ES 在电子银行的安全应用

数据驱动安全 *Data Driven Security*

Gavin Lee

2016 年 12 月

关于

- 关于我
  - Technical Director @HanSight
  - Apache CloudStack Committer
  - HCACD (Hadoop 2.0)
- 关于话题
  - 电子银行日志安全分析
  - Elasticsearch 在此场景中的使用

话题概要

现状

问题

改进

未来……

现状

各种高等级的安全设备及应用

- Firewall
- WAF
- IDS/IPS
- SIEM/SOC

安全团队运维 【7 x 24】

问题 -- 力不从心

- 海量垃圾事件（FW > 90%、SOC > 70%、……）
- 运维疲于奔命
- 分析及回溯内容有限
- 时效性不强
- 潜在未知威胁

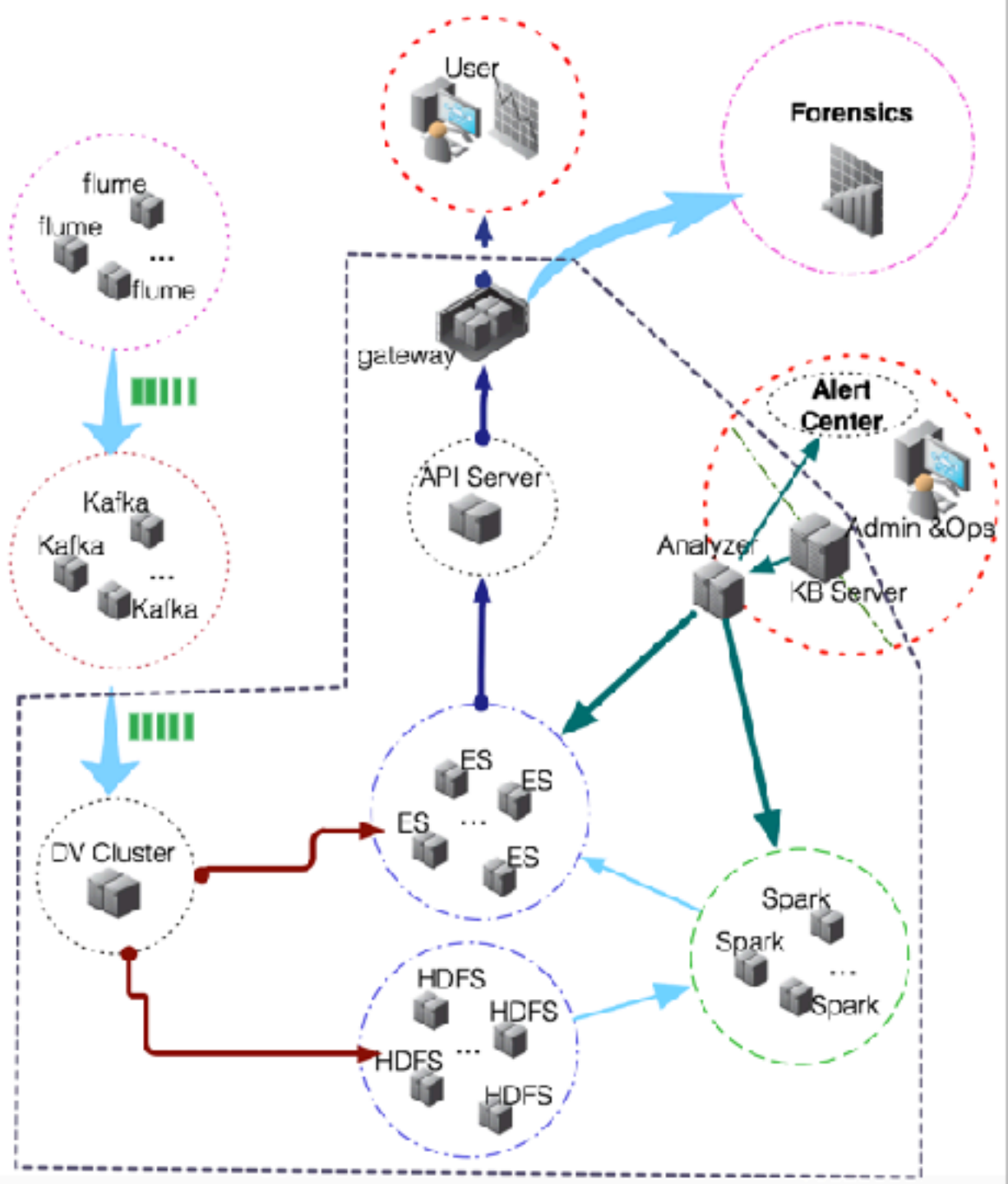
改进

我们的做法

- 大数据一体化平台（ES + Spark +HDFS）
  - 准实时分析（ES）
  - 长周期离线分析（Spark + HDFS）
  - 未知威胁及异常告警（Spark + ES）
    - 机器学习
    - 规则匹配
    - 人工

