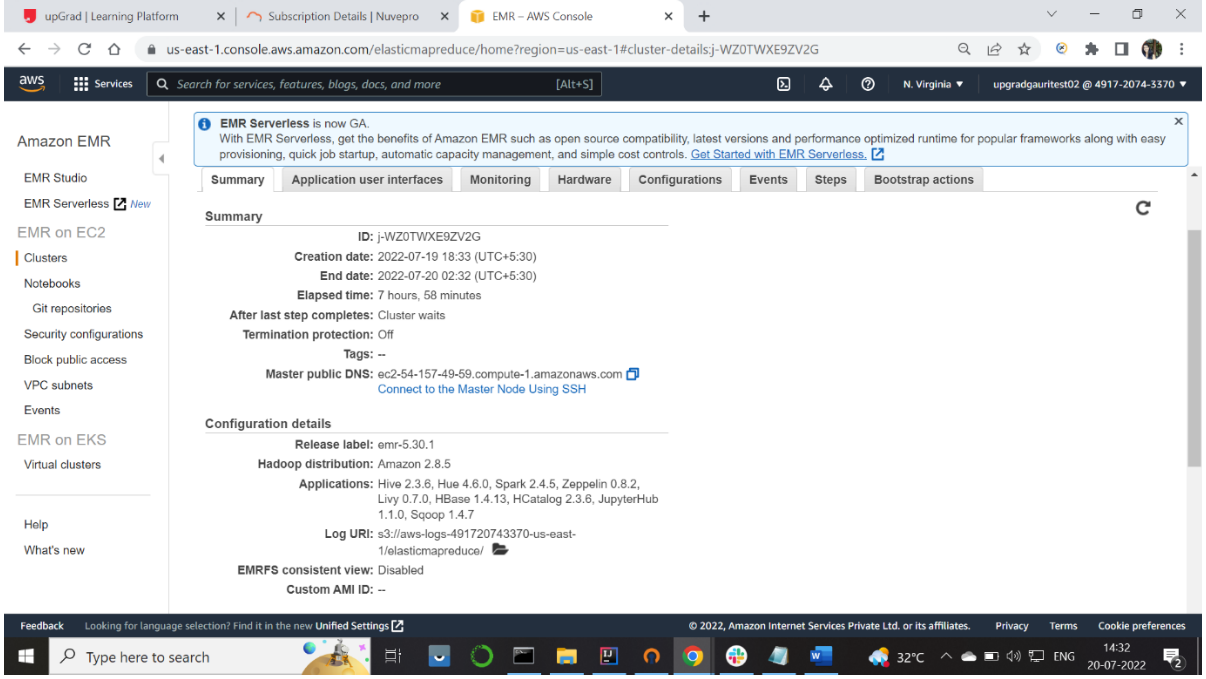
Logic For First Submission

For the capstone project, I have created EMR Cluster with applications as shown in the screenshot. Also configured YARN parameters for this cluster and login with hadoop user in PUTTY terminal


upGrad | Learning Platform
Subscription Details | Nuvepro
EMR – AWS Console
us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-WZ0TWXE9ZV2G
Services
Search for services, features, blogs, docs, and more [Alt+S]
N. Virginia
upgradgauritest02 @ 4917-2074-3370
Amazon EMR
EMR Studio
EMR Serverless New
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
What's new
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
32°C
ENG
14:32
20-07-2022

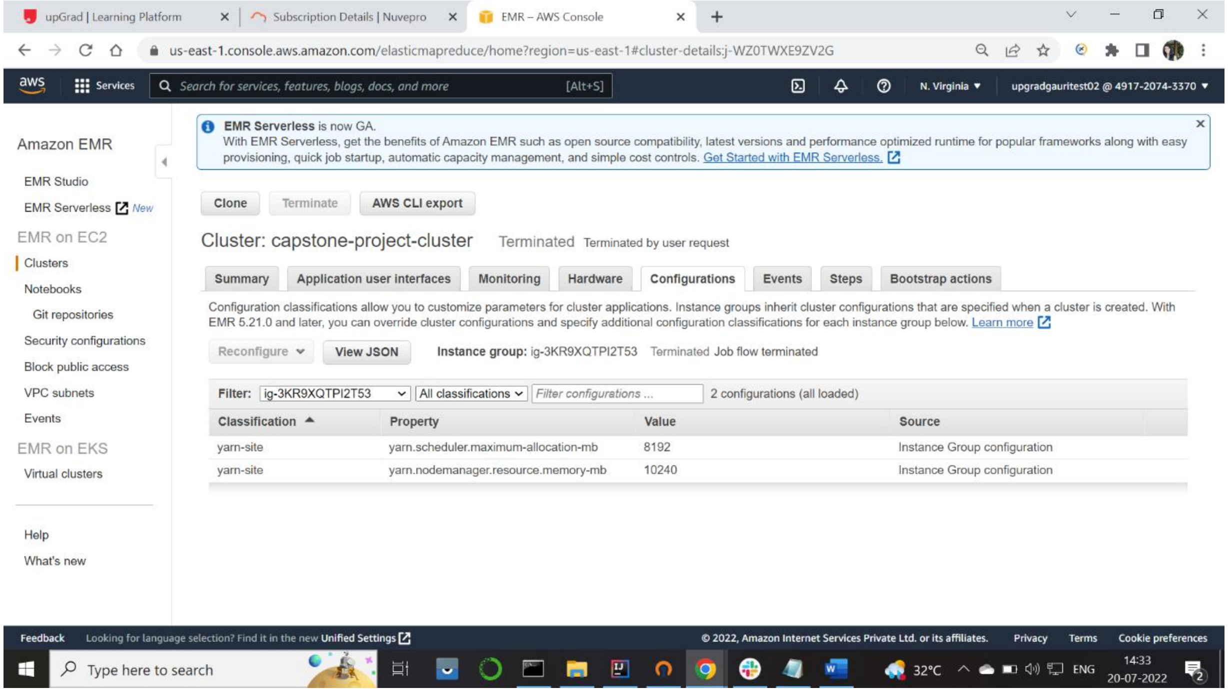
EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Summary

ID: j-WZ0TWXE9ZV2G
Creation date: 2022-07-19 18:33 (UTC+5:30)
End date: 2022-07-20 02:32 (UTC+5:30)
Elapsed time: 7 hours, 58 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-54-157-49-59.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.30.1
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Hue 4.6.0, Spark 2.4.5, Zeppelin 0.8.2, Livy 0.7.0, HBase 1.4.13, HCatalog 2.3.6, JupyterHub 1.1.0, Sqoop 1.4.7
Log URI: s3://aws-logs-491720743370-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --


upGrad | Learning Platform
Subscription Details | Nuvepro
EMR – AWS Console
us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-WZ0TWXE9ZV2G
Services
Search for services, features, blogs, docs, and more [Alt+S]
N. Virginia
upgradgauritest02 @ 4917-2074-3370
Amazon EMR
EMR Studio
EMR Serverless New
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
What's new
Clone
Terminate
AWS CLI export
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Reconfigure
View JSON
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
32°C
ENG
14:33
20-07-2022

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Cluster: capstone-project-cluster Terminated Terminated by user request

Configuration classifications allow you to customize parameters for cluster applications. Instance groups inherit cluster configurations that are specified when a cluster is created. With EMR 5.21.0 and later, you can override cluster configurations and specify additional configuration classifications for each instance group below. Learn more

Instance group: ig-3KR9XQTPI2T53 Terminated Job flow terminated

Filter: ig-3KR9XQTPI2T53 | All classifications | Filter configurations ... 2 configurations (all loaded)

Classification	Property	Value	Source
yarn-site	yarn.scheduler.maximum-allocation-mb	8192	Instance Group configuration
yarn-site	yarn.nodemanager.resource.memory-mb	10240	Instance Group configuration

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

- **Task 1**: To write a job to consume clickstream data from Kafka and ingest to Hadoop.

1. Created a **spark_kafka_to_local.py** file and imported necessary libraries

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

2. Established spark connection

```
spark =
SparkSession.builder.appName("KafkaRead").getOrCreate()
spark.sparkContext.setLogLevel('ERROR')
```

3. Read data from kafka server and topic given

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

4. Casted raw data as string

```
kafkaDF = lines.selectExpr("cast(key as string)","cast(value as string)")
```

5. Wrote Kafka data into json file

```
output = kafkaDF \
        .writeStream \
        .outputMode("append") \
        .format("json") \
        .option("truncate", "false") \
        .option("path","/user/hadoop/clickStreamData/") \
        .option("checkpointLocation", "/user/hadoop/clickstream_checkpoint/") \
        .start()

output.awaitTermination()
```

6. Logged in to the EMR instance and below command is executed to download Spark-SQL-Kafka jar file. This jar is used to run the Spark Streaming-Kafka codes

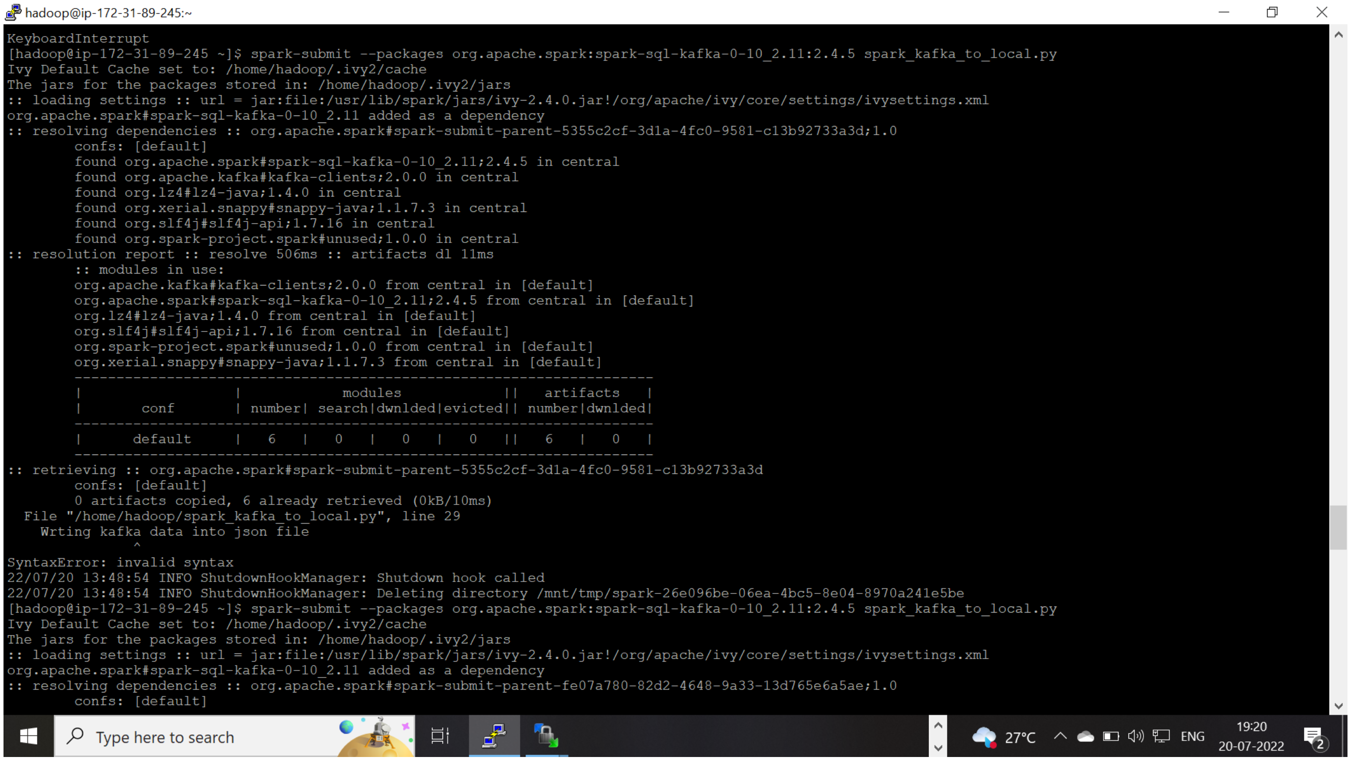
```
wget https://ds-spark-sql-kafka-jar.s3.amazonaws.com/spark-sql-kafka-0- 10_2.11-2.3.0.jar
```

