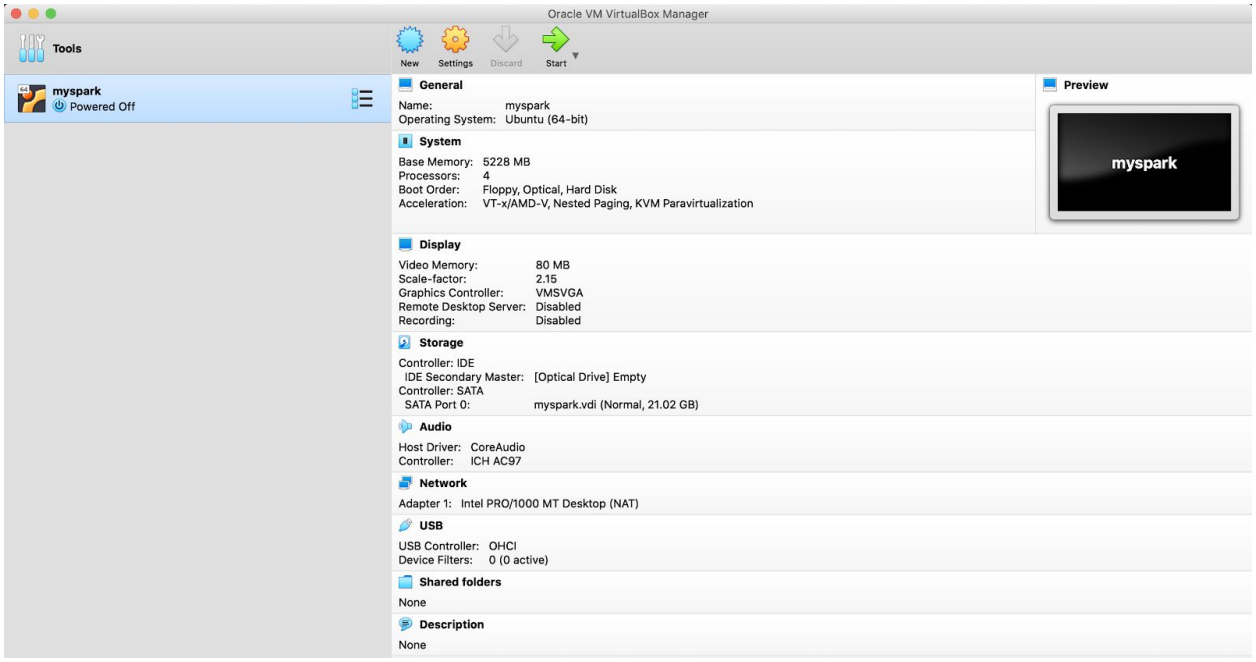
System Setup



Turn on Ubuntu and use Jupyter Notebook

- To open the Jupyter Notebook

  In home directory:

```
export SPARK_HOME='home/ubuntu/spark-3.0.0-bin-hadoop2.7'
export PATH=$SPARK_HOME:$PATH
export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
export PYSPARK_PYTHON=python3
```

- Go to the python folder

```
cd spark-3.0.0-bin-hadoop2.7/python
```

- Run Jupyter Notebook

```
Jupyter Notebook
```

Open jupyter notebook

- Create the notebook, and create spark session

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('traffic').getOrCreate()
df = spark.read.csv('2017-06.csv', inferSchema=True, header=True)
df.show()
df.printSchema()
```

- 2017-06 dataset

```
In [1]: from pyspark.sql import SparkSession

In [2]: spark = SparkSession.builder.appName('traffic').getOrCreate()

In [3]: df = spark.read.csv('2017-06.csv', inferSchema=True, header=True)

In [4]: df.show()
+-------------------+----------+-----------+-----+------+-------
--+------------+-----+----------+-----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime|  vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+------------+-----+----------+-----+------+
|2017-06-01 00:00:00|  01:00:00|     100643|50.92|    49|     0.
65|         180|  0.1|      0.68| 0.56| 28.42|
|2017-06-01 00:00:00|  01:00:00|     100874|62.76|   226|     1.
19|         162|-0.05|       1.1| 4.16|261.03|
```

```
In [7]: df.agg({'volume': 'avg'}).show()
+-----------------+
|      avg(volume)|
+-----------------+
|745.1610920526015|
+-----------------+
```

```
In [6]: df.agg({'speed':'avg'}).show()
+------------------+
|        avg(speed)|
+------------------+
|57.085596295234986|
+------------------+
```

