

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/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 4 items
drwxr-xr-x - hadoop hdfsadmin group 0 2022-01-23 12:56 user/ec2-user/ClickStreamData
drwxr-xr-x - hadoop hdfsadmin group 0 2022-01-23 13:46 user/ec2-user/bookings_data
drwxr-xr-x - hadoop hdfsadmin group 0 2022-01-23 15:01 user/ec2-user/bookings_data_with_header
drwxr-xr-x - hadoop hdfsadmin group 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 hdfsadmin group 0 2022-01-23 15:01 user/ec2-user/bookings_data_with_header/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmin group 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

```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/bookings_data_with_header/part-00000-df625fbb-34aa-4ea0-b40b-7bdd3aa5468a-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]
booking_id,customer_id,driver_id,customer_app_version,customer_phone_no,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,driver_rating,by
B00968087159,51511359,15805646,2,2.14,Android,-49,4319455,183,917851,-86,863875,144,677347,2020-06-23,19:13:18,0,2020-06-06,89:02:10,0,534,63,IMR,black,054-38-4479,4,3,2020-06-23
B002905194,15643218,60872188,3,4.1.108,-83,5480485,68,88885,6,20785,128,367238,2020-06-23,12:22:04,0,2020-06-09,19:02:56,0,126,67,IMR,lime,796-90-6881,5,2,2020-06-23
B007714259,50489789,80274951,4,1.16,IOS,-97,8918645,59,224128,-51,3579,21,974755,2020-06-19,14:14:22,0,2020-06-23,18:00:09,0,292,65,IMR,olive,748-79-5379,1,2,2020-06-19
B00780244235,58238837,45647227,2,4.2,27,Android,13,787887,113,499943,64,3812915,-16,427751,2020-03-24,01:30:15,0,2020-06-19,11:16:46,0,932,32,IMR,white,558-86-4364,3,2,2020-03-24
B00342783255,84232518,86446481,4,1.16,Android,-6,897445,-134,647789,22,844985,76,137827,2020-08-03,19:18:52,0,2020-03-24,08:25:48,0,268,7,IMR,blue,868-72-1637,3,3,2020-08-03
B0041582435,11408246,30642605,2,4.39,IOS,-55,828851,-78,393188,-18,120202,173,672231,2020-07-17,00:33:48,0,2020-06-30,04:04:27,0,907,63,IMR,purple,182-18-5639,5,2,2020-07-17
B00452935584,6087378,78822368,2,1.9,IOS,1,215274,-56,8514983,35,152876,184,324989,2020-01-02,01:48:48,0,2020-02-16,04:28:55,0,647,17,IMR,teal,866-83-4349,2,3,4,2020-01-02
B0072080219,14272719,76427867,3,1.1,Android,-55,4822225,172,362256,65,8212565,61,390751,2020-04-18,05:11:47,0,2020-01-28,21:17:42,0,209,33,IMR,maroon,572-79-6030,3,3,2020-04-18
B0075352667,64487218,42168869,1,3.3,Android,46,885852,-14,825446,7,6120815,-155,428877,2020-06-09,05:54:31,0,2020-03-19,01:53:15,0,707,21,IMR,olive,663-23-5880,2,2,3,2020-06-09
B0001471433,45865173,64786618,3,1.3,IOS,-29,545326,64,843789,64,861899,-49,828835,2020-08-14,20:43:42,0,2020-06-03,09:39:59,0,586,6,IMR,fuchsia,255-52-5654,5,6,1,2020-08-14
B0003148736,37722776,27277776,2,3.3,Android,61,928468,92,249780,0,828189,115,440899,2020-06-07,06:42:10,0,2020-09-29,16:52:41,0,932,86,IMR,green,739-09-9649,2,1,2,2020-06-07
B0243762319,62052966,60677457,3,3.9,IOS,-42,651555,-139,156488,38,829995,-42,8855,2020-07-01,08:34:05,0,2020-09-38,17:40:23,0,821,23,IMR,black,598-44-6613,2,3,4,2020-07-01
