```
val polutionData = sqlContext.read.format("com.databricks.spark.csv").option("header", "true").load(
```

FINISHED

```
polutionData: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [ozone: string, particullate_m
atter: string ... 6 more fields]
```

Took 45 sec. Last updated by anonymous at February 15 2017, 1:53:48 AM.

```
polutionData.toDF().registerTempTable("polData")
```

FINISHED

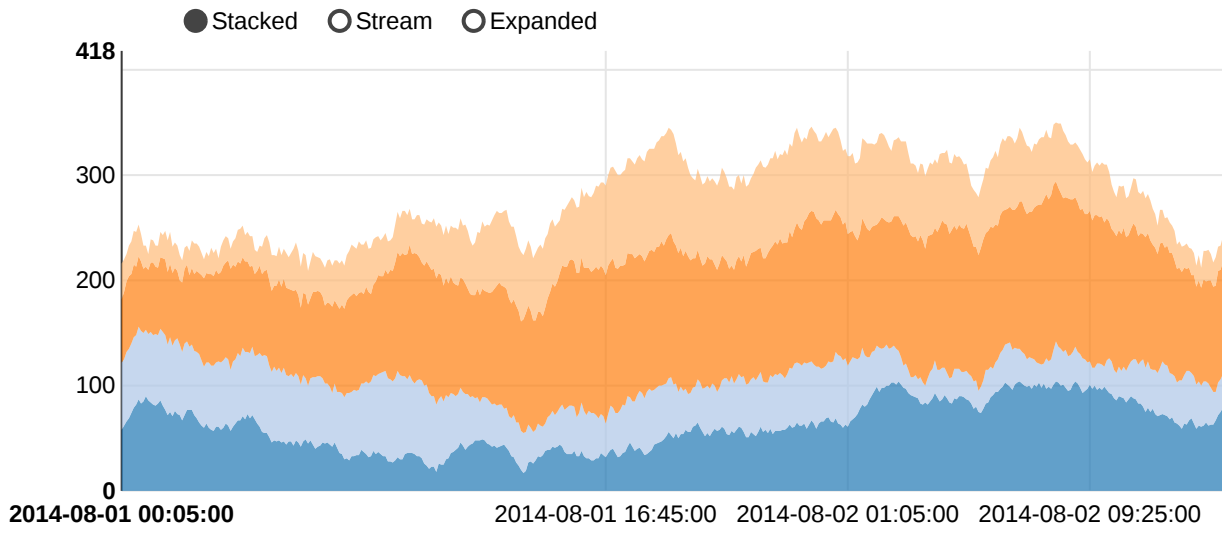
```
warning: there was one deprecation warning; re-run with -deprecation for details
```

Took 1 sec. Last updated by anonymous at February 15 2017, 1:53:53 AM.

```
%sql
select * from polData
```