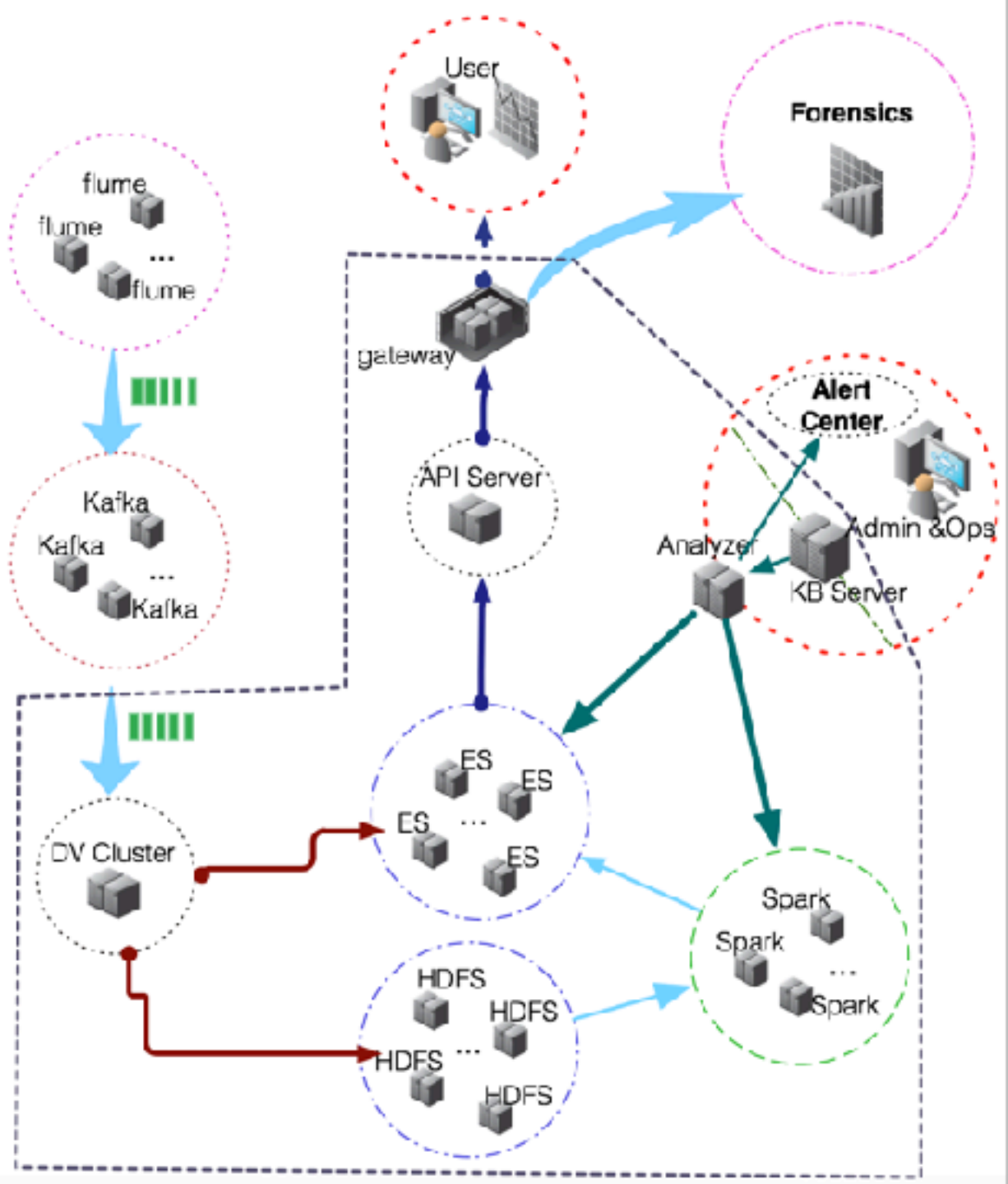
数据特征及架构

- 数据源
  - 电子银行全量访问日志，包括 Web 应用和交易日志
  - 每天 ~500G
  - 实时分析保存近一个月日志
  - 长周期保存近半年日志
- DV Cluster
  - 分布式 ETL
  - 字段丰富化（IP Geo，身份证，手机归属地，设备类型及指纹等）
- Analyzer
  - 准实时分析告警
  - 通过建模数据，固定规则，人工相结合手段产生告警
  - 结果送入统一告警平台（Alert Center）
  - 支撑恶意行为画像及报告



数据特征及架构--续

- ES（1.5.2）
  - 查询及简单聚合
  - 保存近一个月全量日志
  - 告警及事件回溯
  - 准实时图表展现
- Spark（1.5.1 *）
  - 读取 HDFS 历史数据进行长周期建模
  - 模型数据回写 ES
- HDFS（Hadoop 2.5.2）
  - 长周期数据存储
  - Spark 建模读取丰富化后的日志用于建模
- KB Server -- 知识库
  - IP/DNS/URL/手机号等的黑、白名单
  - 规则库（基于人工，建模等产生）

