

Logic For First Submission

Task 1: Write a job to consume clickstream data from Kafka and ingest to Hadoop.

Steps:

1. Import all libraries required for running the code.

```
from pyspark.sql import SparkSession  
from pyspark.sql.functions import *  
from pyspark.sql.types import *  
from pyspark.sql.functions import from_json
```

2. Initialize the spark session and set app name as Kafka to local

```
spark = SparkSession \  
.builder \  
.appName("Kafka-to-local") \  
.getOrCreate()  
spark.sparkContext.setLogLevel('ERROR')
```

3. Read data from Kafka by subscribing to de-capstone3 topic

```
kafka_read = spark \  
.readStream \  
.format("kafka") \  
.option("kafka.bootstrap.servers","18.211.252.152:9092") \  
.option("subscribe","de-capstone3") \  
.option("startingOffsets", "earliest") \  
.load()
```

4. Drop the columns which are not required and change the column name of value to value_str

```
updated_data= kafka_read \  
.withColumn('value_str',kafka_read['value'].cast('string').alias('key_str')).drop('value') \  
.drop('key','topic','partition','offset','timestamp','timestampType')
```

5. Write the data to hadoop using below code. We must use path where the raw clickstream data is stored as well as checkpoint path in HDFS

```
clickstream_data = updated_data.writeStream \  
.format("json") \  
.outputMode("append")
```

```
.option("truncate", "false") \
.option("path", "clickstream_data/") \
.option("checkpointLocation", "clickstream_data/cp/") \
.start()
```

```
clickstream_data.awaitTermination()
```

Cluster: Capstone-2 Waiting Cluster ready to run steps.

```
[hadoop@ip-10-0-0-52 ~]$ hadoop fs -ls clickstream_data
Found 3 items
drwxr-xr-x  - hadoop hdfsadmingroup          0 2022-07-24 19:24 clickstream_data/_spark_metadata
drwxr-xr-x  - hadoop hdfsadmingroup          0 2022-07-24 19:24 clickstream_data/cp
-rw-r--r--  1 hadoop hdfsadmingroup 1267605 2022-07-24 19:24 clickstream_data/part-00000-b34c4c24-11af-4655-8eb3-cd4495aa0f27-c000.json
[hadoop@ip-10-0-0-52 ~]$
```

Cluster: Capstone-2 Waiting Cluster ready to run steps.

```
[hadoop@ip-10-0-0-52 ~]
{"value_str": "{\"customer_id\": \"33090503\", \"app_version\": \"3.1.4\", \"OS_version\": \"Android\", \"lat\": \"-53.8831745\", \"lon\": \"150.937637\", \"page_id\": \"e7b5cfb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_click\": \"No\", \"is_page_view\": \"No\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-02-18 07:12:55\\n\"}"}
{"value_str": "{\"customer_id\": \"83436917\", \"app_version\": \"1.3.19\", \"OS_version\": \"iOS\", \"lat\": \"29.6107005\", \"lon\": \"91.528804\", \"page_id\": \"de545711-3914-4450-8c11-b17b8ddab5e1\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"No\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-07-30 05:04:13\\n\"}"}
{"value_str": "{\"customer_id\": \"68796997\", \"app_version\": \"2.3.29\", \"OS_version\": \"iOS\", \"lat\": \"43.2997955\", \"lon\": \"-42.645686\", \"page_id\": \"b328829-e-17ae-11eb-adc1-0242ac120002\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"Yes\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-10-17 20:00:19\\n\"}"}
{"value_str": "{\"customer_id\": \"84392327\", \"app_version\": \"1.3.12\", \"OS_version\": \"iOS\", \"lat\": \"40.7522335\", \"lon\": \"-105.599661\", \"page_id\": \"de545711-3914-4450-8c11-b17b8ddab5e1\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"No\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-10-17 20:00:19\\n\"}"}
{"value_str": "{\"customer_id\": \"39542434\", \"app_version\": \"2.1.11\", \"OS_version\": \"iOS\", \"lat\": \"-8.8883795\", \"lon\": \"31.780290\", \"page_id\": \"e7bc5fb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"Yes\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-04-05 18:00:34\\n\"}"}
{"value_str": "{\"customer_id\": \"16813062\", \"app_version\": \"3.2.18\", \"OS_version\": \"Android\", \"lat\": \"-88.342371\", \"lon\": \"-176.291730\", \"page_id\": \"b328829e-17ae-11eb-adc1-0242ac120002\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_click\": \"No\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-05-05 23:08:24\\n\"}"}
{"value_str": "{\"customer_id\": \"13307480\", \"app_version\": \"4.3.36\", \"OS_version\": \"iOS\", \"lat\": \"26.6645385\", \"lon\": \"109.842184\", \"page_id\": \"b328829-e-17ae-11eb-adc1-0242ac120002\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_click\": \"No\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"Yes\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-06-12 19:23:21\\n\"}"}
{"value_str": "{\"customer_id\": \"66537285\", \"app_version\": \"3.4.37\", \"OS_version\": \"Android\", \"lat\": \"-43.4804705\", \"lon\": \"108.716577\", \"page_id\": \"de545711-3914-4450-8c11-b17b8ddab5e1\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-09-07 17:10:58\\n\"}"}
{"value_str": "{\"customer_id\": \"48434687\", \"app_version\": \"4.1.28\", \"OS_version\": \"Android\", \"lat\": \"-7.888734\", \"lon\": \"-54.732169\", \"page_id\": \"b328829e-17ae-11eb-adc1-0242ac120002\", \"button_id\": \"e95dd57b-779f-49db-b990483e554\", \"is_button_click\": \"Yes\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-02-18 19:48:48\\n\"}"}
{"value_str": "{\"customer_id\": \"98030364\", \"app_version\": \"3.3.35\", \"OS_version\": \"iOS\", \"lat\": \"59.4450305\", \"lon\": \"-66.155279\", \"page_id\": \"e7bc5fb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-04-15 17:33:44\\n\"}"}
[hadoop@ip-10-0-0-52 ~]$
```

Flattening File:

Running spark_local_flattening.py to clean data for data analysis.

