Data Frames Primer

In [5]:

```python
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName('Dataframe').getOrCreate()

print('Data Frames Primer')

print('\n 1. Input the following data into a data frame called titanic, and displ

data = [('Children', 'First', 6, 0) , ('Children', 'Second', 24, 0) ,
        ('Children', 'Third', 27, 52), ('Men', 'First', 57, 118),
        ('Men', 'Second', 14, 154), ('Men', 'Third', 75, 387),
        ('Men', 'Crew', 192, 693) , ('Women', 'First', 140, 4) ,
        ('Women', 'Second', 80, 13), ('Women', 'Third', 76, 89) ,
        ('Women', 'Crew', 20, 3 )]

columns=['Sex', 'Class', 'Survived', 'Died']

ts = spark.createDataFrame(data = data, schema = columns)

ts.show()
```

```
Data Frames Primer

 1. Input the following data into a data frame called titanic, and display the
entire data frame:

+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|Children| Third|      27|  52|
|     Men| First|      57| 118|
|     Men|Second|      14| 154|
|     Men| Third|      75| 387|
|     Men|  Crew|     192| 693|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
|   Women| Third|      76|  89|
|   Women|  Crew|      20|   3|
+--------+------+--------+----+
```

In [6]:
```
print('2. Delete the crew members from the data.')

ts = ts.filter(ts.Class != "Crew")

ts.show()
```

```
2. Delete the crew members from the data.
+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|Children| Third|      27|  52|
|     Men| First|      57| 118|
|     Men|Second|      14| 154|
|     Men| Third|      75| 387|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
|   Women| Third|      76|  89|
+--------+------+--------+----+
```

In [7]:
```
print('3. Create a new column that is the total number of people for that group (

ts = ts.withColumn("Total", ts.Survived+ ts.Died)

ts.show()
```

```
3. Create a new column that is the total number of people for that group (those
who survived + died).

+--------+------+--------+----+-----+
|     Sex| Class|Survived|Died|Total|
+--------+------+--------+----+-----+
|Children| First|       6|   0|    6|
|Children|Second|      24|   0|   24|
|Children| Third|      27|  52|   79|
|     Men| First|      57| 118|  175|
|     Men|Second|      14| 154|  168|
|     Men| Third|      75| 387|  462|
|   Women| First|     140|   4|  144|
|   Women|Second|      80|  13|   93|
|   Women| Third|      76|  89|  165|
+--------+------+--------+----+-----+
```

In [8]:

```
print('4. Delete the column indicating the total number of people in that group.\

ts = ts.drop("Total")

ts.show()
```

```
4. Delete the column indicating the total number of people in that group.

+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|Children| Third|      27|  52|
|     Men| First|      57| 118|
|     Men|Second|      14| 154|
|     Men| Third|      75| 387|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
|   Women| Third|      76|  89|
+--------+------+--------+----+
```

In [9]:

```
print('5. Only show the rows where more than 80% of the people survived.\n')


ts.filter((ts.Survived/(ts.Survived+ ts.Died))*100 > 80).show()
```

```
5. Only show the rows where more than 80% of the people survived.

+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
+--------+------+--------+----+
```

In [ ]: