Geekbang 极客邦科技 InfoQ

QCon 全球软件开发大会【北京站】2016

Apache Eagle: eBay构建开源分布式实时预警平台实践

陈浩

http://people.apache.org/~hao

Intemational Software Development Conference

Eagle

个人简介 陈浩

Apache Eagle 联合发起人,PMC & Committer
hao@apache.org

eBay分析平台基础架构部门高级工程师
hchen9@ebay.com

全球Hadoop峰会（SJC/SHA/BJ/SZ）特邀讲师
http://people.apache.org/~hao

诸多开源项目贡献者
https://github.com/haoch

议程

- Eagle简介
- Eagle技术架构
- Eagle应用场景
- 关于开源
- Q & A

Including Apache Eagle and More

Marketplace

http://www.ebay.com

291M
MOBILE APP DOWNLOADS GLOBALLY

162M
ACTIVE BUYERS
Q4 2015

800M
ACTIVE LISTINGS
Q4 2015

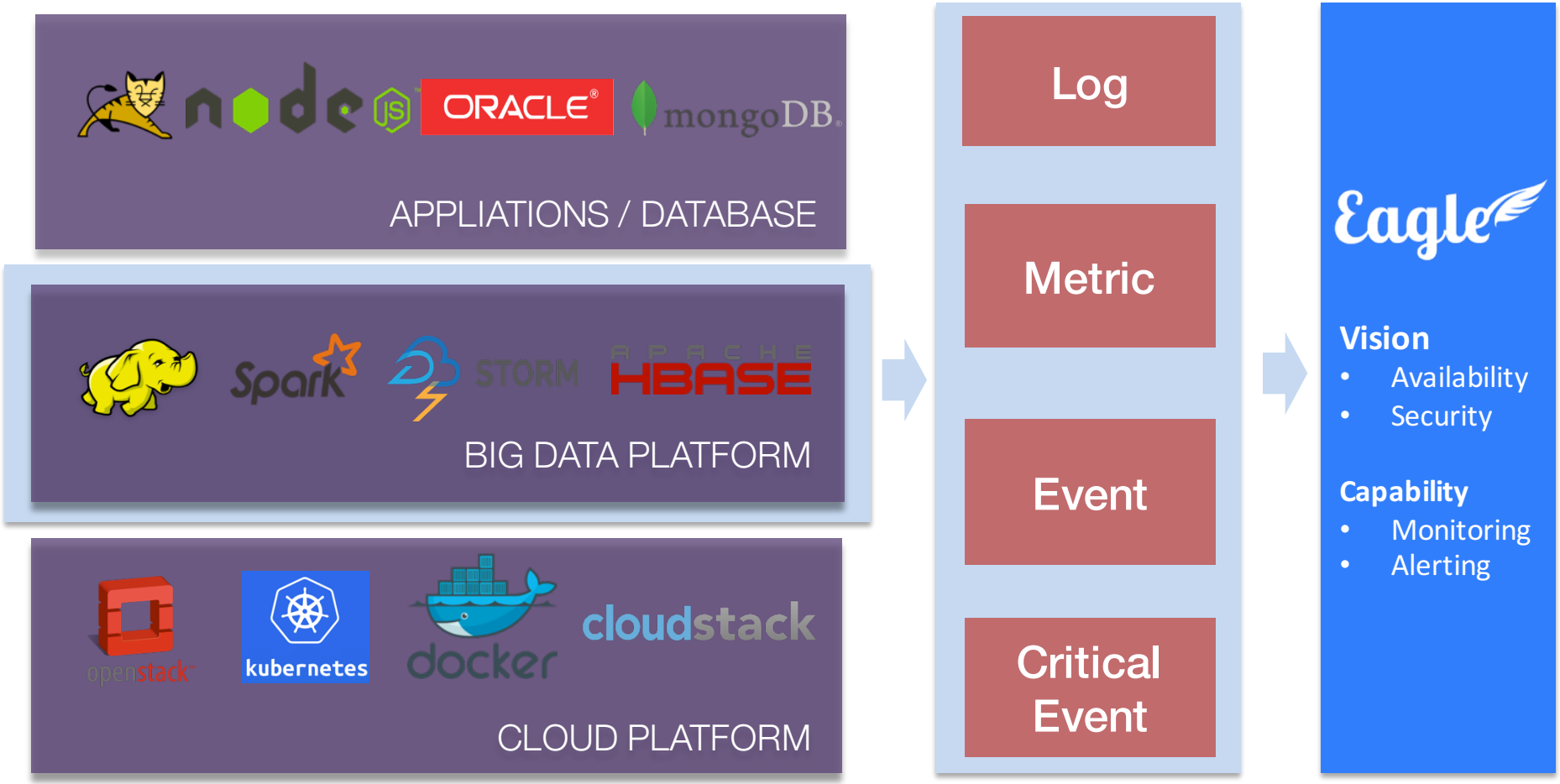
1.3B
LISTINGS CREATED VIA MOBILE EVERY WEEK

25M
ACTIVE SELLERS

8.8M
NEW LISTINGS EVERY WEEK

42%
GMV VIA MOBILE

可信赖的商业基础架构平台



Apache Eagle


是一个由eBay开源并贡献给Apache基金会的分布式实时监控和预警引擎

于2015年10月26日 被接受入Apache 孵化器 http://incubator.apache.org/

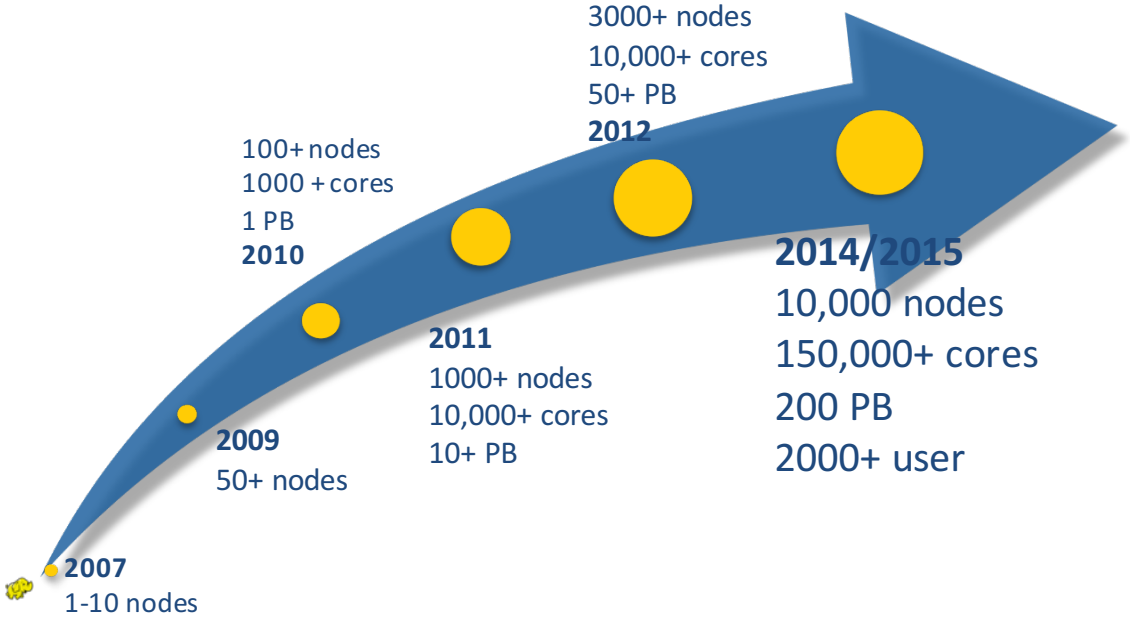
Real-time Anomaly Detection in Big Data Security 作为第一个组件目前主要专注于实时地保证Hadoop的数据安全，致力于通过提供一套相对完善的数据活动监控解决方案，以实现快速地识别敏感数据访问、检测攻击或者恶意行为，并期望在能够及时的触发修复措施

了解更多请参考
http://eagle.incubator.apache.org 或 http://github.com/apache/incubator-eagle

为何构建Eagle？

Eagle was initialized by end of 2013 for hadoop ecosystem monitoring as any existing tool like zabbix, ganglia can not handle the huge volume of metrics/logs generated by hadoop system in eBay.

Hadoop @ eBay Inc



Hadoop Data
- Security
- Activity

Hadoop Platform
- Heath
- Availability
- Performance

主要面临的挑战

1. 海量监控数据分布式实时处理和存储
   - 大量数据吞吐 (实时数据收集以及流处理IO复杂度)
   - 复杂的异常检测规则 (流处理计算复杂度，Window等内存空间复杂度)
   - 多类型semi-structure 数据存储和查询 (Event, Log, Metric)
2. 动态预警策略(Policy)和动态关联模型
   - 复杂可描述的策略模型 (SQL on Streaming)
   - 动态的关联规则 (动态流式Sort, GroupBy, Join, Window)
   - 机器学习预警模型
3. Hadoop生态系统集成
   - 动态实时数据源管理(动态管理Kafka Topic)
   - Hadoop原生预警策略(Policy)集合
4. 多租户平台支持
   - 资源调度与隔离（独立可用级别保证）
   - 高可用与扩展（弹性的资源管理）

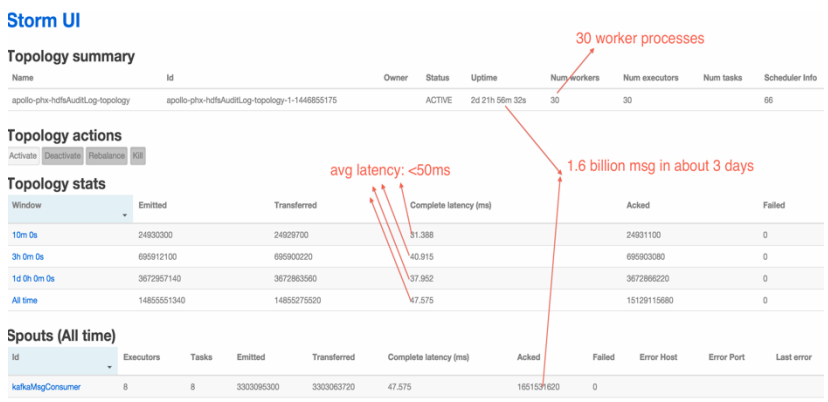
Nagios

Ganglia

Apache Ambari

Eagle @ eBay Inc.