1. Importing required libraries

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
```

2. Initialize Spark Session

```
spark=SparkSession.builder.appName("Kafka-to-HDFS").master("local").getOrCreate()
spark
#Reading data from hdfs
df=spark.read.json("clickstream.json")
df.show(10,truncate=False)
```

3. Selecting the columns from the clickstream data set

```
df=df.select(get_json_object(df['value_str'],"$.customer_id").alias("customer_id"),
            get_json_object(df['value_str'],"$.app_version").alias("app_version"),
            get_json_object(df['value_str'],"$.OS_version").alias("OS_version"),
            get_json_object(df['value_str'],"$.lat").alias("lat"),
            get_json_object(df['value_str'],"$.lon").alias("lon"),
            get_json_object(df['value_str'],"$.page_id").alias("page_id"),
            get_json_object(df['value_str'],"$.button_id").alias("button_id"),
            get_json_object(df['value_str'],"$.is_button_click").alias("is_button_click"),
            get_json_object(df['value_str'],"$.is_page_view").alias("is_page_view"),
            get_json_object(df['value_str'],"$.is_scroll_up").alias("is_scroll_up"),
            get_json_object(df['value_str'],"$.is_scroll_down").alias("is_scroll_down"),
            get_json_object(df['value_str'],"$.timestamp").alias("timestamp")
        )
df.printSchema()
df.show(10)
```

4. Writing the dataset to HDFS

```
df.coalesce(1).write.format('csv').mode('overwrite').save('/user/hadoop/clickstream_flattened',header='true')
```

```
[hadoop@ip-10-0-0-52 ~]$ hadoop fs -ls clickstream_flattened
Found 2 items
-rw-r--r-- 1 hadoop hdfsadmingroup 0 2022-07-24 20:44 clickstream_flattened/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmingroup 403733 2022-07-24 20:44 clickstream_flattened/part-00000-6da4f575-e02-471f-b726-1ad6f743c073-c000.csv
[hadoop@ip-10-0-0-52 ~]$ █
```

```
+-----+
|value_str
+-----+
|{"customer_id": "26564820", "app_version": "3.2.35", "OS_version": "Android", "lat": "16.4454865", "lon": "99.902065", "page_id": "de545711-3914-4450-8c11-b17b8ddbb5e1", "button_id": "fcba68aa-1231-11eb-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "Yes", "timestamp\n": "2020-09-14 09:59:07n"} |
|{"customer_id": "31906387", "app_version": "2.4.7", "OS_version": "Android", "lat": "-64.813749", "lon": "-133.527940", "page_id": "de545711-3914-4450-8c11-b17b8ddbb5e1", "button_id": "095dd57b-779f-49db-819d-b6960483e554", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp\n": "2020-05-16 16:30:21n"} |
|{"customer_id": "25713677", "app_version": "3.4.12", "OS_version": "Android", "lat": "89.943435", "lon": "127.313415", "page_id": "b22829e-176e-11eb-adc1-0242ac120002", "button_id": "fcba68aa-1231-11eb-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "No", "is_scroll_up": "Yes", "is_scroll_down": "No", "timestamp\n": "2020-02-09 00:52:13n"} |
|{"customer_id": "83474293", "app_version": "3.1.8", "OS_version": "Android", "lat": "-69.939070", "lon": "-36.451670", "page_id": "e1e99492-17ae-11eb-adc1-0242ac120002", "button_id": "e1e99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "Yes", "is_page_view": "No", "is_scroll_up": "Yes", "is_scroll_down": "No", "timestamp\n": "2020-06-17 10:42:50n"} |
|{"customer_id": "63727205", "app_version": "2.3.9", "OS_version": "iOS", "lat": "64.082108", "lon": "-81.822078", "page_id": "e1e99492-17ae-11eb-adc1-0242ac120002", "button_id": "fcba68aa-1231-11eb-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp\n": "2020-07-06 02:51:53n"} |
|{"customer_id": "73737477", "app_version": "4.3.19", "OS_version": "Android", "lat": "-18.850508", "lon": "-116.558575", "page_id": "b52829e-17e-11eb-adc1-0242ac120002", "button_id": "e1e99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "Yes", "timestamp\n": "2020-04-26 06:18:16n"} |
|{"customer_id": "36927433", "app_version": "3.2.26", "OS_version": "iOS", "lat": "-84.6857245", "lon": "-146.567671", "page_id": "de545711-3914-4450-8c11-b17b8ddbb5e1", "button_id": "a95dd57b-779f-49db-819d-b6960483e554", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "Yes", "timestamp\n": "2020-02-08 10:40:18n"} |
|{"customer_id": "1269178", "app_version": "3.3.11", "OS_version": "Android", "lat": "38.52025", "lon": "41.411814", "page_id": "de545711-3914-4450-8c11-b17b8ddbb5e1", "button_id": "e1e99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "No", "timestamp\n": "2020-08-08 04:23:56n"} |
|{"customer_id": "22659021", "app_version": "4.4.36", "OS_version": "iOS", "lat": "31.805500", "lon": "159.655550", "page_id": "s70c57b-1231-11eb-adc1-0242ac120002", "button_id": "a95dd57b-779f-49db-819d-b6960483e554", "is_button_click": "No", "is_page_view": "No", "is_scroll_up": "No", "is_scroll_down": "No", "timestamp\n": "2020-08-02 00:33:50n"} |
|{"customer_id": "23593546", "app_version": "1.2.16", "OS_version": "Android", "lat": "8.8918475", "lon": "-83.529878", "page_id": "de545711-3914-4450-8c11-b17b8ddbb5e1", "button_id": "e1e99492-17ae-11eb-adc1-0242ac120002", "is_button_click": "Yes", "is_page_view": "No", "is_scroll_up": "Yes", "is_scroll_down": "No", "timestamp\n": "2020-07-23 23:59:19n"} |
```

Task 2: Write a script to ingest the relevant bookings data from AWS RDS to Hadoop.

Steps:

For performing any analysis on booking data , first we need to import data from AWS RDS to HDFS.

Sqoop Import Command

