SpotProposal

SpotProposal

Abstract

Spot is an open source platform for network telemetry (packet, flow, and proxy at the moment) built on an open data model and Apache Hadoop.

Proposal

Spot (formerly Open Network Insight, or ONI) is an open source solution for network telemetry (packet, flow, and proxy at the moment) built on an open data model and Apache Hadoop. It provides ingestion and transformation of binary data, scalable machine learning, and interactive visualization for identifying threats in network flows and DNS packets.

Spot has a pluggable architecture that can accommodate multiple open data models. Although cybersecurity/network-intrusion analysis is the initial use case for Spot, we are actively encouraging the contribution of new models that will enable other adjacent applications, such as fraud detection or IT-operational analytics such as performance and health monitoring. Because these models are open, users maintain control of their own data.

More information on Spot can be found at the existing project website at http://open-network-insight.org/.

Background

It almost goes without saying that cybersecurity is an acute and paramount concern globally, for organizations of all types and sizes. Fortunately, thanks to the availability of massively scalable (in the PBs) data infrastructure, security professionals can now make authentically data-driven decisions about how they protect their assets. For example, records of network traffic, captured as network flows, are often stored and analyzed for use in network management, and this same information can provide valuable insights into network vulnerabilities.

Cybersecurity is just one example, however: There are other examples of adjacent use cases, such as user fraud detection or IT-operations analytics, that would benefit from the combination of Spot functionality and PB-scale data sets for analysis.

Rationale

Although cybersecurity is its initial use case/data model, Spot is intended to more generally tackle the dual challenges of facilitating the development of big data-driven analytic solutions, while helping vendors avoid having to create one/off infrastructure for each use case. Spot will eliminate issues related to vendor data models that create silos between solutions, and that make it difficult for users to consume these innovations from multiple vendors. In summary, Spot will accelerate the development of new massively scalable analytic applications that give users more flexibility, and more choices.

As an initial effort, we are now seeking to build an ecosystem of developers, data scientists, and security professionals to make Spot the open, community-driven, cybersecurity platform standard it needs to become. By bringing Spot to Apache, we hope to galvanize these groups to cooperate in this highly matrixed effort, and to build a global, and diverse, Spot community.

Initial Goals

Move the existing codebase, website, documentation, and mailing lists to Apache-hosted infrastructure Work with the infrastructure team to implement and approve our build and testing workflows in the context of the ASF Incremental development and releases per Apache guidelines

Current Status

Releases

Spot has undergone one public release (1.0). This initial release was not performed in the typical ASF fashion; we will adopt the ASF source release process upon joining the incubator.

Source

Spot’s source, including core platform and associated submodules, is currently hosted in several GitHub repositories under the indicated licenses:

- Core (Apache License 2.0)
- Oni-ingest (Apache License 2.0)
- Oni-ml (Apache License 2.0
- Oni-oa (BSD & MIT)
- Oni-setup (Apache License 2.0)
- Oni-nfdump (BSD)
- Oni-lda-c (GNU General Public License version 2)

The repositories will be transitioned to Apache’s git hosting during incubation. Issues related to GPL code will be resolved during incubation.

Issue Tracking

Spot’s bug and feature tracking is hosted on Github at:

- https://github.com/Open-Network-Insight/open-network-insight/issues

Issue tracking will be transitioned to Apache’s JIRA instance during incubation.

Code review

Spot maintainers currently use “LGTM” (Looks Good to Me) in comments on the code review to indicate acceptance, with at least three LGTMs required to approve the merge.

Community discussion

A Spot Slack channel is available at:

- https://opennetworkinsights.slack.com/messages/general/ (Invites request via http://open-network-insight.org:3000/)

Community discussion options will be expanded considerably when apache.org mailing lists are available.

Meritocracy

We intend to adhere to a meritocratic approach to electing new committers and PMC members. We also believe that contributions can come in forms other than just code. We will encourage contributions and participation of all types, and ensure that contributors are appropriately recognized and that PMC memberships are appropriately earned.

Community

Though Spot is a relatively new project, it has already seen promising adoption:

- Intel is the original development sponsor for Spot.
- Cloudera is strong advocate for open source cybersecurity solutions and Apache Hadoop, and a supporter of Spot.
- Cloudwick’s OAS cybersecurity solution is built on Spot.
- Accenture’s Cyber Intelligence Platform solution is built on Spot.
- Centrify has announced its intention to contribute identity-based security features to Spot’s network-intrusion detection data model.
- Webroot has announced its intention to contribute endpoint-security functionality.
- Cybraics has announced its intention to contribute network-security functionality.
- Jask has announced its intention to contribute network-security functionality.

As described in the “Rationale” section, we believe that building on and expanding the Spot community will be a key aspect in its success.

Core Developers

Spot was initially developed as a project at Intel, and most of the contributions to date have been from developers employed by that company. By bringing Spot to Apache, we hope to diversify its developer community more rapidly.

Alignment

Spot is built on Apache Hadoop, Apache Kafka, and Apache Spark, and as more functionality is built out, integration with other Apache projects is very likely.

