In [59]:
```
#sc.stop()
```

In [60]:
```
#spark.stop()
```

a) Create a new Spark Session with new SparkConfig

In [61]:
```
# from pyspark import SparkContext, SparkConf
# config = SparkConf().setAppName("PySparkSession").setMaster("local[4]")
# sc = SparkContext(conf = config)
```

In [62]:
```
# from pyspark.sql import SparkSession
# spark = SparkSession.builder.appName("PySparkSession").getOrCreate()
```

In [63]:
```
sc
```

Out[63]:

SparkContext

Spark UI

Version	v2.4.8
Master	local[4]
AppName	PySparkSession

In [64]:
```
spark
```

Out[64]:

SparkSession - hive

SparkContext

Spark UI

Version	v2.4.8
Master	local[4]
AppName	PySparkSession

b) Create new instance of Spark SQL session and define new DataFrame using Flights_Delay.csv dataset.

In [65]:
```
flights_delay_df = spark.read.csv("file:///home/hadoop/Downloads/Flights_D
```

In [66]:
```
flights_delay_df.show()
```

```
+---+----+-----+---+-----------+-------+-------------+-----------+---------
-----+------------------+------------------+--------------+--------------
-+--------+----------+--------------+------------+--------+--------+-------
--+-------+-----------------+------------+-------------+--------+---------+
-------------------+----------------+--------------+-------------+---------
----------+-------------+
| ID|YEAR|MONTH|DAY|DAY_OF_WEEK|AIRLINE|FLIGHT_NUMBER|TAIL_NUMBER|ORIGIN_AI
```

```
RPORT|DESTINATION_AIRPORT|SCHEDULED_DEPARTURE|DEPARTURE_TIME|DEPARTURE_DELA
Y|TAXI_OUT|WHEELS_OFF|SCHEDULED_TIME|ELAPSED_TIME|AIR_TIME|DISTANCE|WHEELS_
ON|TAXI_IN|SCHEDULED_ARRIVAL|ARRIVAL_TIME|ARRIVAL_DELAY|DIVERTED|CANCELLED|
CANCELLATION_REASON|AIR_SYSTEM_DELAY|SECURITY_DELAY|AIRLINE_DELAY|LATE_AIRC
RAFT_DELAY|WEATHER_DELAY|
+---+----+-----+---+----------+-------+------------+----------+---------
-----+------------------+------------------+--------------+-------------
-+--------+----------+--------------+-----------+--------+--------+-------
--+-------+-----------------+-----------+------------+--------+---------+
-------------------+---------------+--------------+-------------+---------
----------+-------------+
|  0|2015|    3|  4|         3|     EV|         5170|     N842AS|
CVG|               XNA|               935|          954|             19|
16|      1010|          115|         129|     108|     562|     1058|
5|             1030|        1103|           33|       0|        0|
null|              14|            0|          19|                  0|
0|
|  1|2015|    2|  2|         1|     MQ|         3584|     N646MQ|
DFW|               SPS|              1240|         1316|             36|
11|      1327|           50|          46|      30|     113|     1357|
5|             1330|        1402|           32|       0|        0|
null|               0|            0|          32|                  0|
0|
|  2|2015|    1| 27|         2|     B6|          716|     N309JB|
JAX|               DCA|              1335|         1505|             90|
16|      1521|          104|         110|      91|     634|     1652|
3|             1519|        1655|           96|       0|        0|
null|               6|            0|          90|                  0|
0|
|  3|2015|    1| 28|         3|     EV|         4289|     N14162|
COS|               IAH|              1442|         1435|             -7|
13|      1448|          139|         127|     101|     809|     1729|
13|             1801|        1742|          -19|       0|        0|
null|            null|         null|         null|               null|
null|
|  4|2015|    2|  5|         4|     EV|         5584|     N851AS|
ATL|               AVL|              1255|         1250|             -5|
25|      1315|           48|          62|      34|     164|     1349|
3|             1343|        1352|            9|       0|        0|
null|            null|         null|         null|               null|
null|
|  5|2015|    2| 15|         7|     UA|          712|     N438UA|
IAH|               SFO|              1535|         1554|             19|
18|      1612|          260|         237|     216|    1635|     1748|
3|             1755|        1751|           -4|       0|        0|
null|            null|         null|         null|               null|
null|
|  6|2015|    2| 19|         4|     OO|         5166|     N746SK|
HDN|               DEN|               928|          924|             -4|
11|       935|           67|          56|      29|     141|     1004|
16|             1035|        1020|          -15|       0|        0|
null|            null|         null|         null|               null|
null|
|  7|2015|    2| 27|         5|     DL|         1571|     N916DN|
ATL|               CAK|              2104|         2103|             -1|
20|      2123|          106|          97|      70|     528|     2233|
7|             2250|        2240|          -10|       0|        0|
null|            null|         null|         null|               null|
null|
|  8|2015|    1| 20|         2|     WN|          518|     N405WN|
HOU|               MEM|              2140|         2150|             10|
8|      2158|           80|          79|      68|     484|     2306|
3|             2300|        2309|            9|       0|        0|
null|            null|         null|         null|               null|
```

```
null|
|  9|2015|    2|  6|          5|     WN|          336|     N663SW|
DAL|                MAF|               1750|          1748|             -2|
7|      1755|            70|           62|      52|     319|     1847|
3|              1900|        1850|           -10|       0|        0|
null|            null|          null|          null|               null|
null|
| 10|2015|    3|  6|          5|     UA|          321|     N455UA|
TPA|                EWR|                950|           947|             -3|
11|       958|           160|          142|     119|     997|     1157|
12|             1230|        1209|           -21|       0|        0|
null|            null|          null|          null|               null|
null|
| 11|2015|    3|  6|          5|     F9|         1343|     N905FR|
TPA|                CLE|               1630|          1622|             -8|
13|      1635|           145|          146|     125|     927|     1840|
8|              1855|        1848|           -7|       0|        0|
null|            null|          null|          null|               null|
null|
| 12|2015|    1| 30|          5|     WN|         2685|     N638SW|
LAS|                SJC|                620|           633|             13|
13|       646|            90|           76|      59|     386|      745|
4|               750|         749|           -1|       0|        0|
null|            null|          null|          null|               null|
null|
| 13|2015|    1| 11|          7|     HA|          371|     N492HA|
ITO|                HNL|               1930|          2000|             30|
10|      2010|            50|           53|      36|     216|     2046|
7|              2020|        2053|           33|       0|        0|
null|               0|            0|           17|                 16|
0|
| 14|2015|    2| 23|          1|     OO|         5416|     N927SW|
ONT|                SFO|               1733|          1727|             -6|
15|      1742|            88|           87|      68|     363|     1850|
4|              1901|        1854|           -7|       0|        0|
null|            null|          null|          null|               null|
null|
| 15|2015|    1| 10|          6|     MQ|         3196|     N607MQ|
SAF|                DFW|               1100|          1051|             -9|
7|      1058|            95|           97|      76|     551|     1314|     1
4|              1335|        1328|           -7|       0|        0|
null|            null|          null|          null|               null|
null|
| 16|2015|    1|  7|          3|     WN|          432|     N961WN|
PDX|                LAS|               1805|          1801|             -4|
7|      1808|           125|          121|     107|     763|     1955|
7|              2010|        2002|           -8|       0|        0|
null|            null|          null|          null|               null|
null|
| 17|2015|    2| 15|          7|     AS|          350|     N557AS|
SEA|                OAK|               2120|          2119|             -1|
25|      2144|           120|          122|      91|     671|     2315|
6|              2320|        2321|            1|       0|        0|
null|            null|          null|          null|               null|
null|
| 18|2015|    2| 24|          2|     OO|         6196|     N751SK|
ONT|                IAH|                701|           806|             65|
24|       830|           189|          180|     145|    1334|     1255|
11|             1210|        1306|           56|       0|        0|
null|               0|            0|           56|                  0|
0|
| 19|2015|    1| 30|          5|     MQ|         3401|     N609MQ|
SHV|                DFW|               1615|          1611|             -4|
8|      1619|            60|           61|      42|     190|     1701|     1
```

```
1|             1715|        1712|           -3|       0|        0|
null|            null|        null|         null|            null|
null|
+---+----+-----+---+----------+------+------------+----------+---------
-----+------------------+------------------+-------------+-------------
-+--------+----------+-------------+-----------+--------+--------+-------
--+-------+-----------------+-----------+-------------+--------+---------+
-------------------+---------------+-------------+------------+---------
----------+-------------+
only showing top 20 rows
```