7+	CLUSTERS
10000+	NODES
200+ PB	DATA
10 B+	EVENTS / DAY
500+	METRIC TYPES
50,000+	JOBS / DAY
50,000,000+	TASKS / DAY

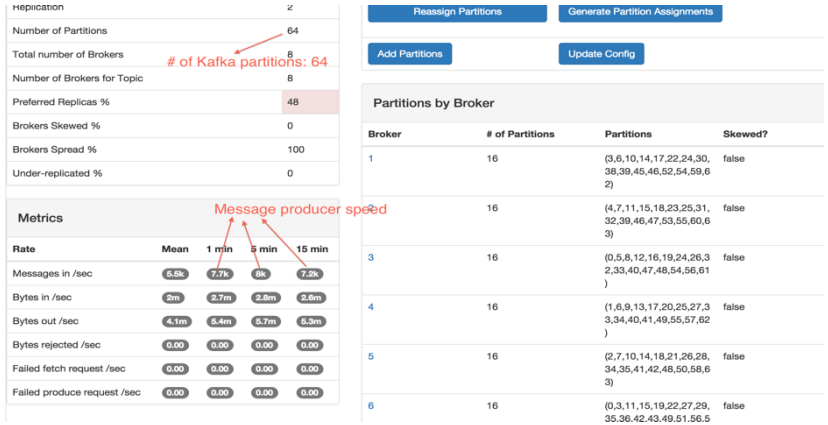


Eagle 在eBay的部署环境

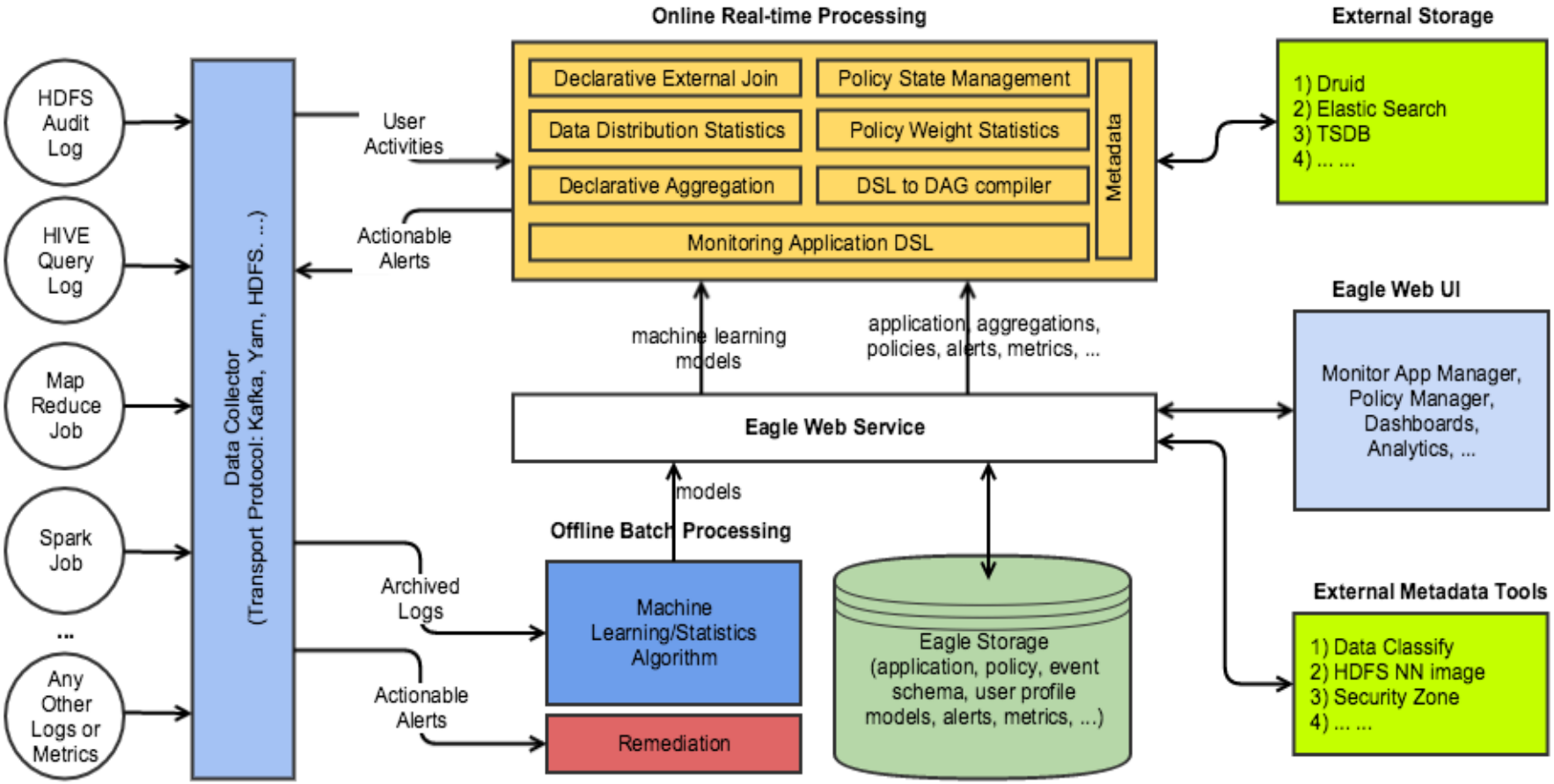
- 近百条安全策略
- 8 物理主机
- 30 工作进程
- 64 kafka分区

Eagle 性能

- 平均延迟（Latency）：~ 50 ms
- 单集群最大吞吐量（Throughput）：300 k /s

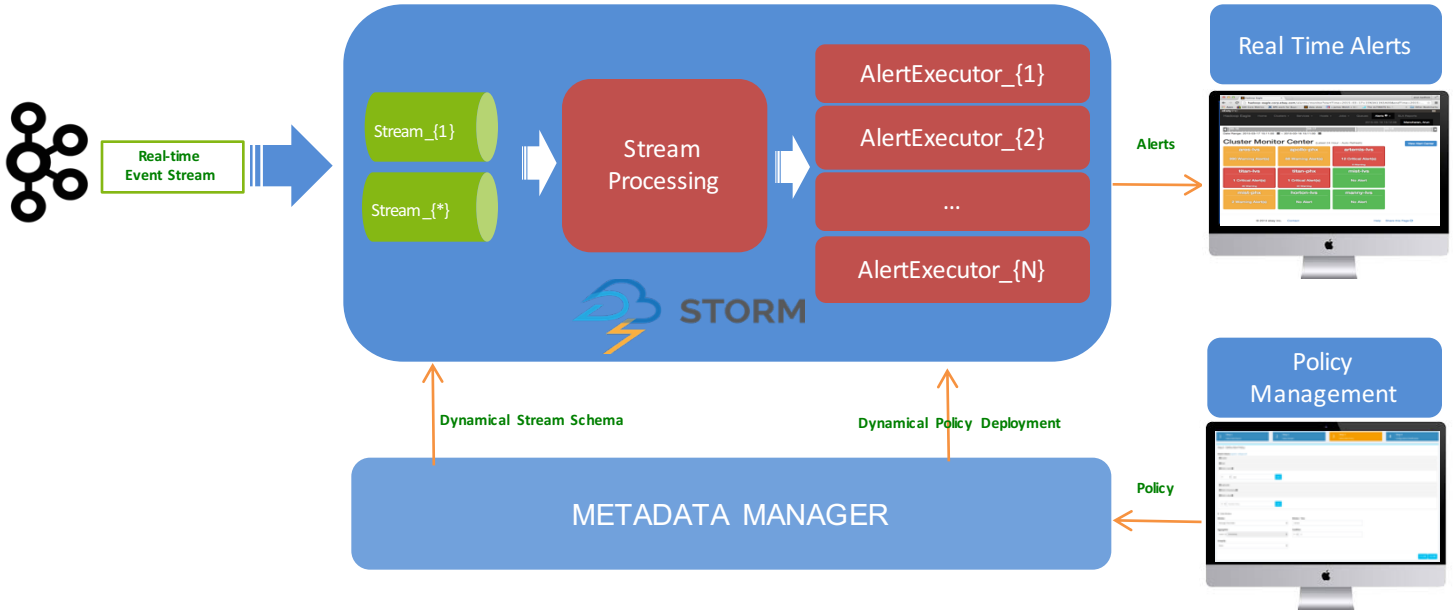


Eagle 架构概览


Online Real-time Processing
External Storage
HDFS Audit Log
HIVE Query Log
Map Reduce Job
Spark Job
...
Any Other Logs or Metrics
Data Collector (Transport Protocol: Kafka, Yarn, HDFS. ...)
User Activities
Actionable Alerts
Declarative External Join
Policy State Management
Data Distribution Statistics
Policy Weight Statistics
Declarative Aggregation
DSL to DAG compiler
Monitoring Application DSL
Metadata
1) Druid
2) Elastic Search
3) TSDB
4) ... ...
machine learning models
application, aggregations, policies, alerts, metrics, ...
Eagle Web UI
Eagle Web Service
Monitor App Manager, Policy Manager, Dashboards, Analytics, ...
models
Offline Batch Processing
Archived Logs
Actionable Alerts
Machine Learning/Statistics Algorithm
Remediation
Eagle Storage (application, policy, event schema, user profile models, alerts, metrics, ...)
External Metadata Tools
1) Data Classify
2) HDFS NN image
3) Security Zone
4) ... ...

