Outputs of the executed functions
HW_01 Streaming Data
CS 467

Venkat Koushik Muthyapu

First command executed:

```
1 spark.conf.set("spark.sql.shuffle.partitions", 5)
2 static = spark.read.json("/databricks-datasets/definitive-guide/data/activity-data")
```

(3) Spark Jobs

static: pyspark.sql.dataframe.DataFrame

```
Arrival_Time: long
Creation_Time: long
Device: string
Index: long
Model: string
User: string
gt: string
x: double
y: double
z: double
```

Command took 16.49 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:03:52 PM on Hw-1

Second command executed:

Cmd 2

```
1 streaming = spark.readStream.schema(static.schema).option("maxFilesPerTrigger", 1)\
2   .json("/databricks-datasets/definitive-guide/data/activity-data")
```

streaming: pyspark.sql.dataframe.DataFrame

```
Arrival_Time: long
Creation_Time: long
Device: string
Index: long
Model: string
User: string
gt: string
x: double
y: double
z: double
```

Command took 0.56 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:05:14 PM on Hw-1

Third command executed:

Cmd 3

```
1 activityCounts = streaming.groupBy("gt").count()
```

activityCounts: pyspark.sql.dataframe.DataFrame

```
gt: string
count: long
```

Command took 0.09 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:06:06 PM on Hw-1

Fourth command executed:

```
activityQuery = activityCounts.writeStream.queryName("activity_counts")\
  .format("memory").outputMode("complete")\
  .start()
```

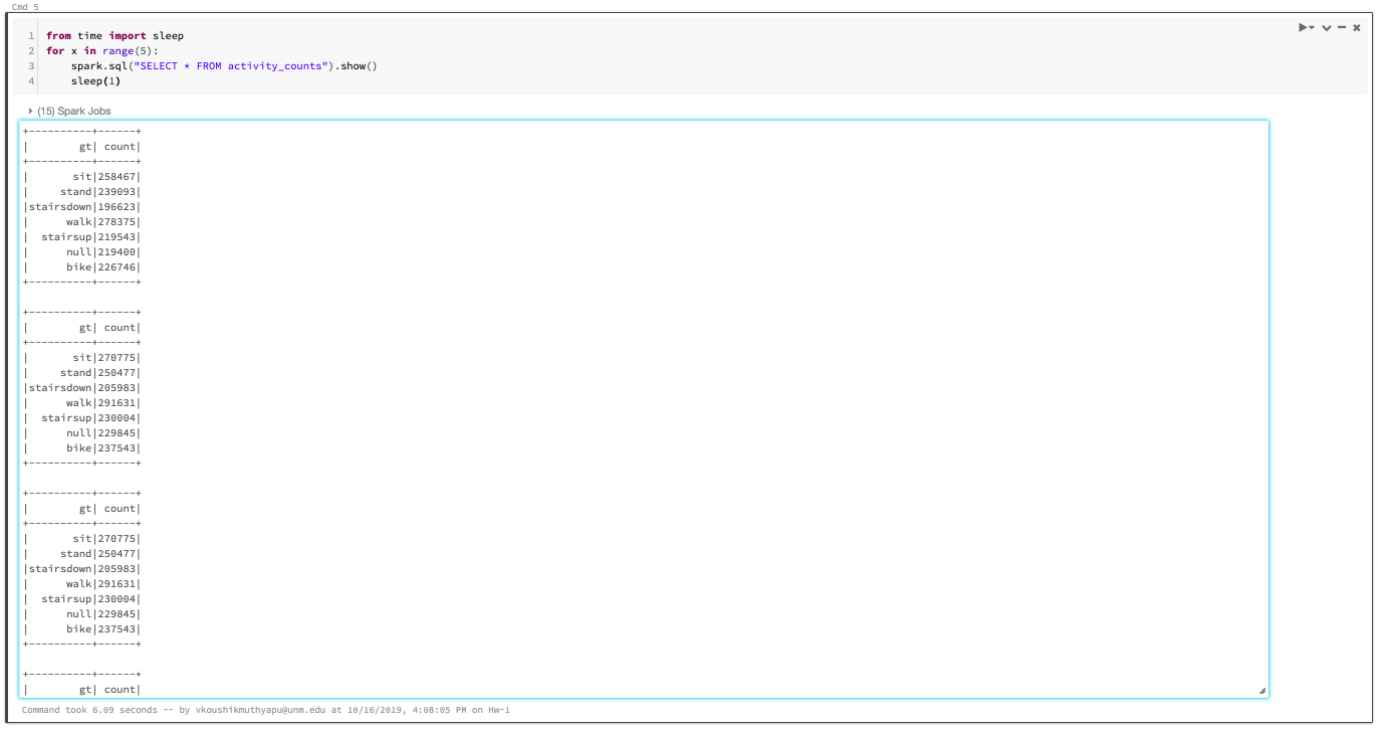
Cancel

(1) Spark Jobs

Job 16 View (2 stages)

activity_counts (id: ac525d34-9b19-4804-b02d-e163247b1403) *Last updated: 5 seconds ago*

Fifth command executed:



Sixth command executed:

Cmd 6

```
from pyspark.sql.functions import expr
simpleTransform = streaming.withColumn("stairs", expr("gt like '%stairs%'"))\
  .where("stairs")\
  .where("gt is not null")\
  .select("gt", "model", "arrival_time", "creation_time")\
  .writeStream\
  .queryName("simple_transform")\
  .format("memory")\
  .outputMode("append")\
  .start()
```

Cancel

(1) Spark Jobs

Job 127 View (1 stages)

simple_transform (id: f4ab409a-01aa-479f-9f40-f9fcca584c20) *Last updated: 5 seconds ago*

Cmd 7

```
spark.sql("SELECT * FROM simple_transform").show()
```

(1) Spark Jobs

```
+--------+------+-------------+-------------------+
|      gt| model| arrival_time|      creation_time|
+--------+------+-------------+-------------------+
|stairsup|nexus4|1424687983719|1424687981726802718|
|stairsup|nexus4|1424687984000|1424687982009853255|
|stairsup|nexus4|1424687984404|1424687982411977009|
|stairsup|nexus4|1424687984805|1424687982814351277|
|stairsup|nexus4|1424687985210|1424687983217500861|
|stairsup|nexus4|1424687985620|1424687983620332892|
|stairsup|nexus4|1424687986016|1424687984023164923|
|stairsup|nexus4|1424687986420|1424687984425874884|
|stairsup|nexus4|1424687986820|1424687984828822915|
|stairsup|nexus4|1424687987225|1424687985231654946|
|stairsup|nexus4|1424687987625|1424687985634469017|
|stairsup|nexus4|1424687987992|1424687986002114280|
|stairsup|nexus4|1424687988191|1424689834237427627|
|stairsup|nexus4|1424687988392|1424689834438660537|
|stairsup|nexus4|1424687988592|1424689834640076553|
|stairsup|nexus4|1424687988794|1424689834841675674|
|stairsup|nexus4|1424687988999|1424689835047943984|
|stairsup|nexus4|1424687989200|1424687987205721701|
|stairsup|nexus4|1424687989409|1424689835458070221|
|stairsup|nexus4|1424687989606|1424687987613772238|
+--------+------+-------------+-------------------+
only showing top 20 rows

Command took 1.33 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:10:16 PM on Hw-1
```

seventh command executed:

```
deviceModelStats = streaming.cube("gt", "model").avg()\
  .drop("avg(Arrival_time)")\
  .drop("avg(Creation_Time)")\
  .drop("avg(Index)")\
  .writeStream.queryName("device_counts").format("memory")\
  .outputMode("complete")\
  .start()
```

Cancel

(1) Spark Jobs

device_counts (id: f60e20df-31d0-4b55-bb91-33aaf35cde49) *Last updated: 5 seconds ago*

