pyspark-interview-qns

January 7, 2024

[0]:

```
from pyspark.sql import functions as F
```

Regex in Pyspark

[0]:

```
data = [(1, 'Rohit', 'J852485741'),
        (2,'Shubham', '6542879845'),
        (3, 'Sam', '854Y698547')]

schema = ['id','Name','Contact']
df = spark.createDataFrame(data, schema)
df.display()
```

[0]:

```
df.select("*").filter(F.col('contact').rlike('^[0-9]*$')).show()
```

```
+---+-------+----------+
| id|   Name|   Contact|
+---+-------+----------+
|  2|Shubham|6542879845|
+---+-------+----------+
```

Count rows in each column where NULLs present

[0]:

```
data = [(1, 'Rohit', 20),
        (2, None, 30),
        (3, 'Sam', None),
        (4, None, None),
        (5, None, 37)  ]

schema = ['id','name','age']

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

[0]:

```
df_cnt = df.select([F.count(F.when(F.col(i).isNull(),i)).alias('null_records in␣
 ↪' + i) for i in df.columns])
df_cnt.display()
```

How to remove delimiters

[0]:
```
schema = ['ID', 'NAME', 'Age', 'Marks']

data = [('1','A',20,'31|32|34'),
        ('2','B',21,'21|32|43'),
        ('3','C',22,'21|32|11'),
        ('4','D',23,'10|12|12')]

df_dlmtr = spark.createDataFrame(data, schema)
df_dlmtr.show()
```

```
+---+----+---+--------+
| ID|NAME|Age|   Marks|
+---+----+---+--------+
|  1|   A| 20|31|32|34|
|  2|   B| 21|21|32|43|
|  3|   C| 22|21|32|11|
|  4|   D| 23|10|12|12|
+---+----+---+--------+
```

[0]:
```
df_dlmtr = df_dlmtr.withColumn('Physics', F.split('Marks','\|')[0])\
                   .withColumn('Math', F.split('Marks','\|')[1])\
                   .withColumn('Eng', F.split('Marks','\|')[2])
df_dlmtr.show()
```

```
+---+----+---+--------+-------+----+---+
| ID|NAME|Age|   Marks|Physics|Math|Eng|
+---+----+---+--------+-------+----+---+
|  1|   A| 20|31|32|34|     31|  32| 34|
|  2|   B| 21|21|32|43|     21|  32| 43|
|  3|   C| 22|21|32|11|     21|  32| 11|
|  4|   D| 23|10|12|12|     10|  12| 12|
+---+----+---+--------+-------+----+---+
```

count Null percentage for each column

[0]:
```
data = [("Raj","Doe",None),
 (None,"Samuel","VIZAG"),
 ("David","Smith", None),
 ("Samson",None, "HYD"),
 ("Immi", "Steve", "BNG"),
 (None, None, None)]

schema = ["Firstname", "Lastname", "City"]

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

```
df.cache()
df.count()
df.show()
```

```
+---------+--------+-----+
|Firstname|Lastname| City|
+---------+--------+-----+
|      Raj|     Doe| null|
|     null|  Samuel|VIZAG|
|    David|   Smith| null|
|   Samson|    null|  HYD|
|     Immi|   Steve|  BNG|
|     null|    null| null|
+---------+--------+-----+
```

[0]:

```
for i in df.columns:
    total_count = df.select(F.col(i)).count()
    null_records = df.filter(F.col(i).isNull()).count()
    percentage = (null_records/total_count)*100
    print(i,round(percentage,2))
```

```
Firstname 33.33
Lastname 33.33
City 50.0
```