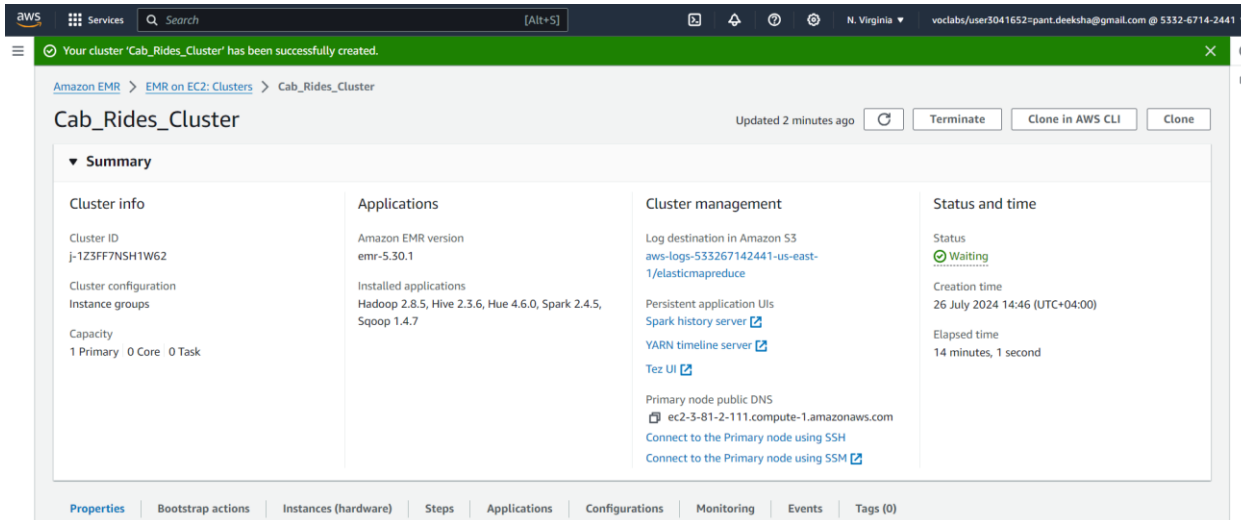


Logic For First Submission

1. Created an EMR cluster('Cab_Rides_Cluster') with Hadoop, Sqoop, Hive, Hue and Spark installed.



2. Loading streaming data from Kafka to HDFS

2.1 Developing a Pyspark file(spark_kafka_to_local.py) to read clickstream data from Kafka topic with below details

- Bootstrap-server - 18.211.252.152
- Port - 9092
- Topic - de-capstone5

2.2 Running the file 'spark_kafka_to_local.py' using 'spark-submit' command with the necessary Kafka package dependency

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

```
hadoop@ip-172-31-32-102:~$ ls -lrt
total 4
-rwxrwxrwx 1 hadoop hadoop 1461 Jul 24 17:23 spark_kafka_to_local.py
hadoop@ip-172-31-32-102:~$ 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-043444dc-5c91-4d98-b056-43b5efe29961;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
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.11/2.4.5/spark-sql-kafka-0-10_2.11-2.4.5.jar ...
[SUCCESSFUL ] org.apache.spark#spark-sql-kafka-0-10_2.11:2.4.5!spark-sql-kafka-0-10_2.11.jar (74ms)
downloading https://repo1.maven.org/maven2/org/apache/kafka/kafka-clients/2.0.0/kafka-clients-2.0.0.jar ...
[SUCCESSFUL ] org.apache.kafka#kafka-clients;2.0.0!kafka-clients.jar (137ms)
downloading https://repo1.maven.org/maven2/org/spark-project/spark/unused/1.0.0/unused-1.0.0.jar ...
[SUCCESSFUL ] org.spark-project.spark#unused;1.0.0!unused.jar (4ms)
downloading https://repo1.maven.org/maven2/org/lz4/lz4-java/1.4.0/lz4-java-1.4.0.jar ...
[SUCCESSFUL ] org.lz4#lz4-java;1.4.0!lz4-java.jar (34ms)
downloading https://repo1.maven.org/maven2/org/xerial/snappy/snappy-java/1.1.7.3/snappy-java-1.1.7.3.jar ...
[SUCCESSFUL ] org.xerial.snappy#snappy-java;1.1.7.3!snappy-java.jar(bundle) (122ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.16/slf4j-api-1.7.16.jar ...
[SUCCESSFUL ] org.slf4j#slf4j-api;1.7.16!slf4j-api.jar (40ms)
```

