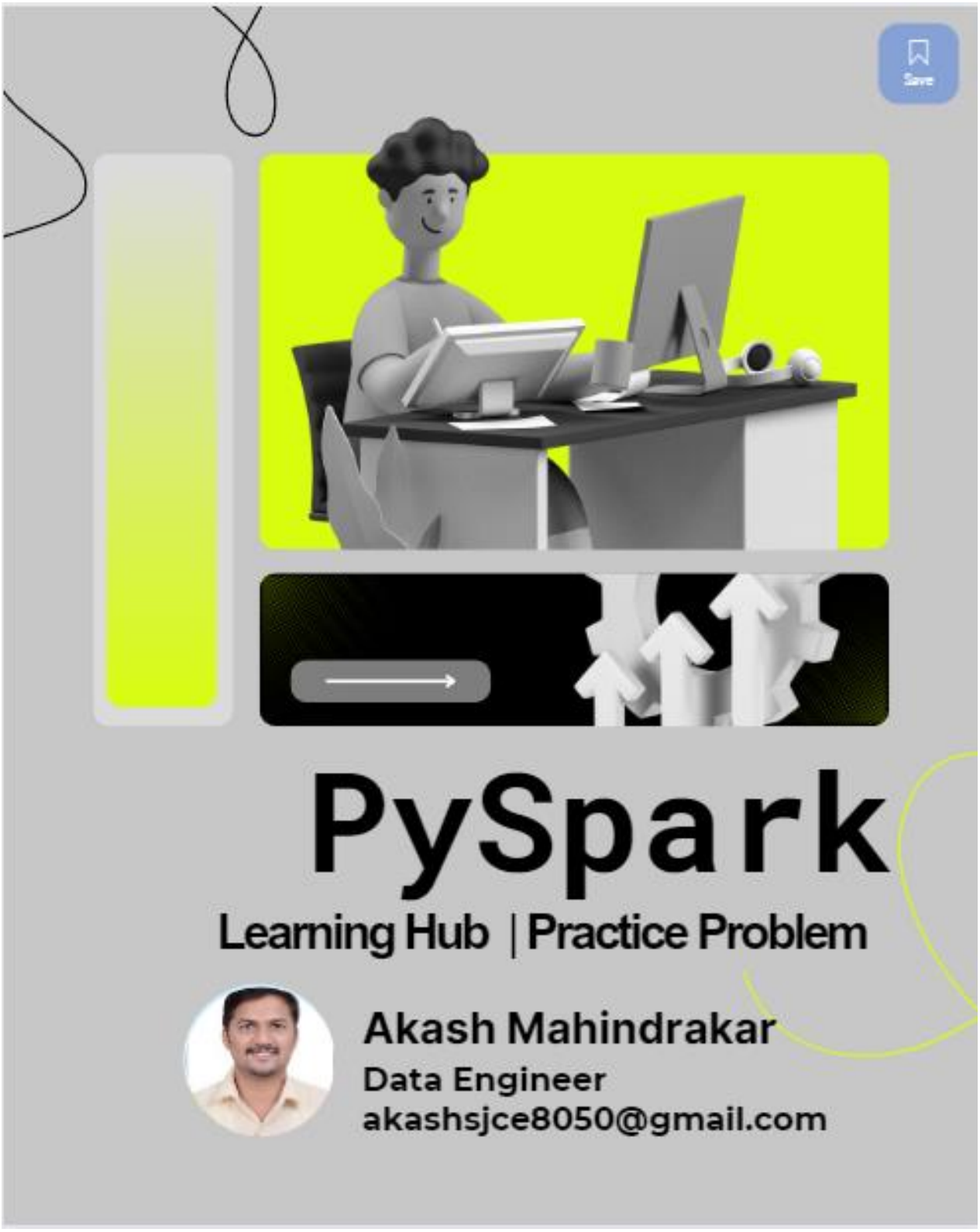
PYSPARK LEARNING HUB : DAY - 15



Step - 1 : Problem Statement

15_Teams Power Users

Write a pyspark code to identify the top 2 Power Users who sent the highest number of messages on Microsoft Teams in August 2022. Display the IDs of these 2 users along with the total number of messages they sent. Output the results in descending order based on the count of the messages.

