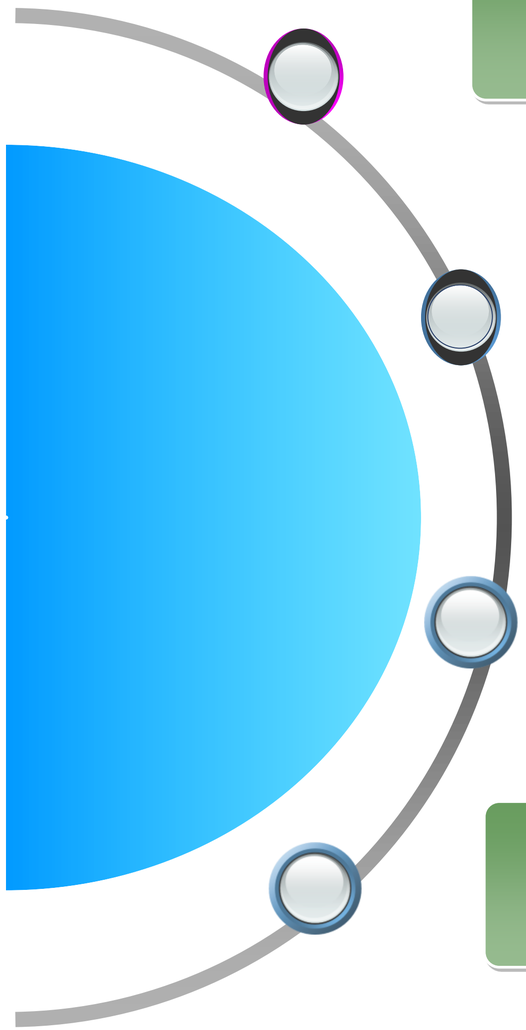
搜索离线平台化

杨孔仕

目录



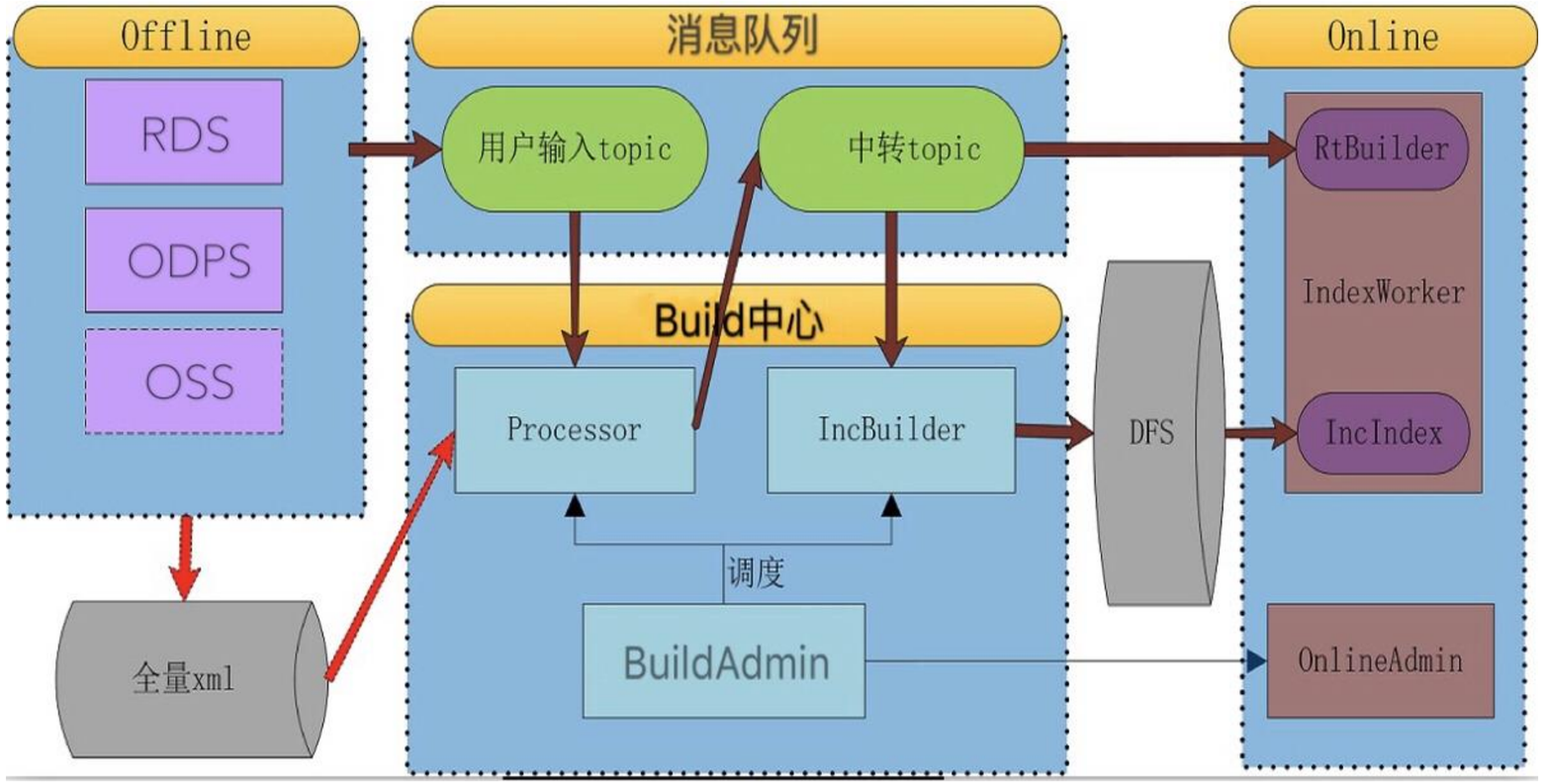
概念扫盲

- 全量索引重建 & 实时更新
  - 全量索引重建在搜索服务中是必不可少的一环
    - 业务本身系统的故障，出现数据的丢失
    - 业务高速发展产生增减字段或者修改分词算法等相关的需求
    - 业务第一次数据导入
  - 实时更新
