



**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'),**

**]**

|  |
| --- |



**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 |



**# 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'),**

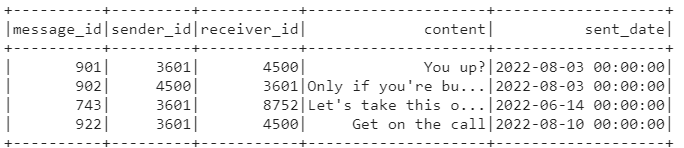
**(743, 3601, 8752, 'Let\'s take this offline', '2022-06-14 00:00:00'),**

**(922, 3601, 4500, 'Get on the call', '2022-08-10 00:00:00'),**

**]**

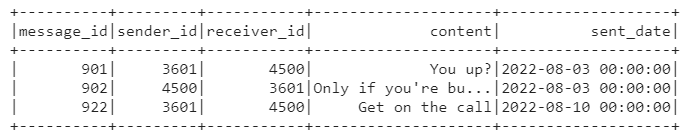
**teams\_df = spark.createDataFrame(data,schema)**

**teams\_df.show()**

****

**filter\_df=teams\_df.filter(teams\_df['sent\_date'].like("2022-08%"))**

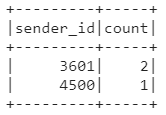
**filter\_df.show()**

****

**result\_df=filter\_df.groupby(filter\_df['sender\_id']).count()**

**result\_df=result\_df.orderBy(desc(result\_df['count'])).limit(2)**

**result\_df.show()**

****

