Cyber aXioms

Ofer Shezaf contemplating the role & value of cybersecurity

Should we normalize security data at all?

I haven't blogged for quite a while. Recently I started spending my time again on security research, and the blogging itch is back. My current research focus is security data normalization, and in the next few posts, I will expand on this topic.

The first question that comes to mind is, why normalize? Should we normalize at all? That is obviously once we agree on what normalization is. So let's start there.

At its core, normalization means that data collected from different sources should be converted to a uniform presentation or schema. Such a uniform schema enables analytics to be source agnostic. It also reduces the learning curve for analysts and enables them to be more productive. The article "SIEM Event Normalization Makes Raw Data Relevant to Both Humans and Machines" provides a good starting point for the rationale.

To deliver on the promise, SIEMs have tried to implement normalization since day one. ArcSight CEF and categorization, Splunk CIM, and QRadar LEEF are all normalization schemes.

Where they successful?

In his seminal blog post, "Security Correlation Then and Now: A Sad Truth About SIEM", Anthon Chuvakin claims that they were not. And I tend to agree. Want proof? If you are a serious security analyst, the number 4624 means something to you. Obviously, it is the Windows Login event. More precisely, successful login (4625 logs failures). You might also know that Login Type 2 is "interactive", or at least you know that you need to consult Randy Franklin Smith's excellent

Ultimate Windows Security. I have certainly used it a lot, as you can see on the right. Or just Google for 4624.

In a perfect world, an analyst would not need to know about event 4624. ArcSight categorization whitepaper mapped it already in the first decade of the millennium to this:

Now, how many people converse in 6424, and how many know the ArcSight categorization. How many systems analyze 4624 events, how many support ArcSight, or an alternative, categorization scheme?

So Anton has a point.

Now back to 2021. My current research at Microsoft, leading the Azure Sentinel Information Model (ASIM) initiative, enables me to get back to the challenge ArcSight started tackling more than 20 years ago. And I hope this time to move the needle. Let there be a generation of security analysts who don't know what 4624 is (and not because Windows will die).

As a starting point, we recently released the ASIM Authentication model, which includes a normalizing 4624 parser. I am sure it is not perfect, and we are already getting ideas for improvement.

In the upcoming blog posts, I will discuss how we try to make normalization work this time. I will address areas such as:

- Value normalization
- Entities, entity IDs, and entity descriptors
- Aliasing

So let's start the journey.

This entry was posted in Normlization, SIEM on July 1, 2021 [https://xiom.com/2021/07/01/should-we-normalize-security-data/] .

3 thoughts on “Should we normalize security data at all?”

Chip Mesec
July 5, 2021 at 4:22 pm

Ofer:
Great point. Another driver is capex/opex. How much bigger are analytic and storage backends because of lack of normalization? I wonder if you could measure this with your tools? It would move some CSOs to your point of view. Good work.

Shezaf Post author
July 5, 2021 at 5:31 pm

Great point, Chip! My experience is that it’s a mixed bag. While you make a great point that normalization reduces the size by eliminating meta information such as column names and waste from the raw data, normalization also adds some. For example, if a source reports an ID and the normalized schema calls to resolve t9 a string value.

Indeed a worthwhile topic to add to my list!

Pingback: SIEM Normalization Dirty Secret: Values | Cyber aXioms