Why Analyze Passive DNS Data And Best Practices To Collect And Process It On Large Networks Abstract

Vinicios Barretos*, Adriano Cansian*, Hugo Koji Kobayashi**
Universidade Estadual Paulista*, NIC.BR – Brazilian Network Information Center**
{vinicios, adriano}@acmesecurity.org*, koji@registro.br**

1. Introduction

In recent years, millions of Internet users have suffered the effects of large-scale cyber-attacks [1]. Many of these attacks were based on the use of DNS as a vector. An attacker can use domain names and multiple IP addresses to orchestrate Phishing, Spam, Distributed Denial of Service (DDoS) attacks and even control botnets through command-and-control (C&C) servers. These actions, in addition to leading to several losses for users and companies, also affect the reputation of the TLD in which the domain was registered.

An efficient way to detect abuses in the domain registration is through the analysis of Passive DNS traffic, which is able to detect Botnets, C&C servers, Fast Flux Networks and Phishing campaigns [3].

Passive DNS analysis has a number of advantages when compared to Active DNS or the use of blacklists. That's because Passive DNS allows you to identify threats with less time and to model more complex attacks [2].

2. Motivation

To develop this project we had the partnership of NIC.br, responsible for the Top Level Domain .br. Our goal was to create a framework to collect, process and analyze DNS traffic to identify malicious domains.

3. Passive DNS and Covid-19

In Brazil, since the government announced monetary aid to the population through an online platform of a government bank, many malicious domain names have been registered in an attempt to appear as the official government website. Many of these cases can be identified quickly through the analysis of the Passive DNS, preventing more people from being harmed.

4. Challenges

Although Passive DNS is very effective for detecting threats, a powerful infrastructure is needed to handle the ingestion and processing of pcaps, especially when applied to a large-scale network.

Working with raw pcaps is not efficient, so the cleaning and pre-processing process is fundamental for Passive DNS. This process must have high performance, scalability, usability, security and also provide support for privacy filters.

For example, in a TLD DNS server infrastructure, it is common to collect dozens of TBs per month. Over time the amount of information becomes so large that traditional databases would not be able to support it and machine learning algorithms would take exorbitant time to run.

5. Architecture

After collecting the pcaps, the data is inserted into instances of the ENTRADA [4], where they go through the cleaning, enrichment and compaction process and then be inserted into a Hadoop Cluster.

In each instance of ENTRADA, each query is combined with the response. The information from the IP, UDP or TCP and DNS headers are stored and then go through the enrichment process, where GeoIP information and the respective ASN are inserted. Finally, data is partitioned by source (authoritative server) and by year, month and day.

After this data processing, files are generated in a compressed format called Parquet, ensuring high performance and less use of storage [4].

6. Results

The generated Parquet files had a reduction of up to 93% in storage when compared to the raw pcaps [4]. Due to this reduction, it is possible to perform complex queries with interactive time, enabling the identification of threats with a low response time. Statistics are obtained using Apache Hive or Apache Spark through SQL queries.

The information obtained is ready to be used with machine learning algorithms, providing integration with data frames format.

During the pcaps conversion process, several statistics are generated and can be viewed in real time by Grafana, providing a report on the use of the servers.

7. Expectations

All the processes used for this work have been well documented and simplified with the use of Docker containers. With this presentation we intend to provide our knowledge base, the scripts for deploying ENTRADA, Hadoop, Grafana and discuss good practices for operating the Passive DNS collection and processing structure.

In other words, through this presentation we intend to encourage the use of Passive DNS and deliver an initial guide to carry out the implementation in a simple way.

References

[1] L. Bilge, S. Sen, D. Balzarotti, E. Kirda, and C. Kruegel. Exposure. *ACM Transactions on Information and System Security*, 16(4):1–28, 2014.

[2] P. Lison and V. Mavroeidis. Neural reputation models learned from passive dns data. *2017 IEEE International Conference on Big Data (Big Data)*, 2017.

[3] S. Torabi, A. Boukhtouta, C. Assi, and M. Debbabi. Detecting internet abuse by analyzing passive dns traffic: A survey of implemented systems. *IEEE Communications Surveys Tutorials*, 20(4):3389–3415, 2018.

[4] M. Wullink, G. C. M. Moura, M. Muller, and C. Hesselman. Entrada: A high-performance network traffic data streaming warehouse. *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, 2016.