c) Create table Spark HIVE table flights_table

In [67]:
```
flights_delay_df.createOrReplaceTempView('flights_table')
```

d) Describe the table schema & show top 10 rows of Dataset

In [68]:
```
from pyspark.sql.types import *
from pyspark.sql.functions import *
```

In [69]:
```
spark.sql("""
desc flights_table
""").show()
```

```
+-------------------+---------+-------+
|           col_name|data_type|comment|
+-------------------+---------+-------+
|                 ID|      int|   null|
|               YEAR|      int|   null|
|              MONTH|      int|   null|
|                DAY|      int|   null|
|        DAY_OF_WEEK|      int|   null|
|            AIRLINE|   string|   null|
|      FLIGHT_NUMBER|      int|   null|
|        TAIL_NUMBER|   string|   null|
|     ORIGIN_AIRPORT|   string|   null|
|DESTINATION_AIRPORT|   string|   null|
|SCHEDULED_DEPARTURE|      int|   null|
|     DEPARTURE_TIME|      int|   null|
|    DEPARTURE_DELAY|      int|   null|
|           TAXI_OUT|      int|   null|
|         WHEELS_OFF|      int|   null|
|     SCHEDULED_TIME|      int|   null|
|       ELAPSED_TIME|      int|   null|
|           AIR_TIME|      int|   null|
|           DISTANCE|      int|   null|
|          WHEELS_ON|      int|   null|
+-------------------+---------+-------+
only showing top 20 rows
```

