

TASK 1

NAME: YEO ZHENG XU ISAAC

STUDENT ID: 6342425

You are to perform exploratory analysis on the given dataset of click events by using Spark's DataFrame API.

The objective is to compute the average time that users stay on items in each category.

For analysis purposes in this task, use the following definitions:

(i) There are 15 item categories in the dataset: S, 0, 1 to 12, and B (for any 8-10 digits number)

(ii) In each session, the time that a user stays on some item is the timestamp difference between a user clicking on this item and the next item (if there is a next item). ¶

```
In [46]: from IPython.display import display
```

```
In [47]: from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "last"
import numpy as np
```

```
In [48]: import findspark
findspark.init()
```

```
In [49]: from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *
from pyspark.sql.window import Window
from pyspark.mllib.linalg.distributed import *
from pyspark.mllib.linalg import *
```

```
In [50]: spark = SparkSession.builder.appName("task1").config("spark-master", "local").getOrCreate()
sc = spark.sparkContext
spark
```

Out[50]: **SparkSession - in-memory**
SparkContext

[Spark UI \(http://WIN-98G7QMLAHN6:4040\)](http://WIN-98G7QMLAHN6:4040)

Version

v2.4.4

Master

local[*]

AppName

task1

Load Data & Defining Dataframe structure

```
In [51]: schema = StructType([
    StructField("session_id", StringType(), True),
    StructField("timestamp", TimestampType(), True),
    StructField("item_id", StringType(), True),
    StructField("category", StringType(), True),
])

timestampFormat= "yyyy-MM-dd'T'HH:mm:ss.SSS'Z'"
data = spark.read.csv("F:\\316\\yoochoose-clicks.dat",
                      schema = schema,
                      timestampFormat=timestampFormat)
```

```
In [52]: data.printSchema()
```

```
root
|-- session_id: string (nullable = true)
|-- timestamp: timestamp (nullable = true)
|-- item_id: string (nullable = true)
|-- category: string (nullable = true)
```

```
In [53]: data.show(10, truncate = False)
```

```
+-----+-----+-----+-----+
|session_id|timestamp          |item_id |category|
+-----+-----+-----+-----+
|1          |2014-04-07 10:51:09.277|214536502|0
|1          |2014-04-07 10:54:09.868|214536500|0
|1          |2014-04-07 10:54:46.998|214536506|0
|1          |2014-04-07 10:57:00.306|214577561|0
|2          |2014-04-07 13:56:37.614|214662742|0
|2          |2014-04-07 13:57:19.373|214662742|0
|2          |2014-04-07 13:58:37.446|214825110|0
|2          |2014-04-07 13:59:50.71 |214757390|0
|2          |2014-04-07 14:00:38.247|214757407|0
|2          |2014-04-07 14:02:36.889|214551617|0
+-----+-----+-----+-----+
only showing top 10 rows
```

Calculate and display the total number of category values

```
In [55]: distinct_category = data.select("category").distinct()
distinct_category.show()
print("Total number category values:", distinct_category.count())
```

```
+-----+
| category|
+-----+
|2088937100|
|2089421594|
|2089318108|
|2088901091|
|2088927353|
|2089329443|
| 5862365|
|      7|
|2089584296|
|2089156240|
|2088902668|
|2088966970|
|2089074648|
|2089155957|
|2089498194|
|2089084286|
|2088937230|
|      11|
|2089282830|
|2089443905|
+-----+
```

only showing top 20 rows

Total number category values: 339

Display number of distinct category values**S, 0, 1 - 12 & B**

```
In [56]: data = data.withColumn("category",when(length("category")>2,"B").otherwise(col("category")))
distinct_category = data.select("category").distinct()
distinct_category.show()
print("Number of distinct category values:",distinct_category.count())
```

```
+-----+
|category|
+-----+
|       7|
|      11|
|       3|
|       8|
|       0|
|       B|
|       5|
|       6|
|       S|
|       9|
|       1|
|      10|
|       4|
|      12|
|       2|
+-----+
```

Number of distinct category valuse: 15

```
In [57]: data = data.orderBy("session_id","timestamp")
data.show(truncate=False)
```

session_id	timestamp	item_id	category
1	2014-04-07 10:51:09.277	214536502	0
1	2014-04-07 10:54:09.868	214536500	0
1	2014-04-07 10:54:46.998	214536506	0
1	2014-04-07 10:57:00.306	214577561	0
10000001	2014-09-08 10:35:38.841	214854230	S
10000001	2014-09-08 10:40:20.143	214556216	S
10000001	2014-09-08 10:40:36.704	214556212	S
10000001	2014-09-08 10:41:12.386	214854230	S
10000001	2014-09-08 10:48:34.245	214854125	S
10000002	2014-09-08 19:10:51.206	214849322	S
10000002	2014-09-08 19:13:31.104	214838094	S
10000002	2014-09-08 19:14:54.518	214714721	S
10000002	2014-09-08 19:33:38.355	214853711	S
10000003	2014-09-05 11:32:15.524	214853090	3
10000003	2014-09-05 11:34:25.159	214851326	3
10000003	2014-09-05 11:37:23.321	214853094	3
10000004	2014-09-05 13:14:45.867	214853090	3
10000004	2014-09-05 13:55:18.886	214851326	3
10000004	2014-09-05 13:56:28.356	214853090	3
10000004	2014-09-05 13:57:08.51	214851326	3