```
In [8]: df.agg({'volume':'sum'}).show()
+-----------+
|sum(volume)|
+-----------+
|   15639441|
+-----------+
```

- Deal with null value

  from pyspark.sql.functions import mean

```
In [9]: from pyspark.sql.functions import mean

In [10]: mean_val = df.select(mean(df['speed'])).collect()

In [11]: mean_val

Out[11]: [Row(avg(speed)=57.085596295234986)]

In [12]: mean_speed = mean_val[0][0]

In [13]: mean_speed

Out[13]: 57.085596295234986


In [13]: mean_speed

Out[13]: 57.085596295234986

In [15]: new_table = df.na.fill(mean_speed, ['speed'])

In [20]: new_table.agg({'speed': 'avg'}).show()
+------------------+
|        avg(speed)|
+------------------+
|57.085596295235014|
+------------------+
```

```
In [21]: mean_val = df.select(mean(df['volume'])).collect()

In [22]: mean_val
Out[22]: [Row(avg(volume)=745.1610920526015)]

In [25]: mean_volume = mean_val[0][0]

In [26]: mean_volume
Out[26]: 745.1610920526015

In [26]: mean_volume
Out[26]: 745.1610920526015

In [27]: new_table = df.na.fill(mean_volume, ['volume'])

In [28]: new_table.agg({'volume': 'avg'}).show()
+-----------------+
|      avg(volume)|
+-----------------+
|745.1610920526015|
+-----------------+

In [29]: new_table.agg({'volume': 'sum'}).show()
+-----------+
|sum(volume)|
+-----------+
|   15639441|
+-----------+
```

- 2018-06 dataset

In [2]:
```
from pyspark.sql import SparkSession
```

In [3]:
```
spark = SparkSession.builder.appName('traffic').getOrCreate()
```

In [4]:
```
df = spark.read.csv('2018-06.csv', inferSchema=True, header=True)
```

In [5]:
```
df.show()
```

```
+-------------------+---------+----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+---------+----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2018-06-01 00:00:00|  01:00:00|    100871|65.49|   217|     1.
29|          157|-0.11|      1.17|4.22|276.68|
```

In [7]:
```
df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|56.61071740047854|
+-----------------+
```

In [8]:
```
df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|681.2837812529264|
+-----------------+
```

In [9]:
```
df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   14550859|
+-----------+
```

- 2019-06 dataset

In [2]:
```
from pyspark.sql import SparkSession
```

In [3]:
```
spark = SparkSession.builder.appName('traffic').getOrCreate()
```

In [18]:
```
df = spark.read.csv('2019-06.csv', inferSchema=True, header=True)
```

In [19]:
```
df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-06-01 00:00:00|  01:00:00|     100648|58.11|   473|     3.
09|          177| 0.02|      0.66|5.21|302.72|
```

In [21]:
```
df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|58.27126614707345|
+-----------------+
```

In [22]:
```
df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|796.5053557903848|
+-----------------+
```

In [23]:
```
df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   22084704|
+-----------+
```

- 2020-06 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [24]: df = spark.read.csv('2020-06.csv', inferSchema=True, header=True)
```

```
In [25]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+-----+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|  vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+-----+
|2020-06-01 00:00:00|  01:00:00|     100681|54.73|    80|     0.
84|           98|  0.1|       1.1|1.47| 80.4|
```

```
In [29]: df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|61.36848311845442|
+-----------------+
```

```
In [30]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|522.8172172928073|
+-----------------+
```

```
In [31]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   13846814|
+-----------+
```

- 2019-01 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [32]: df = spark.read.csv('2019-01.csv', inferSchema=True, header=True)
```

```
In [33]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+-----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime|  vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+-----+------+
|2019-01-01 00:00:00|  01:00:00|     100643|59.33|   147|     0.
73|          179|  0.0|      0.34| 0.84| 49.98|
```

```
In [35]: df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|58.65557559506348|
+-----------------+
```

```
In [36]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|716.6041903643636|
+-----------------+
```

```
In [37]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   19529614|
+-----------+
```

- 2019-02 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [38]: df = spark.read.csv('2019-02.csv', inferSchema=True, header=True)
```

```
In [39]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-02-01 00:00:00|  01:00:00|     102184|70.01|    94|     0.
63|          150|-0.06|      0.36|0.57| 39.95|
```

```
In [41]: df.agg({'speed':'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|57.612613392473925|
+------------------+
```

```
In [42]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|705.7003395889187|
+-----------------+
```

```
In [43]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   19741967|
+-----------+
```

