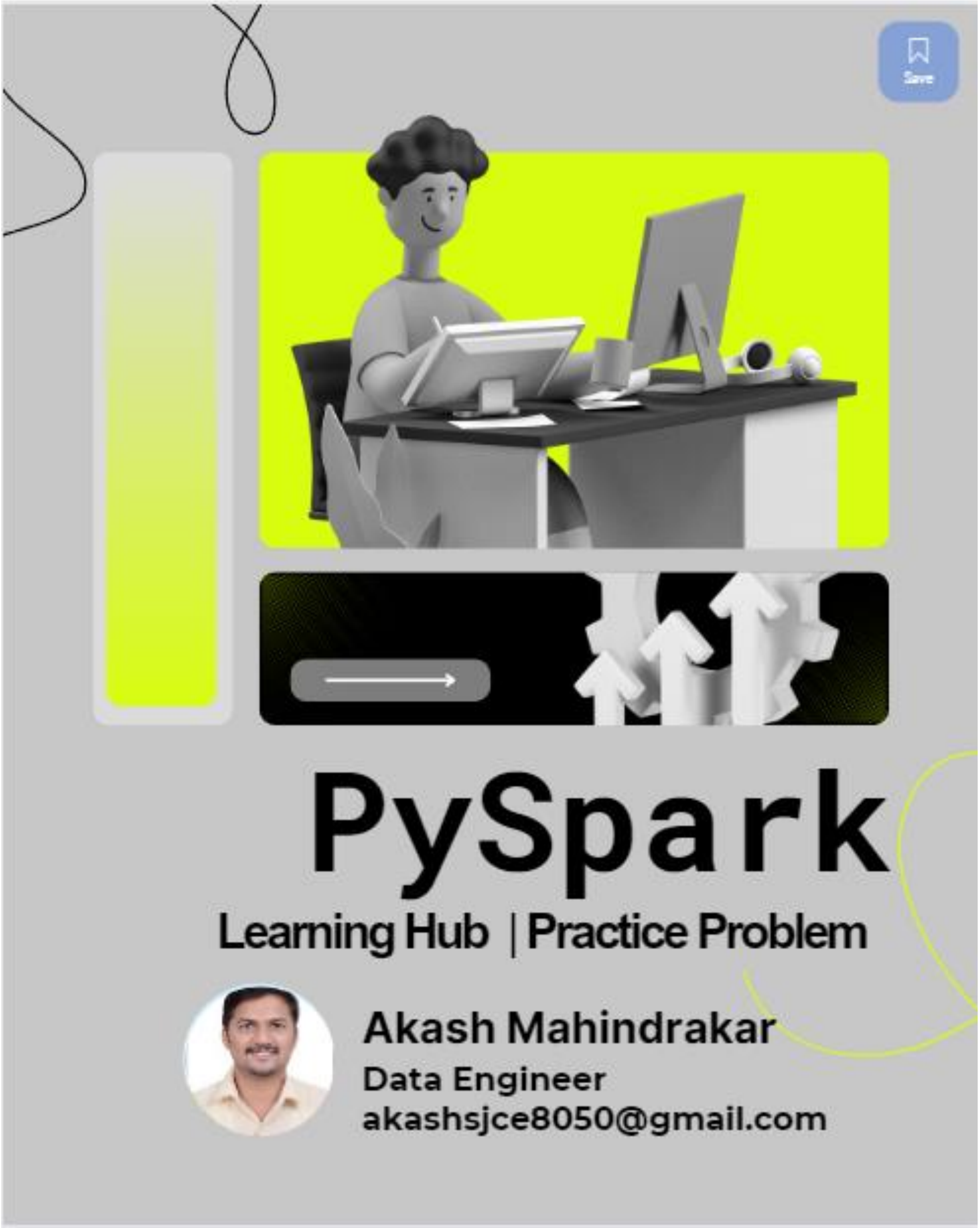
PYSPARK LEARNING HUB : DAY - 9



WWW.LINKEDIN.COM/IN/AKASHMAHINDRAKAR

Step - 1 : Problem Statement

09_Game Play Analysis II

Write a pyspark code that reports the device that is first logged in for each player.

Return the result table in any order.

Difficult Level : EASY

DataFrame:

```
# Define the schema for the "Activity"
activity_schema = StructType([
        StructField("player_id", IntegerType(), True),
        StructField("device_id", IntegerType(), True),
        StructField("event_date", StringType(), True),
        StructField("games_played", IntegerType(), True)
])

# Define data for the "Activity"
activity_data = [
        (1, 2, '2016-03-01', 5),
        (1, 2, '2016-05-02', 6),
        (2, 3, '2017-06-25', 1),
        (3, 1, '2016-03-02', 0),
        (3, 4, '2018-07-03', 5)
]
```

Step - 2 : Identifying The Input Data And Expected Output

INPUT

INPUT			
PLAYER_ID	DEVICE_ID	EVENT_DATE	GAMES_PLAYED
1	2	2016-03-01	5
1	2	2016-05-02	6
2	3	2017-06-25	1
3	1	2016-03-02	0
3	4	2018-07-03	5