    - 实时产生的数据进入Elasticsearch可被搜索

概念扫盲

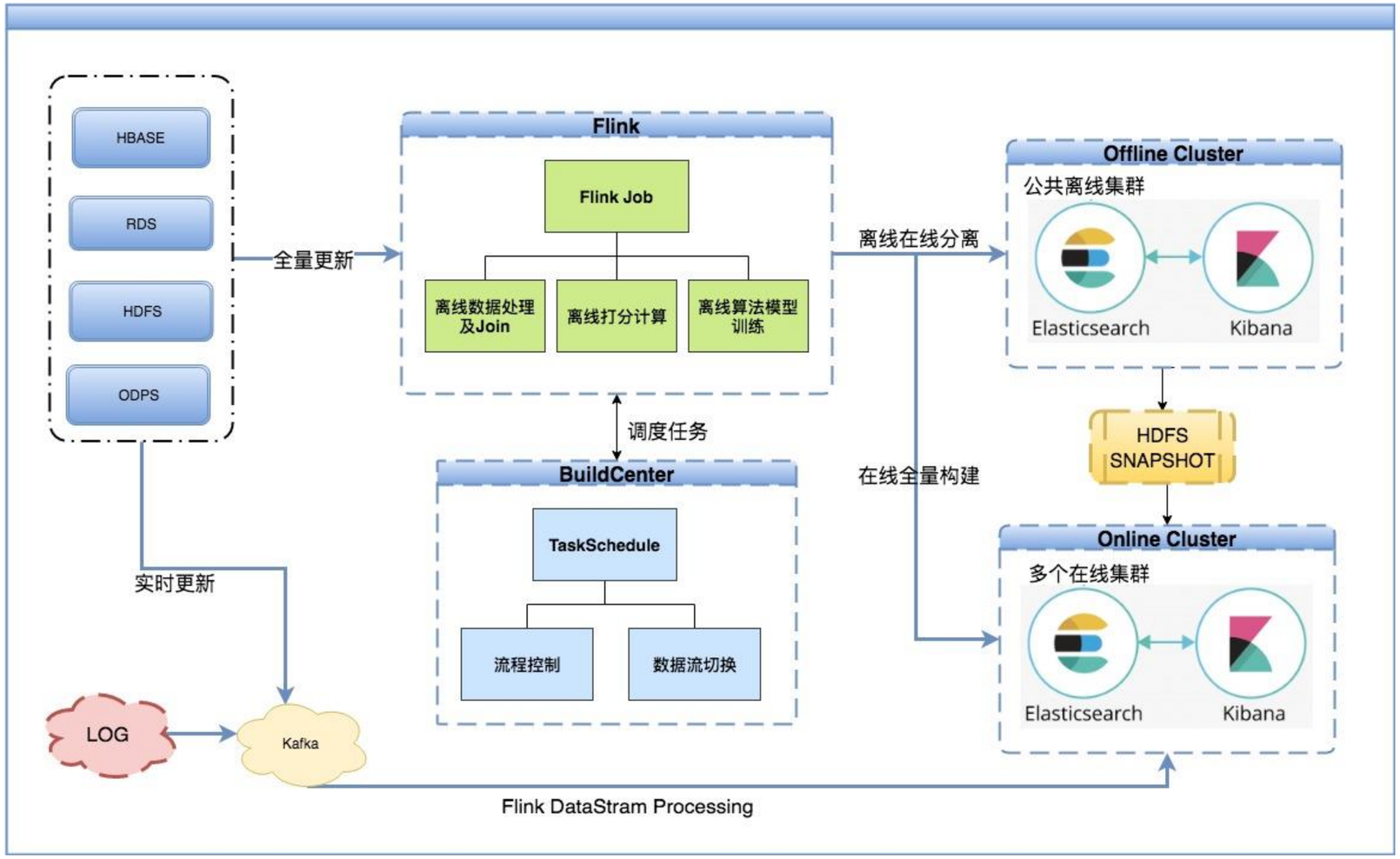
- 消息队列
  - 缓存，为了服务的稳定，通常我们会将实时数据进消息队列，这样我们在数据丢失时重新消费
  - 蓄洪保护，避免将Elasticsearch写挂
  - 追增量，在做全量切换的过程中，会有一个追增量的过程，这时实时数据是双写的，也就是会有两个消费者从不同的时间位点消费数据发向Elasticsearch的两个索引，追完增量才会切换索引