分布式策略(Policy)引擎

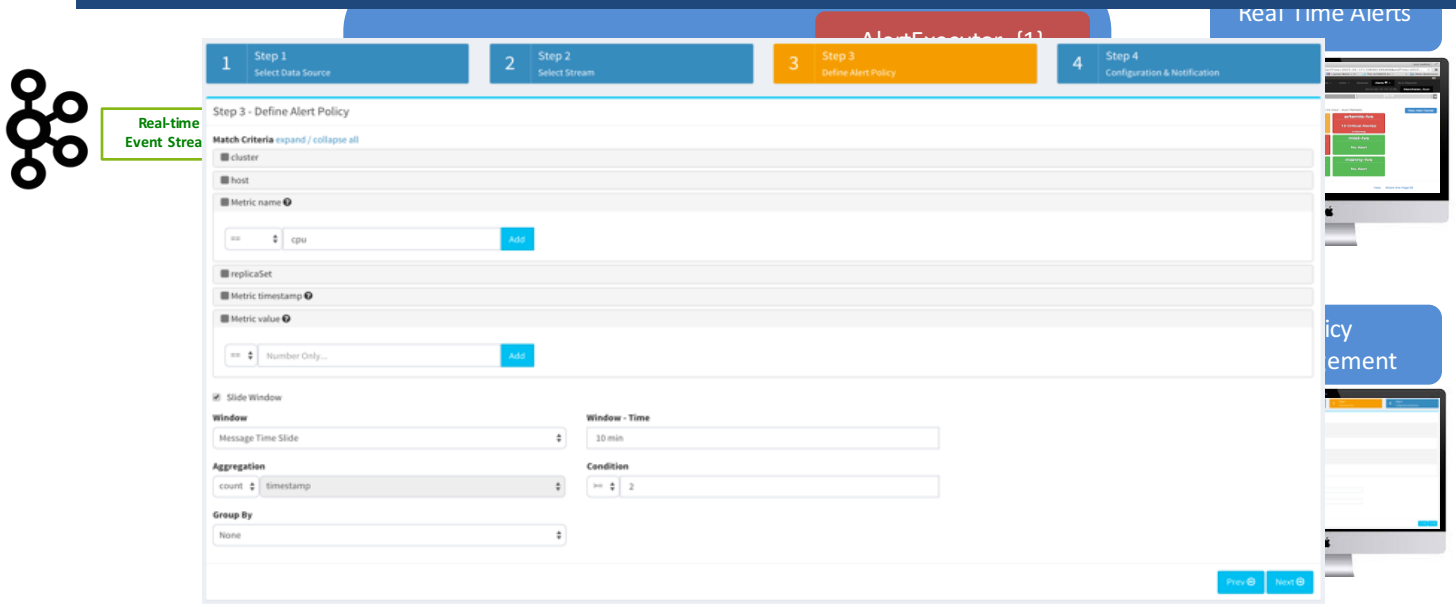
- **实时流处理(Streaming):** Apache Storm (Execution Engine) + Kafka (Message Bus)
- **描述性预警策略(Policy):** SQL (CEP) on Streaming + 动态部署管理
- **线性扩展(Scalability):** 数据量扩展＋计算扩展
- **元数据驱动(Metadata):** Schema 管理和动态协同



分布式策略(Policy)引擎

- **实时流处理(Streaming):** Apache Storm (Execution Engine) + Kafka (Message Bus)
- **描述性预警策略(Policy):** SQL (CEP) on Streaming + 动态部署管理
- **线性扩展(Scalability):** 数据量扩展＋计算扩展
- **元数据驱动(Metadata):** Schema 管理和动态协同

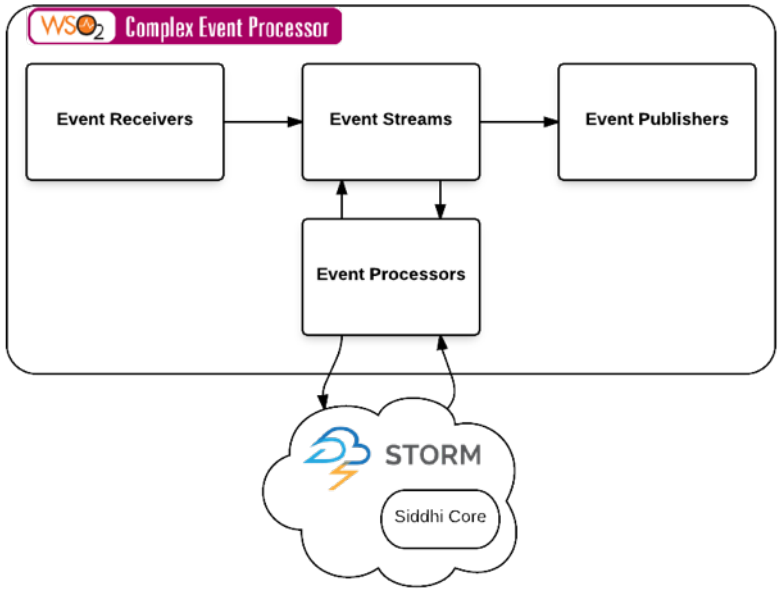
```
from MetricStream[(name == 'ReplLag') and (value > 1000)]
select * insert into outputStream;
```



描述性策略(Policy)规则

```
from MetricStream[(name == 'ReplLag') and (value > 1000)] select * insert into outputStream;
```

基于SQL的分布式流式处理:
Siddhi CEP + Storm by default



- **Filter**
- **Join**
- **Aggregation**: Avg, Sum , Min, Max, etc
- **Group** by
- **Having**
- **Stream handlers for window**: TimeWindow, Batch Window, Length Window
- **Conditions and Expressions**: and, or, not, ==,!=, >=, >, <=, <, and arithmetic operations
- **Pattern Processing**
- **Sequence processing**
- **Event Tables**: intergrate historical data in realtime processing
- **SQL-Like Query**: Query, Stream Definition and Query Plan compilation

https://github.com/wso2/siddhi

描述性策略(Policy)规则 – 示例

示例 1: Alert if hadoop namenode capacity usage exceed 90 percentages

```
from hadoopJmxMetricEventStream
[metric == "hadoop.namenode.fsnamesystemstate.capacityused" and value > 0.9]
select metric, host, value, timestamp, component, site insert into
alertStream;
```

示例 2: Alert if hadoop namenode HA switches

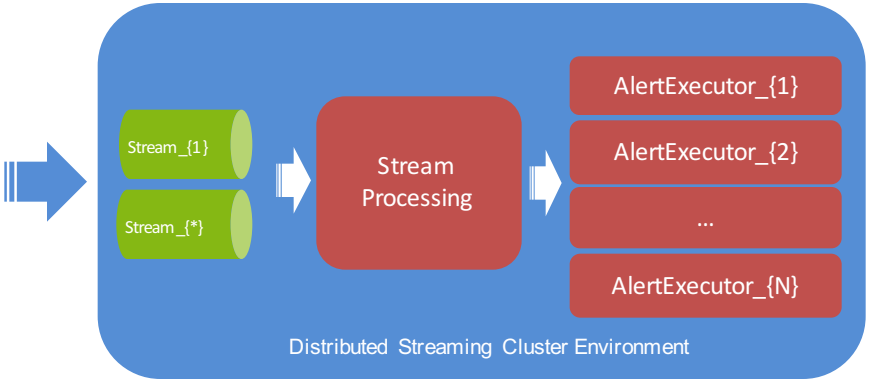
```
from every
a =
hadoopJmxMetricEventStream[metric=="hadoop.namenode.fsnamesystem.hastate"]
->
b = hadoopJmxMetricEventStream[metric==a.metric and b.host == a.host and
a.value != value)]
within 10 min
select a.host, a.value as oldHaState, b.value as newHaState, b.timestamp as
timestamp, b.metric as metric, b.component as component, b.site as site
insert into alertStream;
```

