spark-2

```
import pandas as pd
import numpy as np
import pyspark.pandas as ps
from pyspark.sql import SparkSession
```

```
df=spark.createDataFrame([
    ['red','banana',1,10],['blue','banana',2,20],['red','carrot',3,30],
['blue','grape',4,10],['red','carrot',5,50],['black','carrot',6,60],
['red','banana',7,70],['red','grape',8,80]],schema=
['color','fruit','v1','v2'])
df.show()
```

```
+-----+------+---+---+
|color| fruit| v1| v2|
+-----+------+---+---+
|  red|banana|  1| 10|
| blue|banana|  2| 20|
|  red|carrot|  3| 30|
| blue| grape|  4| 10|
|  red|carrot|  5| 50|
|black|carrot|  6| 60|
|  red|banana|  7| 70|
|  red| grape|  8| 80|
+-----+------+---+---+
```

```
df.groupby('color').avg().show()
```

```
+-----+-------+-------+
|color|avg(v1)|avg(v2)|
+-----+-------+-------+
|  red|    4.8|   48.0|
| blue|    3.0|   15.0|
|black|    6.0|   60.0|
+-----+-------+-------+
```

```
df.groupby('fruit').avg().show()
```

```
+------+------------------+------------------+
| fruit|           avg(v1)|           avg(v2)|
+------+------------------+------------------+
|banana|3.3333333333333335|33.333333333333336|
|carrot| 4.666666666666667|46.666666666666664|
| grape|               6.0|              45.0|
+-------+------------------+------------------+
```

```
df.groupby('color').count().show()
```

```
+-----+-----+
|color|count|
+-----+-----+
|  red|    5|
| blue|    2|
|black|    1|
+-----+-----+
```

```
df.groupby('v1').sum().show()
```

```
+---+-------+-------+
| v1|sum(v1)|sum(v2)|
+---+-------+-------+
|  1|      1|     10|
|  2|      2|     20|
|  3|      3|     30|
|  4|      4|     10|
|  5|      5|     50|
|  6|      6|     60|
|  7|      7|     70|
|  8|      8|     80|
+---+-------+-------+
```

```
df.groupby('fruit').max().show()
```

```
+------+-------+-------+
| fruit|max(v1)|max(v2)|
+------+-------+-------+
|banana|      7|     70|
|carrot|      6|     60|
| grape|      8|     80|
+------+-------+-------+
```

```
def plus_mean(pandas_df):
    return pandas_df.assign(v1= pandas_df.v1.sum())

df.groupby('color').applyInPandas(plus_mean, schema=df.schema).show()
```

```
+-----+------+---+---+
|color| fruit| v1| v2|
+-----+------+---+---+
|black|carrot|  6| 60|
| blue|banana|  6| 20|
| blue| grape|  6| 10|
|  red|banana| 24| 10|
|  red|carrot| 24| 30|
|  red|carrot| 24| 50|
|  red|banana| 24| 70|
|  red| grape| 24| 80|
+-----+------+---+---+
```

```
def plus_mean(pandas_df):
    return pandas_df.assign(v1= pandas_df.v1.max())

df.groupby('color').applyInPandas(plus_mean, schema=df.schema).show()
```

```
+-----+------+---+---+
|color| fruit| v1| v2|
+-----+------+---+---+
|black|carrot|  6| 60|
| blue|banana|  4| 20|
| blue| grape|  4| 10|
|  red|banana|  8| 10|
|  red|carrot|  8| 30|
|  red|carrot|  8| 50|
|  red|banana|  8| 70|
|  red| grape|  8| 80|
+-----+------+---+---+
```

```
def plus_mean(pandas_df):
    return pandas_df.assign(v1= pandas_df.v1.std())

df.groupby('color').applyInPandas(plus_mean, schema=df.schema).show()
```

```
+-----+------+----+---+
|color| fruit|  v1| v2|
+-----+------+----+---+
|black|carrot|null| 60|
| blue|banana|   1| 20|
| blue| grape|   1| 10|
|  red|banana|   2| 10|
|  red|carrot|   2| 30|
|  red|carrot|   2| 50|
|  red|banana|   2| 70|
|  red| grape|   2| 80|
+-----+------+----+---+
def plus_mean(pandas_df):
    return pandas_df.assign(v1= pandas_df.v1.count())

df.groupby('color').applyInPandas(plus_mean, schema=df.schema).show()
```

```
+-----+------+---+---+
|color| fruit| v1| v2|
+-----+------+---+---+
|black|carrot|  1| 60|
| blue|banana|  2| 20|
