


PySpark
Learning Hub | Practice Problem



Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Step - 1 : Problem Statement

06_Customers Who Never Order

Write a Pyspark program to find all customers who never order anything.

Difficult Level : EASY

DataFrame:

```
# Define the schema for the "Customers"
customers_schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("name", StringType(), True)
])

# Define data for the "Customers"
customers_data = [
    (1, 'Joe'),
    (2, 'Henry'),
    (3, 'Sam'),
    (4, 'Max')
]

# Define the schema for the "Orders"
orders_schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("customerId", IntegerType(), True)
])

# Define data for the "Orders"
orders_data = [
    (1, 3),
    (2, 1)
]
```

PYSPARK LEARNING HUB : DAY - 6

Step - 2 : Identifying The Input Data And Expected Output

INPUT

INPUT -1 customers	
ID	NAME
1	Joe
2	Henry
3	Sam
4	Max

INPUT - 2 orders	
ID	CUSTOMERID
1	3
2	1

OUTPUT

OUTPUT
NAME
Max
Henry

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()
```

```
customers_schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("name", StringType(), True)
])
```

Define data for the "Customers"

```
customers_data = [
    (1, 'Joe'),
    (2, 'Henry'),
    (3, 'Sam'),
    (4, 'Max')
]
```

Define the schema for the "Orders"

```
orders_schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("customerId", IntegerType(), True)
])
```

PYSPARK LEARNING HUB : DAY - 6

Define data for the "Orders"

```
orders_data = [  
    (1, 3),  
    (2, 1)  
]
```

Create a PySpark DataFrame

```
cus_df=spark.createDataFrame(customers_data,customers_schema)  
ord_df=spark.createDataFrame(orders_data,orders_schema)
```

```
cus_df.show()  
ord_df.show()
```

```
+---+-----+  
| id| name|  
+---+-----+  
|  1|  Joe|  
|  2|Henry|  
|  3|  Sam|  
|  4|  Max|  
+---+-----+
```

```
+---+-----+  
| id|customerId|  
+---+-----+  
|  1|          3|  
|  2|          1|  
+---+-----+
```

PYSPARK LEARNING HUB : DAY - 6

```
cus_df.join(ord_df,cus_df.id == ord_df.customerId,"left_anti")\
.select("name").show()
```

```
+-----+
|  name  |
+-----+
|   Max  |
| Henry  |
+-----+
```



Save

**Was it
helpful?**
follow for more!



Akash Mahindrakar

Data Engineer

akashsjce8050@gmail.com



Comment

**SHARE YOUR THOUGHTS
IN COMMENT BELOW**



Share