FINISHED

settings

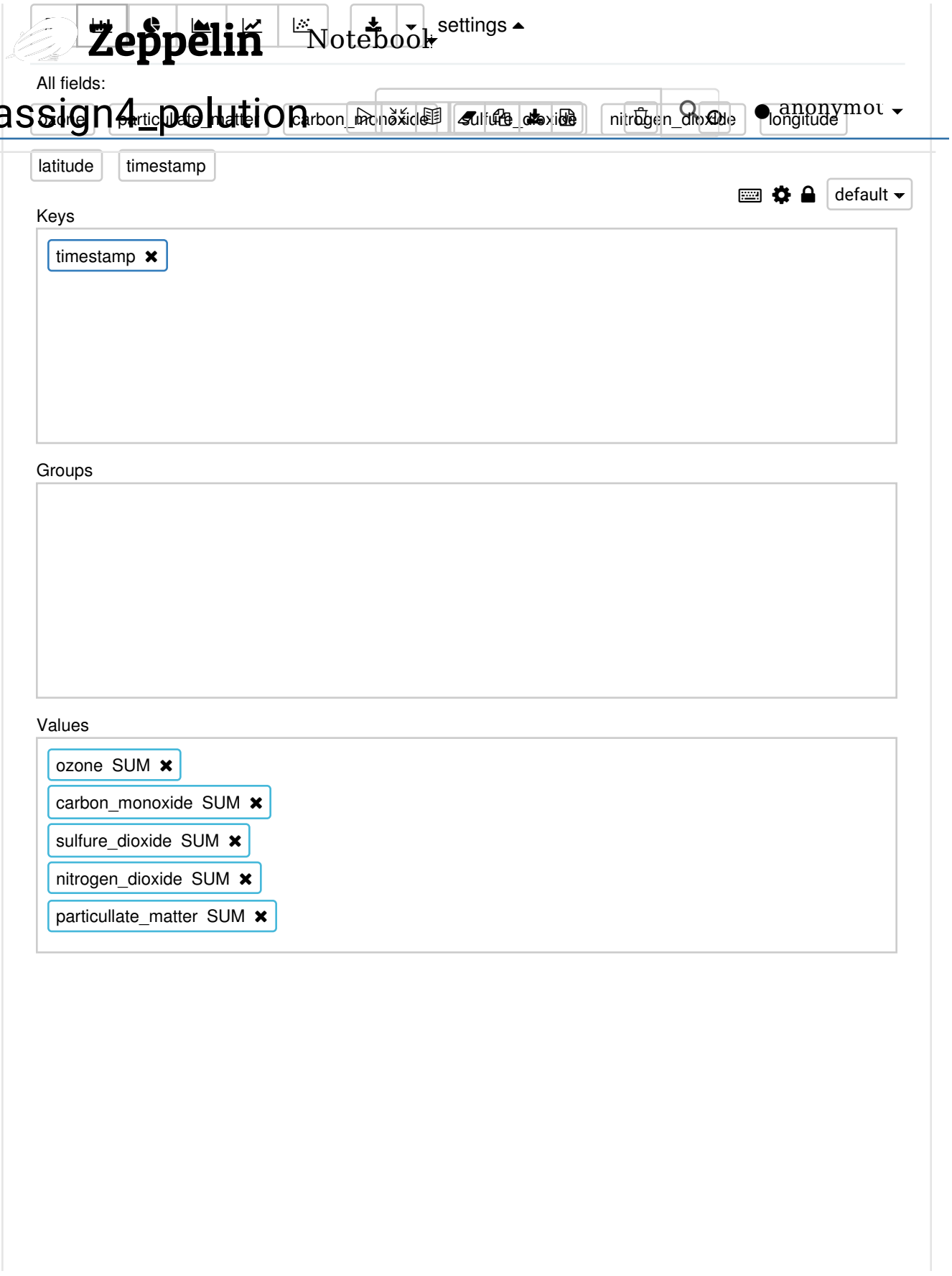



```
Results are limited by 1000.
```

Took 11 sec. Last updated by anonymous at February 15 2017, 1:54:08 AM. (outdated)

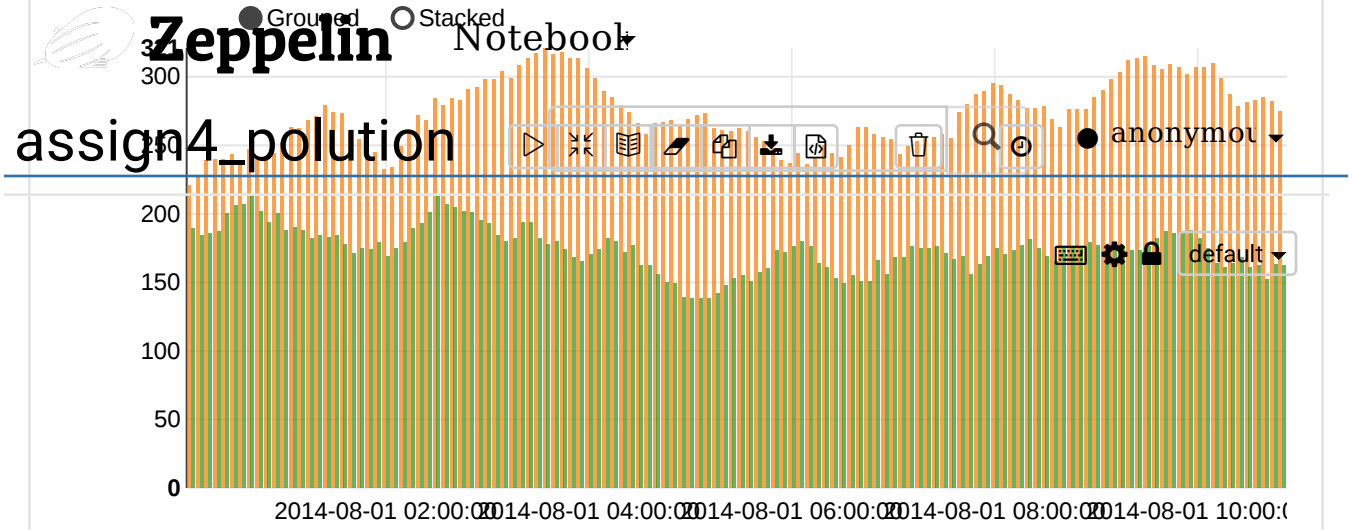
```
%sql
select * from polData where timestamp like "2014-08-01%"
```

