Developing Network Analytics to Support United States Critical Infrastructure

Ethan Kanyid, Timothy Cain

Overview

Background

The Cybersecurity Risk Information Sharing Program (CRISP) provides a platform for cyber-situational awareness among US energy sector entities to expose threats. CRISP analysts bring this to light in a comprehensive cyber threat landscape of the sector through the bidirectional information sharing fostered by the Department of Energy.

Purpose

CRISP, with a desire to advance the program, engaged a team to develop an analytic process to examine domain name system (DNS) traffic and provide additional insights into potential threats facing the energy sector. The proposed methodology cross-profiles DNS traffic with standard network traffic to identify anomalous activity.

Implications

Analysts can use the DNS profiles with other analytics to create informed results. Analysts use these results to discover indicators of compromise, which are formulated into actionable reports and distributed among the energy sector participants. This helps to promote a secure cyber landscape and robust critical infrastructure.

Tech Stack

Kubernetes -> A containerization software that manages how apps, services, and projects run. This platform automatically scales with demand and allows for smooth operations in varying conditions.

Apache NiFi -> A tool for managing, processing, and directing incoming data. It operates on all the received data and logs from the participating networks and stores it in databases for later use.

MinIO -> A high-performance object storage that operates similarly to Amazon S3 (Simple Storage Service). For CRISP, it stores objects in the parquet format for optimal retrieval and long-term storage.

Trino -> A distributed querying software that runs SQL (structured query language) at large scale. It optimizes performance on large data sets to allow for analysis on big data.

Apache Airflow -> An automating software that smoothly schedules tasks and dependencies. It works with complex data pipelines and workflows to programmatically and procedurally execute jobs.

All logos are unofficially used here for illustrative and identification purposes only. No endorsement or affiliation by the respective trademark owners is implied.

Domain Name Systems (DNS)

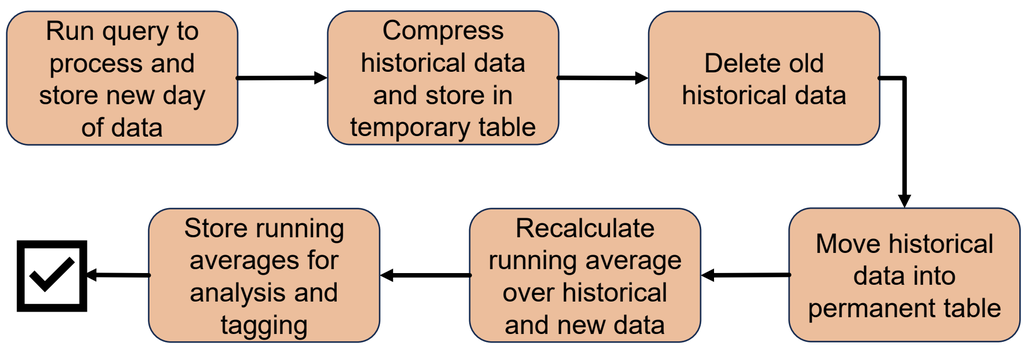
When navigating to a site like www.google.com, how do computers know where to go? They do not without DNS. DNS is like an address book for computers, and it similarly compares names to addresses. This is necessary because the internet does not work on names, but rather numbers, as do all computers. So, www.google.com is translated into an address by a DNS server – specifically, what is called an internet protocol (IP) address. What the CRISP team accomplished is creating a profile of DNS requests to responded IPs. With a few other statistics and measurements, it is possible to analyze this DNS traffic against other internet traffic. Moreover, each conversation with any IP can be fact checked against that IP's profile to indicate legitimate versus malicious traffic.

Results

The analytic process developed on DNS creates many possibilities to further secure the United States Critical Infrastructure. This process has produced these results:

Task	Result
Map every domain name to a specific IP, and if a domain has a known hash (UUID-4 or MD5), truncate it with regular expressions.	Each IP is associated with a known domain, type of service, organization, country, and other statistics calculated on its usage.
Generate statistics and running averages on all previously seen results but only continue storing the results if they have been seen in the last 30 days.	For each IP and corresponding domain, running statistics are regenerated every day if that result was seen in the last 30 days.
Create documented tables within Trino and MinIO that contain the results from the different queries.	Through Trino, an internal table holds the running averages for ease of updating and analysis. An external table is also accessible in MinIO for storing the original, unanalyzed data.
Develop a DAG that runs in Apache Airflow and automates this process.	Utilizing Python and imported libraries from Airflow, a workflow executes these tasks against Trino and MinIO to produce these datasets with fail-safes and logging.

Directed Acyclic Graph (DAG)



Conclusions

The purpose of this project was to develop an analytic process that could be integrated into the CRISP mission and work with the other analytics to provide a better understanding into the activity occurring in the United States energy sector. This work is immediately actionable and available to be incorporated into the CRISP ecosystem where it can begin creating an analysis of IPs and domains. It will be continually providing feedback on the DNS records for analyst use, and in the future, it can be expanded by adding or altering algorithms to provide additional insights. Ideas also may include creating a short-term running analysis on the data to compare against the long-term running data, which allows for calculating a snapshot of the sector. This could be used to identify any changes in trends or variance in the results.

This work was supported by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI)

For additional information, contact:

Ethan Kanyid
(509)-955-3377
Ethan.Kanyid@gmail.com

U.S. DEPARTMENT of ENERGY
BATTELLE
PNNL is operated by Battelle for the U.S. Department of Energy
11/14/2025 | PNNL-SA-213738
Pacific Northwest NATIONAL LABORATORY
www.pnnl.gov