UFRJ - Data Warehouse (MAB 602)

Aluno: Giovani Tricarico Barros - DRE: 118051317

Professor: Geraldo Xexeo

Link do repositório Github: https://github.com/GiovaniTricaricoBarros/Prova-DWH-2021.1
Link do Collab: https://colab.research.google.com/drive/1k1hdJCRobR4goMGTPFMH6lALPGagk8ga#scrollTo=FrxubKdbrSP-

1. *Realize a coleta de dados, obtendo os microdados do ENADE dos anos de 2019, 2018 e 2017, no site do INEP, facilmente encontrados na rede.*

R: Com o Google Collab aberto e com os arquivos do ENADE dos anos de 2017, 2018 e 2019 pré-disponibilizados, foi elaborado um código para coletar automaticamente os dados do Enade através do seu próprio site (nomeados com “ano1, “ano2” e “ano3”), o *myzip* faz o *donwload* e extrai as pastas *.zip* e cria uma pasta.

```
[2]  import pandas as pd
     import matplotlib.pyplot as plt
     import numpy as np
     import zipfile
     import requests
     from io import BytesIO
     import os
     import plotly.graph_objs as go
     import plotly.offline as pyo

     os.makedirs("./Enade", exist_ok=True)

     ano1="https://download.inep.gov.br/microdados/Enade_Microdados/microdados_Enade_2017_portal_2018.10.09.zip"
     filebytes= BytesIO(requests.get(ano1).content)
     myzip = zipfile.ZipFile(filebytes)
     myzip.extractall("./Enade")
     enade2017 = pd.read_csv("/content/Enade/3.DADOS/MICRODADOS_ENADE_2017.txt",sep = ";", decimal = ",")

     ano2="https://download.inep.gov.br/microdados/Enade_Microdados/microdados_enade_2018.zip"
     filebytes= BytesIO(requests.get(ano2).content)
     myzip = zipfile.ZipFile(filebytes)
     myzip.extractall("./Enade")
     enade2018 = pd.read_csv("/content/Enade/2018/3.DADOS/microdados_enade_2018.txt",sep = ";", decimal = ",")

     ano3="https://download.inep.gov.br/microdados/Enade_Microdados/microdados_enade_2019.zip"
     filebytes= BytesIO(requests.get(ano3).content)
     myzip = zipfile.ZipFile(filebytes)
     myzip.extractall("./Enade")
     enade2019 = pd.read_csv("/content/Enade/3.DADOS/microdados_enade_2019.txt",sep = ";", decimal = ",")
```

Imagem 1. Código base

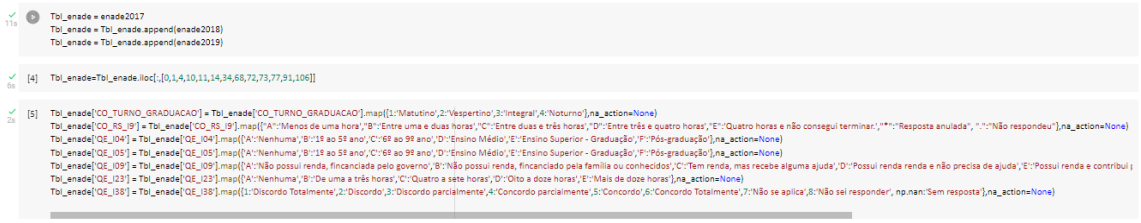

Imagem 2. Filtragem dos dados

2. *Crie um modelo dimensional (estrela) a partir dos dicionários de dados encontrados para os 3 anos.*

R: Analisando o dicionário de Dados, optei por 10 dimensões que se atrelavam as inscrições para fazer a prova do Enade. Partindo da tabela FATO, foi proposto as dimensões: *RegistroEstudante, TempoEnade, Ano, Instituição, Professores, Presença, TempoEstudo, SituaçãoFamília; Escolariadede; Curso.*

