

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19045.3930]
(c) Microsoft Corporation. All rights reserved.

C:\Users\tiana\Desktop\big-data-class\bigdata\cstu_bigdata_handson>docker-compose up -d
request returned Internal Server Error for API route and version http://%2F%2F.%2Fpipe%2Fdocker_engine/v1.24/containers/
json?all=1&filters=%7B%22label%22%3A%7B%22com.docker.compose.config-hash%22%3Atrue%2C%22com.docker.compose.project%3Dcstu_bigdata_handson%22%3Atrue%7D%7D, check if the server supports the requested API version

C:\Users\tiana\Desktop\big-data-class\bigdata\cstu_bigdata_handson>
C:\Users\tiana\Desktop\big-data-class\bigdata\cstu_bigdata_handson>docker ls
docker: 'ls' is not a docker command.
See 'docker --help'

C:\Users\tiana\Desktop\big-data-class\bigdata\cstu_bigdata_handson>docker-compose up -d
[+] Building 0.0s (0/0)                                                                                               docker:default
[+] Running 2/2
   Container spark-master                                                                                          Started 0.0s
   Container cstu_bigdata_handson-spark-worker-1                                                                 Started 0.0s

C:\Users\tiana\Desktop\big-data-class\bigdata\cstu_bigdata_handson>
```

## Create a Spark Core RDD from Different Ways:

```
localhost:8888/notebooks/week2_1.1-rdd.ipynb 110% ☆ ⌵ ⌵ ⌵
Jupyter week2_1.1-rdd Last Checkpoint: yesterday Trusted
File Edit View Run Kernel Settings Help
+ 🔍 📄 ▶ ⏸ ⏹ ⏶ ⏷ ⏸ ⏹ ⏶ ⏷ ⏸ ⏹ Code ▼ JupyterLab Python 3 (ipykernel)

rdd.collect()
Out[4]: ['a', 'b', 'c']

[6]: # read a csv file

rdd = sc.textFile('data/SparkData/mtcars.csv')

rdd.take(5)

Out[5]: ['mpg,cyl,disp,hp,drat,wt,qsec,vs,am,gear,carb',
'Mazda RX4,21.6,160,110,3.9,2.62,16.46,0,1,4,4',
'Mazda RX4 Wag,21.6,160,110,3.9,2.875,17.02,0,1,4,4',
'Datsun 710,22.8,4,108,93,3.85,2.32,18.61,1,1,4,1',
'Hornet 4 Drive,21.4,6,258,110,3.08,3.215,19.44,1,0,3,1']

[7]: # read a txt file

rdd = sc.textFile('data/SparkData/mtcars.csv')

rdd.take(5)

Out[6]: ['Fresh install of XP on new computer. Sweet relief! fuck vista\t1018769417\t1.0',
'Well. Now I know where to go when I want my knives. #ChiChevySXSW http://post.ly/Rv0l\t10284216536\t1.0',
'Literally six weeks before I can take off ""SSC Chair"" off my email. Its like the torturous 4th mile before everything stops hurting."\t102985898',
'Mitsubishi i MIEV - Wikipedia, the free encyclopedia - http://goo.gl/xipe Cutest car ever!\t109017669432377344\t1.0',
'Cheap Eats in SLP - http://t.co/4w8gRp7\t109642968603963392\t1.0']

[ ]:
```

## Create a Spark SQL Dataframe from Different Ways:

```
localhost:8888/notebooks/week2-sql-dataframe.ipynb 110% ☆ ⌵ ⌵ ⌵

Jupyter week2-sql-dataframe Last Checkpoint: 18 minutes ago Trusted

File Edit View Run Kernel Settings Help

+ + + Code

JupyterLab Python 3 (ipykernel)

df.show()

+-----+-----+
|my_column1|my_column2|
+-----+-----+
|      a|      1|
|      b|      2|
+-----+-----+

[40]: df.dtypes

Out[9]: [('my_column1', 'string'), ('my_column2', 'bigint')]

[41]: my_list = [[('a', 1), ('b', 2)]]

df = spark.createDataFrame(my_list, ['x', 'y'])

df.show()

+-----+-----+
|      x|      y|
+-----+-----+
|[a, 1]| [b, 2]|
+-----+-----+

[42]: df.select("x").show()

+-----+
|      x|
+-----+
|[a, 1]|
+-----+
```

## Convert between RDD and Dataframe:

```
localhost:8888/notebooks/week2-conversion-between-rdd-and-dataframe.ipynb 110% ☆ ⌵ ⌵ ⌵