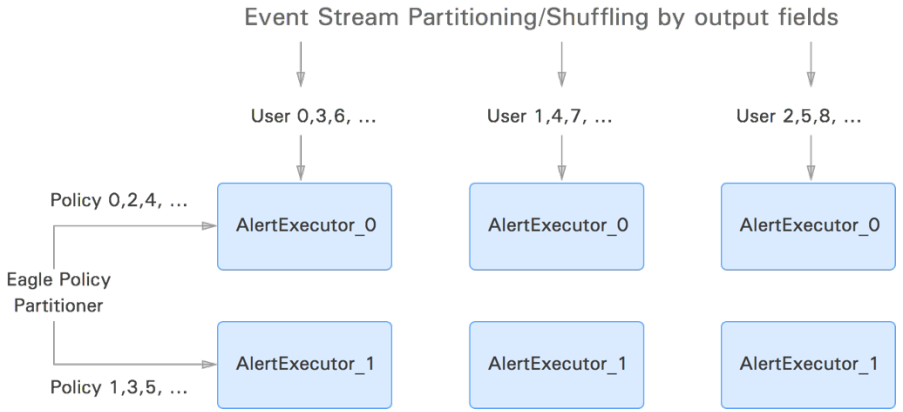
分布式策略(Policy)引擎 – 伸缩性

线性伸缩原理

动态分流{Event}和{Policy}

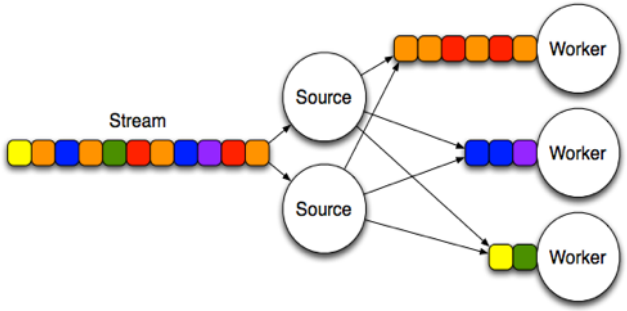
- *N Users with 3 partitions, M policies with 2 partitions, then 3*2 physical tasks*
- *Physical partition + policy-level partition*

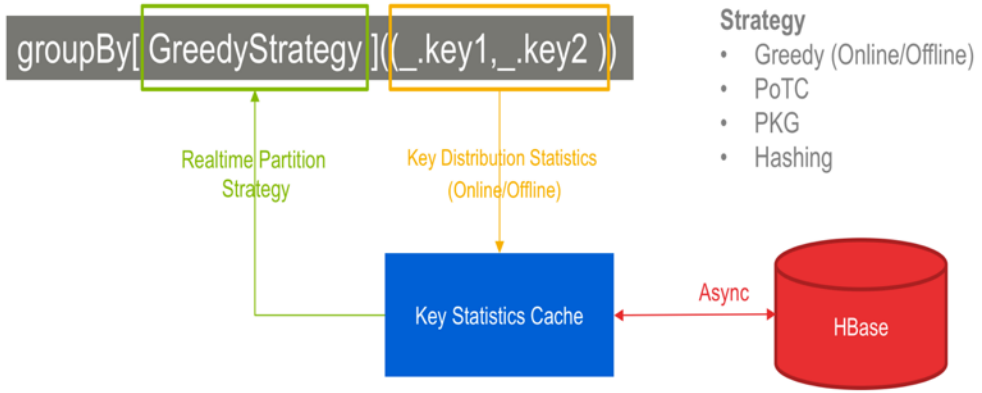




分布式策略(Policy)引擎 – 动态平衡

分流不均衡问题
https://en.wikipedia.org/wiki/Partition_problem





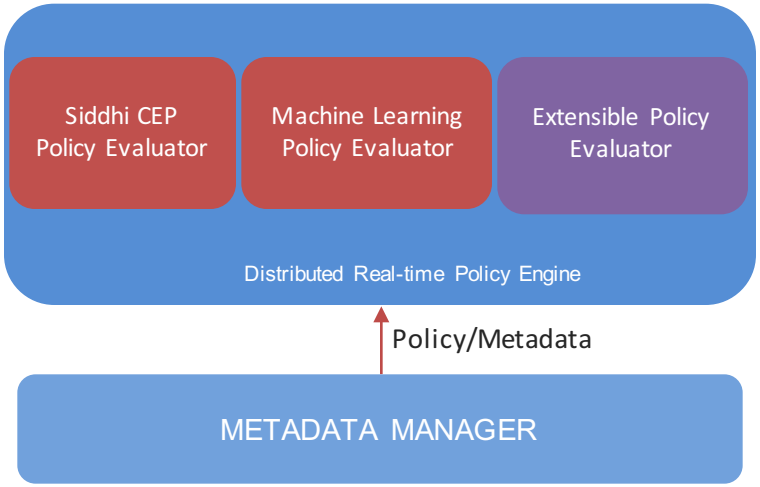
Stream Partition Skew (15:1)

Algorithm	Weights of Executors By Partition User							
Random	0.0484	0.152	0.3535	0.105	0.203	0.072	0.042	0.024
Greedy	0.0837	0.0837	0.0837	0.0837	0.0737	0.0637	0.0437	0.0837

分布式策略(Policy)引擎 – 扩展性

策略引擎的扩展性体现在：

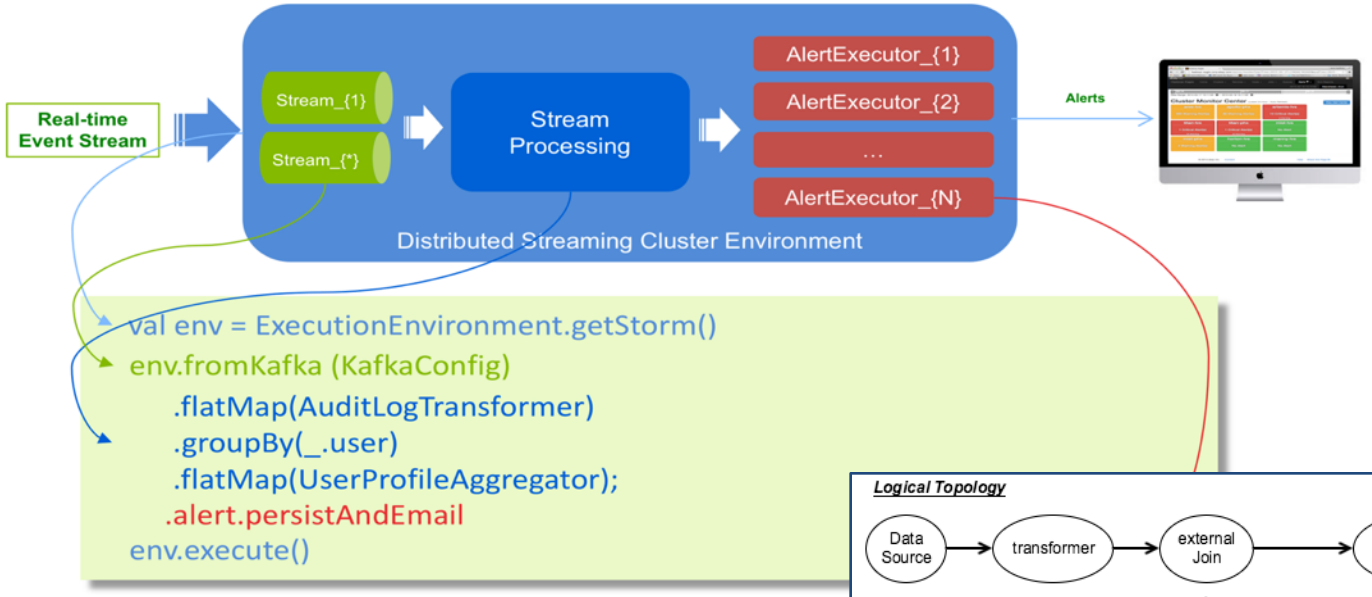
- WSO2 Siddhi CEP as first citizen
- Extensible Policy Engine Implementation
- Extensible Policy Lifecycle Management
- Metadata-based Module Management



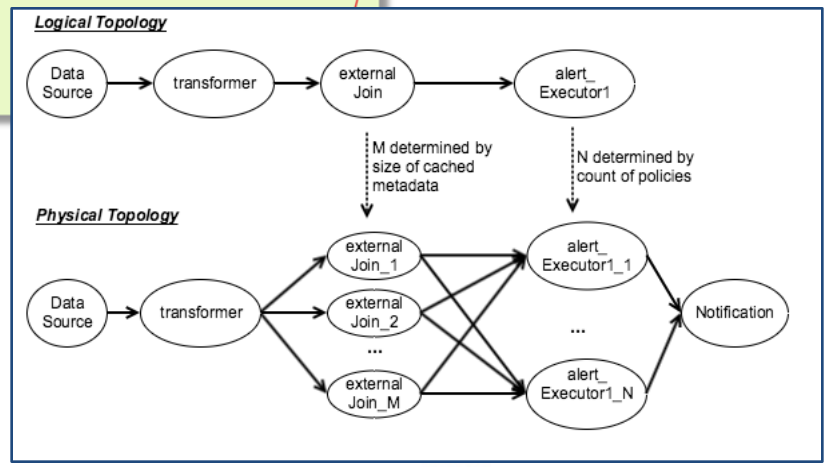

```
public interface PolicyEvaluatorServiceProvider {
     public String getPolicyType(); // literal string to identify one type of policy
     public Class getPolicyEvaluator(); // get policy evaluator implementation
     public List getBindingModules(); // policy text with json format to object mapping
}
```

