Prova MAC0459 - 2021

Professor: Roberto Hirata
Aluno: Daniel Angelo Esteves Lawand **NUSP**: 10297693

14. 01. 2022

1 Questão 1

1.1 Task 2

1.1.1 Descrição do dataset

Tomando o conjunto de dados de acidentes aéreos no Brasil, utilizaremos as planilhas "ocorrencia.csv"e "ocorrencia_tipo.csv"para responder as perguntas previamente estabelecidas. A planilha "ocorrencia.csv"aponta diversos dados sobre as ocorrências aéreas, porém usaremos apenas a data e a localização da ocorrência. A planilha "ocorrencia_tipo.csv"é uma planilha mais enxuta, porém possui dados descritivos sobre o tipo de ocorrência, e estes dados são os que usaremos na nossa análise.

1.1.2 Estratégia

A ideia é popular a base de dados, e depois fazer queries que retornem as informações que buscamos. Para isso, iremos construir dois dataframes com pandas, cada qual correspondendo a uma das tabelas que usaremos. Introduziremos duas novas colunas no dataframe df_ocorrencias, uma indicando o mês e a outra o ano da ocorrência. Após isso, iremos fazer os comandos que criarão os "nodes"no Neo4j. Tendo os comandos, iremos enviar ao Neo4j, para realizar de fato a criação dos "nodes".

Cada "node"do tipo "ocorrencia"terá como atributos o ano, o mês, a UF e a cidade da ocorrência aérea; e cada node "ocorrencia_tipo"terá o tipo de ocorrência como atributo. Após ter criado todos os nós, iremos fazer as queries, que basicamente seguem o mesmo padrão entre si, pois são queries que retornam a quantidade de "nodes"de acordo com cada atributo.

1.2 Task 3

Podemos perceber algumas diferenças entre a abordagem de EDA e de Neo4j. A primeira diferença é a configuração de setup, as ferramentas de EDA são mais simples de serem configuradas e portanto demandam menos tempo de configuração, enquanto as ferramentas de Neo4j exigem um pouco mais de quem está manuseando. Outra diferença a ser ressaltada é a maior iteratividade que a ferramenta de Notebook fornece à abordagem de EDA, podendo executar, ver o resultado e possivelmente alterar o

código com maior facilidade. Por outro lado, Neo4j oferece uma iteratividade similar, após a construção e o populamento do banco de dados é possível fazer queries com certa facilidade.

Outra diferença técnica, para EDA é necessário aprender uma linguagem de programação amplamente usada que é o python, enquanto para Neo4j é necessário aprender uma linguagem que se conecte com o banco de dados - python por exemplo - e é necessário aprender a linguagem Cypher para fazer as queries, e esta linguagem não é amplamente usada.

Além das diferenças técnicas, EDA aborda de forma mais simples, mas eficiente para os nossos problemas, já Neo4j é um banco de dados orientado a grafos capaz de lidar com muitos dados altamente conectados, é uma ferramente muito potente, mas que para a realidade que trabalhamos nessa atividade não era de extrema necessidade.