In [70]:
```
spark.sql("""
select * from flights_table limit 10
""").show()
```

```
+---+----+-----+---+----------+------+------------+----------+---------
-----+------------------+------------------+-------------+-------------
-+--------+----------+-------------+-----------+--------+--------+-------
--+-------+-----------------+-----------+-------------+--------+---------+
```

```
------------------+---------------+-------------+------------+---------
----------+-------------+
| ID|YEAR|MONTH|DAY|DAY_OF_WEEK|AIRLINE|FLIGHT_NUMBER|TAIL_NUMBER|ORIGIN_AI
RPORT|DESTINATION_AIRPORT|SCHEDULED_DEPARTURE|DEPARTURE_TIME|DEPARTURE_DELA
Y|TAXI_OUT|WHEELS_OFF|SCHEDULED_TIME|ELAPSED_TIME|AIR_TIME|DISTANCE|WHEELS_
ON|TAXI_IN|SCHEDULED_ARRIVAL|ARRIVAL_TIME|ARRIVAL_DELAY|DIVERTED|CANCELLED|
CANCELLATION_REASON|AIR_SYSTEM_DELAY|SECURITY_DELAY|AIRLINE_DELAY|LATE_AIRC
RAFT_DELAY|WEATHER_DELAY|
+---+----+-----+---+-----------+-------+-------------+-----------+---------
-----+------------------+------------------+-------------+-------------
-+--------+----------+-------------+------------+--------+--------+-------
--+-------+-----------------+------------+-------------+--------+---------+
------------------+---------------+-------------+------------+---------
----------+-------------+
|  0|2015|    3|  4|          3|     EV|         5170|     N842AS|
CVG|               XNA|                935|           954|             19|
16|      1010|           115|         129|     108|     562|     1058|
5|             1030|        1103|           33|       0|        0|
null|              14|             0|           19|                  0|
0|
|  1|2015|    2|  2|          1|     MQ|         3584|     N646MQ|
DFW|               SPS|               1240|          1316|             36|
11|      1327|            50|          46|      30|     113|     1357|
5|             1330|        1402|           32|       0|        0|
null|               0|             0|           32|                  0|
0|
|  2|2015|    1| 27|          2|     B6|          716|     N309JB|
JAX|               DCA|               1335|          1505|             90|
16|      1521|           104|         110|      91|     634|     1652|
3|             1519|        1655|           96|       0|        0|
null|               6|             0|           90|                  0|
0|
|  3|2015|    1| 28|          3|     EV|         4289|     N14162|
COS|               IAH|               1442|          1435|             -7|
13|      1448|           139|         127|     101|     809|     1729|
13|             1801|        1742|          -19|       0|        0|
null|            null|          null|         null|            null|
null|
|  4|2015|    2|  5|          4|     EV|         5584|     N851AS|
ATL|               AVL|               1255|          1250|             -5|
25|      1315|            48|          62|      34|     164|     1349|
3|             1343|        1352|            9|       0|        0|
null|            null|          null|         null|            null|
null|
|  5|2015|    2| 15|          7|     UA|          712|     N438UA|
IAH|               SFO|               1535|          1554|             19|
18|      1612|           260|         237|     216|    1635|     1748|
3|             1755|        1751|           -4|       0|        0|
null|            null|          null|         null|            null|
null|
|  6|2015|    2| 19|          4|     OO|         5166|     N746SK|
HDN|               DEN|                928|           924|             -4|
11|       935|            67|          56|      29|     141|     1004|
16|             1035|        1020|          -15|       0|        0|
null|            null|          null|         null|            null|
null|
|  7|2015|    2| 27|          5|     DL|         1571|     N916DN|
ATL|               CAK|               2104|          2103|             -1|
20|      2123|           106|          97|      70|     528|     2233|
7|             2250|        2240|          -10|       0|        0|
null|            null|          null|         null|            null|
null|
|  8|2015|    1| 20|          2|     WN|          518|     N405WN|
HOU|               MEM|               2140|          2150|             10|
```

```
8|      2158|            80|          79|      68|     484|     2306|
3|              2300|        2309|            9|       0|        0|
null|            null|          null|         null|               null|
null|
|  9|2015|    2|  6|          5|     WN|          336|     N663SW|
DAL|                 MAF|               1750|          1748|             -2|
7|      1755|            70|          62|      52|     319|     1847|
3|              1900|        1850|          -10|       0|        0|
null|            null|          null|         null|               null|
null|
+---+----+-----+---+----------+------+------------+----------+---------
-----+------------------+------------------+-------------+-------------
-+--------+----------+-------------+-----------+--------+--------+-------
--+-------+-----------------+-----------+-------------+--------+---------+
-------------------+---------------+-------------+-------------+---------
----------+-------------+
```

e) Apply Query performance optimization techniques like – creating Partitioning DataFrame by a specific column, parquet data, caching, predicate pushdown methods etc.

In [71]:
```
flights_delay_df.repartition(3)
```

Out[71]:
```
DataFrame[ID: int, YEAR: int, MONTH: int, DAY: int, DAY_OF_WEEK: int, AIRLI
NE: string, FLIGHT_NUMBER: int, TAIL_NUMBER: string, ORIGIN_AIRPORT: strin
g, DESTINATION_AIRPORT: string, SCHEDULED_DEPARTURE: int, DEPARTURE_TIME: i
nt, DEPARTURE_DELAY: int, TAXI_OUT: int, WHEELS_OFF: int, SCHEDULED_TIME: i
nt, ELAPSED_TIME: int, AIR_TIME: int, DISTANCE: int, WHEELS_ON: int, TAXI_I
N: int, SCHEDULED_ARRIVAL: int, ARRIVAL_TIME: int, ARRIVAL_DELAY: int, DIVE
RTED: int, CANCELLED: int, CANCELLATION_REASON: string, AIR_SYSTEM_DELAY: i
nt, SECURITY_DELAY: int, AIRLINE_DELAY: int, LATE_AIRCRAFT_DELAY: int, WEAT
HER_DELAY: int]
```

In [72]:
```
flights_delay_df.write.parquet("file:///home/hadoop/Downloads/flight/")
```

