



**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)**

**]**



**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** |



**# 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),**

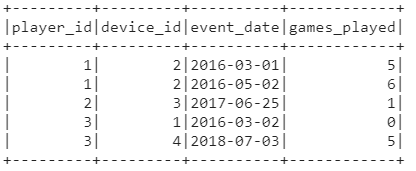
**(3, 4, '2018-07-03', 5)**

**]**

**# Create a PySpark DataFrame**

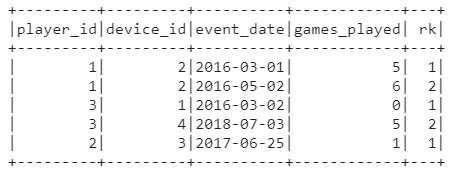
**df=spark.createDataFrame(activity\_data,activity\_schema)**

**df.show()**

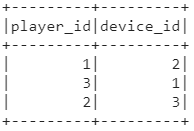
****

**rank\_df=df.withColumn("rk",rank().over(Window.partitionBy(df["player\_id"]).orderBy(df["event\_date"])))**

**rank\_df.show()**

****

**rank\_df.filter(rank\_df["rk"] == 1).select("player\_id","device\_id").show()**

****