2.3 Validating the directory and data of files created by Pyspark file in HDFS:

```
hadoop fs -ls /user/root/clickstream_dump
```

```
[hadoop@ip-172-31-32-102 ~]$  
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/clickstream_dump  
Found 2 items  
drwxr-xr-x   - hadoop  hadoop           0 2024-07-26 13:35 /user/root/clickstream_dump/_spark_metadata  
-rw-r--r--   1 hadoop  hadoop    13421728 2024-07-26 13:35 /user/root/clickstream_dump/part-00000-0698dalb-86e2-4df0-a610-3139f7081ea9-c000  
_json
```

```
hadoop fs -cat /user/root/clickstream_dump/part-00000-0698da1b-86e2-4df0-a610-3139f7081ea9-c000.json | head -n 5
```

```
hadoop@ip-172-31-32-102 ~$
hadoop@ip-172-31-32-102 ~$ hadoop fs -cat /user/root/clickstream_dump/part-00000-0698da1b-86e2-4df0-a610-3139f7081ea9-c000.json | head
n 5
{"value_str": "(\\"customer_id\\": \\"18213492\\", \\"app_version\\": \\"1.4.31\\", \\"OS_version\\": \\"iOS\\", \\"lat\\": \\"5.676960\\", \\"lon\\": \\"16.753712\\", \\"page_id\\": \\"e7bcb5fb2-1231-11eb-adc1-0242ac120002\\", \\"button_id\\": \\"fcba68aa-1231-11eb-adc1-0242ac120002\\", \\"is_button_click\\": \\"No\\", \\"is_page_view\\": \\"No\\", \\"is_scroll_up\\": \\"No\\", \\"is_scroll_down\\": \\"No\\", \\"timestamp\\": \\"2020-01-18 19:44:43\\")\n"}
{"value_str": "(\\"customer_id\\": \\"21531056\\", \\"app_version\\": \\"3.4.16\\", \\"OS_version\\": \\"Android\\", \\"lat\\": \\"-77.6658015\\", \\"lon\\": \\"-28.740867\\", \\"page_id\\": \\"de545711-3914-4450-8c11-b17b8dabb5e1\\", \\"button_id\\": \\"a95d57b-779f-49db-819d-b6960483e554\\", \\"is_button_click\\": \\"Yes\\", \\"is_page_view\\": \\"No\\", \\"is_scroll_up\\": \\"No\\", \\"is_scroll_down\\": \\"No\\", \\"timestamp\\": \\"2020-09-09 23:49:34\\")\n"}
{"value_str": "(\\"customer_id\\": \\"89193671\\", \\"app_version\\": \\"3.1.27\\", \\"OS_version\\": \\"iOS\\", \\"lat\\": \\"68.5232875\\", \\"lon\\": \\"11.807065\\", \\"page_id\\": \\"e7bcb5fb2-1231-11eb-adc1-0242ac120002\\", \\"button_id\\": \\"e1e99492-17ae-11eb-adc1-0242ac120002\\", \\"is_button_click\\": \\"Yes\\", \\"is_page_view\\": \\"Yes\\", \\"is_scroll_up\\": \\"No\\", \\"is_scroll_down\\": \\"No\\", \\"timestamp\\": \\"2020-02-04 02:01:40\\")\n"}
{"value_str": "(\\"customer_id\\": \\"92968159\\", \\"app_version\\": \\"2.2.24\\", \\"OS_version\\": \\"Android\\", \\"lat\\": \\"-33.9557165\\", \\"lon\\": \\"-81.838663\\", \\"page_id\\": \\"de545711-3914-4450-8c11-b17b8dabb5e1\\", \\"button_id\\": \\"fcba68aa-1231-11eb-adc1-0242ac120002\\", \\"is_button_click\\": \\"Yes\\", \\"is_page_view\\": \\"No\\", \\"is_scroll_up\\": \\"Yes\\", \\"is_scroll_down\\": \\"No\\", \\"timestamp\\": \\"2020-05-11 01:22:55\\")\n"}
{"value_str": "(\\"customer_id\\": \\"68273069\\", \\"app_version\\": \\"4.4.30\\", \\"OS_version\\": \\"Android\\", \\"lat\\": \\"42.389278\\", \\"lon\\": \\"-113.983992\\", \\"page_id\\": \\"e7bcb5fb2-1231-11eb-adc1-0242ac120002\\", \\"button_id\\": \\"fcba68aa-1231-11eb-adc1-0242ac120002\\", \\"is_button_click\\": \\"Yes\\", \\"is_page_view\\": \\"Yes\\", \\"is_scroll_up\\": \\"No\\", \\"is_scroll_down\\": \\"Yes\\", \\"timestamp\\": \\"2020-05-22 22:23:13\\")\n"}
cat: Unable to write to output stream.
```