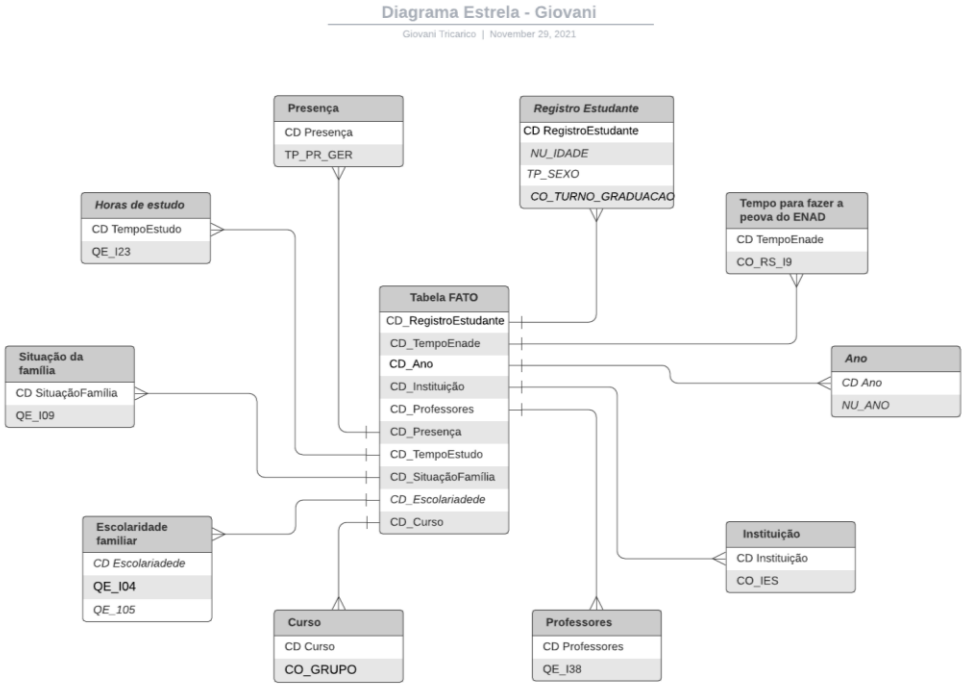

Imagem 2. Diagrama estrela

3. *Crie a base de dados do Data Warehouse em um banco de dados relacional, relatando o script SQL usado para isso.*

R: Foi importada a biblioteca SQLite3 para elaboração do banco de dados relacional, como solicitado no comando da questão.

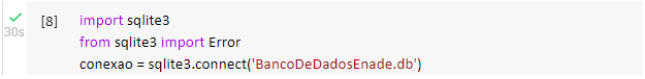

Imagem 3. Criação do Banco de dados

4. *Nesta questão é feita a carga de dados. Todos os dados devem ser alimentados em um banco de dados relacional, a sua escolha, local ou on-line, de acordo com o modelo dimensional planejado.*

R:

Foi elaborado as tabelas que seriam alimentadas dentro do banco de dados, posteriormente a isso, os dados foram devidamente implementados.

```
[7]  RegistroEstudante=Tbl_enade.iloc[:,[3,4,5]]
     TempoEnade=Tbl_enade.iloc[:,[7]]
     Ano=Tbl_enade.iloc[:,[0]]
     Instituição=Tbl_enade.iloc[:,[1]]
     Professores=Tbl_enade.iloc[:,[12]]
     Presença=Tbl_enade.iloc[:,[6]]
     TempoEstudo=Tbl_enade.iloc[:,[11]]
     SituaçãoFamília=Tbl_enade.iloc[:,[10]]
     Escolaridade=Tbl_enade.iloc[:,[8,9]]
     Curso=Tbl_enade.iloc[:,[2]]
```

```
Tbl_enade.to_sql('TabelaFato',conexao,if_exists='replace',index=False)
RegistroEstudante.to_sql('RegistroEstudante',conexao,if_exists='replace',index=False)
TempoEnade.to_sql('TempoEnade',conexao,if_exists='replace',index=False)
Ano.to_sql('Ano',conexao,if_exists='replace',index=False)
Instituição.to_sql('Instituição',conexao,if_exists='replace',index=False)
Professores.to_sql('Professores',conexao,if_exists='replace',index=False)
Presença.to_sql('Presença',conexao,if_exists='replace',index=False)
TempoEstudo.to_sql('TempoEstudo',conexao,if_exists='replace',index=False)
SituaçãoFamília.to_sql('SituaçãoFamília',conexao,if_exists='replace',index=False)
Escolaridade.to_sql('Escolaridade',conexao,if_exists='replace',index=False)
Curso.to_sql('Curso',conexao,if_exists='replace',index=False)
```