| blue| grape|  2| 10|
|  red|banana|  5| 10|
|  red|carrot|  5| 30|
|  red|carrot|  5| 50|
|  red|banana|  5| 70|
|  red| grape|  5| 80|
+-----+------+---+---+
```

```
def plus_mean(pandas_df):
    return pandas_df.assign(v1= pandas_df.v1.var())

df.groupby('color').applyInPandas(plus_mean, schema=df.schema).show()
```

```
+-----+------+----+---+
|color| fruit|  v1| v2|
+-----+------+----+---+
|black|carrot|null| 60|
| blue|banana|   2| 20|
| blue| grape|   2| 10|
|  red|banana|   8| 10|
|  red|carrot|   8| 30|
|  red|carrot|   8| 50|
|  red|banana|   8| 70|
|  red| grape|   8| 80|
+-----+------+----+---+
df1 = spark.createDataFrame(
    [(20000101, 1, 1.0), (20000101, 2, 2.0), (20000102, 1, 3.0), (20000102,
2, 4.0)],
    ('time', 'id', 'v1'))

df2 = spark.createDataFrame(
    [(20000101, 1, 'x'), (20000101, 2, 'y')],
    ('time', 'id', 'v2'))

def asof_join(l, r):#l,r is dataframe instances
    return pd.merge_asof(l, r, on='time', by='id')

df1.groupby('id').cogroup(df2.groupby('id')).applyInPandas(
    asof_join, schema='time int, id int, v1 double, v2 string').show()
```

```
+--------+---+---+---+
|    time| id| v1| v2|
+--------+---+---+---+
|20000101|  1|1.0|  x|
|20000102|  1|3.0|  x|
|20000101|  2|2.0|  y|
|20000102|  2|4.0|  y|
+--------+---+---+---+
```

```
df1 = spark.createDataFrame(
    [(20000101, 1, 1.0), (20000101, 2, 2.0), (20000102, 1, 3.0), (20000102,
2, 4.0)],
    ('time', 'id', 'v1'))

df2 = spark.createDataFrame(
    [(20000101, 1, 'x'), (20000101, 2, 'y')],
    ('time', 'id', 'v2'))

def asof_join(l, r):
    return pd.merge_asof(l, r, on='time', by='id')

df2.groupby('id').cogroup(df1.groupby('id')).applyInPandas(
    asof_join, schema='time int, id int, v1 double, v2 string').show()
```

```
+--------+---+---+---+
|    time| id| v1| v2|
+--------+---+---+---+
|20000101|  1|1.0|  x|
|20000101|  2|2.0|  y|
+--------+---+---+---+
# import pyspark class Row from module sql
from pyspark.sql import *

# Create Example Data - Departments and Employees

# Create the Departments
department1 = Row(id='123456', name='Computer Science')
department2 = Row(id='789012', name='Mechanical Engineering')
department3 = Row(id='345678', name='Theater and Drama')
department4 = Row(id='901234', name='Indoor Recreation')

# Create the Employees
Employee = Row("firstName", "lastName", "email", "salary")
employee1 = Employee('michael', 'armbrust', 'no-reply@berkeley.edu', 100000)
employee2 = Employee('xiangrui', 'meng', 'no-reply@stanford.edu', 120000)
employee3 = Employee('matei', None, 'no-reply@waterloo.edu', 140000)
employee4 = Employee(None, 'wendell', 'no-reply@berkeley.edu', 160000)
employee5 = Employee('michael', 'jackson', 'no-reply@neverla.nd', 80000)

# Create the DepartmentWithEmployees instances from Departments and
Employees
departmentWithEmployees1 = Row(department=department1, employees=[employee1,
employee2])
departmentWithEmployees2 = Row(department=department2, employees=[employee3,
employee4])
departmentWithEmployees3 = Row(department=department3, employees=[employee5,
employee4])
departmentWithEmployees4 = Row(department=department4, employees=[employee2,
employee3])

print(department4)
print(employee3)
print(departmentWithEmployees1.employees[0].email)
```

```
Row(id='901234', name='Indoor Recreation')
Row(firstName='matei', lastName=None, email='no-reply@waterloo.edu', salary=
140000)
no-reply@berkeley.edu
```

```
print(departmentWithEmployees3.employees[0].salary)
```

```
80000
```

```
print(departmentWithEmployees1.employees[0].salary,departmentWithEmployees2.
