python

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 -0.495570 -1.237109    a  one
1 -0.596249  1.226747    a  two
2 -0.146109  0.114855    b  one
3  0.278221  0.341515    b  two
4  0.035725 -1.083888    a  one
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:16:56 PM.

FINISHED

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

grouped
```

```
<pandas.core.groupby.SeriesGroupBy object at 0x7fc1ad4d2390>
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:33:04 PM.

FINISHED

```
%pyspark

grouped.mean()
```

```
key1
a   -0.352031
b    0.066056
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:33:07 PM.

FINISHED

```
%pyspark
```

```
key1  key2
a     one    -0.229922
      two    -0.596249
b     one    -0.146109
      two     0.278221
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:19:52 PM.

FINISHED

```
%pyspark

means.unstack()
```

```
key2       one       two
key1
a    -0.229922 -0.596249
b    -0.146109  0.278221
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:20:15 PM.

FINISHED

```
%pyspark

states = np.array(['Ohio','California','California','Ohio','Ohio'])

years = np.array([2005,2005,2006,2005,2006])

df['data1'].groupby([states,years]).mean()
```

```
California  2005   -0.596249
            2006   -0.146109
Ohio        2005   -0.108675
            2006    0.035725
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:34:00 PM.

FINISHED

```
%pyspark

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

```
         data1     data2
key1
a    -0.352031 -0.364750
```

```
b     0.066056  0.228185
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:34:09 PM.

FINISHED

```
%pyspark

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

```
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, 7:25:44 PM.

FINISHED

```
%pyspark

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

```
a
      data1     data2 key1 key2
0 -0.495570 -1.237109    a  one
1 -0.596249  1.226747    a  two
4  0.035725 -1.083888    a  one
b
      data1     data2 key1 key2
2 -0.146109  0.114855    b  one
3  0.278221  0.341515    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:26:50 PM.

FINISHED

```
%pyspark

for(k1,k2),group in df.groupby(['key1','key2']):
    print k1,k2
    print group
```

```
a one
      data1     data2 key1 key2
0 -0.495570 -1.237109    a  one
4  0.035725 -1.083888    a  one
a two
      data1     data2 key1 key2
1 -0.596249  1.226747    a  two
b one
      data1     data2 key1 key2
```

```
2 -0.146109  0.114855    b  one
b two
      data1     data2 key1 key2
3  0.278221  0.341515    b  two
```

Took 1 sec. Last updated by anonymous at March 09 2017, 7:28:19 PM.

FINISHED

```
%pyspark

pieces = dict(list(df.groupby('key1')))

pieces['b']
```

```
      data1     data2 key1 key2
2 -0.146109  0.114855    b  one
3  0.278221  0.341515    b  two
```

Took 1 sec. Last updated by anonymous at March 09 2017, 7:29:36 PM.

FINISHED

```
%pyspark

df.dtypes
```

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

Took 0 sec. Last updated by anonymous at March 09 2017, 7:29:47 PM.

FINISHED

```
%pyspark

grouped = df.groupby(df.dtypes,axis=1)
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:30:35 PM.

FINISHED

```
%pyspark
dict(list(grouped))
```

```
{dtype('O'):   key1 key2
0    a  one
1    a  two
2    b  one
3    b  two
4    a  one, dtype('float64'):       data1     data2
0 -0.495570 -1.237109
1 -0.596249  1.226747
2 -0.146109  0.114855
3  0.278221  0.341515
```

```
4  0.035725 -1.083888}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:30:48 PM.

```
%pyspark
```

READY