EagleProposal

Eagle

Abstract

Eagle is an Open Source Monitoring solution for Hadoop to instantly identify access to sensitive data, recognize attacks, malicious activities in hadoop and take actions.

Proposal

Eagle audits access to HDFS files, Hive and HBase tables in real time, enforces policies defined on sensitive data access and alerts or blocks user's access to that sensitive data in real time. Eagle also creates user profiles based on the typical access behaviour for HDFS and Hive and sends alerts when anomalous behaviour is detected. Eagle can also import sensitive data information classified by external classification engines to help define its policies.

Overview of Eagle

Eagle has 3 main parts.

1. **Data collection and storage** - Eagle collects data from various hadoop logs in real time using Kafka/Yarn API and uses HDFS and HBase for storage.
2. **Data processing and policy engine** - Eagle allows users to create policies based on various metadata properties on HDFS, Hive and HBase data.
3. **Eagle services** - Eagle services include policy manager, query service and the visualization component. Eagle provides intuitive user interface to administer Eagle and an alert dashboard to respond to real time alerts.

Data Collection and Storage:

Eagle provides programming API for extending Eagle to integrate any data source into Eagle policy evaluation framework. For example, Eagle hdfs audit monitoring collects data from Kafka which is populated from namenode log4j appender or from logstash agent. Eagle hive monitoring collects hive query logs from running job through YARN API, which is designed to be scalable and fault-tolerant. Eagle uses HBase as storage for storing metadata and metrics data, and also supports relational database through configuration change.

Data Processing and Policy Engine:

Processing Engine: Eagle provides stream processing API which is an abstraction of Apache Storm. It can also be extended to other streaming engines. This abstraction allows developers to assemble data transformation, filtering, external data join etc. without physically bound to a specific streaming platform. Eagle streaming API allows developers to easily integrate business logic with Eagle policy engine and internally Eagle framework compiles business logic execution DAG into program primitives of underlying stream infrastructure e.g. Apache Storm. For example, Eagle HDFS monitoring transforms audit log from Namenode to object and joins sensitivity metadata, security zone metadata which are generated from external programs or configured by user. Eagle hive monitoring filters running jobs to get hive query string and parses query string into object and then joins sensitivity metadata.

Alerting Framework: Eagle Alert Framework includes stream metadata API, scalable policy engine framework, extensible policy engine framework. Stream metadata API allows developers to declare event schema including what attributes constitute an event, what is the type for each attribute, and how to dynamically resolve attribute value in runtime when user configures policy. Scalable policy engine framework allows policies to be executed on different physical nodes in parallel. It is also used to define your own policy partitioner class. Policy engine framework together with streaming partitioning capability provided by all streaming platforms will make sure policies and events can be evaluated in a fully distributed way. Extensible policy engine framework allows developer to plugin a new policy engine with a few lines of codes. WSO2 Siddhi CEP engine is the policy engine which Eagle supports as first-class citizen.

Machine Learning module: Eagle provides capabilities to define user activity patterns or user profiles for Hadoop users based on the user behaviour in the platform. These user profiles are modeled using Machine Learning algorithms and used for detection of anomalous users activities. Eagle uses Eigen Value Decomposition, and Density Estimation algorithms for generating user profile models. The model reads data from HDFS audit logs, preprocesses and aggregates data, and generates models using Spark programming APIs. Once models are generated, Eagle uses stream processing engine for near real-time anomaly detection to determine if any user's activities are suspicious or not.

Eagle Services:

Query Service: Eagle provides SQL-like service API to support comprehensive computation for huge set of data on the fly, for e.g. comprehensive filtering, aggregation, histogram, sorting, top, arithmetical expression, pagination etc. HBase is the data storage which Eagle supports as first-class citizen, relational database is supported as well. For HBase storage, Eagle query framework compiles user provided SQL-like query into HBase native filter objects and execute it through HBase coprocessor on the fly.

Policy Manager: Eagle policy manager provides UI and Restful API for user to define policy with just a few clicks. It includes site management UI, policy editor, sensitivity metadata import, HDFS or Hive sensitive resource browsing, alert dashboards etc.

Background

Data is one of the most important assets for today's businesses, which makes data security one of the top priorities of today's enterprises. Hadoop is widely used across different verticals as a big data repository to store this data in most modern enterprises.

At eBay we use hadoop platform extensively for our data processing needs. Our data in Hadoop is becoming bigger and bigger as our user base is seeing an exponential growth. Today there are variety of data sets available in Hadoop cluster for our users to consume. eBay has around 120 PB of data stored in HDFS across 6 different clusters and around 1800+ active hadoop users consuming data thru Hive, HBase and mapreduce jobs everyday to build applications using this data. With this astronomical growth of data there are also challenges in securing sensitive data and monitoring the access to this sensitive data. Today in large organizations HDFS is the defacto standard for storing big data. Data sets which includes and not limited to consumer sentiment, social media data, customer segmentation, web clicks, sensor data, geo-location and transaction data get stored in Hadoop for day to day business needs.

