**Big Data Pipeline for Apache Metron Cyber Security**

**Contents Page No**

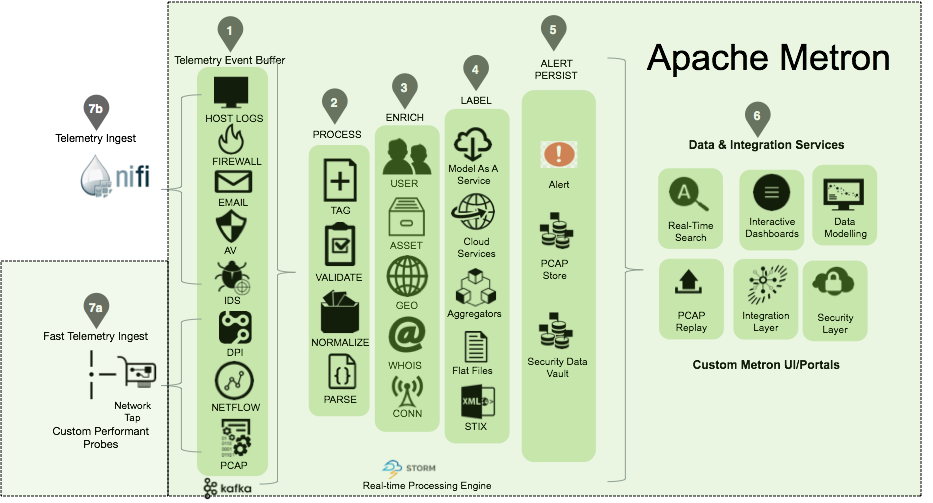
1. Introduction …………………………………………………………………………… 2
2. Big Data Architecture for log processing…………………………………. 2
3. Cluster Setup using Ansible Playbook……………………………………… 3
4. Choose the Services……………………………………………………………….. 3
5. Anomaly Detection for Cybersecurity…………………………………….. 3
6. Pre-Processing the Logs………………………………………………………….. 4
7. Publish the logs to Kafka…………………………………………………………. 4
8. Filter the anomaly from UI……………………………………………………….4
9. Reference……………………………………………………………………………….. 4
10. **Introduction**

Anomaly detection is a process of detecting the unexpected data or events in the datasets which differ from the normal. In real time bank are interested to detecting the fraud transaction, cyber security analysts are interested in detecting unexpected incidents in logs and traffic patterns. In this doc handling the anomaly using Metron Cyber Security approaches.

1. **Big Data Architecture for log processing**

Anomaly detection is usually a big data analytics problem, and is often a real-time streaming analytics problem as well. This obviously calls for a big data analytics pipeline that has streaming capability. One such open source platform that you can build on is the Hortonworks Data Platform (HDP) along with Hortonworks Dataflow (HDF) which are now part of Cloudera. The Hortonworks Cyber Security Platform (HCP) is built on top of HDP and HDF.

Apache Metron is a cyber security application framework that provides organizations the ability to ingest, process and store diverse security data feeds at scale in order to detect cyber anomalies and enable organizations to rapidly respond to them. Metron consists of Nifi, Kafka, Strom, Elastic Search and Kibana.



Apache metron involves 3 processing stages Parsing, Enrichment and Indexing.

1. **Cluster Setup using Ansible Playbook**

For streaming log process its required hadoop cluster (Hortonworks), its required hardware configuration of cluster is RAM, CORE and Hard Disk based on volume of data to be handle. To automate the cluster installation using the ansible playbook.

Ansible is simple open source IT engine which automates application deployment, intra service orchestration, cloud provisioning and many other IT tools. For this install the ansible-playbook to be every machine. In the ansible configuration file update the Master and Slaves Cluster hostname.

**Prerequisites**

* Centos 7
* JDK1.8
* Python 2.7
* SSH Server across the Cluster

1. **Choose the Services**

After successfully launch the cluster setup need to choose the service to be run on each machine using Ambari (Cluster Management). Below URL the Ambari Service will be hosted.

UI: <http://Master_IP:8080>

1. **Anomaly Detection for Cybersecurity**

A Source of Cybersecurity events or logs are existing from **SIEM** (Security Incident and Event Management) tool from which pre-processed events and logs can be streamed into **HDFS** (Hadoop Distributed File System) stored as **Avro file** format. SIEM tools are usually rules based and generate a lot of false positives which need to be filtered. It contains various types of log from authentication and security focused devices, server and applications such as AD servers, firewalls, VPN, IPS/IDS systems etc.

1. **Pre-Processing the Logs**

Using Apache Nifi to monitor the HDFS (Avro file) to notify the file newly create when SIEM tools push the data. The File path pass to Java Application to pre-process the log likes

* Convert the timestamp to custom date format (yyyy-MMM-dd HH:mm:ss.mmm)
* Convert the empty IP to default IP (127.0.0.1)
* Convert the empty port to default Port (0)
* Convert the empty String value to NA.

after that required fields are filter and convert into **CSV File format**.

1. **Publish the logs to Kafka**

Kafka is a distributed streaming platform that is used to publish and subscribe to streams of records. It is used for fault tolerant storage.

Based on each device type (ciscoasa, deep discovery, etc) the record will be publish into each Kafka topics. Apache Storm process the streams data from Kafka it handle the three stage Parser, enrichment and Indexing.

1. **Filter the anomaly from UI**

Indexing record will be stored in Kafka topic and elastic search sync for log management. Kibana is an open source visualization tool mainly used to analyse a large volume of logs in the form of line graph, bar graph, pie charts etc.

1. **Reference**

<https://cwiki.apache.org/confluence/display/METRON/Metron+0.4.0+with+HDP+2.5+bare-metal+install+on+Centos+6+with+MariaDB+for+Metron+REST>

<https://community.hortonworks.com/articles/60805/deploying-a-fresh-metron-cluster-using-ambari-serv.html>