Developing Analytics on Network Traffic to Support United States Critical Infrastructure

Ethan Kanyid, Intern, ethan.kanyid@pnnl.gov

Pacific Northwest National Laboratory, Richland, WA

ABSTRACT. The United States critical infrastructure is the backbone of essential operations that support national security, public health and safety, and the economy. This infrastructure is comprised of systems, assets, and networks that have become increasingly dependent on digital technologies. Though these technologies greatly enhance operations, they also introduce vulnerabilities and security risks. The Cybersecurity Risk Information Sharing Program (CRISP) provides a platform for cyber-situational awareness among US energy sector entities to expose those 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. 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 methodology cross-profiles DNS traffic with standard network traffic to identify anomalous activity. Specifically, structured query language algorithms are run against petabytes of data to create an automated data workflow for further analyst review. The team then refined the process to reduce overhead. As a result, CRISP was able to minimize operations, reduce calculations, and aggregate data to increase speed and accuracy. These informed results help analysts discover indicators of compromise, which are then distributed among the energy sector participants to promote a more secure and robust critical infrastructure.

I. INTRODUCTION

Pacific Northwest National Laboratory, operated by Battelle Memorial Institute for the Department of Energy, has been continually contributing to the Cybersecurity Risk Information Sharing Program (CRISP) for over a decade. The project serves more than 75% of the United States electricity subsector customers and provides informed and actionable reports on the energy sector landscape[1]. Through the extensive scope of this program, CRISP significantly supports both the national security and critical infrastructure of the United States. CRISP analysts and engineers manage the processing and administration of petabytes of CRISP network data over the course of a year. Moreover, complex systems and services ensure the operability of the CRISP infrastructure while enabling the analysis and dissemination of information. This report recounts the production of a new analytic process developed under the CRISP project. This process analyzes domain name system (DNS) data to provide deeper insights into the energy sector landscape, expanding CRISP's DNS analysis in a novel way. This report will detail the background, methodology, and implementation of this process.

II. BACKGROUND

Information from CRISP participants is collected and shared by an information sharing device (ISD) coupled with Zeek, a network monitoring software[2]. Zeek collects many different web protocols including DNS, which is a service to direct web traffic towards appropriate resources. Typically, this involves a query and the response of a corresponding internet protocol (IP) address. Monitoring this information allows traffic to be categorized and processed, which can be used to identify trends and determine whether traffic is legitimate or malicious. Understanding DNS is essential to understanding how it can be used in the CRISP platform to enhance the energy sector landscape.

The CRISP platform manages its ETL (Extract, Transform and Load) pipeline and several open-source services with Kubernetes, a containerization Orchestration software. These services are deployed on Kubernetes for scalability and elastic application deployment. CRISP also

leverages GitLab for its CI/CD (continuous integration and continuous deployment) pipeline to seamlessly deploy this infrastructure with Kubernetes. GitLab is also useful for versioning and hosting the code for CRISP's operations.

The data CRISP collects from its participants is securely transported to the CRISP servers where it enters the system through Apache NiFi. NiFi routes data through decoration software to geotag and add additional information for downstream analytics. After preprocessing, NiFi converts the data into parquet file format and stores it in MinIO, a high-performance object storage solution that operates similarly to Amazon S3 (Simple Storage Service). CRISP uses Trino to run structured query language (SQL) algorithms against data in MinIO. These processes require automation to be able to operate at scale, which CRISP addresses with Apache Airflow. Airflow is a software that streamlines task scheduling and interdependent workflows. It works with complex data pipelines, allowing the analytic process to run daily on the newly ingested data. The aggregated data from the analytic process is then ingested to Elasticsearch, where analysts can generate actionable and insightful reports, which are then distributed to the participants.

III. IMPLEMENTATION

On the CRISP development network, the team iterated through several potential queries to develop a comprehensive analytic that creates a running profile of DNS activity for each of the CRISP participants. The query began with manual interaction with the data, but it eventually progressed to be an automated task in Airflow.