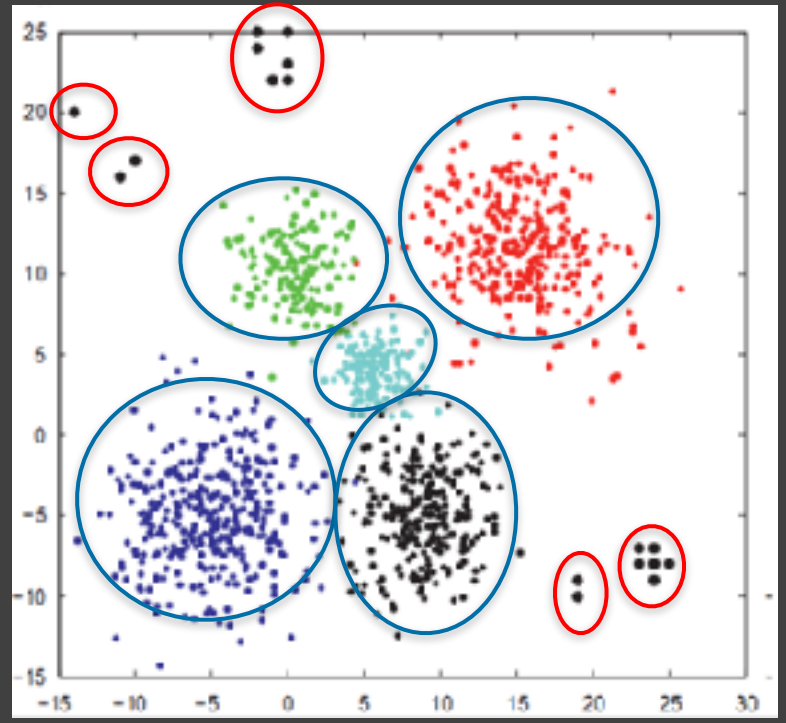


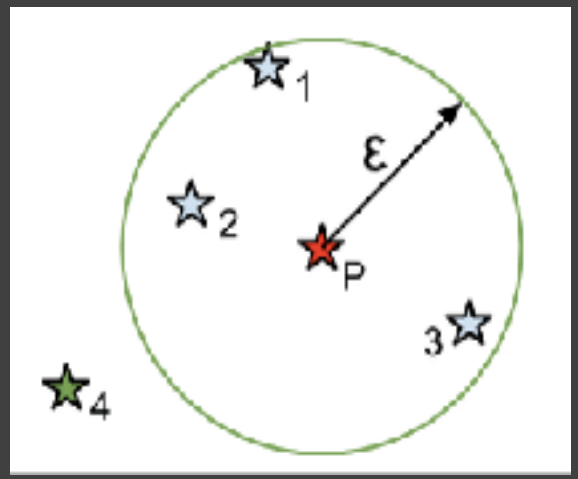
算法浅析 -- 建模及规则

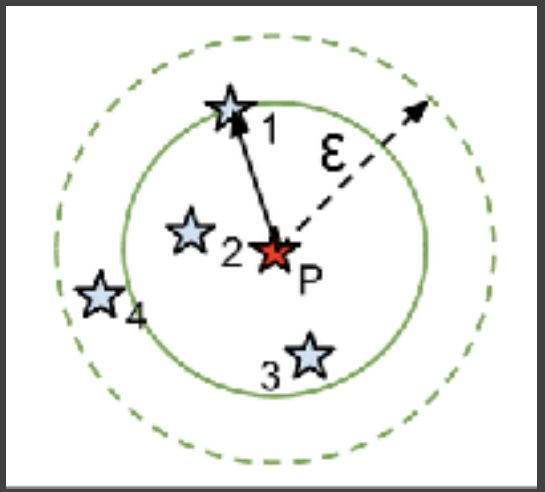
- 聚类算法（无监督，未知威胁，离线建模）
  - K-Means
  - OPTICS
- 时间序列异常算法（预测敏感 URL访问情况，准实时）
  - ARiMA
  - 基线分析（阀值告警）
- 图分析（发现特征关联关系，辅助人工查看，离线）
  - RocksDB

聚类算法：K 均值及 OPTICS

- 输入
  - 全量访问 URL
  - 请求响应状态码
- 算法核心
  - 根据状态码分类 URL
  - 设定不同状态码下各自的 K 值 （K-Means）
  - 设定 e 和 minpts（OPTICS）
  - 同一状态码下不同 URL 依据距离函数聚类
- 输出
  - 不同簇下的 URL
  - 通过调整参数及设定阀值输出模型数据
- 算法应用问题
  - 距离函数选择
  - URL 是否规整
  - K-Means 参数 K 敏感度高

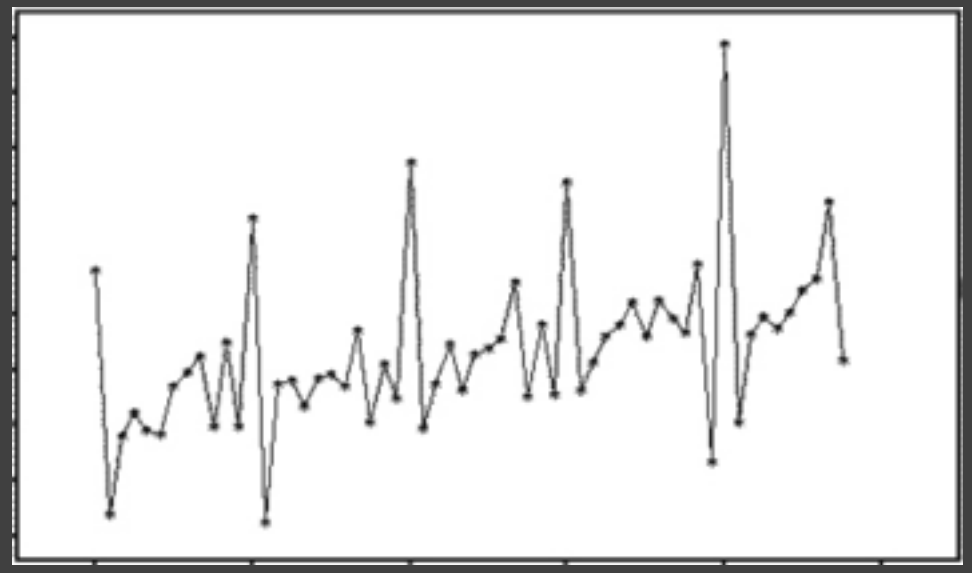


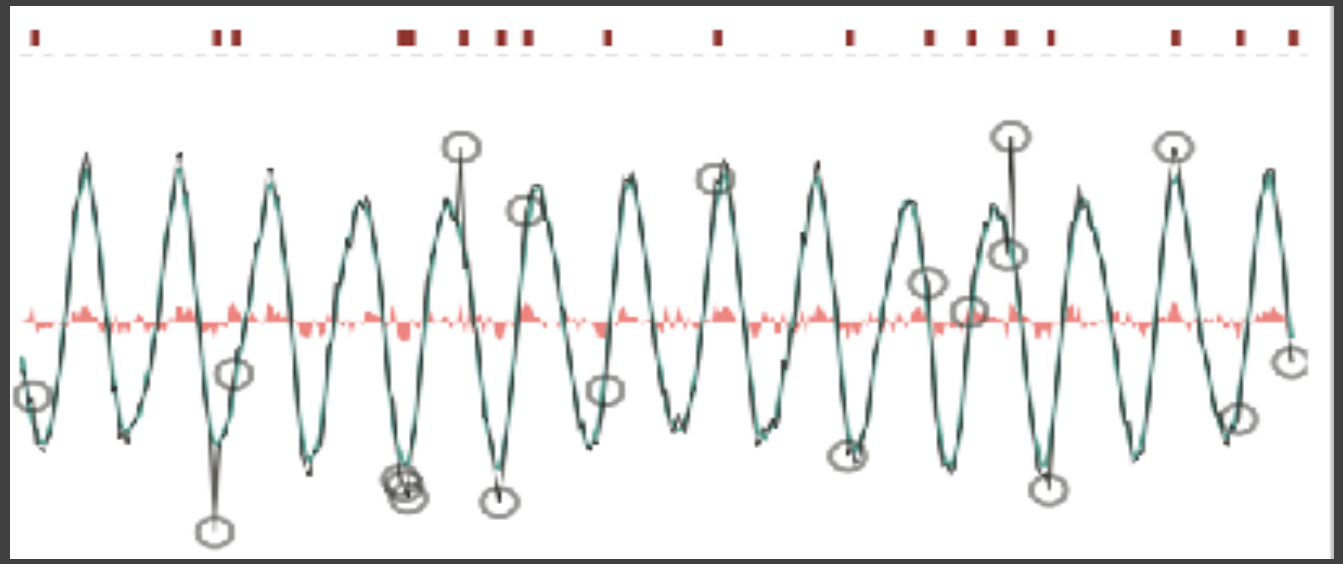




访问异常算法：ARiMA，基线访问

- 输入
  - 过去三周访问的敏感URL
- 算法核心
  - 对于每一个 URL
  - 以每五分钟访问频次计数，每天为单位形成访问向量
  - 过去三周可形成：x | y =>(288 * 21) | $count
  - 使用 SVD 进行奇异值计算
  - 调整迭代周期及阀值
- 输出
  - 异常值及其时段
  - 图表展现
- 算法应用问题
  - 效率
  - 突发情况（双11，电商节等）





请求参数异常：信息熵

- 输入
  - 敏感URL在一段时间的访问情况
  - URL的请求参数
  - 请求 IP
- 算法核心
  - 对于参数中出现不同的字符、频次设置不同的权重和分值
  - 对于 Key-Value 中不同的 Key 和 Value 设置不同的分值
  - 结合URL 及 IP 的情报库
- 输出
  - 异常 访问 IP
- 算法应用问题
  - 误报
  - 参数规整