```
sqoop import \
```

```
--connect jdbc:mysql://upgradetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table bookings \
--username student --password STUDENT123 \
--target-dir /user/hadoop/cab_rides \
-m 1
```

To view imported data

```
hadoop fs -ls /user/hadoop/cab_rides
```

```
[root@ip-10-0-0-52 ~]$ hadoop fs -ls /user/root/cab_rides
hadoop@ip-10-0-0-52:~
```

Found 2 items

```
-rw-r--r-- 1 root hdfsadmingroup 0 2022-07-24 20:15 /user/root/cab_rides/_SUCCESS
-rw-r--r-- 1 root hdfsadmingroup 165678 2022-07-24 20:15 /user/root/cab_rides/part-m-00000
```

[hadoop@ip-10-0-0-52 ~]\$ hadoop fs -cat /user/root/cab_rides/part-m-00000

```
BK8968087150,51811159,15055660,2,2,14,Android,-49,4319655,103,917851,-58,8043875,146,477367,2020-06-23 19:33:10.0,2020-06-06 09:02:10.0,534,83,INR,black,054-38-4479,4,3,3
BK629851904,316635218,60872180,3,4,1,10S,-83,5408405,175,80085,86,20705,128,367238,2020-05-23 12:24:04,0,2020-08-09 19:02:56.0,126,67,INR,lime,796-39-6801,3,2,4
BK1797410350,86869399,94276051,4,1,36,10S,-67,8930645,55,234128,-51,1079,-31,07475,2020-05-19 14:14:32.0,2020-08-23 18:38:39.0,297,63,INR,olive,748-73-1579,1,3,3
BK5788246325,58230837,45457227,2,4,27,Android,13,707887,113,499943,54,3812915,-18,437751,2020-03-24 01:30:15.0,2020-05-19 11:16:45.0,932,32,INR,white,558-80-6346,3,2,2
BK8342703255,84232510,86494681,4,1,34,Android,-6,091461,-114,649789,22,8449505,79,137827,2020-08-03 19:10:52.0,2020-03-24 08:25:40.0,260,7,INR,blue,068-72-1637,3,3,3
BK60155821042,35862658,2,4,39,10S,-18,910034,-70,193103,-10,182921,173,877213,2020-07-17 05:33:48.0,2020-04-30 04:54:27.0,907,53,INR,purple,102-10-5639,3,2,3
BK4529355854,60071878,,8022360,2,1,9,10S,1,215274,-56,014903,35,152876,104,324905,2020-01-02 01:48:40,0,2020-02-16 04:28:55.0,547,17,INR,teal,866-83-4349,2,3,4
BK9720088219,14327312,94427067,3,1,2,Android,-55,4822256,65,0121265,7,126015,-16,826146,7,6126015,-156,428577,2020-06-09 05:56:31.0,2020-03-19 01:53:16.0,787,21,INR,olive,667-23-5880,2,2,3
BK7157532607,464047210,43160003,1,3,4,Android,46,005843,-16,826146,7,6126015,-156,428577,2020-06-09 05:56:31.0,2020-03-19 01:53:16.0,787,21,INR,olive,667-23-5880,2,2,3
BK5014871433,65861573,64708618,1,3,28,iOS,-29,565326,64,843709,84,068109,-49,820835,2020-08-14 20:43:42,0,2020-06-03 09:59.0,586,5,INR,fuchsia,255-52-5654,5,5,1
BK9051488736,37721758,37297770,2,3,13,Android,61,9364605,83,249705,0,0281895,115,4969999,2020-04-07 04:27:59,0,2920-09-29 10:51:41,0,912,80,INR,aqua,739-09-9569,2,1,2
BK243762319,62552969,45877457,3,3,9,10S,-62,6515155,-139,154028,28,0299995,-62,8556,2020-07-01 00:36:05,0,2020-09-30 17:40:23,0,821,23,INR,black,590-44-6613,2,3,4
BK4683595168,56801961,53401707,4,2,34,10S,-5,860265,-100,004839,25,016591,70,471358,2020-05-01 10:17:56,0,2020-06-08 09:11:27,0,71,10,INR,fuchsia,454-04-0608,5,2,3
BK978324253,669909721,40509554,2,2,22,Android,36,1913155,5,686264,88,988393,36,580599,2020-03-01 16:02:01,0,2020-05-29 13:36:15,0,26,81,INR,black,600-17-7043,3,1,3
BK2880021380,50163555,34405420,3,4,23,Android,-83,06591,108,268689,8,8300855,74,872352,2020-01-15 02:00:07,0,2020-05-12 21:53:04,0,571,99,INR,navy,506-09-4981,1,5,3
BK4537426043,91111754,59250769,1,2,19,iOS,-43,1188435,-99,935719,3,702625,46,828716,2020-04-28 05:18:34,0,2020-02-12 11:31:40,0,650,81,INR,white,362-35-8054,5,5,2
BK998130731,67875357,14562526,3,1,15,Android,-10,861959,-111,988853,57,233121,95,469986,2020-01-25 01:37:22,0,2020-04-28 09:42:00,0,590,3,INR,teal,359-51-9362,1,1,4
BK5645323730,18442993,84939946,3,1,29,Android,-81,472235,-88,404916,12,690818,-140,99768,2020-09-24 05:18:31,0,2020-07-14 05:12:24,0,515,1,INR,blue,024-35-8771,1,3,4
BK6163608413,36591778,11946210,4,2,38,Android,60,2036385,120,988501,32,103263,-50,551889,2020-07-26 06:12:56,0,2020-04-23 06:57:20,0,810,58,INR,silver,833-16-1378,3,5,1
BK860373649,28382306,97222676,2,1,18,10S,6,540056,161,083998,-12,943502,-148,232621,2020-09-20 15:52:49,0,2020-09-17 03:13:26,0,927,74,INR,maroon,747-70-5557,2,2,4
BK9764570097,61225539,15265942,4,4,14,10S,80,7211615,179,695812,-33,345655,134,010372,2020-01-26 02:20:39,0,2020-06-12 15:05:49,0,246,72,INR,blue,332-71-7565,5,1,2
BK8362601204,79115927,60281490,3,4,16,10S,-9,4458645,101,745883,80,264612,-46,718991,2020-09-03 13:32:13,0,2020-02-01 21:02:21,0,887,88,INR,blue,225-31-0761,4,1,1
BK6225330481,51110772,58392277,3,3,27,10S,68,1306075,141,450665,-6,926722,-6,24554,2020-05-11 06:25:10,0,2020-06-29 09:31:03,0,429,18,INR,purple,229-41-2152,5,1,3
BK9785297548,13229662,95750789,4,1,14,Android,-57,959954,-172,155546,1,667888,126,729718,2020-04-18 20:12:36,0,2020-01-19 09:07:36,0,35,63,INR,white,681-74-4532,2,3,3
BK4218069991,86269148,4493601,4,2,9,Android,74,050649,-130,95903,-85,015584,106,190804,2020-08-22 18:50:02,0,2020-01-25 09:57:30,0,385,97,INR,aqua,678-19-9649,2,4,2
```

Task 3: Create aggregates for finding date-wise total bookings using the Spark script.

Steps:

1. Import all libraries required for running the code.

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
```

