Trabalho de final de módulo – Mineração de Texto com ELK - INFNET

Professor: Felipe Fink Grael

Aluno: Marcelo Lopes da Silva

Dezembro 2020

Instalação dos pré requisitos

Para a presente solução serão utilizados o Elasticsearch e o Kibana executados via docker. O acesso ao elasticsearch será feito utilizando jupyter notebook. Outras dependências podem ser instaladas utilizando o gerenciador de pacotes pipenv. Os exercícios foram realizados em máquina com sistema operacional Ubuntu versão 20

1 – Instalar o docker/docker compose

https://www.digitalocean.com/community/tutorials/how-to-install-and-use-docker-on-ubuntu-20-04-pt

2 – Instalar o pipenv

Em um terminal digite o comando pip install pipenv

3 – Instalar as dependências utilizadas no exercício

No diretório onde serão executados os notebook digite o comando pipenv install. Note que os arquivo Pipfile e Pipfile.loc tem que estar no mesmo diretório.

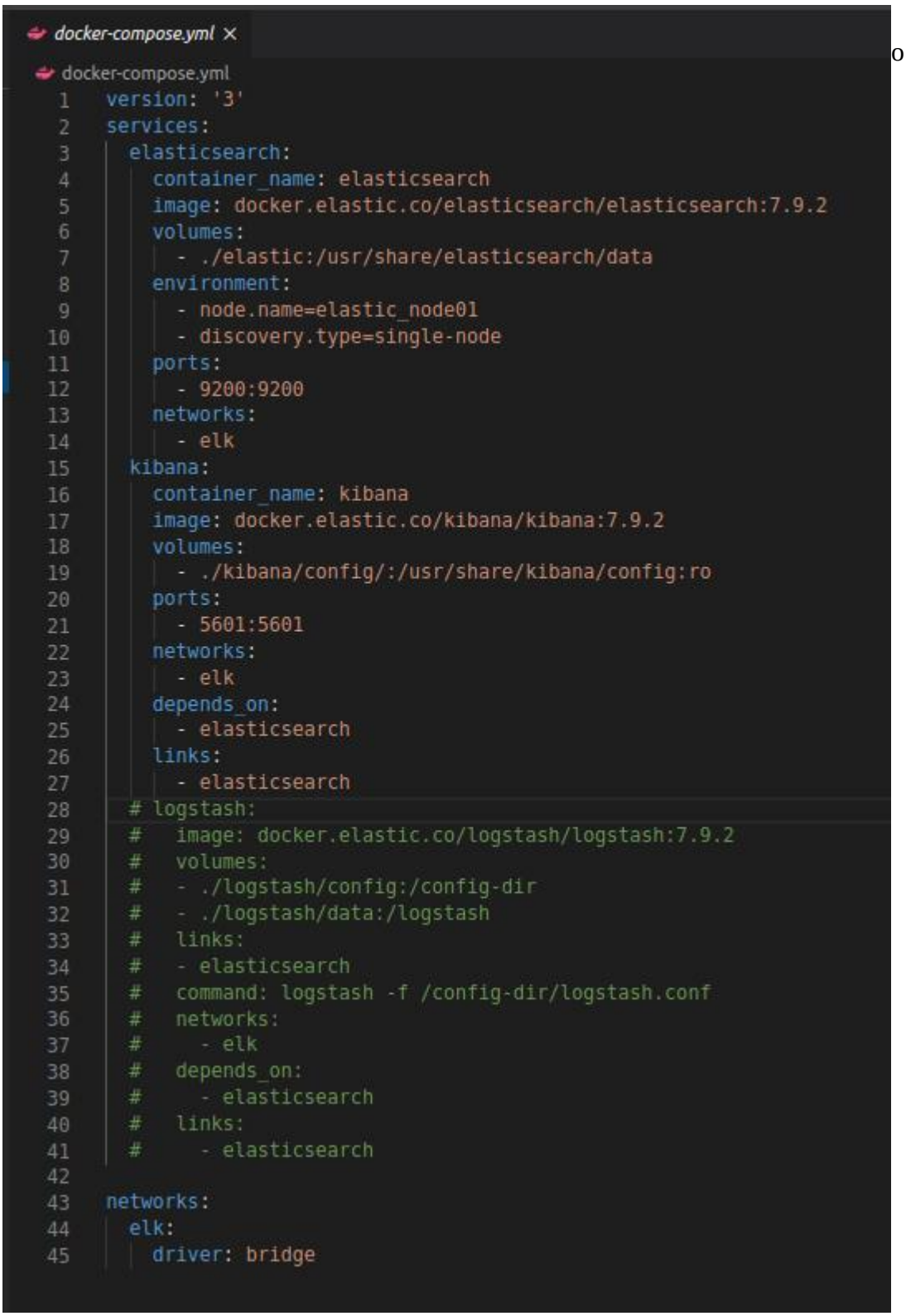
A execução dos notebooks tem que ser feita no ambiente virtual criado pelo pipenv em que as dependências foram instaladas. Para isso no terminal digite o comando pipenv shell.

