Home

About me

## ENSEMBLE BLOGGING

BEYOND DATA, SIGNAL, AND STATISTICS

## hat tip: join two spark dataframe on multiple columns (pyspark)

Author:

Follow @mamhamed

Labels: Big data, Data

Frame, Data Science, Spark

Thursday, September 24,

2015

G+1 Recommend this on Google

Consider the following two spark dataframes:

```
df1.show()
+----+
|id_a|time_a|value_a|
  1
            CA
  1
            CA
```

hat tip: join two spark dataframe on multiple columns (pyspark) | Ensemble Blogging

```
| 2| 1| TX|
| 3| 5| NE|
| 4| 6| WA|
+----+
```

```
df2.show()
+---+----+
|id_b|time_b| value_b|
+---+----+
| 1| 1| San Jose|
| 2| 1|Los Angeles|
| 2| 2| Austin|
+---+----+
```

Now assume, you want to join the two dataframe using both id columns and time columns. This can easily be done in *pyspark*:

```
df = df1.join(df2,(df1.id==df2.id_b)&(df1.time==df2.time),joinType="inner")
```

```
df.show()
+---+----+
|id_a|time_a|value_a|id_b|time_b| value_b|
+---+----+
```

hat tip: join two spark dataframe on multiple columns (pyspark) | Ensemble Blogging

Note that parentheses around the conditions is absolutely necessary.

## 3 COMMENTS:



Shikhar Agarwal January 19, 2016 at 3:15 AM

How to do this in Java?

Reply



Anonymous February 5, 2016 at 6:19 AM

a.col("x").equalTo(b.col("x")).and(a.col("y").equalTo(b.col("y"))

Reply

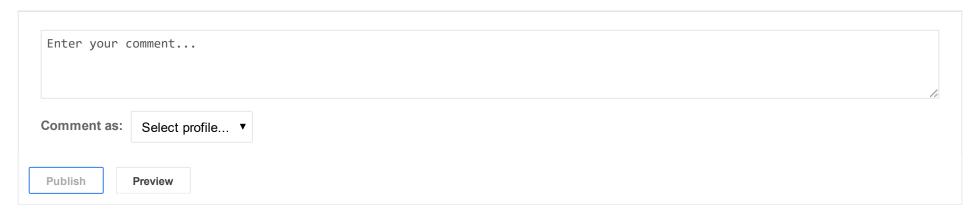


**SSWAP** June 16, 2016 at 4:18 AM

how to do it in Spark 2.0 using Dataset?

Reply

Add comment



Newer Post Older Post

Tags

Big data Data Frame Data Science Forecasting
Graph Query ipython notebook Model Validation R Spark Stats
Visualization

Total Pageviews

Recommended Blogs

R-Bloggers

Data Science central

Base blogs

## **Favorite Quotes**

"I have never thought of writing for reputation and honor. What I have in my heart must out; that is the reason why I compose." —Beethoven

"All models are wrong, but some are useful." --George Box



Copyright © 2015 • Ensemble Blogging

Created By Sora Templates and My Blogger Themes