2. Initialize the spark session and set app name as Kafka to local

```
spark=SparkSession.builder.appName("datewise_bookings_aggregates_spark").master("local")
  .getOrCreate()
spark.sparkContext.setLogLevel('ERROR')
```

3. Read the data from HDFS which was loaded using Sqoop.

```
df=spark.read.csv("/user/hadoop/cab_rides/part-m-00000")
df.show(10)
df.printSchema()
```

```
22/07/24 20:58:40 INFO DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 0.238127 s
22/07/24 20:58:40 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| _c0| _c1| _c2| _c3| _c4| _c5| _c6| _c7| _c8| _c9| _c10| _c11| _c12| _c13| _c14| _c15| _c16| _c17| _c18|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| BK8968087150| 51811359| 15055660| 2.2.14| Android| -49.4319655| 103.917851| -58.8043875| 146.477367| 2020-06-23 19:33:...| 2020-06-06 09:02:...| 534| 83| INR| black| 054-38-4479| 4| 31| 31
| BK29851904| 31663218| 60872180| 3.4.11| IOS| -83.5408405| 175.80085| 86.20705| 128.367238| 2020-05-23 12:22:...| 2020-08-09 19:02:...| 126| 67| INR| lime| 796-39-6801| 3| 21| 41
| BK11797410350| 186869399| 94276051| 4.1.36| IOS| -67.8930645| 55.234128| -51.1079| -31.07475| 2020-05-19 14:14:...| 2020-08-23 18:38:...| 297| 63| INR| olive| 748-73-1579| 3| 31| 31
| BK578246325| 158230837| 745457227| 12.4.27| Android| 13.707887| 113.499943| 54.3812915| -18.437751| 2020-03-24 01:30:...| 2020-05-19 11:16:...| 932| 32| INR| white| 558-80-6346| 3| 21| 21
| BK8342793255| 184232510| 86494681| 4.1.34| Android| -6.091461| -114.649789| 22.8449505| 70.137827| 2020-08-03 19:10:...| 2020-03-24 08:25:...| 260| 7| INR| blue| 068-72-1637| 3| 31| 31
| BK6015582453| 11981042| 358626581| 2.4.39| IOS| -18.910034| -70.19103| -10.182921| 173.877213| 2020-07-17 05:33:...| 2020-04-30 04:54:...| 907| 53| INR| purple| 102-10-5639| 3| 21| 31
| BK4529355854| 160071878| 78022360| 2.1.91| IOS| 1.215274| -56.014903| 35.152876| 104.324905| 2020-01-02 01:48:...| 2020-02-16 04:28:...| 547| 17| INR| teal| 866-83-4349| 2| 31| 41
| BK9720088219| 14327312| 944270671| 3.1.21| Android| -55.482225| 173.362256| 65.0121265| 51.390751| 2020-04-10 15:11:...| 2020-01-20 21:17:...| 259| 33| INR| maroon| 572-73-6526| 3| 31| 21
| BK7157532607| 146407210| 143160003| 1.3.41| Android| 46.005843| -16.826146| 7.61260151| -156.428577| 2020-06-09 05:56:...| 2020-03-19 01:53:...| 787| 21| INR| olive| 667-23-5880| 2| 21| 31
| BK6014871433| 165861573| 647086181| 1.3.28| IOS| -29.565326| 64.843709| 84.068109| -49.820835| 2020-08-14 20:43:...| 2020-06-03 09:39:...| 586| 5| INR| fuchsia| 255-52-5654| 5| 51| 1|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 10 rows

root
|- _c0: string (nullable = true)
|- _c1: string (nullable = true)
|- _c2: string (nullable = true)
|- _c3: string (nullable = true)
|- _c4: string (nullable = true)
|- _c5: string (nullable = true)
|- _c6: string (nullable = true)
|- _c7: string (nullable = true)
|- _c8: string (nullable = true)
|- _c9: string (nullable = true)
|- _c10: string (nullable = true)
|- _c11: string (nullable = true)
|- _c12: string (nullable = true)
|- _c13: string (nullable = true)
|- _c14: string (nullable = true)
|- _c15: string (nullable = true)
|- _c16: string (nullable = true)
|- _c17: string (nullable = true)
|- _c18: string (nullable = true)

22/07/24 20:58:40 INFO ContextCleaner: Cleaned accumulator 2
```

