A3

May 11, 2023

[1]:
```
import requests
import zipfile
```

[2]:
```
url = "https://download.inep.gov.br/microdados/
 ↪microdados_censo_da_educacao_superior_2021.zip"
```

[3]:
```
r=requests.get(url, verify=False)
```

```
/opt/conda/lib/python3.10/site-packages/urllib3/connectionpool.py:1045:
InsecureRequestWarning: Unverified HTTPS request is being made to host
'download.inep.gov.br'. Adding certificate verification is strongly advised.
See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
```

[4]:
```
with open("r.zip","wb") as f:
    f.write(r.content)
```

[5]:
```
with zipfile.ZipFile("r.zip","r") as zip_ref:
    zip_ref.extractall("./data")
```

1 importanto sparksession

[6]:
```
from pyspark.sql import SparkSession
```

[7]:
```
spark = SparkSession \
.builder \
.config("spark.sql.repl.eagerEval.enabled", "True") \
.config("spark.sql.repl.eagerEval.maxNumRows", "10") \
.getOrCreate()
```

[13]:
```
df = spark.read.csv("data/Microdados do Censo da Educaç o Superior 2021/dados/
 ↪MICRODADOS_CADASTRO_IES_2021.CSV", sep=';', header=True, inferSchema=True)
```

[15]:
```
for col in df.columns:
    df = df.withColumnRenamed(col, col.lower())
```

[16]:
```
df.createOrReplaceTempView('df')
```

2 transformando csv em parquet

[17]:
```
df.write.parquet("MICRODADOS_IES_CONSULTA.parquet")
```

[18]:
```
parquet_df = spark.read.parquet("MICRODADOS_IES_CONSULTA.parquet")
```

[19]:
```
parquet_df.createOrReplaceTempView("parquet_df")
```

3 consultas

4 Olhando para o estado de Minas Gerais, quantos municípios têm informações presentes na base de dados?

[20]:
```
p1 = spark.sql("""
select COUNT(DISTINCT no_municipio_ies) from parquet_df
where sg_uf_ies = 'MG'
""")
p1.show()
```

```
+--------------------------------+
|count(DISTINCT no_municipio_ies)|
+--------------------------------+
|                             104|
+--------------------------------+
```

5 Quantos professores doutores existem em cada cidade de Minas Gerais presente na base de dados

[24]:
```
spark.sql("""SELECT no_municipio_ies as municipio, sum(qt_doc_ex_dout) as␣
 ↪Qtd_Doutores from parquet_df WHERE no_uf_ies = 'Minas Gerais'
         group by no_municipio_ies order by Qtd_Doutores DESC""")
```

[24]:
```
+----------------+------------+
|       municipio|Qtd_Doutores|
+----------------+------------+
|  Belo Horizonte|        6615|
|      Uberl ndia|        1939|
|    Juiz de Fora|        1915|
|          Vi osa|        1185|
|   Montes Claros|         925|
```

```
|         Uberaba|         891|
|          Lavras|         812|
|S o Jo o del Rei|         753|
|      Ouro Preto|         746|
|      Diamantina|         673|
+----------------+------------+
only showing top 10 rows
```

6 Qual a quantidade de docentes com deficiência no estado do Paraná?

[23]:
```
spark.sql("SELECT SUM(qt_doc_ex_com_deficiencia) FROM parquet_df WHERE␣
 ↪sg_uf_ies = 'PR' ")
```

[23]:
```
+-----------------------------+
|sum(qt_doc_ex_com_deficiencia)|
+-----------------------------+
|                          160|
+-----------------------------+
```

[ ]:
```

```