7. Kafka version is set using the following command:

```
export SPARK_KAFKA_VERSION=0.10
```

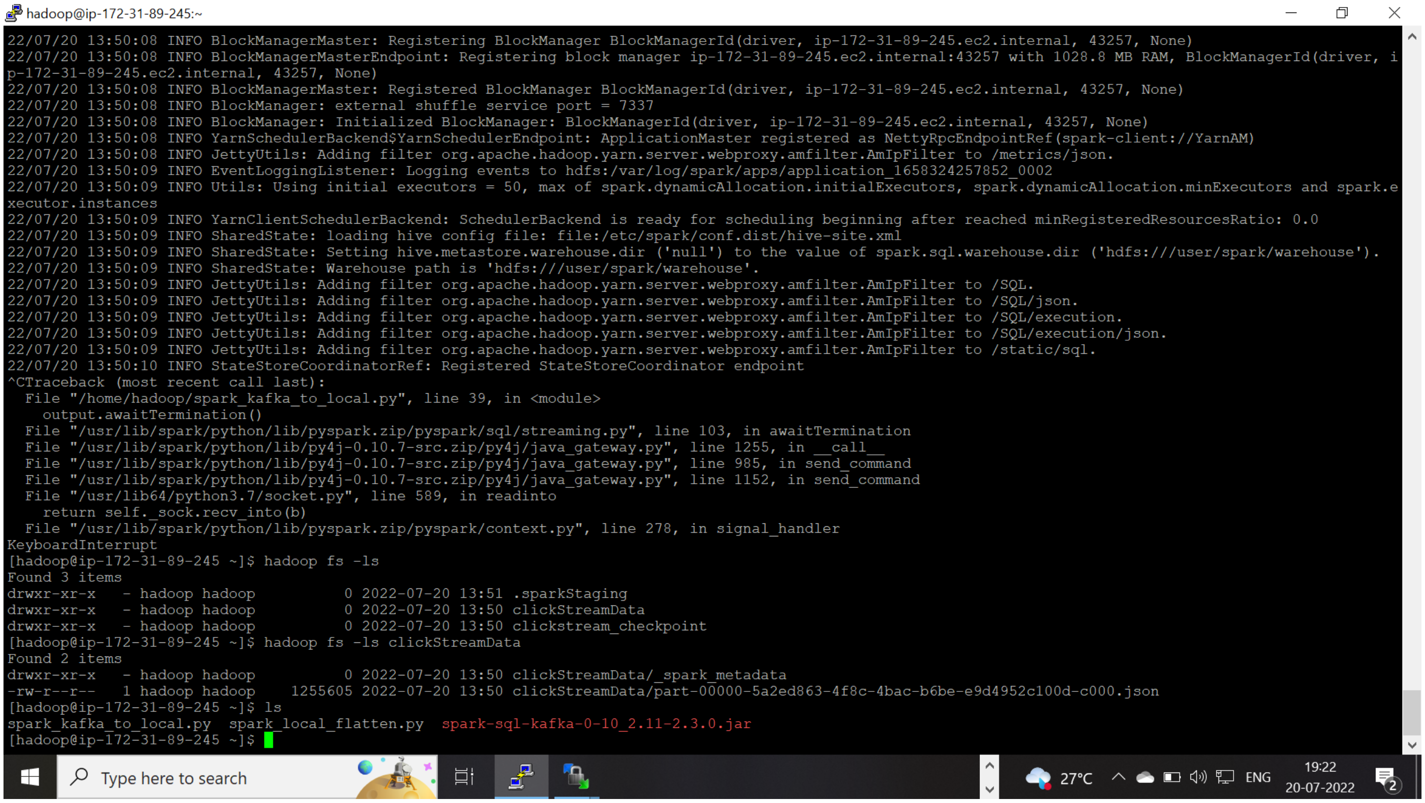
8. Submitted the spark job using command below:

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_kafka_to_local.py
```



```
KeyboardInterrupt
[hadoop@ip-172-31-89-245 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_kafka_to_local.py
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-5355c2cf-3d1a-4fc0-9581-c13b92733a3d;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 in central
        found org.apache.kafka#kafka-clients;2.0.0 in central
        found org.lz4#lz4-java;1.4.0 in central
        found org.xerial.snappy#snappy-java;1.1.7.3 in central
        found org.slf4j#slf4j-api;1.7.16 in central
        found org.spark-project.spark#unused;1.0.0 in central
:: resolution report :: resolve 506ms :: artifacts dl 11ms
        :: modules in use:
        org.apache.kafka#kafka-clients;2.0.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 from central in [default]
        org.lz4#lz4-java;1.4.0 from central in [default]
        org.slf4j#slf4j-api;1.7.16 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.7.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   6   |   0   |   0   |   0   ||   6   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-5355c2cf-3d1a-4fc0-9581-c13b92733a3d
        confs: [default]
        0 artifacts copied, 6 already retrieved (0kB/10ms)
  File "/home/hadoop/spark_kafka_to_local.py", line 29
    Wrting kafka data into json file
            ^
SyntaxError: invalid syntax
22/07/20 13:48:54 INFO ShutdownHookManager: Shutdown hook called
22/07/20 13:48:54 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-26e096be-06ea-4bc5-8e04-8970a241e5be
[hadoop@ip-172-31-89-245 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_kafka_to_local.py
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-fe07a780-82d2-4648-9a33-13d765e6a5ae;1.0
        confs: [default]
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved



```
22/07/20 13:50:08 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-89-245.ec2.internal, 43257, None)
22/07/20 13:50:08 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-89-245.ec2.internal:43257 with 1028.8 MB RAM, BlockManagerId(driver, i
p-172-31-89-245.ec2.internal, 43257, None)
22/07/20 13:50:08 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-89-245.ec2.internal, 43257, None)
22/07/20 13:50:08 INFO BlockManager: external shuffle service port = 7337
22/07/20 13:50:08 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-89-245.ec2.internal, 43257, None)
22/07/20 13:50:08 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
22/07/20 13:50:08 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /metrics/json.
22/07/20 13:50:09 INFO EventLoggingListener: Logging events to hdfs:/var/log/spark/apps/application_1658324257852_0002
22/07/20 13:50:09 INFO Utils: Using initial executors = 50, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.e
xecutor.instances
22/07/20 13:50:09 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
22/07/20 13:50:09 INFO SharedState: loading hive config file: file:/etc/spark/conf.dist/hive-site.xml
22/07/20 13:50:09 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('hdfs:///user/spark/warehouse').
22/07/20 13:50:09 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
22/07/20 13:50:09 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL.
22/07/20 13:50:09 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/json.
22/07/20 13:50:09 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution.
22/07/20 13:50:09 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution/json.
22/07/20 13:50:09 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
22/07/20 13:50:10 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
^CTraceback (most recent call last):
  File "/home/hadoop/spark_kafka_to_local.py", line 39, in <module>
    output.awaitTermination()
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/streaming.py", line 103, in awaitTermination
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1255, in __call__
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 985, in send_command
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1152, in send_command
  File "/usr/lib64/python3.7/socket.py", line 589, in readinto
    return self._sock.recv_into(b)
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/context.py", line 278, in signal_handler
KeyboardInterrupt
[hadoop@ip-172-31-89-245 ~]$ hadoop fs -ls
Found 3 items
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:51 .sparkStaging
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:50 clickStreamData
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:50 clickstream_checkpoint
[hadoop@ip-172-31-89-245 ~]$ hadoop fs -ls clickStreamData
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:50 clickStreamData/_spark_metadata
-rw-r--r--   1 hadoop hadoop    1255605 2022-07-20 13:50 clickStreamData/part-00000-5a2ed863-4f8c-4bac-b6be-e9d4952c100d-c000.json
[hadoop@ip-172-31-89-245 ~]$ ls
spark_kafka_to_local.py  spark_local_flatten.py  spark-sql-kafka-0-10_2.11-2.3.0.jar
[hadoop@ip-172-31-89-245 ~]$ 
```

9. The data extracted from Kafka was in nested json format. Hence wrote a pyspark job in **spark_local_flatten.py** file to flatten the data and load into Hadoop
10. Imported necessary libraries

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql.functions import col
from pyspark.sql.types import *
```

11. Established a spark connection

```
spark=SparkSession \
        .builder \
        .appName('transformKafkaData') \
        .master('yarn') \
        .getOrCreate()
```

12. Read extracted data stored in json format

```
df=spark.read.json('/user/hadoop/clickStreamData/')
```

13. Flattened raw data using regexp_replace function and stored the raw data into respective columns in a dataframe

```
flatten_df=df.withColumn("value",
F.split(F.regexp_replace(F.regexp_replace((F.regexp_replace("value",'\{|}',"")
),'\:',','),'\"|"',"").cast("string"),','))\
.withColumn("customer_id", F.element_at("value",2))\
withColumn("app_version", F.element_at("value",4))\
.withColumn("OS_version",F.element_at("value",6))\
.withColumn("lat",F.element_at("value",8))\
.withColumn("lon", F.element_at("value",10))\
.withColumn("page_id", F.element_at("value",12))\
.withColumn("button_id",F.element_at("value",14))\
.withColumn("is_button_click",F.element_at("value",16))\
.withColumn("is_page_view",F.element_at("value",18))\
.withColumn("is_scroll_up",F.element_at("value",20))\
.withColumn("is_scroll_down",F.element_at("value",22))\
.withColumn("date_hour",F.element_at("value",24))\
.withColumn("minutes",F.element_at("value",25))\
.withColumn("seconds",F.element_at("value",26))\
.drop("value")
```

14. Concatenated **date_hour, minutes and seconds** column to make it into timestamp format:

```
flatten_df=flatten_df.select("*",F.concat(col("date_hour"),F.lit(":")
,col("minutes"),F.lit(":"),col("seconds")).alias("timestamp"))
```

15. Removed extra character \n from timestamp column to make the data more structured