In [73]:
```
spark.read.parquet("file:///home/hadoop/Downloads/flight/").show()
```

```
+---+----+-----+---+----------+------+------------+----------+---------
-----+------------------+------------------+-------------+-------------
-+--------+----------+-------------+-----------+--------+--------+-------
--+-------+-----------------+-----------+-------------+--------+---------+
-------------------+---------------+-------------+-------------+---------
----------+-------------+
| ID|YEAR|MONTH|DAY|DAY_OF_WEEK|AIRLINE|FLIGHT_NUMBER|TAIL_NUMBER|ORIGIN_AI
RPORT|DESTINATION_AIRPORT|SCHEDULED_DEPARTURE|DEPARTURE_TIME|DEPARTURE_DELA
Y|TAXI_OUT|WHEELS_OFF|SCHEDULED_TIME|ELAPSED_TIME|AIR_TIME|DISTANCE|WHEELS_
ON|TAXI_IN|SCHEDULED_ARRIVAL|ARRIVAL_TIME|ARRIVAL_DELAY|DIVERTED|CANCELLED|
CANCELLATION_REASON|AIR_SYSTEM_DELAY|SECURITY_DELAY|AIRLINE_DELAY|LATE_AIRC
RAFT_DELAY|WEATHER_DELAY|
+---+----+-----+---+----------+------+------------+----------+---------
-----+------------------+------------------+-------------+-------------
-+--------+----------+-------------+-----------+--------+--------+-------
--+-------+-----------------+-----------+-------------+--------+---------+
-------------------+---------------+-------------+-------------+---------
----------+-------------+
|  0|2015|    3|  4|          3|     EV|         5170|     N842AS|
CVG|                 XNA|                935|           954|            19|
16|      1010|           115|         129|     108|     562|     1058|
5|              1030|        1103|           33|       0|        0|
null|              14|             0|           19|                    0|
```

```
0|
|  1|2015|    2|  2|          1|     MQ|         3584|     N646MQ|
DFW|                SPS|                1240|          1316|             36|
11|      1327|            50|          46|      30|     113|     1357|
5|             1330|        1402|           32|       0|        0|
null|               0|             0|           32|                  0|
0|
|  2|2015|    1| 27|          2|     B6|          716|     N309JB|
JAX|                DCA|                1335|          1505|             90|
16|      1521|           104|         110|      91|     634|     1652|
3|             1519|        1655|           96|       0|        0|
null|               6|             0|           90|                  0|
0|
|  3|2015|    1| 28|          3|     EV|         4289|     N14162|
COS|                IAH|                1442|          1435|             -7|
13|      1448|           139|         127|     101|     809|     1729|
13|             1801|        1742|          -19|       0|        0|
null|            null|          null|         null|               null|
null|
|  4|2015|    2|  5|          4|     EV|         5584|     N851AS|
ATL|                AVL|                1255|          1250|             -5|
25|      1315|            48|          62|      34|     164|     1349|
3|             1343|        1352|            9|       0|        0|
null|            null|          null|         null|               null|
null|
|  5|2015|    2| 15|          7|     UA|          712|     N438UA|
IAH|                SFO|                1535|          1554|             19|
18|      1612|           260|         237|     216|    1635|     1748|
3|             1755|        1751|           -4|       0|        0|
null|            null|          null|         null|               null|
null|
|  6|2015|    2| 19|          4|     OO|         5166|     N746SK|
HDN|                DEN|                 928|           924|             -4|
11|       935|            67|          56|      29|     141|     1004|
16|             1035|        1020|          -15|       0|        0|
null|            null|          null|         null|               null|
null|
|  7|2015|    2| 27|          5|     DL|         1571|     N916DN|
ATL|                CAK|                2104|          2103|             -1|
20|      2123|           106|          97|      70|     528|     2233|
7|             2250|        2240|          -10|       0|        0|
null|            null|          null|         null|               null|
null|
|  8|2015|    1| 20|          2|     WN|          518|     N405WN|
HOU|                MEM|                2140|          2150|             10|
8|      2158|            80|          79|      68|     484|     2306|
3|             2300|        2309|            9|       0|        0|
null|            null|          null|         null|               null|
null|
|  9|2015|    2|  6|          5|     WN|          336|     N663SW|
DAL|                MAF|                1750|          1748|             -2|
7|      1755|            70|          62|      52|     319|     1847|
3|             1900|        1850|          -10|       0|        0|
null|            null|          null|         null|               null|
null|
| 10|2015|    3|  6|          5|     UA|          321|     N455UA|
TPA|                EWR|                 950|           947|             -3|
11|       958|           160|         142|     119|     997|     1157|
12|             1230|        1209|          -21|       0|        0|
null|            null|          null|         null|               null|
null|
| 11|2015|    3|  6|          5|     F9|         1343|     N905FR|
TPA|                CLE|                1630|          1622|             -8|
13|      1635|           145|         146|     125|     927|     1840|
```

```
8|             1855|        1848|           -7|       0|        0|
null|            null|        null|         null|            null|
null|
| 12|2015|    1| 30|          5|     WN|         2685|     N638SW|
LAS|                SJC|                620|           633|             13|
13|       646|            90|          76|      59|     386|      745|
4|              750|         749|           -1|       0|        0|
null|            null|        null|         null|            null|
null|
| 13|2015|    1| 11|          7|     HA|          371|     N492HA|
ITO|                HNL|                1930|          2000|             30|
10|      2010|            50|          53|      36|     216|     2046|
7|             2020|        2053|           33|       0|        0|
null|               0|            0|          17|                16|
0|
| 14|2015|    2| 23|          1|     OO|         5416|     N927SW|
ONT|                SFO|                1733|          1727|             -6|
15|      1742|            88|          87|      68|     363|     1850|
4|             1901|        1854|           -7|       0|        0|
null|            null|        null|         null|            null|
null|
| 15|2015|    1| 10|          6|     MQ|         3196|     N607MQ|
SAF|                DFW|                1100|          1051|             -9|
7|      1058|            95|          97|      76|     551|     1314|     1
4|             1335|        1328|           -7|       0|        0|
null|            null|        null|         null|            null|
null|
| 16|2015|    1|  7|          3|     WN|          432|     N961WN|
PDX|                LAS|                1805|          1801|             -4|
7|      1808|           125|         121|     107|     763|     1955|
7|             2010|        2002|           -8|       0|        0|
null|            null|        null|         null|            null|
null|
| 17|2015|    2| 15|          7|     AS|          350|     N557AS|
SEA|                OAK|                2120|          2119|             -1|
25|      2144|           120|         122|      91|     671|     2315|
6|             2320|        2321|            1|       0|        0|
null|            null|        null|         null|            null|
null|
| 18|2015|    2| 24|          2|     OO|         6196|     N751SK|
ONT|                IAH|                701|           806|             65|
24|       830|           189|         180|     145|    1334|     1255|
11|             1210|        1306|           56|       0|        0|
null|               0|            0|          56|                   0|
0|
| 19|2015|    1| 30|          5|     MQ|         3401|     N609MQ|
SHV|                DFW|                1615|          1611|             -4|
8|      1619|            60|          61|      42|     190|     1701|     1
1|             1715|        1712|           -3|       0|        0|
null|            null|        null|         null|            null|
null|
+---+----+-----+---+----------+-------+------------+----------+---------
-----+------------------+------------------+-------------+-------------
-+--------+----------+-------------+-----------+--------+--------+-------
--+-------+-----------------+------------+-------------+--------+---------+
-------------------+---------------+-------------+-------------+---------
----------+-------------+
only showing top 20 rows
```