搜索Build离线系统



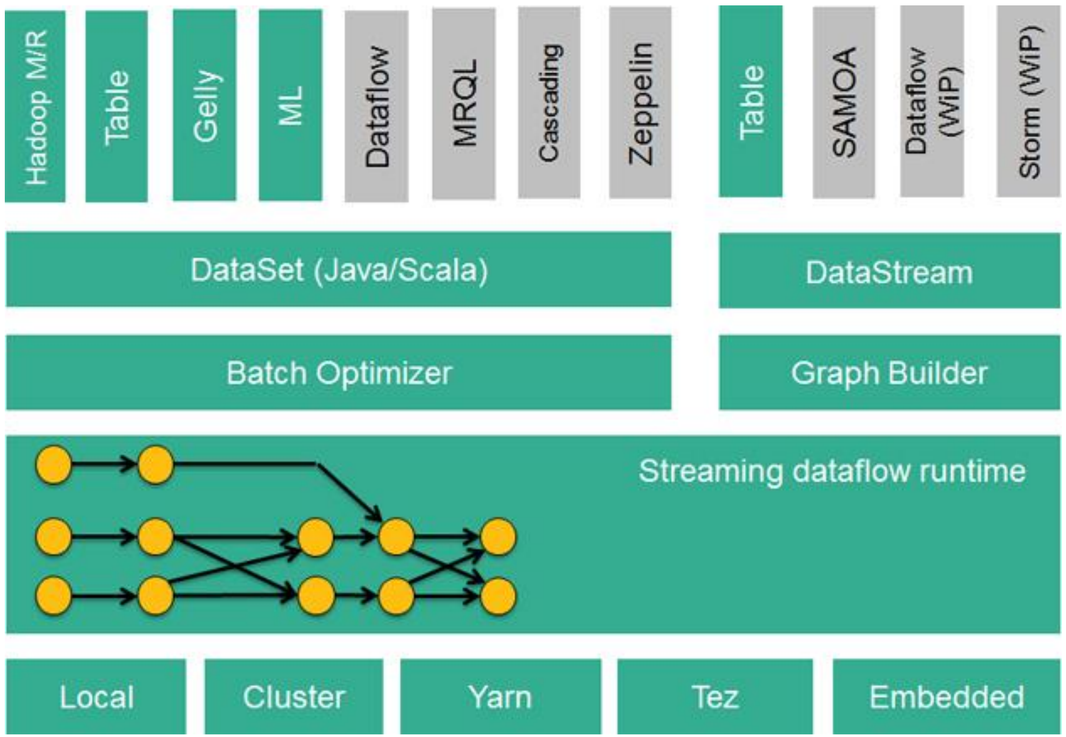
- 离线、在线完全分离（不影响在线查询性能）
  - build / merge 都在离线处理，在线只接受实时的增量，不做Merge操作
  - 增量Merge操作离线做完后存HDFS，在线拷贝后做内存的切换
- 可以类比Elasticsearch在master 、data、ingest等角色外新增了builder和merger角色
- 使得离线全量、增量、实时数据统一入口
- 复杂处理操作在processor完成，多备份的情况下能节约机器资源
- 全量和增量的无缝切换

电商搜索离线平台



- 引入了Flink统一了批次(全量)和流式(实时)任务
- 多表Join和离线算法模型训练产出到在线Elasticsearch
- 可配置化数据接入，全量和增量的无缝切换，自定义Flink Job & Flink SQL
- 重要业务由公共离线集群build全量，保证不影响在线集群
- 研发Flink-Elasticsearch-Connect基于Rest Client实现，不受限于Elasticsearch版本
- 日志场景做Elasticsearch的写入优化
- 自动降速，实时检测ES中队列堆积情况，有堆积则降速，触发背压

