



Mackenzie

Pós-graduação
Cientista de Dados – 2021

Cientista de Dados nas Organizações

Coleta e armazenamento de dados

Atividade de Aprofundamento Trilha 6

“Conhecendo Banco de Dados **NoSQL**”

Aluno: Denyson Tomaz de Lima
Matr. 92174337

1. Índice

1.	Índice	2
2.	Introdução.....	3
3.	Tarefa	3
4.	Resultado:.....	9

2. Introdução

Trabalho referente a trilha 6:

Agora já conhecemos os bancos de dados NoSQL e as suas características. Vamos então trabalhar estes conceitos na nossa última atividade de aprofundamento. Selecione um banco de dados NoSQL de sua preferência. Crie uma instância de teste e faça a inserção de dados a partir de um arquivo csv.

Utilize a base de dados World Happiness disponível no site do Kaggle:
<https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021>

Apresente um notebook com o código utilizado e uma evidência do cadastro realizado. Obs.: Alguns bancos de dados possuem área de sandbox disponível em nuvem. Se você tiver alguma dificuldade em baixar o banco, utilize a opção em nuvem.

Todos os arquivos estão disponíveis no github
Site: <https://github.com/DenysonLima/ColArmDados>

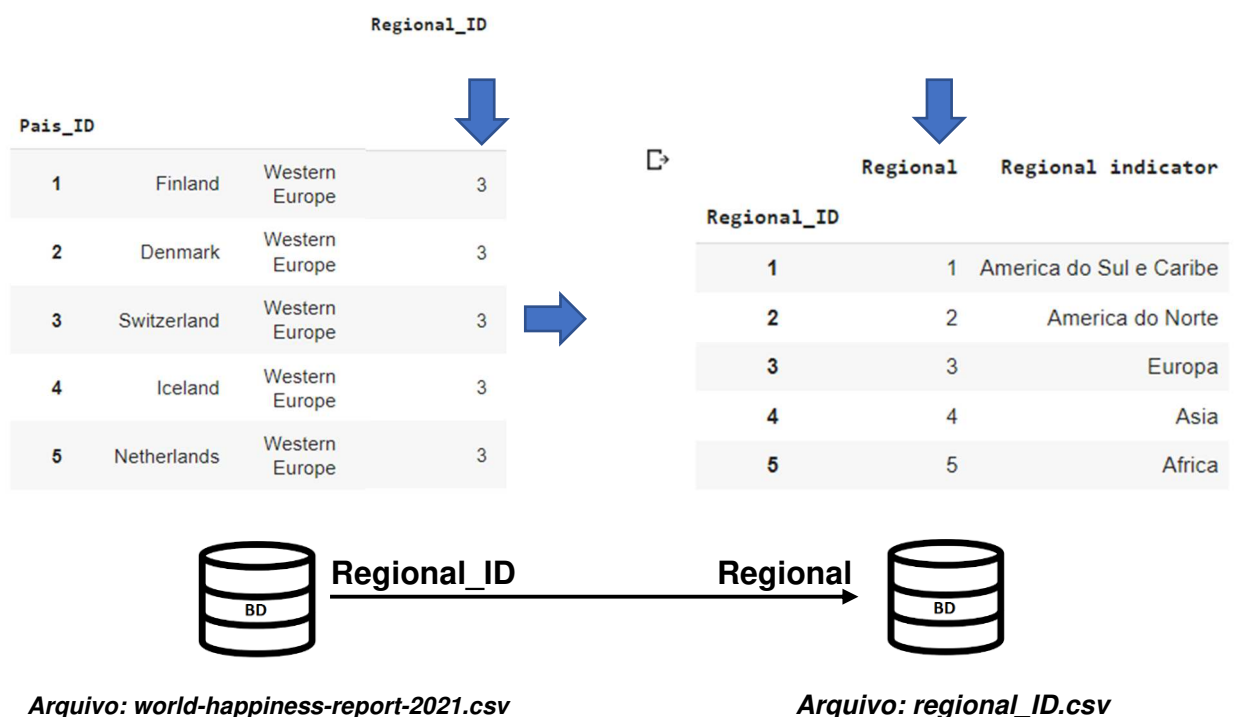
3. Tarefa

Através do site do Kaggle, foi feito o download do arquivo: world-happiness-report-2021.csv.

Este arquivo serviu de base do projeto desta tarefa para ser carregado no Banco de dados NoSQL no site: <https://sandbox.neo4j.com/>.

Antes de inserir o arquivo no Banco de dados no Neo4j, foi feita uma preparação e atualização do Banco de dados world-happiness-report-2021.csv, anexando duas novas colunas.

