Capstone Project - Validation Document

Note: Please note that records retrieved from the Kafka topic and related metrics given below can vary a bit.

Data Ingestion with Sqoop

Please check the number of records that are imported after the Sqoop Job

```
Number of records retrieved - 1000
```

Bookings Table Count

Please check the number of records in the bookings table

```
Number of records - 1000
```

Clickstream Table Count

Please check the number of records in the clickstream table

```
Number of records - 2984
```

Bookings Aggregates Table Count

Please check the number of records in the bookings aggregates table

```
Number of records - 289
```

Hive Queries

1. When you run the query to calculate the total number of different drivers for each customer, you would get an output as shown below:

```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-11-17 12:23:06,034 Stage-1 map = 0%,  reduce = 0%
2020-11-17 12:23:12,394 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.27 sec
2020-11-17 12:23:20,727 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.69 sec
MapReduce Total cumulative CPU time: 7 seconds 690 msec
Ended Job = job_1605615116654_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.69 sec   HDFS Read: 43007 HDFS Write: 11000 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 690 msec
OK
10022393        1
10058402        1
10339567        1
10435129        1
10555335        1
10592274        1
10614890        1
10678994        1
11264797        1
11353346        1
11418437        1
11438890        1
11454977        1
11479815        1
11518953        1
11580321        1
11596512        1
11608791        1
11655671        1
11757536        1
11764909        1
11860278        1
11981042        1
12106105        1
12142182        1
12312603        1
12334699        1
12367832        1
12856708        1
12885363        1
12913608        1
12914577        1
12966909        1
13015449        1
13229062        1
```

2. When you run the query to calculate the total rides taken by each customer, you would get an output as shown below:

```
Ended Job = job_1605615116654_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.65 sec   HDFS Read: 38721 HDFS Write: 11000 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 650 msec
OK
10022393        1
10058402        1
10339567        1
10435129        1
10555335        1
10592274        1
10614890        1
10678994        1
11264797        1
11353346        1
11418437        1
11438890        1
11454977        1
11479815        1
11518953        1
11580321        1
11596512        1
11608791        1
11655671        1
11757536        1
11764909        1
11860278        1
11981042        1
12106105        1
12142182        1
12312603        1
```

3. When you run the query to get the conversion ratio, you should get the conversion ratio as **0.9688.**
4. Count of all trips done on black cabs - **72.**
5. When you run the query to get the total amount of tips given date wise to all drivers by customers, you would get an output as shown below:

2020-01-01	59
2020-01-02	95
2020-01-03	11
2020-01-04	123
2020-01-05	134
2020-01-06	189
2020-01-07	148
2020-01-08	111
2020-01-09	48
2020-01-10	77
2020-01-11	81
2020-01-12	109
2020-01-14	142
2020-01-15	338
2020-01-16	155
2020-01-17	296
2020-01-18	240
2020-01-20	210
2020-01-21	5
2020-01-23	148
2020-01-24	472
2020-01-25	98
2020-01-26	209
2020-01-27	231
2020-01-28	567
2020-01-29	123
2020-01-30	112
2020-01-31	256
2020-02-01	317
2020-02-02	338
2020-02-03	191
2020-02-04	258
2020-02-05	212
2020-02-06	154
2020-02-07	91
2020-02-08	270

6. When you run the query to get the total count of all the bookings with ratings lower than 2 as given by customers in a particular month, you would get an output as shown below:

```
Total MapReduce CPU Time Spent: 7 seconds 970 msec
OK
2020-01 26
2020-02 16
2020-03 16
2020-04 21
2020-05 21
2020-06 14
2020-07 20
2020-08 32
2020-09 21
2020-10 15
```

7. You should get the count of all iOS users as **1503**.