平台独立的流式处理框架

通过高度抽象的Eagle DSL调用Alert Library



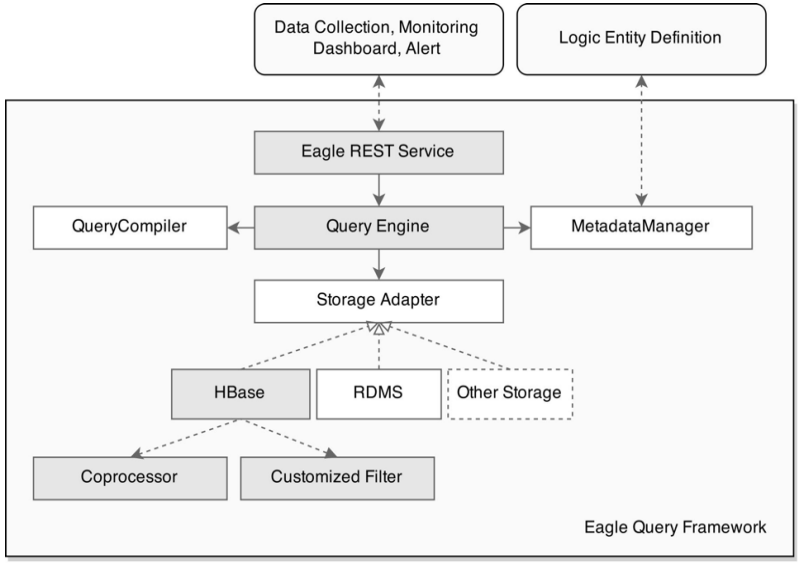
1. Development　　2. Optimization　　3. Compile to native app



海量监控数据存储和查询框架

- 轻量级ORM框架支持HBase和RDBMS
- 功能强大的统一类描述性REST WebService 接口
- 针对监控数据特征进行了特别优化的存储结构和Rokwey
- 提供原生Coprocessor 支持，极大降低大数据量聚合时的延迟
- 应用层二级索引的支持

```
@Table("alertdef")
@ColumnFamily("f")
@Prefix("alertdef")
@Service(AlertConstants.ALERT_DEFINITION_SERVICE_ENDPOINT_NAME)
@JsonIgnoreProperties(ignoreUnknown = true)
@TimeSeries(false)
@Tags({"site", "dataSource", "alertExecutorId", "policyId",
"policyType"})
@Indexes({
   @Index(name="Index_1_alertExecutorId", columns = {
"alertExecutorID" }, unique = true),
})
public class AlertDefinitionAPIEntity extends
TaggedLogAPIEntity{
   @Column("a")
   private String desc;
   @Column("b")
   private String policyDef;
   @Column("c")
   private String dedupeDef;
```




```
Query=AlertDefinitionService[@dataSource="hiveQueryLog"]{@policyDef}
```

海量监控数据存储和查询框架

统一针对监控数据的HBase Rowkey 设计

```
Rowkey ::= Prefix | Partition Keys | timestamp | tagName | tagValue | …
```

- Metric

```
Rowkey ::= Metric Name | Partition Keys | timestamp | tagName | tagValue | …
```

- Entity

```
Rowkey ::= Default Prefix | Partition Keys | timestamp | tagName | tagValue | …
```

- Log

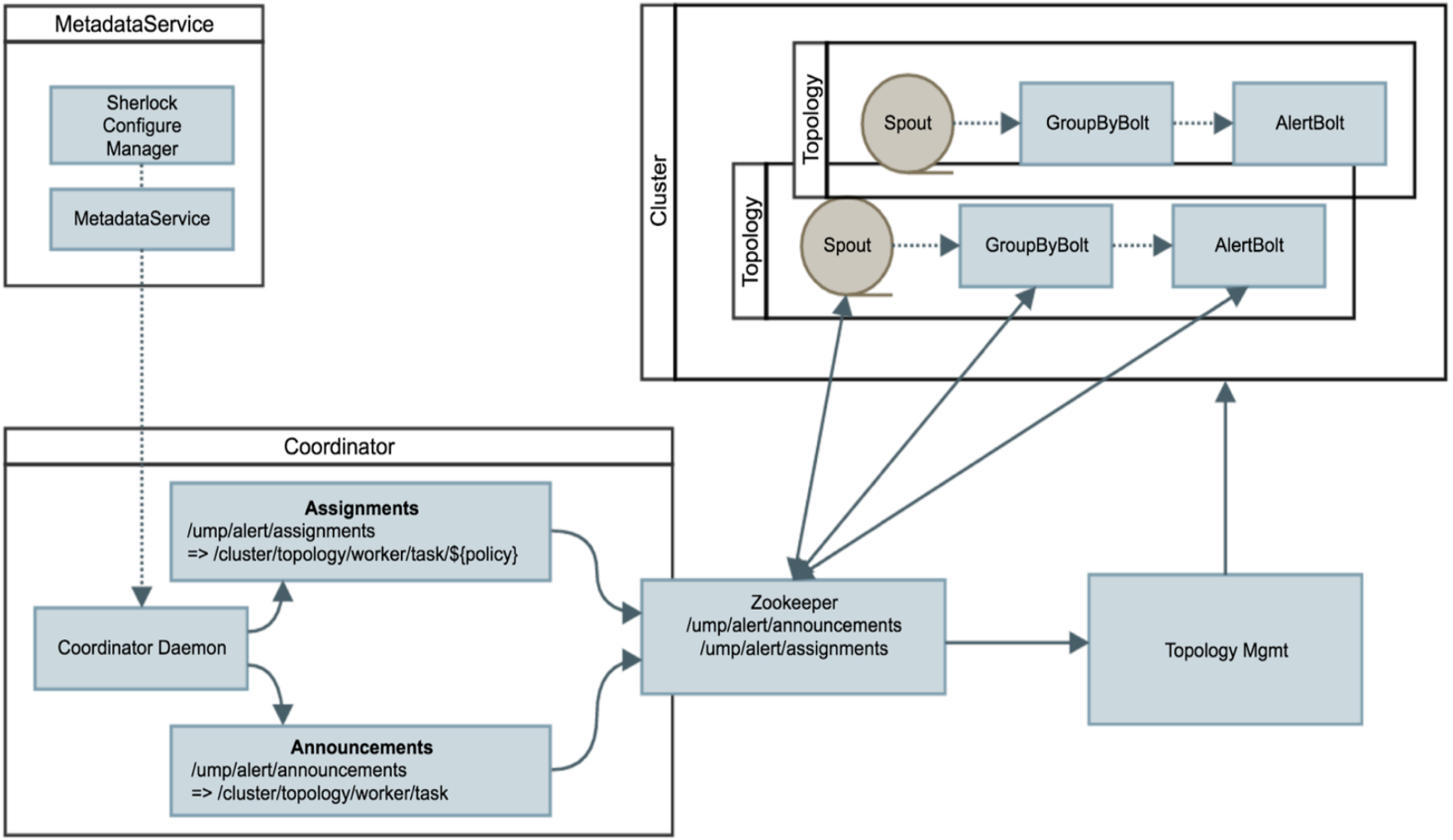
```
Rowkey ::= Log Type | Partition Keys | timestamp | tagName | tagValue | …
Rowvalue ::= Log Content
```

OPENTSDB *http://opentsdb.net*

多租户支持

- 元数据驱动动态Topology管理
  - 动态数据源（Kafka Topic）加载
  - 动态GroupBy/Join路由
  - 动态Policy部署与执行
  - 可定制预警发布/通知
- 资源调度和隔离
  - 资源隔离单元：CEP Runtime, Bolt, Topology, Cluster
  - 通过实时统计不同资源单位性能指标（IO,CPU,Memory）以动态优化Policy分布
- 高可用和容错
  - 流式计算状态管理：Message WAL + Siddhi (Checkpoint Snapshot) + Storm 状态管理
  - HA: 通过多Topology实例实现HA，同时支持无宕机维护或升级
- 可伸缩性
  - IO伸缩：针对单一数据源基于PartitionKey分流
  - 计算伸缩：针对Policy 计算复杂度或者内存消耗分布
  - 跨Topology伸缩：不同Topology仅执行部分Policy
  - 跨集群伸缩：不同Cluster仅执行部分(Scalability)或全部Policy (DR)