No arquivo world-happiness-report-2021.csv foi acrescentado uma coluna que se refere ao continente e a região na qual apelidamos de Regional_ID. Esta coluna contém um número de 1 a 6 que representa a região do continente, identificado como:



Foi criado um arquivo chamado de regional_ID que contém a associação das regiões continentais (1 a 6).

O arquivo world-happiness-report-2021.csv foi preparado para receber a nova coluna (Regional_ID) identificando a região. Foi aplicado o programa PYTHON para realizar esta inserção da nova coluna, veja abaixo o programa em PYTHON.

Iniciação:

```
import pandas as pd
import numpy as np
mundo_feliz = pd.read_csv('world-happiness-report-2021.csv')
```

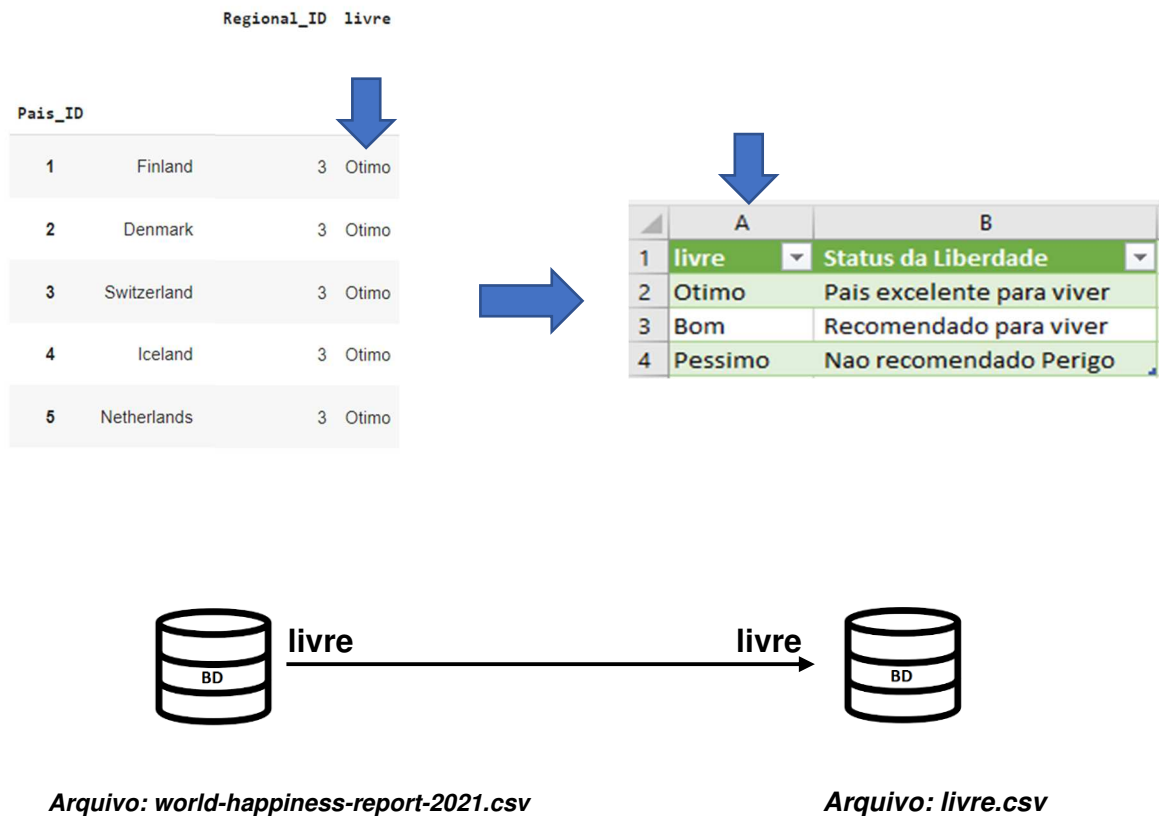
Criando uma nova coluna que reflete a da tabela da região continental (America, Europa, Asia, Africa e Estados Independentes)

```
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Latin America and
Caribbean', 'Regional_ID'] = '1'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'North America and
ANZ', 'Regional_ID'] = '2'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Central and Eastern Europe', 'Regional_ID'] = '3'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Western Europe', 'Regional_ID'] = '3'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'East Asia', 'Regional_ID'] = '4'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Southeast Asia', 'Regional_ID'] = '4'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'South Asia', 'Regional_ID'] = '4'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Sub-Saharan Africa', 'Regional_ID'] = '5'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Middle East and North Africa', 'Regional_ID'] = '5'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Commonwealth of Independent States', 'Regional_ID'] = '6'
mundo_feliz.index = mundo_feliz.index + 1
mundo_feliz.index.names = ['Pais_ID']

mundo_feliz.to_csv('mundo_feliz_2021.csv')
mundo_feliz.head()
```

whisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual	Regional_ID
7.780	10.775	0.954	72.0	0.949	-0.098	0.186	2.43	1.446	1.106	0.741	0.691	0.124	0.481	3.253	3
7.552	10.933	0.954	72.7	0.946	0.030	0.179	2.43	1.502	1.108	0.763	0.686	0.208	0.485	2.868	3
7.500	11.117	0.942	74.4	0.919	0.025	0.292	2.43	1.566	1.079	0.816	0.653	0.204	0.413	2.839	3
7.438	10.878	0.983	73.0	0.955	0.160	0.673	2.43	1.482	1.172	0.772	0.698	0.293	0.170	2.967	3
7.410	10.932	0.942	72.4	0.913	0.175	0.338	2.43	1.501	1.079	0.753	0.647	0.302	0.384	2.798	3

No arquivo world-happiness-report-2021.csv também foi acrescentado outra coluna que se refere uma indexação de liberdade de livre escolha do povo. Foi mapeado como: Péssimo, Bom e Ótimo, apenas para manipular o Banco de dados, sem objetivo de realizar uma pesquisa série com dados relevante. O intuito modesto tarefa é de apenas manipular e visualizar o Banco e dados através do NoSQL, aqui representado pelo Neo4j.




O arquivo world-happiness-report-2021.csv revisão A foi preparado para receber também uma nova coluna (livre) identificando a satisfação da liberdade do povo. Foi aplicado o programa PYTHON para realizar esta inserção da nova coluna, veja abaixo o programa em PYTHON.

Criando uma coluna que reflete a da tabela da satisfação de liberdade de uma região continental (America, Europa, Asia, Africa e Estados Independentes)

```

mundo_feliz['livre'] = None
mundo_feliz.loc[mundo_feliz['Explained by: Freedom to make life choices'] < 0.3, 'livre'] = 'Pessimo'
mundo_feliz.loc[(mundo_feliz['Explained by: Freedom to make life choice s'] >= 0.3) & (mundo_feliz['Explained by: Freedom to make life choices'] < 0.6), 'livre'] = 'Bom'
mundo_feliz.loc[mundo_feliz['Explained by: Freedom to make life choices'] >= 0.6, 'livre'] = 'Otimo'
mundo_feliz.to_csv('mundo_feliz_2021.csv')
mundo_feliz.head()

```



	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual	Regional_ID	livre
30	10.775	0.954	72.0	0.949	-0.098	0.186	2.43	1.446	1.106	0.741	0.691	0.124	0.481	3.253	3	Otimo
32	10.933	0.954	72.7	0.946	0.030	0.179	2.43	1.502	1.108	0.763	0.686	0.208	0.485	2.868	3	Otimo
30	11.117	0.942	74.4	0.919	0.025	0.292	2.43	1.566	1.079	0.816	0.653	0.204	0.413	2.839	3	Otimo
38	10.878	0.983	73.0	0.955	0.160	0.673	2.43	1.482	1.172	0.772	0.698	0.293	0.170	2.967	3	Otimo
10	10.932	0.942	72.4	0.913	0.175	0.338	2.43	1.501	1.079	0.753	0.647	0.302	0.384	2.798	3	Otimo

Após a preparação dos arquivos e criação de dois novos podemos carregar no site <https://sandbox.neo4j.com/> e manipulá-los. Veja o comando no site. Siga os passos abaixo:

- 1- Conecte ao site <https://sandbox.neo4j.com/> com sua chave e senha.
- 2- Crie um projeto do tipo blank
- 3- Abra o projeto criado
- 4- No comando abaixo digite a linha de carga (load) para realizar o download dos arquivos csv. (Nota: os arquivos deverão estar disponibilizados na web. Neste trabalho utilizamos a facilidade do sistema GitHub).