Execução do Elastic e Kibana

Como informado anteriormente a execução é feita utilizando docker. Um arquivo docker compose é fornecido junto a solução.

docker-compose.yml

execute

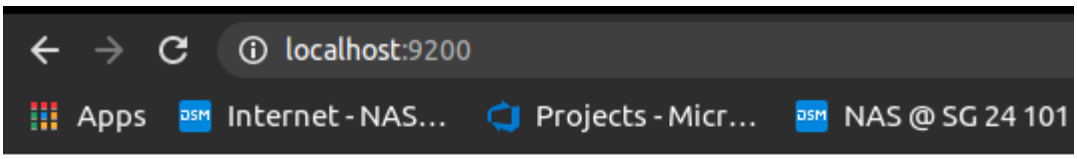


comando docker-compose up para iniciar os serviços

```
marcelo@BD-DESKTOP-UBUNTU:~$ cd Documentos/infnet/mineracao_elk/
marcelo@BD-DESKTOP-UBUNTU:~/Documentos/infnet/mineracao_elk$ docker-compose up
Starting elasticsearch ... done
Starting kibana        ... done
Attaching to elasticsearch, kibana
elasticsearch    | {"type": "server", "timestamp": "2020-12-08T10:07:19,454Z", "level": "INFO", "component": "o.e.n.Node
", "cluster.name": "docker-cluster", "node.name": "elastic_node01", "message": "version[7.9.2], pid[6], build[default/do
cker/d34da0ea4a966c4e49417f2da2f244e3e97b4e6e/2020-09-23T00:45:33.626720Z], OS[Linux/5.4.0-56-generic/amd64], JVM[AdoptO
penJDK/OpenJDK 64-Bit Server VM/15/15+36]" }
elasticsearch    | {"type": "server", "timestamp": "2020-12-08T10:07:19,484Z", "level": "INFO", "component": "o.e.n.Node
", "cluster.name": "docker-cluster", "node.name": "elastic_node01", "message": "JVM home [/usr/share/elasticsearch/jdk]"
 }
elasticsearch    | {"type": "server", "timestamp": "2020-12-08T10:07:19,485Z", "level": "INFO", "component": "o.e.n.Node
", "cluster.name": "docker-cluster", "node.name": "elastic_node01", "message": "JVM arguments [-Xshare:auto, -Des.networ
kaddress.cache.ttl=60, -Des.networkaddress.cache.negative.ttl=10, -XX:+AlwaysPreTouch, -Xss1m, -Djava.awt.headless=true,
 -Dfile.encoding=UTF-8, -Djna.nosys=true, -XX:-OmitStackTraceInFastThrow, -XX:+ShowCodeDetailsInExceptionMessages, -Dio.
netty.noUnsafe=true, -Dio.netty.noKeySetOptimization=true, -Dio.netty.recycler.maxCapacityPerThread=0, -Dio.netty.alloca
tor.numDirectArenas=0, -Dlog4j.shutdownHookEnabled=false, -Dlog4j2.disable.jmx=true, -Djava.locale.providers=SPI,COMPAT,
 -Xms1g, -Xmx1g, -XX:+UseG1GC, -XX:G1ReservePercent=25, -XX:InitiatingHeapOccupancyPercent=30, -Djava.io.tmpdir=/tmp/ela
sticsearch-8339672847912995948, -XX:+HeapDumpOnOutOfMemoryError, -XX:HeapDumpPath=data, -XX:ErrorFile=logs/hs_err_pid%p.
log, -Xlog:gc*,gc+age=trace,safepoint:file=logs/gc.log:utctime,pid,tags:filecount=32,filesize=64m, -Des.cgroups.hierarch
y.override=/, -XX:MaxDirectMemorySize=536870912, -Des.path.home=/usr/share/elasticsearch, -Des.path.conf=/usr/share/elas
ticsearch/config, -Des.distribution.flavor=default, -Des.distribution.type=docker, -Des.bundled_jdk=true]" }
kibana           | {"type":"log","@timestamp":"2020-12-08T10:07:24Z","tags":["warning","plugins-discovery"],"pid":6,"mes
```

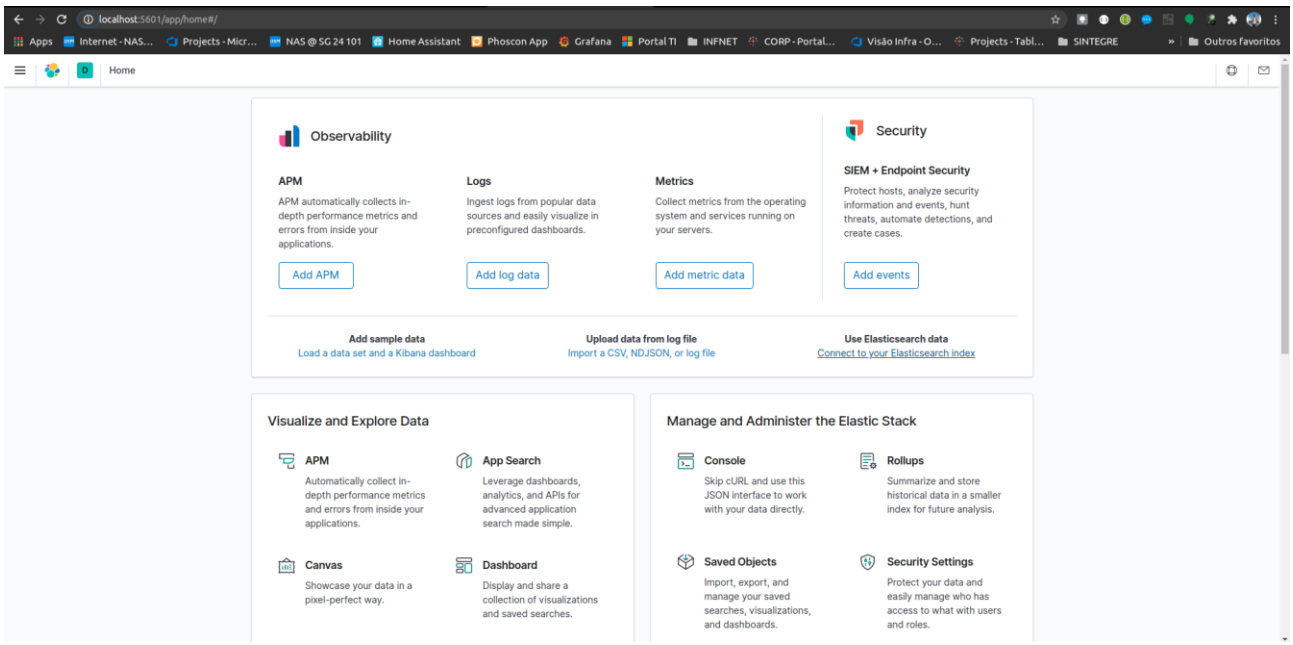
Após algum tempo verifique se o elasticsearch está respondendo, abrindo um browser e apontando para localhost:9200

Depois




```
{
  "name": "elastic_node01",
  "cluster_name": "docker-cluster",
  "cluster_uuid": "Bui7DUIOSd-FSP9v4pfrlw",
  "version": {
    "number": "7.9.2",
    "build_flavor": "default",
    "build_type": "docker",
    "build_hash": "d34da0ea4a966c4e49417f2da2f244e3e97b4e6e",
    "build_date": "2020-09-23T00:45:33.626720Z",
    "build_snapshot": false,
    "lucene_version": "8.6.2",
    "minimum_wire_compatibility_version": "6.8.0",
    "minimum_index_compatibility_version": "6.0.0-beta1"
  },
  "tagline": "You Know, for Search"
}
```