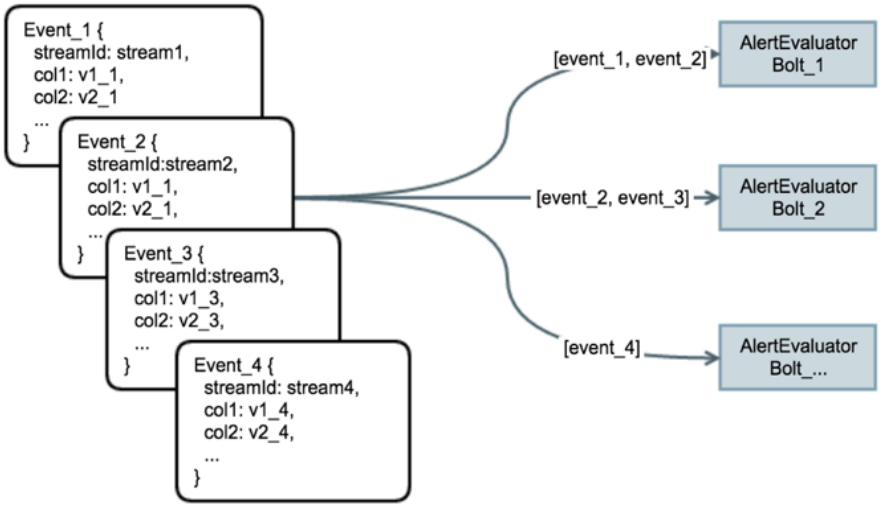
多租户支持 – Topology管理与资源调度



多租户支持 – 预警策略(Policy)动态关联

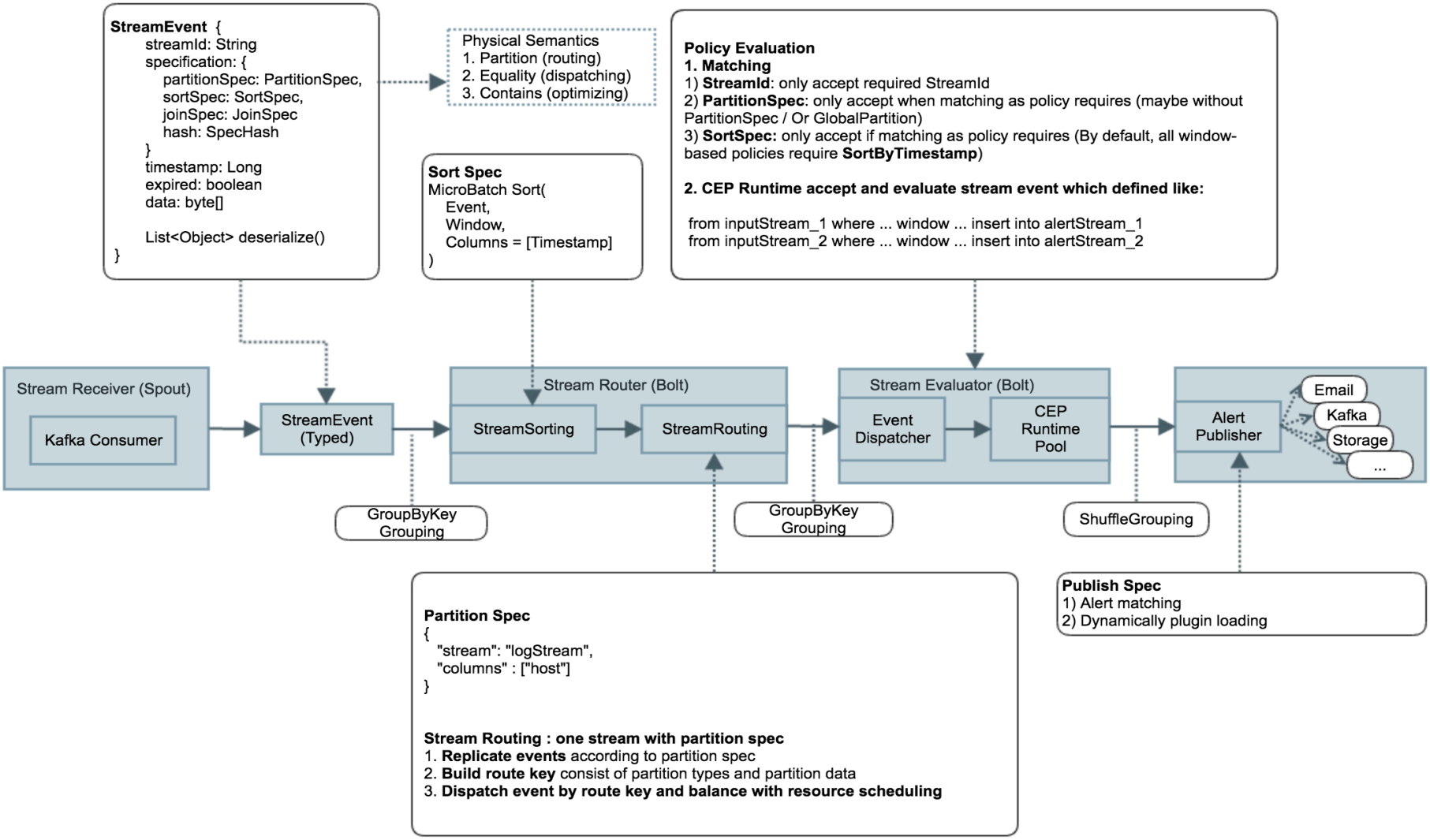
- 单一Runtime关联
  - Sort, Groupby, Join, Window
  - Policy动态部署(Hot Deploy)
  - 生命周期管理(Lifecycle)

- 单一数据流多重关联
  - 单一数据流进行多种不同Group
  - 单一数据流进行多种不同Sort
  - 单一数据流进行多种不同Join



- 多数据流多重关联
  - 多数据流（Stream）JOIN
  - 实时（Real-time）与历史（Historical）数据流JOIN
  - Alert多重关联去燥

多租户支持 –预警策略(Policy)动态关联



Eagle应用场景

1. 大数据安全：实时异常数据行为安全监控
   Secure Hadoop in Realtime a data activity monitoring solution to instantly identify access to sensitive data, recognize attacks/ malicious activity and block access in real time.
2. Hadoop性能监控：JoB性能监控与异常检测
   Hadoop, Spark Job Profiling & Performance Monitoring, Cluster Health Anomaly Detection
3. eBay全球统一监控系统平台预警引擎
   Shared multi-tenant alert engine of global unified monitoring platform
4. 其他通用分布式实时Anomaly Detection/Alerting场景

大数据安全：异常数据行为安全监控

数据丢失保护

Get alerted and stop a malicious user trying to copy, delete, move sensitive data from the Hadoop cluster.

异常登录或授权

Detect login when malicious user tries to guess password. Eagle creates user profiles using machine learning algorithm to detect anomalies

未授权访问

Detect and stop a malicious user trying to access classified data without privilege.

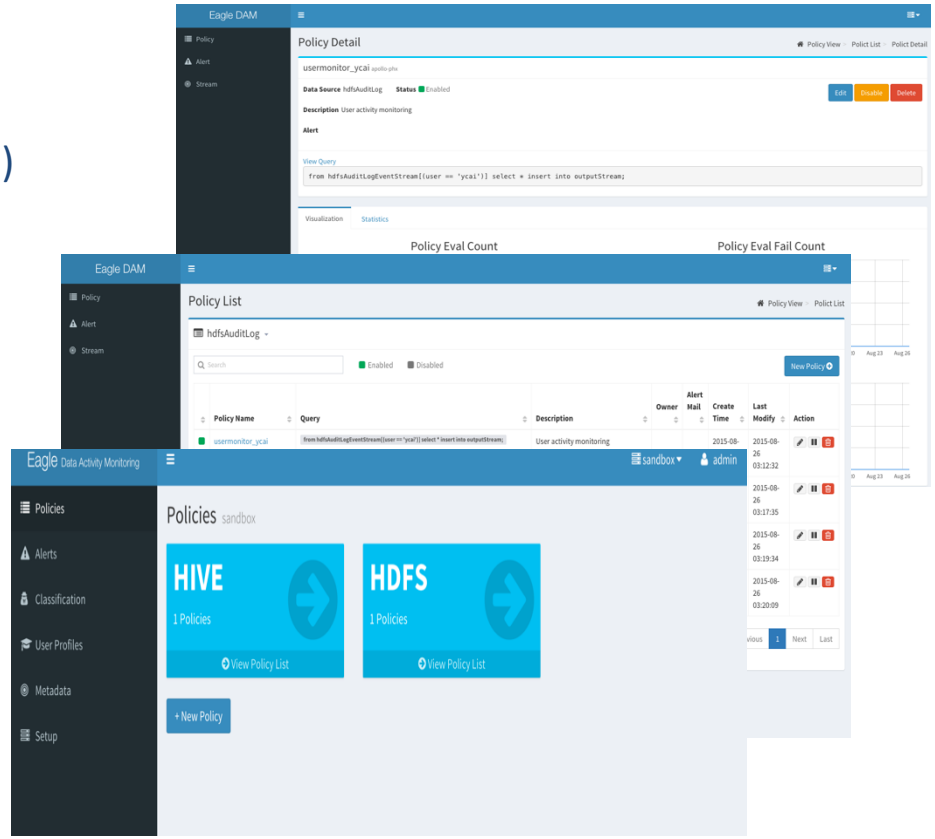
异常数据操作

Detect and stop a malicious user trying to delete large amount of data. Operation type is one parameter of Eagle user profiles. Eagle supports multiple native operation types.