4.1 Carregando o arquivo mundo_feliz_2021.csv (World Happiness Report)



```
1 LOAD CSV WITH HEADERS FROM "https://raw.githubusercontent.com/DenysonLima/ColArmDados/main/mundo_feliz_2021.csv" AS row
2 CREATE (w:WorldHappinessReport)
3 SET w = row
```

Comando:

```
LOAD CSV WITH HEADERS FROM
"https://raw.githubusercontent.com/DenysonLima/ColArmDados/main/mundo_feliz_2021.csv" AS row
CREATE (w:WorldHappinessReport)
SET w = row
```

4.2 Carregando o arquivo livre.csv



```
1 LOAD CSV WITH HEADERS FROM "https://raw.githubusercontent.com/DenysonLima/ColArmDados/main/regional_ID.csv" AS row
2 CREATE (r:Regional)
3 SET r = row
```

Comando:

```
LOAD CSV WITH HEADERS FROM
"https://raw.githubusercontent.com/DenysonLima/ColArmDados/main/livre.csv"
AS row
CREATE (l:livre)
SET l = row
```

4.3 Carregando o arquivo regional_ID.csv

```
1 LOAD CSV WITH HEADERS FROM "https://raw.githubusercontent.com/DenysonLima/ColArmDados/main/regional_ID.csv" AS row
2 CREATE (r:Regional)
3 SET r = row
```

Comando:

```
LOAD CSV WITH HEADERS FROM
"https://raw.githubusercontent.com/DenysonLima/ColArmDados/main/regional_ID.csv" AS row
CREATE (r:Regional)
SET r = row
```

- 5- Após o carregamento dos três arquivos, devemos indexar todos eles, veja abaixo o comando:

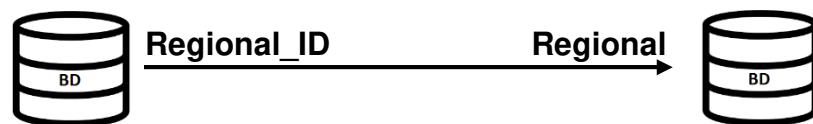
Comando:

```
CREATE INDEX ON :WorldHappinessReport(Pais_ID)
CREATE INDEX ON :WorldHappinessReport(livre)
CREATE INDEX ON :RegionalIndicator(Regional)
```

```
neo4j$ CREATE INDEX ON :RegionalIndicator(Regional)
```

- 6- Depois de carregado e indexado devemos realizar a operação do tipo “Join” ou relacionamentos entre os arquivos identificando os ID’s comuns entre os arquivos.

```
1 MATCH (w:WorldHappinessReport),(r:Regional)
2 WHERE w.Regional_ID = r.Regional
3 CREATE (w)-[:PART_OF]->(r)
```



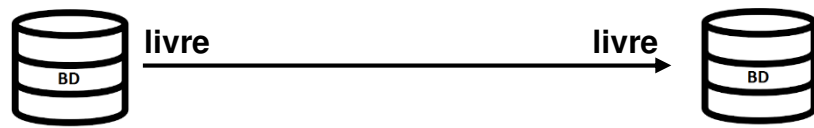
Arquivo: world-happiness-report-2021.csv

Arquivo: regional_ID.csv

Comando:

```
MATCH (w:WorldHappinessReport),(r:Regional)
WHERE w.Regional_ID = r.Regional
CREATE (w)-[:PART_OF]->(r)
```

```
https://6bd6554e9b674e2b0b8b39c8b958.neo4j sandbox.com/browser/?token=pwfetch6bd6554e9b674e2b0b8b39c8b958ey/hbGoOUsJat1NlslmR5cCi6kgXVClsimtpZCi6lIFubENPRV4U...  
MATCH (w:WorldHappinessReport),(l:livre)  
WHERE w.livre = l.livre  
CREATE (w)-[:PART_OF]->(l)
```



