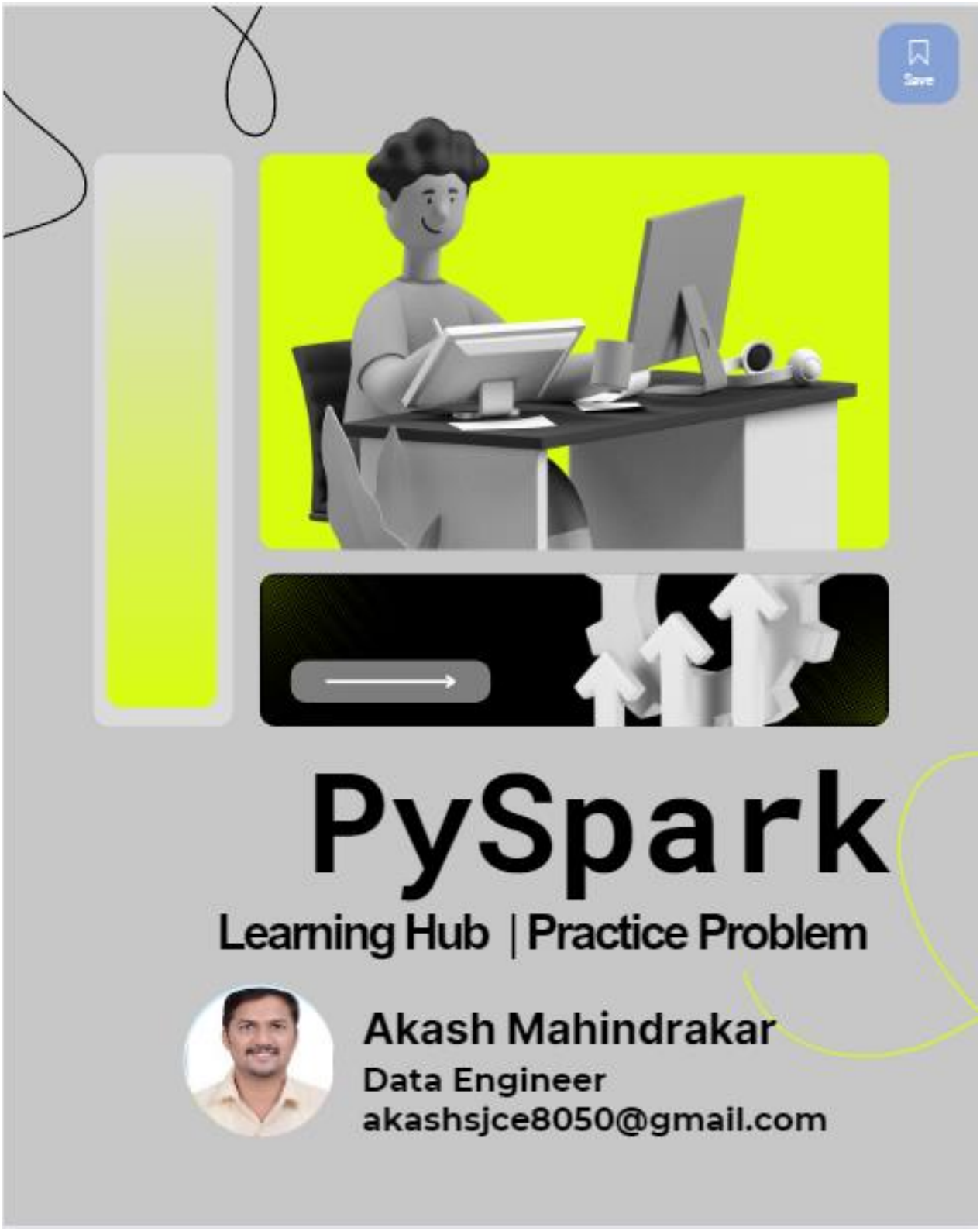
PYSPARK LEARNING HUB : DAY - 2



Step - 1 : Problem Statement

Ads Performance

Write an pyspark code to find the ctr of each Ad.Round ctr to 2 decimal points. Order the result table by ctr in descending order and by ad_id in ascending order in case of a tie.

