```
pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10
```

```
from pyspark.sql import SparkSession


spark = SparkSession.builder.appName("NYSE and Airlines data analysis").getOrCreate()


from pyspark.sql.types import StructType, StringType, IntegerType, DoubleType, LongType


schema_nyse = StructType().add("exchange_name",StringType(),True).add("stock_id",StringType(),True).add("stock_dt



print(schema_nyse)
```

```
StructType([StructField('exchange_name', StringType(), True), StructField('stock_id', StringType(), True), St
```

```
df_with_schema = spark.read.format("csv").option("header","False").schema(schema_nyse).load("/content/sample_data



df_with_schema.printSchema()
```

```
root
 |-- exchange_name: string (nullable = true)
 |-- stock_id: string (nullable = true)
 |-- stock_dt: string (nullable = true)
 |-- open: double (nullable = true)
 |-- high: double (nullable = true)
 |-- low: double (nullable = true)
 |-- close: double (nullable = true)
 |-- volume: long (nullable = true)
 |-- adj_close: double (nullable = true)
```

```
df_with_schema.count()
```

```
735026
```

```
df_with_schema.rdd.getNumPartitions()
```

```
2
```

```
df_with_schema.show(30)
```

```
+-------------+--------+----------+----+----+----+-----+------+---------+
|exchange_name|stock_id|  stock_dt|open|high| low|close|volume|adj_close|
+-------------+--------+----------+----+----+----+-----+------+---------+
|         NYSE|     AEA|2010-02-08|4.42|4.42|4.21| 4.24|205500|     4.24|
|         NYSE|     AEA|2010-02-05|4.42|4.54|4.22| 4.41|194300|     4.41|
|         NYSE|     AEA|2010-02-04|4.55|4.69|4.39| 4.42|233800|     4.42|
|         NYSE|     AEA|2010-02-03|4.65|4.69| 4.5| 4.55|182100|     4.55|
|         NYSE|     AEA|2010-02-02|4.74| 5.0|4.62| 4.66|222700|     4.66|
|         NYSE|     AEA|2010-02-01|4.84|4.92|4.68| 4.75|194800|     4.75|
|         NYSE|     AEA|2010-01-29|4.97|5.05|4.76| 4.83|222900|     4.83|
|         NYSE|     AEA|2010-01-28|5.12|5.22|4.81| 4.98|283100|     4.98|
|         NYSE|     AEA|2010-01-27|4.82|5.16|4.79| 5.09|243500|     5.09|
|         NYSE|     AEA|2010-01-26|5.18|5.18|4.81| 4.84|554800|     4.84|
|         NYSE|     AEA|2010-01-25|5.42|5.48| 5.2| 5.22|257300|     5.22|
|         NYSE|     AEA|2010-01-22|5.52|5.59|5.31| 5.37|260800|     5.37|
|         NYSE|     AEA|2010-01-21|5.67|5.74|5.37| 5.51|264300|     5.51|
|         NYSE|     AEA|2010-01-20|5.65| 5.7|5.53| 5.66|244600|     5.66|
```

```
|         NYSE|     AEA|2010-01-19|5.54| 5.7|5.54| 5.69|368000|     5.69|
|         NYSE|     AEA|2010-01-15|5.48|5.55|5.33| 5.54|435500|     5.54|
|         NYSE|     AEA|2010-01-14|5.41| 5.5|5.39| 5.41|272200|     5.41|
|         NYSE|     AEA|2010-01-13| 5.5| 5.5|5.41| 5.45|176400|     5.45|
|         NYSE|     AEA|2010-01-12|5.47|5.51|5.41| 5.46|233100|     5.46|
|         NYSE|     AEA|2010-01-11|5.64|5.64|5.49| 5.55|178900|     5.55|
|         NYSE|     AEA|2010-01-08|5.61|5.68|5.52| 5.59|144200|     5.59|
|         NYSE|     AEA|2010-01-07|5.47|5.65| 5.4| 5.62|228900|     5.62|
|         NYSE|     AEA|2010-01-06|5.56| 5.7|5.44| 5.49|208900|     5.49|
|         NYSE|     AEA|2010-01-05|5.55|5.62|5.51| 5.55|267000|     5.55|
|         NYSE|     AEA|2010-01-04|5.65|5.66|5.49| 5.55|335500|     5.55|