In [74]:

```
flights_delay_df.cache()
```

Out[74]: DataFrame[ID: int, YEAR: int, MONTH: int, DAY: int, DAY_OF_WEEK: int, AIRLI

```
NE: string, FLIGHT_NUMBER: int, TAIL_NUMBER: string, ORIGIN_AIRPORT: strin
g, DESTINATION_AIRPORT: string, SCHEDULED_DEPARTURE: int, DEPARTURE_TIME: i
nt, DEPARTURE_DELAY: int, TAXI_OUT: int, WHEELS_OFF: int, SCHEDULED_TIME: i
nt, ELAPSED_TIME: int, AIR_TIME: int, DISTANCE: int, WHEELS_ON: int, TAXI_I
N: int, SCHEDULED_ARRIVAL: int, ARRIVAL_TIME: int, ARRIVAL_DELAY: int, DIVE
RTED: int, CANCELLED: int, CANCELLATION_REASON: string, AIR_SYSTEM_DELAY: i
nt, SECURITY_DELAY: int, AIRLINE_DELAY: int, LATE_AIRCRAFT_DELAY: int, WEAT
```

In [75]:

```
#persistance of dataframe with a specific storage level
from pyspark import StorageLevel
flights_delay_df.persist(StorageLevel.MEMORY_AND_DISK)
```

Out[75]:

```
DataFrame[ID: int, YEAR: int, MONTH: int, DAY: int, DAY_OF_WEEK: int, AIRLI
NE: string, FLIGHT_NUMBER: int, TAIL_NUMBER: string, ORIGIN_AIRPORT: strin
g, DESTINATION_AIRPORT: string, SCHEDULED_DEPARTURE: int, DEPARTURE_TIME: i
nt, DEPARTURE_DELAY: int, TAXI_OUT: int, WHEELS_OFF: int, SCHEDULED_TIME: i
nt, ELAPSED_TIME: int, AIR_TIME: int, DISTANCE: int, WHEELS_ON: int, TAXI_I
N: int, SCHEDULED_ARRIVAL: int, ARRIVAL_TIME: int, ARRIVAL_DELAY: int, DIVE
RTED: int, CANCELLED: int, CANCELLATION_REASON: string, AIR_SYSTEM_DELAY: i
nt, SECURITY_DELAY: int, AIRLINE_DELAY: int, LATE_AIRCRAFT_DELAY: int, WEAT
HER_DELAY: int]
```

Write Spark SQL queries to show following analysis with Visualization on Databricks Community Edition.

f) Average arrival delay caused by airlines

In [76]:

```
spark.sql("""
    select
        airline,
        round(avg(CASE when arrival_delay> 0 then arrival_delay else NULL e
        from flights_table
        group by airline
        order by Average_arrival_delay desc
""").show()
```

```
+-------+---------------------+
|airline|Average_arrival_delay|
+-------+---------------------+
|     F9|                47.37|
|     B6|                42.78|
|     MQ|                42.57|
|     NK|                39.85|
|     OO|                37.54|
|     DL|                36.48|
|     AA|                36.29|
|     EV|                36.21|
|     VX|                35.14|
|     UA|                34.13|
|     US|                29.41|
|     WN|                27.64|
|     AS|                24.83|
|     HA|                16.05|
+-------+---------------------+
```

g) Days of months with respected to average of arrival delays

In [77]:
```
spark.sql("""
    select
        day,
        round(avg(CASE when arrival_delay> 0 then arrival_delay else NULL
        from flights_table
        group by day
        order by day
""").show()
```

```
+---+--------------------+
|day|Average_arrival_delay|
+---+--------------------+
|  1|               43.55|
|  2|               37.13|
|  3|               41.83|
|  4|               42.05|
|  5|               42.09|
|  6|               39.52|
|  7|               31.25|
|  8|               34.96|
|  9|               31.23|
| 10|               24.73|
| 11|               33.65|
| 12|               36.25|
| 13|               28.53|
| 14|               27.39|
| 15|               26.97|
| 16|               33.15|
| 17|               37.84|
| 18|               27.11|
| 19|                24.6|
| 20|               29.92|
+---+--------------------+
only showing top 20 rows
```

h) Arrange weekdays with respect to the average arrival delays caused

In [78]:
```
spark.sql("""
    select
        day_of_week ,
        round(avg(CASE when arrival_delay> 0 then arrival_delay else NULL
        from flights_table
        group by day_of_week
        order by Average_arrival_delay desc
""").show()
```

```
+-----------+--------------------+
|day_of_week|Average_arrival_delay|
+-----------+--------------------+
|          7|               38.42|
|          2|               36.64|
|          1|               36.38|
|          6|               34.05|
|          4|               32.85|
|          3|               32.78|
|          5|               31.19|
+-----------+--------------------+
```

