Sessionize

October 26, 2017

In [1]:
```
from pyspark.sql import SparkSession

# Build the SparkSession
spark = SparkSession.builder \
    .master("local") \
    .appName("Sessionize IP addresses") \
    .config("spark.executor.memory", "1gb") \
    .getOrCreate()
sc = spark.sparkContext
```

In [2]:
```
# Load the data by creating rdd
rdd = sc.textFile('/home/hassan/Side_Projects/WeblogChallenge/data/2015_07_22_mktplace_s
# split the data into columns
rdd = rdd.map(lambda line: line.split(" "))
```

In [3]:
```
# ====================================
# Manipulating data
# ====================================
from pyspark.sql import Row
from pyspark.sql.types import *
from pyspark.sql.functions import *

#Map the RDD to a DF for better performance
mainDF = rdd.map(lambda line: Row(timestamp=line[0], ipaddress=line[2].split(':')[0],url
mainDF.show(20)
```

```
+---------------+--------------------+--------------------+
|      ipaddress|           timestamp|                 url|
+---------------+--------------------+--------------------+
|123.242.248.130|2015-07-22T09:00:...|https://paytm.com...|
|  203.91.211.44|2015-07-22T09:00:...|https://paytm.com...|
|    1.39.32.179|2015-07-22T09:00:...|https://paytm.com...|
| 180.179.213.94|2015-07-22T09:00:...|https://paytm.com...|
| 120.59.192.208|2015-07-22T09:00:...|https://paytm.com...|
| 117.239.195.66|2015-07-22T09:00:...|https://paytm.com...|
|  101.60.186.26|2015-07-22T09:00:...|https://paytm.com...|
|   59.183.41.47|2015-07-22T09:00:...|https://paytm.com...|
| 117.239.195.66|2015-07-22T09:00:...|https://paytm.com...|
```

```
|  183.83.237.83|2015-07-22T09:00:...|https://paytm.com...|
|  117.195.91.36|2015-07-22T09:00:...|https://paytm.com...|
|122.180.245.251|2015-07-22T09:00:...|https://paytm.com...|
| 117.198.215.20|2015-07-22T09:00:...|https://paytm.com...|
| 223.176.154.91|2015-07-22T09:00:...|https://paytm.com...|
|223.225.236.110|2015-07-22T09:00:...|https://paytm.com...|
| 117.241.97.140|2015-07-22T09:00:...|https://paytm.com...|
|117.205.247.140|2015-07-22T09:00:...|https://paytm.com...|
|   14.102.53.58|2015-07-22T09:00:...|https://paytm.com...|
|  203.200.99.67|2015-07-22T09:00:...|https://paytm.com...|
|107.167.109.204|2015-07-22T09:00:...|https://paytm.com...|
+---------------+--------------------+--------------------+
only showing top 20 rows
```

```
In [4]: # convert timestamps from string to timestamp datatype
        mainDF = mainDF.withColumn('timestamp', mainDF['timestamp'].cast(TimestampType()))

In [5]: # sessionizing data based on 15 min fixed window time
        # assign an Id to each session
        SessionDF = mainDF.select(window("timestamp", "15 minutes").alias('FixedTimeWindow'),'ti
        SessionDF = SessionDF.withColumn("SessionId", monotonically_increasing_id())
        SessionDF.show(20,False)
```

```
+---------------------------------------------+--------------+------------------------+--------
|FixedTimeWindow                              |ipaddress     |NumberHitsInSessionForIp|SessionI
+---------------------------------------------+--------------+------------------------+--------
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|1.38.17.231   |14                      |0
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|161.51.16.10  |1                       |1
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|117.213.93.103|3                       |2
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|165.225.104.65|35                      |3
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|1.39.46.218   |7                       |4
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|122.160.168.148|2                      |5
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|106.219.13.17 |2                       |6
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|106.76.90.62  |14                      |7
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|182.74.140.218|2                       |8
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|117.237.13.128|10                      |9
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|49.14.48.156  |1                       |10
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|125.16.14.134 |4                       |11
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|120.60.31.116 |3                       |12
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|183.82.99.148 |102                     |13
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|117.242.229.95|3                       |14
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|117.203.165.166|4                      |15
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|182.75.0.219  |1                       |16
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|90.216.134.197|2                       |17
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|115.117.83.101|9                       |18
|[2015-07-22 05:00:00.0,2015-07-22 05:15:00.0]|119.235.48.219|3                       |19
```

```
+----------------------------------------------+--------------+----------------------+--------
only showing top 20 rows
```

```
In [6]: # join the time stamps and url to the Sessionized DF
        dfWithTimeStamps = mainDF.select(window("timestamp", "15 minutes").alias('FixedTimeWindo
        SessionDF = dfWithTimeStamps.join(SessionDF,['FixedTimeWindow','ipaddress'])
        SessionDF.show(20)
```

```
+--------------------+--------------+--------------------+--------------------+----------------
|     FixedTimeWindow|     ipaddress|           timestamp|                 url|NumberHitsInSess
+--------------------+--------------+--------------------+--------------------+----------------
|[2015-07-21 22:30...|  106.51.141.73|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|107.167.109.115|2015-07-21 22:43:...|http://www.paytm...|
|[2015-07-21 22:30...|113.193.203.163|2015-07-21 22:41:...|https://paytm.com...|
|[2015-07-21 22:30...|  115.184.19.68|2015-07-21 22:43:...|https://paytm.com...|
|[2015-07-21 22:30...|  115.250.103.3|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|  115.250.103.3|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|  115.250.103.3|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|116.203.129.121|2015-07-21 22:43:...|https://paytm.com...|
|[2015-07-21 22:30...|116.203.129.121|2015-07-21 22:43:...|https://paytm.com...|
|[2015-07-21 22:30...|116.203.129.121|2015-07-21 22:43:...|https://www.paytm...|
|[2015-07-21 22:30...|116.203.129.121|2015-07-21 22:43:...|https://paytm.com...|
|[2015-07-21 22:30...|116.203.129.121|2015-07-21 22:43:...|https://paytm.com...|
|[2015-07-21 22:30...|  117.198.45.19|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|  117.198.45.19|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|  117.198.45.19|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|117.199.132.124|2015-07-21 22:43:...|https://paytm.com...|
|[2015-07-21 22:30...|117.253.213.104|2015-07-21 22:43:...|https://paytm.com...|
|[2015-07-21 22:30...|117.253.213.104|2015-07-21 22:44:...|https://paytm.com...|
|[2015-07-21 22:30...|  120.60.191.85|2015-07-21 22:41:...|http://paytm.com:...|
|[2015-07-21 22:30...|  120.60.191.85|2015-07-21 22:41:...|https://paytm.com...|
+--------------------+--------------+--------------------+--------------------+----------------
only showing top 20 rows
```

```
In [8]: # Finding the first hit time of each ip for each session and join in to our session df
        FirstHitTimeStamps = SessionDF.groupBy("SessionId").agg(min("timestamp").alias('FristHit
        SessionDF = FirstHitTimeStamps.join(SessionDF,['SessionId'])
        SessionDF.select(col("SessionId"),col("ipaddress"),col("FristHitTime")).show(20)
```

```
+---------+--------------+--------------------+
|SessionId|     ipaddress|        FristHitTime|
+---------+--------------+--------------------+
|       26|  218.248.82.9|2015-07-22 05:02:...|
|       26|  218.248.82.9|2015-07-22 05:02:...|
|       26|  218.248.82.9|2015-07-22 05:02:...|
```

```
|       26|  218.248.82.9|2015-07-22 05:02:...|
|       29|  27.62.30.188|2015-07-22 05:02:...|
|       29|  27.62.30.188|2015-07-22 05:02:...|
|       29|  27.62.30.188|2015-07-22 05:02:...|
|       29|  27.62.30.188|2015-07-22 05:02:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
|      474|101.221.128.95|2015-07-22 06:35:...|
+---------+--------------+--------------------+
only showing top 20 rows
```

