Advanced Topics in Database Systems
Semester Assignment
Group 19
Vlachakis Nikos, el18441
Apostolos Garos, el18198

Github Link: https://github.com/ApostolisGaros/Spark_Hadoop_Project_AdvancedDB

Comment: We figured that almost every query was executed slightly faster when utilizing 2 workers, which was expected. However, in 2 cases (Q3_RDD and Q5) we recorder slightly worse performance. Additionally, the performances were really close in every test. These are probably a result of the dataset being relatively small and the distribution to 2 workers did not prove significantly faster. This is due to the overhead that accompanies such distribution.

1) Q1: Find the route with the biggest tip in March and the arrival point "Battery Park"

Q1 time taken: 14.336193323135376 seconds. (2 workers)
Q1 time taken: 15.106910299156738 seconds. (1 worker)

Attribute	Value
VendorID	2
tpep_pickup_datetime	2022-03-17 12:27:47
tpep_dropoff_datetime	2022-03-17 12:27:58
passenger_count	1.0
trip_distance	0.0

RatecodeID	1.0
store_and_fwd_flag	N
PULocationID	12
DOLocationID	12
payment_type	1
fare_amount	2.5
extra	0.0
mta_tax	0.5
tip_amount	40.0
tolls_amount	0.0
improvement_surcharge	0.3
total_amount	45.8
congestion_surcharge	2.5

airport_fee	0.0

2) Q2: Find, for each month, the route with the highest amount of tolls. Ignore zero amounts.

Q2 time taken: 40.03586554527283 seconds. (2 workers)
Q2 time taken: 41.10699152946472 seconds. (1 worker)

Attributes	January	March	May	February	April	June
VendorID	1	1	1	1	1	1
tpep_pickup_datetime	2022-1-22 11:39:07	2022-3-11 20:08:32	2022-5-21 16:47:48	2022-2-18 02:33:30	2022-4-29 04:31:21	2022-6-12 16:51:46
tpep_dropoff_datetime	2022-1-22 12:31:09	2022-3-11 20:09:45	2022-5-21 17:05:47	2022-2-18 02:35:28	2022-4-29 04:32:30	2022-6-12 17:56:48
passenger_count	1	1	1	1	2	9
trip_distance	33.4	0	2.4	1.3	0	22
RatecodeID	1	1	3	1	1	1

store_and_fwd_flag	Y	N	N	N	N	N
PULocationID	70	265	239	265	249	142
DOLocationID	265	265	246	265	249	132
payment_type	4	1	3	1	3	2
fare_amount	88	2.5	31.5	3	3	67.5
extra	0	1	0	0.5	3	2.5
mta_tax	0.5	0.5	0	0.5	0.5	0.5
tip_amount	0	48	0	19.85	0	0
tolls_amount	193.3	235.7	813.75	95	911.87	800.09
improvement_surcharge	0.3	0.3	0.3	0.3	0.3	0.3
total_amount	282.1	288	845.55	119.15	918.67	870.89
congestion_surcharge	0	0	0	0	2.5	2.5

airport_fee	0	0	0	0	0	0

3) Q3: Find, per 15 days, the average distance and cost for all the routes and the average cost for all the routes with a departure point different from the arrival point.

Q3_DF time taken: 21.311964750289917 seconds. (2 workers)
Q3_DF time taken: 22.71215009689331 seconds. (1 worker)

Q3_RDD time taken: 284.08192801475525 seconds. (2 workers)
Q3_RDD time taken: 282.54228043556213 seconds. (1 worker)

Group	15-Day Average Trip Distance	15-Day Average Total Amount
0	5.58	20.04
1	5.23	19.02
2	6.0	19.6
3	6.17	20.2
4	6.69	20.72
5	5.51	21.15

6	5.67	21.54
7	5.81	21.43
8	6.26	21.94
9	7.89	22.79
10	6.38	22.46
11	6.16	22.36
12	5.81	22.17

4) Q4: Find the top three (top 3) peak hours per day of the week, meaning the hours (e.g., 7-8am, 3-4pm, etc.) of the day with the highest number of passengers in a taxi trip. The calculation applies to all months

Q4 time taken: 20.776392221450806 seconds. (2 workers)
Q4 time taken: 21.249610662460327 seconds. (1 worker)

Day of Week	Hour	Avg Passenger Count	Index
1	0	1.52995	1
1	1	1.52784	2
1	2	1.50807	3
2	0	1.468	1
2	1	1.44429	2
2	2	1.4232	3
3	0	1.42003	1
3	1	1.41751	2

Day of Week	Hour	Avg Passenger Count	Index
3	2	1.41045	3
4	1	1.40885	1
4	0	1.40123	2
4	2	1.40115	3
5	23	1.40538	1
5	1	1.40259	2
5	0	1.40104	3
6	23	1.47558	1
6	22	1.44481	2
6	2	1.42306	3
7	23	1.5226	1
7	22	1.50682	2

Day of Week	Hour	Avg Passenger Count	Index
7	0	1.49932	3

5) Q5: Find the top five (top 5) days per month on which the races had the highest percentage of tips. For example, if the race cost 10$ (fare_amount) and the tip was $5, the percentage is 50%

Q5 time taken: 21.779485940933228 seconds. (2 workers)
Q5 time taken: 20.235378421578932 seconds. (1 worker)

Month	Day of Month	Tip Percentage	Index
1	29	0.2154833669	1
1	15	0.1953226616	2
1	22	0.1933725570	3
1	30	0.1928065002	4
1	21	0.1927674958	5
2	4	0.195575922	1
2	5	0.195341226	2
2	6	0.194006362	3
2	10	0.1935517543	4
2	17	0.1929098458	5
3	9	0.1955577609	1

3	12	0.193920178	2
3	30	0.1932932451	3
3	24	0.1927850108	4
3	10	0.192740558	5
4	1	0.191378334	1
4	7	0.1912479413	2
4	6	0.190912209	3
4	27	0.1903199258	4
4	28	0.1893686874	5
5	12	0.1921416227	1
5	4	0.1913877678	2
5	11	0.1902909360	3
5	15	0.1893044270	4

5	29	0.1879759500	5
6	16	0.1904	1
6	8	0.1897	2
6	23	0.1892	3
6	9	0.1891	4
6	17	0.1883	5