Cmd 9

```
spark.sql("SELECT * FROM device_counts").show()
```

(3) Spark Jobs

```
+----------+------+--------------------+--------------------+--------------------+
|        gt| model|              avg(x)|              avg(y)|              avg(z)|
+----------+------+--------------------+--------------------+--------------------+
|       sit|  null|-4.84874225480320...|3.237740430277438...|-4.65213713798297E-5|
|     stand|  null|-3.19865608911239...|4.070066943871052E-4|1.027885098598942...|
|       sit|nexus4|-4.84874225480320...|3.237740430277438...|-4.65213713798297E-5|
|     stand|nexus4|-3.19865608911239...|4.070066943871052E-4|1.027885098598942...|
|      null|  null|-0.00601179783148...|-7.22352747272202...|0.003847526633724...|
|      null|  null|0.001116290402170147|-0.00657669929797...|-0.00918707169429...|
|      walk|  null|-0.00371092404649...|0.003353258755646...|0.001107451361787...|
|      null|nexus4|-0.00601179783148...|-7.22352747272202...|0.003847526633724...|
|      null|nexus4|0.001116290402170147|-0.00657669929797...|-0.00918707169429...|
|      bike|  null|0.020961064903431637|-0.00915222693840...|-0.08538726699652664|
|  stairsup|  null|-0.024425885716177797|-0.011502523241 03...|-0.09947070802664394|
|stairsdown|  null|0.025383337512333622| -0.0362580763332372| 0.12712694345857248|
|      bike|nexus4|0.020961064903431637|-0.00915222693840...|-0.08538726699652664|
|      walk|nexus4|-0.00371092404649...|0.003353258755646...|0.001107451361787...|
|stairsdown|nexus4|0.025383337512333622| -0.0362580763332372| 0.12712694345857248|
|  stairsup|nexus4|-0.024425885716177797|-0.011502523241 03...|-0.09947070802664394|
+----------+------+--------------------+--------------------+--------------------+

Command took 0.46 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:12:36 PM on Hw-1
```

Eighth command executed:

Cmd 10

```
historicalAgg = static.groupBy("gt", "model").avg()
deviceModelStats = streaming.drop("Arrival_Time", "Creation_Time", "Index")\
  .cube("gt", "model").avg()\
  .join(historicalAgg, ["gt", "model"])\
  .writeStream.queryName("device_counts").format("memory")\
  .outputMode("complete")\
  .start()
```

Cancel

(1) Spark Jobs

device_counts (id: 1aadfb82-e19d-4b17-ab8b-d5d7c203fc1a) *Last updated: 25 seconds ago*

historicalAgg: pyspark.sql.dataframe.DataFrame = [gt: string, model: string ... 6 more fields]

Cmd 11

```
spark.sql("SELECT * FROM device_counts").show()
```

(3) Spark Jobs

gt	model	avg(x)	avg(y)	avg(z)	avg(Arrival_Time)	avg(Creation_Time)	avg(Index)	avg(x)	avg(y)	avg(z)
bike	nexus4	0.023512544470933663	-0.01304747996973...	-0.08360475809007027	1.424751134339985...	1.424752127369589...	326459.6867328154	0.02268875955086685	-0.00877912156368...	-0.08251001663412344
null	nexus4	-0.003025012215O6...	-0.00410754501410...	0.005961452067049477	1.424749002876339...	1.424749919482127...	219276.9663669269	-0.00847688860109...	-7.30455258739188...	0.003090601491419...
stairsdown	nexus4	0.028103791071500104	-0.03570080351911373	0.12203047970606548	1.424744591412857E12	1.424745503635636...	230452.44623187225	0.021613908669165436	-0.03249018824752617	0.12035922691504071
stand	nexus4	-3.00989804137386...	4.133303473120163...	-2.86960196767402...	1.424743637921209...	1.424744579547459...	31317.877585550017	-3.110821896917l1...	3.2184616659753G1...	2.141300040636498E-4
walk	nexus4	0.001970352086338262	7.489666845957323E-5	-0.00149828380428...	1.424746420641789...	1.424747351060674...	149760.09974990616	-0.00390116006094...	0.001052508689953...	-6.95435553042997...
sit	nexus4	-4.92132825379803...	3.757376157445784...	-4.428633462507Z9...	1.424741207868231...	1.424742112220356...	74577.84690275553	-5.494332440395S7...	2.79144628170004E-4	-2.33994461689905...
stairsup	nexus4	-0.026233013188633781	-0.0138593176529181	-0.0939500972801951	1.424745961101163E12	1.424746915892737...	227912.96550673083	-0.024799652877716421	-0.00800392344379...	-0.10034088415060395

Command took 4.17 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:14:30 PM on Hw-1

Ninth command executed:

Cmd 12

```
streaming = spark\
.readStream\
.schema(static.schema)\
.option("maxFilesPerTrigger", 10)\
.json("/databricks-datasets/definitive-guide/data/activity-data")
```

streaming: pyspark.sql.dataframe.DataFrame

```
Arrival_Time: long
Creation_Time: long
Device: string
Index: long
Model: string
User: string
gt: string
x: double
y: double
z: double
```

Command took 0.38 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:15:32 PM on Hw-1

Tenth command executed:

Cmd 13

```
withEventTime = streaming.selectExpr(
  "*",
  "cast(cast(Creation_Time as double)/1000000000 as timestamp) as event_time")
```

withEventTime: pyspark.sql.dataframe.DataFrame

```
Arrival_Time: long
Creation_Time: long
Device: string
Index: long
Model: string
User: string
gt: string
x: double
y: double
z: double
event_time: timestamp
```

Command took 0.12 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:16:28 PM on Hw-1

Eleventh command executed:

Cmd 14

```
from pyspark.sql.functions import window, col
withEventTime.groupBy(window(col("event_time"), "10 minutes")).count()\
  .writeStream\
  .queryName("pyevents_per_window")\
  .format("memory")\
  .outputMode("complete")\
  .start()
```

Cancel

(1) Spark Jobs

pyevents_per_window (id: 894d3535-3633-4709-823c-e69678563ff0) *Last updated: 5 seconds ago*

Out[14]: <pyspark.sql.streaming.StreamingQuery at 0x7fc99bcc5dd8>

Cmd 15

```
spark.sql("SELECT * FROM pyevents_per_window").show()
```

(2) Spark Jobs

```
+--------------------+-----+
|              window|count|
+--------------------+-----+
|[2015-02-24 11:50...|56549|
|[2015-02-24 13:00...|50017|
|[2015-02-23 12:30...|37826|
|[2015-02-23 10:20...|37283|
|[2015-02-24 12:30...|47147|
|[2015-02-24 13:10...|39531|
|[2015-02-23 10:30...|37621|
|[2015-02-23 10:40...|33200|
|[2015-02-23 13:20...|39798|
|[2015-02-22 00:40...|   11|
|[2015-02-24 11:20...|42637|
|[2015-02-24 12:20...|50131|
|[2015-02-24 14:00...|56297|
|[2015-02-24 14:10...|63438|
|[2015-02-24 13:40...|49484|
|[2015-02-24 13:50...|36033|
|[2015-02-23 14:30...|35486|
|[2015-02-23 13:40...|62875|
|[2015-02-24 12:00...|75137|
|[2015-02-23 12:20...|39988|
+--------------------+-----+
only showing top 20 rows
```

Command took 0.42 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:17:59 PM on Hw-1

Twelfth command executed:

Cmd 16

```
  from pyspark.sql.functions import window, col
  withEventTime.groupBy(window(col("event_time"), "10 minutes")).count()\
    .writeStream\
    .queryName("pyevents_per_window")\
    .format("console")\
    .outputMode("complete")\
    .start()
```