- 2019-03 dataset

In [3]:
```
spark = SparkSession.builder.appName('traffic').getOrCreate()
```

In [44]:
```
df = spark.read.csv('2019-03.csv', inferSchema=True, header=True)
```

In [45]:
```
df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-03-01 00:00:00|  01:00:00|     101039|59.18|    85|     0.
59|          144| 0.02|      1.23|1.75|103.28|
```

In [47]:
```
df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|58.20532706972573|
+-----------------+
```

In [48]:
```
df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|810.3327544621321|
+-----------------+
```

In [49]:
```
df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   25197297|
+-----------+
```

- 2019-04 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [50]: df = spark.read.csv('2019-04.csv', inferSchema=True, header=True)
```

```
In [51]: df.show()
```

```
+-------------------+---------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+---------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-04-01 00:00:00|  01:00:00|     102191|69.16|    83|     0.
33|          180|-0.11|      0.69|0.96|  66.4|
```

```
In [53]: df.agg({'speed':'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|58.230850134074785|
+------------------+
```

```
In [55]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|779.4638827305516|
+-----------------+
```

```
In [56]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   23556178|
+-----------+
```

- 2019-05 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [57]: df = spark.read.csv('2019-05.csv', inferSchema=True, header=True)
```

```
In [58]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-05-01 00:00:00|  01:00:00|     102152|54.77|   241|     2.
45|          120| 0.07|      0.76|3.04|166.29|
```

```
In [60]: df.agg({'speed':'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|58.497939953958685|
+------------------+
```

```
In [61]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|778.1844703626778|
+-----------------+
```

```
In [62]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   24353283|
+-----------+
```

- 2019-06 dataset
  Done previously

- 2019-07 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [63]: df = spark.read.csv('2019-07.csv', inferSchema=True, header=True)
```

```
In [64]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-07-01 00:00:00|  01:00:00|     100867|59.88|    82|     0.
74|          126|  0.0|      1.59|2.17|129.97|
```

```
In [66]: df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|58.10923116598405|
+-----------------+
```

```
In [67]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|801.8246947782109|
+-----------------+
```

```
In [68]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   24891044|
+-----------+
```

- 2019-08 dataset

In [3]:
```
spark = SparkSession.builder.appName('traffic').getOrCreate()
```

In [69]:
```
df = spark.read.csv('2019-08.csv', inferSchema=True, header=True)
```

In [70]:
```
df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-08-01 00:00:00|  01:00:00|     102162| 60.6|    97|     0.
95|          180|  0.0|       0.5| 0.8|  48.5|
```

In [72]:
```
df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.47981326699833|
+-----------------+
```

In [73]:
```
df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|823.8053502171247|
+-----------------+
```

In [74]:
```
df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   24851736|
+-----------+
```

- 2019-09 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [75]: df = spark.read.csv('2019-09.csv', inferSchema=True, header=True)
```

```
In [76]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-09-01 00:00:00|  01:00:00|     102152|54.26|   446|     3.
27|          179| 0.07|      0.76|5.67|307.74|
```

```
In [78]: df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.52269430572795|
+-----------------+
```

```
In [79]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|768.8318761933474|
+-----------------+
```

```
In [80]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   22951938|
+-----------+
```

- 2010-10 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()

In [81]: df = spark.read.csv('2019-10.csv', inferSchema=True, header=True)

In [82]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-10-01 00:00:00|  01:00:00|     100870|63.13|   243|     1.
42|          180|-0.06|      1.21|4.91|309.83|
```

```
In [84]: df.agg({'speed':'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|57.489340013936236|
+------------------+
```

```
In [85]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|764.3207096667219|
+-----------------+
```

```
In [86]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   23048091|
+-----------+
```

- 2019-11 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()
```

```
In [87]: df = spark.read.csv('2019-11.csv', inferSchema=True, header=True)
```

```
In [88]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-11-01 00:00:00|  01:00:00|     101040|64.44|    18|     0.
12|          138|-0.08|      1.13|0.34| 21.87|
```

```
In [90]: df.agg({'speed':'avg'}).show()
```

```
+----------------+
|      avg(speed)|
+----------------+
|57.8025591794723|
+----------------+
```

```
In [91]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|782.7144022703086|
+-----------------+
```

```
In [92]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   22064719|
+-----------+
```

- 2019-12 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()

In [93]: df = spark.read.csv('2019-12.csv', inferSchema=True, header=True)

