DNS Big Data Analytics

DNS-OARC Fall 2015 Workshop
October 4th 2015
Maarten Wullink, SIDN
SIDN

• Domain name registry for .nl ccTLD

• > 5.6 million domain names

• 2.45 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 query's 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 for analyzing network data do not work well with large datasets and have limited analytical capabilities.
- **Main requirement**: high-performance, near real-time data warehouse
- **Approach**: avoid expensive pcap analysis:
  - Convert pcap data to a performance-optimized format (key)
  - Perform analysis with tools/engines that leverage that
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 (HDFS + Impala + Parquet)
SQL on Hadoop

Best fit for our requirements

• Eerste niveau
• Tweede niveau

1
2
3

100%
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

Node A: 2 1 1 2 5
Node B: 4 2 3 3 5
Node C: 2 2 4 5 1
Node D: 1 3 3 4 1
Node E: 5 2 5 4 3
```

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
Impala query engine

- **MPP** (massively parallel processing)
- Inspired by Google Dremel paper
- Provides low latency and high concurrency for BI/analytic queries on Hadoop
- Excellent performance when compared to other Hadoop based query engines.

![Single-User Response Time/Impala Times Faster Than](chart.png)
Impala (2)

**Data formats**
- Text
- Hadoop formats
- Apache Avro
- Apache Parquet

**Interfaces**
- Web-based GUI
- Command line (impala-shell)
- Python (Impyla)
- JDBC
Apache Parquet

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

- Columnar storage format
Apache Parquet (2)

- Columnar storage allows for efficient encoding/compression
  - multiple encoding schemes
  - support for Snappy compression

- Partition data (e.g. by year, month, day and server)
  - Partition pruning allows Impala to skip data we are not interested in

- Other analytical engines such as Apache Spark can use the same Parquet data.
ENTRADA Architecture

• ‘DNS big data’ system

• Goal: develop applications and services that further enhance the security and stability of .nl, the DNS, and the Internet at large

• ENTRADA main components
  • Applications and services
  • Platform
  • Data sources
  • Privacy framework
ENTRADA 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)

Download paper: http://goo.gl/GvsfzQ
Cluster Design

nano sized

location I
management node

location II
data nodes

location III
data nodes

2Gb/s network
Hardware

Management node
HP ProLiant DL380
Xeon 1.9 GHz 12 core CPU
64GB RAM
3 TB storage

Data node
HP ProLiant DL380
Xeon 1.9 GHz 12 core CPU
64GB RAM
6 TB storage

Scaling
• Vertical by adding more resources
• Horizontal by adding more data nodes
**Workflow**

- **Name server**
  - PCAP staging
  - PCAP decode
  - Join
  - Filter
  - Enrich
  - Metrics
  - Import

**Hadoop**
- Impala
- Parquet

**Monitoring**

*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
<table>
<thead>
<tr>
<th>ENTRADA Status</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name server feeds</td>
<td>1</td>
</tr>
<tr>
<td>Queries per day</td>
<td>~150M</td>
</tr>
<tr>
<td>Daily PCAP volume (gzipped)</td>
<td>~33GB</td>
</tr>
<tr>
<td>Daily Parquet volume</td>
<td>~6GB</td>
</tr>
<tr>
<td>Months operational</td>
<td>18</td>
</tr>
<tr>
<td>Total # queries stored</td>
<td>&gt; 71B</td>
</tr>
<tr>
<td>Total Parquet volume</td>
<td>&gt; 3TB</td>
</tr>
<tr>
<td>HDFS (3x replication)</td>
<td>&gt; 9TB</td>
</tr>
<tr>
<td>Cluster capacity</td>
<td>~150B-200B tuples</td>
</tr>
</tbody>
</table>
Use Cases
Focussed on increasing the security and stability of .nl

- Visualize DNS patterns (visualize traffic patterns for phishing domain names)
- Detect botnet infections
- Real-time Phishing detection
- Statistics (stats.sidnlabs.nl)
- Scientific research (collaboration with Dutch Universities)
- Operational support for DNS operators
Example Applications

- DNS security scoreboard
- Resolver reputation
DNS Security Dcoreboard

**Goal:** Visualize DNS patterns for malicious activity

**How:** Combine external phishing feeds with DNS data
Architecture

Security feed I

Security feed II

new event

Event Analyzer

new event

enrich with DNS data

enriched event

Hadoop

PostgreSQL

REST API

Web UI

save enriched event

retrieve event data
Traffic Visualization

Overview

Network

Top 10 event  average

<table>
<thead>
<tr>
<th>ASN</th>
<th>#</th>
<th>ASN</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS15169</td>
<td>56</td>
<td>AS15169</td>
<td>10</td>
</tr>
<tr>
<td>AS383406</td>
<td>38</td>
<td>AS8737</td>
<td>3</td>
</tr>
<tr>
<td>AS202109</td>
<td>22</td>
<td>AS31334</td>
<td>2</td>
</tr>
<tr>
<td>AS12352</td>
<td>19</td>
<td>AS38022</td>
<td>1</td>
</tr>
<tr>
<td>AS202118</td>
<td>18</td>
<td>AS7819</td>
<td>1</td>
</tr>
<tr>
<td>AS43350</td>
<td>10</td>
<td>AS3786</td>
<td>1</td>
</tr>
<tr>
<td>AS48859</td>
<td>9</td>
<td>UNKN</td>
<td>1</td>
</tr>
<tr>
<td>AS14009</td>
<td>9</td>
<td>AS69X9</td>
<td>1</td>
</tr>
</tbody>
</table>

Location

![Map of location with color-coded areas](image)
Resolver Reputation (RESREP)

**Goal:** Try to detect malicious activity by assigning reputation scores to resolvers

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

Malicious activity:
- Spam-runs
- Botnets like Cutwail
- 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.


4. De azeelding is nu te verschuiven, door met een linkermuisklik vast te houden op de azeelding en de muis naar de gewenste richting te bewegen.
Conclusions

Technical:
• Hadoop HDFS + Parquet + Impala is a winning combination!

Contributions:
• Research by SIDN Labs and universities
• Identified malicious domain names and botnets
• External data feed to the Abuse Information Exchange
• Insight into DNS query data
Future Work

• Combine data from .nl authoritative name server with scans of the complete .nl zone and ISP data.

• Get data from more name servers and resolvers

• Expand Open Data program
Questions and Feedback

Maarten Wullink
Senior Research Engineer
maarten.wullink@sidn.nl
@wulliak

www.sidnlabs.nl

https://stats.sidnlabs.nl