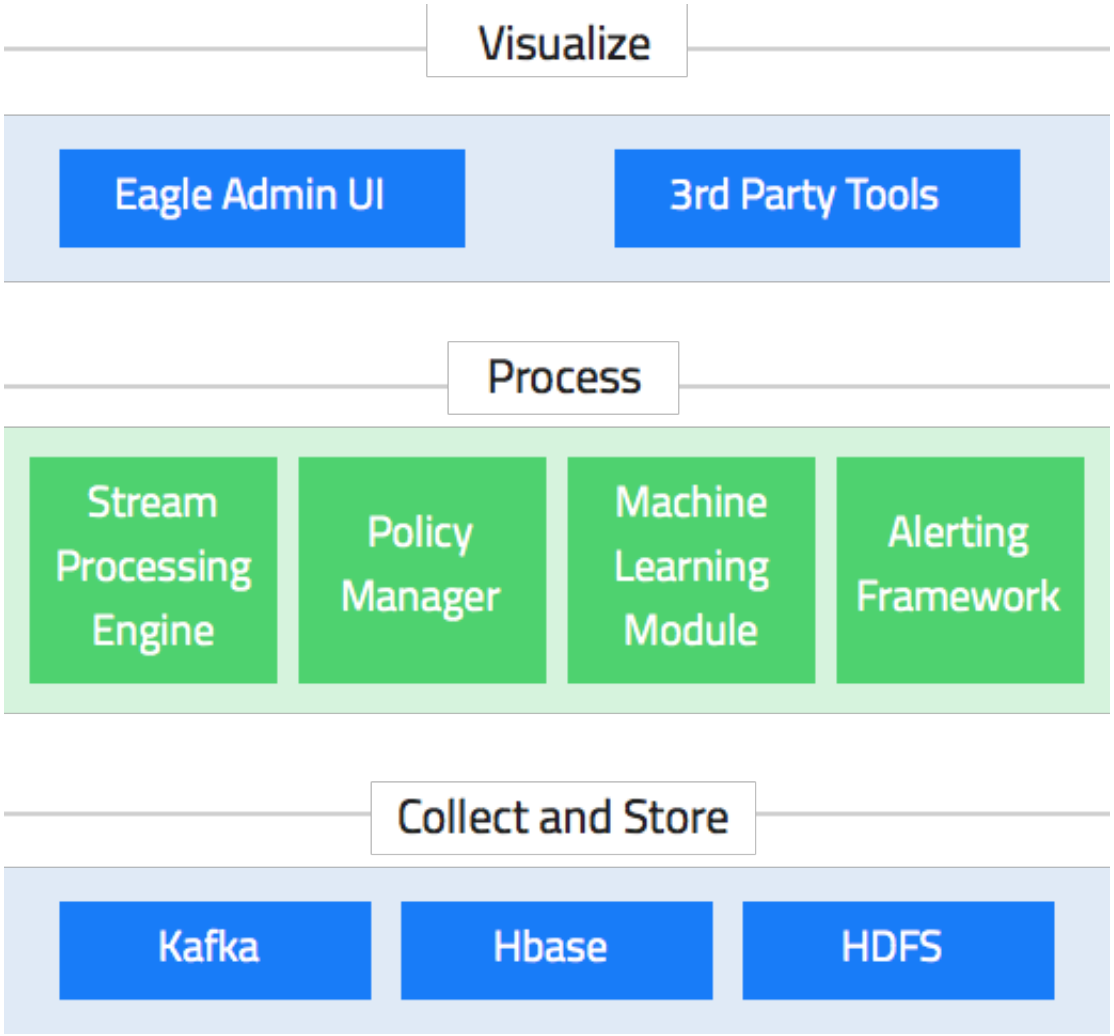
Apache Eagle: 分布式实时Hadoop数据安全方案

Apache Eagle 是由 eBay 公司开源的一个识别大数据平台上的安全和性能问题的开源解决方案。该项目于2017年1月10日正式成为 Apache 顶级项目。 Apache Eagle 提供一套高效分布式的流式策略引擎，具有高实时、可伸缩、易扩展、交互友好等特点，同时集成机器学习对用户行为建立Profile以实现实时智能实时地保护 Hadoop 生态系统中大数据的安全。
Apache Eagle 主要包括三大层：