FINISHED


Zeppelin
Notebook
settings
assign4_polution
All fields:
ozone
particullate_matter
carbon_monoxide
sulfure_dioxide
nitrogen_dioxide
longitude
anonymou
latitude
timestamp
default
Keys
timestamp
Groups
Values
ozone SUM
carbon_monoxide SUM
sulfure_dioxide SUM
nitrogen_dioxide SUM
particullate_matter SUM

Zeppelin Notebook

assign4_polution




```
Results are limited by 1000.
```

Took 1 sec. Last updated by anonymous at February 15 2017, 1:55:06 AM. (outdated)

FINISHED

```
%pyspark
import glob
files = glob.glob("/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/*.csv")
print (files[:20])
```

```
['/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData192866.csv', '/home/scarface/D
esktop/sem-3/capstone/Prithvi/pollution/pollutionData206475.csv', '/home/scarface/Desktop/sem-3/capst
one/Prithvi/pollution/pollutionData193402.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollut
ion/pollutionData195312.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData2
01828.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData197896.csv', '/home
/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData181331.csv', '/home/scarface/Desktop/
sem-3/capstone/Prithvi/pollution/pollutionData201722.csv', '/home/scarface/Desktop/sem-3/capstone/Pri
thvi/pollution/pollutionData194960.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pol
lutionData197355.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData190799.c
sv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData179228.csv', '/home/scarfa
ce/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData187006.csv', '/home/scarface/Desktop/sem-3/c
apstone/Prithvi/pollution/pollutionData187297.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/po
llution/pollutionData185396.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionD
ata192946.csv', '/home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData158983.csv', '/
home/scarface/Desktop/sem-3/capstone/Prithvi/pollution/pollutionData197734.csv', '/home/scarface/Desk
top/sem-3/capstone/Prithvi/pollution/pollutionData188039.csv', '/home/scarface/Desktop/sem-3/capstone
/Prithvi/pollution/pollutionData195552.csv']
```

Took 0 sec. Last updated by anonymous at March 27 2017, 10:41:44 PM.

FINISHED

```
%pyspark
import csv
dic={}
for each in files:
    oz=0
```

Zeppelin Notebook

assign4_polution

anonymou

default

```
17568
1674447
17568
1856742
```

Took 26 sec. Last updated by anonymous at March 27 2017, 8:21:32 PM. (outdated)

FINISHED

```
%pyspark
oxM=0
pmM=0
cmM=0
sdM=0
ndM=0
o,p,c,s,n="","","","",""
for j in dic:
    if dic[j][0] > oxM:
        o=j
        oxM=dic[j][0]
    if dic[j][1] > pmM:
        p=j
        pmM=dic[j][1]
    if dic[j][2] > cmM:
        c=j
        cmM=dic[j][2]
    if dic[j][3] > sdM:
        s=j
        sdM=dic[j][3]
    if dic[j][4] > ndM:
        n=j
        ndM=dic[j][4]
li= [o,p,c,s,n]
```

Zeppelin Notebook

assign4_polution

anonymous

```
['10.250786139881143+56.20251117218925', '10.149547918402732+56.170886798880105', '10.171611868717264
+56.174917180024806', '10.10711200000003+56.21731711429131', '10.269519082174156+56.2107989833621']
```

Took 0 sec. Last updated by anonymous at March 27 2017, 9:20:15 PM. (outdated)

default

FINISHED

```
%pyspark
headLi=["ozone", "particullate_matter", "carbon_monoxide", "sulfure_dioxide", "nitrogen_dioxide"]
for l in li:
    latLon=l.split("+")
    print "Highest percentage of ozone levels are detected at location " + latLon[0] + ", "+ latLon[
```

```
Highest percentage of ozone levels are detected at location 10.250786139881143, 56.20251117218925
Highest percentage of ozone levels are detected at location 10.149547918402732, 56.170886798880105
Highest percentage of ozone levels are detected at location 10.171611868717264, 56.174917180024806
Highest percentage of ozone levels are detected at location 10.10711200000003, 56.21731711429131
Highest percentage of ozone levels are detected at location 10.269519082174156, 56.2107989833621
```

Took 0 sec. Last updated by anonymous at March 27 2017, 9:24:49 PM.