B00483591568,56891961,53461747,4,2.36,IOS,-5,867365,-188,884339,35,816591,78,471358,2020-05-03,18:17:56,0,2020-06-08,09:11:27,0,71,18,IMR,fuchsia,454-84-8688,5,2,3,2020-05-03
B0073264253,64989721,4889581,2,2.22,Android,36,351315,5,484264,68,988353,36,588599,2020-03-05,16:02:18,0,2020-05-29,13:36:18,0,26,81,IMR,black,686-17-7843,1,1,3,2020-03-05
B0088821389,59135555,58436429,3,4.23,Android,-83,66599,186,268889,6,8388855,74,872352,2020-01-15,02:00:07,0,2020-05-12,21:53:04,0,571,99,IMR,navy,586-89-4981,1,5,3,2020-01-15
B0057428463,91511754,8928979,3,2.19,IOS,-43,118845,-99,939719,3,7826225,-44,208716,2020-06-28,01:18:34,0,2020-02-12,11:31:08,0,608,26,IMR,white,382-36-8684,6,2,2020-06-28
B0001387351,67875357,14654524,3,1.1,Android,-18,861959,-11,988953,57,233121,95,465994,2020-01-26,01:37:22,0,2020-04-28,09:42:08,0,999,3,IMR,teal,399-81-9362,1,1,4,2020-01-26
B00454323738,18447993,8497996,3,1.19,Android,-81,472235,-88,484916,12,698818,-148,99748,2020-09-24,08:18:31,0,2020-07-16,05:12:24,0,615,1,IMR,blue,824-35-8771,1,3,4,2020-09-24
B0016488433,36891778,15146219,4,2.38,Android,48,2615385,128,988881,32,182763,-58,501889,2020-07-26,06:12:06,0,2020-04-23,06:07:28,0,610,58,IMR,silver,833-16-3376,3,5,1,2020-07-26
B0003973649,28832866,9722676,2,1.18,IOS,-9,848860,161,839399,-12,943802,-149,236231,2020-09-28,10:02:49,0,2020-09-17,03:13:20,0,927,74,IMR,maroon,747-78-5057,2,2,4,2020-09-17
B0076478997,42725239,45284992,4,4.16,IOS,80,721451,179,498931,-32,340450,134,818322,2020-01-26,02:20:39,0,2020-06-12,11:05:04,0,246,72,IMR,blue,332-77-7648,5,1,2,2020-01-26
B00262681386,79159277,80265198,3,1.15,IOS,-9,445646,181,758883,80,354412,-46,718991,2020-09-03,11:32:13,0,2020-02-01,21:02:21,0,887,68,IMR,blue,225-31-8761,4,1,1,2020-09-03
B00625138481,51118772,58392277,3,3.27,IOS,68,1386763,141,658665,-6,926722,-64,24854,2020-05-11,06:25:18,0,2020-06-29,09:31:03,0,429,18,IMR,purple,229-41-2152,5,1,3,2020-05-11
B00785197448,13227964,3227964,4,2,Android,-57,695954,-172,155564,1,662888,126,729718,2020-06-18,20:11:26,0,2020-01-16,09:07:26,0,356,63,IMR,white,682-74-6532,3,3,3,2020-06-18
B02178492118,76255897,74117427,3,4.11,Android,14,3796175,97,889236,49,1656959,67,210169,2020-10-01,18:44:07,0,2020-03-27,22:38:08,0,927,68,IMR,silver,677-83-5852,5,4,3,2020-10-01
B0021806991,86209148,49983481,4,2,Android,76,808849,-138,99988,-85,855584,186,130884,2020-08-22,18:18:02,0,2020-01-20,09:57:58,0,385,97,IMR,green,478-19-9649,2,4,2,2020-08-22
B0076489479,60311458,30958015,4,1.12,IOS,61,89693,-161,53221,-24,864364,-133,519783,2020-02-06,01:53:27,0,2020-09-12,03:20:41,0,145,79,IMR,yellow,152-83-7438,1,2,3,2020-02-06
B003848116,66586834,9192598,3,3.3,IOS,-83,97133,4,531571,46,7478225,99,769593,2020-01-38,22:00:11,0,2020-06-28,11:56:17,0,61,16,IMR,gray,270-86-5864,2,5,3,2020-01-38
B0062438997,30489921,36339637,3,2,Android,-36,679446,68,821439,16,634643,2020-06-29,12:14:44,0,2020-06-29,12:14:44,0,818,59,IMR,silver,772-44-4863,1,1,3,2020-06-29
B0054734972,72812799,94442625,2,4.28,IOS,73,6442765,-7,728244,-41,581288,-174,587515,2020-09-08,08:12:51,0,2020-06-18,09:59:36,0,888,42,IMR,blue,854-85-1821,2,5,3,2020-09-08
B00456845437,83782838,98483882,2,4.1,IOS,-77,755876,-76,74789,-47,728688,-157,247688,2020-03-38,23:29:07,0,2020-06-13,06:03:29,0,169,88,IMR,silver,343-22-7214,4,2,2020-03-38
B0077880774,31623623,80780746,4,1.14,Android,-57,695954,-172,155564,1,662888,126,729718,2020-06-18,20:11:26,0,2020-01-16,09:07:26,0,356,63,IMR,white,682-74-6532,3,3,3,2020-06-18
B0055199415,3299782,60843642,3,4.1,IOS,2,915335,-187,961587,76,189977,58,82885,2020-03-31,14:07:36,0,2020-08-03,19:49:01,0,976,6,IMR,silver,714-38-3326,1,2,3,2020-03-31
B0072259222,13874749,4397934,4,2.31,IOS,-25,95059,28,56389,-28,686862,-87,074888,2020-02-12,08:24:07,0,2020-08-08,15:04:24,0,684,86,IMR,green,128-71-7729,4,1,2,2020-02-12
B0039701558,69278912,70954657,1,4.14,Android,17,9299495,96,843572,22,8251935,-165,408847,2020-06-38,22:18:15,0,2020-05-21,06:27:18,0,581,69,IMR,fuchsia,518-71-4452,1,2,2,2020-05-21
B0076423346,51944487,47858521,3,1.7,IOS,63,875193,162,669719,-76,595888,-148,563838,2020-10-18,18:15:09,0,2020-03-15,19:38:49,0,684,82,IMR,yellow,869-14-5564,1,2,2,2020-10-18
B0046497126,30828961,30828961,3,2,Android,-36,679446,68,821439,16,634643,2020-06-29,12:14:44,0,2020-06-29,12:14:44,0,818,59,IMR,silver,772-44-4863,1,1,3,2020-06-29
B0048160268,6486783,1481119,1,2.14,Android,64,644839,1,639925,68,722219,9,987899,2020-05-23,03:18:32,0,2020-05-12,20:03:36,0,226,70,IMR,maroon,163-71-4848,1,5,2,2020-05-23
B0022483943,84855446,9714591,2,4.28,Android,-33,421588,-157,971794,-13,214283,24,888849,2020-10-17,08:43:27,0,2020-06-06,06:13:49,0,483,59,IMR,black,557-78-8182,5,1,2,2020-10-17
B008977286,25594864,12,476233,-6,897261,15,981532,-144,238737,2020-01-26,02:57:53,0,2020-07-01,02:45:18,0,846,9,IMR,blue,448-20-8933,3,1,2,2020-01-26
B0081133848,85297366,1883426,1,3.6,Android,66,929865,-115,864259,-22,9164335,-61,674738,2020-09-27,04:15:08,0,2020-05-24,04:44:57,0,614,35,IMR,purple,389-83-9649,1,4,1,2020-09-27
B0079288452,94685732,4681752,3,4.48,Android,51,483234,9,8232,-10,815154,-149,914774,2020-01-02,06:15:09,0,2020-06-18,08:48:17,0,999,82,IMR,silver,538-71-4299,1,1,2,2020-01-02
B0029284849,67888586,4843423,3,1.2,Android,39,7187256,68,478476,3,1289235,-65,204863,2020-04-02,21:53:25,0,2020-05-01,12:24:31,0,999,34,IMR,black,854-98-3749,1,1,1,2020-04-02
B0045343355,68288837,44473375,4,3.7,IOS,43,895155,-144,937484,84,13471,135,991656,2020-03-24,04:34:28,0,2020-07-02,16:18:48,0,448,69,IMR,black,899-58-2088,1,5,1,2020-03-24
B0082314923,56328887,4796113,4,1.18,Android,87,784970,-129,133879,-67,9980495,43,705080,2020-07-01,09:44:38,0,2020-03-03,11:04:42,0,638,72,IMR,green,392-18-7882,3,3,3,2020-07-01
B00822787821,66842962,46776485,4,1.1,IOS,-38,28461,147,798468,167,399283,167,399283,2020-08-22,04:17:06,0,2020-05-18,12:03:45,0,64,48,IMR,fuchsia,185-44-2718,4,4,2,2020-08-22
```

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-b40b-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 hdfsadmin group 0 2022-01-23 15:01 user/ec2-user/date_aggregated_bookings/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmin group 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
```