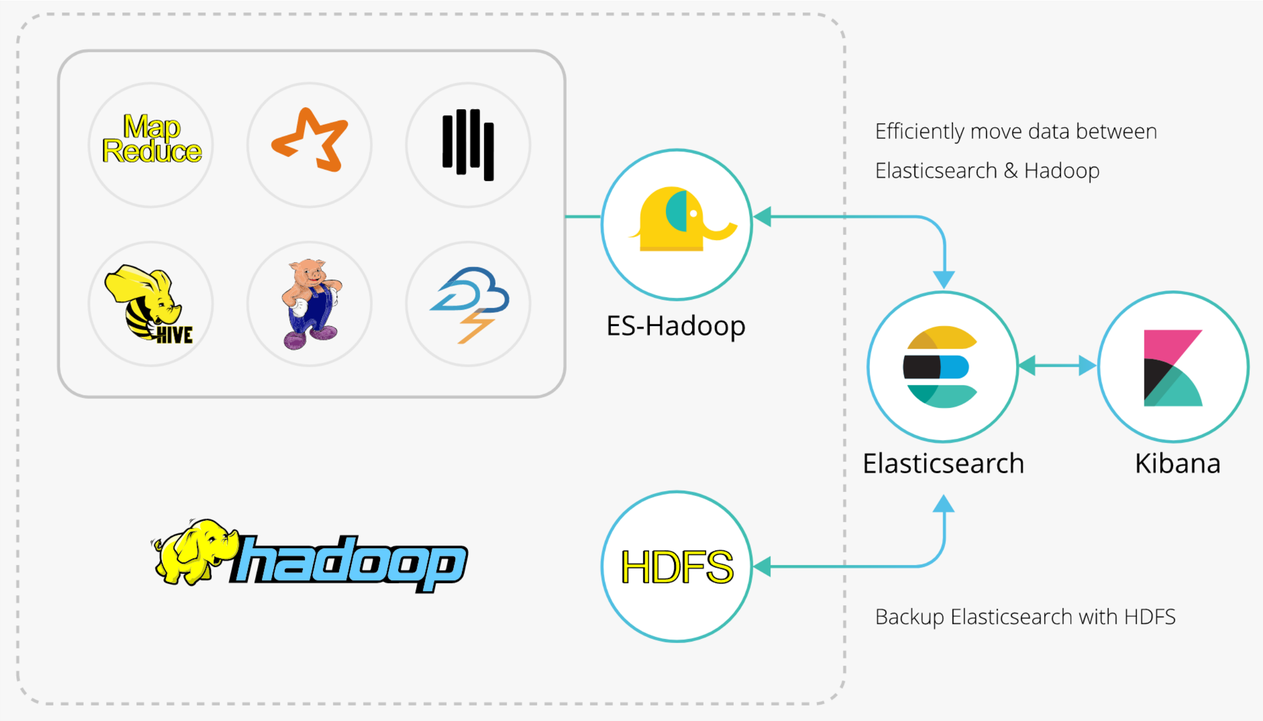
Why Flink



- 统一的流式和批处理流程
- Table & SQL API
- Flink 认为 Batch 是 Streaming 的一个特例，所以 Flink 底层引擎是一个流式引擎
- 基于WaterMark实现时间窗口和Event Time消息乱序的处理

Hadoop与ES数据互通

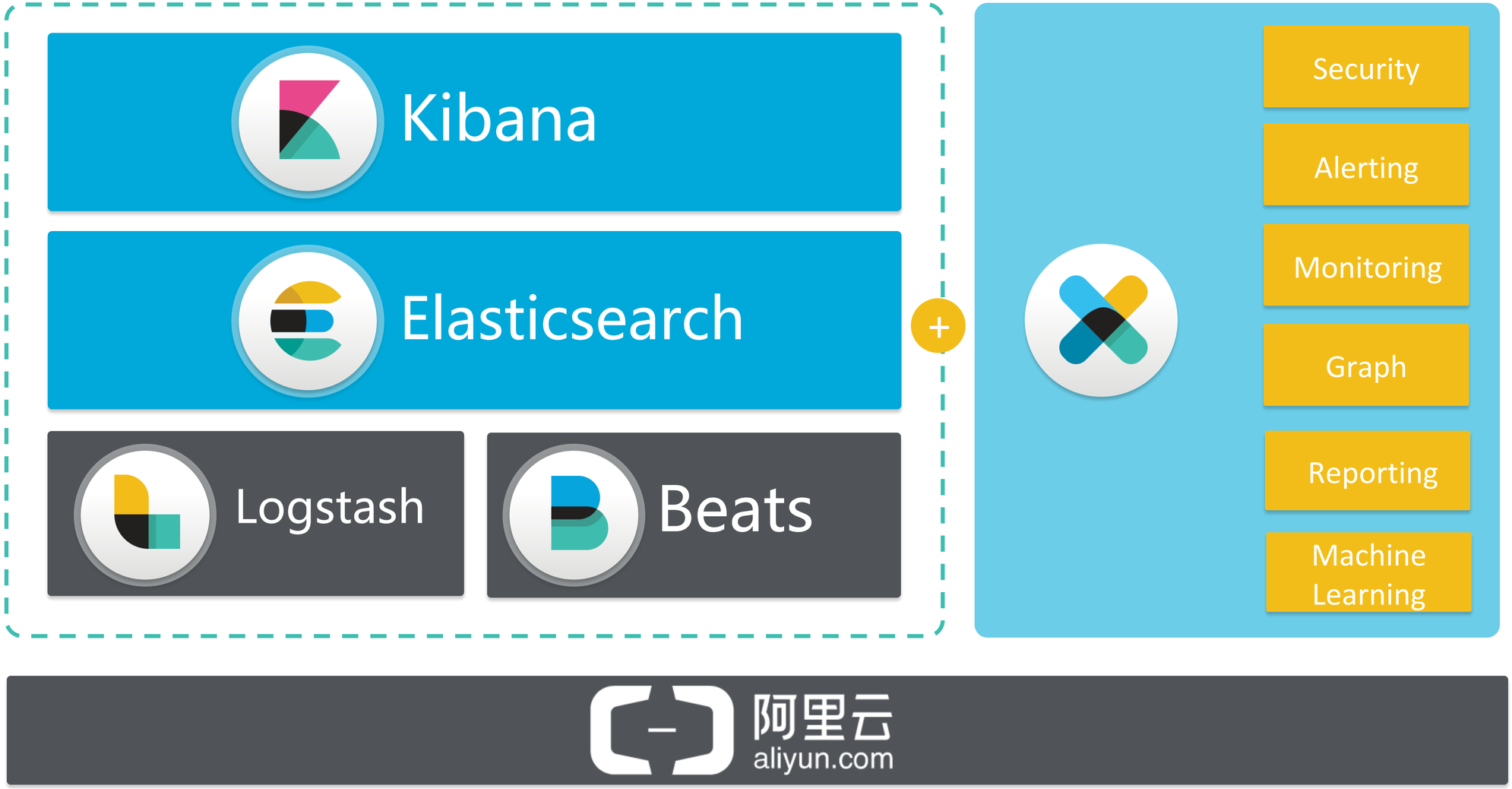
除了 Flink 之外, 使用 Spark 和 ES-Hadoop 的组合也是不错的选择 ，可以使用 ES-Hadoop 将 Hadoop 数据索引到 Elasticsearch，以充分利用快速的 Elasticsearch 引擎和美观的 Kibana 进行可视化。



