3-9-2017 Data Aggr...

FINISHED

```
%pyspark
from pandas import Series, DataFrame
import numpy as np, pandas as pd
df = DataFrame({'key1': ['a','a','b','b','a'],
                'key2': ['one','two','one','two','one',],
                'data1' : np.random.randn (5),
                'data2' : np.random.randn (5)})
df
```

```
      data1     data2 key1 key2
0 -0.826471  0.996799    a  one
1  0.913979  0.859096    a  two
2 -0.035160 -0.920278    b  one
3  1.230343 -0.520285    b  two
4 -1.552572  0.345240    a  one
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:04:04 PM.

FINISHED

```
%pyspark
grouped = df['data1'].groupby(df['key1'])
grouped
```

```
<pandas.core.groupby.SeriesGroupBy object at 0x10e589b90>
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:07:21 PM.

FINISHED

```
%pyspark
grouped.mean()
```

```
key1
a   -0.488355
b    0.597591
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:10:22 PM.

FINISHED

```
%pyspark
means = df['data1'].groupby([df['key1'], df['key2']]).mean()
means
```

```
key1  key2
a     one    -1.189521
      two     0.913979
b     one    -0.035160
      two     1.230343
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:15:41 PM.

```
%pyspark
means.unstack()
```

FINISHED

```
key2       one       two
key1
a    -1.189521  0.913979
b    -0.035160  1.230343
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:16:09 PM.

```
%pyspark
states = np.array(['Ohio','California','California','Ohio','Ohio'])
years = np.array(['2005','2005','2006','2005','2006'])
df['data1'].groupby([states, years]).mean()
```

FINISHED

```
California  2005    0.913979
            2006   -0.035160
Ohio        2005    0.201936
            2006   -1.552572
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:21:46 PM.

```
%pyspark
df.groupby('key1').mean()
```

FINISHED

```
         data1     data2
key1
a    -0.488355  0.733711
b     0.597591 -0.720281
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:23:02 PM.

```
%pyspark
df.groupby(['key1','key2']).mean()
```

FINISHED

```
              data1     data2
key1 key2
a    one  -1.189521  0.671019
     two   0.913979  0.859096
b    one  -0.035160 -0.920278
     two   1.230343 -0.520285
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:24:19 PM.

```
%pyspark
df.groupby(['key1','key2']).size()
```

FINISHED

```
key1  key2
a     one     2
      two     1
b     one     1
      two     1
dtype: int64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:24:40 PM.

```
%pyspark
for name, group in df.groupby('key1'):
    print name
    print group
```

FINISHED

```
a
      data1     data2 key1 key2
0 -0.826471  0.996799    a  one
1  0.913979  0.859096    a  two
4 -1.552572  0.345240    a  one
b
      data1     data2 key1 key2
2 -0.035160 -0.920278    b  one
3  1.230343 -0.520285    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:27:24 PM.

```
%pyspark
for (k1,k2), group in df.groupby(['key1','key1']):
    print k1, k2
    print group
```

FINISHED

```
a a
      data1     data2 key1 key2
0 -0.826471  0.996799    a  one
1  0.913979  0.859096    a  two
4 -1.552572  0.345240    a  one
b b
      data1     data2 key1 key2
2 -0.035160 -0.920278    b  one
3  1.230343 -0.520285    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:29:38 PM.

```
%pyspark
pieces = dict(list(df.groupby('key1')))
pieces['b']
```

FINISHED

```
      data1     data2 key1 key2
2 -0.035160 -0.920278    b  one
3  1.230343 -0.520285    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:32:05 PM.

```
%pyspark
df.dtypes
```

FINISHED

```
data1    float64
data2    float64
key1      object
key2      object
dtype: object
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:33:11 PM.

```
%pyspark
grouped = df.groupby(df.dtypes, axis =1)
dict(list(grouped))
```

FINISHED

```
{dtype('O'):   key1 key2
0    a  one
1    a  two
2    b  one
3    b  two
4    a  one, dtype('float64'):       data1     data2
0 -0.826471  0.996799
1  0.913979  0.859096
2 -0.035160 -0.920278
3  1.230343 -0.520285
4 -1.552572  0.345240}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:35:50 PM.

```
%pyspark
```

READY