Known Risks

Orphaned products

The risk of Spot being abandoned is low. Intel has made substantial investments already, Cloudera has publicly expressed the importance of Spot as a “killer app” for Apache Hadoop, and Cloudwick and Accenture both have offerings that are built on Spot/CDH.

Inexperience with Open Source

Most of Spot’s initial committers have experience in open source development, although not necessarily within the ASF. Those Spot developers who have little open source experience or are not Apache committers are eager to learn ASF practices as a means of improving project governance and diversifying the developer community.

Homogenous Developers

As mentioned previously, Intel developers are mostly responsible for what Spot code exists, to date. As a benefit of ASF governance, we hope to scale-up contributions from new developers and community members and eventually, develop them into committers by adhering to Apache’s meritocratic principles.

Reliance on Salaried Developers

To date, all Spot code has been written by salaried developers (chiefly employed by Intel).

Relationships with Other Apache Products

Spot is currently related to the following other Apache projects:

- Apache Hadoop
- Apache Spark
- Apache Kafka

We look forward to continuing to integrate and collaborate with these communities.

A Excessive Fascination with the Apache Brand

Although most (not all) of the initial committers are not currently Apache committers, they are resolved to learning, with the help of the more experienced committers/project mentors/champion, the Apache Way. We believe that adhering to these principles will be of great value with respect to meeting long-term project goals, including facilitating widespread adoption.

Documentation

Spot functionality is divided into different repositories, with each repository containing the relevant developer documentation:

- oni-ingest
- oni-ml
- oni-oa
- oni-setup
- oni-nfdump
- oni-lda-c

An Installation Guide is published in the project wiki:

- https://github.com/Open-Network-Insight/open-network-insight/wiki

The Spot (currently Open Network Insight) website is managed via a Wordpress instance hosted by Bluehost:

- http://open-network-insight.org/

A Docker-based demo is available via Docker Hub:

- https://hub.docker.com/r/opennetworkinsight/oni-demo/

Initial Source

The Spot codebase is currently hosted on GitHub and will be transitioned to the ASF repositories during incubation. Spot and its submodules are currently licensed under several different licenses.

No trademarks or domain names for Spot have been registered to date, and it will be up to the ASF's discretion to do so. The project's current website at open-network-insight.org will be redirected to spot.incubator.apache.org during incubation.

Some portions of the code are imported from other open source projects under the Apache 2.0, BSD, or MIT licenses.

External Dependencies

The full set of dependencies and licenses are:

- Jupyter: BSD
- D3js: BSD
- Nfdump: BSD
- Wireshark: GNU General Public License version 2
- Apache Hadoop: Apache License 2.0
- Apache Spark: Apache License 2.0
- JQuery: MIT
- ReactJS: BSD
- Bootstrap: MIT

Issues related to GPL dependencies will be resolved during incubation.

Cryptography

Spot does not currently include any cryptography-related code.

Required Resources

Developer and user mailing lists

- private@spot.incubator.apache.org (PMC)
- commits@spot.incubator.apache.org (git push emails)
- issues@spot.incubator.apache.org (JIRA issue feed)
- dev@spot.incubator.apache.org (code reviews plus dev discussion)
- user@spot.incubator.apache.org (user questions)

Repository

- git://git.apache.org/spot

Issue Tracker

We would like to import our current JIRA project into the ASF JIRA, such that our historical commit messages and code comments continue to reference the appropriate bug numbers.

Initial Committers

- Grant Babb
- Ricardo Barona
- Cesar Berho
- Jarek Jarcec Cecho
- Michael Czerny
- Nick Gamb
- Sai Ganji
- Gabriela Lima Garza
- Victor Gonzalez
- Mark Grover
- Morris Hicks
- Ritu Kama
- Austin Leahy
- Ashrith Mekala
- Diego Ortiz
- Sudharshan Rao PakalaSai
- Srinivasa Reddy
- Alan Ross
- Everardo Lopez Sandoval
- Nathan Segerlind
- Vartika Singh
- Nathanael Smith
- Carlos Villavicencio

Affiliations

- Grant Babb: Jask
- Ricardo Barona : Intel
- Cesar Berho: Intel
- Jarek Jarcec Cecho: StreamSets
- Michael Czerny: Cybraics
- Nick Gamb: Centrify
- Sai Ganji: Cloudwick
- Gabriela Lima Garza: Intel
- Victor Gonzalez: Intel
- Mark Grover: Cloudera
- Morris Hicks: Cloudera
- Ritu Kama: Intel
- Austin Leahy: eBay
- Ashrith Mekala: Cloudwick
- Diego Ortiz: Intel
- Sudharshan Rao PakalaSai: Cloudwick
- Srinivasa Reddy: Cloudera
- Alan Ross: Intel
- Everardo Lopez Sandoval: Intel
- Nathan Segerlind: Intel
- Vartika Singh: Cloudera
- Nathanael Smith: Intel
- Carlos Villavicencio: Intel

Sponsors

Champion

- Doug Cutting - Cloudera

Nominated Mentors

- Brock Noland - ASF Member, phData
- Jarek Jarcec Cecho - ASF Member, StreamSets
- Andrei Savu - Cloudera
- Uma Maheswara Rao G - Intel

Sponsoring Entity

The Apache Incubator.