


PySpark
Learning Hub | Practice Problem



Akash Mahindrakar
Data Engineer
akashsjce8050@gmail.com

Step - 1 : Problem Statement

13_Page With No Likes

Write a pyspark code to return the IDs of the Facebook pages that have zero likes. The output should be sorted in ascending order based on the page IDs.

Difficult Level : EASY

DataFrame:

```
# Define the schema for the pages
pages_schema = StructType([
    StructField("page_id", IntegerType(), True),
    StructField("page_name", StringType(), True)
])
# Define the schema for the page_likes table
page_likes_schema = StructType([
    StructField("user_id", IntegerType(), True),
    StructField("page_id", IntegerType(), True),
    StructField("liked_date", StringType(), True)
])
# Create an RDD with the data for pages
pages_data = [
    (20001, 'SQL Solutions'),
    (20045, 'Brain Exercises'),
    (20701, 'Tips for Data Analysts')
]
# Create an RDD with the data for page_likes table
page_likes_data = [
    (111, 20001, '2022-04-08 00:00:00'),
    (121, 20045, '2022-03-12 00:00:00'),
    (156, 20001, '2022-07-25 00:00:00')
]
```

Step - 2 : Identifying The Input Data And Expected

INPUT

INPUT - 1 PAGES	
PAGE_ID	PAGE_NAME
20001	SQL Solutions
20045	Brain Exercises
20701	Tips for Data Analysts

INPUT - 2 PAGES_LIEKS		
USER_ID	PAGE_ID	LIKED_DATE
111	20001	2022-04-08 0:00:00
121	20045	2022-03-12 0:00:00
156	20001	2022-07-25 0:00:00

OUTPUT

OUTPUT
PAGE_ID
20701

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()
```

Define the schema for the pages

```
pages_schema = StructType([
    StructField("page_id", IntegerType(), True),
    StructField("page_name", StringType(), True)
])
```

Define the schema for the page_likes table

```
page_likes_schema = StructType([
    StructField("user_id", IntegerType(), True),
    StructField("page_id", IntegerType(), True),
    StructField("liked_date", StringType(), True)
])
```

Create an RDD with the data for pages

```
pages_data = [
    (20001, 'SQL Solutions'),
    (20045, 'Brain Exercises'),
    (20701, 'Tips for Data Analysts')
]
```

Create an RDD with the data for page_likes table

```
page_likes_data = [
    (111, 20001, '2022-04-08 00:00:00'),
    (121, 20045, '2022-03-12 00:00:00'),
]
```

PYSPARK LEARNING HUB : DAY - 13

```
(156, 20001, '2022-07-25 00:00:00')  
]
```

```
page_df=spark.createDataFrame(pages_data,pages_schema)  
page_like_df=spark.createDataFrame(page_likes_data,page_likes_  
schema)  
page_df.show()  
page_like_df.show()
```

```
+-----+-----+  
|page_id|      page_name|  
+-----+-----+  
|  20001|      SQL Solutions|  
|  20045|      Brain Exercises|  
|  20701|Tips for Data Ana...|  
+-----+-----+
```

```
+-----+-----+-----+  
|user_id|page_id|      liked_date|  
+-----+-----+-----+  
|    111|  20001|2022-04-08 00:00:00|  
|    121|  20045|2022-03-12 00:00:00|  
|    156|  20001|2022-07-25 00:00:00|  
+-----+-----+-----+
```

```
# Perform a left anti join to get pages with zero likes  
zero_likes_pages = page_df.join(page_like_df, 'page_id',  
'left_anti')  
# Select and sort the result  
result = zero_likes_pages.select("page_id").orderBy("page_id")  
# Show the result  
result.show()
```

```
+-----+  
|page_id|  
+-----+  
| 20701|
```



Save

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



Akash Mahindrakar

Data Engineer

akashsjce8050@gmail.com



Comment

**SHARE YOUR THOUGHTS
IN COMMENT BELOW**



Share

[WWW.LINKEDIN.COM/IN/AKASHMAHINDRAKAR](https://www.linkedin.com/in/akashmahindrakar)