Ctr=Clicked/(Clicked+Viewed)

Difficult Level : EASY

DataFrame:

```
# Define the schema for the Ads table
schema=StructType([
        StructField('AD_ID',IntegerType(),True)
        ,StructField('USER_ID',IntegerType(),True)
        ,StructField('ACTION',StringType(),True)
])

# Define the data for the Ads table
data = [
        (1, 1, 'Clicked'),
        (2, 2, 'Clicked'),
        (3, 3, 'Viewed'),
        (5, 5, 'Ignored'),
        (1, 7, 'Ignored'),
        (2, 7, 'Viewed'),
        (3, 5, 'Clicked'),
        (1, 4, 'Viewed'),
        (2, 11, 'Viewed'),
        (1, 2, 'Clicked')
]
```

Step - 2 : Identifying The Input Data And Expected

INPUT

INPUT		
AD_ID	USER_ID	ACTION
1	1	Clicked
2	2	Clicked
3	3	Viewed
5	5	Ignored
1	7	Ignored
2	7	Viewed
3	5	Clicked
1	4	Viewed
2	11	Viewed
1	2	Clicked

OUTPUT

OUTPUT	
AD_ID	CTR
1	0.67
3	0.5
2	0.33
5	0

Step - 3 : Writing the pyspark code to solve

```
# Creating Spark Session

from pyspark.sql import SparkSession
from pyspark.sql.types import
StructType,StructField,IntegerType,StringType
from pyspark.sql.functions import when
from pyspark.sql import functions as F
from pyspark.sql.window import Window

#creating spark session
spark = SparkSession. \
builder. \
config('spark.shuffle.useOldFetchProtocol', 'true'). \
config('spark.ui.port','0'). \
config("spark.sql.warehouse.dir", "/user/itv008042/warehouse"). \
enableHiveSupport(). \
master('yarn'). \
getOrCreate()


# Define the schema for the Ads table
schema=StructType([
        StructField('AD_ID',IntegerType(),True)
        ,StructField('USER_ID',IntegerType(),True)
        ,StructField('ACTION',StringType(),True)
])
```

```
# Define the data for the Ads table
data = [
        (1, 1, 'Clicked'),
        (2, 2, 'Clicked'),
        (3, 3, 'Viewed'),
        (5, 5, 'Ignored'),
        (1, 7, 'Ignored'),
        (2, 7, 'Viewed'),
        (3, 5, 'Clicked'),
        (1, 4, 'Viewed'),
        (2, 11, 'Viewed'),
        (1, 2, 'Clicked')
]

# Create a PySpark DataFrame
df=spark.createDataFrame(data,schema)
df.show()
```

```
+-----+-------+-------+
|AD_ID|USER_ID| ACTION|
+-----+-------+-------+
|    1|      1|Clicked|
|    2|      2|Clicked|
|    3|      3| Viewed|
|    5|      5|Ignored|
|    1|      7|Ignored|
|    2|      7| Viewed|
|    3|      5|Clicked|
|    1|      4| Viewed|
|    2|     11| Viewed|
|    1|      2|Clicked|
+-----+-------+-------+
```

```
ctr_df = (
        ads_df.groupBy("ad_id")
        .agg(
        F.sum(F.when(ads_df["action"] == "Clicked",
1).otherwise(0)).alias("click_count"),
        F.sum(F.when(ads_df["action"] == "Viewed",
1).otherwise(0)).alias("view_count")
        )
        .withColumn("ctr", F.round(F.col("click_count") /
(F.col("click_count") + F.col("view_count")), 2))
)

# Order the result table by CTR in descending order and by ad_id in
ascending order
window_spec = Window.orderBy(F.col("ctr").desc(),
F.col("ad_id").asc())
result_df = ctr_df.withColumn("rank", F.rank().over(window_spec))

# Show the result DataFrame
result_df.select('ad_id','ctr').show()
```

```
+-----+----+
|ad_id| ctr|
+-----+----+
|    1|0.67|
|    3| 0.5|
|    5|null|
|    2|0.33|
+-----+----+
```

PYSPARK LEARNING HUB : DAY - 2

Save

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share