Arquivo: world-happiness-report-2021.csv

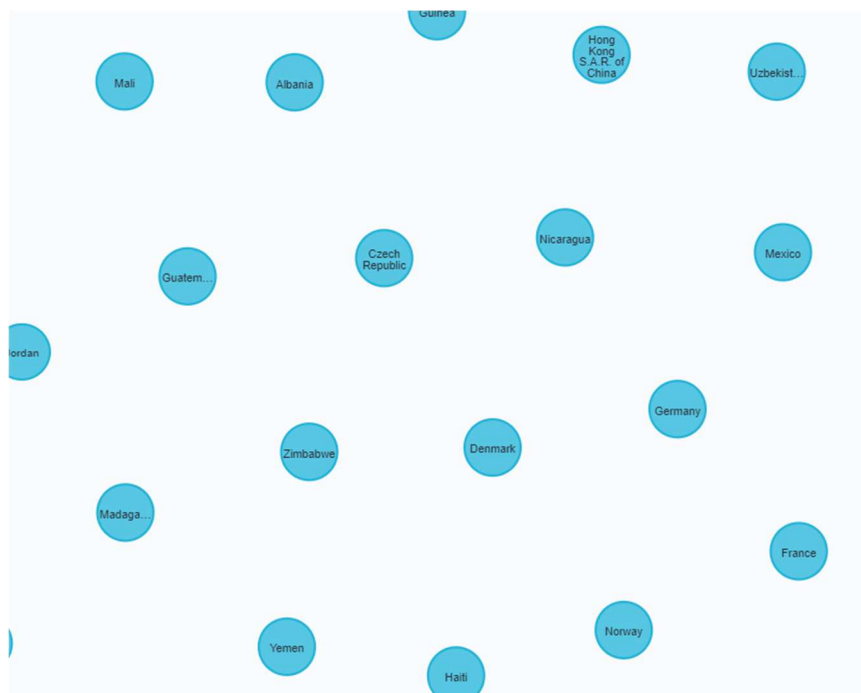
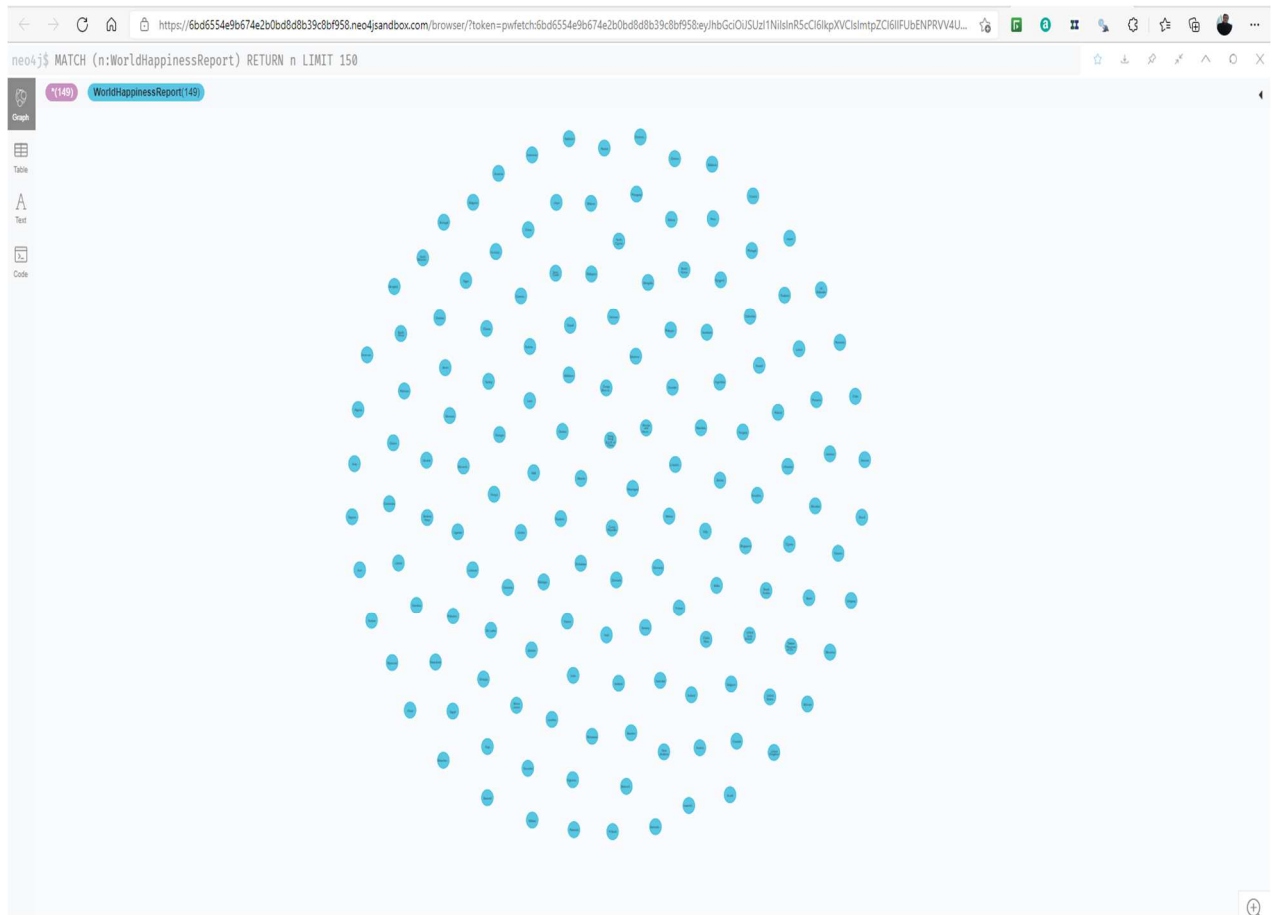
Arquivo: livre.csv

