Setting up a security analysis toolbox

Introducing serverless in security

By: Hatitye Chindove

Presentation approach

- Discussions and Labs
  - Covering basics of discovery of network activity against common external intelligence feeds.
  - Limiting death by powerpoint
  - I reserve the right to say “I don’t know” :-)
  - Freebies :D - Sample code made available on Github
    - https://github.com/hatityechindove/Analysis-Tool-Box-Demo
- Not a one size fits all demonstration
  - There are a 100 ways to skin the cat
  - Ask questions that can relate to your context

Agenda:

1. Data driven security approach
2. Toolbox setup
    a. Tools and frameworks primer
    b. Harvester (Collection)
    c. Data Lake (Storage)
    d. C.I.A triage (Data Lab)
3. Project approach and scoping
4. Project run (Analytics and Reporting)
5. Recommendations
6. Others :-)

Data driven security

What is data driven security?

Data driven security is a concept utilized in organizations operating in a constantly changing environment to effectively manage the dynamic risks which challenge them through applying a data centric approach to information security management.

Why data driven security?

“Out with the gut feelings, in with the analysis. Fact based decisions enable accurate definition of strategy.”

Success of security can be achieved through a carefully balanced act of orchestration:

1. Selecting and deploying effective security measures
2. Working within a budget ($1000 fence for $100 donkey)
3. Reducing liability of exposure

Data driven security refers to using measurable factors to drive a security program.

Framework Primer

Serverless Framework:

Serverless is your toolkit for deploying and operating serverless architectures. Focus on your application, not your infrastructure.

[https://serverless.com/]

Terraform:

Terraform is a tool for building, changing, and versioning infrastructure safely and efficiently.

[https://www.terraform.io/]

Anaconda:

Anaconda is a freemium open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment.

[https://www.anaconda.com/]

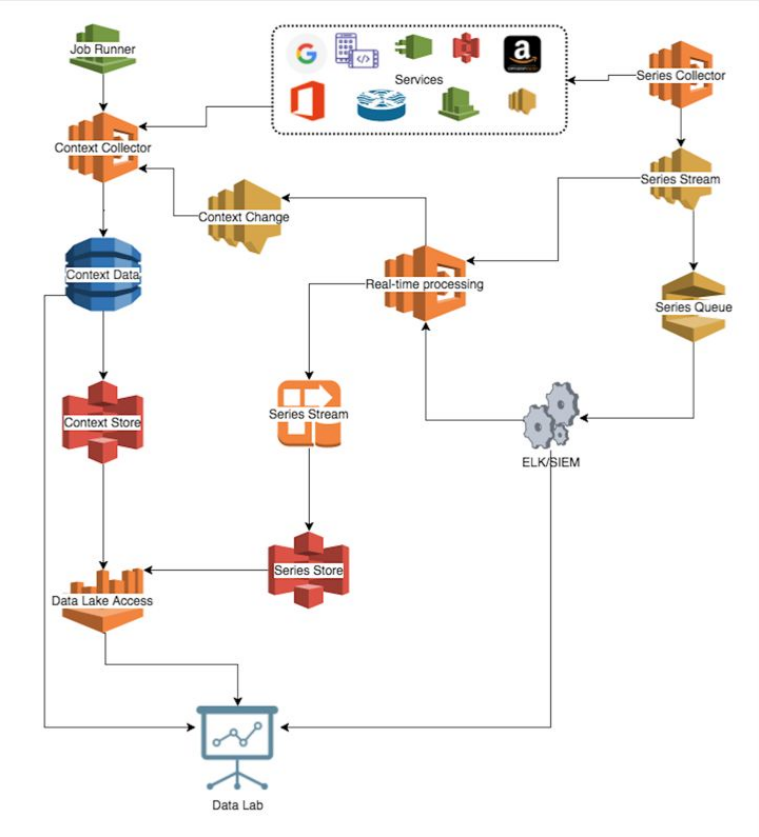
Starting an investigation

- Finding and collecting analysis data (Data lake and other sources)
  - VPC flow logs (Series Data)
  - EC2 instance descriptors and Network card descriptors (Context Data)
  - Alienvault reputation data sets (External IP Reputation data sets)
- Conducting analysis
  - Scoping analysis and working towards a question
  - Exporting results

Components

- Scraper/Harvester (Collection)
- Data Lake (Storage)
- Data Lab (Analysis)
- Presentation (Reporting)

Setup


Job Runner
Services
Series Collector
Context Collector
Series Stream
Context Change
Context Data
Real-time processing
Series Queue
Context Store
Series Stream
ELK/SIEM
Series Store
Data Lake Access
Data Lab

Data Lake

- Definition
  - A central location where all security data is collected and stored
    - Using a data lake is similar to log management or security information and event management (SIEM) but not a replacement
    - Typically not used for real time analytics
- Four goals:
  - Provide a single way or a process to collect all data
  - Process, clean, and enrich the data in one location
  - Only store data once
  - Access the data using a standard interface

Data Lake key capabilities

1. Automated collection
2. Security context
3. Time series data to contextual data mapping
4. Security analysis and reporting interface
5. Scale out architecture (horizontal/vertical)

Types of data

- Time series data (a.k.a log data)
  - Mostly single line records with timestamps such as IDS, firewalls, web servers, net flow logs.
  - In some instances they are called alerts or events.
- Contextual data (a.k.a context)
  - Mostly records that provide information about specific objects of time series data e.g. Configuration management systems, directories, special purpose applications (such as HR systems)

Data Lake usage scenario 1

Sample scenario to ask a data lake:

A new CVE (e.g. Heartbleed, Spectre, CVE-2017-0143 for SMB remote exploit) has been filed for a dependency which you have found on Github insights or a something you found online.

