

# Apache Metron

## Lessons Learned

**Casey Stella**  
@casey\_stella



2018

## Introduction

---

Hi, I'm Casey Stella!

# Apache Metron: A Cybersecurity Analytics Platform

---

- Metron provides a scalable, advanced security analytics framework to offer a centralized tool for security monitoring and analysis.

# Apache Metron: A Cybersecurity Analytics Platform

---

- Metron provides a scalable, advanced security analytics framework to offer a centralized tool for security monitoring and analysis.
- Metron was initiated at Cisco in 2014 as OpenSOC.

# Apache Metron: A Cybersecurity Analytics Platform

---

- Metron provides a scalable, advanced security analytics framework to offer a centralized tool for security monitoring and analysis.
- Metron was initiated at Cisco in 2014 as OpenSOC.
- Metron was submitted to the Apache Incubator in December 2015

# Apache Metron: A Cybersecurity Analytics Platform

---

- Metron provides a scalable, advanced security analytics framework to offer a centralized tool for security monitoring and analysis.
- Metron was initiated at Cisco in 2014 as OpenSOC.
- Metron was submitted to the Apache Incubator in December 2015
- Metron graduated to a top level project in April 2017

## Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus



# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework
  - HBase provides a low latency key/value lookup store for enrichments and profiles

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework
  - HBase provides a low latency key/value lookup store for enrichments and profiles
  - Zookeeper provides a distributed configuration store

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework
  - HBase provides a low latency key/value lookup store for enrichments and profiles
  - Zookeeper provides a distributed configuration store
- Ingested network telemetry can be enriched pluggably

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework
  - HBase provides a low latency key/value lookup store for enrichments and profiles
  - Zookeeper provides a distributed configuration store
- Ingested network telemetry can be enriched pluggably
  - New enrichments can be done live on running topologies without restart

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework
  - HBase provides a low latency key/value lookup store for enrichments and profiles
  - Zookeeper provides a distributed configuration store
- Ingested network telemetry can be enriched pluggably
  - New enrichments can be done live on running topologies without restart
  - New enrichment capabilities can be added via user defined functions

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework
  - HBase provides a low latency key/value lookup store for enrichments and profiles
  - Zookeeper provides a distributed configuration store
- Ingested network telemetry can be enriched pluggably
  - New enrichments can be done live on running topologies without restart
  - New enrichment capabilities can be added via user defined functions
  - Enrichments can be composed through a domain specific language called **Stellar**

# Characteristics of Metron

---

- Metron is built atop the Apache Hadoop ecosystem handle capturing, ingesting, enriching and storing streaming data at scale
  - Kafka provides a unified data bus
  - Storm providing a distributed streaming framework
  - HBase provides a low latency key/value lookup store for enrichments and profiles
  - Zookeeper provides a distributed configuration store
- Ingested network telemetry can be enriched pluggably
  - New enrichments can be done live on running topologies without restart
  - New enrichment capabilities can be added via user defined functions
  - Enrichments can be composed through a domain specific language called **Stellar**
- Data stored in HBase can be the source of enrichments



# Characteristics of Metron

---

- Enriched telemetry can be indexed into a Security data lake
  - Indexes supported are pluggable and include HDFS, Solr and Elasticsearch

# Characteristics of Metron

---

- Enriched telemetry can be indexed into a Security data lake
  - Indexes supported are pluggable and include HDFS, Solr and Elasticsearch
- Advanced analytics can be done on streaming data

# Characteristics of Metron

---

- Enriched telemetry can be indexed into a Security data lake
  - Indexes supported are pluggable and include HDFS, Solr and Elasticsearch
- Advanced analytics can be done on streaming data
  - Probabalistic data structures (e.g. sketches) can sketch streaming data across time and enable approximate distribution, set existence and distinct count queries

# Characteristics of Metron

---

- Enriched telemetry can be indexed into a Security data lake
  - Indexes supported are pluggable and include HDFS, Solr and Elasticsearch
- Advanced analytics can be done on streaming data
  - Probabalistic data structures (e.g. sketches) can sketch streaming data across time and enable approximate distribution, set existence and distinct count queries
  - Models can be deployed using Yarn, autodiscovered via Zookeeper and interrogated via Stellar functions

# Stellar

---

Metron needed the ability to allow users to pluggably and consistently enrich and transform streaming data. Out of this need, we created **Stellar**:

# Stellar

---

Metron needed the ability to allow users to pluggably and consistently enrich and transform streaming data. Out of this need, we created **Stellar**:

- Interact with the various enabling Hadoop components in a unified manner

# Stellar

---

Metron needed the ability to allow users to pluggably and consistently enrich and transform streaming data. Out of this need, we created **Stellar**:

- Interact with the various enabling Hadoop components in a unified manner
- Compose a rich set of built-in functions with user defined functions

# Stellar

---

Metron needed the ability to allow users to pluggably and consistently enrich and transform streaming data. Out of this need, we created **Stellar**:

- Interact with the various enabling Hadoop components in a unified manner
- Compose a rich set of built-in functions with user defined functions
- Provide simple primitives around the functions: boolean operations, conditionals, numerical computation.