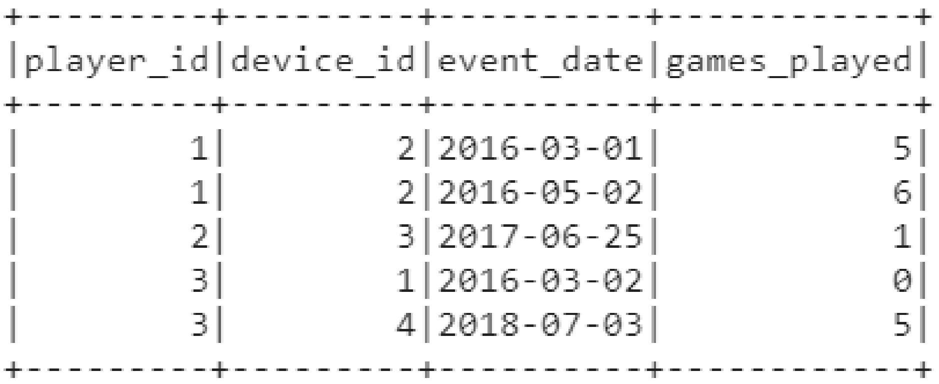
OUTPUT

OUTPUT	
PLAYER_ID	DEVICE_ID
1	2
2	3
3	1

Step - 3 : Writing the pyspark code to solve

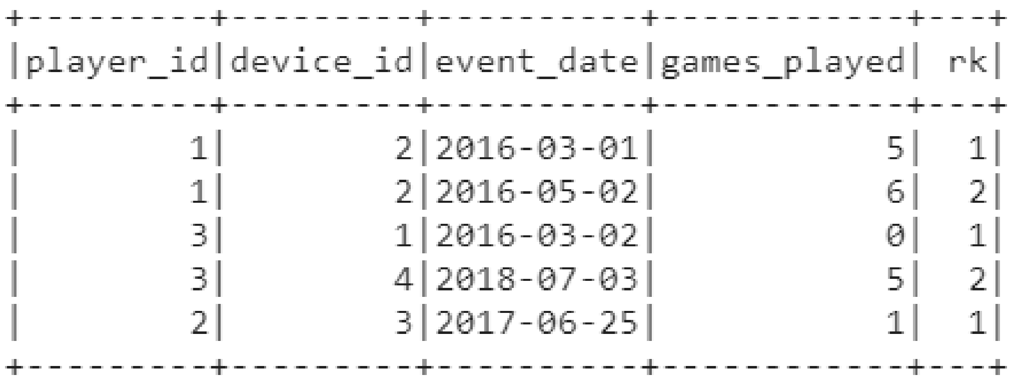
```
# Creating Spark Session
from pyspark.sql import SparkSession
from pyspark.sql.types import
StructType,StructField,IntegerType,StringType

#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 "Activity"
activity_schema = StructType([
        StructField("player_id", IntegerType(), True),
        StructField("device_id", IntegerType(), True),
        StructField("event_date", StringType(), True),
        StructField("games_played", IntegerType(), True)
])

# Define data for the "Activity"
activity_data = [
        (1, 2, '2016-03-01', 5),
        (1, 2, '2016-05-02', 6),
        (2, 3, '2017-06-25', 1),
        (3, 1, '2016-03-02', 0),
        (3, 4, '2018-07-03', 5)
]
```

```
# Create a PySpark DataFrame

df=spark.createDataFrame(activity_data,activity_schema)
df.show()
```

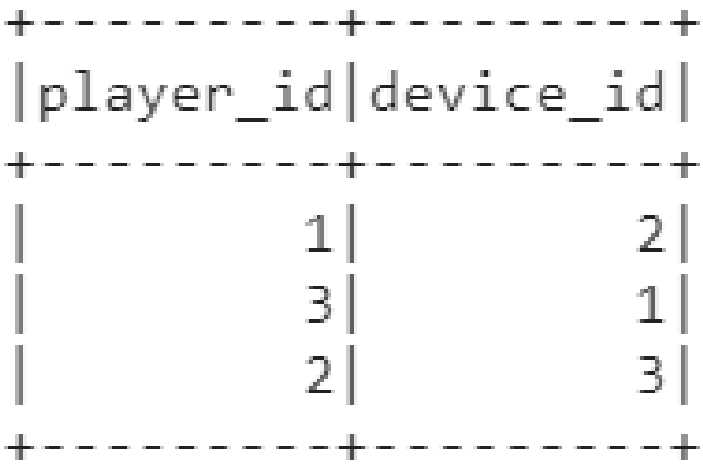


```
+---------+---------+----------+------------+
|player_id|device_id|event_date|games_played|
+---------+---------+----------+------------+
|        1|        2|2016-03-01|           5|
|        1|        2|2016-05-02|           6|
|        2|        3|2017-06-25|           1|
|        3|        1|2016-03-02|           0|
|        3|        4|2018-07-03|           5|
+---------+---------+----------+------------+
```

```
rank_df=df.withColumn("rk",rank().over(Window.partitionBy(df["player_id"]).orderBy(df["event_date"])))
rank_df.show()
```



```
+---------+---------+----------+------------+---+
|player_id|device_id|event_date|games_played| rk|
+---------+---------+----------+------------+---+
|        1|        2|2016-03-01|           5|  1|
|        1|        2|2016-05-02|           6|  2|
|        3|        1|2016-03-02|           0|  1|
|        3|        4|2018-07-03|           5|  2|
|        2|        3|2017-06-25|           1|  1|
+---------+---------+----------+------------+---+
rank_df.filter(rank_df["rk"] ==
1).select("player_id","device_id").show()
```



```
+---------+---------+
|player_id|device_id|
+---------+---------+
|        1|        2|
|        3|        1|
|        2|        3|
+---------+---------+
```

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

SHARE YOUR THOUGHTS IN COMMENT BELOW