Hadoop 数据行为监控规则

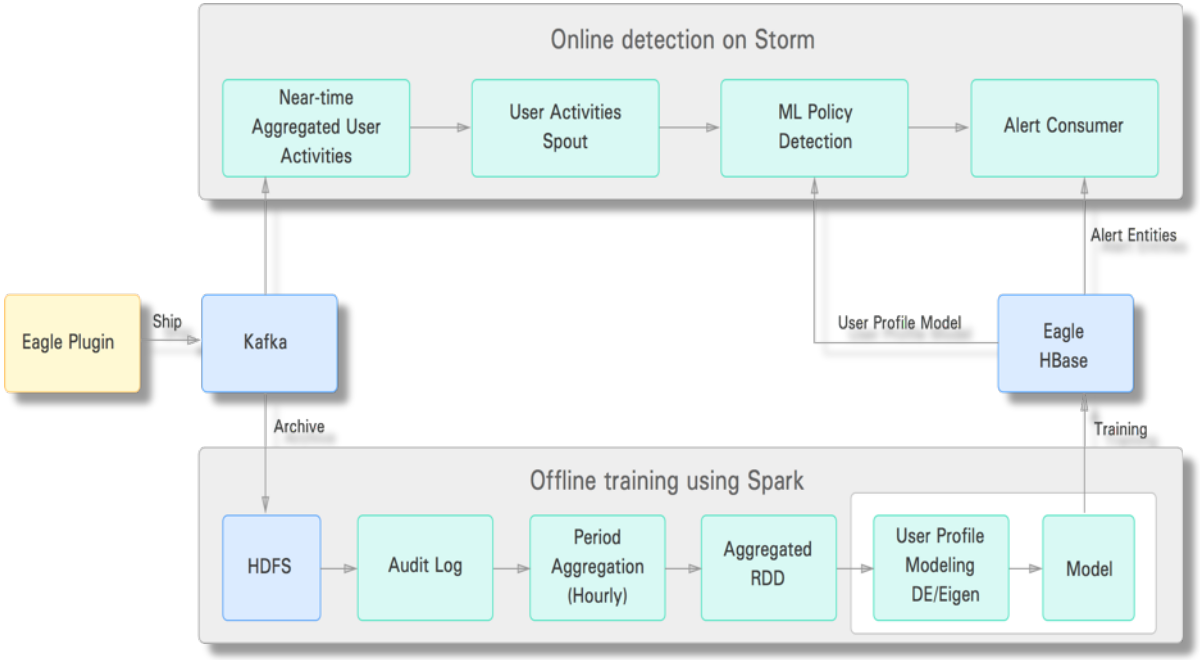
Detect anomalies in accessing HDFS and Hive

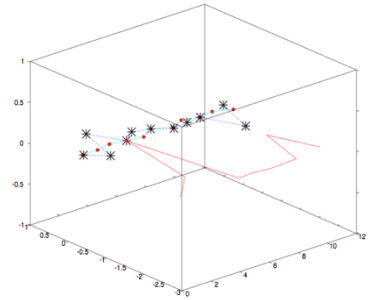
- HDFS 监控策略
  - 敏感文件访问
  - HDFS命令的执行 (read, write, update…)
  - 客户端主机地址
  - 目标文件地址
  - HDFS文件安全区域（Zone）
- Hive 监控策略
  - 访问数据仓库表中 PII数据
  - Hive SQL查询Profile
  - 客户端主机地址
  - Hive表数据安全区域（Zone ）
- 数据安全分类以及敏感表识
  - 支持REST导入或者通过UI标注
  - 支持DataGuise敏感信息元数据集成



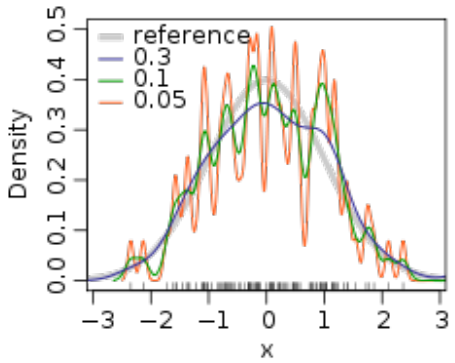
基于机器学习的用户画像(User Profile)

离线训练: Determine bandwidth from training dataset the kernel density function parameters (KDE)
在线探测: If a test data point lies outside the trained bandwidth, it is anomaly (Policy)





PCs(Principle Components) in EVD (Eigenvalue Value Decomposition)



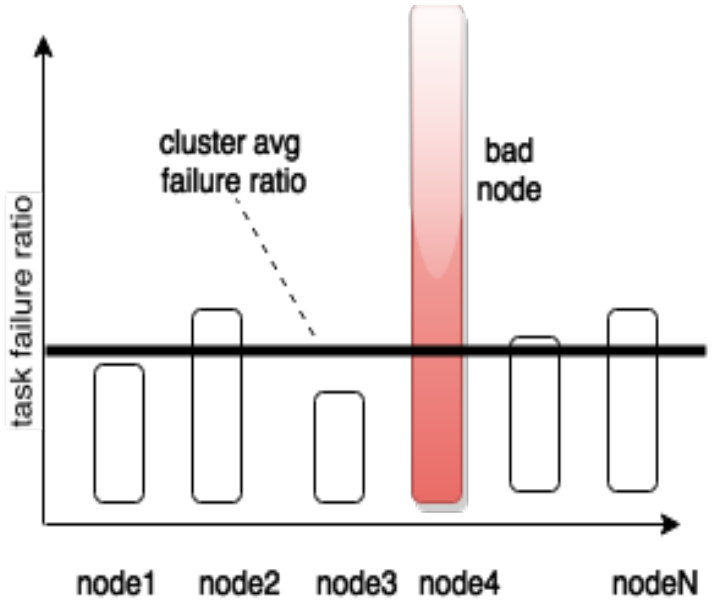
Kernel Density Function

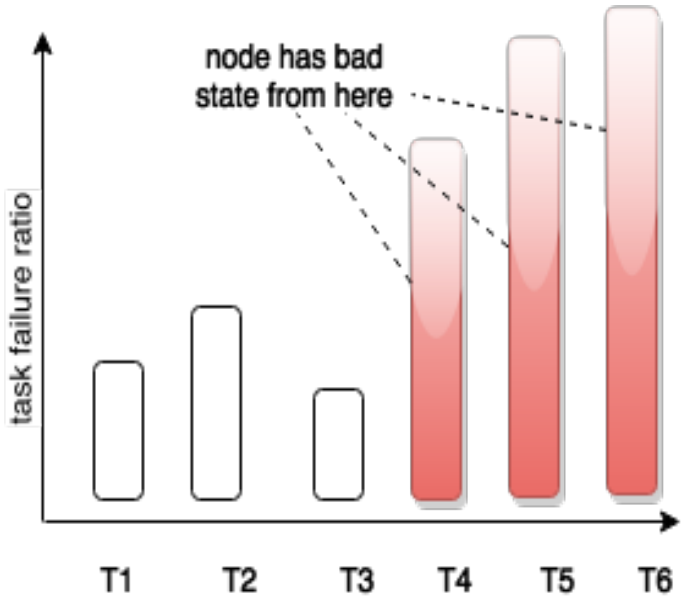
Hadoop性能监控：Job/Node异常监测

场景 *Detect node anomaly by analyzing task failure ratio across all nodes*

假设 *Task failure ratio for every node should be approximately equal*

算法 *Node by node compare (symmetry violation) and per node trend*



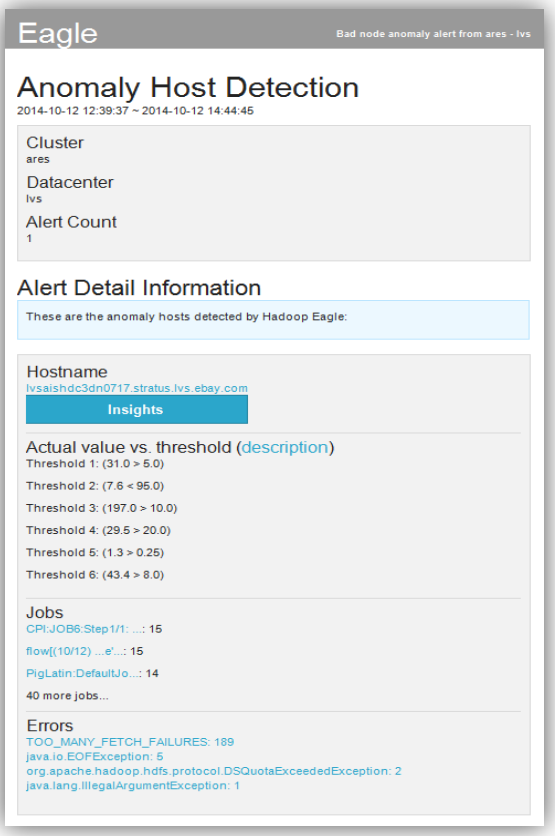


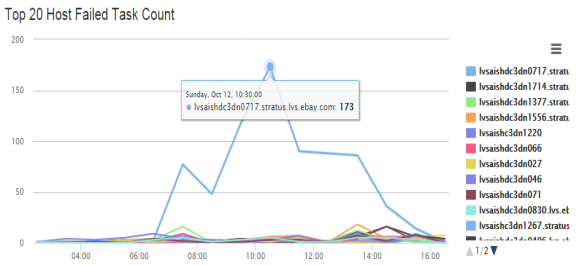
基于Task统计模型的节点异常预警与分析

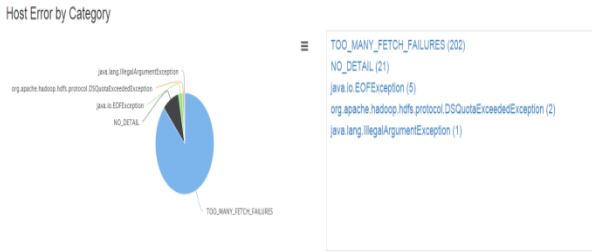
预警:
Anomaly Detection Alerting

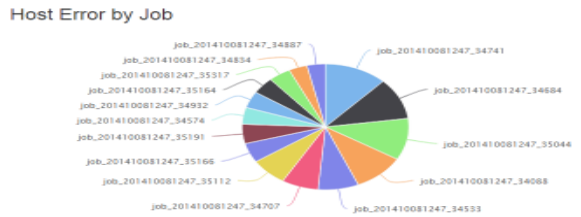
Trouble-Shooting:
Task failure drill-down