2.4 Developing a Pyspark file 'spark_local_flatten.py' for transforming/flattening the loaded Kafka data to a more structured format in HDFS

2.5 Executing the file 'spark_local_flatten.py' to flatten .json file in HDFS to .csv format with the necessary Kafka package dependency:

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

```
hadoop@ip-172-31-32-102:~$  
hadoop@ip-172-31-32-102:~$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_local_flatten.py
```

```
hadoop@ip-172-31-32-102:~$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_local_flatten.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-f561d178-502d-4b74-bbb5-581e53735359;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 358ms :: artifacts dl 9ms
  :: 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-f561d178-502d-4b74-bbb5-581e53735359
  confs: [default]
  0 artifacts copied, 6 already retrieved (0kB/11ms)
```

2.6 Validating the directory and contents of flattened file (.csv from .json) in HDFS:

```
hadoop fs -ls /user/root/clickstream
```

```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/clickstream
Found 1 items
drwxr-xr-x - hadoop hadoop 0 2024-07-26 13:40 /user/root/clickstream/_temporary
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/clickstream
Found 2 items
-rw-r--r-- 1 hadoop hadoop 0 2024-07-26 13:40 /user/root/clickstream/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 85468739 2024-07-26 13:40 /user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv
[hadoop@ip-172-31-32-102 ~]$
```

```
hadoop fs -cat /user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv| head -n 5
```

```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -cat /user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv| head -n 5
customer_id,app version,OS version,lat,lon,page_id,button_id,is button click,is page view,is scroll up,is scroll down,timestamp
18213492,1.4.31,iOS,5.676968,16.753712,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,No,No,""
21531056,3.4.16,Android,-77.6658015,28.740867,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,Yes,No,No,No,""
39193671,3.1.27,iOS,68.5232875,111.807065,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,Yes,No,No,""
92968159,2.2.24,Android,-33.9557165,-81.838663,de545711-3914-4450-8c11-b17b8dabb5e1,fcba68aa-1231-11eb-adc1-0242ac120002,Yes,No,Yes,No,""
cat: Unable to write to output stream.
[hadoop@ip-172-31-32-102 ~]$
```

3. Loading data from AWS RDS to Hadoop

3.1 Installing MySQL connector

```
wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
```

```
tar -xvf mysql-connector-java-8.0.25.tar.gz
```

```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
--2024-07-26 11:23:00-- https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
Resolving de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)... 3.5.29.246, 52.217.197.161, 52.217.49.164, ...
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)|3.5.29.246|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4079310 (3.9M) [application/x-gzip]
Saving to: 'mysql-connector-java-8.0.25.tar.gz.1'

100%[=====] 4,079,310 --.-K/s in 0.1s

2024-07-26 11:23:00 (28.6 MB/s) - 'mysql-connector-java-8.0.25.tar.gz.1' saved [4079310/4079310]

[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ tar -xvf mysql-connector-java-8.0.25.tar.gz
mysql-connector-java-8.0.25/
mysql-connector-java-8.0.25/src/
mysql-connector-java-8.0.25/src/build/
mysql-connector-java-8.0.25/src/build/java/
mysql-connector-java-8.0.25/src/build/java/documentation/
mysql-connector-java-8.0.25/src/build/java/instrumentation/
mysql-connector-java-8.0.25/src/build/misc/
mysql-connector-java-8.0.25/src/build/misc/debian.in/
mysql-connector-java-8.0.25/src/build/misc/debian.in/source/
mysql-connector-java-8.0.25/src/demo/
mysql-connector-java-8.0.25/src/demo/java/
mysql-connector-java-8.0.25/src/demo/java/demo/
mysql-connector-java-8.0.25/src/demo/java/demo/x/
mysql-connector-java-8.0.25/src/demo/java/demo/x/devapi/
mysql-connector-java-8.0.25/src/generated/
mysql-connector-java-8.0.25/src/generated/java/
```

```
cd mysql-connector-java-8.0.25/
```

```
sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib
```

```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ cd mysql-connector-java-8.0.25/
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
```

