databricks Csv - Parquet

(https://databricks.com)

Convertento arquivo csv em parquet

- **spark.read.format("csv") Definindo o formato para leitura**
- **.option("header", "True") Defindo cabeçalho**
- **.option("delimiter", ";") Definir separador de colunas no arquivo**
- **.option("inferSchema", "True") Inferir de forma automatica o tipo de dado de cada coluna**
- **.load("/FileStore/tables/siconv_etapa_crono_fisico.csv") Carregar arquivo e especificar o caminho**

3

```
# Lendo arquivos csv do dbfs com Spark
df = (
    spark.read.format("csv")\
    .option("header", "True")\
    .option("delimiter", ";")\
    .option("inferSchema", "True")\
    .load("/FileStore/tables/siconv_etapa_crono_fisico.csv")
)
```

▸ df: pyspark.sql.dataframe.DataFrame = [ID_META: string, ID_ETAPA: integer ... 11 more fields]

4

```
# Verificando o Schema do dataframe
df.printSchema()
```

```
root
 |-- ID_META: string (nullable = true)
 |-- ID_ETAPA: integer (nullable = true)
 |-- NR_ETAPA: integer (nullable = true)
 |-- DESC_ETAPA: string (nullable = true)
 |-- DATA_INICIO_ETAPA: string (nullable = true)
 |-- DATA_FIM_ETAPA: string (nullable = true)
 |-- UF_ETAPA: string (nullable = true)
 |-- MUNICIPIO_ETAPA: string (nullable = true)
 |-- ENDERECO_ETAPA: string (nullable = true)
 |-- CEP_ETAPA: string (nullable = true)
 |-- QTD_ETAPA: string (nullable = true)
 |-- UND_FORNECIMENTO_ETAPA: string (nullable = true)
 |-- VL_ETAPA: string (nullable = true)
```

5

```
# Verificar as 10 primeiras linhas
display(df.show(10))
```

```
neiro Fran...|87700-000|     null|                null|  150000|
|   1234|    1111|       1|palestras educativas|       18/08/2008|    24/11/2008|      PR|      PARANAVAI|Em dive
```

```
|   1242|    1150|       1|Compra da prensa ...|       01/09/2008|    01/09/2009|      MG|     TAIOBEIRAS|
null|     null|     null|                  null|   10000|
|   1243|    1120|       1|Aquisição de pren...|       01/09/2008|    01/09/2009|      MG|     TAIOBEIRAS|Praça d
a Matriz, ...|39550-000|     null|                  null|    5800|
|   1243|    1119|       2|Construção do gal...|       01/09/2008|    01/09/2009|      MG|     TAIOBEIRAS|Praça d
a Matriz, ...|39550-000|     null|                  null|   84200|
+-------+--------+--------+--------------------+----------------+--------------+--------+---------------+-------
------------+---------+---------+----------------------+--------+
only showing top 10 rows
```

6

```
# Contando a quantidade de linhas
df.count()
```

```
Out[23]: 3011901
```

Leva os dados convertidos para Processing Zone

- **df.write.format("parquet") Define o formato de saída do arquivo**
- **.mode("overwrite") Especifica o modo de gravação e sobrescrever os arquivo no diretório**
- **.save("/FileStore/tables/csv to parquet") Define o caminho onde os arquivos serão salvos**

9

```
# Converte para o formato parquet
df.write.format("parquet")\
.mode("overwrite")\
.save("/FileStore/tables/csv to parquet")
```

- **df_parquet = spark.read.format("parquet")\ Definindo um novo DataFrame**
- **.load("/FileStore/tables/csv to parquet" Define o caminho onde sera carregado**

11

```
# Lendo arquivos parquet
df_parquet = spark.read.format("parquet")\
.load("/FileStore/tables/csv to parquet")
```

```
▸ df_parquet:  pyspark.sql.dataframe.DataFrame = [ID_META: string, ID_ETAPA: integer ... 11 more fields]
```

12

```
# conta a quantidade de linhas
df_parquet.count()
```

```
Out[29]: 3011901
```

13

Table