PySpark: Tratamento de Dados e Big Data

O objetivo desse projeto é demonstrar conhecimentos em PySpark. Para esse projeto, utilizei um dataset de domínio público de estatísticas do Youtube: https://www.kaggle.com/datasets/advaypatil/youtube-statistics/ /

PySpark: Data Processing and Big Data

The objective of this project is to demonstrate knowledges in PySpark. For this project, I utilized a public domain Youtube statistics Dataset: https://www.kaggle.com/datasets/advaypatil/youtube-statistics/

```
# Instalando o PySpark / Installing PySpark
! pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.12/dist-packages (4.0.1)
Requirement already satisfied: py4j==0.10.9.9 in /usr/local/lib/python3.12/dist-packages (from pyspark) (0.10.9.9
```

```
# Importando PySpark e SparkSession / Importing PySpark and SparkSession
import pyspark
from pyspark.sql import SparkSession
```

```
# Criando uma sessão Spark / Creating a Spark session
spark = SparkSession.builder.getOrCreate()
```

```
# Lendo os dados do arquivo “videos-stats.csv” / Reading data from the “videos-stats.csv” file
df = spark.read.option('header', 'true').csv('videos-stats.csv')
```

```
# Visualizando os primeiros 8 registros do arquivo / Viewing the first 8 records of the file
df.show(8)
```

```
+---+--------------------+-----------+------------+-------+--------+--------+---------+
|_c0|               Title|   Video ID|Published At|Keyword|   Likes|Comments|    Views|
+---+--------------------+-----------+------------+-------+--------+--------+---------+
|  0|Apple Pay Is Kill...|wAZZ-UWGVHI|  2022-08-23|   tech|  3407.0|   672.0| 135612.0|
|  1|The most EXPENSIV...|b3x28s61q3c|  2022-08-24|   tech| 76779.0|  4306.0|1758063.0|
|  2|My New House Gami...|4mgePWWCAmA|  2022-08-23|   tech| 63825.0|  3338.0|1564007.0|
|  3|Petrol Vs Liquid ...|kXiYSI7H2b0|  2022-08-23|   tech| 71566.0|  1426.0| 922918.0|
|  4|Best Back to Scho...|ErMwWXQxHp0|  2022-08-08|   tech| 96513.0|  5155.0|1855644.0|
|  5|Brewmaster Answer...|18fwz9Itbvo|  2021-11-05|   tech| 33570.0|  1643.0| 943119.0|
|  6|Tech Monopolies: ...|jXf04bhcjbg|  2022-06-13|   tech|135047.0|  9367.0|5937790.0|
|  7|I bought the STRA...|2TqOmtTAMRY|  2022-08-07|   tech|216935.0| 12605.0|4782514.0|
+---+--------------------+-----------+------------+-------+--------+--------+---------+
only showing top 8 rows
```

```
# Visualizando o esquema do arquivo / Viewing the file schema
df.printSchema()
```

```
root
 |-- _c0: string (nullable = true)
 |-- Title: string (nullable = true)
 |-- Video ID: string (nullable = true)
 |-- Published At: string (nullable = true)
 |-- Keyword: string (nullable = true)
 |-- Likes: string (nullable = true)
 |-- Comments: string (nullable = true)
 |-- Views: string (nullable = true)
```

Para os propósito desse projeto será utilizado o inferSchema, mas aqui está um exemplo de como eu definitira o schema manualmente:

