month&gt;6) or</td>
</tr>
<tr>
<td>year=2015)</td>
</tr>
<tr>
<td>group by country</td>
</tr>
</tbody>
</table>

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>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
### 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>11,83,129</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 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
Questions?

Maarten Wullink
Sr. Research Engineer

maarten.wullink@sidn.nl
@wulliak
www.sidnlabs.nl

Ga naar de tab 'Start'.
Hier vind je 2 knopjes zoals onderstaande afbeelding.
D.m.v. deze knopjes kun je eenvoudig tussen tekstniveaus wisselen.

ENTRADA
An open source platform for network data analytics

Performance
Analyze the Parquet data equivalent of about 50 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/SIP/TCP/UDP/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 ENTRAÐA processes and ingested network data.

Storage
Automatic conversion to a columnar data format with efficient compressors 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 ENTRAÐA 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.

Questions?

Maarten Wullink
Sr. Research Engineer

maarten.wullink@sidn.nl
@wulliak
www.sidnlabs.nl

Ga naar de tab 'Start'.
Hier vind je 2 knopjes zoals onderstaande afbeelding.
D.m.v. deze knopjes kun je eenvoudig tussen tekstniveaus wisselen.

ENTRADA
An open source platform for network data analytics

Performance
Analyze the Parquet data equivalent of about 50 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/SIP/TCP/UDP/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 ENTRAÐA processes and ingested network data.

Storage
Automatic conversion to a columnar data format with efficient compressors 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 ENTRAÐA 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.

Questions?

Maarten Wullink
Sr. Research Engineer

maarten.wullink@sidn.nl
@wulliak
www.sidnlabs.nl

Ga naar de tab 'Start'.
Hier vind je 2 knopjes zoals onderstaande afbeelding.
D.m.v. deze knopjes kun je eenvoudig tussen tekstniveaus wisselen.

ENTRADA
An open source platform for network data analytics

Performance
Analyze the Parquet data equivalent of about 50 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/SIP/TCP/UDP/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 ENTRAÐA processes and ingested network data.

Storage
Automatic conversion to a columnar data format with efficient compressors 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 ENTRAÐA 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.

Questions?

Maarten Wullink
Sr. Research Engineer

maarten.wullink@sidn.nl
@wulliak
www.sidnlabs.nl

Ga naar de tab 'Start'.
Hier vind je 2 knopjes zoals onderstaande afbeelding.
D.m.v. deze knopjes kun je eenvoudig tussen tekstniveaus wisselen.

ENTRADA
An open source platform for network data analytics

Performance
Analyze the Parquet data equivalent of about 50 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/SIP/TCP/UDP/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 ENTRAÐA processes and ingested network data.

Storage
Automatic conversion to a columnar data format with efficient compressors 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 ENTRAÐA 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.

Questions?

Maarten Wullink
Sr. Research Engineer

maarten.wullink@sidn.nl
@wulliak
www.sidnlabs.nl

Ga naar de tab 'Start'.
Hier vind je 2 knopjes zoals onderstaande afbeelding.
D.m.v. deze knopjes kun je eenvoudig tussen tekstniveaus wisselen.

ENTRADA
An open source platform for network data analytics

Performance
Analyze the Parquet data equivalent of about 50 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/SIP/TCP/UDP/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 ENTRAÐA processes and ingested network data.

Storage
Automatic conversion to a columnar data format with efficient compressors 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 ENTRAÐA 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.

Questions?

Maarten Wullink
Sr. Research Engineer

maarten.wullink@sidn.nl
@wulliak
www.sidnlabs.nl

Ga naar de tab 'Start'.
Hier vind je 2 knopjes zoals onderstaande afbeelding.
D.m.v. deze knopjes kun je eenvoudig tussen tekstniveaus wisselen.