So for our security team's sanity, we would like to answer these questions:

1. Are we or have we been exposed to this vulnerability?
2. If yes, what was exposed to this vulnerability?
3. If something was exposed, did someone exploit it?
4. If yes, how did they exploit it?

Data Lake usage scenario 2

Sample scenario to ask a data lake:

The security team has been getting a lot of alerts from our SIEM about brute-force attacks from known compromised networks that are using repupation data set. So for the security team's sanity we would like to answer these questions:

1. What is exposed on our network to these threats?
2. What are these compromised networks really targeting?
3. Which of these alerts should we really be paying attention to?
4. How can we improve our network to counter these attacks?

Sample collection stream

Lab 1 - Manually creating collection stream

- Manually setting up collection and storage for netflow log series

Lab 2 - Creating collection stream as code

- Setting up collection and storage for netflow log series as code
  - Using Serverless and Terraform

Setting up the data lake

Zappi Store

Lab 3 - Pushing series to data lake

- Create VPC flow logs series data lake pool
- Create a consumer service for flow log publisher
- Create a push to flow logs Series pool

C.I.A Triage

- Information security is primarily focused on the preservation of confidentiality, integrity and availability of information through risk management
  - Will adding this security feature enhance any of the core security principles(Question I always have to ask myself before jumping on a project)?
- Information risk management in the scope of C.I.A
  - What risk do I have (Probability and Impact)?
  - What controls for mitigation do I have?
  - What is the inherent risk and what is the residual risk?
  - Are the risks acceptable?

Sample project setup

Lab 4 - Conducting analysis for ingested data

- Security analysis project setup
- Setting up questions

Target:

1. *Risk to information confidentiality and integrity*

Question:

1. *Which attackers should we take on to address the concerns regarding information confidentiality and integrity on our network?*
2. *What sort of threats are we facing and how can we mitigate against them?*

Presenting findings

Lab 5 - Presenting findings

- Reporting
- Visual presentation
  - https://us-east-1.quicksight.aws.amazon.com/sn/dashboards/37b06479-c3d0-4c11-8e5b-3ba8f4339c57

Recommendations

- Threat modelling
- Lockheed cyber kill chain

Threat Modelling

Ideally threat modelling should be done before rollout

- Justifiable as part of the continuous improvement process

DREAD – Risk Ranking Methodology

With the DREAD model, you arrive at the risk rating for a given threat by asking the following questions:

- Damage potential: How great is the damage if the vulnerability is exploited?
- Reproducibility: How easy is it to reproduce the attack?
- Exploitability: How easy is it to launch an attack?
- Affected users: As a rough percentage, how many users are affected?
- Discoverability: How easy is it to find the vulnerability?

Risk = (Damage + Reproducibility + Exploitability + Affected Users + Discoverability) / 5

What Are You Going To Do About It?

Spoofing	Authentication	• Passwords, multi-factor authN • Digital signatures
Tampering	Integrity	• Permissions/ACLs • Digital signatures
Repudiation	Non-Repudiation	• Secure logging and auditing • Digital Signatures
Information Disclosure	Confidentiality	• Encryption • Permissions/ACLS
Denial of Service	Availability	• Permissions/ACLs • Filtering • Quotas
Elevation of privilege	Authorization	• Permissions/ACLs • Input validation

Cyber Kill Chain

Detect	Deny	Disrupt	Degrade	Deceive
Web analytics	Firewall ACL			
NIDS	NIPS			
Vigilant User	Proxy filter	Inline AV	Email Queuing	
HIDS	Vendor Patch	EMET, DEP		
HIDS		AV		
NIDS	Firewall ACL	NIPS	Tarpit	DNS redirect
Audit log			Quality of Service throttle	Honeypot

Figure 1: Countermeasures

Other external resources

Ipinfo.io

udger.com

maxmind.com

shodan.io

censys.io

pastebin.com

mxtoolbox.com

http://iplists.firehol.org/

Other great resources

- https://www.streamalert.io/overview.html [AirBnB Alerting]
- http://mybinder.readthedocs.io/en/latest/using.html [Notebook Runtime]
- https://lockheedmartin.com/content/dam/lockheed/data/corporate/documents/Seven_Ways_to_Apply_the_Cyber_Kill_Chain_with_a_Threat_Intelligence_Platform.PDF

AWS knowledge areas covered

- VPC flow logs - https://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/flow-logs.html#flow-logs-basics
- Athena - https://aws.amazon.com/athena/ (You are charged $5 per terabyte scanned by your queries.)
- Lambda - https://aws.amazon.com/lambda/ (First 1M requests per month are free. $0.20 PER 1M REQUESTS THEREAFTER $0.0000002 per request.)
- S3 - https://aws.amazon.com/s3/ (First 50 TB / Month $0.023 per GB)
- Cloud Watch Logs - https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/WhatIsCloudWatchLogs.html ($0.50 per GB ingested)
- SNS - https://aws.amazon.com/sns/ (No charge for deliveries to Lambda and SQS)
- Firehose -https://aws.amazon.com/kinesis/data-firehose/ (First 500 TB / month at $0.029 per GB i.e 12TB at approx $350/month)

Extra beef up:

- Cloudformation (Used by Serverless framework)
- Inspector (Analysis) - https://docs.aws.amazon.com/inspector/latest/userguide/inspector_introduction.html
- Guard Duty (Threat Detection) - https://aws.amazon.com/guardduty/

Special thanks

- Mark Regensberg
- Zac Blazic
- Simbarashe Nyatsanga
- Nicholas Moorcroft