ENTRADA:
An *Open Source* Platform for Network Data Analysis

Moritz Müller | 2nd YETI DNS Workshop
2016-11-12 in Seoul, South Korea
SIDN

• Domain name registry for .nl ccTLD

• > 5.6 million domain names

• 2.5 million domain names secured with DNSSEC

• SIDN Labs is the research 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!
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

Hadoop Node N
- IMPALA
  - PARQUET
  - HDFS

Hadoop Node N+1
- IMPALA
  - PARQUET

Hadoop Node N+2
- IMPALA
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
## 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
ENTRADA Architecture

Main components

• Data sources
• Platform
• Applications and services
• Privacy framework
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

- Name server
- PCAP staging
- PCAP decode
  - Join
  - Filter
  - Enrich
  - Import
- Hadoop
  - Impala
  - Parquet
- Analyst
- Monitoring
- Metrics
Performance

Example query, count # ipv4 queries per day.

Select day, month, year, count(1)
from dns.queries
where ipv=4
group by day, month, year

1 Year of data is 2.2TB Parquet ~ 52TB of PCAP
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

nDEWS Architecture

Every day workflow

Newly Registered Domains
Registry DB
ς PReq: popularity
ς PIPs: resolver diversity
ς PCC: country diversity
ς PASes: AS diversity

Get Query Characteristics
ENTRADA

Cluster Domains

Legit Domains

Suspicious Domains

Share with Registrar
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
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 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.
Conclusions and Future Work

• Hadoop HDFS + Parquet + Impala is a winning combination!
• Running for over 2 years
• > 150 billion queries stored

• Develop more use cases
  • DNS Abuse
  • Operational support
• Increase the number of ENTRADA users
It’s open source!

- Since January 2016

- Project site: entrada.sidnlabs.nl

- GitHub: github.com/SIDN/ENTRADA/

- 6 registries are using it
Questions? Feedback?

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

www.sidnlabs.nl

entrada.sidnlabs.nl