ES-Hadoop

凭借现有 Hadoop API 的动态扩展，ES-Hadoop 让您能够在 Elasticsearch 和 Hadoop 之间轻松地双向移动数据，同时借助 HDFS 作为存储库，进行长期存档。分区感知、故障处理、类型转换和数据共置均可透明地完成。

阿里云Elasticsearch产品架构



产品优势

产品功能

- Elasticsearch 5.5.3 ；
- Kibana；
- X-Pack；
  - Security;
  - Alerting；
  - Monitoring；
  - Reporting & Graph;
  - Machine Learning;
- 预置分词库及自定义分词
- OSS SNAPSHOT

安全稳定性

- 数据冗余备份；
- 多可用区；
- 弹性扩容；
- Dedicated Master；
- 专有网络VPC安全隔离；
- 优化资源利用率；
- 云监控报警

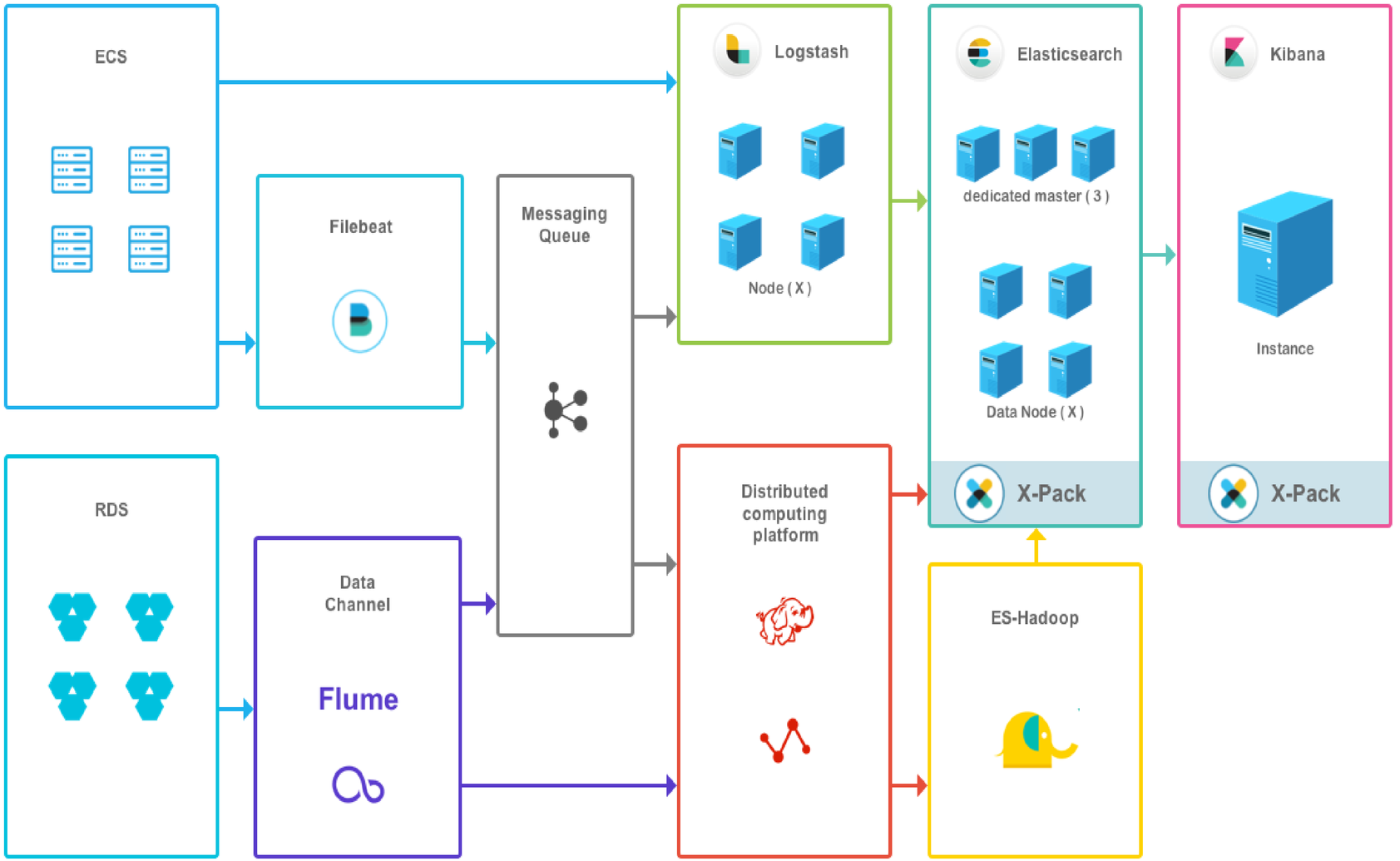
技术支持

- 7*24*365在线技术支持；
- 网络、电话、钉钉支持；
- Elastic开源社区合作支持；

常见痛点

自建痛点	产品优势
无灾备能力和容错机制	阿里云IaaS支持解决容灾
部署运维成本高	自动化部署，0成本运维
技术瓶颈解决不了线上问题	Elastic官方技术支持
分析功能缺失	可视化分析功能
搜索结果无法优化	优化中文分词及搜索结果
性能瓶颈无法弹性扩容	扩容简单
无产品组合	云上产品组合使用