|         NYSE|     AEA|2009-12-31|5.57|5.71|5.54| 5.56|418600|     5.56|
|         NYSE|     AEA|2009-12-30|5.65|5.67| 5.5| 5.57|226400|     5.57|
|         NYSE|     AEA|2009-12-29|5.67|5.74|5.66| 5.67|115100|     5.67|
|         NYSE|     AEA|2009-12-28|5.81|5.86|5.63| 5.67|326600|     5.67|
|         NYSE|     AEA|2009-12-24|5.92|5.94|5.81| 5.84|111900|     5.84|
+-------------+--------+----------+----+----+----+-----+------+---------+
only showing top 30 rows
```

```
df_with_schema.createOrReplaceTempView("nyse")
```

```
df_StockVol = spark.sql("select stock_id, sum(volume) as total from nyse group by stock_id")
```

```
df_StockVol.count()
```

```
203
```

```
df_StockVol.show(203)
```

```
|     ACL| 1176516200|
|     AGO| 1356870600|
|     ARJ|  289810600|
|     ACG| 1481168200|
|     AXR|  107629900|
|     ATK|  933991800|
|     ASX| 1045139800|
|     ALJ|  428456900|
|     ABC|11439581700|
|     AGP| 1425712200|
|     AZO| 3366821200|
|     AUY|11034706100|
|     AWC|  259152600|
|     AVF|  129141600|
|     AIQ|  387333900|
|      AF| 2789196400|
+--------+-----------+
```

```
df_StockVol.rdd.getNumPartitions()
```

```
1
```

```
df_StockVol.write.csv("/content/sample_data/spark-sql1")
```

```
# get the same output using pyspark rdd code
```

```
sc = spark.sparkContext
```

```
rdd1 = sc.textFile("/content/sample_data/NYSE.csv")
rdd2 = rdd1.map(lambda a : (a.split(",")[1], int(a.split(",")[7])   ))
rdd3 = rdd2.reduceByKey(lambda a,b : a+b)
```

```
for line in rdd3.collect():
  print(line)
```

```
('AA', 42061448400)
('ATU', 1226088700)
('ARG', 1713739100)
('AEB', 53273300)
('AEO', 14731442100)
('APL', 364876100)
('AKT', 41654000)
('AGD', 100765300)
('AFB', 98894100)
('AVT', 3427089600)
('APD', 5601186900)
('ATT', 99347600)
('ADI', 14597316000)
('ALV', 1339964100)
('AVK', 123961500)
('AHS', 615786600)
('ARD', 691227500)
('AMX', 11000819500)
('AOL', 147580700)
('APH', 3807963100)
('ADM', 15354593500)
('ANH', 1407062000)
('AP', 158385300)
('AZN', 3418077300)
```

```
year = 1995
print(type(year))
```

```
<class 'int'>
```

```
# airlines use case
```

```
airlines_DF = spark.read.format("csv").option("header", "True").option("inferSchema", "True").load("/content/sampl
```

```
airlines_DF.printSchema()
```

```
root
 |-- Year: integer (nullable = true)
 |-- Quarter: integer (nullable = true)
 |-- Avg_rev_per_seat: double (nullable = true)
 |-- booked_seats: integer (nullable = true)
```

```
airlines_DF.count()
```

```
84
```

```
airlines_DF.show()
```

```
+----+-------+----------------+------------+
|Year|Quarter|Avg_rev_per_seat|booked_seats|
+----+-------+----------------+------------+
|1995|      1|           296.9|       46561|
|1995|      2|           296.8|       37443|
|1995|      3|          287.51|       34128|
|1995|      4|          287.78|       30388|
|1996|      1|          283.97|       47808|
|1996|      2|          275.78|       43020|
|1996|      3|          269.49|       38952|
|1996|      4|          278.33|       37443|
|1997|      1|           283.4|       35067|
|1997|      2|          289.44|       46565|
|1997|      3|          282.27|       38886|
|1997|      4|          293.51|       37454|
|1998|      1|          304.74|       31315|
|1998|      2|          300.97|       30852|
|1998|      3|          315.25|       38118|
