

Analysis of cyber security data sets

Maciej Korczyński, Simon Fernandez

The assessment will take two sessions. It does not require any final report. It will be evaluated at the end of the second session by the teacher.

A good way to uncover some of the threats that exist on the Internet is by using honeypots. In this assignment, you will analyze data collected using several honeypots located on Amazon Web Services. Your main goal is to make sense of it and figure out what type of questions you are able to answer.

Instructions:

1. Download the data file from Chamilo :
Documents > Session 1 Analysis of cyber security data sets > data.csv

datetime	Timestamp
host	Name of the honeypot
src	Source IP address as long number
proto	Protocol
type	ICMP type
spt	Source port
dpt	Destination port
srcstr	Source IP address as string
cc	Country code
country	Country name

DATA FIELDS

The analysis of the provided cybersecurity data will require statistical software; During the course, it is recommended to use Jupyter Notebook (previously referred to as IPython Notebook). Please take time to get familiar with the Jupyter environment and Python using the following tutorial for beginners: <https://www.dataquest.io/blog/jupyter-notebook-tutorial>. Once, you feel comfortable with the environment, analyze the provided data.

To help you with the analysis, answer the following questions:

1. What are the traffic differences among the different honeypots? For instance, what is the honeypot that received more packets? Which country sent the highest number of packets to this honeypot?
2. The volume of incoming traffic varies per day. What is the highest number of received packets that a honeypot received in a day? Is this an outlier or the common trend? Note that sometimes the same IP sends more than one packet per day to the same honeypot. Why is that?
3. Analyze the different protocols. What is the least common protocol in terms of packets received? What do you think is the purpose of these packets?
4. Take a look at the top10 destination ports. What is the main destination port? Why do you think that the attackers are interested in these ports?
5. Finally, we want to analyze the temporal evolution of the top10 TCP destination ports. Is there any trend in the data that suggest that some ports are starting to become popular targets?

If you have additional observations don't hesitate to present them to the teacher at the end of the session.