4. Rename the columns

```
new_col =
["booking_id","customer_id","driver_id","customer_app_version","customer_phone_os_version",
"pickup_lat","pickup_lon","drop_lat",
"drop_lon","pickup_timestamp","drop_timestamp","trip_fare","tip_amount","currency_code","cab_color",
"cab_registration_no","customer_rating_by_driver","rating_by_customer","passenger_count"]
```

```
new_df = df.toDF(*new_col)
```

`new_df.show(truncate=False)`

booking_id	customer_id	driver_id	customer_app_version	customer_phone_os_version	pickup_lat	pickup_lon	drop_lat	drop_lon	pickup_timestamp	drop_timestamp	trip_fare	tip_amount	currency_code	cab_color	cab_registration_no	customer_rating_by_driver	rating_by_customer	passenger_count
1B8968087150151811359	115055660	12.2.14	iOS	iOS	-49.4319655	183.917851	-58.8043875	146.477367	2020-06-23 19:33:10.0	2020-06-06 09:02:10.0	1534	183	INR	black				
1054-38-4479	115055660	12.2.14	iOS	iOS	-48.5408405	175.80085	-86.29705	128.567238	2020-05-23 12:22:04.0	2020-08-09 19:02:56.0	126	167	INR	lime				
1030-39-6801	115055660	12.2.14	iOS	iOS	-67.8930645	155.234128	-51.1079	-31.07475	2020-05-19 14:14:32.0	2020-08-23 18:38:39.0	1297	163	INR	olive				
1796-39-6801	115055660	12.2.14	iOS	iOS	-13.707887	113.49994	154.3812915	-18.437751	2020-03-24 01:30:15.0	2020-05-19 11:16:45.0	1932	132	INR	white				
1748-73-1579	115055660	12.2.14	iOS	iOS	-6.091461	-114.6497891	22.8449505	70.137827	2020-08-03 19:10:52.0	2020-03-24 08:25:40.0	1260	17	INR	blue				
1558-80-6346	115055660	12.2.14	iOS	iOS	-18.910034	-70.193103	10.182921	1173.877213	2020-07-17 05:33:48.0	2020-04-30 04:54:27.0	1907	153	INR	purple				
1061-38-8054	115055660	12.2.14	iOS	iOS	1.215274	1-56.014903	135.152876	1104.324905	2020-01-02 01:48:40.0	2020-02-16 04:28:55.0	1547	117	INR	teal				
1061-38-8054	115055660	12.2.14	iOS	iOS	-55.4822251	173.362256	165.0121265	151.390751	2020-04-10 15:11:07.0	2020-01-20 21:17:42.0	1259	133	INR	maroon				
1061-38-8054	115055660	12.2.14	iOS	iOS	146.005843	1-16.826146	17.6126015	-156.428577	2020-06-09 05:56:31.0	2020-03-19 01:53:16.0	187	121	INR	olive				
1061-38-8054	115055660	12.2.14	iOS	iOS	-29.565326	164.843709	184.068109	-49.820835	2020-08-14 20:43:42.0	2020-06-03 09:39:59.0	1586	15	INR	fuchsia				
1061-38-8054	115055660	12.2.14	iOS	iOS	161.9364605	183.249705	10.0281895	115.469099	2020-04-07 04:27:59.0	2020-09-29 10:51:41.0	1912	180	INR	lqua				
1739-09-9569	115055660	12.2.14	iOS	iOS	-62.65151551	-139.1540281	28.0299995	-62.8556	2020-07-01 00:36:05.0	2020-09-30 17:40:23.0	1821	123	INR	black				
18K243762319	162552969	145877457	iOS	iOS	-62.65151551	-139.1540281	28.0299995	-62.8556	2020-07-01 00:36:05.0	2020-09-30 17:40:23.0	1821	123	INR	black				
1590-44-6613	162552969	145877457	iOS	iOS	-5.860265	1-100.0048391	25.016591	170.471358	2020-05-03 10:17:56.0	2020-06-08 09:11:27.0	171	110	INR	fuchsia				
1030-38-8054	162552969	145877457	iOS	iOS	136.1913155	15.686264	188.988393	136.580859	2020-03-05 16:02:01.0	2020-05-29 13:36:15.0	126	181	INR	black				
1057-38-8054	162552969	145877457	iOS	iOS	183.06599	186.266869	18.8308855	174.872352	2020-01-15 02:00:07.0	2020-05-12 21:53:04.0	1571	199	INR	navy				
1057-38-8054	162552969	145877457	iOS	iOS	143.11884351	-99.935719	13.702625	-46.828716	2020-04-28 05:18:34.0	2020-02-12 11:31:40.0	1560	181	INR	white				
1059-51-9362	162552969	145877457	iOS	iOS	-10.861959	1-111.9898531	57.233121	195.469986	2020-01-25 01:37:22.0	2020-04-28 09:42:00.0	1590	13	INR	teal				
1059-51-9362	162552969	145877457	iOS	iOS	1-81.472235	1-88.404916	12.690818	-140.99768	2020-09-24 08:18:31.0	2020-07-16 05:12:24.0	1515	11	INR	blue				
1024-35-8771	162552969	145877457	iOS	iOS	160.2036385	120.988501	32.103263	-50.51889	2020-07-26 06:12:56.0	2020-04-23 06:57:20.0	1810	158	INR	silver				
1033-16-1378	162552969	145877457	iOS	iOS	16.540056	161.083998	12.943502	-148.232621	2020-09-20 15:52:49.0	2020-09-17 03:13:26.0	1927	174	INR	maroon				
1047-78-5557	162552969	145877457	iOS	iOS														
only showing top 20 rows																		

