GigaSOC: Machine Learning applied to detect cyberattacks on MRENs

Marcos F. Caetano, Paulo Angelo A. Resende and Lucas R. Costa.

GigaCandanga Network (*https://gigacandanga.net.br*)

THE PROBLEM

Due to the impact and recurrence of cyberattacks over the last years, cybersecurity has been a relevant concern for society. Timely detecting cyberattacks on Metropolitan Research and Education Networks (MRENs) may play an important role in protecting networks and systems.
However, detecting attacks on large networks is not trivial due to the volume and the distribution of data over the network. Collecting and processing network packages in such an environment is challenging.

PROPOSAL: GigaSOC

The GigaSOC project aims to apply machine learning and big-data technologies to store and process network flows in order to detect attacks against the backbone and against networks connected to the MREN.

Network flows (Netflow, IPFIX or SFlow) are extracted on routers and sent to a collector system which stores data into a NoSQL database. Flows are used to provide network visibility and also are submitted to intrusion detection approaches. The development is based on the big data paradigm, which enables scalability to process large amounts of data.

TECHNOLOGIES

The main used software technologies are described below:

- **Apache Spark:** an open-source engine for parallel data processing, including machine learning algorithms that are used to detect attacks in network flows.
- **HBase:** is an open-source NoSQL distributed database that runs on top of Hadoop/HDFS, used to store data for processing.
- **Hadoop/HDFS:** an open-source distributed file system that supports HBase.
- **Elasticsearch:** a distributed search engine used to store and search network flows for visualization.
- **Logstash:** an application used to parse network flows (sFlow, Netflow, etc) into a structured JSON.
- **Kibana:** a web interface used to visualize data retrieved from Elasticsearch.

ARCHITECTURE

The core project's architecture was developed through 8 servers with Ubuntu 20.04, 4-core processors, and 8Gb of RAM. The ELK stack (Elasticsearch, Logstash, and Kibana) was used. Clustered installation on Elasticsearch was used on three servers. In addition, a server is used for Kibana and another for Logstash. The remaining three machines were left with clustered installations of Apache Spark, HBase, and Hadoop/HDFS.

Figure 1 represents the environment's implementation architecture. The arrows show the relationship between the different mechanisms of the project. The developer can interact with the environment through the Jupyter Notebook in which he can develop machine learning models via PySpark language.

RESULTS

The main result of the project is the creation of an open and extensible cloud-based infrastructure using open softwares for Metropolitan Research and Education Networks (MRENs). This infrastructure is capable of offering and making available means for the development of cyberattack detection algorithms based on real network flows. This brings the development of mechanisms much more effective, as the tests are developed in realistic network environments.

GIGASOC ARCHITECTURE