For the purposes of this project inferSchema will be utilized, but here's and example of how I would define the schema manually:

```
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DateType, LongType

schema = StructType([
  StructType("_c0", IntegerType(), True),
  StructType("Title", StringType(), True),
  StructType("Video ID", StringType(), True),
  StructType("Published At", DateType(), True),
  StructType("Keyword", StringType(), True),
  StructType("Likes", IntegerType(), True),
  StructType("Comments", IntegerType(), True),
  StructType("Views", LongType(), True),
])
```

```
# Lendo novamente o arquivo inferindo o esquema e visualizando o esquema novamente / Reading the file again, inf
df = spark.read.option('header', 'true').option('inferSchema', 'true').csv('videos-stats.csv')
df.printSchema()
```

```
root
 |-- _c0: integer (nullable = true)
 |-- Title: string (nullable = true)
 |-- Video ID: string (nullable = true)
 |-- Published At: date (nullable = true)
 |-- Keyword: string (nullable = true)
 |-- Likes: double (nullable = true)
 |-- Comments: double (nullable = true)
 |-- Views: double (nullable = true)
```

```
# Salvando o arquivo como 'videos-parquet' no formato parquet e adicionando o cabeçalho nos dados / Saving the f
df.write.option('header', 'true').option('inferSchema', 'true').save('output/videos-parquet')
```

```
# Lendo e visualizando o arquivo 'videos-parquet' com cabeçalho nos dados / Reading and viewing the 'videos-parq
df = spark.read.option('header', 'true').option('inferSchema', 'true').parquet('output/videos-parquet')
df.show(10)
```

```
+---+--------------------+-----------+------------+-------+--------+--------+---------+
|_c0|               Title|   Video ID|Published At|Keyword|   Likes|Comments|    Views|
+---+--------------------+-----------+------------+-------+--------+--------+---------+
|  0|Apple Pay Is Kill...|wAZZ-UWGVHI|  2022-08-23|   tech|  3407.0|   672.0| 135612.0|
|  1|The most EXPENSIV...|b3x28s61q3c|  2022-08-24|   tech| 76779.0|  4306.0|1758063.0|
|  2|My New House Gami...|4mgePWWCAmA|  2022-08-23|   tech| 63825.0|  3338.0|1564007.0|
|  3|Petrol Vs Liquid ...|kXiYSI7H2b0|  2022-08-23|   tech| 71566.0|  1426.0| 922918.0|
|  4|Best Back to Scho...|ErMwWXQxHp0|  2022-08-08|   tech| 96513.0|  5155.0|1855644.0|
|  5|Brewmaster Answer...|18fwz9Itbvo|  2021-11-05|   tech| 33570.0|  1643.0| 943119.0|
|  6|Tech Monopolies: ...|jXf04bhcjbg|  2022-06-13|   tech|135047.0|  9367.0|5937790.0|
|  7|I bought the STRA...|2TqOmtTAMRY|  2022-08-07|   tech|216935.0| 12605.0|4782514.0|
|  8|15 Emerging Techn...|wLlL46pYcg4|  2021-12-08|   tech| 45565.0|  2882.0|7001236.0|
|  9|Toxicologist Answ...|R7qsau3X6Ks|  2022-07-14|   tech| 24252.0|  1068.0| 667767.0|
+---+--------------------+-----------+------------+-------+--------+--------+---------+
only showing top 10 rows
```

```
# Salvando o arquivo do exec. anterior como tabela chamada ‘tb_videos’ no banco de dados default do spark catalo
df.write.option('header', 'true').option('inferSchema', 'true').saveAsTable('tb_videos')
```

```
# Listando as tabelas do spark catalog para verificar a tabela / Listing the Spark catalog tables to verify the
spark.catalog.listTables()
```

```
[Table(name='tb_videos', catalog='spark_catalog', namespace=['default'], description=None, tableType='MANAGED',
isTemporary=False)]
```

```
# Utilizando o spark SQL para ler a tabela ‘tb_videos’ / Using Spark SQL to read the ‘tb_videos’ table
tab_df = spark.sql('SELECT * FROM tb_videos')
tab_df.show()
```

```
+---+--------------------+-----------+------------+-------+--------+--------+-----------+
|_c0|               Title|   Video ID|Published At|Keyword|   Likes|Comments|      Views|
+---+--------------------+-----------+------------+-------+--------+--------+-----------+
|  0|Apple Pay Is Kill...|wAZZ-UWGVHI|  2022-08-23|   tech|  3407.0|   672.0|   135612.0|