only showing top 20 rows

```
In [58]: my_window = Window.partitionBy("session_id").orderBy("timestamp")
data = data.withColumn("next_timestamp",lead(col("timestamp"),1, None).over(my_
window))
data = data.orderBy("session_id","timestamp")
data.show(truncate=False)
```

```
+-----+-----+-----+-----+-----+
-+
|session_id|timestamp                |item_id |category|next_timestamp
|
+-----+-----+-----+-----+-----+
-+
|1          |2014-04-07 10:51:09.277|214536502|0        |2014-04-07 10:54:09.86
8|
|1          |2014-04-07 10:54:09.868|214536500|0        |2014-04-07 10:54:46.99
8|
|1          |2014-04-07 10:54:46.998|214536506|0        |2014-04-07 10:57:00.30
6|
|1          |2014-04-07 10:57:00.306|214577561|0        |null
|
|10000001   |2014-09-08 10:35:38.841|214854230|S        |2014-09-08 10:40:20.14
3|
|10000001   |2014-09-08 10:40:20.143|214556216|S        |2014-09-08 10:40:36.70
4|
|10000001   |2014-09-08 10:40:36.704|214556212|S        |2014-09-08 10:41:12.38
6|
|10000001   |2014-09-08 10:41:12.386|214854230|S        |2014-09-08 10:48:34.24
5|
|10000001   |2014-09-08 10:48:34.245|214854125|S        |null
|
|10000002   |2014-09-08 19:10:51.206|214849322|S        |2014-09-08 19:13:31.10
4|
|10000002   |2014-09-08 19:13:31.104|214838094|S        |2014-09-08 19:14:54.51
8|
|10000002   |2014-09-08 19:14:54.518|214714721|S        |2014-09-08 19:33:38.35
5|
|10000002   |2014-09-08 19:33:38.355|214853711|S        |null
|
|10000003   |2014-09-05 11:32:15.524|214853090|3        |2014-09-05 11:34:25.15
9|
|10000003   |2014-09-05 11:34:25.159|214851326|3        |2014-09-05 11:37:23.32
1|
|10000003   |2014-09-05 11:37:23.321|214853094|3        |null
|
|10000004   |2014-09-05 13:14:45.867|214853090|3        |2014-09-05 13:55:18.88
6|
|10000004   |2014-09-05 13:55:18.886|214851326|3        |2014-09-05 13:56:28.35
6|
|10000004   |2014-09-05 13:56:28.356|214853090|3        |2014-09-05 13:57:08.51
|
|10000004   |2014-09-05 13:57:08.51 |214851326|3        |2014-09-05 13:57:59.65
9|
+-----+-----+-----+-----+-----+
-+
only showing top 20 rows
```

```
In [59]: data = data.where(col("next_timestamp").isNotNull())
data.show(truncate=False)
```

```
+-----+-----+-----+-----+-----+
-+
|session_id|timestamp                |item_id |category|next_timestamp
|
+-----+-----+-----+-----+-----+
-+
|1          |2014-04-07 10:51:09.277|214536502|0        |2014-04-07 10:54:09.86
8|
|1          |2014-04-07 10:54:09.868|214536500|0        |2014-04-07 10:54:46.99
8|
|1          |2014-04-07 10:54:46.998|214536506|0        |2014-04-07 10:57:00.30
6|
|10000001   |2014-09-08 10:35:38.841|214854230|S        |2014-09-08 10:40:20.14
3|
|10000001   |2014-09-08 10:40:20.143|214556216|S        |2014-09-08 10:40:36.70
4|
|10000001   |2014-09-08 10:40:36.704|214556212|S        |2014-09-08 10:41:12.38
6|
|10000001   |2014-09-08 10:41:12.386|214854230|S        |2014-09-08 10:48:34.24
5|
|10000002   |2014-09-08 19:10:51.206|214849322|S        |2014-09-08 19:13:31.10
4|
|10000002   |2014-09-08 19:13:31.104|214838094|S        |2014-09-08 19:14:54.51
8|
|10000002   |2014-09-08 19:14:54.518|214714721|S        |2014-09-08 19:33:38.35
5|
|10000003   |2014-09-05 11:32:15.524|214853090|3        |2014-09-05 11:34:25.15
9|
|10000003   |2014-09-05 11:34:25.159|214851326|3        |2014-09-05 11:37:23.32
1|
|10000004   |2014-09-05 13:14:45.867|214853090|3        |2014-09-05 13:55:18.88
6|
|10000004   |2014-09-05 13:55:18.886|214851326|3        |2014-09-05 13:56:28.35
6|
|10000004   |2014-09-05 13:56:28.356|214853090|3        |2014-09-05 13:57:08.51
|
|10000004   |2014-09-05 13:57:08.51 |214851326|3        |2014-09-05 13:57:59.65
