How to use Except function with spark Dataframe

Asked 2 years, 3 months ago Active 2 years, 3 months ago Viewed 11k times

2

3

I would like to get differences between two dataframe but returning the row with the different fields only. For example, I have 2 dataframes as follow:

```
val DF1 = Seq(
    (3,"Chennai",  "rahman",9846, 45000,"SanRamon"),
    (1,"Hyderabad","ram",9847, 50000,"SF")
).toDF("emp_id","emp_city","emp_name","emp_phone","emp_sal","emp_site")

val DF2 = Seq(
    (3,"Chennai",  "rahman",9846, 45000,"SanRamon"),
    (1,"Sydney","ram",9847, 48000,"SF")
).toDF("emp_id","emp_city","emp_name","emp_phone","emp_sal","emp_site")
```

The only difference between these two dataframe is `emp_city` and `emp_sal` for the second row. Now, I am using the `except` function which gives me the entire row as follow:

```
DF1.except(DF2)

+------+---------+--------+---------+-------+--------+
|emp_id| emp_city|emp_name|emp_phone|emp_sal|emp_site|
+------+---------+--------+---------+-------+--------+
|     1|Hyderabad|     ram|     9847|  50000|      SF|
+------+---------+--------+---------+-------+--------+
```

However, I need the output to be like this:

```
+---------+--------+-----+
|emp_id| emp_city|emp_sal|
+------+---------+-------+
|     1|Hyderabad|  50000|
+------+---------+-------+
```

Which shows the different cells as well as `emp_id` .

Edit : if there is change in column then it should appear if there is no change then it should be hidden or Null

scala apache-spark

Share Improve this question Follow

edited Dec 5 '18 at 4:59

Amin Mohebi
194 1 2 14

asked Dec 4 '18 at 6:51

milad ahmadi
365 3 5 15

How would the expected output look if the `emp_id=3` rows differ in, for example, the `emp_name` column? – Shaido Dec 4 '18 at 9:55

Also, different rows can differ in different columns. emp_name or emp_phone. How should the final result look like in that

case ? – user238607 Dec 4 '18 at 11:13

problem statement is not much clear – vikrant rana Dec 4 '18 at 13:57

if there is change in column then it should appear if there is no change then it should be hidden – milad ahmadi Dec 4 '18 at 23:47

2 Answers

Active | Oldest | Votes

1

I found this solution which seems to be working fine :

```
val cols = DF1.columns.filter(_ != "emp_id").toList
val DF3 = DF1.except(DF2)
def mapDiffs(name: String) = when($"l.$name" === $"r.$name", null
).otherwise(array($"l.$name", $"r.$name")).as(name)
val result = DF2.as("l").join(DF3.as("r"), "emp_id").select($"emp_id" ::
cols.map(mapDiffs): _*)
```

It generates the output as follow :

```
+------+------------------+--------+---------+--------------+--------+
|emp_id|          emp_city|emp_name|emp_phone|       emp_sal|emp_site|
+------+------------------+--------+---------+--------------+--------+
|     1|[Sydney, Hyderabad]|    null|     null|[48000, 50000]|    null|
|
+------+------------------+--------+---------+--------------+--------+
```

Share Improve this answer Follow

answered Dec 5 '18 at 1:09

Amin Mohebi

194 1 2 14

1

You should consider the comment from @user238607 as we cannot predict which columns are going to differ,

Still you can try this workaround.

I'm assuming `emp_id` is unique,

```
scala> val diff = udf((col: String, c1: String, c2: String) => if (c1 == c2) "" else
col )

scala> DF1.join(DF2, DF1("emp_id") === DF2("emp_id"))
res15: org.apache.spark.sql.DataFrame = [emp_id: int, emp_city: string ... 10 more
fields]

scala> res15.withColumn("diffcolumn", split(concat_ws(",",DF1.columns.map(x =>
diff(lit(x), DF1(x), DF2(x))):_*),","))
res16: org.apache.spark.sql.DataFrame = [emp_id: int, emp_city: string ... 11 more
fields]

scala> res16.show(false)
+------+---------+--------+---------+-------+--------+------+--------+--------+--------
-+-------+--------+--------------------------+
```

```
|emp_id|emp_city
|emp_name|emp_phone|emp_sal|emp_site|emp_id|emp_city|emp_name|emp_phone|emp_sal|emp_site|(
                |
+------+---------+--------+---------+------+--------+------+--------+--------+--------
-+-------+--------+--------------------------+
|3     |Chennai  |rahman  |9846     |45000  |SanRamon|3     |Chennai |rahman  |9846
|45000  |SanRamon|[, , , , , ]               |
|1     |Hyderabad|ram     |9847     |50000  |SF      |1     |Sydney  |ram     |9847
|48000  |SF      |[, emp_city, , , emp_sal, ]|
+------+---------+--------+---------+------+--------+------+--------+--------+--------
-+-------+--------+--------------------------+

scala> val diff_cols = res16.select(explode($"diffcolumn")).filter("col !=
''").distinct.collect.map(a=>col(a(0).toString))

scala> val exceptOpr = DF1.except(DF2)

scala> exceptOpr.select(diff_cols:_*).show

+-------+---------+
|emp_sal| emp_city|
+-------+---------+
|  50000|Hyderabad|
+-------+---------+
```

Share Improve this answer Follow

edited Dec 4 '18 at 12:32

answered Dec 4 '18 at 12:25

Sathiyan S

888 5 13