In [94]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+------------+-----+----------+----+------+
|2019-12-01 00:00:00|  01:00:00|     100646| 58.7|   197|
0.9|          174| 0.01|      0.35|1.14| 66.98|
```

```
In [96]: df.agg({'speed':'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.72480320773076|
+-----------------+
```

```
In [97]: df.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|774.4716671137339|
+-----------------+
```

```
In [98]: df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   23097843|
+-----------+
```

- Week-2020 dataset

```
In [3]: spark = SparkSession.builder.appName('traffic').getOrCreate()

In [99]: df = spark.read.csv('week-2020.csv', inferSchema=True, header=Tru

In [100]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2020-06-01 00:00:00|  01:00:00|     101041|59.35|    79|     0.
84|          134| 0.02|      1.61|2.12|126.01|
```

- 0601

```
In [107]: table0601 = df.filter(df['starttime'].contains('2020-06-01'))

In [110]: table0601.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|61.32254341164454|
+-----------------+
```

```
In [111]: table0601.agg({'volume': 'avg'}).show()
```

```
+------------------+
|       avg(volume)|
+------------------+
|476.75714285714287|
+------------------+
```

```
In [112]:  1 table0601.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     467222|
+-----------+
```

- 0602

In [113]:
```
table0602 = df.filter(df['starttime'].contains('2020-06-02'))
```

In [114]:
```
table0602.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|61.09919202518366|
+-----------------+
```

In [115]:
```
table0602.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|478.4230366492147|
+-----------------+
```

In [116]:
```
1 table0602.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     456894|
+-----------+
```

- 0603

In [117]:
```
table0603 = df.filter(df['starttime'].contains('2020-06-03'))
```

In [118]:
```
table0603.agg({'speed': 'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|61.386080508474556|
+------------------+
```

In [119]:
```
table0603.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|506.0010582010582|
+-----------------+
```

In [120]:
```
1 table0603.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     478171|
+-----------+
```

- 0604

In [121]:
```
table0604 = df.filter(df['starttime'].contains('2020-06-04'))
```

In [122]:
```
table0604.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|61.20460824742269|
+-----------------+
```

In [123]:
```
table0604.agg({'volume': 'avg'}).show()
```

```
+------------------+
|       avg(volume)|
+------------------+
|506.13917525773195|
+------------------+
```

In [124]:
```
1 table0604.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     490955|
+-----------+
```

- 0605

In [125]:
```
table0605 = df.filter(df['starttime'].contains('2020-06-05'))
```

In [126]:
```
table0605.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|60.95950668036996|
+-----------------+
```

In [127]:
```
table0605.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|562.4640657084188|
+-----------------+
```

In [128]:
```
1 table0605.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     547840|
+-----------+
```

- 0606

In [129]:
```
table0606 = df.filter(df['starttime'].contains('2020-06-06'))
```

In [130]:
```
table0606.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|62.47982817869418|
+-----------------+
```

In [131]:
```
table0606.agg({'volume': 'avg'}).show()
```

```
+------------------+
|       avg(volume)|
+------------------+
|290.32114285714283|
+------------------+
```

In [132]:
```
1 table0606.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     254031|
+-----------+
```

- 0607

In [176]:
```
table0607 = df.filter(df['starttime'].contains('2020-06-07'))
```

In [177]:
```
table0607.agg({'speed': 'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|63.262249999999995|
+------------------+
```

In [178]:
```
table0607.agg({'volume': 'avg'}).show()
```

```
+------------------+
|       avg(volume)|
+------------------+
|252.97505422993493|
+------------------+
```

```
In [179]: 1 table0607.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     233243|
+-----------+
```

- Week-2019 dataset

```
In [137]: df = spark.read.csv('week-2019.csv', inferSchema=True, header=Tru
```