|1998|      4|          316.18|       35393|
|1999|      1|          331.74|       47453|
```

```
|1999|      2|          329.34|       38243|
|1999|      3|          317.22|       33048|
|1999|      4|          317.93|       31256|
+----+-------+-----------------+------------+
only showing top 20 rows
```

```
airlines_DF.createOrReplaceTempView("airlines")
```

```
YrWiseRev = spark.sql("select year, sum(Avg_rev_per_seat*booked_seats) as total from airlines group by year order
```

```
YrWiseRev.show()
```

```
+----+--------------------+
|year|               total|
+----+--------------------+
|2013|       6.636320871E7|
|2014| 6.262417585000001E7|
|2015|       6.237899057E7|
|2012|       6.219912728E7|
|2008|5.7653170760000005E7|
|2007|       5.730921607E7|
|2001| 5.553377999999999E7|
|2010|       5.486152129E7|
|2000|5.2342926550000004E7|
|2011|       5.188828622E7|
|2004|5.0631364949999996E7|
|2006|5.0437898419999994E7|
|2003|       4.927321083E7|
|1999|       4.875771448E7|
|2002|        4.74991465E7|
|2009|       4.674644659E7|
|2005|       4.637678624E7|
|1996|       4.635877803E7|
|1997|       4.538523616E7|
|1995|       4.349424322E7|
+----+--------------------+
only showing top 20 rows
```

```
YrWiseRev = spark.sql("select year, sum(Avg_rev_per_seat*booked_seats)/1000000 as total_in_mill from airlines grc
```

```
YrWiseRev.show()
```

```
+----+------------------+
|year|     total_in_mill|
+----+------------------+
|2013|       66.36320871|
|2014| 62.62417585000001|
|2015|       62.37899057|
|2012|       62.19912728|
|2008|       57.65317076|
|2007|       57.30921607|
|2001| 55.53377999999999|
|2010|       54.86152129|
|2000|       52.34292655|
|2011|       51.88828622|
|2004|       50.63136495|
|2006|50.437898419999996|
|2003|       49.27321083|
|1999|       48.75771448|
|2002|        47.4991465|
|2009|46.746446590000005|
|2005|       46.37678624|
|1996|       46.35877803|
|1997|       45.38523616|
|1995|       43.49424322|
+----+------------------+
only showing top 20 rows
```

```
YrWiseRev = spark.sql("select year, round(sum(Avg_rev_per_seat*booked_seats)/1000000,2) as total_in_mill from air
```

```
YrWiseRev.show()
```

```
+----+-------------+
|year|total_in_mill|
+----+-------------+
|2013|        66.36|
|2014|        62.62|
|2015|        62.38|
|2012|         62.2|
|2008|        57.65|
|2007|        57.31|
|2001|        55.53|
|2010|        54.86|
|2000|        52.34|
|2011|        51.89|
|2004|        50.63|
|2006|        50.44|
|2003|        49.27|
|1999|        48.76|
|2002|         47.5|
|2009|        46.75|
|2005|        46.38|
|1996|        46.36|
|1997|        45.39|
|1995|        43.49|
+----+-------------+
only showing top 20 rows
```

```
YrWisePsx = spark.sql("select year, sum(booked_seats) as total_psx from airlines group by year order by total_psx
```

```
YrWisePsx.show()
```

```
+----+---------+
|year|total_psx|
+----+---------+
|2007|   176299|
|2013|   173676|
|2001|   173598|
|1996|   167223|
|2008|   166897|
|2012|   166076|
|2015|   165438|
|2004|   164800|
|2010|   163741|
|2014|   159823|
|1997|   157972|
|2003|   156153|
|2000|   154376|
|2006|   153789|
|2002|   152195|
|2005|   150610|
|2009|   150308|
|1999|   150000|
|1995|   148520|
|2011|   142647|
+----+---------+
only showing top 20 rows
```