场景实例 | 日志处理

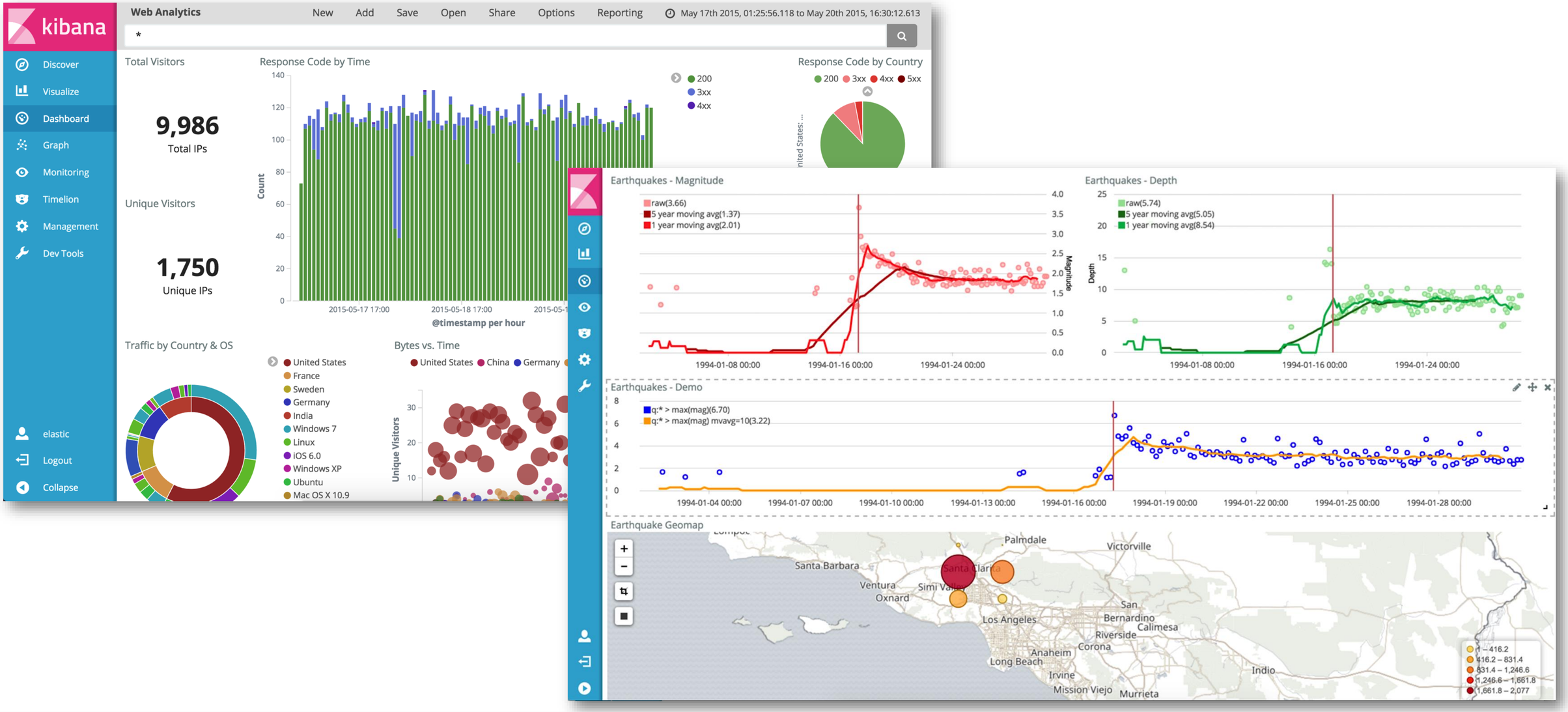


Aliyun-DataHub

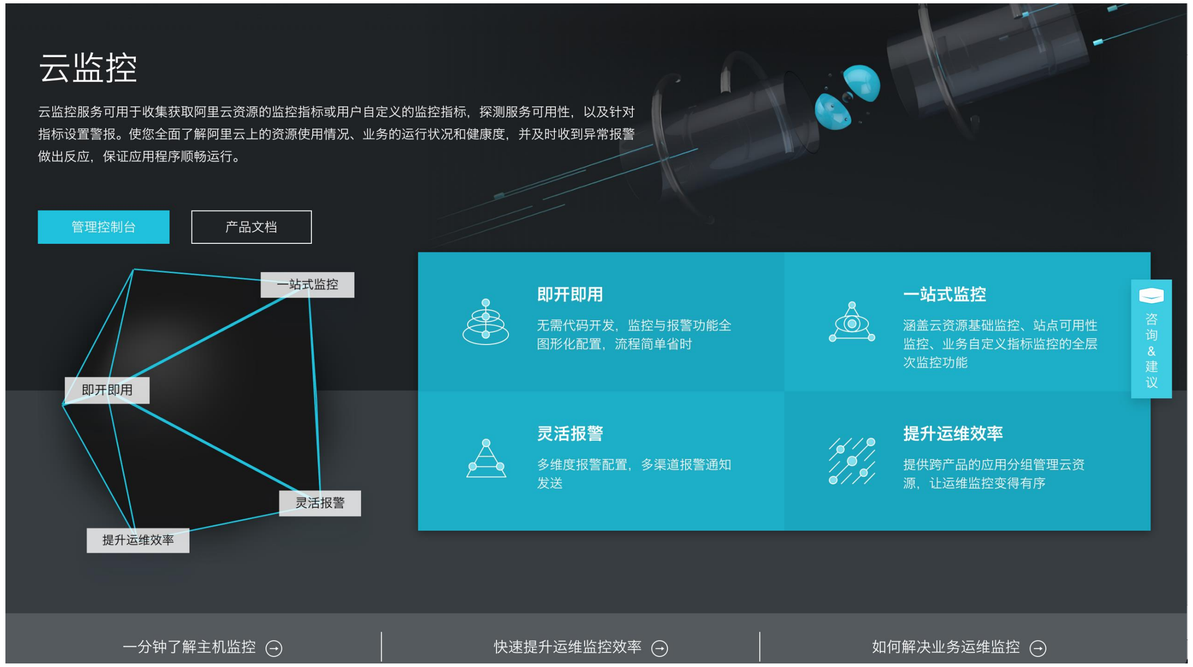
Aliyun-SLS

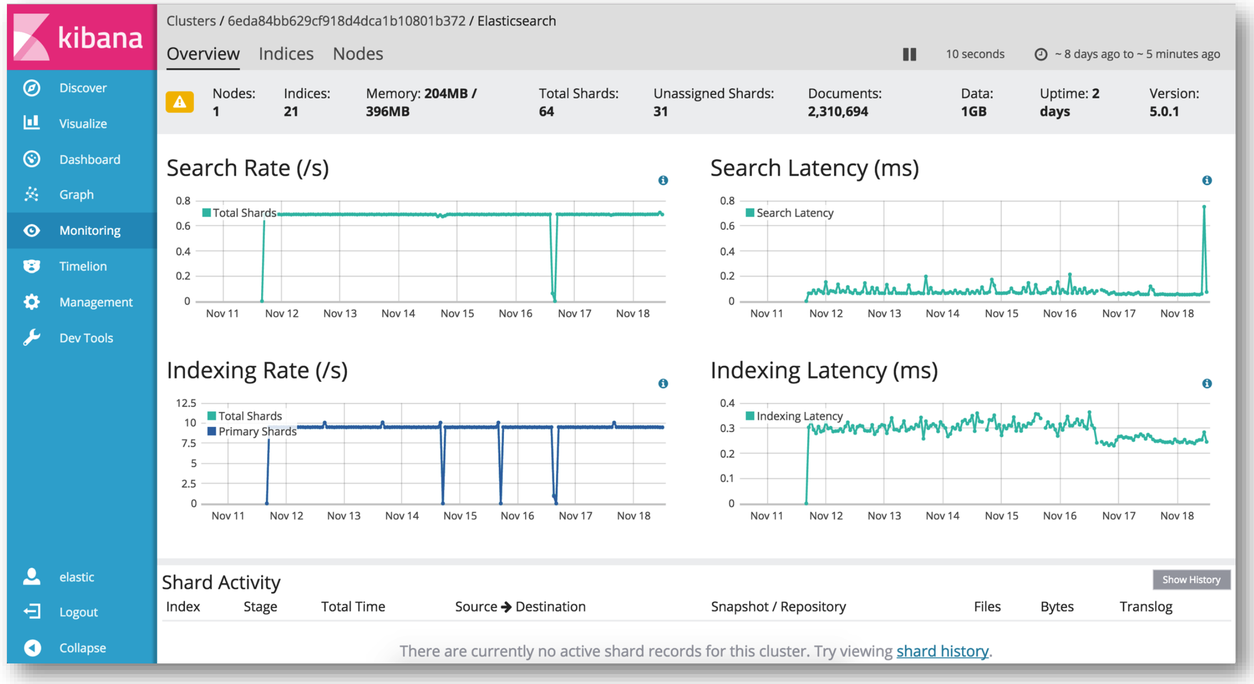
Aliyun-DTS

处理日志 | 分析聚合及可视化