3.2 Ingesting batch data (bookings data stored in the RDS) to Hadoop using Sqoop import command

```
sqoop import \
--connect jdbc:mysql://upgradtest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table bookings \
--username student --password STUDENT123 \
--target-dir /user/root/bookings \
-m 1
```

Where:

RDS Connection String-> jdbc:mysql://upgradtest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase

Username-> student

Password-> STUDENT123

Table Name-> bookings

Target directory-> /user/root/bookings

No of mappers-> 1

```
hadoop@ip-172-31-32-102:~/mysql-connector-java-8.0.25
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ sqoop import \
> --connect jdbc:mysql://upgradtest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
> --table bookings \
> --username student --password STUDENT123 \
> --target-dir /user/root/bookings \
> -m 1
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
24/07/26 11:25:26 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/awk/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/07/26 11:25:26 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/07/26 11:25:27 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/07/26 11:25:27 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
24/07/26 11:25:27 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'bookings' AS t LIMIT 1
24/07/26 11:25:27 INFO orm.CompilationManager: HADOOP MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/a78b2d7187129f09e71e95225cb3ca59/bookings.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/07/26 11:25:30 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/a78b2d7187129f09e71e95225cb3ca59/bookings.jar
24/07/26 11:25:30 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/07/26 11:25:30 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/07/26 11:25:30 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/07/26 11:25:30 INFO mapreduce.ImportJobBase: Beginning import of bookings
24/07/26 11:25:31 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
```

```
hadoop@ip-172-31-32-102:~/mysql-connector-java-8.0.25
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)=205728
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=4286
  Total vcore-milliseconds taken by all map tasks=4286
  Total megabyte-milliseconds taken by all map tasks=6583296
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)=63
  CPU time spent (ms)=2450
  Physical memory (bytes) snapshot=281841664
  Virtual memory (bytes) snapshot=3303337984
  Total committed heap usage (bytes)=246939648
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=165678
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Transferred 161.7949 KB in 18.8134 seconds (8.6 KB/sec)
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Retrieved 1000 records.
hadoop@ip-172-31-32-102:~/mysql-connector-java-8.0.25$
```

**** No of records ingested is matching is matching with the validation documents**

Data Ingestion with Sqoop

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

Number of records retrieved - 1000

3.3 Validating the directory and files created in HDFS

hadoop fs -ls /user/root/bookings

```
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Transferred 161.7949 KB in 18.8134 seconds (8.6 KB/sec)
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Retrieved 1000 records.
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ hadoop fs -ls /user/root/bookings
Found 2 items
-rw-r--r-- 1 hadoop hadoop 0 2024-07-26 11:25 /user/root/bookings/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 165678 2024-07-26 11:25 /user/root/bookings/part-m-00000
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
```

Checking the first 5 records of the file generated :

hadoop fs -cat /user/root/bookings/part-m-00000 | head -n 5

```
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ hadoop fs -cat /user/root/bookings/part-m-00000 | head -n 5
BK8968087150,51811359,15055660,2.2.14,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
BK629851904,31663218,60872180,3.4.1,iOS,-83.5408405,175.80085,86.20705,128.367238,2020-05-23 12:22:04.0,2020-08-09 19:02:56.0,126,67,INR,lime,
796-39-6801,3,2,4
BK1797410350,86869399,94276051,4.1.36,iOS,-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,IN
R,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 19:10:52.0,2020-03-24 08:25:40.0,260,7,INR
,blue,068-72-1637,3,3,3
cat: Unable to write to output stream.
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
```

4. Loading Aggregate bookings data into HDFS

4.1 Creating Pyspark file “datewise_bookings_aggregates_spark.py” to find date-wise total bookings data.

4.2 Executing the file “datewise_bookings_aggregates_spark.py to generate aggregated data (.csv format) in HDFS using spark-submit command with the necessary Kafka package dependency:

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5
/home/hadoop/datewise_bookings_aggregates_spark.py
```

```
hadoop@ip-172-31-32-102:~$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 datewise_bookings_aggregates_spark.py
```