Comando:

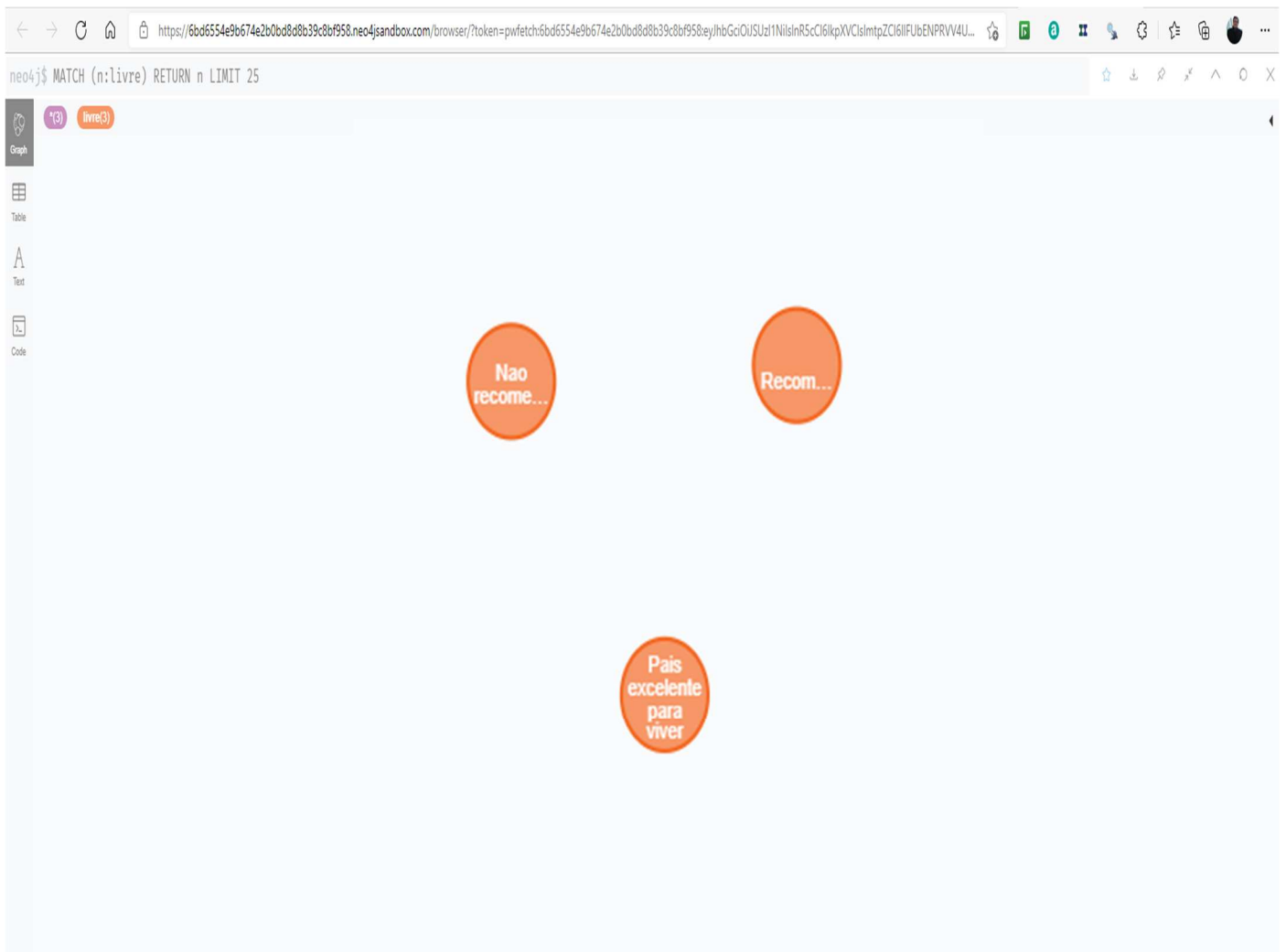
```
MATCH (w:WorldHappinessReport),(l:livre)  
WHERE w.livre = l.livre  
CREATE (w)-[:PART_OF]->(l)
```


