ENTRADA: An *open source* platform for network data analysis

Moritz Müller | Netnod Spring Meeting 2016-04-17
SIDN

• Domain name registry for .nl ccTLD

• > 5.6 million domain names

• 2.5 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 queries 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

• **Problem**: Existing solutions do not work well with large datasets and have limited analytical capabilities.
Requirements

• SQL support
• Scalability
• High performance
• Capacity for >1 year of DNS data
• Extensibility
• Stability
• Don’t spend too much money!
ENTRADA Architecture

Main components

• Data sources
• Platform
• Applications and services
• Privacy framework
Query Engine Options

Engines galore!

Evaluated SQL and NoSQL solutions

• Relational SQL (PostgreSQL)
• MongoDB
• Cassandra
• Elasticsearch
• Hadoop (HBASE + Apache Phoenix or Hive)

→ SQL on Hadoop (Impala + Parquet +HDFS)
SQL on Hadoop

Best fit for our requirements
Impala query engine

- MPP (massively parallel processing)
- Low latency and high concurrency for BI/analytic queries on Hadoop
- Excellent performance compared to other Hadoop based query engines

![Single-User Response Time](chart.png)

<table>
<thead>
<tr>
<th>Category</th>
<th>Impala 1.4</th>
<th>Spark SQL</th>
<th>Presto</th>
<th>Tez</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive</td>
<td>6</td>
<td>4.6x</td>
<td>6.5x</td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>18</td>
<td>2.1x</td>
<td>13.0x</td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td>30</td>
<td>3.8x</td>
<td>6.3x</td>
<td>7.2x</td>
</tr>
</tbody>
</table>
## Impala (2)

<table>
<thead>
<tr>
<th>Data formats</th>
<th>Interfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Text</td>
<td>• Web-based GUI</td>
</tr>
<tr>
<td>• Hadoop formats</td>
<td>• Command line (impala-shell)</td>
</tr>
<tr>
<td>• Apache Avro</td>
<td>• Python (Impyla)</td>
</tr>
<tr>
<td>• Apache Parquet</td>
<td>• JDBC</td>
</tr>
</tbody>
</table>
Apache Parquet

- Why not just use the PCAP files?
  - Reading (compressed) PCAP data is just too slow
  - Analytical engines cannot read PCAP files
HDFS

- Distributed file system for storing large volumes of data
- High availability through replication of data blocks
- Scalable to hundreds of PB’s and thousands of servers

HDFS Data Distribution

```
Input File

1 2 3 4 5
```

```
Node A

2 1 4 2 5

Node B

1 2 3 3 4

Node C

1 1 3 4 5

Node D

2 3 3 5 4

Node E

1 1 3 4 5
```
Cluster Design

nano sized

location I
management node

location II
data nodes

location III
data nodes

2Gb/s network
Workflow

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

Query response times
Use Cases
Focussed on increasing the security and stability of .nl

- Visualize DNS patterns
- Statistics (stats.sidnlabs.nl)
- Scientific research
- Support for operators
- Real-time Phishing detection
- Detect botnet infections
Use Cases
Focussed on increasing the security and stability of .nl

- Visualize DNS patterns
- Statistics (stats.sidnlabs.nl)
- Scientific research
- Support for operators
- **Real-time Phishing detection**
- Detect botnet infections
Malicious Domain Detection with nDEWS

**Observation:** Phishing domains have unique query patterns
Every day

nDEWS Architecture

Newly Registered Domains

Registry DB

Get Query Characteristics

ENTRADA

Cluster Domains

Suspicious Domains

Share with Registrar

Legit Domains
Resolver Reputation (RESREP)

**Goal:** Detect malicious activity by assigning reputation scores to resolvers

**How:** “fingerprinting” resolver behaviour
RESREP Concept

Malicious activity:
- Spam-runs
- Botnets
- DNS-amplification attacks
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.

RESREP Architecture
Conclusions

Technical:
- Hadoop HDFS + Parquet + Impala is a winning combination!
- Running since almost 2 years
- > 115 billion queries stored

Contributions:
- Research by SIDN Labs and universities
- Identified malicious domain names and botnets
- Insight into DNS queries, shared at stats.sidnlabs.nl
It’s open source!

• Since January 2016

• Project site:
  entrada.sidnlabs.nl

• GitHub:
  github.com/SIDN/ENTRADA
Questions? Feedback?

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

www.sidnlabs.nl

entrada.sidnlabs.nl