In [9]:
```
#2. Determine the average session time
# Among all the hits in a session the last one has the max diff with first hit
# we define the time difference of first and last hit in a session to be the duration of
# if there is only one hit in a session the duration is zero
timeDiff = (unix_timestamp(SessionDF.timestamp)-unix_timestamp(SessionDF.FristHitTime))
SessionDF = SessionDF.withColumn("timeDiffwithFirstHit", timeDiff)
tmpdf = SessionDF.groupBy("SessionId").agg(max("timeDiffwithFirstHit").alias("SessionDur
SessionDF = SessionDF.join(tmpdf,['SessionId'])
SessionDF.select(col("SessionId"),col("ipaddress"),col("SessionDuration")).show(20)
```

```
+---------+--------------+---------------+
|SessionId|     ipaddress|SessionDuration|
+---------+--------------+---------------+
|       26|  218.248.82.9|             13|
|       26|  218.248.82.9|             13|
|       26|  218.248.82.9|             13|
|       26|  218.248.82.9|             13|
|       29|  27.62.30.188|             33|
|       29|  27.62.30.188|             33|
|       29|  27.62.30.188|             33|
|       29|  27.62.30.188|             33|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
```

```
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
|      474|101.221.128.95|            226|
+---------+--------------+---------------+
only showing top 20 rows
```

```
In [ ]: # showing the mean session duration
        # the printed number is secconds
        meandf = SessionDF.groupBy().avg('SessionDuration')
        meandf.show()
```

```
In [36]: #3. Determine unique URL visits per session. To clarify, count a hit to a unique URL on
         dfURL = SessionDF.groupBy("SessionId","URL").count().distinct().withColumnRenamed('coun
         dfURL.show(20)
```

```
+----------+--------------------+-----------+
| SessionId|                 URL|hitURLcount|
+----------+--------------------+-----------+
|        26|https://paytm.com...|          2|
|        26|http://www.paytm...|          2|
|        29|https://paytm.com...|          1|
|        29|https://paytm.com...|          1|
|        29|https://paytm.com...|          1|
|        29|https://paytm.com...|          1|
|       474|https://paytm.com...|          2|
|       474|https://paytm.com...|          2|
|       474|https://paytm.com...|          2|
|       474|https://paytm.com...|          5|
|       474|https://paytm.com...|          3|
|       474|https://paytm.com...|          2|
|       474|https://paytm.com...|          1|
|       474|https://paytm.com...|          1|
|       474|https://paytm.com...|          1|
|       474|https://paytm.com...|          1|
|8589934658|https://paytm.com...|          1|
|8589934965|https://paytm.com...|          1|
|8589934965|https://paytm.com...|          1|
|8589935171|https://paytm.com...|          1|
+----------+--------------------+-----------+
only showing top 20 rows
```

In [37]:
```
#4. Find the most engaged users, ie the IPs with the longest session times
EngagedUsers = SessionDF.select("ipaddress","SessionID","SessionDuration").sort(col("Se
EngagedUsers.show(2)
```

```
+--------------+------------+---------------+
|     ipaddress|   SessionID|SessionDuration|
+--------------+------------+---------------+
|164.100.96.254|249108103236|            847|
|111.119.199.22|283467841590|            839|
+--------------+------------+---------------+
only showing top 2 rows
```

In [ ]: