

TECNOLOGIA EM SISTEMAS PARA INTERNET

Kálita Rodrigues de Souza
Lucas Nascimento Verdam de Araújo
Nikolle de Lacerda Nascimento
Pedro Henrique Fernandes de Santana

**RELATÓRIO DE PRÁTICA INTEGRADA
DE
CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA**

Brasília - DF

26/01/2022

Sumário

1. Objetivos	3
2. Descrição do problema	4
3. Desenvolvimento	5
3.1 Código implementado	5
4. Considerações finais	6
Referências	7

1. Objetivos

Os objetivos dessa sprint são armazenar os dados que foram coletados, explorados e preparados na *sprint* 1 e após isso será feita uma análise dos dados para retirar algum valor dos dados.

2. Descrição do problema

A problemática possui as seguintes principais etapas:

- Criar uma conta no *MongoDB Atlas*;
- Entender como o *pymongo* e o *mongoDB* funcionam;
- Conectar-se com o banco de dados pelo *pymongo*;
- Criar o banco de dados;
- Criar coleção;
- Pegar os dados da preparação da *sprint* 1 e inserir na coleção criada;

3. Desenvolvimento

As principais tecnologias utilizadas no desenvolvimento desta tarefa foram:

Python

Python é uma linguagem de programação de alto nível, do termo em inglês, *high level language*. Estruturas desse tipo são, geralmente, classificadas como orientadas a objetos. (COUTINHO, 2019). De acordo com Roveda (2021), ao trabalhar com ciência de dados, desenvolver em *Python* pode ser uma alternativa, pois o crescente número de bibliotecas disponíveis em *Python* voltadas à análise de dados oferece funções e métodos de otimização para praticamente quaisquer objetivos.

Google Colaboratory

O *Google Colaboratory*, ou *Colab*, é um serviço de nuvem gratuito hospedado pelo *Google* para incentivar a pesquisa de Aprendizado de Máquina e Inteligência Artificial (SANTOS, 2020). A ferramenta funciona com uma estruturação em células. Os códigos rodados dentro dessas células podem ser rodados separadamente ou em conjunto, permitindo uma programação mais dinâmica (NOLETO, 2020). A plataforma roda linguagens como *Python* e *Jupyter*.

Github

É uma plataforma online de trabalho colaborativo. A primeira parte do nome, “*Git*”, é por causa da utilização do sistema de controle de versão e a segunda parte, “*Hub*”, tem a ver com a conexão entre profissionais de programação de qualquer lugar do mundo. (SOUZA, 2020).

Criamos um *notebook* para o projeto na plataforma *Google Colab*, pois ela possibilita o *upload* dos arquivos de dados bem como o desenvolvimento compartilhado dos códigos necessários para se realizar as tarefas solicitadas. Neste *notebook*, para a codificação dos *scripts* necessários para coleta, preparação e exploração dos dados utilizou-se a linguagem *Python* por ser uma das mais apropriadas e usadas para tarefas relacionadas à ciência de dados. Além é claro, dos módulos e bibliotecas necessários para tal, sendo eles: a biblioteca *Pandas* e o módulo ‘os’ do *python* para a coleta e preparação; a biblioteca *Matplotlib* e *PandasSQL* para a exploração.

MongoDB

Esse banco de dados não relacional e orientado a documentos é um dos mais utilizados no mercado. Foi lançado em 2009 e é escrito de C++ o que o faz estar disponível em várias plataformas (MONGO, 2022).

3.1 Código implementado

A seguir será apresentado o código que foi implementado nessas *sprint*:

1.7 - Armazenamento:

Primeiro, instalar o *mongoDB* e importá-lo, assim como importar o *Pandas*

```
pip install pymongo[srv]
!pip install dnspython
import pymongo
import pandas as pd
```

Conexão com o MongoDB e criando um banco de dados:

```
myclient =
pymongo.MongoClient("mongodb://lucanis12t:lucanis12t@cluster-shard-00-
00
.as99a.mongodb.net:27017,cluster-shard-00-01.as99a.mongodb.net:27017,c
lu
ster-shard-00-02.as99a.mongodb.net:27017/myFirstDatabase?ssl=true&repl
ic
aSet=atlas-fbdlnt-shard-0&authSource=admin&retryWrites=true&w=majority
")
mydb = myclient["Cluster"]
```