```
In [138]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2019-06-03 00:00:00|  01:00:00|     100870|61.45|   250|     1.
42|          172|-0.03|      1.24|5.19|318.75|
```

- 20190603

In [140]:
```
table190603 = df.filter(df['starttime'].contains('2019-06-03'))
```

In [141]:
```
table190603.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.79483380816711|
+-----------------+
```

In [142]:
```
table190603.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|843.0132701421801|
+-----------------+
```

In [143]:
```
1 table190603.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     889379|
+-----------+
```

- 20190604

In [144]:
```
table190604 = df.filter(df['starttime'].contains('2019-06-04'))
```

In [145]:
```
table190604.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|56.94217224880378|
+-----------------+
```

In [146]:
```
table190604.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|866.5397129186603|
+-----------------+
```

In [147]:
```
1 table190604.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     905534|
+-----------+
```

- 20190605

In [148]:
```
table190605 = df.filter(df['starttime'].contains('2019-06-05'))
```

In [149]:
```
table190605.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.46585482330473|
+-----------------+
```

In [150]:
```
table190605.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|889.1982840800763|
+-----------------+
```

In [151]:
```
1 table190605.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     932769|
+-----------+
```

- 20190606

In [152]:
```
table190606 = df.filter(df['starttime'].contains('2019-06-06'))
```

In [153]:
```
table190606.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.14722592945662|
+-----------------+
```

In [154]:
```
table190606.agg({'volume': 'avg'}).show()
```

```
+----------------+
|     avg(volume)|
+----------------+
|722.161105815062|
+----------------+
```

In [155]:
```
1 table190606.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     757547|
+-----------+
```

- 20190607

In [156]:
```
table190607 = df.filter(df['starttime'].contains('2019-06-07'))
```

In [157]:
```
table190607.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.39788571428571|
+-----------------+
```

In [158]:
```
table190607.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|750.5271170313987|
+-----------------+
```

In [159]:
```
1 table190607.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     788804|
+-----------+
```

- 20190608

In [160]:
```
table190608 = df.filter(df['starttime'].contains('2019-06-08'))
```

In [161]:
```
table190608.agg({'speed': 'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|60.114465832531316|
+------------------+
```

In [162]:
```
table190608.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|680.6807692307692|
+-----------------+
```

In [163]:
```
1 table190608.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     707908|
+-----------+
```

- 20190609

In [170]:
```
table190609 = df.filter(df['starttime'].contains('2019-06-09'))
```

In [171]:
```
table190609.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|60.66832038834943|
+-----------------+
```

In [172]:
```
table190609.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|603.1485436893204|
+-----------------+
```

```
In [173]: 1 table190609.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     621243|
+-----------+
```

- Week-2018 dataset

```
In [180]: df = spark.read.csv('week-2018.csv', inferSchema=True, header=Tru
```

```
In [181]: df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2018-06-04 00:00:00|  01:00:00|     100645|61.09|   188|     1.
11|          131|-0.01|      0.57|1.78|109.04|
```

- 20180604

```
In [183]: table180604 = df.filter(df['starttime'].contains('2018-06-04'))
```

```
In [184]: table180604.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|56.87918395573996|
+-----------------+
```

```
In [185]: table180604.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|565.1590594744122|
+-----------------+
```

In [186]:
```
1 table180604.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     408610|
+-----------+
```

- 20180605

In [187]:
```
table180605 = df.filter(df['starttime'].contains('2018-06-05'))
```

In [188]:
```
table180605.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|56.14689226519336|
+-----------------+
```

In [189]:
```
table180605.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|664.4502762430939|
+-----------------+
```

In [190]:
```
1 table180605.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     481062|
+-----------+
```

- 20180606

In [191]:
```
table180606 = df.filter(df['starttime'].contains('2018-06-06'))
```

In [192]:
```
table180606.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|55.94876901798057|
+-----------------+
```

In [193]:
```
table180606.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|733.7348066298342|
+-----------------+
```

In [194]:
```
1 table180606.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     531224|
+-----------+
```

- 20180607

In [195]:
```
table180607 = df.filter(df['starttime'].contains('2018-06-07'))
```

In [196]:
```
table180607.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|56.40349381017882|
+-----------------+
```

In [197]:
```
table180607.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|668.7736625514403|
+-----------------+
```

In [198]:
```
1 table180607.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     487536|
+-----------+
```

- 20180608

In [199]:
```
table180608 = df.filter(df['starttime'].contains('2018-06-08'))
```

In [200]:
```
table180608.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|54.08467123287674|
+-----------------+
```

In [201]:
```
table180608.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|643.6747967479674|
+-----------------+
```

In [202]:
```
1 table180608.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     475032|
+-----------+
```

- 20180609

In [203]:
```
table180609 = df.filter(df['starttime'].contains('2018-06-09'))
```

In [204]:
```
table180609.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|57.75826388888892|
+-----------------+
```

In [205]:
```
table180609.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|251.8370883882149|
+-----------------+
```

In [206]:
```
1 table180609.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     145310|
+-----------+
```

- 20180610

```
In [207]: table180610 = df.filter(df['starttime'].contains('2018-06-10'))