employees[0].salary)
```

```
100000 140000
```

```
print(departmentWithEmployees1.employees[0].email,departmentWithEmployees1.e
mployees[0].email)
print(departmentWithEmployees1.employees[0].firstName,departmentWithEmployee
s2.employees[0].firstName)
```

```
no-reply@berkeley.edu no-reply@berkeley.edu
michael matei
```

```
print(departmentWithEmployees1.employees[0].email)
```

```
no-reply@berkeley.edu
```

```
for i in range(0,len(departmentWithEmployees1)):

print(departmentWithEmployees1.employees[i].firstName.departmentWithEmployee
s1)
```

```
  AttributeError: 'str' object has no attribute 'departmentWithEmployees1'
```

```
departmentsWithEmployeesSeq1 = [departmentWithEmployees1,
departmentWithEmployees2]
df1 = spark.createDataFrame(departmentsWithEmployeesSeq1)

df1.show(truncate=False)

departmentsWithEmployeesSeq2 = [departmentWithEmployees3,
departmentWithEmployees4]
df2 = spark.createDataFrame(departmentsWithEmployeesSeq2)

df2.show(truncate=False)
```

```
+--------------------------------+------------------------------------------
------------------------------------------------------------+
|department                      |employees
|
+--------------------------------+------------------------------------------
------------------------------------------------------------+
|{123456, Computer Science}      |[{michael, armbrust, no-reply@berkeley.ed
u, 100000}, {xiangrui, meng, no-reply@stanford.edu, 120000}]|
|{789012, Mechanical Engineering}|[{matei, null, no-reply@waterloo.edu, 1400
00}, {null, wendell, no-reply@berkeley.edu, 160000}]        |
+--------------------------------+------------------------------------------
------------------------------------------------------------+
```

```
+---------------------------+-----------------------------------------------
-------------------------------------------------+
|department                 |employees
|
+---------------------------+-----------------------------------------------
-------------------------------------------------+
|{345678, Theater and Drama}|[{michael, jackson, no-reply@neverla.nd, 8000
0}, {null, wendell, no-reply@berkeley.edu, 160000}]|
```

```
unionDF = df1.union(df2)
unionDF.show(truncate=False)
```

```
+---------------------------------+------------------------------------------
-----------------------------------------------------------+
|department                       |employees
|
+---------------------------------+------------------------------------------
-----------------------------------------------------------+
|{123456, Computer Science}       |[{michael, armbrust, no-reply@berkeley.ed
u, 100000}, {xiangrui, meng, no-reply@stanford.edu, 120000}]|
|{789012, Mechanical Engineering}|[{matei, null, no-reply@waterloo.edu, 1400
00}, {null, wendell, no-reply@berkeley.edu, 160000}]       |
|{345678, Theater and Drama}      |[{michael, jackson, no-reply@neverla.nd, 8
0000}, {null, wendell, no-reply@berkeley.edu, 160000}]     |
|{901234, Indoor Recreation}      |[{xiangrui, meng, no-reply@stanford.edu, 1
20000}, {matei, null, no-reply@waterloo.edu, 140000}]      |
+---------------------------------+------------------------------------------
-----------------------------------------------------------+
```

```
# Remove the file if it exists
dbutils.fs.rm("/tmp/databricks-df-example.parquet", True)
df.write.format("parquet").save("/tmp/databricks-df-example.parquet")