```
hadoop@ip-172-31-32-102:~$
ss_spark.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-ac38adc0-af7d-493b-a61c-101dcd86d73a;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 341ms :: artifacts dl 10ms
  :: 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]
  -----
  | conf | modules | artifacts | | | | |
|---|---|---|---|---|---|---|
  | default | 6 | 0 | 0 | 0 | 6 | 0 |
  -----
:: retrieving :: org.apache.spark:spark-submit-parent-ac38adc0-af7d-493b-a61c-101dcd86d73a
  confs: [default]
  0 artifacts copied, 6 already retrieved (0kB/10ms)
```

4.3 Validating the directory and file contents containing the aggregated bookings data in HDFS

hadoop fs -ls /user/root/datewise_bookings_agg/

Checking the first few records of the file generated:

`hadoop fs -cat /user/root/datewise_bookings_agg/part-00000-18cdabab-83e6-4637-8841-cf5ccb3fae86-c000.csv| head -n 5`

```
hadoop@ip-172-31-32-102:~$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/
Found 5 items
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:25 /user/root/bookings
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:16 /user/root/clickstream
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:00 /user/root/clickstream_dump
drwxr-xr-x - hadoop hadoop 0 2024-07-26 10:58 /user/root/clickstream_dump_checkpoint
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:31 /user/root/datewise_bookings_agg
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/datewise_bookings_agg/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 0 2024-07-26 11:31 /user/root/datewise_bookings_agg/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 3776 2024-07-26 11:31 /user/root/datewise_bookings_agg/part-00000-18cdabab-83e6-4637-8841-cf5ccb3fae86-c000.csv
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -cat /user/root/datewise_bookings_agg/part-00000-18cdabab-83e6-4637-8841-cf5ccb3fae86-c000.csv| head -n 5
pickup_date,count
2020-01-01,1
2020-01-02,3
2020-01-03,2
2020-01-04,2
[hadoop@ip-172-31-32-102 ~]$
```

5. Creating and loading the data in Hive-Managed Database & Tables.

5.1 Connecting to Hive instance and creating the database named 'cab_rides'

create database cab_rides;

use cab_rides;

```
[hadoop@ip-172-31-32-102 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database cab_rides;
OK
Time taken: 1.201 seconds
hive>
> use cab_rides;
OK
Time taken: 0.052 seconds
hive>
>
```

5.2 Creating tables in Hive

CREATE TABLE IF NOT EXISTS clickstream_data (

customer_id INT,

app_version STRING,

os_version STRING,

lat DOUBLE,

lon DOUBLE,

page_id STRING,

button_id STRING,

is_button_click BOOLEAN,

is_page_view BOOLEAN,

is_scroll_up BOOLEAN,

is_scroll_down BOOLEAN,

time_stamp TIMESTAMP

) ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

tblproperties('skip.header.line.count' = '1');

//To load the file with comma as delimiter

//To skip the header present in the file


```
hive>
>
>
> CREATE TABLE IF NOT EXISTS clickstream_data (
>   customer_id INT,
>   app_version STRING,
>   os_version STRING,
>   lat DOUBLE,
>   lon DOUBLE,
>   page_id STRING,
>   button_id STRING,
>   is_button_click BOOLEAN,
>   is_page_view BOOLEAN,
>   is_scroll_up BOOLEAN,
>   is_scroll_down BOOLEAN,
>   time_stamp TIMESTAMP
> ) ROW FORMAT DELIMITED
>   FIELDS TERMINATED BY ','
>   tblproperties('skip.header.line.count' = '1');
OK
Time taken: 0.078 seconds
```

```
CREATE TABLE IF NOT EXISTS bookings_detail (
  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 DECIMAL(10, 2),
  tip_amount DECIMAL(10, 2),
  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 ',';
```

//To load the file with comma as delimiter

```
hive>
> CREATE TABLE IF NOT EXISTS bookings_detail (
>   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 DECIMAL(10, 2),
>   tip_amount DECIMAL(10, 2),
>   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 ',' ;
OK
Time taken: 0.331 seconds
hive>
```

```
CREATE TABLE IF NOT EXISTS datewise_total_bookings (
  pickup_date DATE,
  total_bookings INT
) ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' //To load the file with comma as delimiter
tblproperties('skip.header.line.count'='1'); //To skip the header present in the file
```

