In [2]:
```
# import findspark
# findspark.init()
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
from pyspark.sql.types import *

spark = SparkSession\
    .builder\
    .appName("chapter-19-perf")\
    .getOrCreate()

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

In [3]:
```
file_path = SPARK_BOOK_DATA_PATH + "/data/flight-data/csv/2015-summary.
# Original loading code that does *not* cache DataFrame
DF1 = spark.read.format("csv")\
  .option("inferSchema", "true")\
  .option("header", "true")\
  .load(file_path)
```

In [4]:
```
DF1.show(5)
```

```
+-----------------+-------------------+-----+
|DEST_COUNTRY_NAME|ORIGIN_COUNTRY_NAME|count|
+-----------------+-------------------+-----+
|    United States|            Romania|   15|
|    United States|            Croatia|    1|
|    United States|            Ireland|  344|
|            Egypt|      United States|   15|
|    United States|              India|   62|
+-----------------+-------------------+-----+
only showing top 5 rows
```

In [5]:
```
%%time
DF2 = DF1.groupBy("DEST_COUNTRY_NAME").count().collect()
```

```
CPU times: user 15.5 ms, sys: 5.25 ms, total: 20.8 ms
Wall time: 2.41 s
```

In [6]:
```
%%time
DF3 = DF1.groupBy("ORIGIN_COUNTRY_NAME").count().collect()
```

```
CPU times: user 6.22 ms, sys: 1.2 ms, total: 7.42 ms
Wall time: 867 ms
```

In [8]:
```
%%time
DF4 = DF1.groupBy("count").count().collect()
```

```
CPU times: user 6.49 ms, sys: 824 µs, total: 7.31 ms
Wall time: 935 ms
```

In [9]:
```
# COMMAND ----------

DF1.cache()
```

Out[9]:
```
DataFrame[DEST_COUNTRY_NAME: string, ORIGIN_COUNTRY_NAME: string, coun
t: int]
```

In [10]:
```
%%time
DF1.count()
```

```
CPU times: user 29 µs, sys: 2.72 ms, total: 2.75 ms
Wall time: 389 ms
```

Out[10]:
```
256
```

In [11]:
```
%%time
DF2 = DF1.groupBy("DEST_COUNTRY_NAME").count().collect()
```

```
CPU times: user 6.28 ms, sys: 735 µs, total: 7.02 ms
Wall time: 740 ms
```

In [12]:
```
%%time
DF3 = DF1.groupBy("ORIGIN_COUNTRY_NAME").count().collect()
```

```
CPU times: user 7.31 ms, sys: 1.13 ms, total: 8.44 ms
Wall time: 596 ms
```

In [13]:
```
DF4 = DF1.groupBy("count").count().collect()
```

In [14]:
```
DF4[:10]
```

Out[14]:
```
[Row(count=31, count=1),
 Row(count=2025, count=1),
 Row(count=588, count=1),
 Row(count=53, count=1),
 Row(count=853, count=1),
 Row(count=362, count=1),
 Row(count=1468, count=1),
 Row(count=155, count=1),
 Row(count=108, count=1),
 Row(count=211, count=1)]
```

In [15]:
```
DF1.is_cached
```

Out[15]:
```
True
```

In [16]:
```
DF1.unpersist()
```

Out[16]:
```
DataFrame[DEST_COUNTRY_NAME: string, ORIGIN_COUNTRY_NAME: string, coun
t: int]
```

In [ ]:
```

```