9|
|10000004   |2014-09-05 13:57:59.659|214853248|S        |2014-09-05 13:59:33.96
|
|10000004   |2014-09-05 13:59:33.96 |214851326|3        |2014-09-05 14:00:05.95
5|
|10000004   |2014-09-05 14:00:05.955|214853094|3        |2014-09-05 14:06:42.48
9|
|10000006   |2014-09-05 17:37:18.748|214829261|1        |2014-09-05 17:37:59.43
5|
+-----+-----+-----+-----+-----+
-+
only showing top 20 rows
```

```
In [61]: data = data.withColumn("stay_time",
                                col("next_timestamp").cast("double")-col("timestamp").cast("double"))
data.show(truncate=False)
```

```
+-----+-----+-----+-----+-----+
+-----+
|session_id|timestamp          |item_id |category|next_timestamp
|stay_time      |
+-----+-----+-----+-----+-----+
+-----+
|1          |2014-04-07 10:51:09.277|214536502|0        |2014-04-07 10:54:09.86
8|180.59100008010864|
|1          |2014-04-07 10:54:09.868|214536500|0        |2014-04-07 10:54:46.99
8|37.12999987602234 |
|1          |2014-04-07 10:54:46.998|214536506|0        |2014-04-07 10:57:00.30
6|133.30800008773804|
|10000001   |2014-09-08 10:35:38.841|214854230|S        |2014-09-08 10:40:20.14
3|281.3019998073578 |
|10000001   |2014-09-08 10:40:20.143|214556216|S        |2014-09-08 10:40:36.70
4|16.561000108718872|
|10000001   |2014-09-08 10:40:36.704|214556212|S        |2014-09-08 10:41:12.38
6|35.681999921798706|
|10000001   |2014-09-08 10:41:12.386|214854230|S        |2014-09-08 10:48:34.24
5|441.8589999675751 |
|10000002   |2014-09-08 19:10:51.206|214849322|S        |2014-09-08 19:13:31.10
4|159.89800000190735|
|10000002   |2014-09-08 19:13:31.104|214838094|S        |2014-09-08 19:14:54.51
8|83.4139997959137 |
|10000002   |2014-09-08 19:14:54.518|214714721|S        |2014-09-08 19:33:38.35
5|1123.837000131607 |
|10000003   |2014-09-05 11:32:15.524|214853090|3        |2014-09-05 11:34:25.15
9|129.63499999046326|
|10000003   |2014-09-05 11:34:25.159|214851326|3        |2014-09-05 11:37:23.32
1|178.16200017929077|
|10000004   |2014-09-05 13:14:45.867|214853090|3        |2014-09-05 13:55:18.88
6|2433.018999814987 |
|10000004   |2014-09-05 13:55:18.886|214851326|3        |2014-09-05 13:56:28.35
6|69.47000002861023 |
|10000004   |2014-09-05 13:56:28.356|214853090|3        |2014-09-05 13:57:08.51
4|40.15400004386902 |
|10000004   |2014-09-05 13:57:08.51 |214851326|3        |2014-09-05 13:57:59.65
9|51.1489999294281 |
|10000004   |2014-09-05 13:57:59.659|214853248|S        |2014-09-05 13:59:33.96
|94.30100011825562 |
|10000004   |2014-09-05 13:59:33.96 |214851326|3        |2014-09-05 14:00:05.95
5|31.994999885559082|
|10000004   |2014-09-05 14:00:05.955|214853094|3        |2014-09-05 14:06:42.48
9|396.53400015830994|
|10000006   |2014-09-05 17:37:18.748|214829261|1        |2014-09-05 17:37:59.43
5|40.687000036239624|
+-----+-----+-----+-----+-----+
+-----+
only showing top 20 rows
```



```
In [62]: data.groupBy("category").agg((sum("stay_time") / countDistinct("session_id")).  
alias("avg_stay_time")).orderBy("category").show()
```

```
+-----+-----+  
|category|    avg_stay_time|  
+-----+-----+  
|      0| 414.4709083946925|  
|      1| 442.1116800819731|  
|     10| 398.3940570482952|  
|     11| 287.9408092237587|  
|     12| 376.0636118335499|  
|      2| 414.30921255268976|  
|      3| 338.9622899890425|  
|      4| 401.3212670098933|  
|      5| 402.2364771293502|  
|      6| 406.7753709809426|  
|      7| 400.57711884434644|  
|      8| 403.7298656270004|  
|      9| 337.38195319053904|  
|      B| 366.37772847468267|  
|      S| 379.07649009551193|  
+-----+-----+
```

```
In [ ]:
```