22/07/24 20:58:42 INFO FileSourceStrategy: Pruning directories with:
 22/07/24 20:58:42 INFO FileSourceStrategy: Post-Scan Filters:

5. Converting pickup_timestamp to date by extracting date from pickup_timestamp for aggregation

```
new_df=new_df.select("booking_id","customer_id","driver_id","customer_app_version","customer_phone_os_version","pickup_lat","pickup_lon","drop_lat",
"drop_lon",to_date(col('pickup_timestamp')).alias('pickup_date').cast("date"),"drop_timestamp","trip_fare","tip_amount","currency_code","cab_color","cab_registration_no","customer_rating_by_driver",
"rating_by_customer","passenger_count")
```

`new_df.show()`

```

22/07/24 20:58:43 INFO Taskscheduleimpl: Removed taskset 6.0, whose tasks have all completed, from pool
22/07/24 20:58:43 INFO DAGScheduler: ResultStage 6 (showString at NativeMethodAccessorImpl.java:0) finished in 0.081 s
22/07/24 20:58:43 INFO DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 0.092930 s
+-----+
| booking_id|customer_id|driver_id|customer_app_version|customer_phone_os_version|pickup_latl|pickup_lonl|drop_latl|drop_lonl|pickup_date|drop_timestamp|trip_fare|tip_amount|currency_code|cab_color|cab_registration_no|customer_rating_by_driver|rating_by_customer|passenger_count|
+-----+
| 1BK89680871501 | 518113591 | 158566001 | 2.2.141 | 31 | Android|49.4319651|103.9178511-58.80438751 | 146.4773671 | 2020-06-23|2020-06-06 09:02:... | 5341 | 831 | INRI | blackl | 0 |
| 54-38-44791 | 41 | 31 | 31 | 1051-83.54084051 | 175.800851 | 86.287051 | 128.3672381 | 2020-05-23|2020-08-09 19:02:... | 1261 | 671 | INRI | lime1 | 7 |
| 1BK89680871501 | 518113591 | 608721801 | 3.4.11 | 21 | 1051-83.54084051 | 175.800851 | 86.287051 | 128.3672381 | 2020-05-23|2020-08-09 19:02:... | 1261 | 671 | INRI | lime1 | 7 |
| 96-39-68011 | 316532181 | 31 | 31 | 1051-67.89306451 | 55.2341281 | -51.10791 | -31.074751 | 2020-05-19|2020-08-23 18:38:... | 2971 | 631 | INRI | olive1 | 7 |
| 1BK17974103501 | 868693991 | 942760511 | 4.1.361 | 11 | 1051-67.89306451 | 55.2341281 | -51.10791 | -31.074751 | 2020-05-19|2020-08-23 18:38:... | 2971 | 631 | INRI | olive1 | 7 |
| 48-73-15791 | 31 | 31 | 21 | 1051-13.7078871 | 113.4999431 | 54.38129151 | -18.4377511 | 2020-03-24|2020-05-19 11:16:... | 9321 | 321 | INRI | whitel | 5 |
| 1BK57882463251 | 582308371 | 454572271 | 2.4.271 | 21 | 1051-13.7078871 | 113.4999431 | 54.38129151 | -18.4377511 | 2020-03-24|2020-05-19 11:16:... | 9321 | 321 | INRI | whitel | 5 |
| 58-80-63461 | 31 | 31 | 21 | 1051-6..0.014611-114.6497891 | 22.84495051 | 70.1378271 | 2020-08-03|2020-03-24 08:25:... | 2601 | 71 | INRI | blue1 | 0 |
| 1BK83427032551 | 842325101 | 864946811 | 4.1.341 | 31 | 1051-6..0.014611-114.6497891 | 22.84495051 | 70.1378271 | 2020-08-03|2020-03-24 08:25:... | 2601 | 71 | INRI | blue1 | 0 |
| 68-72-62361 | 31 | 31 | 21 | 1051-18.9100341 | -70.1931031 | -10.1829211 | 173.8772131 | 2020-07-17|2020-04-30 04:54:... | 9071 | 531 | INRI | purple1 | 1 |
| 1BK83427032551 | 119810421 | 358626581 | 2.4.359 | 31 | 1051-18.9100341 | -70.1931031 | -10.1829211 | 173.8772131 | 2020-07-17|2020-04-30 04:54:... | 9071 | 531 | INRI | purple1 | 1 |
| 02-10-56391 | 31 | 31 | 21 | 1051-18.9100341 | -70.1931031 | -10.1829211 | 173.8772131 | 2020-07-17|2020-04-30 04:54:... | 9071 | 531 | INRI | purple1 | 1 |
| 1BK45293558541 | 600718781 | 788223601 | 2.1.91 | 31 | 1051-1.2152741 | -56.0149031 | 35.1528761 | 104.3249051 | 2020-01-02|2020-02-16 04:28:... | 5471 | 171 | INRI | teal1 | 8 |
| 66-83-43491 | 21 | 31 | 41 | 1051-1.2152741 | -56.0149031 | 35.1528761 | 104.3249051 | 2020-01-02|2020-02-16 04:28:... | 5471 | 171 | INRI | teal1 | 8 |
| 1BK97200882191 | 143273121 | 944270671 | 3.1.21 | 31 | 1051-173.3622561 | 65.01212651 | 51.3907511 | 2020-04-10|2020-01-20 21:17:... | 2591 | 331 | INRI | maroon1 | 5 |
| 72-73-65261 | 31 | 31 | 21 | 1051-173.3622561 | 65.01212651 | 51.3907511 | 2020-04-10|2020-01-20 21:17:... | 2591 | 331 | INRI | maroon1 | 5 |