```
flatten_df =
flatten_df.select("*").withColumn("timestamp",F.expr("substring(timestamp, 1,
length(timestamp)-2)")).drop("date_hour").drop("minutes").drop("seconds")
```

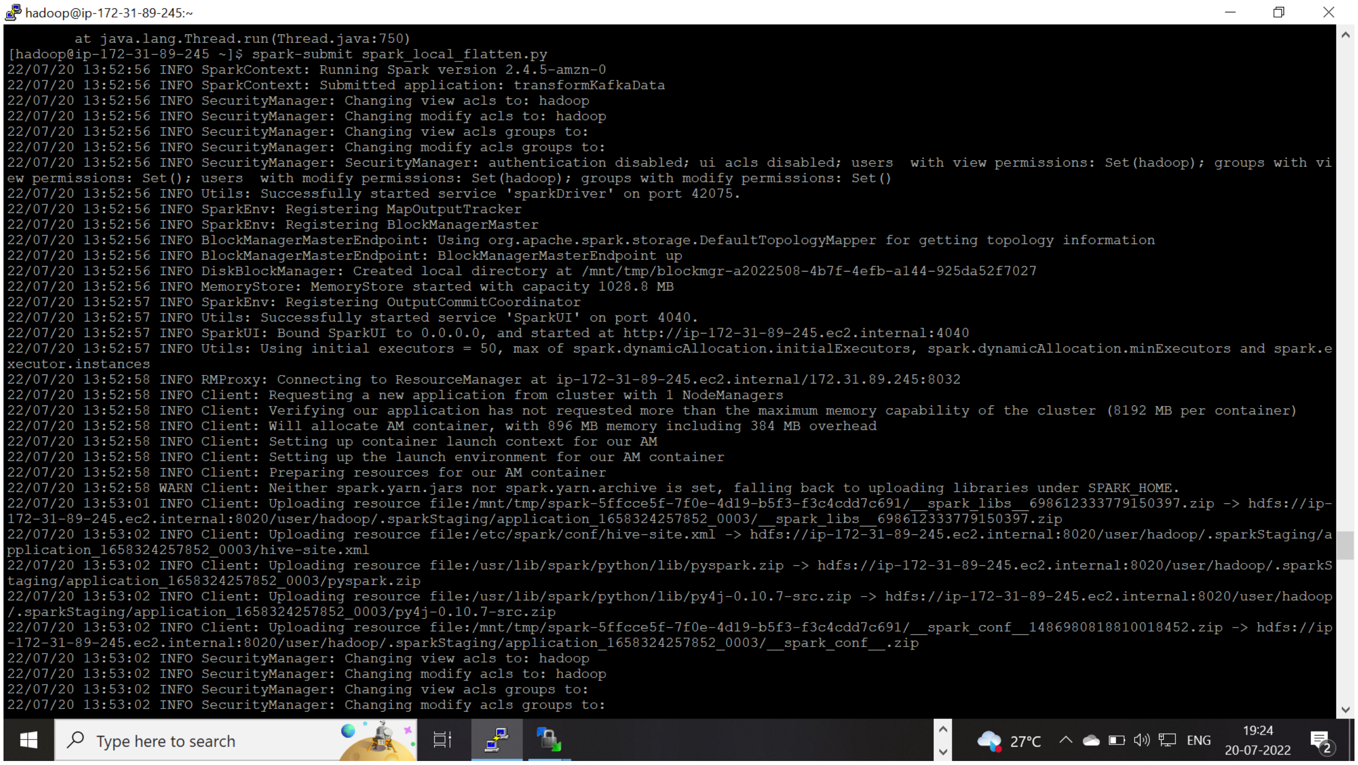
16. Wrote the flattened dataframe in csv file

```
flatten_df.write.option("header","true").csv('/user/hadoop/
clickStream_flatten_data/')
```

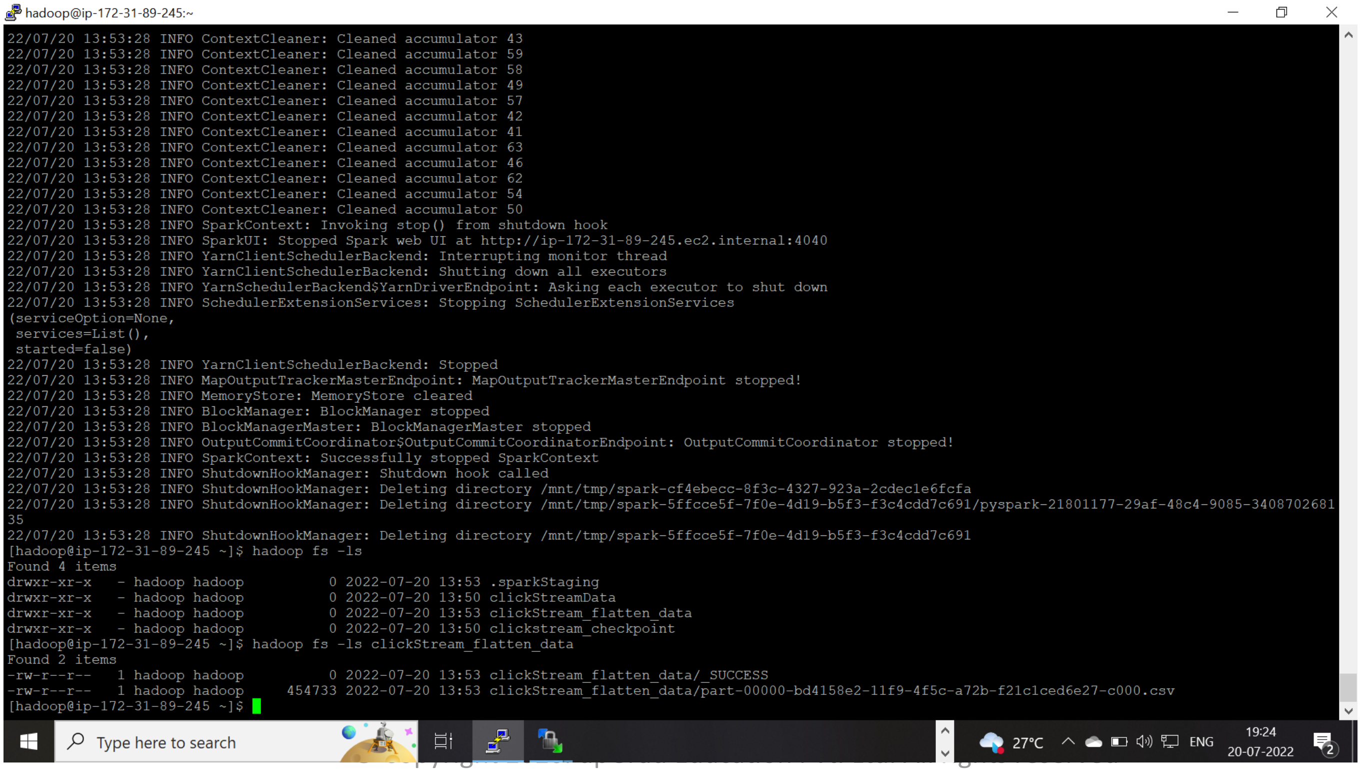
© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

17. Executed spark_local_flatten.py file using command:

```
spark-submit spark_local_flatten.py
```

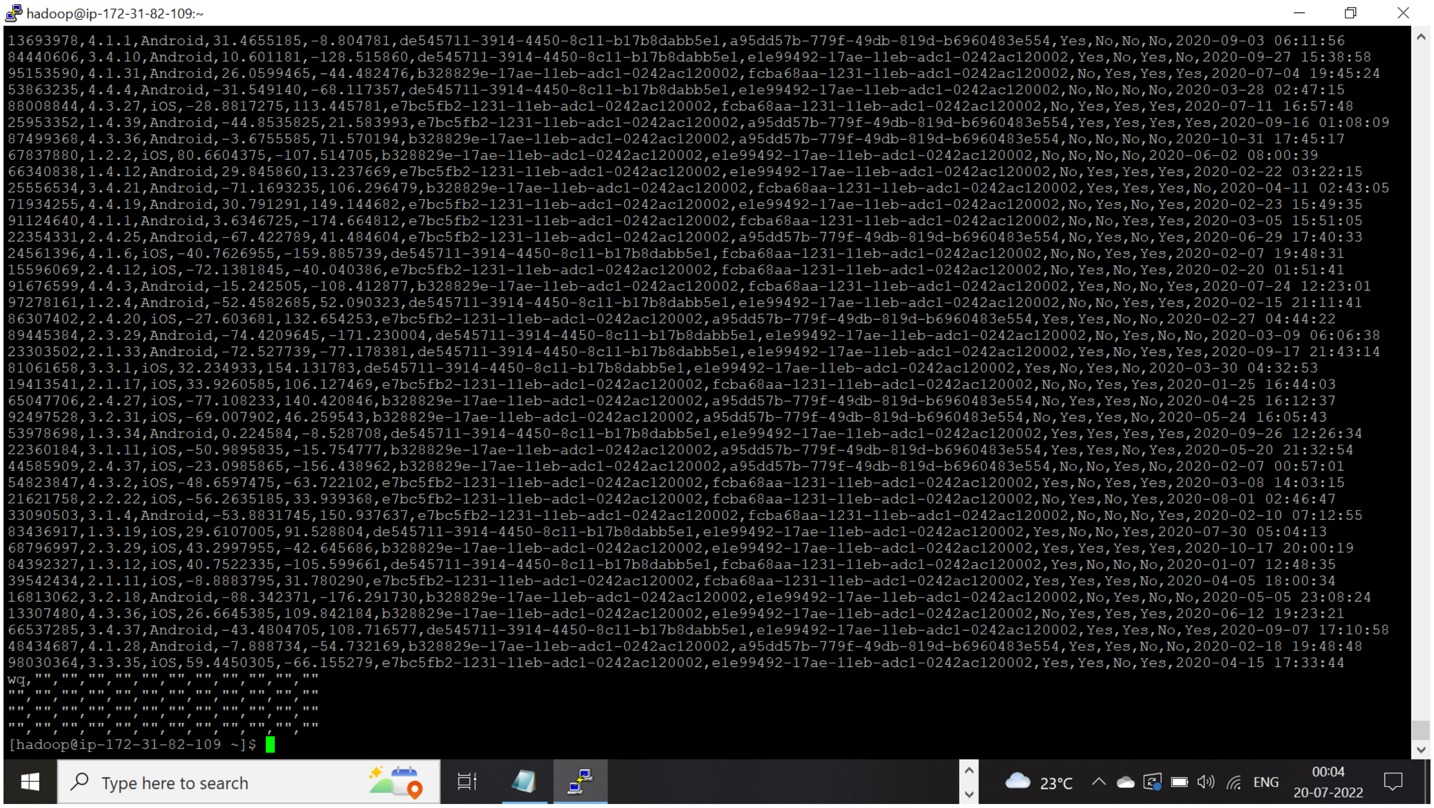


```
        at java.lang.Thread.run(Thread.java:750)
[hadoop@ip-172-31-89-245 ~]$ spark-submit spark_local_flatten.py
22/07/20 13:52:56 INFO SparkContext: Running Spark version 2.4.5-amzn-0
22/07/20 13:52:56 INFO SparkContext: Submitted application: transformKafkaData
22/07/20 13:52:56 INFO SecurityManager: Changing view acls to: hadoop
22/07/20 13:52:56 INFO SecurityManager: Changing modify acls to: hadoop
22/07/20 13:52:56 INFO SecurityManager: Changing view acls groups to:
22/07/20 13:52:56 INFO SecurityManager: Changing modify acls groups to:
22/07/20 13:52:56 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with vi
ew permissions: Set(); users  with modify permissions: Set(hadoop); groups with modify permissions: Set()
22/07/20 13:52:56 INFO Utils: Successfully started service 'sparkDriver' on port 42075.
22/07/20 13:52:56 INFO SparkEnv: Registering MapOutputTracker
22/07/20 13:52:56 INFO SparkEnv: Registering BlockManagerMaster
22/07/20 13:52:56 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/07/20 13:52:56 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/07/20 13:52:56 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-a2022508-4b7f-4efb-a144-925da52f7027
22/07/20 13:52:56 INFO MemoryStore: MemoryStore started with capacity 1028.8 MB
22/07/20 13:52:57 INFO SparkEnv: Registering OutputCommitCoordinator
22/07/20 13:52:57 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/07/20 13:52:57 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://ip-172-31-89-245.ec2.internal:4040
22/07/20 13:52:57 INFO Utils: Using initial executors = 50, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.e
xecutor.instances
22/07/20 13:52:58 INFO RMProxy: Connecting to ResourceManager at ip-172-31-89-245.ec2.internal/172.31.89.245:8032
22/07/20 13:52:58 INFO Client: Requesting a new application from cluster with 1 NodeManagers
22/07/20 13:52:58 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (8192 MB per container)
22/07/20 13:52:58 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
22/07/20 13:52:58 INFO Client: Setting up container launch context for our AM
22/07/20 13:52:58 INFO Client: Setting up the launch environment for our AM container
22/07/20 13:52:58 INFO Client: Preparing resources for our AM container
22/07/20 13:52:58 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
22/07/20 13:53:01 INFO Client: Uploading resource file:/mnt/tmp/spark-5ffcce5f-7f0e-4d19-b5f3-f3c4cdd7c691/__spark_libs__6986123337791503970.zip -> hdfs://ip-
172-31-89-245.ec2.internal:8020/user/hadoop/.sparkStaging/application_1658324257852_0003/__spark_libs__6986123337791503970.zip
22/07/20 13:53:02 INFO Client: Uploading resource file:/etc/spark/conf/hive-site.xml -> hdfs://ip-172-31-89-245.ec2.internal:8020/user/hadoop/.sparkStaging/a
pplication_1658324257852_0003/hive-site.xml
22/07/20 13:53:02 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-89-245.ec2.internal:8020/user/hadoop/.sparkS
taging/application_1658324257852_0003/pyspark.zip
22/07/20 13:53:02 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.7-src.zip -> hdfs://ip-172-31-89-245.ec2.internal:8020/user/hadoop
/.sparkStaging/application_1658324257852_0003/py4j-0.10.7-src.zip
22/07/20 13:53:02 INFO Client: Uploading resource file:/mnt/tmp/spark-5ffcce5f-7f0e-4d19-b5f3-f3c4cdd7c691/__spark_conf__1486980818810018452.zip -> hdfs://ip
-172-31-89-245.ec2.internal:8020/user/hadoop/.sparkStaging/application_1658324257852_0003/__spark_conf__.zip
22/07/20 13:53:02 INFO SecurityManager: Changing view acls to: hadoop
22/07/20 13:53:02 INFO SecurityManager: Changing modify acls to: hadoop
22/07/20 13:53:02 INFO SecurityManager: Changing view acls groups to:
22/07/20 13:53:02 INFO SecurityManager: Changing modify acls groups to:
```



```
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 43
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 59
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 58
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 49
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 57
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 42
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 41
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 63
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 46
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 62
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 54
22/07/20 13:53:28 INFO ContextCleaner: Cleaned accumulator 50
22/07/20 13:53:28 INFO SparkContext: Invoking stop() from shutdown hook
22/07/20 13:53:28 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-89-245.ec2.internal:4040
22/07/20 13:53:28 INFO YarnClientSchedulerBackend: Interrupting monitor thread
22/07/20 13:53:28 INFO YarnClientSchedulerBackend: Shutting down all executors
22/07/20 13:53:28 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/07/20 13:53:28 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
22/07/20 13:53:28 INFO YarnClientSchedulerBackend: Stopped
22/07/20 13:53:28 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/07/20 13:53:28 INFO MemoryStore: MemoryStore cleared
22/07/20 13:53:28 INFO BlockManager: BlockManager stopped
22/07/20 13:53:28 INFO BlockManagerMaster: BlockManagerMaster stopped
22/07/20 13:53:28 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/07/20 13:53:28 INFO SparkContext: Successfully stopped SparkContext
22/07/20 13:53:28 INFO ShutdownHookManager: Shutdown hook called
22/07/20 13:53:28 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-cf4ebecc-8f3c-4327-923a-2cdec1e6fcfa
22/07/20 13:53:28 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-5ffcce5f-7f0e-4d19-b5f3-f3c4cdd7c691/pyspark-21801177-29af-48c4-9085-3408702681
35
22/07/20 13:53:28 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-5ffcce5f-7f0e-4d19-b5f3-f3c4cdd7c691
[hadoop@ip-172-31-89-245 ~]$ hadoop fs -ls
Found 4 items
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:53 .sparkStaging
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:50 clickStreamData
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:53 clickStream_flatten_data
drwxr-xr-x   - hadoop hadoop          0 2022-07-20 13:50 clickstream_checkpoint
[hadoop@ip-172-31-89-245 ~]$ hadoop fs -ls clickStream_flatten_data
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2022-07-20 13:53 clickStream_flatten_data/_SUCCESS
-rw-r--r--   1 hadoop hadoop     454733 2022-07-20 13:53 clickStream_flatten_data/part-00000-bd4158e2-11f9-4f5c-a72b-f21c1ced6e27-c000.csv
[hadoop@ip-172-31-89-245 ~]$ 
```

18. Screenshot of the flattened data



```
13693978,4.1.1,Android,31.4655185,-8.804781,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,Yes,No,No,No,2020-09-03 06:11:56
84440606,3.4.10,Android,10.601181,-128.515860,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,No,Yes,No,2020-09-27 15:38:58
95153590,4.1.31,Android,26.0599465,-44.482476,b328829e-17ae-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,Yes,Yes,2020-07-04 19:45:24
53863235,4.4.4,Android,-31.549140,-68.117357,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,No,No,No,No,2020-03-28 02:47:15
88008844,4.3.27,iOS,-28.8817275,113.445781,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,Yes,Yes,2020-07-11 16:57:48
25953352,1.4.39,Android,-44.8535825,21.583993,e7bc5fb2-1231-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,Yes,Yes,Yes,Yes,2020-09-16 01:08:09
87499368,4.3.36,Android,-3.6755585,71.570194,b328829e-17ae-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,No,No,No,No,2020-10-31 17:45:17
67837880,1.2.2,iOS,80.6604375,-107.514705,b328829e-17ae-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,No,No,No,2020-06-02 08:00:39
66340838,1.4.12,Android,29.845860,13.237669,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,Yes,Yes,2020-02-22 03:22:15
25556534,3.4.21,Android,-71.1693235,106.296479,b328829e-17ae-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,Yes,Yes,Yes,No,2020-04-11 02:43:05
71934255,4.4.19,Android,30.791291,149.144682,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,No,Yes,2020-02-23 15:49:35
91124640,4.1.1,Android,3.6346725,-174.664812,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,Yes,Yes,2020-03-05 15:51:05
22354331,2.4.25,Android,-67.422789,41.484604,e7bc5fb2-1231-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,No,Yes,No,Yes,2020-06-29 17:40:33
24561396,4.1.6,iOS,-40.7626955,-159.885739,de545711-3914-4450-8c11-b17b8dabb5e1,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,Yes,Yes,2020-02-07 19:48:31
15596069,2.4.12,iOS,-72.1381845,-40.040386,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,No,Yes,2020-02-20 01:51:41
91676599,4.4.3,Android,-15.242505,-108.412877,b328829e-17ae-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,Yes,No,No,Yes,2020-07-24 12:23:01
97278161,1.2.4,Android,-52.4582685,52.090323,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,No,No,Yes,Yes,2020-02-15 21:11:41
86307402,2.4.20,iOS,-27.603681,132.654253,e7bc5fb2-1231-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,Yes,Yes,No,No,2020-02-27 04:44:22
89445384,2.3.29,Android,-74.4209645,-171.230004,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,No,No,2020-03-09 06:06:38
23303502,2.1.33,Android,-72.527739,-77.178381,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,No,Yes,Yes,2020-09-17 21:43:14
81061658,3.3.1,iOS,32.234933,154.131783,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,No,Yes,No,2020-03-30 04:32:53
19413541,2.1.17,iOS,33.9260585,106.127469,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,Yes,Yes,2020-01-25 16:44:03
65047706,2.4.27,iOS,-77.108233,140.420846,b328829e-17ae-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,No,Yes,Yes,No,2020-04-25 16:12:37
92497528,3.2.31,iOS,-69.007902,46.259543,b328829e-17ae-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,No,Yes,Yes,No,2020-05-24 16:05:43
53978698,1.3.34,Android,0.224584,-8.528708,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,Yes,Yes,Yes,2020-09-26 12:26:34
22360184,3.1.11,iOS,-50.9895835,-15.754777,b328829e-17ae-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,Yes,Yes,No,Yes,2020-05-20 21:32:54
44585909,2.4.37,iOS,-23.0985865,-156.438962,b328829e-17ae-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,No,No,Yes,No,2020-02-07 00:57:01
54823847,4.3.2,iOS,-48.6597475,-63.722102,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,Yes,No,Yes,Yes,2020-03-08 14:03:15
21621758,2.2.22,iOS,-56.2635185,33.939368,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,No,Yes,2020-08-01 02:46:47
33090503,3.1.4,Android,-53.8831745,150.937637,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,No,Yes,2020-02-10 07:12:55
83436917,1.3.19,iOS,29.6107005,91.528804,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,No,No,Yes,2020-07-30 05:04:13
68796997,2.3.29,iOS,43.2997955,-42.645686,b328829e-17ae-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,Yes,Yes,Yes,2020-10-17 20:00:19
84392327,1.3.12,iOS,40.7522335,-105.599661,de545711-3914-4450-8c11-b17b8dabb5e1,fcba68aa-1231-11eb-adc1-0242ac120002,Yes,No,No,No,2020-01-07 12:48:35
39542434,2.1.11,iOS,-8.8883795,31.780290,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,Yes,Yes,Yes,No,2020-04-05 18:00:34
16813062,3.2.18,Android,-88.342371,-176.291730,b328829e-17ae-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,No,No,2020-05-05 23:08:24
13307480,4.3.36,iOS,26.6645385,109.842184,b328829e-17ae-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,Yes,Yes,2020-06-12 19:23:21
66537285,3.4.37,Android,-43.4804705,108.716577,de545711-3914-4450-8c11-b17b8dabb5e1,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,Yes,No,Yes,2020-09-07 17:10:58
48434687,4.1.28,Android,-7.888734,-54.732169,b328829e-17ae-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,Yes,Yes,No,No,2020-02-18 19:48:48
98030364,3.3.35,iOS,59.4450305,-66.155279,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,Yes,No,Yes,2020-04-15 17:33:44
wq,"","","","","","","","","","",""
"","","","","","","","","","","",""
"","","","","","","","","","","",""
"","","","","","","","","","","",""
[hadoop@ip-172-31-82-109 ~]$
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

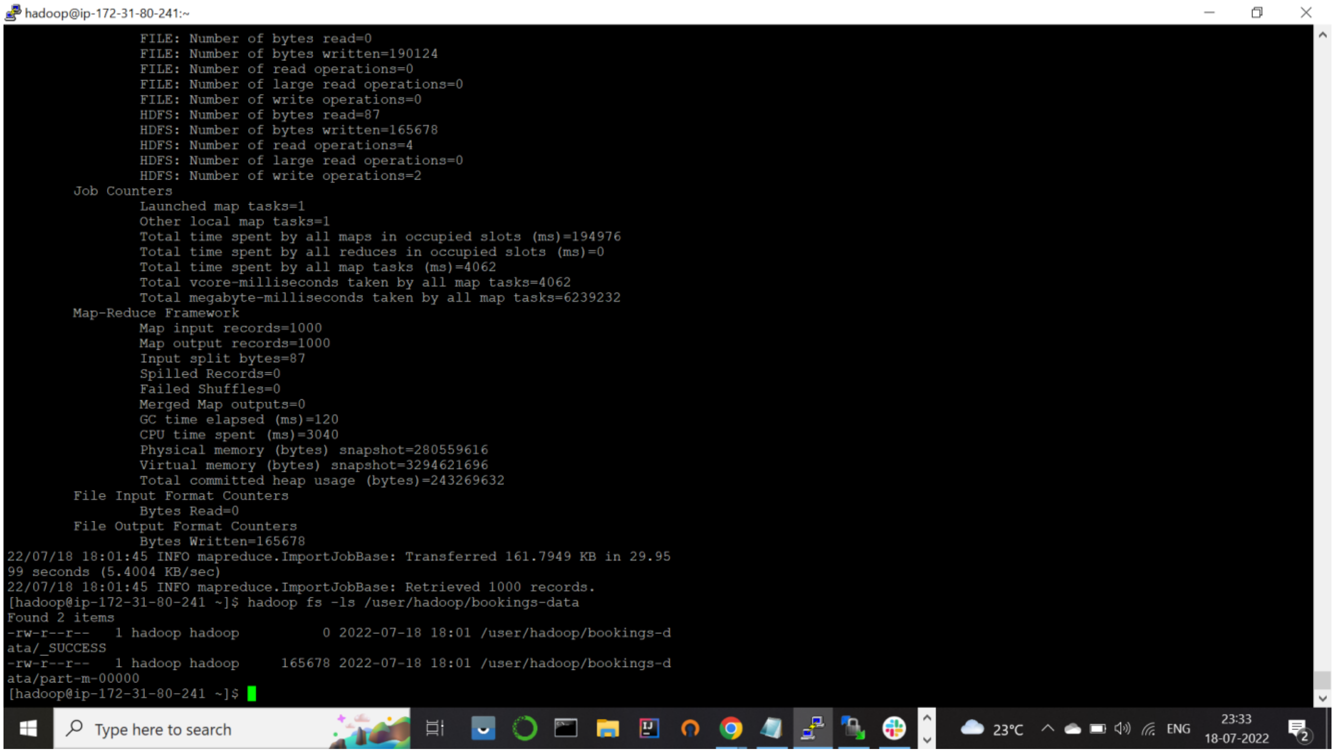
- **Task 2**: To write a script to ingest the relevant bookings data from AWS RDS to Hadoop.

