Charting a Course to Navigate the Waters of a Cybersecurity Data Lake

Flocon 2024

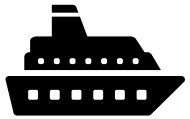
Mobile, Alabama

January 9, 2024

Rosalie Bakken, PhD; Matthew Spitzer, PhD; Jennifer Marr, BA

Development and Implementation of a Cybersecurity Data Lake sets the Stage for Adventure

- Utilizing a lake to its full extent is not necessarily intuitive
- The sheer volume and variety of data types available make many things possible
- It is easy to become overwhelmed with the endless questions that could be asked of the data
- Each use case can be explored in multiple ways, following many alternative pathways
- Each result opens another set of doors for possible exploration
- It is very easy to get lost, run amuck, or become overwhelmed trying to boil the ocean
- **A solid, well-disciplined process is needed to efficiently draw value from a cybersecurity lake**



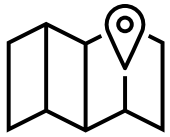
“I wisely started with a map.” – J.R.R. Tolkien

Step 1: Create a Map

- The breadth and depth of a cybersecurity data lake can be overwhelming
  - Identifying destinations in a lake is vital to pursual of targeted analytic efforts
  - We are not here for sight-seeing expeditions
- Define the problems to be addressed
  - This should be a collaborative effort
  - Focus on prioritized risks and specific use cases related to those risks
  - Annually, create a plan for specific use cases against a quarterly timeline
  - Build flexibility into the plan to accommodate emerging prioritized use cases throughout the year

Step 2: Describe the Destinations in Detail

- Frame the hypotheses that will drive the analyses. Ask:
  - What is the concern/risk being addressed?
  - What does “success” look like?
  - Does the lake contain the data needed to address the hypothesis?
  - What are the parameters of the hypothesis (i.e., timeframe, entry parameters, etc.)?



What constitutes a well-developed hypothesis or question to address?

Which of these are actionable?

- Look for reconnaissance efforts on back ups
- Identify anomalous connections to crown jewels
- Build a model to identify connections to IOCs
- Determine whether there are compromised software versions running on our network
- Look for successful SSH sessions initiated by external sources that meet business criteria x/y/z

What to avoid?

- Too broad
- Unrealistic
- Don't really have the right data to address this question
- Won't actually address the goal/business problem
- Result set will contain too much noise
- Result set will likely be far too large to be actionable

“If you don’t know what port you are sailing to, no wind is favorable.” – Seneca the Younger

Step 3: Map a Course to a Destination, and Know the Water Along the Way

Perform the following two efforts in parallel:

- Collectively walk through and document your logic
  - If the logic used to navigate the data is in error or biased, the results will be irrelevant and/or misleading. ***This step cannot be expedited!***
  - Provide guardrails against an analyst’s potential biases
  - Circle back to the logic to correct errors in approach
- Scope and verify availability of all data sources needed to support the logic
  - Does the data source align with the timeframe of the hypothesis?
  - What criteria can be used to eliminate noise from the data source?
    - What should be included, what should be excluded?
    - For example, are there any data points that could be considered “known good” so that more relevant signals are returned?
  - How can the data from each source be accurately joined together to create a meaningful result?

“When the well is dry, we know the worth of water.” – Benjamin Franklin

Step 4: Launch the Boat

- Proceed to apply predefined logic, taking the “temperature” of the water periodically
  - Collect general statistics to scope the data included in the analysis
  - Validate that the data available can answer each question being posed
  - Build a library of repeatable logic and queries

Step 5: Obtain Results and Correct Course as Needed

- Iterative approach to the analytic process
- New avenues of investigation will arise
- New discoveries about the data and relationships within it will become apparent
- Some of these will reveal errors in prior logic, and corrections will be required
- Document these in the logic
  - Provides an audit trail
  - Provides feedback to involved teams to increase understanding
  - Keeps everyone aligned through the evolving course of an analytic endeavor

Example of Logic Step #1: Initial “XX” Server Analysis

Step	Reasoning	Result
Identify each candidate XX server in production to analyze for this hypothesis	Focus on internal servers for which our lake can historically confirm IP address assignment	27 XX servers identified
Identify the full IP address history for each specific XX server, complete with start and end timestamps	Historically accurate IP address/timeframe combinations are required to reasonably attribute connections meeting those criteria to the XX servers.	Server 1: • 10.10.10.1, 2/2/2022 11:14 – 4/18/2022 15:20 • 10.10.10.40, 4/18/2022 15:21 – 9/1/2022 7:39 Server 2: • 10.10.10.2, 3/8/2022 9:20 – 12/18/2022 23:59 …

Conclusion

Relevant Netflow records can be identified using the start and end timestamps of the IP address assignment for each specific server, which results in the correct attribution of network connections to servers at the time their IP addresses were valid. Analysis can then be performed on the relevant connection activity.

“Perhaps some detours aren’t detours at all. Perhaps they are actually the path.” – Katherine Wolf

Side Trip: Unexpected Discoveries and Detours

- New discoveries will be made while executing the pre-defined logic
  - Constant distractions derail progress and potentially introduce chaos
  - Document discoveries so they are not lost, but kept in context to prevent re-charting a new course of logic

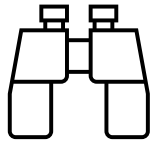
Step 6: Reaching the Destination

- Ask questions about the results
  - Do the results directly address the hypothesis?
  - Do the results meet the success criteria as defined in the hypothesis?
- Hand off results to customers thoughtfully and in context
  - Establish a secure process up front for sharing results
  - Translate to a scale absorbable by the audience (avoid death-by-PowerPoint)
  - Document, share, and act on learnings
  - Communicate policy-related items with appropriate stakeholders
  - Use insights to frame and prioritize plans for the analytic next cycle

“One’s destination is no longer a place, rather a new way of seeing.” – Henry Miller

Beyond proving or disproving any specific hypotheses, a data lake brings opportunities for organizational self-reflection and evolution

- A large-scale cybersecurity data lake provides a much more powerful environment for analysts
  - It can be seen as a larger hammer
  - Or it can be leveraged as an entirely different tool with much deeper, broader, and more integrative capabilities delivering multiple functionality
- To realize the greater value of the data lake, its collective users need to be willing to change
  - Work with new presumptions and think beyond the status quo
  - Leverage existing collaborations and create new workflows



“So, throw off the bowlines, sail away from the safe harbor, catch the trade winds in your sails.” – Mark Twain

The greatest value of the lake comes from its ability to inspire creativity in the quest to thwart cyber attacks

- Supports complex and purpose-tuned analytic efforts
- Its use can be customized to organizational needs
- More powerful than a data repository, and includes the precision, flexibility, interconnected data, and sophistication to facilitate answering questions that no one could envision previously

Wade into this space thoughtfully, with attention to the people and the process aspects that need to be as well-developed as the technical aspects of using the data lake