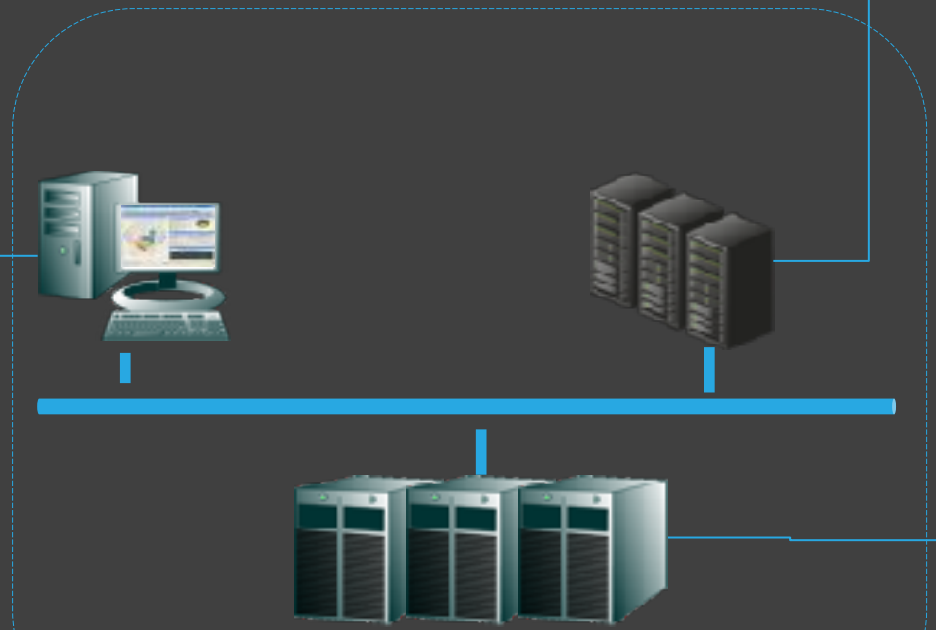
ES物理部署示意图

Elasticsearch 集群部署架构

- Master：3
- Client：3
- Data：15

系统应用管理服务器（例）

- 计算和存储＊2台
  - ✓ 4核
  - ✓ 16GB
  - ✓ 200G存储



日志采集服务器集群（例）

- 计算和存储＊4台

✓ 16核
✓ 16GB
✓ 200G存储

- 性能

✓ 每天可采集1.5T数据；
✓ 单台采集器性能~20000EPS；

历史日志检索分析服务器集群

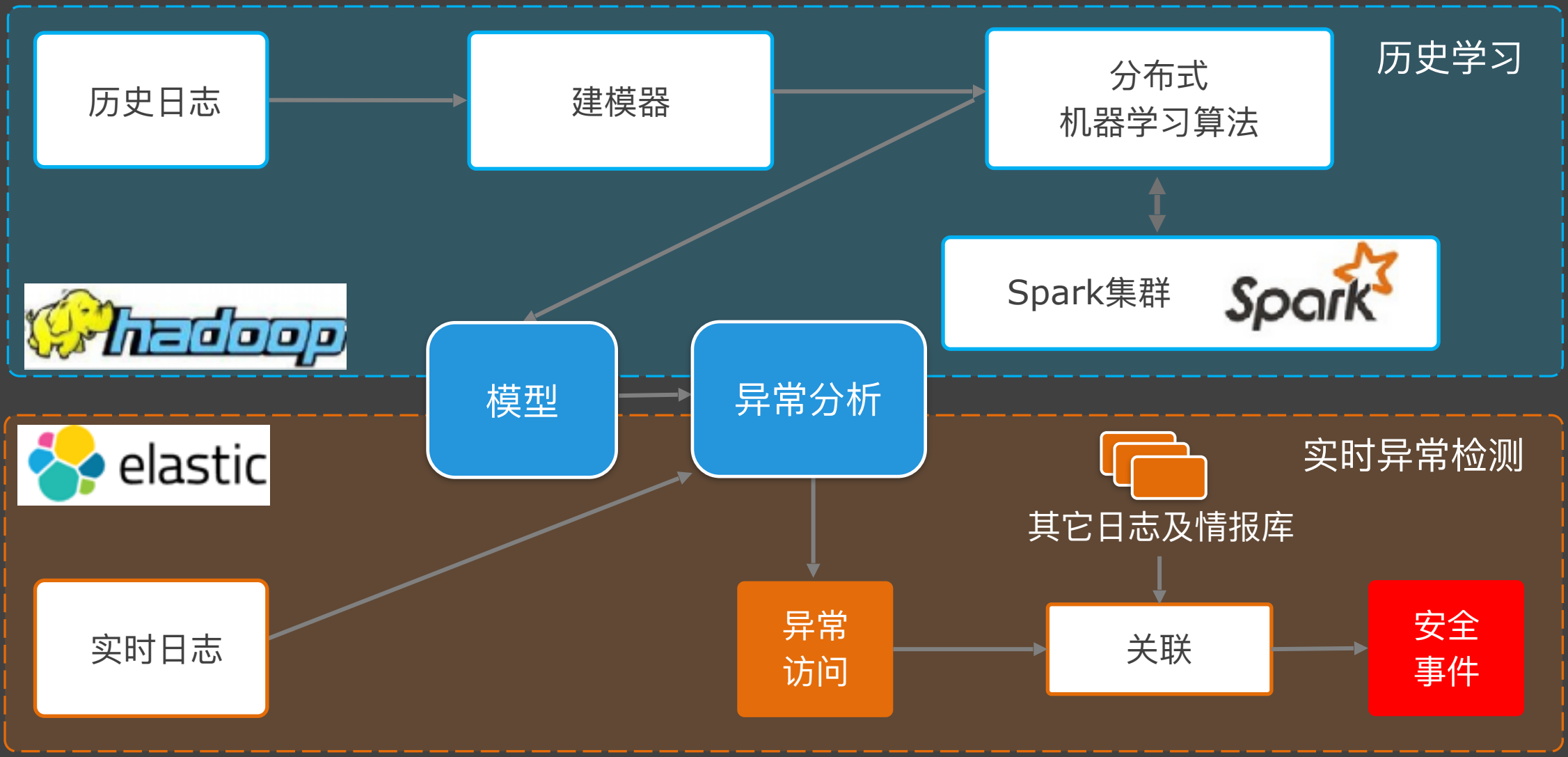
- 计算和存储＊21台

✓ 8核
✓ 64GB
✓ 3*1TB存储

- 性能

✓ 可存储40T数据，30天滚动窗口；
✓ 1天~600G数据，搜索返回时间秒级；
✓ 1周5T数据，搜索聚合返回3秒左右；

系统异常检测流程图



使用场景

- 恶意扫描（基于 Web 访问日志）
- 撞库分析（基于交易日志）
- 养号分析（基于交易日志）
- 恶意行为画像（全量日志）
- 推荐告警 IP（及设备指纹）
- 全球告警地图展现
- 告警回溯/下钻

ES 的使用

- 数据存储基础
  - 规整后的原始日志
  - 模型数据
  - 告警数据
  - 部分情报数据
- 使用的功能
  - 准实时报表展现
  - 多级聚合查询
  - Spark on ES 分析， elasticsearch-hadoop
  - 告警溯源

ES 使用中的林林总总

- 大量聚合查询导致的 OOME
- 磁盘满导致的错误
- Spark On ES 导致 ES 宕机
- 字段引起的 Mapping 错误（简单/复合，时间，IP 等）
- 入库性能调优（当前峰值~30000EPS，数据节点15）
- Index 策略与业务的关系
- 热数据和冷数据（冷数据→HDFS，热数据→多份复本）
- 物理机向虚拟化平台迁移
- 安全性（防火墙，Kerberos，隔离）

未来

下一步……

- 升级（ES 2.3、Spark 2.0）
- 增加 Spark-Streaming 流计算
- 集成并自定义 Kibana 4
- 更多建模分析场景
- 对接反欺诈平台
- 丰富情报库

http://www.hansight.com/hr.html

jobs@HanSight.com

谢谢！ HanSight 瀚思

www.HanSight.com

微信公众号：瀚思安信

北京市海淀区中关村软件园9号楼2区306A