In [208]: table180610.agg({'speed': 'avg'}).show()
+-----------------+
|       avg(speed)|
+-----------------+
|58.24217213114749|
+-----------------+

In [209]: table180610.agg({'volume': 'avg'}).show()
+-----------------+
|      avg(volume)|
+-----------------+
|724.8422131147541|
+-----------------+

In [210]: 1 table180610.agg({'volume': 'sum'}).show()
+-----------+
|sum(volume)|
+-----------+
|     353723|
+-----------+
```

- Week-2017 dataset

```
In [211]: df = spark.read.csv('week-2017.csv', inferSchema=True, header=Tru

In [212]: df.show()
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2017-06-05 00:00:00|  01:00:00|     100676|51.98|   206|     1.
84|          171| 0.04|      0.31|1.07| 55.62|
```

- 20170605

In [213]:
```
table170605 = df.filter(df['starttime'].contains('2017-06-05'))
```

In [214]:
```
table170605.agg({'speed': 'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|57.299324137931016|
+------------------+
```

In [215]:
```
table170605.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|790.5613793103448|
+-----------------+
```

In [216]:
```
1 table170605.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     573157|
+-----------+
```

- 20170606

In [217]:
```
table170606 = df.filter(df['starttime'].contains('2017-06-06'))
```

In [218]:
```
table170606.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|56.34639502762432|
+-----------------+
```

In [219]:
```
table170606.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|782.1401098901099|
+-----------------+
```

In [220]:
```
1 table170606.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     569398|
+-----------+
```

- 20170607

In [221]:
```
table170607 = df.filter(df['starttime'].contains('2017-06-07'))
```

In [222]:
```
table170607.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|56.35655266757862|
+-----------------+
```

In [223]:
```
table170607.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|836.2311901504788|
+-----------------+
```

In [224]:
```
1 table170607.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     611285|
+-----------+
```

- 20170608

In [225]:
```
table170608 = df.filter(df['starttime'].contains('2017-06-08'))
```

In [226]:
```
table170608.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|55.61853591160218|
+-----------------+
```

In [227]:
```
table170608.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|840.5455801104972|
+-----------------+
```

In [228]:
```
1 table170608.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     608555|
+-----------+
```

- 20170609

In [229]:
```
table170609 = df.filter(df['starttime'].contains('2017-06-09'))
```

In [230]:
```
table170609.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|55.40690934065928|
+-----------------+
```

In [231]:
```
table170609.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|831.0947802197802|
+-----------------+
```

In [232]:
```
1 table170609.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     605037|
+-----------+
```

- 20170610

In [233]:
```
table170610 = df.filter(df['starttime'].contains('2017-06-10'))
```

In [235]:
```
table170610.agg({'speed': 'avg'}).show()
```

```
+-----------------+
|       avg(speed)|
+-----------------+
|59.27706849315063|
+-----------------+
```

In [236]:
```
table170610.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|730.0849315068493|
+-----------------+
```

In [237]:
```
1 table170610.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     532962|
+-----------+
```

- 20170611

In [238]:
```
table170611 = df.filter(df['starttime'].contains('2017-06-11'))
```

In [239]:
```
table170611.agg({'speed': 'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|60.323017241379226|
+------------------+
```

In [240]:
```
table170611.agg({'volume': 'avg'}).show()
```

```
+-----------------+
|      avg(volume)|
+-----------------+
|631.5316091954023|
+-----------------+
```

In [242]:
```
1 table170611.agg({'volume': 'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|     439546|
+-----------+
```

2016

In [2]:
```
spark = SparkSession.builder.appName('traffic').getOrCreate()
```

In [3]:
```
df = spark.read.csv('2016-06.csv', inferSchema=True, header=True)
```

In [4]:
```
df.show()
```

```
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime| vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+----+------+
|2016-06-01 00:00:00|  01:00:00|     101041| 58.1|    82|     0.
78|          124| 0.05|      1.64|2.24|130.38|
```

In [6]:
```
df.agg({'speed':'avg'}).show()
```

```
+------------------+
|        avg(speed)|
+------------------+
|57.254966478015035|
+------------------+
```

In [7]:
```
df.agg({'volume': 'avg'}).show()
```

```
+----------------+
|     avg(volume)|
+----------------+
|681.778715503418|
+----------------+
```

In [9]:
```
df.agg({'volume':'sum'}).show()
```

```
+-----------+
|sum(volume)|
+-----------+
|   16156792|
+-----------+
```