# Create aggregates for finding date-wise total bookings using the Spark script

**<Steps to create aggregation>**

1. Create a python script which will do the task.
    a. Command: `vi datewise_bookings_aggregates_spark.py` (File is added in the folder)
2. Run spark-submit command to run the above file
    a. Command: `spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2 datewise_bookings_aggregates_spark.py`

**<Command to move the csv file to HDFS>**

hdfs dfs -put command can be used to move file from local to HDFS.

Note: since the script is storing the file directly in the HDFS, this command isn't required.

**<Screenshot of the file in HDFS>**

Files in HDFS:

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBi
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/Stati
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 4 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 12:56 user/ec2-user/ClickStreamData
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 13:46 user/ec2-user/bookings_data
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/bookings_data_with_header
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/date_aggregated_bookings
[hadoop@ip-172-31-31-78 ~]$
```

LS bookings data with header

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/bookings_data_with_header
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/bookings_data_with_header/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup     176961 2022-01-23 15:01 user/ec2-user/bookings_data_with_header/part-00000-df625fbb-34aa-4ea0-b4db-7bdd3aa5468a-c000.csv
[hadoop@ip-172-31-31-78 ~]$
```

cat bookings data with header


Number of lines in bookings data with header

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/bookings_data_with_header/part-00000-df625fbb-34aa-4ea0-b4db-7bdd3aa5468a-c000.csv | wc -l
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
1001
```

Ls cat aggregated data

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/date_aggregated_bookings
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-01-23 15:01 user/ec2-user/date_aggregated_bookings/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       3769 2022-01-23 15:01 user/ec2-user/date_aggregated_bookings/part-00000-d9ea8de7-41c1-4f0d-84cc-964db6285270-c000.csv
[hadoop@ip-172-31-31-78 ~]$
```

cat date aggregated data

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/date_aggregated_bookings/part-00000-d9ea8de7-41c1-4f0d-84cc-964db6285270-c000.csv
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
date,count
2020-06-20,1
2020-04-20,3
2020-05-14,3
2020-04-22,2
2020-03-16,2
2020-09-16,2
2020-05-16,5
2020-01-18,4
2020-10-04,5
2020-03-05,5
2020-04-25,4
2020-05-26,2
2020-01-10,2
2020-06-07,3
2020-10-05,4
2020-08-04,7
2020-02-02,6
2020-05-18,2
2020-08-23,3
2020-08-17,4
2020-03-11,2
2020-10-08,4
2020-03-28,1
2020-08-22,6
```

Number of lines in date aggregated data

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/date_aggregated_bookings/part-00000-d9ea8de7-41c1-4f0d-84cc-964db6285270-c000.csv | wc -l
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.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]
290
```