i) Arrange Days of month as per cancellations done in Descending

In [79]:

```
# flights_delay_df.printSchema()
spark.sql("""
    select
        day,
        sum(case when cancelled = 1 then 1 else 0 end) Cancellations
        from flights_table
        group by day
        order by cancellations desc
""").show()
```

```
+---+-------------+
|day|Cancellations|
+---+-------------+
|  1|          237|
|  5|          215|
|  2|          195|
| 27|          185|
| 26|          114|
|  4|          113|
| 28|           98|
|  9|           89|
|  3|           88|
| 15|           83|
| 23|           69|
| 16|           63|
| 25|           61|
| 21|           61|
|  8|           61|
| 17|           59|
| 24|           57|
|  6|           53|
| 22|           41|
|  7|           31|
+---+-------------+
only showing top 20 rows
```

j) Find Top 10 busiest airports with respect to day of week

In [106…

```
spark.sql("""
    with flightCount AS(
    select
        day_of_week,
        origin_airport airport,
        count(*) total_count
    from flights_table
    group by day_of_week, origin_airport
    UNION ALL
    select
        day_of_week,
        destination_airport airport,
        count(*) total_count
    from flights_table
    group by day_of_week, destination_airport
    ),
    totalAirports AS(
    select
        day_of_week,
        airport,
        sum(total_count) totalFlights
    from flightCount
    group by airport, day_of_week
    ),
    airportRanks AS(
    select
        day_of_week,
        airport,
        totalFlights,
    RANK() OVER( PARTITION BY day_of_week order by totalFlights desc) AS f
    from totalAirports
    )
    select
        day_of_week,
        airport,
        flightRank,
        totalFlights
        from airportRanks
        where flightRank<=10
        --order by day_of_week

""").show()
```

```
+-----------+-------+----------+------------+
|day_of_week|airport|flightRank|totalFlights|
+-----------+-------+----------+------------+
|          1|    ATL|         1|        1106|
|          1|    ORD|         2|         844|
|          1|    DFW|         3|         818|
|          1|    LAX|         4|         631|
|          1|    DEN|         5|         613|
|          1|    IAH|         6|         494|
|          1|    PHX|         7|         485|
|          1|    SFO|         8|         466|
|          1|    LAS|         9|         398|
|          1|    MSP|        10|         382|
|          6|    ATL|         1|         817|
|          6|    DFW|         2|         722|
|          6|    ORD|         3|         711|
|          6|    DEN|         4|         514|
|          6|    LAX|         5|         509|
|          6|    PHX|         6|         427|
```

```
|          6|    SFO|         7|         395|
|          6|    LAS|         8|         373|
|          6|    MCO|         9|         361|
|          6|    IAH|        10|         350|
+-----------+-------+----------+------------+
only showing top 20 rows
```

k) Finding airlines that make the maximum number of cancellations

In [81]:

```
spark.sql("""
    select
        airline,
        count(*) total_count
        from flights_table
        where cancelled = 1
        group by airline
        order by total_count desc
""").show()
```

```
+-------+-----------+
|airline|total_count|
+-------+-----------+
|     MQ|        414|
|     WN|        358|
|     EV|        312|
|     AA|        241|
|     DL|        177|
|     US|        169|
|     OO|        153|
|     B6|        145|
|     UA|        122|
|     NK|         21|
|     VX|         13|
|     AS|         12|
|     F9|         11|
|     HA|          3|
+-------+-----------+
```

l) Find and order airlines in descending that make the most number of diversions

In [82]:

```
spark.sql("""
    select
        airline,
        count(*) total_count
        from flights_table
        where diverted = 1
        group by airline
        order by total_count desc
""").show()
```

```
+-------+-----------+
|airline|total_count|
+-------+-----------+
|     WN|         35|
|     OO|         25|
|     EV|         22|
|     DL|         18|
|     B6|         16|
|     AA|         12|
```

```
|     US|          9|
|     UA|          8|
|     MQ|          5|
|     HA|          1|
+-------+-----------+
```

m) Finding days of month that see the most number of diversion

In [83]:

```
spark.sql("""
    select
        day,
        sum(diverted) as Total_diversion
        from flights_table
        group by day
        order by Total_diversion desc
""").show()
```

```
+---+---------------+
|day|Total_diversion|
+---+---------------+
|  2|             15|
|  1|             13|
|  4|             12|
|  5|             11|
|  9|              9|
| 14|              8|
|  6|              7|
|  7|              6|
| 23|              6|
| 11|              5|
| 30|              5|
|  3|              5|
|  8|              5|
| 18|              5|
| 28|              4|
| 16|              4|
| 12|              4|
| 20|              4|
| 21|              4|
| 17|              3|
+---+---------------+
only showing top 20 rows
```

In [84]:

```
spark.sql("""
select month, day , count(*) total_count from flights_table
    where diverted = 1
    group by  month, day order by count(*) desc
""").show()
```

```
+-----+---+-----------+
|month|day|total_count|
+-----+---+-----------+
|    2|  2|          9|
|    3|  1|          7|
|    2| 14|          7|
|    3|  5|          7|
|    3|  2|          6|
|    3|  4|          6|
|    1| 30|          5|
|    2| 23|          5|
|    1|  7|          5|
```

```
|    2|  1|          5|
|    1| 11|          5|
|    1| 18|          5|
|    1|  8|          4|
|    2|  9|          4|
|    2| 21|          4|
|    2|  5|          3|
|    2| 26|          3|
|    2| 28|          3|
|    2|  4|          3|
|    1|  9|          3|
+-----+---+-----------+
only showing top 20 rows
```

n) Calculating mean and standard deviation of departure delay for all flights in minutes