```
hive> CREATE TABLE IF NOT EXISTS datewise_total_bookings (
>   pickup_date DATE,
>   total_bookings INT
> ) ROW FORMAT DELIMITED
>   FIELDS TERMINATED BY ',' ;
OK
Time taken: 0.059 seconds
```

- 5.3 Loading the Clickstream data into Hive tables using files in HDFS & validating the contents of the tables created.
- ```
LOAD DATA INPATH '/user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv'
OVERWRITE INTO TABLE clickstream_data;
```

Checking the first few records of the table loaded:  
 Select \* from clickstream\_data limit 2;



```
> LOAD DATA INPATH '/user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv' OVERWRITE
INTO TABLE clickstream_data;
Loading data to table cab_rides.clickstream_data
OK
Time taken: 1.005 seconds
hive>
>
> select * from clickstream_data limit 2;
OK
18213492 1.4.31 iOS 5.676968 16.753712 e7bc5fb2-1231-11eb-adc1-0242ac120002 fcba68aa-
231-11eb-adc1-0242ac120002 NULL NULL NULL NULL NULL
21531056 3.4.16 Android -77.6658015 28.740867 de545711-3914-4450-8c11-b17b8dabb5e1 a95dd57b-
79f-49db-819d-b6960483e554 NULL NULL NULL NULL NULL
Time taken: 0.187 seconds, Fetched: 2 row(s)
```

Checking the count of records ingested into the table:

SELECT COUNT(DISTINCT(CUSTOMER\_ID)) FROM CLICKSTREAM\_DATA ;

```
hive>
>
>
>
> SELECT COUNT(DISTINCT(CUSTOMER_ID))
> FROM CLICKSTREAM_DATA ;
Query ID = hadoop_20240726145111_b0b7c1ba-fd86-4ee1-ba10-d998966bdb7f
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_1721991376046_0023)

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED

Map 1 container SUCCEEDED 1 1 0 0 0 0
Reducer 2 container SUCCEEDED 2 2 0 0 0 0
Reducer 3 container SUCCEEDED 1 1 0 0 0 0

VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 6.45 s

OK
3000
Time taken: 13.464 seconds, Fetched: 1 row(s)
hive>
```

**\*\* No of records ingested is having slight difference of 16 records with the validation documents due to live streaming data.**

## Clickstream Table Count

Please check the number of records in the clickstream table

```
Number of records - 2984
```

- 5.4 Loading the Bookings data into Hive tables using files in HDFS & validating the contents of the tables created:  
LOAD DATA INPATH '/user/root/bookings/part-m-00000' OVERWRITE INTO TABLE bookings\_detail;

Checking the count of records ingested into the table:

select count(\*) from bookings\_detail;

```
hadoop@ip-172-31-32-102:~
hive>
>
> LOAD DATA INPATH '/user/root/bookings/part-m-00000'
> OVERWRITE INTO TABLE bookings_detail;
Loading data to table cab_rides.bookings_detail
OK
Time taken: 0.454 seconds
hive>
> select count(*) from bookings_detail;
Query ID = hadoop_20240726115232_12e64393-99df-4cd1-9a01-0bffa7ac6d35d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0005)

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: 4.86 s

OK
1000
Time taken: 5.423 seconds, Fetched: 1 row(s)
hive>
```

**\*\* No of records ingested is matching is matching with the validation documents**

5.5 Loading the Date-wise total bookings data into Hive tables using files in HDFS & validating the contents of the tables created

LOAD DATA INPATH '/user/root/datewise\_bookings\_agg/part-\*.csv' OVERWRITE INTO TABLE  
datewise\_total\_bookings;

```
hive>
> LOAD DATA INPATH '/user/root/datewise_bookings_agg/part-*.csv' OVERWRITE INTO TABLE datewise_total_bookings;
Loading data to table cab_rides.datewise_total_bookings
OK
Time taken: 1.026 seconds
hive>
> select * from datewise_total_bookings limit 2;
OK
2020-01-01 1
2020-01-02 3
Time taken: 0.147 seconds, Fetched: 2 row(s)
hive>
```

Checking the count of records ingested into the table

select count(\*) from datewise\_total\_bookings;

```
hive>
> select count(*) from datewise_total_bookings;
Query ID = hadoop_20240726142239_f2361e51-786c-4b61-899a-d846ab36d5b9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1721991376046_0021)

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: 4.77 s

OK
289
Time taken: 6.808 seconds, Fetched: 1 row(s)
hive> █
```

**\*\* No of records ingested is matching is matching with the validation documents**

## Bookings Aggregates Table Count

Please check the number of records in the bookings aggregates table

Number of records - 289