Difficult Level : EASY

DataFrame:

```
schema = StructType([
        StructField("message_id", IntegerType(), True),
        StructField("sender_id", IntegerType(), True),
        StructField("receiver_id", IntegerType(), True),
        StructField("content", StringType(), True),
        StructField("sent_date", StringType(), True),
])

# Define the data
data = [
        (901, 3601, 4500, 'You up?', '2022-08-03 00:00:00'),
        (902, 4500, 3601, 'Only if you\'re buying', '2022-08-03 00:00:00'),
        (743, 3601, 8752, 'Let\'s take this offline', '2022-06-14 00:00:00'),
        (922, 3601, 4500, 'Get on the call', '2022-08-10 00:00:00'),
]
```

Step - 2 : Identifying The Input Data And Expected Output

INPUT

INPUT				
MESSAGE_ID	SENDER_ID	RECEIVER_ID	CONTENT	SENT_DATE
901	3601	4500	You up?	2022-08-03 0:00:00
902	4500	3601	Only if you're buying	2022-08-03 0:00:00
743	3601	8752	Let's take this offline	2022-06-14 0:00:00
922	3601	4500	Get on the call	2022-08-10 0:00:00

OUTPUT

OUTPUT	
SENDER_ID	COUNT(*)
3601	2
4500	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()

schema = StructType([
        StructField("message_id", IntegerType(), True),
        StructField("sender_id", IntegerType(), True),
        StructField("receiver_id", IntegerType(), True),
        StructField("content", StringType(), True),
        StructField("sent_date", StringType(), True),
])

# Define the data
data = [
        (901, 3601, 4500, 'You up?', '2022-08-03 00:00:00'),
        (902, 4500, 3601, 'Only if you\'re buying', '2022-08-03 00:00:00'),
        (743, 3601, 8752, 'Let\'s take this offline', '2022-06-14 00:00:00'),
        (922, 3601, 4500, 'Get on the call', '2022-08-10 00:00:00'),
]
```

```
teams_df = spark.createDataFrame(data,schema)
teams_df.show()
```

```
+----------+---------+-----------+--------------------+-------------------+
|message_id|sender_id|receiver_id|             content|          sent_date|
+----------+---------+-----------+--------------------+-------------------+
|       901|     3601|       4500|            You up?|2022-08-03 00:00:00|
|       902|     4500|       3601|Only if you're bu...|2022-08-03 00:00:00|
|       743|     3601|       8752|Let's take this o...|2022-06-14 00:00:00|
|       922|     3601|       4500|    Get on the call|2022-08-10 00:00:00|
+----------+---------+-----------+--------------------+-------------------+
```

```
filter_df=teams_df.filter(teams_df['sent_date'].like("2022-08%"))
filter_df.show()
```

```
+----------+---------+-----------+--------------------+-------------------+
|message_id|sender_id|receiver_id|             content|          sent_date|
+----------+---------+-----------+--------------------+-------------------+
|       901|     3601|       4500|            You up?|2022-08-03 00:00:00|
|       902|     4500|       3601|Only if you're bu...|2022-08-03 00:00:00|
|       922|     3601|       4500|    Get on the call|2022-08-10 00:00:00|
+----------+---------+-----------+--------------------+-------------------+
```

```
result_df=filter_df.groupby(filter_df['sender_id']).count()
result_df=result_df.orderBy(desc(result_df['count'])).limit(2)
result_df.show()
```

```
+---------+-----+
|sender_id|count|
+---------+-----+
|     3601|    2|
|     4500|    1|
+---------+-----+
```

PYSPARK LEARNING HUB : DAY - 15

Save

Was it helpful?

follow for more!

Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Comment

SHARE YOUR THOUGHTS IN COMMENT BELOW

Share