Cancel

(1) Spark Jobs

pyevents_per_window (id: dd3c8ba6-4ae7-4dfe-a9dd-574fec27f1bd) *Last updated: 5 seconds ago*

Out[16]: <pyspark.sql.streaming.StreamingQuery at 0x7fc99bccc4a8>

Cmd 17

```
spark.sql("SELECT * FROM pyevents_per_window").show()
```

(2) Spark Jobs

```
+--------------------+------+
|              window| count|
+--------------------+------+
|[2015-02-24 11:50...|150773|
|[2015-02-24 13:00...|133323|
|[2015-02-23 12:30...|100853|
|[2015-02-23 10:20...| 99178|
|[2015-02-24 12:30...|125679|
|[2015-02-24 13:10...|105494|
|[2015-02-23 10:30...|100443|
|[2015-02-23 10:40...| 88681|
|[2015-02-23 13:20...|106075|
|[2015-02-22 00:40...|    35|
|[2015-02-24 11:20...|113768|
|[2015-02-24 12:20...|133623|
|[2015-02-24 14:00...|150225|
|[2015-02-24 14:10...|169064|
|[2015-02-24 13:40...|132243|
|[2015-02-24 13:50...| 96023|
|[2015-02-23 14:30...| 94669|
|[2015-02-23 13:40...|167565|
|[2015-02-24 12:00...|200133|
|[2015-02-23 12:20...|106291|
+--------------------+------+
only showing top 20 rows
```

Command took 0.25 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:19:33 PM on Hw-1

Thirteenth command executed:

Cmd 18

```
from pyspark.sql.functions import window, col
withEventTime.groupBy(window(col("event_time"), "10 minutes"), "User").count()\
  .writeStream\
  .queryName("pyevents_per_window")\
  .format("memory")\
  .outputMode("complete")\
  .start()
```

Cancel

(1) Spark Jobs

pyevents_per_window (id: 8f3b4b62-453d-4ae1-9a1e-f9c62373620e) *Last updated: 10 seconds ago*

Out[20]: <pyspark.sql.streaming.StreamingQuery at 0x7fc99bcb55c0>

Cmd 19

```
spark.sql("SELECT * FROM pyevents_per_window").show()
```

(2) Spark Jobs

```
+--------------------+----+-----+
|              window|User|count|
+--------------------+----+-----+
|[2015-02-23 10:20...|   g|12403|
|[2015-02-24 13:30...|   b|10175|
|[2015-02-24 12:20...|   f|16712|
|[2015-02-24 15:00...|   e|12468|
|[2015-02-24 13:00...|   f| 4146|
|[2015-02-23 13:40...|   a|13947|
|[2015-02-24 14:50...|   e|15818|
|[2015-02-23 14:30...|   h|11841|
|[2015-02-24 14:10...|   e| 8440|
|[2015-02-24 13:00...|   d|12490|
|[2015-02-22 00:40...|   a|    5|
|[2015-02-24 14:20...|   b|12860|
|[2015-02-24 14:20...|   e|14222|
|[2015-02-23 13:50...|   h|11780|
|[2015-02-23 12:30...|   c|12617|
|[2015-02-23 11:10...|   g|11464|
|[2015-02-23 14:00...|   h|12165|
|[2015-02-23 13:30...|   a|12026|
|[2015-02-24 13:10...|   d|13113|
|[2015-02-24 12:00...|   i|11945|
+--------------------+----+-----+
only showing top 20 rows
```

Command took 1.19 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:20:54 PM on Hw-1

Fourteenth Command Executed:

Cmd 20

```
from pyspark.sql.functions import window, col
withEventTime.groupBy(window(col("event_time"), "10 minutes", "5 minutes"))\
  .count()\
  .writeStream\
  .queryName("pyevents_per_window")\
  .format("memory")\
  .outputMode("complete")\
  .start()
```

Cancel

(1) Spark Jobs

pyevents_per_window (id: 50f4b778-56c7-4b2e-87d5-5d84e1368597) *Last updated: 5 seconds ago*

Out[22]: <pyspark.sql.streaming.StreamingQuery at 0x7fc99bc74a20>

Cmd 21

```
spark.sql("SELECT * FROM pyevents_per_window").show()
```

(2) Spark Jobs

```
+--------------------+------+
|              window| count|
+--------------------+------+
|[2015-02-23 14:15...| 80710|
|[2015-02-24 11:50...|113104|
|[2015-02-24 13:00...| 99825|
|[2015-02-22 00:35...|    19|
|[2015-02-23 12:30...| 75668|
|[2015-02-23 10:20...| 74470|
|[2015-02-23 13:25...| 68783|
|[2015-02-24 14:25...|153030|
|[2015-02-23 12:55...| 85415|
|[2015-02-22 00:40...|    19|
|[2015-02-23 12:35...| 68457|
|[2015-02-23 13:05...|154443|
|[2015-02-24 11:20...| 85275|
|[2015-02-24 13:35...|131110|
|[2015-02-24 14:00...|112625|
|[2015-02-24 12:30...| 94285|
|[2015-02-24 13:10...| 79134|
|[2015-02-23 10:30...| 75357|
|[2015-02-24 11:45...|103765|
|[2015-02-23 10:40...| 66463|
+--------------------+------+
only showing top 20 rows
```

Command took 0.25 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:22:36 PM on Hw-1

Fifteenth Command Executed:

Cmd 22

```
from pyspark.sql.functions import window, col
withEventTime\
  .withWatermark("event_time", "30 minutes")\
  .groupBy(window(col("event_time"), "10 minutes", "5 minutes"))\
  .count()\
  .writeStream\
  .queryName("pyevents_per_window")\
  .format("memory")\
  .outputMode("complete")\
  .start()
```

Cancel

(1) Spark Jobs

pyevents_per_window (id: 0a21f66f-25a1-4947-9099-5f915a22b77e) *Last updated: 5 seconds ago*

Out[24]: <pyspark.sql.streaming.StreamingQuery at 0x7fc99eb1abe0>

Cmd 23

```
spark.sql("SELECT * FROM pyevents_per_window").show()
```

(2) Spark Jobs

```
+--------------------+-----+
|              window|count|
+--------------------+-----+
|[2015-02-23 14:15...|13523|
|[2015-02-24 11:50...|18854|
|[2015-02-24 13:00...|16636|
|[2015-02-22 00:35...|    5|
|[2015-02-23 12:30...|12617|
|[2015-02-23 10:20...|12403|
|[2015-02-23 13:25...|11465|
|[2015-02-24 12:30...|15731|
|[2015-02-24 13:10...|13113|
|[2015-02-24 14:25...|25431|
|[2015-02-23 10:30...|12589|
|[2015-02-24 11:45...|17232|
|[2015-02-23 10:40...|11035|
|[2015-02-23 12:55...|14146|
|[2015-02-23 13:20...|13269|
|[2015-02-22 00:40...|    5|
|[2015-02-23 12:35...|11361|
|[2015-02-23 13:05...|25748|
|[2015-02-24 11:20...|14160|
|[2015-02-24 12:20...|16712|
+--------------------+-----+
only showing top 20 rows

Command took 0.19 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:49:38 PM on Hw-1
```

Sixteenth Command Executed:

Cmd 24

```
from pyspark.sql.functions import expr

withEventTime\
  .withWatermark("event_time", "5 seconds")\
  .dropDuplicates(["User", "event_time"])\
  .groupBy("User")\
  .count()\
  .writeStream\
  .queryName("pydeduplicated")\
  .format("memory")\
  .outputMode("complete")\
  .start()
```

Cancel

(1) Spark Jobs

pydeduplicated (id: b1bd2732-792c-4883-b5ab-063e276e2020) *Last updated: 5 seconds ago*

Out[26]: <pyspark.sql.streaming.StreamingQuery at 0x7fc99bcccc50>

Cmd 25

```
spark.sql("SELECT * FROM pydeduplicated").show()
```

(3) Spark Jobs

```
+----+-----+
|User|count|
+----+-----+
|   a|80850|
|   b|91230|
|   c|77150|
|   g|91679|
|   h|77330|
|   e|96030|
|   f|92060|
|   d|81240|
|   i|92550|
+----+-----+

Command took 1.01 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 4:50:40 PM on Hw-1
```

My Execution:

Cmd 8

```
deviceModelStats = streaming.cube("gt", "device").avg()\
  .drop("avg(x)")\
  .drop("avg(y)")\
  .drop("avg(z)")\
  .writeStream.queryName("My_Execution").format("memory")\
  .outputMode("Complete")\
  .start()
```

Cancel

(1) Spark Jobs

My_Execution (id: 19931dc5-14ca-4c04-8dc5-68035a5b9bea) *Last updated: 5 seconds ago*

Cmd 9

```
spark.sql("SELECT * FROM device_counts").show()
```

(3) Spark Jobs

```
+----------+--------+--------------------+--------------------+------------------+
|        gt|  device|   avg(Arrival_Time)|  avg(Creation_Time)|        avg(Index)|
+----------+--------+--------------------+--------------------+------------------+
|       sit|    null|1.424741207868242...|1.424742112220356...| 74577.84690275553|
|     stand|    null|1.424743637921220...|1.424744579547463...|31317.877585550017|
|      null|nexus4_1|1.424745948293702E12|1.424745947996588...|172877.10556693148|
|      null|nexus4_2|1.424748432585106...|1.424750278191438...| 221445.9827479782|
|       sit|nexus4_1|1.424741711833447...|1.424741712789980...|  73824.0260887563|
|  stairsup|nexus4_2|1.424746435113130...|1.424748280733558...|227282.96971718763|
|      null|    null|1.424749002876356...|1.424749919482132...| 219276.9663669269|
|      null|    null|1.424745913733565...|1.424746844881817...|174536.91224919248|
|      walk|    null|1.424746420641803E12|1.424747351060679...|149760.09974990616|
|      null|nexus4_2|1.424745879803032...|1.424747725428376...|176166.48202865507|
|stairsdown|nexus4_2|1.424745052412463...|1.424746898052936...| 230126.7973287425|
|      walk|nexus4_1|1.424746463839954...|1.424746465071124...| 150063.2682104983|
|  stairsup|    null|1.424745996101175...|1.424746915892740...|227912.96550673083|
|       sit|nexus4_2|1.424740682751081...|1.424742528415252...| 75363.30637839901|
|      bike|    null|1.424751134340001...|1.424752127369580...| 326459.6867328154|
|      null|nexus4_1|1.424749564691623...|1.424749566104161...|217140.18710138695|
|stairsdown|    null|1.424744591412865E12|1.424745503635642...|230452.44623187225|
|      walk|nexus4_2|1.424746378093942E12|1.424748223712842...|149461.49508631244|
|     stand|nexus4_2|1.424743176833233...|1.424745022474491...|31667.805502842235|
|stairsdown|nexus4_1|1.424744141399292...|1.424744142448540...|230770.33463484643|
+----------+--------+--------------------+--------------------+------------------+
only showing top 20 rows

Command took 0.21 seconds -- by vkoushikmuthyapu@unm.edu at 10/16/2019, 8:45:04 PM on Hw-1
```

- In my execution, I am executing a summary table by implementing aggregations, I have taken the cube on activities,device and average of the values of Arrival_Time, Creation_Time and Index by dropping average of x, y, z.
- I have used memory as my output destination and Complete as my output-mode.
- We will see a summary table with columns gt, device, avg(Arrival_Time), avg(Creation_Time), avg(Index),