4. Resultado:

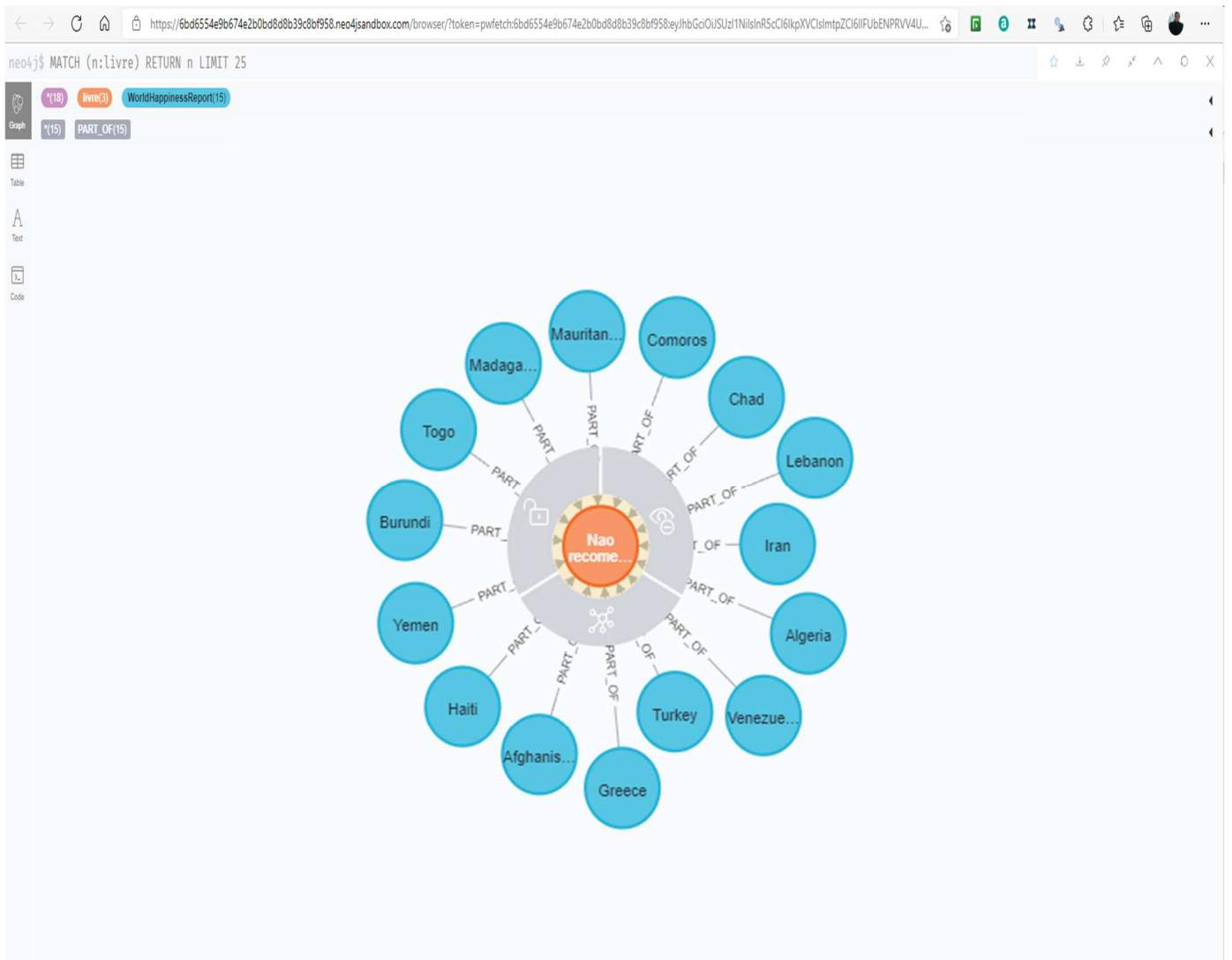
Após o carregamento dos arquivos através do comando LOAD CSV WITH HEADERS FROM <https://www...> e a criação da indexação dos arquivos pertinentes através do comando CREATE INDEX ON <nome do arquivo.csv> e finalmente elaborado todos os relacionamentos entre os CSV's, teremos o seguinte resultado da visualização.



