UNIVERSIDAD EAFIT
DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS
ST1612 SISTEMAS INTENSIVOS EN DATOS

2021-2

Lab 6 Streaming Real-time con ElasticSearch-Kibana-LogStash-Kafka

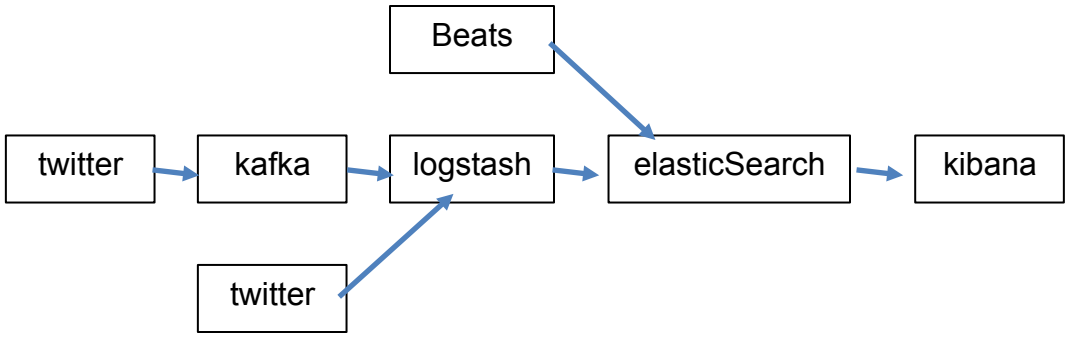
1. Descripción del caso de estudio

(ver guía de instalación de Kafka, elasticsearch, kibana, logstash del github: https://github.com/st1612eafit/st1612_20212.git

Realizar la instalación básica de ELK (Beats, LogStash, ElasticSearch, Kibana), realizar los tutoriales básicos en el github de LogStash (fuentes File y Twitter), ElasticSearch en https://assets.contentstack.io/v3/assets/bltefdd0b53724fa2ce/blt56ad3f4e2c755f29/5d37c1602a506857d64eff48/es_commands.txt y los diferentes tutoriales en línea que hay de kibana y en general de ELK en la página: https://www.elastic.co/ y Elasticsearch: Primeros pasos | Elastic Videos (estos tutoriales los puedes hacer en la nube de elasticsearch o configurando las opciones de seguridad mínima de ElasticSearch en la versión standalone).

Se desea diseñar e implementar un sistema de análisis y visualización de datos de twitter.

Ver arquitectura:



Como fuente se tendrá Twitter, con al menos 2 palabras claves: ‘quinterocalle’ y ‘eafit’.

Como fuente utilizará el agente Beats para Archivos o Métricas:

Esta fuente de datos (twitter) será adquirida por 2 medios:

1. Desde LogStash – twitter directo.
2. Desde LogStash – Kafka - twitter

Desde LogStash, deberá ejecutar la captura y ingesta a Elastic Search con:

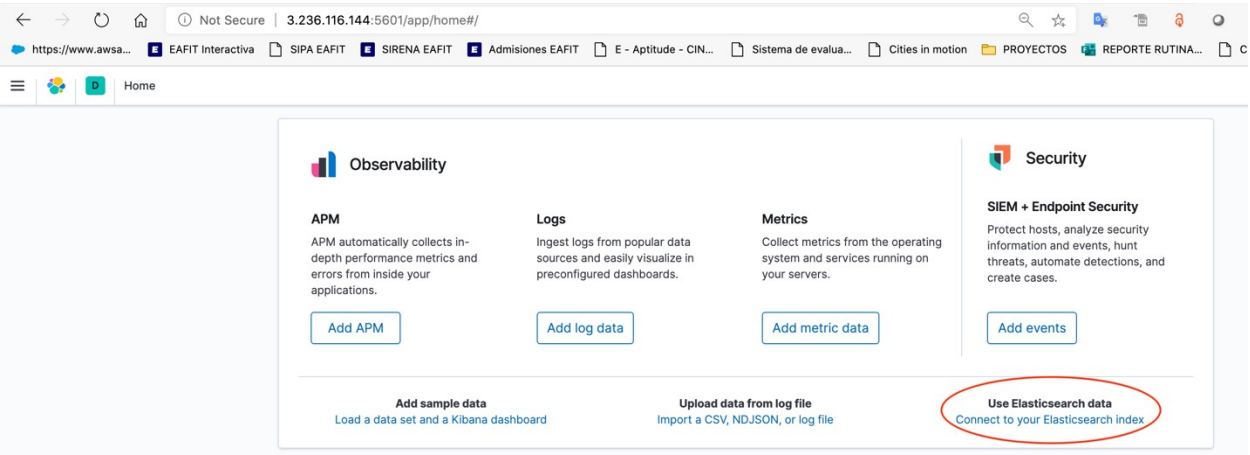
bin/logstash -f etl-from-twitter.conf

o

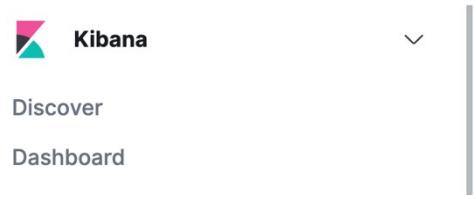
bin/logstash -f etl-from-kafka.conf

Cada una de las fuentes anteriores, deberá enviar los datos a un servidor ElasticSearch.

Los datos en el servidor ElasticSearch, deberán ser visualizados a nivel exploratorio (realizar gráficos generales y exploratorios de acuerdo con los datos), lo primero es relacionar los datos desde Kibana desde ElasticSearch:



Ya para hacer el análisis y visualización se sugiere: la opción de ‘Discovery’ y ‘Dashboard’.



Ambiente de implementación:

1. Montarlo en una máquina AWS EDUCATE EC2 – t2.medium – 20 GB DD.
2. Allí montar: Kafka, beatsfile, beatsmetric, elasticsearch, kibana, logstash

Entregable:

1. Documento básico a modo de bitácora del proceso (3 pag máx)
2. IP pública con puerto abierto 5601 para poder revisar.
3. Clave .pem para poder ingresar a la máquina de revisar.