- 数据收集及存储层（Data Collection and Storage）
- 数据处理层（Data Processing）
- 可视化层（Visualize）

整个组成如下：

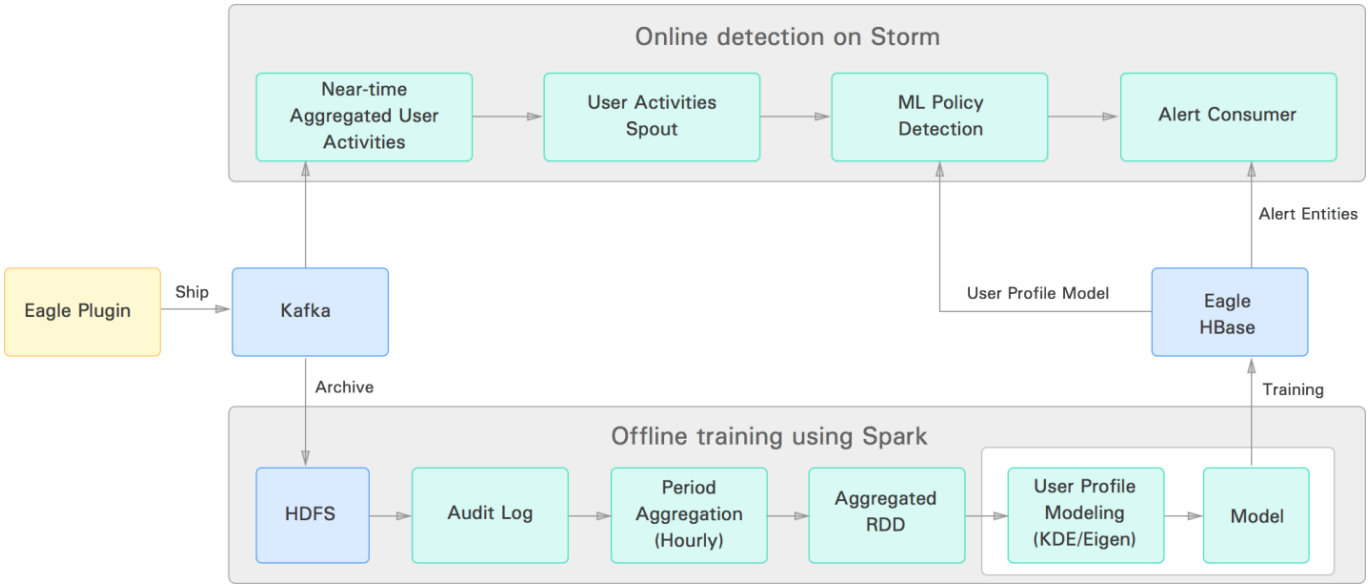


如果想及时了解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

Apache Eagle 依赖于 Apache Storm 来进行数据活动和操作日志的流处理，并且可以执行基于策略的检测和报警。它提供多个API：作为基于Storm API上的一层抽象的流式处理API和 policy engine provider API的抽象，它将WSO2的开源Siddhi CEP engine作为第一类对象。Siddhi CEP engine支持报警规则的热部署，并且警报可以使用属性过滤和基于窗口的规则（例如，在10分钟内三次以上的访问）来定义。

Eagle 支持根据用户在Hadoop平台上历史使用行为习惯来定义行为模式或用户Profile的能力。拥有了这个功能，不需要在系统中预先设置固定临界值的情况下，也可以实现智能地检测出异常的行为。Eagle中用户Profile是通过机器学习算法生成，用于在用户当前实时行为模式与其对应的历史模型模式存在一定程度的差异时识别用户行为是否为异常。目前，Eagle
内置提供以下两种算法来检测异常，分别为特征值分解（Eigen-Value Decomposition）和
密度估计（Density Estimation）。这些算法从HDFS
审计日志中读取数据，对数据进行分割、审查、交叉分析，周期性地为每个用户依次创建Profile行为模型。一旦模型生成，Eagle的实时流策略引擎能够近乎实时地识别出异常，分辨当前用户的行为可疑的或者与他们的历史行为模型不相符。

下图简单描述了目前Eagle中用户Profile的离线训练建模和在线实时监测的数据流：



如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

Apache Eagle 官方网址：https://eagle.apache.org/