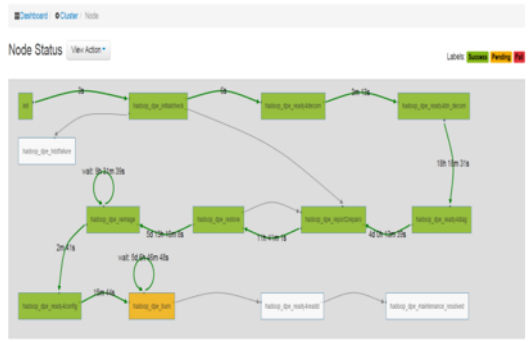
分析:
Task failure drill-down











Hadoop Job数据倾斜预警

场景 Detect data skew by statistics and distributions for attempt execution durations and counters

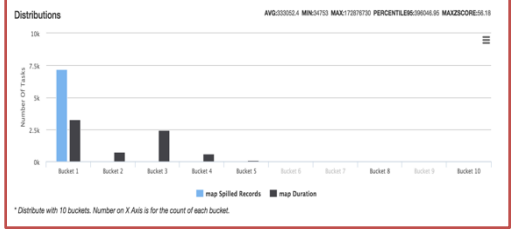
假设 Duration and counters should be in normal distribution

Counters & Features

mapDuration
reduceDuration
mapInputRecords
reduceInputRecords
combineInputRecords
mapSpilledRecords
reduceShuffleRecords
mapLocalFileBytesRead
reduceLocalFileBytesRead
mapHDFSBytesRead
reduceHDFSBytesRead

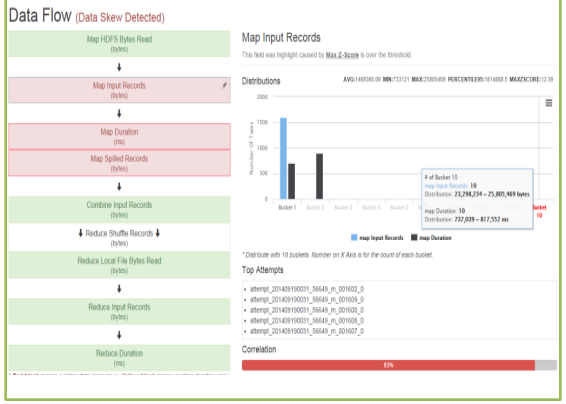
Modeling & Statistics

Avg
Min
Max
Distributions
Max z-score
Top-N
Correlation



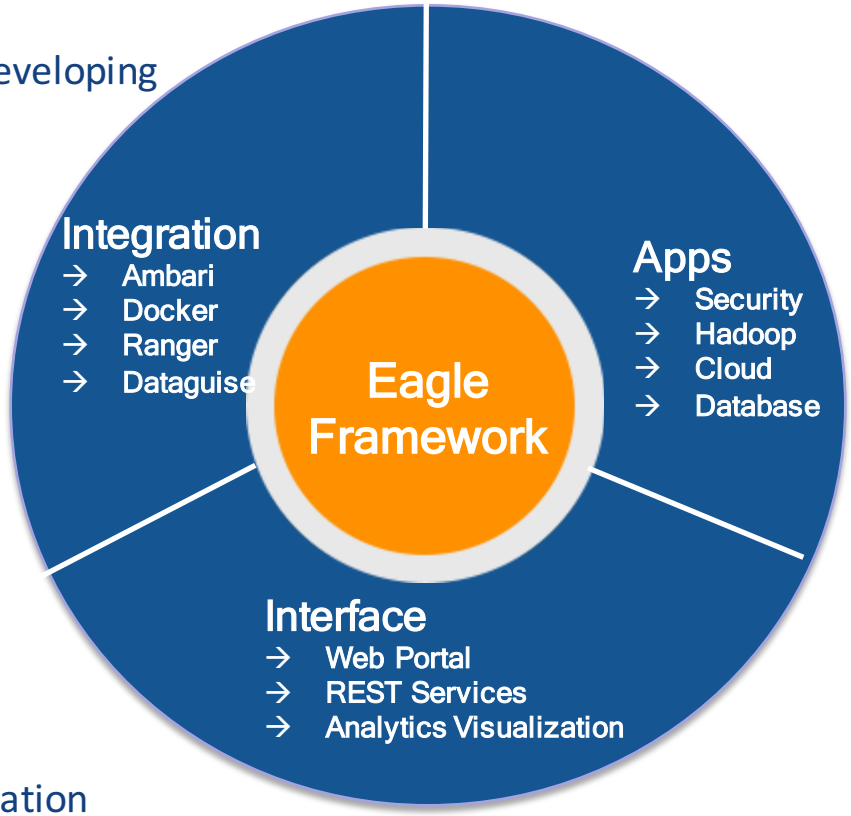
Threshold & Detection

Correlation > 0.9
& Max(Z-Score) > 90%



Eagle生态系统

1. Eagle 框架
   Distributed real-time framework for efficiently developing highly scalable monitoring applications
2. Eagle 应用
   Security/ Hadoop/ Operational Intelligence / …
3. Eagle 接口
   REST Service / Management UI / Customizable Analytics Visualization
4. Eagle 集成
   Ambari / Docker / Ranger / Dataguise
5. 开源与社区
   Community-driven and Cross-community cooperation



了解更多

Apache Eagle社区

- 网站: http://eagle.incubator.apache.org
- Github: http://github.com/apache/incubator-eagle
- 邮件列表: dev@eagle.incubator.apache.org

论文发表

- EAGLE: USER PROFILE-BASED ANOMALY DETECTION IN HADOOP CLUSTER (IEEE)
- EAGLE: DISTRIBUTED REALTIME MONITORING FRAMEWORK FOR HADOOP CLUSTER

会议或活动 (既将开始)

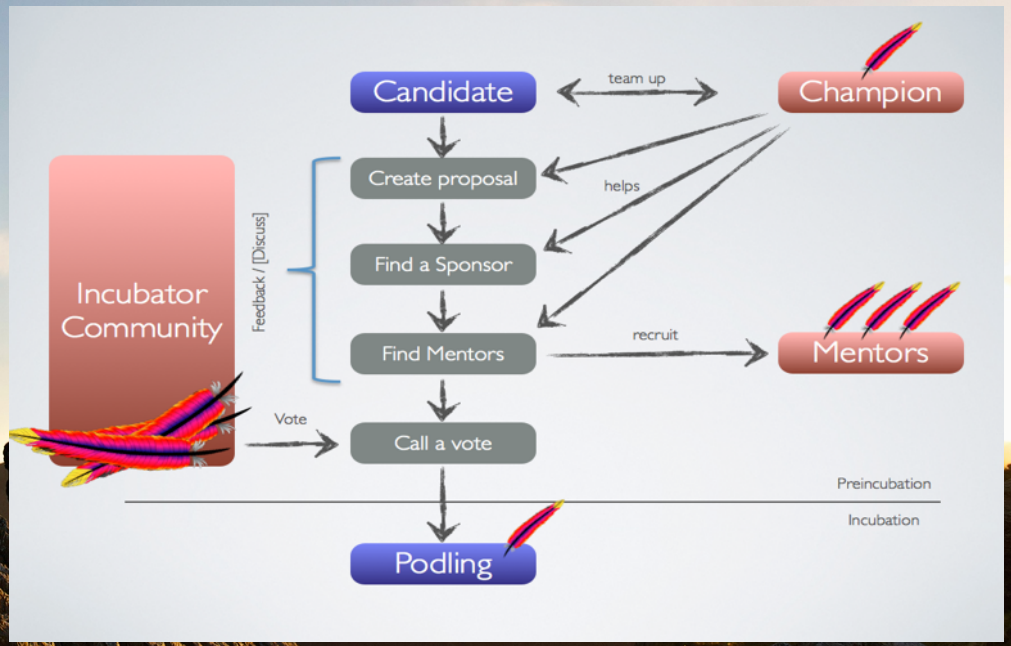
- Hadoop Summit 2016 at Dublin (http://hadoopsummit.org/dublin/)
- Hadoop Summit 2016 at San Jose (http://hadoopsummit.org/san-jose/)
- Hadoop Strata World 2016 at London (http://conferences.oreilly.com/strata/hadoop-big-data-eu)
- DTCC 2016 中国数据库大会 (http://dtcc.it168.com)

关于开源

If you want to go fast, go alone.
If you want to go far, go together.

-- African Proverb

THE APACHE SOFTWARE FOUNDATION



PULSAR

Open Sourced By ebay

Q & A

http://eagle.incubator.apache.org

dev@eagle.incubator.apache.org

apache/incubator-eagle

@TheApacheEagle

Geekbang 极客邦科技 InfoQ

THANKS!

ebay THE APACHE SOFTWARE FOUNDATION