In [85]:

```
spark.sql("""
    select
        round(mean(departure_delay),3) Mean_departure_delay,
        round(stddev(departure_delay),3) STD_Dev_departure_delay
        from flights_table
""").show()
```

```
+--------------------+-----------------------+
|Mean_departure_delay|STD_Dev_departure_delay|
+--------------------+-----------------------+
|              11.329|                 39.621|
+--------------------+-----------------------+
```

o) Calculating mean and standard deviation of arrival delay for all flights in minutes

In [86]:

```
spark.sql("""
    select
        round(mean(arrival_delay),3) Mean_arrival_delay,
        round(stddev(arrival_delay),3) STD_Dev_arrival_delay
        from flights_table
""").show()
```

```
+------------------+---------------------+
|Mean_arrival_delay|STD_Dev_arrival_delay|
+------------------+---------------------+
|             7.545|               42.378|
+------------------+---------------------+
```

p) Finding all diverted Route from a source to destination Airport & which route is the most diverted

In [87]:
```
spark.sql("""
    select
        ORIGIN_AIRPORT,
        DESTINATION_AIRPORT,
        COUNT(*) as diverted_count
        from flights_table
        where DIVERTED = 1
        group by ORIGIN_AIRPORT, DESTINATION_AIRPORT
        ORDER BY diverted_count DESC
""").show()
```

```
+--------------+-------------------+--------------+
|ORIGIN_AIRPORT|DESTINATION_AIRPORT|diverted_count|
+--------------+-------------------+--------------+
|           HOU|                DAL|             2|
|           PHL|                SAN|             2|
|           STT|                PHL|             2|
|           TPA|                LGA|             2|
|           IAH|                ASE|             2|
|           JFK|                EGE|             2|
|           JFK|                SEA|             2|
|           ORD|                ASE|             2|
|           CLT|                IAH|             2|
|           KOA|                SFO|             1|
|           SNA|                SFO|             1|
|           FLL|                PVD|             1|
|           ATL|                LGA|             1|
|           FLL|                BWI|             1|
|           BOS|                LAS|             1|
|           ASE|                LAX|             1|
|           IAH|                ISN|             1|
|           LAX|                ASE|             1|
|           ATL|                GTR|             1|
|           MCO|                BWI|             1|
+--------------+-------------------+--------------+
only showing top 20 rows
```

q) Finding AIRLINES with its total flight count, total number of flights arrival delayed by more than 30 Minutes, % of such flights delayed by more than 30 minutes when it is not Weekends with minimum count of flights from Airlines by more than 10. Also Exclude some of Airlines 'AK', 'HI', 'PR', 'VI' and arrange output in descending order by % of such count of flights.

In [88]:
```
spark.sql("""
    SELECT
        airline,
        count(*) total_flights,
        SUM(CASE WHEN arrival_delay > 30 THEN 1 ELSE 0 END) AS flight_delay
        SUM(CASE WHEN arrival_delay > 30 AND day_of_week NOT IN (6, 7) THEN
        round((SUM(CASE WHEN arrival_delay > 30 AND day_of_week NOT IN (6,
        FROM flights_table WHERE airline NOT IN ('AK', 'HI', 'PR', 'VI')
        GROUP BY airline
        HAVING total_flights > 10
        ORDER BY percentage_delay DESC
""").show()
```

```
+-------+-------------+-------------+--------------------------------+----
------------+
|airline|total_flights|flight_delays|delayed_flight_excluding_weekends|Perc
entage_delay|
```

```
+-------+-------------+-------------+---------------------------------+----
------------+
|     F9|          794|          198|                              139|
17.51|
|     MQ|         3502|          775|                              601|
17.16|
|     B6|         2548|          485|                              360|
14.13|
|     NK|         1048|          186|                              139|
13.26|
|     EV|         5916|          874|                              665|
11.24|
|     OO|         5708|          859|                              633|
11.09|
|     UA|         4701|          653|                              497|
10.57|
|     AA|         5250|          700|                              484|
9.22|
|     VX|          573|           67|                               47|
8.2|
|     US|         3925|          452|                              310|
7.9|
|     DL|         7989|          746|                              592|
7.41|
|     WN|        11738|         1235|                              869|
7.4|
|     AS|         1586|          100|                               64|
4.04|
|     HA|          722|           38|                               23|
3.19|
+-------+-------------+-------------+---------------------------------+----
------------+
```

r) Finding AIRLINES with its total flight count with total number of flights departure delayed by less than 30 Minutes, % of such flights delayed by less than 30 minutes when it is Weekends with minimum count of flights from Airlines by more than 10. Also Exclude some of Airlines 'AK', 'HI', 'PR', 'VI' and arrange output in descending order by % of such count of flights.