FINISHED

```
%pyspark
oxMi=100
pmMi=100
cmMi=100
sdMi=100
ndMi=100
o,p,c,s,n="","","","",""
for j in dic:
    if dic[j][0] < oxMi:
        o=j
        oxMi=dic[j][0]
    if dic[j][1] < pmMi:
        p=j
        pmMi=dic[j][1]
    if dic[j][2] < cmMi:
        c=j
        cmMi=dic[j][2]
    if dic[j][3] < sdMi:
        s=j
        sdMi=dic[j][3]
    if dic[j][4] < ndMi:
        n=j
        ndMi=dic[j][4]
lim= [o,p,c,s,n]
print lim
```

```
['10.173023480985648+56.21071820426365', '10.141663381614649+56.124304277059785', '10.175338207340246
+56.179322409643085', '10.149874309192683+56.148242094184255', '10.212459373016372+56.18696229571611'
]
```

Took 1 sec. Last updated by anonymous at March 27 2017, 9:42:40 PM.

FINISHED

```
%pyspark
```

Zeppelin Notebook

assign4_polution

anonymou

default

```
Lowest percentage of ozone levels are detected at location 10.173023480985648, 56.21071820426365
Lowest percentage of ozone levels are detected at location 10.141663381614649, 56.124304277059785
Lowest percentage of ozone levels are detected at location 10.175338207340246, 56.179322409643085
Lowest percentage of ozone levels are detected at location 10.149874309192683, 56.148242094184255
Lowest percentage of ozone levels are detected at location 10.212459373016372, 56.18696229571611
```

Took 0 sec. Last updated by anonymous at March 27 2017, 9:45:06 PM. (outdated)

FINISHED

```
%pyspark
print headLi
print li
print lim
```

```
['ozone', 'particullate_matter', 'carbon_monoxide', 'sulfure_dioxide', 'nitrogen_dioxide']
['10.250786139881143+56.20251117218925', '10.149547918402732+56.170886798880105', '10.171611868717264
+56.174917180024806', '10.10711200000003+56.21731711429131', '10.269519082174156+56.2107989833621']
['10.173023480985648+56.21071820426365', '10.141663381614649+56.124304277059785', '10.175338207340246
+56.179322409643085', '10.149874309192683+56.148242094184255', '10.212459373016372+56.18696229571611'
]
```

Took 0 sec. Last updated by anonymous at March 27 2017, 9:49:08 PM.

FINISHED

```
%pyspark
from sklearn import datasets, linear_model
```

Took 5 sec. Last updated by anonymous at March 27 2017, 10:05:04 PM.

FINISHED

```
%pyspark
ozLi=[]
pmLi=[]
cmLi=[]
sdLi=[]
ndLi=[]
with open(files[0], "rb") as f:
    reader = csv.reader(f)
    a=reader.next()
    for row in reader:
        ozLi.append(int(row[0]))
        pmLi.append(int(row[1]))
        cmLi.append(int(row[2]))
        sdLi.append(int(row[3]))
        ndLi.append(int(row[4]))
print ozLi[:100]
```

```
[30, 29, 34, 35, 34, 39, 43, 41, 41, 39, 35, 37, 35, 38, 37, 38, 42, 37, 38, 38, 37, 33, 36, 36, 37,
35, 40, 38, 36, 31, 31, 31, 31, 26, 27, 26, 26, 31, 35, 31, 34, 39, 40, 40, 41, 36, 37, 37, 37, 33, 3
1, 36, 32, 33, 33, 38, 39, 37, 37, 32, 34, 39, 42, 46, 44, 39, 35, 31, 26, 25, 21, 25, 23, 19, 28, 32
```

```
, 32, 31, 31, 26, 27, 26, 26, 25, 28, 30, 35, 37, 41, 40, 43, 47, 44, 42, 37, 35, 39, 42, 47, 49]
```

Took 0 sec. Last updated by anonymous at March 27 2017, 10:43:10 PM.

Zeppelin Notebook

assign4_polution

anonymous

```
%pyspark
sd_train=ozLi[:-20]
sd_test=ozLi[-20:]

pm_train=ozLi[:-20]
pm_test=ozLi[-20:]

pm_train=pd.DataFrame(pm_train)
sd_train=pd.DataFrame(sd_train)
```

FINISHED

default

Took 0 sec. Last updated by anonymous at March 27 2017, 10:35:11 PM.

```
%pyspark
reg = linear_model.LinearRegression()
reg.fit(sd_train, pm_train)
```

FINISHED

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Took 0 sec. Last updated by anonymous at March 27 2017, 10:35:15 PM.

```
%pyspark
import numpy as np
print('Coefficients: ', reg.coef_)
print("Mean squared error: 35.07" % np.mean((reg.predict(sd_train) - pm_train) ** 2))
```

FINISHED

```
('Coefficients: ', array([[ 1.]]))
Mean squared error: 35.07
```

Took 0 sec. Last updated by anonymous at March 27 2017, 10:43:38 PM.

READY