Pyspark changing type of column from date to string

I have the following dataframe:

```
corr_temp_df
[('vacationdate', 'date'),
 ('valueE', 'string'),
 ('valueD', 'string'),
 ('valueC', 'string'),
 ('valueB', 'string'),
 ('valueA', 'string')]
```

Now I would like to change the datatype of the column vacationdate to String, so that also the dataframe takes this new type and overwrites the datatype data for all of the entries. E.g. after writing:

```
corr_temp_df.dtypes
```

The datatype of vacationdate should be overwritten.

I already used functions like cast, StringType or astype, but I was not successful. Do you know how to do that?

python   apache-spark   apache-spark-sql   pyspark

1 Answer

Lets create some dummy data:

```python
import datetime
from pyspark.sql import Row
from pyspark.sql.functions import col

row = Row("vacationdate")

df = sc.parallelize([
    row(datetime.date(2015, 10, 07)),
    row(datetime.date(1971, 01, 01))
]).toDF()
```

If you Spark >= 1.5.0 you can use `date_format` function:

```python
from pyspark.sql.functions import date_format

(df
    .select(date_format(col("vacationdate"), "dd-MM-YYYY")
    .alias("date_string"))
    .show())
```

In Spark < 1.5.0 it can be done using Hive UDF:

```python
df.registerTempTable("df")
sqlContext.sql(
    "SELECT date_format(vacationdate, 'dd-MM-YYYY') AS date_string FROM df")
```

It is of course still available in Spark >= 1.5.0.

If you don't use `HiveContext` you can mimic `date_format` using UDF:

```
from pyspark.sql.functions import udf, lit
my_date_format = udf(lambda d, fmt: d.strftime(fmt))

df.select(
    my_date_format(col("vacationdate"), lit("%d-%m-%Y")).alias("date_string")
).show()
```

Please note it is using C standard format not a Java simple date format

edited Oct 7 '15 at 8:29   answered Oct 7 '15 at 8:21

zero323
65.9k   16   77   136