



**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')**

**]**

|  |
| --- |

|  |
| --- |



**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** |



**# 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'),**

**(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()**

| **# 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|**

**+-------+**