处理日志 | 监控

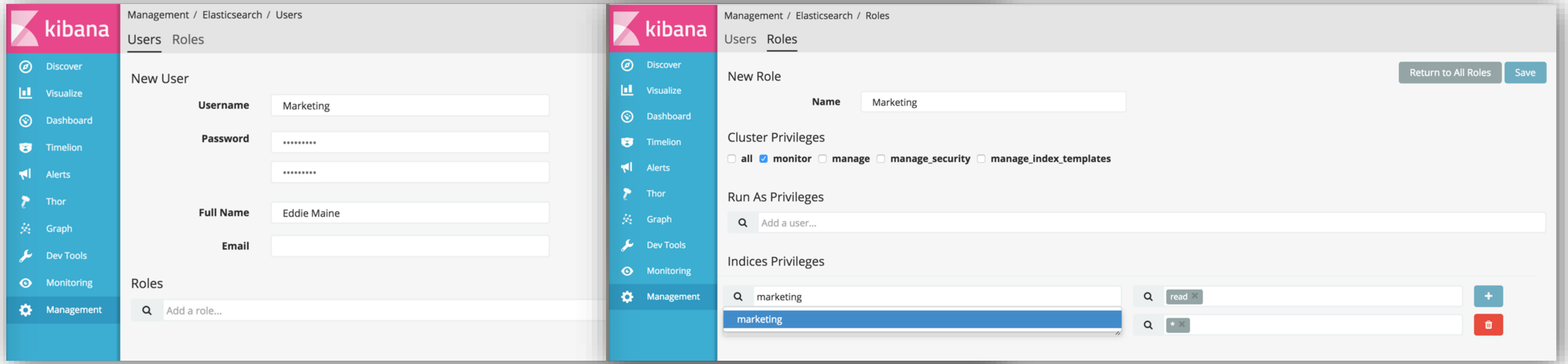




【安全监控】Kibana&X-Pack

- 实时集群健康监控，云监控对接；
- 提供实时告警；
- 丰富可视化交互界面；

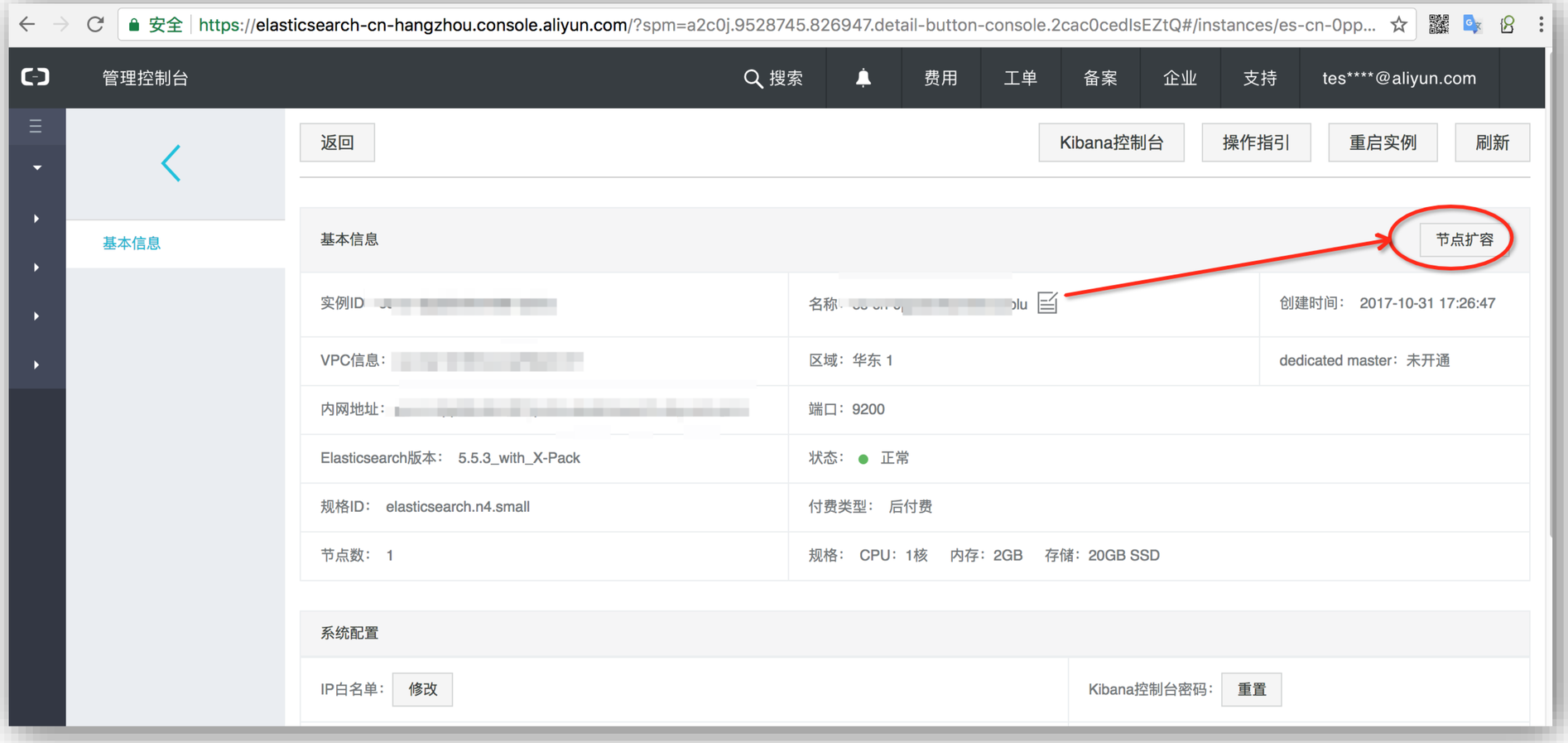
处理日志 | 安全及角色管理



【角色管理】Kibana&X-Pack

- RBAC权限模型，可以自定义user及权限，admin及user区分权限，适应企业级多人协同工作；
- 访问的权限密码校验，admin和普通user区分校验；
- 文档级别的安全保护；
- 用户级别租户隔离；
- 审计日志；

处理日志 | 弹性伸缩



【弹性伸缩】

阿里云IaaS

- Cluster级别快速扩容
- Cluster配置
- Node级别存储扩容

为了无法计算的价值 | 阿里云