In the initial steps of development, research focused on general statistics about the dataset rather than the endpoints involved in a conversation. Statistics revealed how often an IP address appeared within specific hourly or daily intervals. In addition, different measurements counted which and how many IP addresses were interfacing with local vs external DNS servers. Further examination brought in how many bytes and packets were sent by different IP addresses. This resulted in the calculation of several benchmarks: the mean, standard deviation, z-score, coefficient of variation, and skewness. Each metric served a distinct role in understanding the dataset. The mean offered an overall insight into the

average value of each column, serving as a central point of reference. The standard deviation highlighted the degree of variability, indicating how spread out the values were around the mean. The z-score provided a standardized, quantifiable measurement of how far a specific data point deviated from the mean – allowing for comparisons across different points in the dataset. The coefficient of variation offered a perspective on relative variability by comparing the standard deviation to the mean, which made it useful for assessing consistency across datasets with different scales. Finally, skewness portrayed the symmetry of the distribution. It clarified the relationship between the mean and the median; a skew near zero suggested a more symmetrical (or normal) distribution, while a skew diverged from zero indicated potential outliers or asymmetry in the data. Although these metrics offered valuable insights, they were excluded from the final query due to the highly unpredictable nature of DNS traffic.

The final query incorporates similar but refined measurements that are calculated based on grouped data. The data is grouped by DNS conversations, ISD device, organization, country, and other network information so each record is unique to the CRISP participant network. This query runs daily on the previous full day of Zeek DNS data. It joins this data with the monitored traffic where the IP addresses match to create a running profile of all observed DNS requests and their corresponding traffic.

One challenge in this process was the effective management of potentially unmatched records. This arises when multiple answers are returned for the same request, indicating alternate paths to the same resource. However, not all of these answers are utilized, leading to unmatched records where no corresponding traffic is observed. Despite this, monitoring these supplementary responses was essential to provide deeper insights. As such, properly unnesting this array and associating each IP address with the affiliated domain was a critical aspect to this query. In addition, some optimizations were made so that the query was less computationally expensive. For instance, some records contained MD5 or UUID4 hashes – likely unique sessions, fingerprints, or telemetry – yet they shared the same subdomain. Therefore, these hashes were able to be filtered with regex and merged at the

subdomain, which greatly reduced overhead and more accurately described the traffic. Furthermore, preprocessing and additional filtering steps pared down the dataset before undergoing computationally intensive calculations. One such process includes a mechanism that removes records older than 30 days if they remain inactive. This ensures statistics are current and prevents cluttering in the data. As a result, the new query includes the requested domain, returned IP address, geotagging information, number of times seen, and average duration, packets, and payload for each source and destination endpoint.

The next step in the process involved establishing a reliable storage solution for the data. MinIO partitions the data by both date and information sharing device ID, which not only simplifies access to specific datasets but also ensures efficient organization allowing for concurrency. After the analytic process is finished, the running statistics are stored in an internal Trino table for easier processing and updating, while larger, long-term datasets are saved in MinIO for extended storage needs. This data is then also available through Elasticsearch, which gives analysts access to different data visualization software for reporting.

The last step in this analytic process was to integrate the different steps and automate the workflow. An Airflow DAG (directed acyclic graph), written in Python, executes the tasks against Trino and MinIO to produce the analytic dataset with fail-safes and logging. The DAG includes the following actions in order: locking the resources while the task runs, executing the SQL code to store data in a temporary table, writing the processed data back to the main tables, and freeing the resources.

IV. CONCLUSIONS

The purpose of this project was to design an analytic process that could be integrated into the CRISP mission and collaborate with other analytics to provide a deeper understanding into the activity occurring in the United States energy sector. The project involved querying DNS traffic data to create a profile of all previously observed data, managing storage of the data, and automating the workflows. The final touch was making the process production-ready, including thorough documentation and final adjustments. The result of this process is a dataset that can be used for tagging to identify trends and determine whether traffic is legitimate or malicious.

Such ideas include a tag for a record that just appeared for the first time, or a domain or IP address pair that had not previously been associated. The analytic is fully operational and available in the CRISP ecosystem where it is creating an ongoing analysis of IP addresses and domains for analyst use. In the future, it can be expanded by adding or altering algorithms to provide additional insights. This may include creating a short-term running analysis of 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. In conclusion, this project successfully developed an analytic to monitor DNS traffic and provide the CRISP platform with insightful datasets.

V. ACKNOWLEDGEMENTS

This work was supported in part 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).

VI. REFERENCES

[1] Energy Sector Cybersecurity Preparedness, https://www.energy.gov/ceser/energy-sector-cybersecurity-preparedness (Accessed 3 July 2025)

[2] DOE, Cybersecurity Risk Information Sharing Program (CRISP), https://www.energy.gov/sites/default/files/2021-12/CRISP%20Fact%20Sheet_508.pdf (Accessed 3 July 2025)