In [1]:
```python
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
from pyspark.sql.types import *

spark = SparkSession\
    .builder\
    .appName("chapter-31-pandas")\
    .getOrCreate()

import os
SPARK_BOOK_DATA_PATH = os.environ['SPARK_BOOK_DATA_PATH']
```

In [2]:
```python
import pandas as pd
```

In [3]:
```python
df = pd.DataFrame({"first":range(200), "second":range(50,250)})
df.head()
```

Out[3]:

	first	second
0	0	50
1	1	51
2	2	52
3	3	53
4	4	54

convert Pandas DataFrame to Spark DataFrame

```
sparkDF = spark.createDataFrame(df)
```

In [4]:
```python
sparkDF = spark.createDataFrame(df)
sparkDF.show(5)
```

```
+-----+------+
|first|second|
+-----+------+
|    0|    50|
|    1|    51|
|    2|    52|
|    3|    53|
|    4|    54|
+-----+------+
only showing top 5 rows
```

convert Pandas DataFrame to Python

```
obj = df.to_dict()

obj = df.to_list()
```

In [10]:
```
obj = df.to_dict()
```

In [14]:
```
obj
```

```
 14: 14,
 15: 15,
 16: 16,
 17: 17,
 18: 18,
 19: 19,
 20: 20,
 21: 21,
 22: 22,
 23: 23,
 24: 24,
 25: 25,
 26: 26,
 27: 27,
 28: 28,
 29: 29,
 30: 30,
 31: 31,
 32: 32,
```

Why xarray

https://xarray.pydata.org/en/stable/why-xarray.html (https://xarray.pydata.org/en/stable/why-xarray.html)

In [17]:
```
!pip install xarray
```

```
Collecting xarray
  Downloading xarray-0.17.0-py3-none-any.whl (759 kB)
     |[redacted]| 759 kB 2.2 MB/s eta 0:00:01
Requirement already satisfied: setuptools>=40.4 in /usr/lib/python3/di
st-packages (from xarray) (45.2.0)
Requirement already satisfied: numpy>=1.15 in /home/wengong/.local/li
b/python3.8/site-packages (from xarray) (1.19.4)
Requirement already satisfied: pandas>=0.25 in /home/wengong/.local/li
b/python3.8/site-packages (from xarray) (1.1.4)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/lib/pyth
on3/dist-packages (from pandas>=0.25->xarray) (2.7.3)
Requirement already satisfied: pytz>=2017.2 in /usr/lib/python3/dist-p
ackages (from pandas>=0.25->xarray) (2019.3)
Installing collected packages: xarray
Successfully installed xarray-0.17.0
```

In [19]:
```
obj = df.to_xarray()
obj
```

Out[19]:

xarray.Dataset

▸ Dimensions: (**index**: 200)

▾ Coordinates:

index	(index)	int64	0 1 2 3 4 5 ... 195 196 197 198 199

▾ Data variables:

first	(index)	int64	0 1 2 3 4 5 ... 195 196 197 198 199
second	(index)	int64	50 51 52 53 54 ... 246 247 248 249

▸ Attributes: (0)

In [21]:
```
obj.first[:5]
```

Out[21]:

xarray.DataArray 'first' (**index**: 5)

```
array([0, 1, 2, 3, 4])
```

▾ Coordinates:

index	(index)	int64	0 1 2 3 4

▸ Attributes: (0)

In [22]:
```
narray = df.to_numpy()
```

In [24]:
```
narray[:5]
```

Out[24]:
```
array([[ 0, 50],
       [ 1, 51],
       [ 2, 52],
       [ 3, 53],
       [ 4, 54]])
```

convert Spark DataFrame to Pandas DataFrame

```
pandasDF = sparkDF.toPandas()
```

In [5]:
```
pandasDF = sparkDF.toPandas()
pandasDF.head()
```

Out[5]:

	first	second
0	0	50
1	1	51
2	2	52
3	3	53
4	4	54

convert Spark DataFrame to Python

```
dfobj = sparkDF.collect()
```

In [8]:
```
dfobj = sparkDF.collect()
dfobj[:5]
```

Out[8]:
```
[Row(first=0, second=50),
 Row(first=1, second=51),
 Row(first=2, second=52),
 Row(first=3, second=53),
 Row(first=4, second=54)]
```

In [ ]:
```

```