|  1|The most EXPENSIV...|b3x28s61q3c|  2022-08-24|   tech| 76779.0|  4306.0|  1758063.0|
|  2|My New House Gami...|4mgePWWCAmA|  2022-08-23|   tech| 63825.0|  3338.0|  1564007.0|
|  3|Petrol Vs Liquid ...|kXiYSI7H2b0|  2022-08-23|   tech| 71566.0|  1426.0|   922918.0|
|  4|Best Back to Scho...|ErMwWXQxHp0|  2022-08-08|   tech| 96513.0|  5155.0|  1855644.0|
|  5|Brewmaster Answer...|18fwz9Itbvo|  2021-11-05|   tech| 33570.0|  1643.0|   943119.0|
|  6|Tech Monopolies: ...|jXf04bhcjbg|  2022-06-13|   tech|135047.0|  9367.0|  5937790.0|
|  7|I bought the STRA...|2TqOmtTAMRY|  2022-08-07|   tech|216935.0| 12605.0|  4782514.0|
|  8|15 Emerging Techn...|wLlL46pYcg4|  2021-12-08|   tech| 45565.0|  2882.0|  7001236.0|
|  9|Toxicologist Answ...|R7qsau3X6Ks|  2022-07-14|   tech| 24252.0|  1068.0|   667767.0|
| 10|Dope Tech: The Mo...|MEiq0oCUb_8|  2022-08-15|   tech|118001.0|  4123.0|  2359142.0|
| 11|Cool Tech Under $...|pT_9hntWj34|  2022-08-06|   tech| 20999.0|  3091.0|   413179.0|
| 12|Cool Back to Scho...|cj4lxmHQV0o|  2022-08-13|   tech| 15322.0|   547.0|   389114.0|
| 13|Best Tech/EDC Gif...|d-BdIo8_wpA|  2021-12-15|   tech| 17866.0|   157.0|   444953.0|
| 14|My Massive Tech U...|eFhhW6fsAbQ|  2022-07-09|   tech| 13217.0|   442.0|   371563.0|
| 15|Why Millennials A...|N88OE2ZCHBM|  2021-06-04|   tech| 26890.0|  4999.0|  1633059.0|
| 16|10 Coolest Gadget...|PKATJiyz0iI|  2021-08-13|   tech|  9562.0|   199.0|   760249.0|
| 17|17 Coolest Gadget...|qiMnSaZWf3M|  2022-07-21|   tech| 11743.0|   143.0|  1086568.0|
| 18|Almost EVERYONE i...|4AnyhHl3_tE|  2022-08-14|   tech|146978.0| 11105.0|  3186890.0|
| 19|I bought the THIN...|nmY2kgWYwyQ|  2022-03-25|   tech|363771.0| 13609.0|1.1422924E7|
+---+--------------------+-----------+------------+-------+--------+--------+-----------+
only showing top 20 rows
```

```
# Lendo o arquivo ‘comments.csv' inferindo o esquema e visualizando / Reading the ‘comments.csv' file, inferring
df = spark.read.option('header', 'true').option('inferSchema', 'true').csv('comments.csv')
```

```
df.show()
```

```
+--------------+-----------+--------------------+------+---------+
|           _c0|   Video ID|             Comment| Likes|Sentiment|
+--------------+-----------+--------------------+------+---------+
|             0|wAZZ-UWGVHI|Let's not forget ...|  95.0|      1.0|
|             1|wAZZ-UWGVHI|Here in NZ 50% of...|  19.0|      0.0|
|             2|wAZZ-UWGVHI|I will forever ac...| 161.0|      2.0|
|             3|wAZZ-UWGVHI|Whenever I go to ...|   8.0|      0.0|
|             4|wAZZ-UWGVHI|Apple Pay is so c...|  34.0|      2.0|
|             5|wAZZ-UWGVHI|We’ve been houndi...|   8.0|      1.0|
|             6|wAZZ-UWGVHI|We only got Apple...|  29.0|      2.0|
|             7|wAZZ-UWGVHI|For now, I need b...|   7.0|      1.0|
|             8|wAZZ-UWGVHI|In the United Sta...|   2.0|      2.0|
|             9|wAZZ-UWGVHI|In Cambodia, we h...|  28.0|      1.0|
|            10|b3x28s61q3c|Wow, you really w...|1344.0|      2.0|
|            11|b3x28s61q3c|The lab is the mo...| 198.0|      2.0|
|            12|b3x28s61q3c|Linus, I'm an eng...| 365.0|      2.0|
|            13|b3x28s61q3c|There used to be ...| 211.0|      2.0|
|            14|b3x28s61q3c|Holy crap. I was ...| 821.0|      0.0|
|            15|b3x28s61q3c|I love the direct...| 150.0|      2.0|
|            16|b3x28s61q3c|I am more excited...|  49.0|      2.0|
|            17|b3x28s61q3c|I adore the worki...|  19.0|      2.0|
|            18|b3x28s61q3c|LMGs growth is ho...|  NULL|     NULL|
|More technical|   in depth| engineering orie...|  17.0|      2.0|
+--------------+-----------+--------------------+------+---------+
only showing top 20 rows
```

```
# Salvando arquivo como ‘comments-parquet' no formato parquet e adicionando o cabeçalho nos dados / Saving the f
df.write.option('header', 'true').option('inferSchema', 'true').save('comments-parquet')
```