```

```
parquetDF = spark.read.format("parquet").load("/tmp/databricks-df-
example.parquet")
parquetDF.show(truncate=False)
```

```
+-----+------+---+---+
|color|fruit |v1 |v2 |
+-----+------+---+---+
|black|carrot|6  |60 |
|blue |banana|2  |20 |
|red  |carrot|5  |50 |
|red  |banana|7  |70 |
|red  |banana|1  |10 |
|red  |carrot|3  |30 |
|blue |grape |4  |10 |
|red  |grape |8  |80 |
+-----+------+---+---+
```

```
dbutils.fs.rm("/tmp/databricks-df-example.parquet", True)
unionDF.write.format("parquet").save("/tmp/databricks-df-example.parquet")
```

```
parquetDF = spark.read.format("parquet").load("/tmp/databricks-df-
example.parquet")
parquetDF.show(truncate=False)
```

```
+--------------------------------+------------------------------------------
-----------------------------------------------------------+
|department                      |employees
|
+--------------------------------+------------------------------------------
-----------------------------------------------------------+
|{789012, Mechanical Engineering}|[{matei, null, no-reply@waterloo.edu, 1400
00}, {null, wendell, no-reply@berkeley.edu, 160000}]       |
|{345678, Theater and Drama}     |[{michael, jackson, no-reply@neverla.nd, 8
0000}, {null, wendell, no-reply@berkeley.edu, 160000}]     |
|{123456, Computer Science}      |[{michael, armbrust, no-reply@berkeley.ed
u, 100000}, {xiangrui, meng, no-reply@stanford.edu, 120000}]|
|{901234, Indoor Recreation}     |[{xiangrui, meng, no-reply@stanford.edu, 1
20000}, {matei, null, no-reply@waterloo.edu, 140000}]      |
+--------------------------------+------------------------------------------
-----------------------------------------------------------+
```

```
from pyspark.sql.functions import explode

explodeDF = unionDF.select(explode("employees").alias("e"))
flattenDF = explodeDF.selectExpr("e.firstName", "e.lastName", "e.email",
"e.salary")

flattenDF.show(truncate=False)
```

```
+---------+--------+---------------------+------+
|firstName|lastName|email                |salary|
+---------+--------+---------------------+------+
|michael  |armbrust|no-reply@berkeley.edu|100000|
|xiangrui |meng    |no-reply@stanford.edu|120000|
|matei    |null    |no-reply@waterloo.edu|140000|
|null     |wendell |no-reply@berkeley.edu|160000|
|michael  |jackson |no-reply@neverla.nd  |80000 |
|null     |wendell |no-reply@berkeley.edu|160000|
|xiangrui |meng    |no-reply@stanford.edu|120000|
|matei    |null    |no-reply@waterloo.edu|140000|
+---------+--------+---------------------+------+
```

```
filterDF = flattenDF.filter(flattenDF.firstName ==
"michael").sort(flattenDF.salary)
filterDF.show(truncate=False)
```

```
+---------+--------+--------------------+------+
|firstName|lastName|email               |salary|
+---------+--------+--------------------+------+
|michael  |jackson |no-reply@neverla.nd |80000 |
|michael  |armbrust|no-reply@berkeley.edu|100000|
+---------+--------+--------------------+------+
```

```
+---------+--------+--------------------+------+
|firstName|lastName|email               |salary|
+---------+--------+--------------------+------+
|michael  |armbrust|no-reply@berkeley.edu|100000|
|michael  |jackson |no-reply@neverla.nd |80000 |
|xiangrui |meng    |no-reply@stanford.edu|120000|
|xiangrui |meng    |no-reply@stanford.edu|120000|
+---------+--------+--------------------+------+
```