


**PySpark**  
Learning Hub | Practice Problem



**Akash Mahindrakar**  
Data Engineer  
akashsjce8050@gmail.com

## Step - 1 : Problem Statement

### 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)
]
```

## Step - 2 : Identifying The Input Data And Expected

### INPUT

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

### OUTPUT

OUTPUT
EMAIL
a@b.com

## 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
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()
```

```
+-----+-----+
| email|count|
+-----+-----+
|a@b.com|    2|
+-----+-----+
```



Save

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



**Akash Mahindrakar**

Data Engineer

akashsjce8050@gmail.com



Comment

**SHARE YOUR THOUGHTS  
IN COMMENT BELOW**



Share