Zeppelin

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  0.476316  0.753479    a  one
1 -0.197014 -0.786007    a  two
2  0.254329  1.256762    b  one
3  0.918940 -1.499794    b  two
4 -1.060057  0.512586    a  one
```

FINISHED

```
%pyspark

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

grouped
```

```
<pandas.core.groupby.SeriesGroupBy object at 0x104e19750>
```

FINISHED

```
%pyspark

grouped.mean()
```

```
key1
a   -0.260252
b    0.586635
Name: data1, dtype: float64
```

FINISHED

```
%pyspark

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

FINISHED

```
%pyspark
```

```
means
```

```
key1  key2
a     one    -0.291870
      two    -0.197014
b     one     0.254329
      two     0.918940
Name: data1, dtype: float64
```

FINISHED

```
%pyspark

means.unstack()
```

```
key2       one       two
key1
a    -0.291870 -0.197014
b     0.254329  0.918940
```

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   -0.197014
            2006    0.254329
Ohio        2005    0.697628
            2006   -1.060057
Name: data1, dtype: float64
```

FINISHED

```
%pyspark

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

```
         data1     data2
key1
a    -0.260252  0.160019
b     0.586635 -0.121516
```

FINISHED

```
%pyspark
```

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

```
              data1     data2
key1 key2
a    one  -0.291870  0.633032
     two  -0.197014 -0.786007
b    one   0.254329  1.256762
     two   0.918940 -1.499794
```

FINISHED

```
%pyspark

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

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

FINISHED

```
%pyspark

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

```
a
      data1     data2 key1 key2
0  0.476316  0.753479    a  one
1 -0.197014 -0.786007    a  two
4 -1.060057  0.512586    a  one
b
      data1     data2 key1 key2
2  0.254329  1.256762    b  one
3  0.918940 -1.499794    b  two
```

FINISHED

```
%pyspark

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

```
a one
      data1     data2 key1 key2
0  0.476316  0.753479    a  one
4 -1.060057  0.512586    a  one
a two
      data1     data2 key1 key2
1 -0.197014 -0.786007    a  two
b one
      data1     data2 key1 key2
2  0.254329  1.256762    b  one
b two
     data1     data2 key1 key2
3  0.91894 -1.499794    b  two
```

FINISHED

```
%pyspark

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

pieces['b']
```

```
      data1     data2 key1 key2
2  0.254329  1.256762    b  one
3  0.918940 -1.499794    b  two
```

FINISHED

```
%pyspark

df.dtypes
```

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

FINISHED

```
%pyspark

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

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  0.476316  0.753479
1 -0.197014 -0.786007
2  0.254329  1.256762
3  0.918940 -1.499794
4 -1.060057  0.512586}
```

READY