



**05\_Duplicate Emails**

Write a Pyspark program to report all the duplicate emails. Note that it's guaranteed that the email field is not NULL.

**Difficult Level :** EASY

**DataFrame:**

# Define the schema for the "employees"

employees\_schema **=** StructType**([**

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

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

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

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

**])**

# Define data for the "employees"

employees\_data **= [**

**(1,** 'Joe'**, 70000, 3),**

**(2,** 'Henry'**, 80000, 4),**

**(3,** 'Sam'**, 60000, None),**

**(4,** 'Max'**, 90000, None)**

**]**



**INPUT**

| **INPUT** | |
| --- | --- |
| **ID** | **EMAIL** |
| **1** | **a@b.com** |
| **2** | **c@d.com** |
| **3** | **a@b.com** |

**OUTPUT**

| **OUTPUT** |
| --- |
| **EMAIL** |
| **a@b.com** |



**# Creating Spark Session**

**from pyspark.sql import SparkSession**

**from pyspark.sql.types import StructType,StructField,IntegerType,StringType**

**from pyspark.sql.functions import when**

**from pyspark.sql import functions as F**

**from pyspark.sql.window import Window**

**#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 "emails" table**

**emails\_schema = StructType([**

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

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

**])**

**# Define data for the "emails" table**

**emails\_data = [**

**(1, 'a@b.com'),**

**(2, 'c@d.com'),**

**(3, 'a@b.com')**

**]**

**# Create a PySpark DataFrame**

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

**df.show()**

**df\_group=df.groupby("email").count()**

**df\_group.filter(df\_group["count"] > 1).show()**

****