We at eBay want to make sure the sensitive data and data platforms are completely protected from security breaches. So we partnered very closely with our Information Security team to understand the requirements for Eagle to monitor sensitive data access on hadoop:

1. Ability to identify and stop security threats in real time 2. Scale for big data (Support PB scale and Billions of events) 3. Ability to create data access policies 4. Support multiple data sources like HDFS, HBase, Hive 5. Visualize alerts in real time 6. Ability to block malicious access in real time

We did not find any data access monitoring solution that available today and can provide the features and functionality that we need to monitor the data access in the hadoop ecosystem at our scale. Hence with an excellent team of world class developers and several users, we have been able to bring Eagle into production as well as open source it.

Rationale

In today's world; data is an important asset for any company. Businesses are using data extensively to create amazing experiences for users. Data has to be protected and access to data should be secured from security breaches. Today Hadoop is not only used to store logs but also stores financial data, sensitive data sets, geographical data, user click stream data sets etc. which makes it more important to be protected from security breaches. To secure a data platform there are multiple things that need to happen. One is having a strong access control mechanism which today is provided by Apache Ranger and Apache Sentry. These tools provide the ability to provide fine grain access control mechanism to data sets on hadoop. But there is a big gap in terms of monitoring all the data access events and activities in order to securing the hadoop data platform. Together with strong access control, perimeter security and data access monitoring in place data in the hadoop clusters can be secured against breaches. We looked around and found following:

Existing data activity monitoring products are designed for traditional databases and data warehouse. Existing monitoring platforms cannot scale out to support fast growing data and petabyte scale. Few products in the industry are still very early in terms of supporting HDFS, Hive, HBase data access monitoring.

As mentioned in the background, the business requirement and urgency to secure the data from users with malicious intent drove eBay to invest in building a real time data access monitoring solution from scratch to offer real time alerts and remediation features for malicious data access.

With the power of open source distributed systems like Hadoop, Kafka and much more we were able to develop a data activity monitoring system that can scale, identify and stop malicious access in real time.

Eagle allows admins to create standard access policies and rules for monitoring HDFS, Hive and HBase data. Eagle also provides out of box machine learning models for modeling user profiles based on user access behaviour and use the model to alert on anomalies.

Current Status

Meritocracy

Eagle has been deployed in production at eBay for monitoring billions of events per day from HDFS and Hive operations. From the start; the product has been built with focus on high scalability and application extensibility in mind and Eagle has demonstrated great performance in responding to suspicious events instantly and great flexibility in defining policy.

Community

Eagle seeks to develop the developer and user communities during incubation.

Core Developers

Eagle is currently being designed and developed by engineers from eBay Inc. – Edward Zhang, Hao Chen, Chaitali Gupta, Libin Sun, Jilin Jiang, Qingwen Zhao, Senthil Kumar, Hemanth Dendukuri, Arun Manoharan. All of these core developers have deep expertise in developing monitoring products for the Hadoop ecosystem.

Alignment

The ASF is a natural host for Eagle given that it is already the home of Hadoop, HBase, Hive, Storm, Kafka, Spark and other emerging big data projects. Eagle leverages lot of Apache open-source products. Eagle was designed to offer real time insights into sensitive data access by actively monitoring the data access on various data sets in hadoop and an extensible alerting framework with a powerful policy engine. Eagle compliments the existing Hadoop platform area by providing a comprehensive monitoring and alerting solution for detecting sensitive data access threats based on preset policies and machine learning models for user behaviour analysis.

Known Risks

Orphaned Products

The core developers of Eagle team work full time on this project. There is no risk of Eagle getting orphaned since eBay is extensively using it in their production Hadoop clusters and have plans to go beyond hadoop. For example, currently there are 7 hadoop clusters and 2 of them are being monitored using Hadoop Eagle in production. We have plans to extend it to all hadoop clusters and eventually other data platforms. There are 10's of policies onboarded and actively monitored with plans to onboard more use case. We are very confident that every hadoop cluster in the world will be monitored using Eagle for securing the hadoop ecosystem by actively monitoring for data access on sensitive data. We plan to extend and diversify this community further through Apache. We presented Eagle at the hadoop summit in china and garnered interest from different companies who use hadoop extensively.