Imagem 4 e 5. Carga de Dados

5. *Nesta questão deve ser feita a análise de dados.*

R:

<u>5 perguntas a serem propostas:</u>

01- Quanto tempo a maioria dos alunos levou para fazer a prova anualmente?

02 - Qual perfil de estudo dos estudantes por ano, em relação ao período de horas?

03 - Quantas presenças confirmadas pelo Enade por edição?

04 – Qual o perfil médio de um estudante que presta a prova do Enade?

05 – Qual curso e instituição possui maior presença na prova do Enade?

Escolhi a pergunta de número 02 para responder, dela foi elaborada uma tabela e um gráfico, como solicitado no comando da questão.

QE_I23	Anos	De uma a três horas	Mais de doze horas	Nenhuma	Oito a doze horas	Quatro a sete horas
0	2017	192297	55742	18576	60636	140405
1	2018	228068	36497	42051	46075	132818
2	2019	161944	52859	15023	55217	122829

Imagem 6. Tabela da Pergunta 02

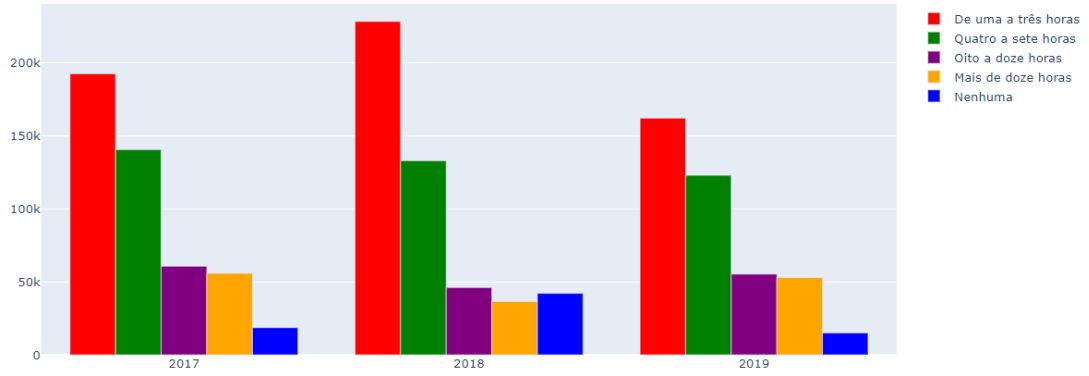

Imagem 7. Gráfico da Pergunta 02

Após análise do gráfico e da tabela, foi concluído que a maioria dos estudantes estuda de uma a três horas por semana.

6. *Nessa questão deve ser feita uma tentativa de aprender algo a partir dos dados.*

R: Não consegui elaborar uma tentativa de aprendizado de dados.

7. *Liste todas as ferramentas utilizadas, indicando o motivo da escolha. Inclua um link de referência para cada ferramenta.*

R: Para elaboração do trabalho, foi utilizado o Google Collab pelo fato de não ter o Python no computador utilizado (estou utilizando um computador emprestado) e não gostaria de baixar programas nesse. Link da ferramenta utilizada: https://colab.research.google.com/

Para elaboração do diagrama estrela, foi utilizado o Lucidchart, ferramenta já utilizada em outros trabalhos enviados para a disciplina e também pelo fato de ser bastante dedutiva e acessível. Link da ferramenta: https://lucid.app/