| 1BK741573326071 | 464072101 | 431600031 | 1.3.41 | 31 | 1051-16.4822251 | 7.61260151-156.4285771 | 2020-06-09|2020-03-19 01:53:... | 7871 | 211 | INRI | olive1 | 6 |
| 07-78-65261 | 31 | 31 | 21 | 1051-16.4822251 | 7.61260151-156.4285771 | 2020-06-09|2020-03-19 01:53:... | 7871 | 211 | INRI | olive1 | 6 |
| 1BK50148714331 | 658615731 | 647086181 | 1.3.281 | 31 | 1051-29.5653261 | 64.8437091 | 84.0681091 | -49.8208351 | 2020-08-14|2020-06-03 09:39:... | 5861 | 51 | INRI | fuchsia1 | 2 |
| 55-52-56541 | 51 | 51 | 11 | 1051-29.5653261 | 64.8437091 | 84.0681091 | -49.8208351 | 2020-08-14|2020-06-03 09:39:... | 5861 | 51 | INRI | fuchsia1 | 2 |
| 1BK909514887361 | 377217581 | 272977701 | 2.3.131 | 31 | 1051-61.93646051 | 83.2497051 | 0.02818951 | 115.4609091 | 2020-04-07|2020-09-29 10:51:... | 9121 | 801 | INRI | aqua1 | 7 |
| 39-09-95691 | 21 | 31 | 21 | 1051-61.93646051 | 83.2497051 | 0.02818951 | 115.4609091 | 2020-04-07|2020-09-29 10:51:... | 9121 | 801 | INRI | aqua1 | 7 |
| 1BK2437623191 | 625529691 | 458774571 | 3.3.91 | 31 | 1051-62.65151551-139.1540281 | 28.02999951 | -62.85561 | 2020-07-01|2020-09-30 17:40:... | 8211 | 231 | INRI | black1 | 5 |
| 90-44-66131 | 21 | 31 | 41 | 1051-62.65151551-139.1540281 | 28.02999951 | -62.85561 | 2020-07-01|2020-09-30 17:40:... | 8211 | 231 | INRI | black1 | 5 |
| 1BK97200882191 | 568019611 | 534017071 | 4.2.341 | 31 | 1051-5..0.0048391 | 25.0165911 | 70.4713581 | 2020-05-03|2020-06-08 09:11:... | 711 | 101 | INRI | fuchsia1 | 4 |
| 52-04-66881 | 51 | 51 | 21 | 1051-5..0.0048391 | 25.0165911 | 70.4713581 | 2020-05-03|2020-06-08 09:11:... | 711 | 101 | INRI | fuchsia1 | 4 |
| 1BK97932842531 | 669097211 | 405095541 | 2.2.221 | 31 | 1051-36.19131551 | 5.6862641 | 88.9883931 | 36.5809591 | 2020-03-05|2020-05-29 13:36:... | 261 | 811 | INRI | black1 | 6 |
| 00-17-79431 | 31 | 31 | 31 | 1051-36.19131551 | 5.6862641 | 88.9883931 | 36.5809591 | 2020-03-05|2020-05-29 13:36:... | 261 | 811 | INRI | black1 | 6 |
| 1BK28800213801 | 501635551 | 340854201 | 3.4.231 | 31 | 1051-83.065991 | 106.2686891 | 8.8308851 | 74.8723521 | 2020-01-15|2020-05-12 21:53:... | 5711 | 991 | INRI | navy1 | 5 |
| 06-09-49811 | 11 | 51 | 31 | 1051-83.065991 | 106.2686891 | 8.8308851 | 74.8723521 | 2020-01-15|2020-05-12 21:53:... | 5711 | 991 | INRI | navy1 | 5 |
| 1BK45374260431 | 911117541 | 592507691 | 1.2.191 | 31 | 1051-118.884351 | -99.9357191 | -3.70262251 | -46.8287161 | 2020-04-28|2020-02-12 11:31:... | 6501 | 811 | INRI | whitel | 3 |
| 62-35-88954 | 51 | 51 | 21 | 1051-118.884351 | -99.9357191 | -3.70262251 | -46.8287161 | 2020-04-28|2020-02-12 11:31:... | 6501 | 811 | INRI | whitel | 3 |
| 1BK97200882191 | 678753571 | 14562561 | 3.1.151 | 31 | 1051-10..8619591-111.9898531 | 57.2331211 | 95.4699861 | 2020-01-25|2020-08-04 09:42:... | 5901 | 31 | INRI | teal1 | 3 |
| 59-51-93621 | 31 | 31 | 41 | 1051-10..8619591-111.9898531 | 57.2331211 | 95.4699861 | 2020-01-25|2020-08-04 09:42:... | 5901 | 31 | INRI | teal1 | 3 |
| 1BK56453237301 | 184429931 | 840399461 | 3.1.291 | 31 | 1051-81.4722351 | -88.4049161 | 12.6908181 | -140.997681 | 2020-09-24|2020-07-16 05:12:... | 5151 | 11 | INRI | blue1 | 0 |
| 24-35-87711 | 31 | 31 | 41 | 1051-81.4722351 | -88.4049161 | 12.6908181 | -140.997681 | 2020-09-24|2020-07-16 05:12:... | 5151 | 11 | INRI | blue1 | 0 |
| 1BK61636084131 | 365917781 | 119462101 | 4.2.281 | 31 | 1051-60.20363851 | 120.9885011 | 32.1032631 | -50.5518891 | 2020-07-26|2020-04-23 06:57:... | 8101 | 581 | INRI | silver1 | 8 |
| 33-16-13781 | 31 | 51 | 11 | 1051-60.20363851 | 120.9885011 | 32.1032631 | -50.5518891 | 2020-07-26|2020-04-23 06:57:... | 8101 | 581 | INRI | silver1 | 8 |
| 1BK68033736491 | 283823061 | 972226761 | 2.1.181 | 31 | 1051-6..5.6400561 | 161.0839981 | -12.9435021-148.2326211 | 2020-09-20|2020-09-17 03:13:... | 9271 | 741 | INRI | maroon1 | 7 |
| 47-70-55571 | 21 | 21 | 41 | 1051-6..5.6400561 | 161.0839981 | -12.9435021-148.2326211 | 2020-09-20|2020-09-17 03:13:... | 9271 | 741 | INRI | maroon1 | 7 |
+-----+
only showing top 20 rows
  
