Big Data Analytic Techniques and Applications
Homework2 report

310551158 林鑫伯

colab source code:
https://colab.research.google.com/drive/197faFqVs9DDihtP8r6gid0ASLw9wnW7t?usp=sharing
github: https://github.com/gleeshot/NYCU_BIGDATA_HW2

development environment

Operating System: Ubuntu 20.04 LTS
Memory Size: 16 GiB
CPU: Intel core i7-6700
tools: pyspark, colab

Questions

1. Find the maximal delays (you should consider both ArrDelay and DepDelay) for each month of 2007.

For each month, I assigned the column of ArrDelay and DepDelay to two dataframes. Both of them cut the null values then did descending sort.

```
months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']
for i in range(1, 13):
  print(months[i - 1] + ': ')
  print('DepDelay')
  df_month = df.filter(df.Month == i)
  df_month_DepDelay = df_month.filter("DepDelay != 'NA'")
  df_month_DepDelay = df_month_DepDelay.withColumn("DepDelay", df_month_DepDelay["DepDelay"].cast('int'))
  df_month_DepDelay.select(["DepDelay"]).orderBy(desc("DepDelay")).show(5)

  print('ArrDelay')
  df_month_ArrDelay = df_month.filter("ArrDelay != 'NA'")
  df_month_ArrDelay = df_month_ArrDelay.withColumn("ArrDelay", df_month_ArrDelay["ArrDelay"].cast('int'))
  df_month_ArrDelay.select(["ArrDelay"]).orderBy(desc("ArrDelay")).show(5)
```

Month	ArrDelay	DepDelay
Jan	1426	1406
Feb	1359	1340
Mar	1564	1547
Apr	1402	1415
May	1429	1416
Jun	1351	1360
Jul	1386	1369
Aug	1472	1449
Sep	1665	1689
Oct	2598	2601
Nov	1146	1137
Dec	1942	1956

2. How many flights were delayed caused by security between 2000 ~ 2005? Please show the counting for each year.

For csv file each year, I cleared the null value in the file,and made a filter to select rows that security delay > 0, and then counted the number of rows.

Since it is impossible for no security delay during the period of 2000-2002, I guessed there may be no record of security delay in those files.

```
files = ['2000.csv', '2001.csv', '2002.csv', '2003.csv', '2004.csv', '2005.csv']
total_SecurityDelay_count = 0
for file in files:
  fp = "/content/gdrive/MyDrive/bigdata_hw2/" + file
  df_SecurityDelay = spark.read.csv(fp, header=True, inferSchema=True)
  df_SecurityDelay.na.fill(0)
  df_SecurityDelay = df_SecurityDelay.withColumn("SecurityDelay", df_SecurityDelay["SecurityDelay"].cast('int'))
  df_SecurityDelay = df_SecurityDelay.filter("SecurityDelay > 0")

  tmp = df_SecurityDelay.count()
  total_SecurityDelay_count += tmp
  print(file[0:4] + ' has ' + str(tmp) + ' SecurityDelay.')
  if tmp != 0:
    df_SecurityDelay.select(["Year", "Month", "SecurityDelay"]).orderBy(desc("SecurityDelay")).show(5)

print("there are " + str(total_SecurityDelay_count) + " SecurityDelays between 2000 - 2005.")
```

```
2000 has 0 SecurityDelay.
2001 has 0 SecurityDelay.
2002 has 0 SecurityDelay.
2003 has 3740 SecurityDelay.
+----+-----+-------------+
|Year|Month|SecurityDelay|
+----+-----+-------------+
|2003|   12|          230|
|2003|    6|          218|
|2003|    6|          214|
|2003|   11|          208|
|2003|    6|          204|
+----+-----+-------------+
only showing top 5 rows

2004 has 8158 SecurityDelay.
+----+-----+-------------+
|Year|Month|SecurityDelay|
+----+-----+-------------+
|2004|    8|          533|
|2004|    8|          451|
|2004|    8|          382|
|2004|    8|          380|
|2004|    9|          312|
+----+-----+-------------+
only showing top 5 rows

2005 has 6627 SecurityDelay.
+----+-----+-------------+
|Year|Month|SecurityDelay|
+----+-----+-------------+
|2005|    8|          326|
|2005|    7|          294|
|2005|    8|          282|
|2005|   12|          261|
|2005|    7|          259|
+----+-----+-------------+
only showing top 5 rows

there are 18525 SecurityDelays between 2000 - 2005.
```

3. List Top 5 airports which occur delays most and least in 2008. (Please show the IATA airport code)

At first, I did the same thing just as Q1 and Q2, clearing the null value in the dataframes, making filters to select rows we need, and then replaced the column names both “ArrDelay” and “DepDelay” to “IATA”.

```
fp = "/content/gdrive/MyDrive/bigdata_hw2/2008.csv"
df = spark.read.csv(fp, header=True, inferSchema=True)

df_DepDelay = df.select(["Origin", "DepDelay"])
df_ArrDelay = df.select(["Dest", "ArrDelay"])

# clear na
df_DepDelay.na.fill(0)
df_ArrDelay.na.fill(0)

df_DepDelay = df_DepDelay.filter("DepDelay > 0")
df_ArrDelay = df_ArrDelay.filter("ArrDelay > 0")

origin_delays = df_DepDelay.withColumn("IATA", df_DepDelay["Origin"])
print("most 5 origin delays")
top_origin = origin_delays.groupBy("IATA").count().orderBy(desc("count")).show(5)
print("least 5 origin delays")
least_origin = origin_delays.groupBy("IATA").count().orderBy("count").show(5)

dest_delays = df_ArrDelay.withColumn("IATA", df_ArrDelay["Dest"])
print("most 5 dest delays")
top_dest = dest_delays.groupBy("IATA").count().orderBy(desc("count")).show(5)
print("least 5 dest delays")
least_dest = dest_delays.groupBy("IATA").count().orderBy("count").show(5)
```

Second, I used union to concatenate the information in the dataframes we just got. And then, counting the times of delay to get the answer (next page).

```
total_delays = origin_delays.union(dest_delays)
print("most 5 total delays")
top_total = total_delays.groupBy("IATA").count().orderBy(desc("count")).show(5)
print("least 5 total delays")
least_total = total_delays.groupBy("IATA").count().orderBy("count").show(5)
```

The top 5 airports that most delay occur are ATL, ORD, DFW, DEN, LAX and the least 5 are PUB, TUP, PIR, BJI, INL.

```
most 5 total delays
+----+------+
|IATA| count|
+----+------+
| ATL|362260|
| ORD|311298|
| DFW|247566|
| DEN|207095|
| LAX|180766|
+----+------+
only showing top 5 rows

least 5 total delays
+----+-----+
|IATA|count|
+----+-----+
| PUB|    2|
| TUP|    3|
| PIR|    7|
| BJI|   16|
| INL|   18|
+----+-----+
only showing top 5 rows
```