



**11\_Find Customer Referee**

Find the names of the customer that are not referred by the customer with id = 2.

Return the result table in any order

**Difficult Level :** EASY

**DataFrame:**

**# Define the schema for the Customer table**

**schema = StructType([**

**StructField("id", IntegerType(), True),**

**StructField("name", StringType(), True),**

**StructField("referee\_id", IntegerType(), True)**

**])**

**# Create an RDD with the data**

**data = [**

**(1, 'Will', None),**

**(2, 'Jane', None),**

**(3, 'Alex', 2),**

**(4, 'Bill', None),**

**(5, 'Zack', 1),**

**(6, 'Mark', 2)**

**]**



**INPUT**

| **INPUT** | | |
| --- | --- | --- |
| **ID** | **NAME** | **REFEREE\_ID** |
| **1** | **Will** |  |
| **2** | **Jane** |  |
| **3** | **Alex** | **2** |
| **4** | **Bill** |  |
| **5** | **Zack** | **1** |
| **6** | **Mark** | **2** |

**OUTPUT**

| **OUTPUT** |
| --- |
| **NAME** |
| **Will** |
| **Jane** |
| **Bill** |
| **Zack** |



**# 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 Customer table**

**schema = StructType([**

**StructField("id", IntegerType(), True),**

**StructField("name", StringType(), True),**

**StructField("referee\_id", IntegerType(), True)**

**])**

**# Create an RDD with the data**

**data = [**

**(1, 'Will', None),**

**(2, 'Jane', None),**

**(3, 'Alex', 2),**

**(4, 'Bill', None),**

**(5, 'Zack', 1),**

**(6, 'Mark', 2)**

**]**

**# Create a PySpark DataFrame**

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

| **# Filter customers not referred by customer with id = 2**  **result\_df = customer\_df.filter((col("referee\_id").isNull()) | (col("referee\_id") != 2))**    **# Select only the 'name' column**  **result\_df = result\_df.select("name")** |
| --- |

**+-----+**

**| name|**

**+-----+**

**| Will|**

**| Jane|**

**| Bill|**

**| Zack|**

**+-----+**