```

22/07/24 20:58:43 INFO ContextCleaner: Cleaned accumulator 85
 22/07/24 20:58:43 INFO ContextCleaner: Cleaned accumulator 130

6. Aggregation on pickup date.

```
agg_df=new_df.groupBy("pickup_date").count().orderBy("pickup_date")
```

```
agg_df.show()
```

```

22/07/24 20:58:44 INFO DAGScheduler: ResultStage 9 (showString at NativeMethodAccessorImpl.java:0) finished in 0.000 ms
22/07/24 20:58:44 INFO DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0, took 0.000 ms
22/07/24 20:58:44 INFO CodeGenerator: Code generated in 19.87628 ms
+-----+
| pickup_date|count|
+-----+
| 2020-01-01 | 11 |
| 2020-01-02 | 31 |
| 2020-01-03 | 21 |
| 2020-01-04 | 21 |
| 2020-01-05 | 21 |
| 2020-01-06 | 31 |
| 2020-01-07 | 21 |
| 2020-01-08 | 41 |
| 2020-01-09 | 21 |
| 2020-01-10 | 21 |
| 2020-01-11 | 31 |
| 2020-01-12 | 31 |
| 2020-01-14 | 21 |
| 2020-01-15 | 51 |
| 2020-01-16 | 31 |
| 2020-01-17 | 41 |
| 2020-01-18 | 41 |
| 2020-01-20 | 41 |
| 2020-01-21 | 11 |
| 2020-01-23 | 41 |
+-----+
only showing top 20 rows
  
```

22/07/24 20:58:44 INFO FileSourceStrategy: Pruning directories with:
 22/07/24 20:58:44 INFO FileSourceStrategy: Post-Scan Filters:
 22/07/24 20:58:44 INFO FileSourceStrategy: Output Data Schema: struct<_c9: string>
 22/07/24 20:58:44 INFO FileSourceScanExec: Pushed Filters:
 22/07/24 20:58:44 INFO FileSourceScanExec: Pushed Filters:

7. Command to move above grouped data into HDFS.

```
agg_df.coalesce(1).write.format('csv').mode('overwrite').save('/user/hadoop/datewise_bookings_'
agg',header='true')
```

```
[hadoop@ip-10-0-0-52 ~]$ hadoop fs -ls datewise_bookings_agg
Found 2 items
-rw-r--r-- 1 hadoop hdfsadmingroup          0 2022-07-24 20:58 datewise_bookings_agg/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmingroup 3776 2022-07-24 20:58 datewise_bookings_agg/part-00000-8bff7fc8-bf32-4cbd-bc83-5b82c09640cc-c000.csv
[hadoop@ip-10-0-0-52 ~]$ █
```

8. Copy file to local

```
hadoop fs -get /user/hadoop/datewise_bookings_agg/part-00000-8bff7fc8-bf32-4cbd-bc83-
5b82c09640cc-c000.csv /home/hadoop/
```

Task 4:

Steps:

First, we must create database with capstone1 name. The tables which I have created for all three tasks are clickstream_data, booking_data, aggregate_datewise. Here the data is stored in csv format, so while creating tables, we have to use fields terminated by comma (,) .

Create a Hive-managed table for clickstream data.

```
create table if not exists clickstream_data (
```

```
    customer_id int,  
    app_version string,  
    os_version string,  
    lat double,  
    lon double,  
    page_id varchar(100),  
    button_id varchar(100),  
    is_button_click string,  
    is_page_view string,  
    is_scroll_up string,  
    is_scroll_down string,  
    timestamp timestamp )
```

```
ROW FORMAT DELIMITED
```

```
FIELDS TERMINATED BY ','
```

```
stored as textfile;
```

```
load data local inpath '/hadoop/clickstream/clickstream_flattened.csv' into table  
clickstream_data;
```

Create a Hive-managed table for bookings data.

```
create table if not exists booking_data (
```

```
    booking_id string,  
    customer_id int,  
    driver_id int,  
    customer_app_version string,  
    customer_phone_os_version string,  
    pickup_lat double,  
    pickup_lon double,  
    drop_lat double,  
    drop_lon double,  
    pickup_timestamp timestamp,  
    drop_timestamp timestamp,  
    trip_fare int,  
    tip_amount int,
```

```
currency_code string,  
cab_color string,  
cab_registration_no int,  
customer_rating_by_driver varchar(100),  
rating_by_customer int,  
passenger_count int)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
stored as textfile;  
  
load data local inpath '/hadoop/booking_data/part-m-00000' into table booking_data;
```

Create a Hive-managed table for aggregated data in Task 3.

```
create table if not exists aggregate_datewise(  
    pickup_date date,  
    booking_id_count int  
)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
stored as textfile;  
  
load data local inpath '/hadoop/agg_datewise' into table aggregate_datewise;
```