Jupyter week2-conversion-between-rdd-and-dataframe Last Checkpoint: 7 minutes ago Trusted

File Edit View Run Kernel Settings Help

+ + + Code

JupyterLab Python 3 (ipykernel)

Out[8]: Row(a=1, b=2, c=3)
<----->

[20]: #Let's define a function

def list_to_row(keys, values):

    row_dict = dict(zip(keys, values))

    return Row(**row_dict)

[21]: rdd_rows = rdd.map(lambda x: list_to_row(header, x))

rdd_rows.take(3)

[21]: [Row(model='Mazda RX4', mpg='21', cyl='6', disp='160', hp='110', drat='3.9', wt='2.62', qsec='16.46', vs='0', am='1', gear='4', carb='4'),
Row(model='Mazda RX4 Wag', mpg='21', cyl='6', disp='160', hp='110', drat='3.9', wt='2.875', qsec='17.02', vs='0', am='1', gear='4', carb='4'),
Row(model='Datsun 710', mpg='22.8', cyl='4', disp='108', hp='93', drat='3.85', wt='2.32', qsec='18.61', vs='1', am='1', gear='4', carb='1')]

[22]: #Now we can convert the RDD to a DataFrame.

df = spark.createDataFrame(rdd_rows)

df.show(5)

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      model|mpg|cyl|disp|hp|drat|  wt|  qsec|  vs|am|gear|carb|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      Mazda RX4|  21|  6| 160|110| 3.9| 2.62|16.46|  0| 1|  4|  4|
|      Mazda RX4 Wag|  21|  6| 160|110| 3.9| 2.875|17.02|  0| 1|  4|  4|
|      Datsun 710| 22.8|  4| 108| 93| 3.85| 2.32|18.61|  1| 1|  4|  1|
|      Hornet 4 Drive| 21.4|  6| 258|110| 3.08| 3.215|19.44|  1| 0|  3|  1|
|      Hornet Sportabout| 18.7|  8| 360|175| 3.15| 3.44|17.02|  0| 0|  3|  2|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

## Map and other RDD methods

```
localhost:8888/notebooks/week2-map-function.ipynb 110% ☆ ⓘ

Jupyter week2-map-function Last Checkpoint: 9 minutes ago

File Edit View Run Kernel Settings Help Trusted

+ ✂ 📄 ▶ ■ 🔁 ⏪ Code ▼ JupyterLab Python 3 (pykernel)

header = map_exp_rdd_1.first()

# the filter method apply a function to each elemnts. The function output is a boolean value (TRUE or FALSE)

# elements that have output TRUE will be kept.

map_exp_rdd_2 = map_exp_rdd_1.filter(lambda x: x != header)

map_exp_rdd_2.take(4)

[5]: [('Mazda RX4',
      ['21', '6', '160', '110', '3.9', '2.62', '16.46', '0', '1', '4', '4']),
      ('Mazda RX4 Wag',
      ['21', '6', '160', '110', '3.9', '2.875', '17.02', '0', '1', '4', '4']),
      ('Datsun 710',
      ['22.8', '4', '108', '93', '3.85', '2.32', '18.61', '1', '1', '4', '1']),
      ('Hornet 4 Drive',
      ['21.4', '6', '258', '110', '3.00', '3.215', '19.44', '1', '0', '3', '1'])]

[6]: # convert string values to numeric values

map_exp_rdd_3 = map_exp_rdd_2.map(lambda x: (x[0], list(map(float, x[1]))))

map_exp_rdd_3.take(4)

[6]: [('Mazda RX4',
      [21.0, 6.0, 160.0, 110.0, 3.9, 2.62, 16.46, 0.0, 1.0, 4.0, 4.0]),
      ('Mazda RX4 Wag',
      [21.0, 6.0, 160.0, 110.0, 3.9, 2.875, 17.02, 0.0, 1.0, 4.0, 4.0]),
      ('Datsun 710',
      [22.8, 4.0, 108.0, 93.0, 3.85, 2.32, 18.61, 1.0, 1.0, 4.0, 1.0]),
      ('Hornet 4 Drive',
      [21.4, 6.0, 258.0, 110.0, 3.00, 3.215, 19.44, 1.0, 0.0, 3.0, 1.0])]

[7]:

mapValues_exp_rdd = map_exp_rdd_3

mapValues_exp_rdd.take(4)

[7]: {'Mazda RX4':
```