In [89]:

```
spark.sql("""
select
    airline,
    count(*) total_flights,
    SUM(CASE WHEN departure_delay < 30 THEN 1 ELSE 0 END) AS low_departure_
    SUM(CASE WHEN departure_delay < 30 AND day_of_week IN (6, 7) THEN 1 ELS
    round((SUM(CASE WHEN departure_delay < 30 AND DAY_OF_WEEK IN (6, 7) THE
    FROM flights_table
    WHERE airline NOT IN ('AK', 'HI', 'PR', 'VI')
    GROUP BY airline
    HAVING total_flights > 10
    ORDER BY Percentage_departure_delay DESC""").show()
```

```
+-------+-------------+------------------+------------------------------
-------+-------------------------+
|airline|total_flights|low_departure_delay|weekend_flight_with_low_departur
e_delay|Percentage_departure_delay|
+-------+-------------+------------------+------------------------------
-------+-------------------------+
|     AS|         1586|              1468|
412|                    25.98|
|     HA|          722|               692|
179|                    24.79|
```

```
|     NK|         1048|                839|
253|                      24.14|
|     AA|         5250|               4342|
1214|                      23.12|
|     DL|         7989|               7010|
1814|                      22.71|
|     VX|          573|                490|
129|                      22.51|
|     WN|        11738|               9945|
2636|                      22.46|
|     US|         3925|               3356|
867|                      22.09|
|     OO|         5708|               4736|
1244|                      21.79|
|     B6|         2548|               1947|
543|                      21.31|
|     EV|         5916|               4819|
1203|                      20.33|
|     UA|         4701|               3903|
950|                      20.21|
|     MQ|         3502|               2443|
622|                      17.76|
|     F9|          794|                585|
133|                      16.75|
+-------+-------------+-------------------+------------------------------
-------+--------------------------+
```

s) When is the best time of day/day of week/time of a year to fly with minimum delays?

In [90]:

```
from pyspark.sql.functions import *

# corresponding hours

new_flight_df = flights_delay_df.withColumn("SCHEDULED_DEPARTURE_HR",\
(flights_delay_df["SCHEDULED_DEPARTURE"]/100 ).cast("int"))\
.withColumn("SCHEDULED_ARRIVAL_HR",(flights_delay_df["SCHEDULED_ARRIVAL"]/

# average departure and arrival delay by hour of the day

new_flight_df.groupBy("SCHEDULED_DEPARTURE_HR")\
.agg(avg(when(col("DEPARTURE_DELAY") > 0, col("DEPARTURE_DELAY"))).alias("
avg(when(col("ARRIVAL_DELAY") > 0,col("ARRIVAL_DELAY"))).alias("avg_arr_de
).orderBy("avg_dep_delay","avg_arr_delay").show()
```

```
+---------------------+------------------+------------------+
|SCHEDULED_DEPARTURE_HR|     avg_dep_delay|     avg_arr_delay|
+---------------------+------------------+------------------+
|                    2|             13.75|              11.0|
|                    4|20.333333333333332|              15.0|
|                    0|              27.8|23.606060606060606|
|                   22| 30.91695501730104|31.971119133574007|
|                    5| 30.97902097902098|30.914141414141415|
|                   11|31.459610027855152| 33.02514367816092|
|                   13| 31.47463768115942| 32.56007509386733|
|                    7|31.937923250564335| 29.16142735768904|
|                   10|  32.0852314474651|32.733853797019165|
|                   16|            32.515|34.525513585155736|
|                    9|33.173761946133794|32.386656557998485|
|                   23| 33.55797101449275| 30.81203007518797|
|                   14|33.801867911941294| 35.57083042568039|
|                    3|              34.0|              51.4|
```

```
|                    12| 34.24120234604106| 34.89728096676737|
|                     8| 34.25353283458021| 33.96315028901734|
|                    15| 34.75278810408922| 37.62823061630219|
|                    17| 35.29085140137494| 38.11011235955056|
|                    19| 35.48476992871031|  36.2907133243607|
|                    21|35.525916561314794| 36.87735849056604|
+--------------------+------------------+------------------+
only showing top 20 rows
```

t) Which airlines are best airline to travel considering number of cancellations, arrival, departure delays and all reasons affecting performance of airline industry.

In [91]:

```
spark.sql("""
SELECT
    airline,
    count(*) total_flights,
    SUM(cancelled) total_cancellation,
    round(AVG(CASE WHEN departure_delay > 0 THEN departure_delay ELSE NULL
    round(AVG(CASE WHEN arrival_delay > 0 THEN arrival_delay ELSE NULL END
    round(AVG(CASE WHEN air_system_delay > 0 THEN air_system_delay ELSE NUl
    round(AVG(CASE WHEN security_delay > 0 THEN security_delay ELSE NULL El
    rounD(AVG(CASE WHEN airline_delay > 0 THEN airline_delay ELSE NULL END
    round(AVG(CASE WHEN late_aircraft_delay > 0 THEN late_aircraft_delay El
    round(AVG(CASE WHEN weather_delay > 0 THEN weathEr_delay ELSE NULL END
    FROM flights_table
    GROUP BY airline
    order by average_departure_delay, average_arrival_delay , average_air_s
        average_airline_delay, average_late_aircraft_delay, average_weathe
    limit 5
""").show()
```

```
+-------+-------------+-----------------+----------------------+---------
------------+------------------------+----------------------+-------------
-------+---------------------------+--------------------+
|airline|total_flights|total_cancellation|average_departure_delay|average_a
rrival_delay|average_air_system_delay|average_security_delay|average_airlin
e_delay|average_late_aircraft_delay|average_weather_delay|
+-------+-------------+-----------------+----------------------+---------
------------+------------------------+----------------------+-------------
-------+---------------------------+--------------------+
|     HA|          722|                3|                 18.34|
16.05|                    10.0|                null|                28.6
9|                   24.61|               15.86|
|     WN|        11738|              358|                 25.84|
27.64|                    16.54|                44.5|                22.8
6|                   33.78|               64.75|
|     UA|         4701|              122|                 28.14|
34.13|                    25.91|                null|                 30.
9|                   44.02|               38.42|
|     AS|         1586|               12|                  28.5|
24.83|                    20.57|                27.5|                36.1
9|                   57.73|               84.73|
|     US|         3925|              169|                 30.17|
29.41|                    24.36|                28.89|                31.6
1|                   39.85|               29.91|
+-------+-------------+-----------------+----------------------+---------
------------+------------------------+----------------------+-------------
-------+---------------------------+--------------------+
```

In [ ]: