In [1]:
```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
from pyspark.sql.types import *

spark = SparkSession\
    .builder\
    .appName("chapter-21-streaming")\
    .getOrCreate()

import os
SPARK_BOOK_DATA_PATH = os.environ['SPARK_BOOK_DATA_PATH']
```

dataset - Heterogeneity Human Activity Recognition

The data consists of smartphone and smartwatch sensor readings from a variety of devices (such as accelerometer, gyroscope), sampled at the highest possible frequency supported by the devices. Readings from these sensors weree recorded while users perfomed activities like biking, sitting, standing, walking, and so on

In [2]:
```
file_path = SPARK_BOOK_DATA_PATH + "/data/activity-data/"
static = spark.read.json(file_path)
dataSchema = static.schema
```

In [3]:
```
static.show(5)
```

```
+------------+------------------+--------+-----+------+----+-----+--
----------+------------+------------+
| Arrival_Time|      Creation_Time|  Device|Index| Model|User|   gt|
x|           y|           z|
+------------+------------------+--------+-----+------+----+-----+--
----------+------------+------------+
|1424686735090|1424686733090638193|nexus4_1|   18|nexus4|   g|stand|
3.356934E-4|-5.645752E-4|-0.018814087|
|1424686735292|1424688581345918092|nexus4_2|   66|nexus4|   g|stand|-
0.005722046| 0.029083252| 0.005569458|
|1424686735500|1424686733498505625|nexus4_1|   99|nexus4|   g|stand|
0.0078125|-0.017654419| 0.010025024|
|1424686735691|1424688581745026978|nexus4_2|  145|nexus4|   g|stand|-
3.814697E-4|   0.0184021|-0.013656616|
|1424686735890|1424688581945252808|nexus4_2|  185|nexus4|   g|stand|-
3.814697E-4|-0.031799316| -0.00831604|
+------------+------------------+--------+-----+------+----+-----+--
----------+------------+------------+
only showing top 5 rows
```

In [4]:
```
## Extract

streaming = spark.readStream.schema(dataSchema)\
  .option("maxFilesPerTrigger", 1)\
  .json(file_path)
```

In [6]:
```
type(static), type(streaming), streaming.isStreaming
```

Out[6]:
```
(pyspark.sql.dataframe.DataFrame, pyspark.sql.dataframe.DataFrame, Tru
e)
```

In [8]:
```
# watch stream

from time import sleep

def show_streaming(SQL_stmt, ntimes=5, sleep_sec=1):
    for x in range(ntimes):
        spark.sql(SQL_stmt).show()
        sleep(sleep_sec)
```

use format = `memory` , other formats are: `console, socket, kafka`

In [9]:
```
## Load (action)

activityQuery = (
    streaming.groupBy("gt")   # Transform
    .count()
    .writeStream
    .queryName("activity_counts")
    .format("memory")
    .outputMode("complete")
    .start()
)

# activityQuery.awaitTermination()
```

In [11]:
```
show_streaming(SQL_stmt="SELECT * FROM activity_counts", ntimes=10,
```

```
+----------+------+
|        gt| count|
+----------+------+
|  stairsup|156805|
|       sit|184618|
|     stand|170782|
|      walk|198839|
|      bike|161964|
|stairsdown|140453|
|      null|156718|
+----------+------+

+----------+------+
|        gt| count|
+----------+------+
|  stairsup|167258|
|       sit|196925|
|     stand|182167|
|      walk|212095|
```

In [12]:
```
spark.streams.active
```

Out[12]:
```
[<pyspark.sql.streaming.StreamingQuery at 0x7fa730b8d940>]
```

In [13]:
```
# COMMAND ----------

from pyspark.sql.functions import expr
simpleTransform = streaming.withColumn("stairs", expr("gt like '%sta
  .where("stairs")\
  .where("gt is not null")\
  .select("gt", "model", "arrival_time", "creation_time")\
  .writeStream\
  .queryName("simple_transform2")\
  .format("memory")\
  .outputMode("append")\
  .start()
```

In [14]:
```
1 show_streaming(SQL_stmt="select * from simple_transform2")
```

```
+---+-----+------------+-------------+
| gt|model|arrival_time|creation_time|
+---+-----+------------+-------------+
+---+-----+------------+-------------+

+---+-----+------------+-------------+
| gt|model|arrival_time|creation_time|
+---+-----+------------+-------------+
+---+-----+------------+-------------+

+--------+------+-------------+-------------------+
|      gt| model| arrival_time|      creation_time|
+--------+------+-------------+-------------------+
|stairsup|nexus4|1424687983730|1424687981736873519|
|stairsup|nexus4|1424687984021|1424687982023708236|
|stairsup|nexus4|1424687984422|1424687982431691365|
|stairsup|nexus4|1424687984826|1424687982835622029|
|stairsup|nexus4|1424687985228|1424687983237459357|
|stairsup|nexus4|1424687985634|1424687983640474493|
|stairsup|nexus4|1424687986036|1424687984043306525|
|stairsup|nexus4|1424687986438|1424687984441042120|
|stairsup|nexus4|1424687986841|1424687984849086587|
|stairsup|nexus4|1424687987244|1424687985251949135|
|stairsup|nexus4|1424687987645|1424687985654671653|
|stairsup|nexus4|1424687988001|1424689834045929824|
|stairsup|nexus4|1424687988203|1424689834247498428|
|stairsup|nexus4|1424687988408|1424687986415017112|
|stairsup|nexus4|1424687988602|1424689834650147354|
|stairsup|nexus4|1424687988805|1424687986812996849|
|stairsup|nexus4|1424687989006|1424687987014412864|
|stairsup|nexus4|1424687989210|1424687987215792501|
|stairsup|nexus4|1424687989409|1424687987407290304|
|stairsup|nexus4|1424687989611|1424687987618807638|
+--------+------+-------------+-------------------+
only showing top 20 rows

+--------+------+-------------+-------------------+
|      gt| model| arrival_time|      creation_time|
+--------+------+-------------+-------------------+
|stairsup|nexus4|1424687983730|1424687981736873519|
|stairsup|nexus4|1424687984021|1424687982023708236|
|stairsup|nexus4|1424687984422|1424687982431691365|
|stairsup|nexus4|1424687984826|1424687982835622029|
|stairsup|nexus4|1424687985228|1424687983237459357|
|stairsup|nexus4|1424687985634|1424687983640474493|
|stairsup|nexus4|1424687986036|1424687984043306525|
|stairsup|nexus4|1424687986438|1424687984441042120|
|stairsup|nexus4|1424687986841|1424687984849086587|
|stairsup|nexus4|1424687987244|1424687985251949135|
|stairsup|nexus4|1424687987645|1424687985654671653|
|stairsup|nexus4|1424687988001|1424689834045929824|
|stairsup|nexus4|1424687988203|1424689834247498428|
|stairsup|nexus4|1424687988408|1424687986415017112|
|stairsup|nexus4|1424687988602|1424689834650147354|
|stairsup|nexus4|1424687988805|1424687986812996849|
```

```
|stairsup|nexus4|1424687989006|1424687987014412864|
|stairsup|nexus4|1424687989210|1424687987215792501|
|stairsup|nexus4|1424687989409|1424687987407290304|
|stairsup|nexus4|1424687989611|1424687987618807638|
+--------+------+-------------+-------------------+
only showing top 20 rows

+--------+------+-------------+-------------------+
|      gt| model| arrival_time|      creation_time|
+--------+------+-------------+-------------------+
|stairsup|nexus4|1424687983730|1424687981736873519|
|stairsup|nexus4|1424687984021|1424687982023708236|
|stairsup|nexus4|1424687984422|1424687982431691365|
|stairsup|nexus4|1424687984826|1424687982835622029|
|stairsup|nexus4|1424687985228|1424687983237459357|
|stairsup|nexus4|1424687985634|1424687983640474493|
|stairsup|nexus4|1424687986036|1424687984043306525|
|stairsup|nexus4|1424687986438|1424687984441042120|
|stairsup|nexus4|1424687986841|1424687984849086587|
|stairsup|nexus4|1424687987244|1424687985251949135|
|stairsup|nexus4|1424687987645|1424687985654671653|
|stairsup|nexus4|1424687988001|1424689834045929824|
|stairsup|nexus4|1424687988203|1424689834247498428|
|stairsup|nexus4|1424687988408|1424687986415017112|
|stairsup|nexus4|1424687988602|1424689834650147354|
|stairsup|nexus4|1424687988805|1424687986812996849|
|stairsup|nexus4|1424687989006|1424687987014412864|
|stairsup|nexus4|1424687989210|1424687987215792501|
|stairsup|nexus4|1424687989409|1424687987407290304|
|stairsup|nexus4|1424687989611|1424687987618807638|
+--------+------+-------------+-------------------+
only showing top 20 rows
```