Verificar se existe um banco de dados no sistema:

```
print(myclient.list_database_names())
```

Criando coleção:

```
mycol = mydb["Moves"]
print(mydb.list_collection_names())
```

Ler o arquivo dados_preparados e insere seus dados na coleção do banco de dados:

```
df = pd.read_csv('dados_preparados.csv')

for i in range(len(df)):
    mydict =
    {"X":int(df['X'][i]),"Y":int(df['Y'][i]),"Z":int(df['Z'][i]),"Ano":int(df
['Ano'][i]),"Mes":int(df['Mes'][i]),"Dia":int(df['Dia'][i]),"Horas":int(d
f['Horas'][i]),"Minutos":int(df['Minutos'][i]),"Segundos":int(df['Segundo
s'][i]),"Tipo_movimento":df['Tipo_movimento'][i],"Genero":df['Genero'][i]
,"Num_voluntario":int(df['Num_voluntario'][i]),"Media":int(df['Media'][i]
)}
    x = mycol.insert_one(mydict)
```

1.8 - Análise:

Importar *MongoClient*, *Pandas* e arquivos do *google.colab*

```
from pymongo import MongoClient
import pandas as pd
from google.colab import files
```

Conexão com o *mongoDB*:

```
myclient =
MongoClient("mongodb://lucanis12t:lucanis12t@cluster-shard-00-00.as99a
.m
ongodb.net:27017,cluster-shard-00-01.as99a.mongodb.net:27017,cluster-s
ha
rd-00-02.as99a.mongodb.net:27017/myFirstDatabase?ssl=true&replicaSet=a
tl as-fbdlnt-shard-0&authSource=admin&retryWrites=true&w=majority")
```

Salva a *database* e a coleção em uma variável:

```
banco = myclient["Cluster"]  
coll = banco["Moves"]
```

Lista todos os documentos da coleção e passa-os para o *dataframe*:

```
mydoc = coll.find()  
  
df = pd.DataFrame(list(mydoc))
```

Excluindo a coluna `_id` dos dados:

```
df = df.drop(columns=['_id'])
```

Transforma o *dataframe* em um arquivo .csv:

```
df.to_csv('Medidas.csv', encoding='utf-8', index=False)
```


4. Considerações finais

A Sprint foi realizada com desempenho satisfatório. O grupo conseguiu realizar as tarefas sem grandes problemas.

Referências

Pandas- <https://pandas.pydata.org/docs/>

PandaSql- <https://pypi.org/project/pandasql/>

Python- <https://docs.python.org/3/>

Matplotlib- <https://matplotlib.org/>

Sobre correlação

<https://escoladedados.org/tutoriais/correlacao-nao-e-causalidade-mas-o-que-e-entao/>

COUTINHO, Thiago. O que é Python e pra que serve?. **Voitto**, 2019. Disponível em: <<https://www.voitto.com.br/blog/artigo/python>>. Acesso em: 01 de aug. de 2021.

NOLETO, Cairo. Google Colab: saiba o que é essa ferramenta e como usar. **Trybe**, 2020. Disponível em: <<https://blog.betrybe.com/carreira/google-colab/>>. Acesso em: 29 de jul. de 2021.

ROVEDA, Ugo. O que é Python, para que serve e por que aprender?. **Kenzie**, 2021. Disponível em: <<https://kenzie.com.br/blog/o-que-e-python/>>. Acesso em: 01 de aug. de 2021.

SANTOS, Thiago G. Google Colab: O que é, Tutorial de Como Usar e Criar Códigos. **Alura**, 2020. Disponível em: <<https://www.alura.com.br/artigos/google-colab-o-que-e-e-como-usar>>. Acesso em: 29 de jul. de 2021.

SOUZA, Ivan de. Entenda de uma vez o que é Github e a importância dele num negócio. **Rock Content**, 2020. Disponível em: <<https://rockcontent.com/br/blog/o-que-e-github/>>. Acesso em: 01 de aug. de 2021.

MONGO, db. O que é o MongoDB?. **MongoDB**, 2022. Disponível em: <[O que é o MongoDB? | MongoDB](#)>. Acesso em 26 de Jan. de 2022.