Parquet - Read, Write / Show() / with Column :-

Read parquet file data into Dataframe

Using read.parquet("path") or read.format("parquet").load("path") you can read a parquet file into a PySpark DataFrame

```
df = spark.read.parquet('dbfs:/FileStore/data/userdata1.parquet')
display(df)
```

```
df = spark.read.parquet('dbfs:/FileStore/parquetdata/')
display(df)
print('count of dataframe is ' + str(df.count()))
```

(5) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [registration_dttm: timestamp, id

Table +

	registration_dttm	id	first_name
1	2016-02-03T07:55:29.000+0000	1	Amanda
2	2016-02-03T17:04:03.000+0000	2	Albert
3	2016-02-03T01:09:31.000+0000	3	Evelyn
4	2016-02-03T00:36:21.000+0000	4	Denise
5	2016-02-03T05:05:31.000+0000	5	Carlos
6	2016-02-03T07:22:34.000+0000	6	Kathryn
7	2016-02-03T08:33:08.000+0000	7	Samuel

Truncated results, showing first 1,000 rows. | 1.30 seconds runtime

```
count of dataframe is 2000
```

writeDFtoParquet Python

File Edit View Run Help Last edit was 1 minute ago Give feedback

1	1	Maheer
2	2	Wafa

Showing all 2 rows. | 6.83 seconds runtime

Cmd 2

```
df.write.parquet('/FileStore/parquetoutput', mode='overwrite')
```

(1) Spark Jobs

Command took 1.22 seconds -- by maheer3222@gmail.com at 10/25/2022, 2:30:42 PM

Cmd 3

```
df1 = spark.read.parquet('/FileStore/parquetoutput')
display(df1)
```

(4) Spark Jobs

df1: pyspark.sql.dataframe.DataFrame = [id: long, name: string]

Table +

	id	name
1	1	Maheer
2	1	Maheer

Write Dataframe into Parquet

Use the DataFrameWriter object to write PySpark DataFrame to a parquet file.

```
df.write.parquet('/FileStore/data/parquetoutput')
```

Show()

```
data = [(1, 'fdsajfsajfoiewanaoicnwcoewncoewvnwoav'),
        (2, 'vsavavewaompewmvowvv waoeviv wkev'),
        (3, 'qewqfewqfewbrwebebeb'),
        (4, 'noinoipninoinoinewcanwanv')]
schema = ['id','comments']

df = spark.createDataFrame(data=data,schema=schema)

df.show()
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [id: long, comments: string]

```
+---+--------------------+
| id|            comments|
+---+--------------------+
|  1|fdsajfsajfoiewana...|
|  2|vsavavewaompewmvo...|
|  3|qewqfewqfewbrwebebeb|
|  4|noinoipninoinoine...|
+---+--------------------+
```

In default the Show function only show 20 height is more than 20 it will trunket.

```
data = [(1, 'fdsajfsajfoiewanaoicnwcoewncoewvnwoav'),
        (2, 'vsavavewaompewmvowvv waoeviv wkev'),
        (3, 'qewqfewqfewbrwebebeb'),
        (4, 'noinoipninoinoinewcanwanv')]
schema = ['id','comments']

df = spark.createDataFrame(data=data,schema=schema)

df.show(n=2, truncate=False)
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [id: long, comments: string]

```
+---+----------------------------------+
|id |comments                          |
+---+----------------------------------+
|1  |fdsajfsajfoiewanaoicnwcoewncoewvnwoav|
|2  |vsavavewaompewmvowvv waoeviv wkev |
+---+----------------------------------+
only showing top 2 rows
```

(n=2, truncate = False)

```
data = [(1, 'fdsajfsajfoiewanaoicnwcoewncoewvnwoav'),
        (2, 'vsavavewaompewmvowvv waoeviv wkev'),
        (3, 'qewqfewqfewbrwebebeb'),
        (4, 'noinoipninoinoinewcanwanv')]
schema = ['id','comments']

df = spark.createDataFrame(data=data,schema=schema)

df.show(n=4, truncate=False, vertical=True)
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [id: long, comments: string]

```
-RECORD 0----------------------------------
id       | 1
comments | fdsajfsajfoiewanaoicnwcoewncoewvnwoav
-RECORD 1----------------------------------
id       | 2
comments | vsavavewaompewmvowvv waoeviv wkev
-RECORD 2----------------------------------
id       | 3
comments | qewqfewqfewbrwebebeb
-RECORD 3----------------------------------
id       | 4
comments | noinoipninoinoinewcanwanv
```

(n=4, Truncate = False, vertical = True)

n = 4 → no. of rows.

Truncate false → Show all char like ∞ length can.

Show()

- DataFrame show() is used to display the contents of the DataFrame in a Table Row and Column Format.
- By default, it shows only 20 Rows, and the column values are truncated at 20 characters.
- To display full column contents, we should use "truncate=False" parameter inside show(). We can also truncate column value to desired length.

```
# Show full contents of column
df.show(truncate=False)
```

```
df.show(truncate=8)
```

- We can control rows to display as well as shown below.

```
df.show(n=1,truncate=False)
```

- We can display content vertically too as show below.

```
df.show(truncate=False,vertical=True)
```

withColumn()

- PySpark withColumn() is a transformation function of DataFrame which is used to change the value, convert the datatype of an existing column, create a new column, and many more

```
from pyspark.sql.functions import col

data = [(1,'Maheer','3000'),(2,'Wafa','4000')]
schema = ['id','name','salary']
df = spark.createDataFrame(data,schema)

#Change column datatype

df1 = df.withColumn("salary",col("salary").cast("Integer"))

df1.printSchema()
```

```
df.withColumn("salary",col("salary")*100).show()
```

```
df.withColumn("CopiedColumn",col("salary")* -1).show()
```

```
df.withColumn("Country", lit("USA"))
```

```
3  data = [(1, 'Maheer','3000'),(2, 'Wafa','4000')]
4  columns = ['id','name','salary']
5
6  df = spark.createDataFrame(data=data,schema=columns)
7
8  df1 = df.withColumn(colName='salary',col=col('salary').cast('Integer'))
9
10 df2 = df1.withColumn('salary', col('salary') * 2)
11
12 df3 = df2.withColumn('country',lit('india'))
13 df3.show()
```

Convert the data type to 'Integer'

Multiply the 'Salary' Column value with '2'

Create new 'Country' Column with 'India' Value in the table

▸ (3) Spark Jobs

▸ df: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 1 more field]
▸ df1: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 1 more field]
▸ df2: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 1 more field]
▸ df3: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 2 more fields]

```
+---+------+------+-------+
| id|  name|salary|country|
+---+------+------+-------+
|  1|Maheer|  6000|  india|
|  2|  Wafa|  8000|  india|
+---+------+------+-------+
```