In [15]:

```
# COMMAND ----------

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()
```

In [16]:
```
1 show_streaming(SQL_stmt="select * from device_counts")
```

```
+---+-----+------+------+------+
| gt|model|avg(x)|avg(y)|avg(z)|
+---+-----+------+------+------+
+---+-----+------+------+------+

+---+-----+------+------+------+
| gt|model|avg(x)|avg(y)|avg(z)|
+---+-----+------+------+------+
+---+-----+------+------+------+

+---+-----+------+------+------+
| gt|model|avg(x)|avg(y)|avg(z)|
+---+-----+------+------+------+
+---+-----+------+------+------+

+---+-----+------+------+------+
| gt|model|avg(x)|avg(y)|avg(z)|
+---+-----+------+------+------+
+---+-----+------+------+------+

+----------+------+--------------------+--------------------+---------
-----------+
|        gt| model|              avg(x)|              avg(y)|
avg(z)|
+----------+------+--------------------+--------------------+---------
-----------+
|      null|nexus4|-0.00786799708513591|-0.00148733897879...|0.0065174
00118233163|
|      null|nexus4|0.002493916706910...|-0.00693672737540...|-0.009995
28491813...|
|      null|  null|0.002493916706910...|-0.00693672737540...|-0.009995
28491813...|
|      bike|nexus4| 0.02756607566542555|-0.01192289439331294|-0.080142
16329739748|
|     stand|  null|-4.34675125278877...|5.469481713131301E-4|3.3709793
98682453E-4|
|       sit|nexus4|-5.24316487285726...|1.505994925989123...|-4.254466
21983686...|
|     stand|nexus4|-4.34675125278877...|5.469481713131301E-4|3.3709793
98682453E-4|
|stairsdown|  null| 0.02934167046623987| -0.0372512654500534| 0.122021
76476404009|
|  stairsup|  null|-0.02644122577495918|-0.01006484684695...| -0.10257
45279937134|
|       sit|  null|-5.24316487285726...|1.505994925989123...|-4.254466
21983686...|
|  stairsup|nexus4|-0.02644122577495918|-0.01006484684695...| -0.10257
45279937134|
|      walk|  null|-5.99343870247432...|0.003705911656842188|-0.004313
96913741...|
|stairsdown|nexus4| 0.02934167046623987| -0.0372512654500534| 0.122021
76476404009|
|      bike|  null| 0.02756607566542555|-0.01192289439331294|-0.080142
16329739748|
|      walk|nexus4|-5.99343870247432...|0.003705911656842188|-0.004313
```

```
96913741...|
|      null|  null|-0.00786799708513591|-0.00148733897879...|0.0065174
00118233163|
+----------+------+--------------------+--------------------+---------
-----------+
```

In [ ]:

```
1 # COMMAND ----------
2
3 historicalAgg = static.groupBy("gt", "model").avg()
```

In [20]:

```
1 historicalAgg.show()
```

```
+----------+------+--------------------+--------------------+---------
---------+--------------------+--------------------+------------------
--+
|        gt| model|   avg(Arrival_Time)|  avg(Creation_Time)|        a
vg(Index)|              avg(x)|              avg(y)|              avg
(z)|
+----------+------+--------------------+--------------------+---------
---------+--------------------+--------------------+------------------
--+
|      bike|nexus4|1.424751134339985E12|1.424752127369588...| 326459.6
867328154|  0.0226887595508668|-0.00877912156368...|-0.082510016634123
43|
|      walk|nexus4|1.424746420641789...|1.424747351060674...|149760.09
974990616|-0.00390116006094...|0.001052508689953...|-6.9543555304299
7...|
|stairsdown|nexus4|1.424744591412857...|1.424745503635636...|230452.44
623187225|0.021613908669165474|-0.03249018824752616| 0.120359226915040
75|
|       sit|nexus4|1.424741207868231E12|1.424742112220355...| 74577.84
690275553|-5.49433244039557...|2.791446281700046E-4|-2.3399446168990
4...|
|     stand|nexus4|1.424743637921209...|1.424744579547460...|31317.877
585550017|-3.11082189691711...|3.218461665975360...|2.14130004063649
8...|
|      null|nexus4|1.424749002876339...|1.424749919482127...| 219276.9
663669269|-0.00847688860109...|-7.30455258739191...|0.0030906014914199
28|
|  stairsup|nexus4|1.424745996101162...|1.424746915892737...|227912.96
550673083|-0.02479965287771642|-0.00800392344379...|-0.100340884150604
02|
+----------+------+--------------------+--------------------+---------
---------+--------------------+--------------------+------------------
--+
```

