# Data Aggregation

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  0.830010  0.098991    a  one
1  0.655552 -0.140033    a  two
2  1.805540 -1.066978    b  one
3 -0.733145  0.282225    b  two
4  0.389095  0.305841    a  one
```

FINISHED

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

```
key1
a    0.624885
b    0.536198
Name: data1, dtype: float64
```

FINISHED

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

```
key1  key2
a     one     0.609552
      two     0.655552
b     one     1.805540
      two    -0.733145
Name: data1, dtype: float64
```

FINISHED

```
%pyspark
```

```
means.unstack()
```

```
key2       one       two
key1
a     0.609552  0.655552
b     1.805540 -0.733145
```

```
%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    0.655552
            2006    1.805540
Ohio        2005    0.048432
            2006    0.389095
Name: data1, dtype: float64
```

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

FINISHED

```
         data1     data2
key1
a     0.624885  0.088267
b     0.536198 -0.392376
```

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

FINISHED

```
              data1     data2
key1 key2
a    one   0.609552  0.202416
     two   0.655552 -0.140033
b    one   1.805540 -1.066978
     two  -0.733145  0.282225
```

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

FINISHED

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

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

FINISHED

```
a
      data1     data2 key1 key2
0  0.830010  0.098991    a  one
1  0.655552 -0.140033    a  two
4  0.389095  0.305841    a  one
b
      data1     data2 key1 key2
2  1.805540 -1.066978    b  one
3 -0.733145  0.282225    b  two
```

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

FINISHED

```
a one
      data1     data2 key1 key2
0  0.830010  0.098991    a  one
4  0.389095  0.305841    a  one
a two
      data1     data2 key1 key2
1  0.655552 -0.140033    a  two
b one
     data1     data2 key1 key2
2  1.80554 -1.066978    b  one
b two
      data1     data2 key1 key2
3 -0.733145  0.282225    b  two
```

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

FINISHED

```
pieces['b']
```

```
      data1     data2 key1 key2
2  1.805540 -1.066978    b  one
3 -0.733145  0.282225    b  two
```

```
%pyspark
df.dtypes
```

FINISHED

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

```
%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  0.830010  0.098991
1  0.655552 -0.140033
2  1.805540 -1.066978
3 -0.733145  0.282225
4  0.389095  0.305841}
```

```
%pyspark
```

READY