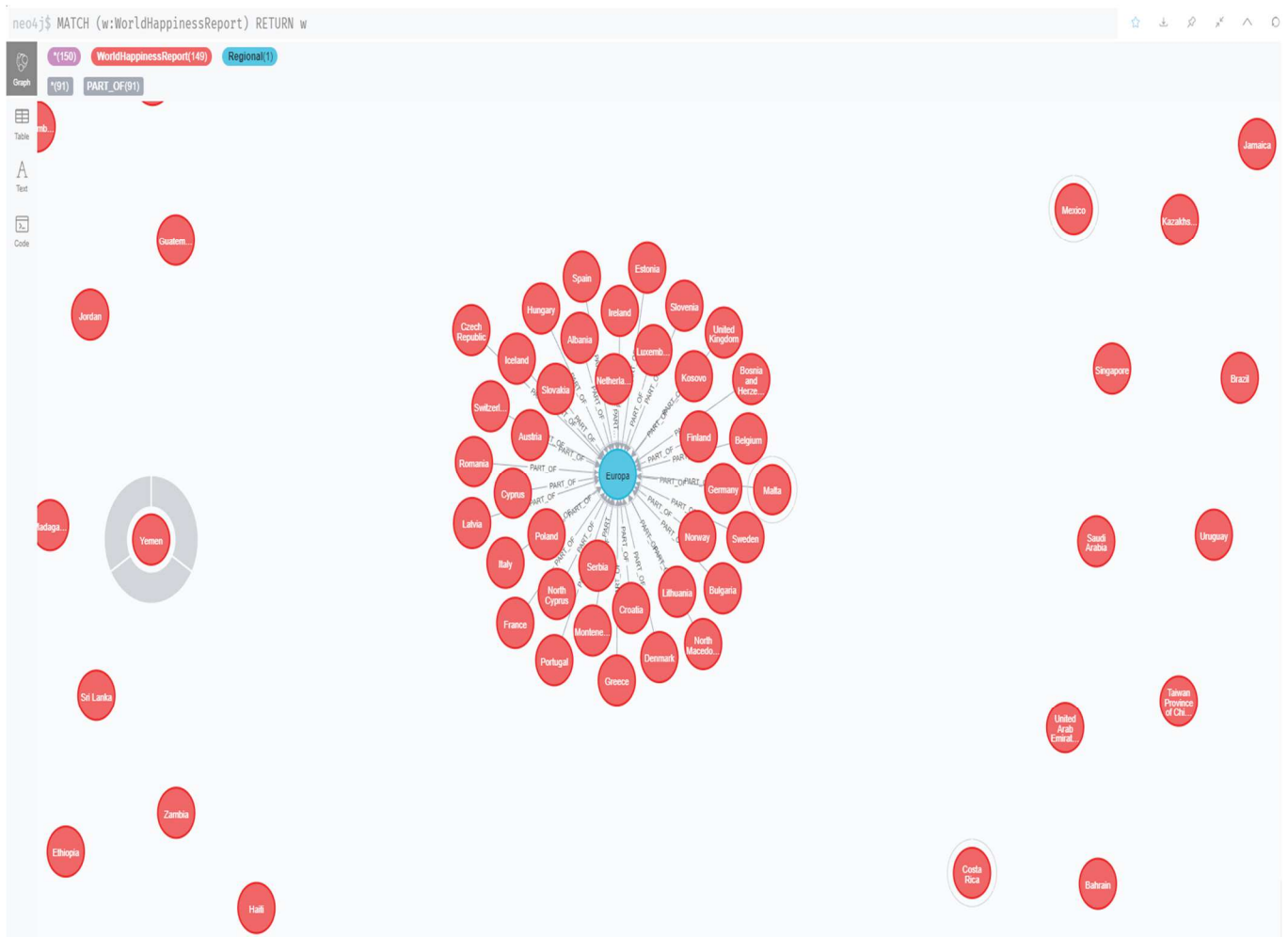
Todos os países. (zoom) (mundo_feliz_2021.csv)



Categoria Liberdade para fazer escolhas (livre.csv)



Países Não recomendado para viver conforme coluna no
arquivo mundo_feliz_2021.csv



Países agrupado conforme o relacionamento regional.

Adotando a regra elaborado no PYTHON:

```

mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Latin America and Caribbean', 'Regional_ID'] = '1'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'North America and ANZ', 'Regional_ID'] = '2'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Central and Eastern Europe', 'Regional_ID'] = '3'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Western Europe', 'Regional_ID'] = '3'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'East Asia', 'Regional_ID'] = '4'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Southeast Asia', 'Regional_ID'] = '4'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'South Asia', 'Regional_ID'] = '4'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Sub-Saharan Africa', 'Regional_ID'] = '5'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Middle East and North Africa', 'Regional_ID'] = '5'
mundo_feliz.loc[mundo_feliz['Regional indicator'] == 'Commonwealth of Independent States', 'Regional_ID'] = '6'

```

Regional_ID	Regional indicator
1	1 America do Sul e Caribe
2	2 America do Norte
3	3 Europa
4	4 Asia
5	5 Africa