In [17]:
```
deviceModelStats = (
    streaming.drop("Arrival_Time", "Creation_Time", "Index")
  .cube("gt", "model").avg()
  .join(historicalAgg, ["gt", "model"])
  .writeStream.queryName("join_hist")
  .format("memory")
  .outputMode("complete")
  .start()
)
```

In [19]:
```
show_streaming(SQL_stmt="select * from join_hist", ntimes=10)
```

```
+----------+------+--------------------+--------------------+--------
------------+--------------------+--------------------+-------------
----+--------------------+--------------------+--------------------+
|        gt| model|              avg(x)|              avg(y)|
avg(z)|   avg(Arrival_Time)|  avg(Creation_Time)|        avg(Index)|
avg(x)|              avg(y)|              avg(z)|
+----------+------+--------------------+--------------------+--------
------------+--------------------+--------------------+-------------
----+--------------------+--------------------+--------------------+
|      bike|nexus4| 0.02756607566542555|-0.01192289439331294|-0.08014
216329739748|1.424751134339985E12|1.424752127369588...| 326459.686732
8154|  0.0226887595508668|-0.00877912156368...|-0.08251001663412343|
|      walk|nexus4|-5.99343870247432...|0.003705911656842188|-0.00431
396913741...|1.424746420641789...|1.424747351060674...|149760.0997499
0616|-0.00390116006094...|0.001052508689953...|-6.95435553042997...|
|stairsdown|nexus4| 0.02934167046623987| -0.0372512654500534| 0.12202
176476404009|1.424744591412857...|1.424745503635636...|230452.4462318
7225|0.021613908669165474|-0.03249018824752616| 0.12035922691504075|
|       sit|nexus4|-5.24316487285726...|1.505994925989123...|-4.25446
```

In [ ]:
```

```

In [ ]:
```

```

see `chapter-21-stream-kafka.ipynb` for example using streaming with Kafka

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```
# Subscribe to 1 topic
df1 = spark.readStream.format("kafka")\
  .option("kafka.bootstrap.servers", "host1:port1,host2:port2")\
  .option("subscribe", "topic1")\
  .load()
```

In [ ]:

```
# Subscribe to multiple topics
df2 = spark.readStream.format("kafka")\
  .option("kafka.bootstrap.servers", "host1:port1,host2:port2")\
  .option("subscribe", "topic1,topic2")\
  .load()
# Subscribe to a pattern
df3 = spark.readStream.format("kafka")\
  .option("kafka.bootstrap.servers", "host1:port1,host2:port2")\
  .option("subscribePattern", "topic.*")\
  .load()


# COMMAND ----------

df1.selectExpr("topic", "CAST(key AS STRING)", "CAST(value AS STRING
  .writeStream\
  .format("kafka")\
  .option("kafka.bootstrap.servers", "host1:port1,host2:port2")\
  .option("checkpointLocation", "/to/HDFS-compatible/dir")\
  .start()
df1.selectExpr("CAST(key AS STRING)", "CAST(value AS STRING)")\
  .writeStream\
  .format("kafka")\
  .option("kafka.bootstrap.servers", "host1:port1,host2:port2")\
  .option("checkpointLocation", "/to/HDFS-compatible/dir")\
  .option("topic", "topic1")\
  .start()


# COMMAND ----------

socketDF = spark.readStream.format("socket")\
  .option("host", "localhost").option("port", 9999).load()


# COMMAND ----------

activityCounts.writeStream.trigger(processingTime='5 seconds')\
  .format("console").outputMode("complete").start()


# COMMAND ----------

activityCounts.writeStream.trigger(once=True)\
  .format("console").outputMode("complete").start()


# COMMAND ----------
```