Think of Stellar as Excel functions that we can run on streaming data.



```
window := PROFILE_WINDOW('...')
profile := PROFILE_GET('attempts_by_user', user, window)
distinct_auth_attempts := HLLP_CARDINALITY(GET_LAST(profile))
distribution_profile := PROFILE_GET('auth_distribution', 'global', window)
stats := STATS_MERGE(distribution_profile)
distinct_auth_attempts_median := STATS_PERCENTILE(stats, 0.5)
distinct_auth_attempts_stddev := STATS_SD(stats)
```

## Streaming Technologies are Still Immature

---

Building Metron on streaming technologies (e.g. Storm) came with a host of lessons learned:

- The Bad

# Streaming Technologies are Still Immature

---

Building Metron on streaming technologies (e.g. Storm) came with a host of lessons learned:

- The Bad
  - It has more knobs to tune than a jet airplane, but without the autopilot

# Streaming Technologies are Still Immature

---

Building Metron on streaming technologies (e.g. Storm) came with a host of lessons learned:

- The Bad
  - It has more knobs to tune than a jet airplane, but without the autopilot
  - Sometimes it fails at the simple stuff

# Streaming Technologies are Still Immature

---

Building Metron on streaming technologies (e.g. Storm) came with a host of lessons learned:

- The Bad
  - It has more knobs to tune than a jet airplane, but without the autopilot
  - Sometimes it fails at the simple stuff
- The Good

# Streaming Technologies are Still Immature

---

Building Metron on streaming technologies (e.g. Storm) came with a host of lessons learned:

- The Bad
  - It has more knobs to tune than a jet airplane, but without the autopilot
  - Sometimes it fails at the simple stuff
- The Good
  - We chose Storm, in part, because it was the most battle-tested of the lot

# Streaming Technologies are Still Immature

---

Building Metron on streaming technologies (e.g. Storm) came with a host of lessons learned:

- The Bad
  - It has more knobs to tune than a jet airplane, but without the autopilot
  - Sometimes it fails at the simple stuff
- The Good
  - We chose Storm, in part, because it was the most battle-tested of the lot
  - Storm's abstractions are sufficient to solve the business problem in a linearly scalable manner

# Abstractions Cost Time and Money

---

- Big data technologies are abstraction bleeding machines



# Abstractions Cost Time and Money

---

- Big data technologies are abstraction bleeding machines
- Understand the cost of the operations that you depend on

# Abstractions Cost Time and Money

---

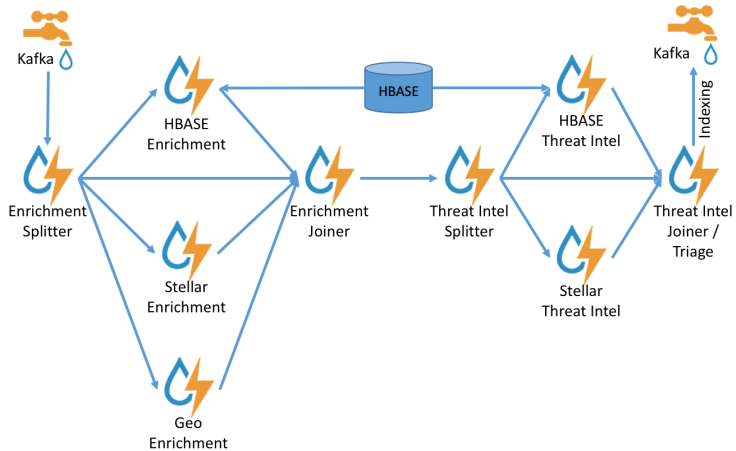
- Big data technologies are abstraction bleeding machines
- Understand the cost of the operations that you depend on
- What can seem perfectly logical on the whiteboard can be a dog on the cluster

# Abstractions Cost Time and Money

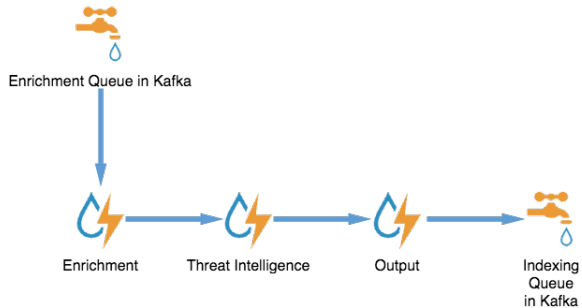
---

- Big data technologies are abstraction bleeding machines
- Understand the cost of the operations that you depend on
- What can seem perfectly logical on the whiteboard can be a dog on the cluster
- Even so, swapping out technology almost always just shuffles problems around.

## Enrichment: Old & Busted



## Enrichment: New Hotness



# Kappa Is a Great Tool But a Poor Master

---

- We started with a Kappa architecture

# Kappa Is a Great Tool But a Poor Master

---

- We started with a Kappa architecture
- Pretty much immediately we were asked to rerun data in batch

# If you take away three things from this talk...

---

- Monitor, monitor, monitor



# If you take away three things from this talk...

---

- Monitor, monitor, monitor
- Streaming analytics, marriage and diplomacy are exercises in compromise.

## If you take away three things from this talk...

- Monitor, monitor, monitor
- Streaming analytics, marriage and diplomacy are exercises in compromise.
- Effective streaming analytics is about bringing to bear as much context into the data streaming by as you possibly can computationally. Find compromises accordingly.

# Questions

---

Thanks for your attention!

- Find me at <http://caseystella.com>
- Twitter handle: @casey\_stella
- Email address: [cestella@gmail.com](mailto:cestella@gmail.com)