Inexperience with Open Source

The core developers are all active users and followers of open source. They are already committers and contributors to the Eagle Github project. All have been involved with the source code that has been released under an open source license, and several of them also have experience developing code in an open source environment. Though the core set of Developers do not have Apache Open Source experience, there are plans to onboard individuals with Apache open source experience on to the project. Apache Kylin PMC members are also in the same ebay organization. We work very closely with Apache Ranger committers and are looking forward to find meaningful integrations to improve the security of hadoop platform.

Homogenous Developers

The core developers are from eBay. Today the problem of monitoring data activities to find and stop threats is a universal problem faced by all the businesses. Apache Incubation process encourages an open and diverse meritocratic community. Eagle intends to make every possible effort to build a diverse, vibrant and involved community and has already received substantial interest from various organizations.

Reliance on Salaried Developers

eBay invested in Eagle as the monitoring solution for Hadoop clusters and some of its key engineers are working full time on the project. In addition, since there is a growing need for securing sensitive data access we need a data activity monitoring solution for Hadoop, we look forward to other Apache developers and researchers to contribute to the project. Additional contributors, including Apache committers have plans to join this effort shortly. Also key to addressing the risk associated with relying on Salaried developers from a single entity is to increase the diversity of the contributors and actively lobby for Domain experts in the security space to contribute. Eagle intends to do this.

Relationships with Other Apache Products

Eagle has a strong relationship and dependency with Apache Hadoop, HBase, Spark, Kafka and Storm. Being part of Apache's Incubation community, could help with a closer collaboration among these projects and as well as others. An Excessive Fascination with the Apache Brand Eagle is proposing to enter incubation at Apache in order to help efforts to diversify the committer-base, not so much to capitalize on the Apache brand. The Eagle project is in production use already inside eBay, but is not expected to be an eBay product for external customers. As such, the Eagle project is not seeking to use the Apache brand as a marketing tool.

Documentation

Information about Eagle can be found at https://github.com/eBay/Eagle. The following link provide more information about Eagle http://goeagle.io.

Initial Source

Eagle has been under development since 2014 by a team of engineers at eBay Inc. It is currently hosted on Github.com under an Apache license 2.0 at https://github.com/eBay/Eagle. Once in incubation we will be moving the code base to apache git library.

External Dependencies

Eagle has the following external dependencies.

Basic

- JDK 1.7+
- Scala 2.10.4
- Apache Maven
- JUnit
- Log4j
- Slf4j
- Apache Commons
- Apache Commons Math3
- Jackson
- Siddhi CEP engine

Hadoop

- Apache Hadoop
- Apache HBase
- Apache Hive
- Apache Zookeeper
- Apache Curator

Apache Spark

- Spark Core Library

REST Service

- Jersey

Query

- Antlr

Stream processing

- Apache Storm
- Apache Kafka

Web

- AngularJS
- jQuery
- Bootstrap V3
- Moment JS
- Admin LTE
- html5shiv
- respond
- Fastclick
- Date Range Picker
- Flot JS

Cryptography

Eagle will eventually support encryption on the wire. This is not one of the initial goals, and we do not expect Eagle to be a controlled export item due to the use of encryption. Eagle supports but does not require the Kerberos authentication mechanism to access secured Hadoop services.

Required Resources

Mailing List

- eagle-private for private PMC discussions
- eagle-dev for developers
- eagle-commits for all commits
- eagle-users for all eagle users

Subversion Directory

- Git is the preferred source control system.

Issue Tracking

- JIRA Eagle (Eagle)

Other Resources

The existing code already has unit tests so we will make use of existing Apache continuous testing infrastructure. The resulting load should not be very large.

Initial Committers

- Seshu Adunuthula <sadunuthula at ebay dot com>
- Arun Manoharan <armanoharan at ebay dot com>
- Edward Zhang <yonzhang at ebay dot com>
- Hao Chen <hchen9 at ebay dot com>
- Chaitali Gupta <cgupta at ebay dot com>
- Libin Sun <libsun at ebay dot com>
- Jilin Jiang <jiljiang at ebay dot com>
- Qingwen Zhao <qingwzhao at ebay dot com>
- Hemanth Dendukuri <hdendukuri at ebay dot com>
- Senthil Kumar <senthilkumar at ebay dot com>

Affiliations

The initial committers are employees of eBay Inc.

Sponsors

Champion

- Henry Saputra <hsaputra at apache dot org> - Apache IPMC member

Nominated Mentors

- Owen O'Malley < omalley at apache dot org > - Apache IPMC member, Hortonworks
- Henry Saputra <hsaputra at apache dot org> - Apache IPMC member
- Julian Hyde <jhyde at apache dot org> - Apache IPMC member, Hortonworks
- Taylor Goetz < ptgoetz at apache dot org> - Apache IPMC member
- Amareshwari Sriramdasu <amareshwari at apache dot org> - Apache IPMC member

Sponsoring Entity

We are requesting the Incubator to sponsor this project.