verifique se o kibana está respondendo, apontando o browser para localhost:5601

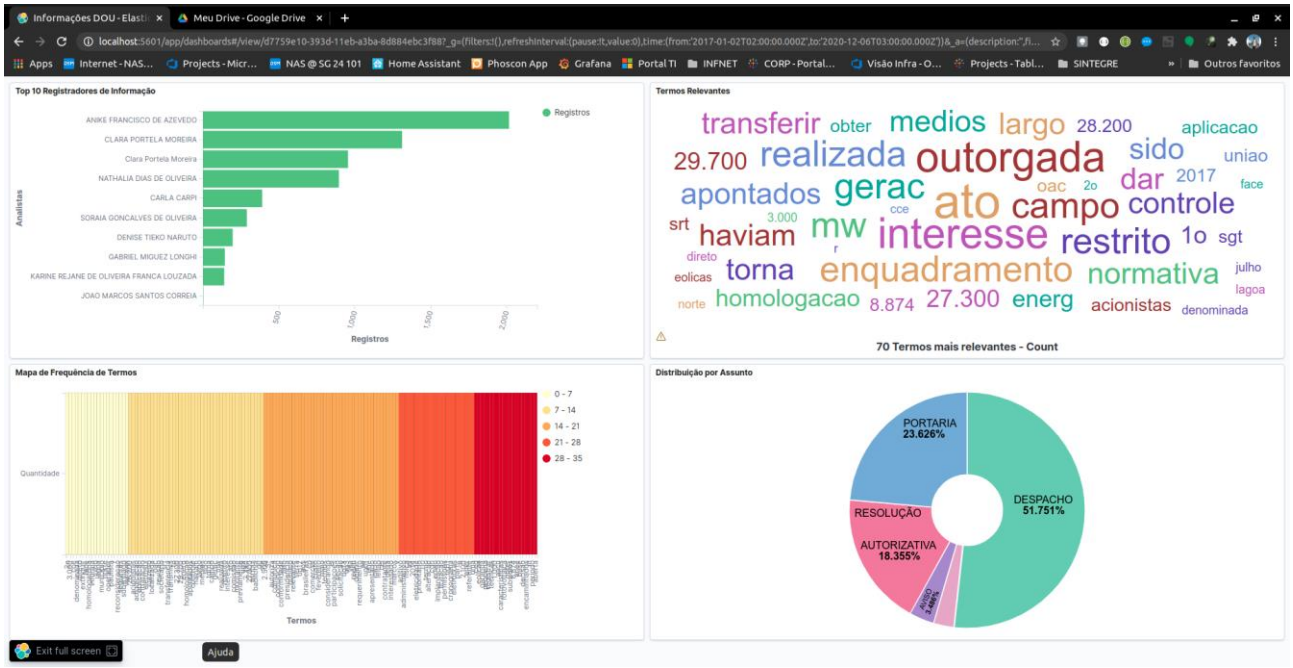


Notebooks

Os notebooks de carga de informações e consultas podem ser observados nos arquivos: 1_carga_informacoes_dou_es.ipynb e 2_consultas_informacoes_dou_es.ipynb

Dashboard

Dashboard criado no kibana com visualizações dos top 10 registradores de informações, distribuição de documentos por tipo, frequência de termos e nuvem de palavras com os 70 termos mais relevantes.


Informações DOU - Elasti
Meu Drive - Google Drive
localhost:5601/app/dashboards#/view/d7759e10-393d-11eb-a3ba-8d884ebc3f88?_g=(filters:!(),refreshInterval:(pause:!t,value:0),time:(from:'2017-01-02T02:00:00.000Z',to:'2020-12-06T03:00:00.000Z'))&_a=(description:'',fi...
Apps
Internet - NAS...
Projects - Micr...
NAS @ SG 24 101
Home Assistant
Phoscon App
Grafana
Portal TI
INFNET
CORP - Portal...
Visão Infra - O...
Projects - Tabl...
SINTEGRE
Outros favoritos
Top 10 Registradores de Informação
Registros
ANIKE FRANCISCO DE AZEVEDO
CLARA PORTELA MOREIRA
Clara Portela Moreira
NATHALIA DIAS DE OLIVEIRA
CARLA CARPI
SORAIA GONCALVES DE OLIVEIRA
DENISE TIEKO NARUTO
GABRIEL MIGUEZ LONGHI
KARINE REJANE DE OLIVEIRA FRANCA LOUZADA
JOAO MARCOS SANTOS CORREIA
Analistas
500
1,000
1,500
2,000
Registros
Termos Relevantes
transferir
obter
medios
largo
28.200
aplicacao
29.700
realizada
outorgada
sido
uniao
oac
2o
dar
2017
face
apontados
gerac
ato
campo
controle
srt
3.000
cce
haviam
mw
r
interesse
restrito
1o
sgt
direto
eolicas
torna
enquadramento
normativa
julho
lagoa
norte
homologacao
8.874
27.300
energ
acionistas
denominada
70 Termos mais relevantes - Count
Mapa de Frequência de Termos
Quantidade
0 - 7
7 - 14
14 - 21
21 - 28
28 - 35
Termos
Distribuição por Assunto
PORTARIA
23.626%
DESPACHO
51.751%
RESOLUÇÃO
AUTORIZATIVA
18.355%
Exit full screen
Ajuda