1. For this task, first I installed mysql connector and MySQL on EMR cluster
2. Checked if the directory to load the data is already present in Hadoop

```
hadoop fs -rm -r /user/hadoop/bookings-data
```

3. Imported data from AWS RDS to Hadoop using command:

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

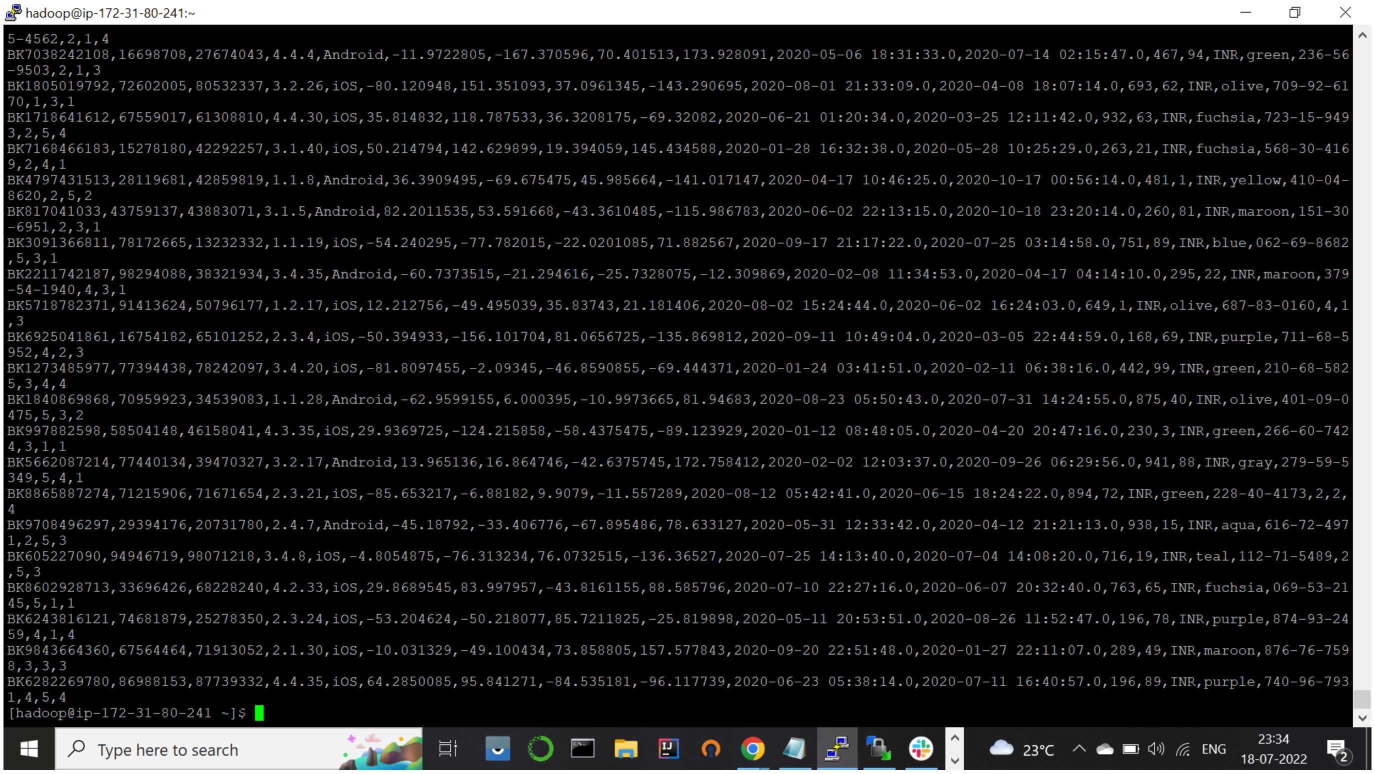


```
                FILE: Number of bytes read=0
                FILE: Number of bytes written=190124
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=165678
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=194976
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4062
                Total vcore-milliseconds taken by all map tasks=4062
                Total megabyte-milliseconds taken by all map tasks=6239232
        Map-Reduce Framework
                Map input records=1000
                Map output records=1000
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=120
                CPU time spent (ms)=3040
                Physical memory (bytes) snapshot=280559616
                Virtual memory (bytes) snapshot=3294621696
                Total committed heap usage (bytes)=243269632
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=165678
22/07/18 18:01:45 INFO mapreduce.ImportJobBase: Transferred 161.7949 KB in 29.95
99 seconds (5.4004 KB/sec)
22/07/18 18:01:45 INFO mapreduce.ImportJobBase: Retrieved 1000 records.
[hadoop@ip-172-31-80-241 ~]$ hadoop fs -ls /user/hadoop/bookings-data
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2022-07-18 18:01 /user/hadoop/bookings-d
ata/_SUCCESS
-rw-r--r--   1 hadoop hadoop     165678 2022-07-18 18:01 /user/hadoop/bookings-d
ata/part-m-00000
[hadoop@ip-172-31-80-241 ~]$
```

4. Viewed imported data in hdfs using command

```
hadoop fs -ls /user/hadoop/bookings-data
```

```
hadoop fs -cat /user/hadoop/bookings-data/part-m-00000
```



```
5-4562,2,1,4
BK7038242108,16698708,27674043,4.4.4,Android,-11.9722805,-167.370596,70.401513,173.928091,2020-05-06 18:31:33.0,2020-07-14 02:15:47.0,467,94,INR,green,236-56
-9503,2,1,3
BK1805019792,72602005,80532337,3.2.26,iOS,-80.120948,151.351093,37.0961345,-143.290695,2020-08-01 21:33:09.0,2020-04-08 18:07:14.0,693,62,INR,olive,709-92-61
70,1,3,1
BK1718641612,67559017,61308810,4.4.30,iOS,35.814832,118.787533,36.3208175,-69.32082,2020-06-21 01:20:34.0,2020-03-25 12:11:42.0,932,63,INR,fuchsia,723-15-949
3,2,5,4
BK7168466183,15278180,42292257,3.1.40,iOS,50.214794,142.629899,19.394059,145.434588,2020-01-28 16:32:38.0,2020-05-28 10:25:29.0,263,21,INR,fuchsia,568-30-416
9,2,4,1
BK4797431513,28119681,42859819,1.1.8,Android,36.3909495,-69.675475,45.985664,-141.017147,2020-04-17 10:46:25.0,2020-10-17 00:56:14.0,481,1,INR,yellow,410-04-
8620,2,5,2
BK817041033,43759137,43883071,3.1.5,Android,82.2011535,53.591668,-43.3610485,-115.986783,2020-06-02 22:13:15.0,2020-10-18 23:20:14.0,260,81,INR,maroon,151-30
-6951,2,3,1
BK3091366811,78172665,13232332,1.1.19,iOS,-54.240295,-77.782015,-22.0201085,71.882567,2020-09-17 21:17:22.0,2020-07-25 03:14:58.0,751,89,INR,blue,062-69-8682
,5,3,1
BK2211742187,98294088,38321934,3.4.35,Android,-60.7373515,-21.294616,-25.7328075,-12.309869,2020-02-08 11:34:53.0,2020-04-17 04:14:10.0,295,22,INR,maroon,379
-54-1940,4,3,1
BK5718782371,91413624,50796177,1.2.17,iOS,12.212756,-49.495039,35.83743,21.181406,2020-08-02 15:24:44.0,2020-06-02 16:24:03.0,649,1,INR,olive,687-83-0160,4,1
,3
BK6925041861,16754182,65101252,2.3.4,iOS,-50.394933,-156.101704,81.0656725,-135.869812,2020-09-11 10:49:04.0,2020-03-05 22:44:59.0,168,69,INR,purple,711-68-5
952,4,2,3
BK1273485977,77394438,78242097,3.4.20,iOS,-81.8097455,-2.09345,-46.8590855,-69.444371,2020-01-24 03:41:51.0,2020-02-11 06:38:16.0,442,99,INR,green,210-68-582
5,3,4,4
BK1840869868,70959923,34539083,1.1.28,Android,-62.9599155,6.000395,-10.9973665,81.94683,2020-08-23 05:50:43.0,2020-07-31 14:24:55.0,875,40,INR,olive,401-09-0
475,5,3,2
BK997882598,58504148,46158041,4.3.35,iOS,29.9369725,-124.215858,-58.4375475,-89.123929,2020-01-12 08:48:05.0,2020-04-20 20:47:16.0,230,3,INR,green,266-60-742
4,3,1,1
BK5662087214,77440134,39470327,3.2.17,Android,13.965136,16.864746,-42.6375745,172.758412,2020-02-02 12:03:37.0,2020-09-26 06:29:56.0,941,88,INR,gray,279-59-5
349,5,4,1
BK8865887274,71215906,71671654,2.3.21,iOS,-85.653217,-6.88182,9.9079,-11.557289,2020-08-12 05:42:41.0,2020-06-15 18:24:22.0,894,72,INR,green,228-40-4173,2,2,
4
BK9708496297,29394176,20731780,2.4.7,Android,-45.18792,-33.406776,-67.895486,78.633127,2020-05-31 12:33:42.0,2020-04-12 21:21:13.0,938,15,INR,aqua,616-72-497
1,2,5,3
BK605227090,94946719,98071218,3.4.8,iOS,-4.8054875,-76.313234,76.0732515,-136.36527,2020-07-25 14:13:40.0,2020-07-04 14:08:20.0,716,19,INR,teal,112-71-5489,2
,5,3
BK8602928713,33696426,68228240,4.2.33,iOS,29.8689545,83.997957,-43.8161155,88.585796,2020-07-10 22:27:16.0,2020-06-07 20:32:40.0,763,65,INR,fuchsia,069-53-21
45,5,1,1
BK6243816121,74681879,25278350,2.3.24,iOS,-53.204624,-50.218077,85.7211825,-25.819898,2020-05-11 20:53:51.0,2020-08-26 11:52:47.0,196,78,INR,purple,874-93-24
59,4,1,4
BK9843664360,67564464,71913052,2.1.30,iOS,-10.031329,-49.100434,73.858805,157.577843,2020-09-20 22:51:48.0,2020-01-27 22:11:07.0,289,49,INR,maroon,876-76-759
8,3,3,3
BK6282269780,86988153,87739332,4.4.35,iOS,64.2850085,95.841271,-84.535181,-96.117739,2020-06-23 05:38:14.0,2020-07-11 16:40:57.0,196,89,INR,purple,740-96-793
1,4,5,4
[hadoop@ip-172-31-80-241 ~]$
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

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

1. For this task, created a **datewise_bookings_aggregates_spark.py** file.
2. Imported necessary libraries

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql.functions import col
from pyspark.sql.types import *
```

3. Established spark connection

```
spark = SparkSession \
        .builder \
        .appName('aggregateBatchData') \
        .master('yarn') \
        .getOrCreate()
```

4. Read data from csv filed extracted from AWS RDS and stored in HDFS

```
df=spark.read.csv('/user/hadoop/bookings-
data/',header=False,inferSchema = True)
```

5. Added column data according to given data

```
new_columns=["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_cus
tomer","passenger_count"]

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

6. Created a new column with date extracted from **pickup_timestamp** column

```
new_df = new_df.withColumn("date", F.to_date(F.col("pickup_timestamp")))
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

7. Datewise bookings aggregate using groupBy function

```
aggregate_df = new_df.groupby('date').count()
```

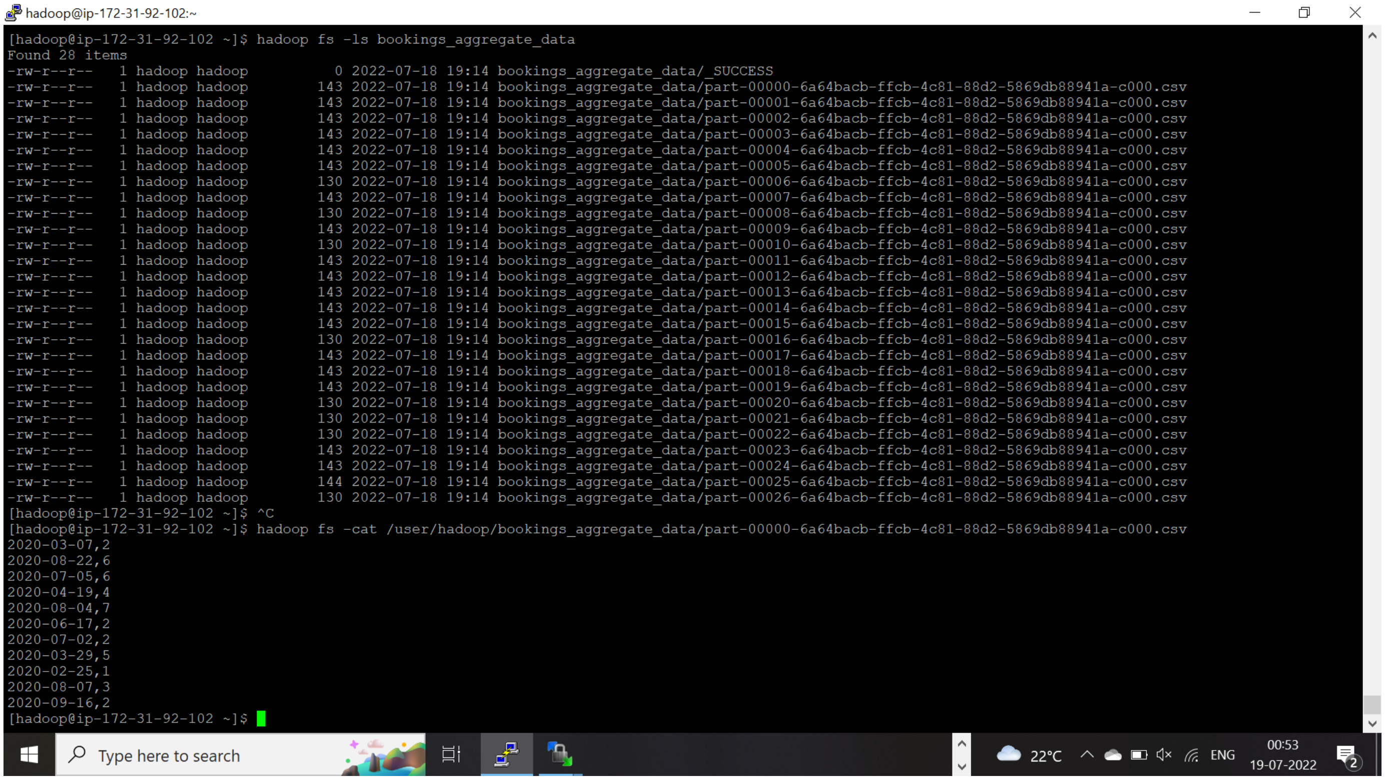
8. Wrote **aggregate_df** dataframe in csv files in HDFS

```
aggregate_df.write.csv('/user/hadoop/bookings_aggregate_data/')
```

9. Executed **datewise_bookings_aggregates_spark.py** using command:

```
spark-submit datewise_bookings_aggregates_spark.py
```

10. Screenshot of the csv files and data



```
[hadoop@ip-172-31-92-102 ~]$ hadoop fs -ls bookings_aggregate_data
Found 28 items
-rw-r--r--   1 hadoop hadoop          0 2022-07-18 19:14 bookings_aggregate_data/_SUCCESS
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00000-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00001-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00002-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00003-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00004-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00005-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00006-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00007-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00008-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00009-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00010-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00011-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00012-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00013-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00014-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00015-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00016-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00017-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00018-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00019-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00020-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00021-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00022-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00023-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00024-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        144 2022-07-18 19:14 bookings_aggregate_data/part-00025-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00026-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
[hadoop@ip-172-31-92-102 ~]$ ^C
[hadoop@ip-172-31-92-102 ~]$ hadoop fs -cat /user/hadoop/bookings_aggregate_data/part-00000-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
2020-03-07,2
2020-08-22,6
2020-07-05,6
2020-04-19,4
2020-08-04,7
2020-06-17,2
2020-07-02,2
2020-03-29,5
2020-02-25,1
2020-08-07,3
2020-09-16,2
[hadoop@ip-172-31-92-102 ~]$
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

- **Task 4**: To create a Hive-managed table for clickstream data, bookings data and aggregated data:

1. Below command is used to launch hive CLI

```
hive
```

2. Database **cab_rides_data** is created using command:

```
create database if not exists cab_rides_data;
```

3. Command to create clickStreamData table:
   As the clickStreamData has column header, I have used command to skip the first row

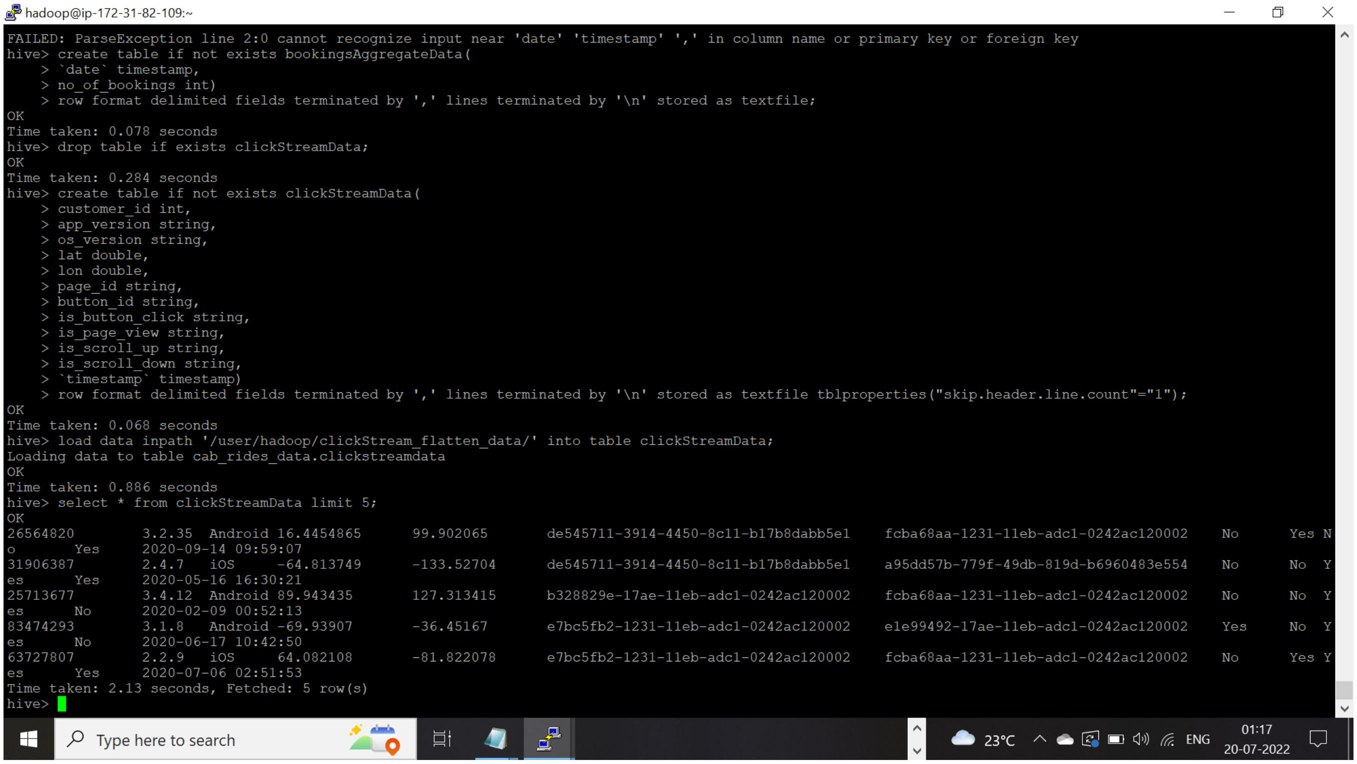
```
create table if not exists clickStreamData(
customer_id int,
app_version string,
os_version string,
lat double,
lon double,
page_id string,
button_id string,
is_button_click string,
is_page_view string,
is_scroll_up string,
is_scroll_down string,
`timestamp` timestamp)
row format delimited fields terminated by ',' lines
terminated by '\n' stored as textfile
tblproperties("skip.header.line.count"="1");
```

4. Command to load data from HDFS to **clickStreamData** table:

```
load data inpath '/user/hadoop/clickStream_flatten_data/' into
table clickStreamData;
```

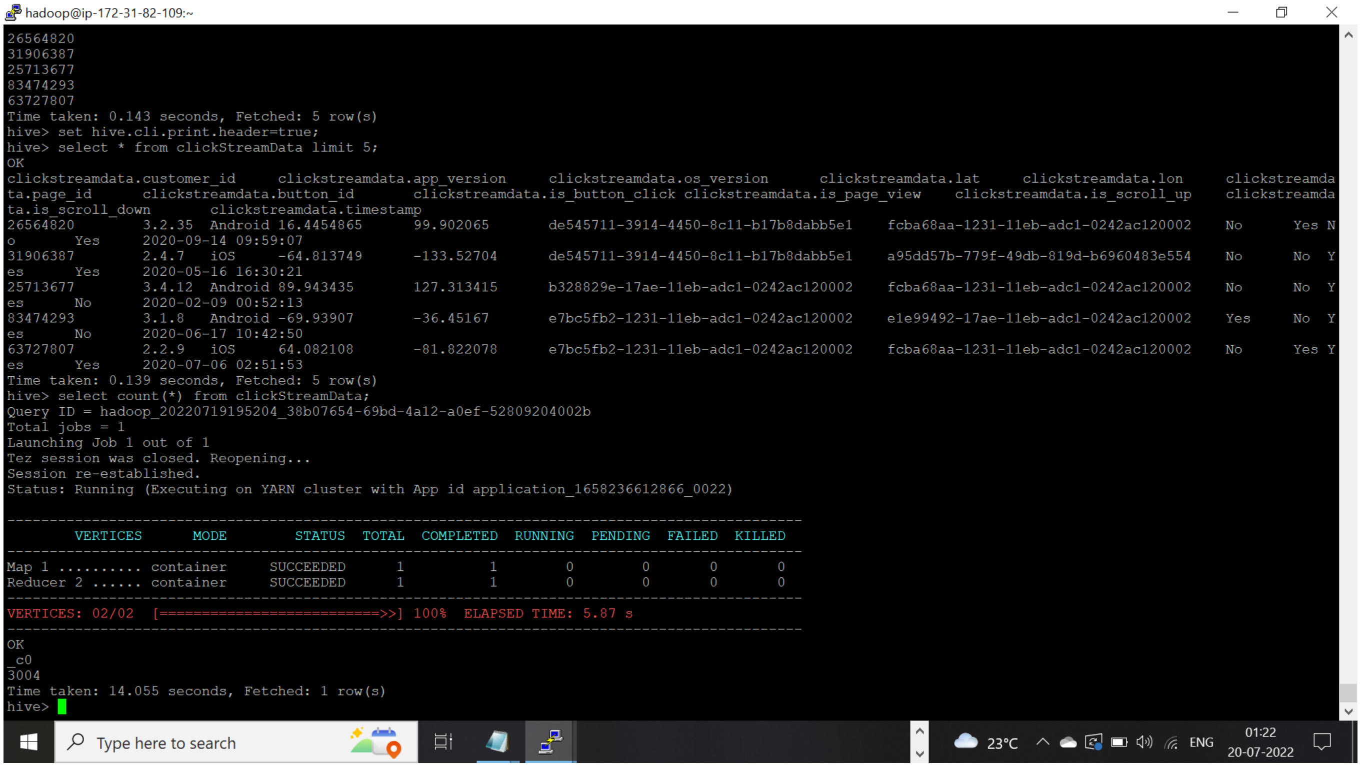
© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

5. Screenshot of the loaded data in clickStreamData table:



```
FAILED: ParseException line 2:0 cannot recognize input near 'date' 'timestamp' ',' in column name or primary key or foreign key
hive> create table if not exists bookingsAggregateData(
    > `date` timestamp,
    > no_of_bookings int)
    > row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.078 seconds
hive> drop table if exists clickStreamData;
OK
Time taken: 0.284 seconds
hive> create table if not exists clickStreamData(
    > customer_id int,
    > app_version string,
    > os_version string,
    > lat double,
    > lon double,
    > page_id string,
    > button_id string,
    > is_button_click string,
    > is_page_view string,
    > is_scroll_up string,
    > is_scroll_down string,
    > `timestamp` timestamp)
    > row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.068 seconds
hive> load data inpath '/user/hadoop/clickStream_flatten_data/' into table clickStreamData;
Loading data to table cab_rides_data.clickstreamdata
OK
Time taken: 0.886 seconds
hive> select * from clickStreamData limit 5;
OK
26564820        3.2.35  Android 16.4454865      99.902065       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes N
o       Yes     2020-09-14 09:59:07
31906387        2.4.7   iOS     -64.813749      -133.52704      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      No  Y
es      Yes     2020-05-16 16:30:21
25713677        3.4.12  Android 89.943435       127.313415      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No  Y
es      No      2020-02-09 00:52:13
83474293        3.1.8   Android -69.93907       -36.45167       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No  Y
es      No      2020-06-17 10:42:50
63727807        2.2.9   iOS     64.082108       -81.822078      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes Y
es      Yes     2020-07-06 02:51:53
Time taken: 2.13 seconds, Fetched: 5 row(s)
hive> 
```

6. To check the number of rows in **clickStreamData** table:



```
26564820
31906387
25713677
83474293
63727807
Time taken: 0.143 seconds, Fetched: 5 row(s)
hive> set hive.cli.print.header=true;
hive> select * from clickStreamData limit 5;
OK
clickstreamdata.customer_id     clickstreamdata.app_version     clickstreamdata.os_version      clickstreamdata.lat     clickstreamdata.lon     clickstreamda
ta.page_id      clickstreamdata.button_id       clickstreamdata.is_button_click clickstreamdata.is_page_view    clickstreamdata.is_scroll_up    clickstreamda
ta.is_scroll_down       clickstreamdata.timestamp
26564820        3.2.35  Android 16.4454865      99.902065       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes N
o       Yes     2020-09-14 09:59:07
31906387        2.4.7   iOS     -64.813749      -133.52704      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      No  Y
es      Yes     2020-05-16 16:30:21
25713677        3.4.12  Android 89.943435       127.313415      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No  Y
es      No      2020-02-09 00:52:13
83474293        3.1.8   Android -69.93907       -36.45167       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No  Y
es      No      2020-06-17 10:42:50
63727807        2.2.9   iOS     64.082108       -81.822078      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      Yes Y
es      Yes     2020-07-06 02:51:53
Time taken: 0.139 seconds, Fetched: 5 row(s)
hive> select count(*) from clickStreamData;
Query ID = hadoop_20220719195204_38b07654-69bd-4a12-a0ef-52809204002b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1658236612866_0022)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.87 s
----------------------------------------------------------------------------------------------
OK
_c0
3004
Time taken: 14.055 seconds, Fetched: 1 row(s)
hive> 
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

7. Command to create **bookingsData** table:

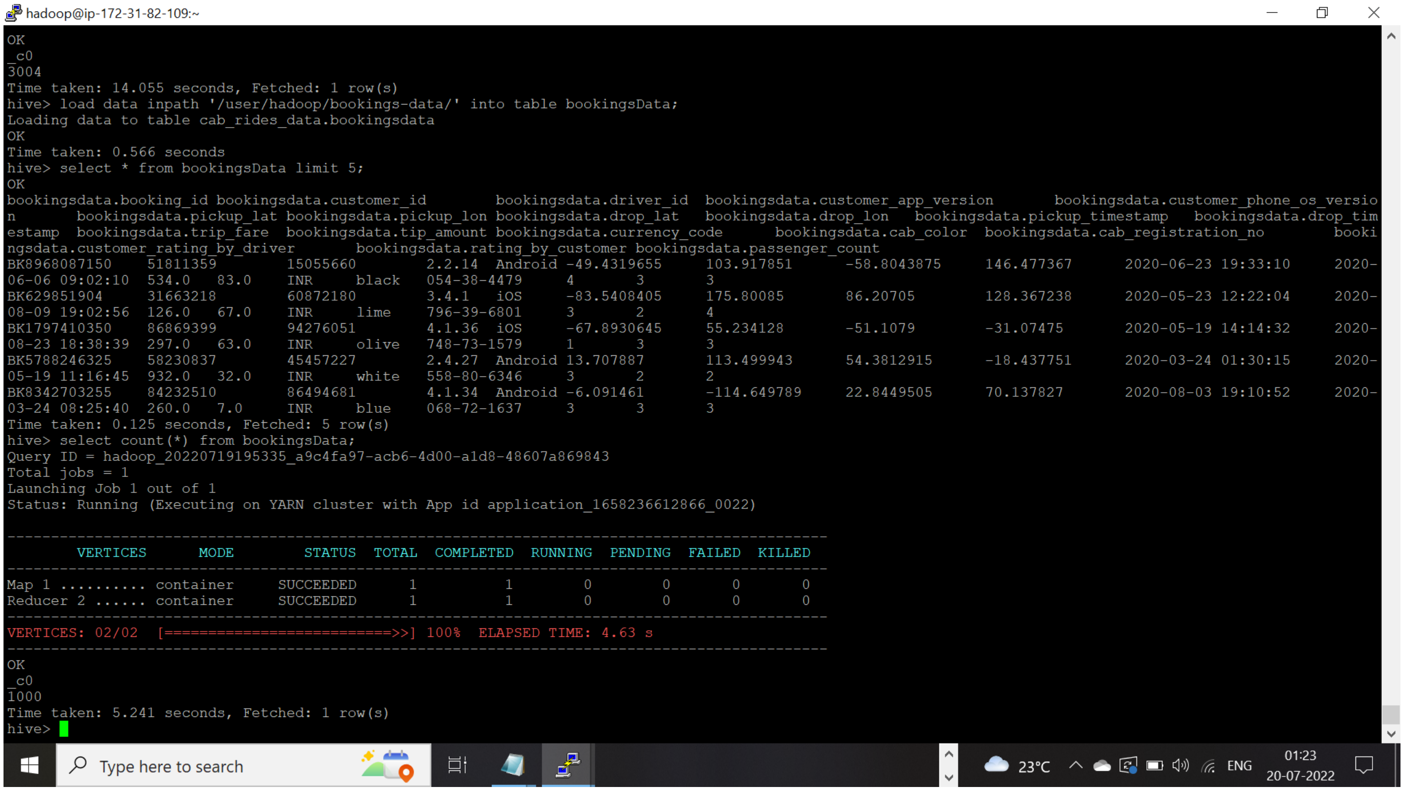
```
create table if not exists bookingsData(
booking_id string,
customer_id int,
driver_id int,
customer_app_version string,
customer_phone_os_version string,
pickup_lat double,
pickup_lon double,
drop_lat double,
drop_lon double,
pickup_timestamp timestamp,
drop_timestamp timestamp,
trip_fare double,
tip_amount double,
currency_code string,
cab_color string,
cab_registration_no string,
customer_rating_by_driver int,
rating_by_customer int,
passenger_count int)
row format delimited fields terminated by ',' lines terminated
by '\n' stored as textfile;
```

8. Command to load the data into **bookingsData** table:

```
load data inpath '/user/hadoop/bookings-
data/' into table bookingsData;
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

9. Screenshot of the loaded data and count of rows in **bookingsData** table:



```
OK
_c0
3004
Time taken: 14.055 seconds, Fetched: 1 row(s)
hive> load data inpath '/user/hadoop/bookings-data/' into table bookingsData;
Loading data to table cab_rides_data.bookingsdata
OK
Time taken: 0.566 seconds
hive> select * from bookingsData limit 5;
OK
bookingsdata.booking_id bookingsdata.customer_id        bookingsdata.driver_id  bookingsdata.customer_app_version       bookingsdata.customer_phone_os_versio
n       bookingsdata.pickup_lat bookingsdata.pickup_lon bookingsdata.drop_lat   bookingsdata.drop_lon   bookingsdata.pickup_timestamp   bookingsdata.drop_tim
estamp  bookingsdata.trip_fare  bookingsdata.tip_amount bookingsdata.currency_code       bookingsdata.cab_color  bookingsdata.cab_registration_no        booki
ngsdata.customer_rating_by_driver       bookingsdata.rating_by_customer bookingsdata.passenger_count
BK8968087150    51811359        15055660        2.2.14  Android -49.4319655     103.917851      -58.8043875     146.477367      2020-06-23 19:33:10     2020-
06-06 09:02:10  534.0   83.0    INR     black   054-38-4479     4       3       3
BK629851904     31663218        60872180        3.4.1   iOS     -83.5408405     175.80085       86.20705        128.367238      2020-05-23 12:22:04     2020-
08-09 19:02:56  126.0   67.0    INR     lime    796-39-6801     3       2       4
BK1797410350    86869399        94276051        4.1.36  iOS     -67.8930645     55.234128       -51.1079        -31.07475       2020-05-19 14:14:32     2020-
08-23 18:38:39  297.0   63.0    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     2020-
05-19 11:16:45  932.0   32.0    INR     white   558-80-6346     3       2       2
BK8342703255    84232510        86494681        4.1.34  Android -6.091461       -114.649789     22.8449505      70.137827       2020-08-03 19:10:52     2020-
03-24 08:25:40  260.0   7.0     INR     blue    068-72-1637     3       3       3
Time taken: 0.125 seconds, Fetched: 5 row(s)
hive> select count(*) from bookingsData;
Query ID = hadoop_20220719195335_a9c4fa97-acb6-4d00-a1d8-48607a869843
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658236612866_0022)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.63 s
----------------------------------------------------------------------------------------------
OK
_c0
1000
Time taken: 5.241 seconds, Fetched: 1 row(s)
hive>
```

10. Command to create **bookingsAggregateData** table and load the data:
In bookings_aggregate_data, we have **date** column with **date** as **datatype**. Hence to cast the column in date type, I have created a temporary table named as **testAggregateData** and loaded with data from HDFS:

Command to create testAggregateData table:

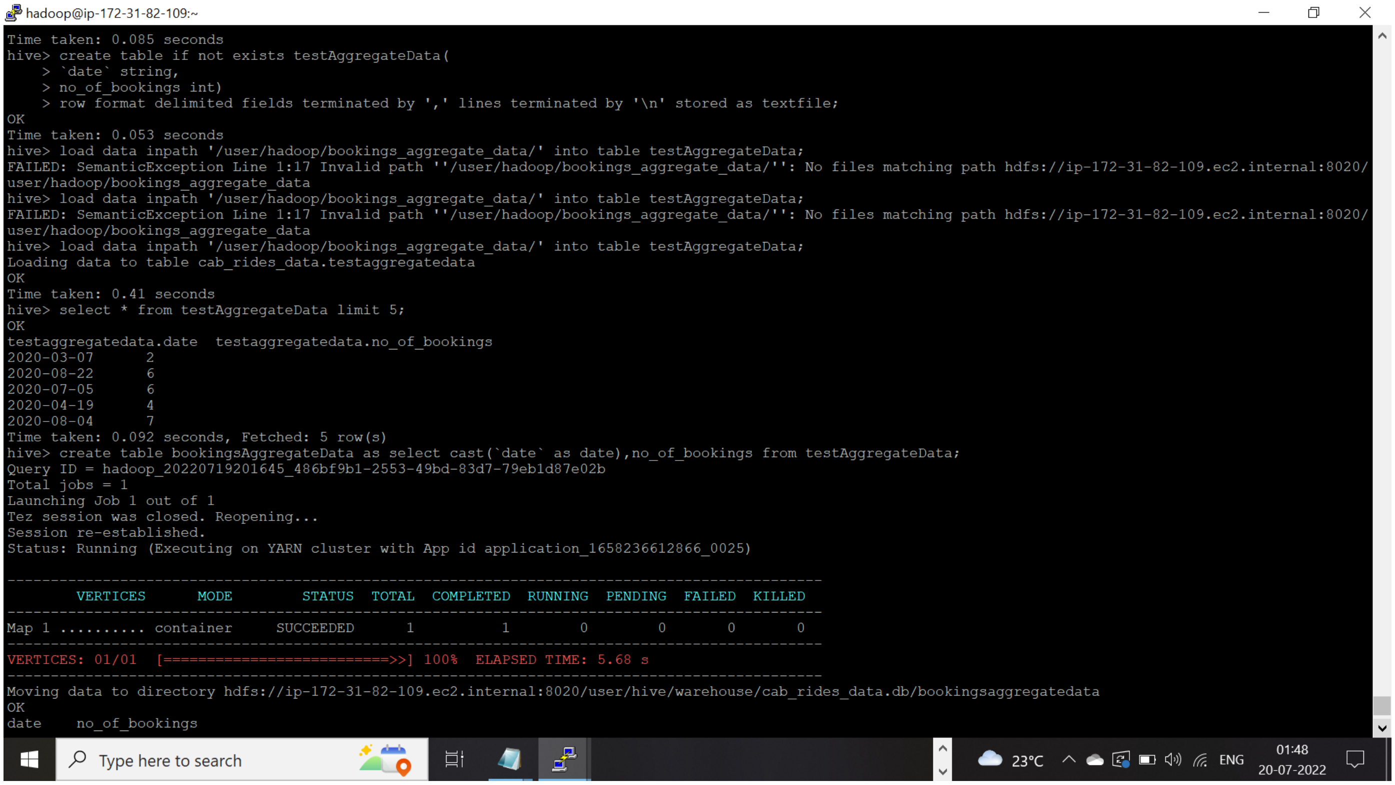
```
create table if not exists testAggregateData(
`date` string,
no_of_bookings int)
row format delimited fields terminated by ','
lines terminated by '\n' stored as textfile;
```

Command to load data from HDFS:

```
load data inpath '/user/hadoop/bookings_aggregate_data/'
into table testAggregateData;
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

Screenshot of the loaded data



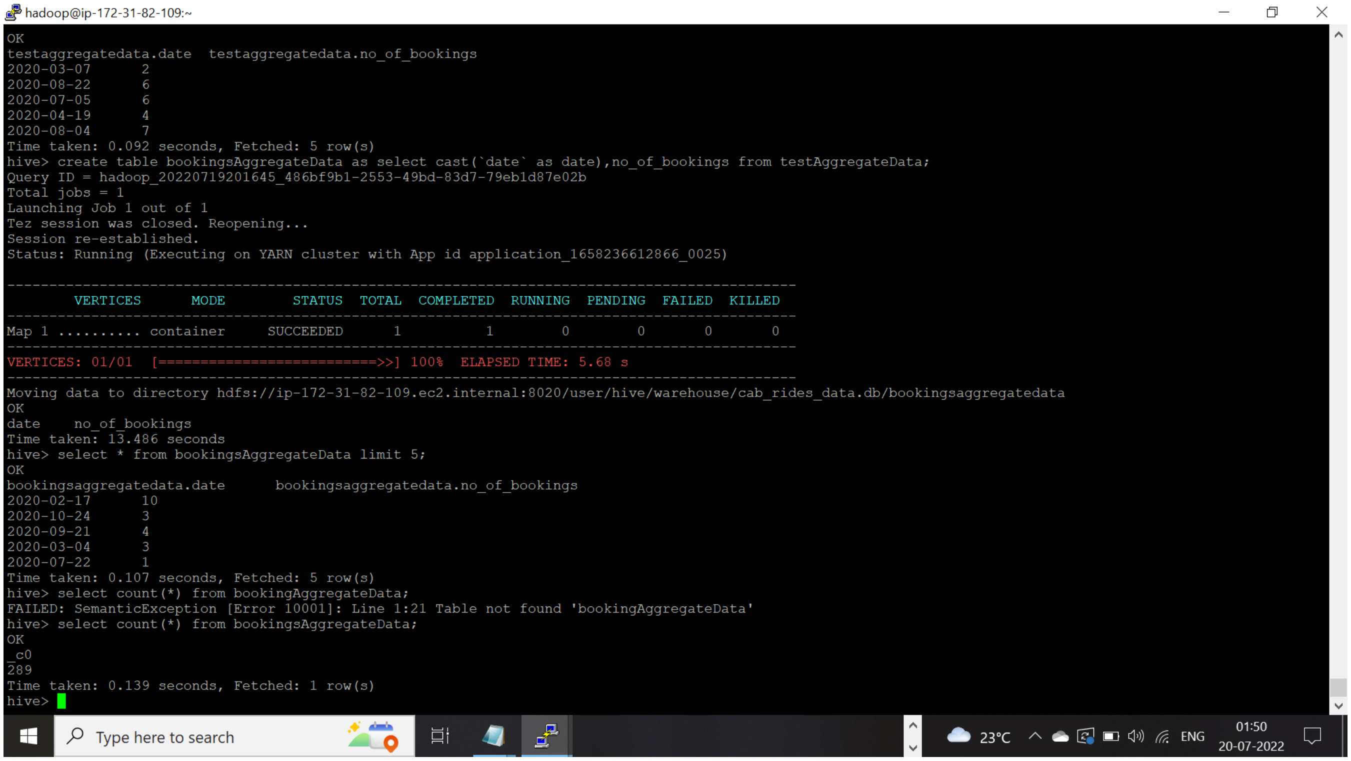
```
Time taken: 0.085 seconds
hive> create table if not exists testAggregateData(
    > `date` string,
    > no_of_bookings int)
    > row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.053 seconds
hive> load data inpath '/user/hadoop/bookings_aggregate_data/' into table testAggregateData;
FAILED: SemanticException Line 1:17 Invalid path ''/user/hadoop/bookings_aggregate_data/'': No files matching path hdfs://ip-172-31-82-109.ec2.internal:8020/
user/hadoop/bookings_aggregate_data
hive> load data inpath '/user/hadoop/bookings_aggregate_data/' into table testAggregateData;
FAILED: SemanticException Line 1:17 Invalid path ''/user/hadoop/bookings_aggregate_data/'': No files matching path hdfs://ip-172-31-82-109.ec2.internal:8020/
user/hadoop/bookings_aggregate_data
hive> load data inpath '/user/hadoop/bookings_aggregate_data/' into table testAggregateData;
Loading data to table cab_rides_data.testaggregatedata
OK
Time taken: 0.41 seconds
hive> select * from testAggregateData limit 5;
OK
testaggregatedata.date  testaggregatedata.no_of_bookings
2020-03-07      2
2020-08-22      6
2020-07-05      6
2020-04-19      4
2020-08-04      7
Time taken: 0.092 seconds, Fetched: 5 row(s)
hive> create table bookingsAggregateData as select cast(`date` as date),no_of_bookings from testAggregateData;
Query ID = hadoop_20220719201645_486bf9b1-2553-49bd-83d7-79eb1d87e02b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1658236612866_0025)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.68 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-82-109.ec2.internal:8020/user/hive/warehouse/cab_rides_data.db/bookingsaggregatedata
OK
date    no_of_bookings
```

In the next step, I have created a table named as **bookingsAggregateData** to cast the column date into date datatype.
Command to create **bookingsAggregateData** table**:**

```
create table bookingsAggregateData as select
cast(`date` as date),no_of_bookings from
testAggregateData;
```

Screenshot of the data loaded in bookingsAggregateData table and total count of rows:



```
OK
testaggregatedata.date  testaggregatedata.no_of_bookings
2020-03-07      2
2020-08-22      6
2020-07-05      6
2020-04-19      4
2020-08-04      7
Time taken: 0.092 seconds, Fetched: 5 row(s)
hive> create table bookingsAggregateData as select cast(`date` as date),no_of_bookings from testAggregateData;
Query ID = hadoop_20220719201645_486bf9b1-2553-49bd-83d7-79eb1d87e02b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1658236612866_0025)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.68 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-82-109.ec2.internal:8020/user/hive/warehouse/cab_rides_data.db/bookingsaggregatedata
OK
date    no_of_bookings
Time taken: 13.486 seconds
hive> select * from bookingsAggregateData limit 5;
OK
bookingsaggregatedata.date      bookingsaggregatedata.no_of_bookings
2020-02-17      10
2020-10-24      3
2020-09-21      4
2020-03-04      3
2020-07-22      1
Time taken: 0.107 seconds, Fetched: 5 row(s)
hive> select count(*) from bookingAggregateData;
FAILED: SemanticException [Error 10001]: Line 1:21 Table not found 'bookingAggregateData'
hive> select count(*) from bookingsAggregateData;
OK
_c0
289
Time taken: 0.139 seconds, Fetched: 1 row(s)
hive>
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved