Remove Duplicates {# 1. SQL / 2. Python / 3.Pandas /4. Pyspark}

SQL

In [1]:

```
#pip install ipython-sql 1)Load the Extension, 2)Check SQL Cell
%load_ext sql
%sql sqlite://
```

In [2]:

```
%%sql

-- Create a table
CREATE TABLE employees (
    employee_id INT PRIMARY KEY,
    first_name TEXT,
    last_name TEXT,
    department TEXT,
    salary INT
);

-- Insert sample data
INSERT INTO employees (employee_id, first_name, last_name, department,
VALUES
    (1, 'John', 'Doe', 'HR', 50000),
    (2, 'Jane', 'Williams', 'Finance', 60000),
    (3, 'Alice', 'Johnson', 'IT', 55000),
    (4, 'John', 'Brown', 'IT', 60000),
    (5, 'John', 'Brown', 'HR', 60000),
    (6, 'Eve', 'Williams', 'Finance', 62000);
```

```
 * sqlite://
Done.
Done.
```

Out[2]: []

In [4]:

```
%%sql
SELECT * FROM employees;
```

```
 * sqlite://
Done.
```

Out[4]:

employee_id	first_name	last_name	department	salary
1	John	Doe	HR	50000
2	Jane	Williams	Finance	60000
3	Alice	Johnson	IT	55000
4	John	Brown	IT	60000
5	John	Brown	HR	60000
6	Eve	Williams	Finance	62000

In [6]:

```
%%sql
SELECT first_name,count(*)
FROM employees
Group BY first_name
Having count(*) =1;
```

```
 * sqlite://
Done.
```

Out[6]:

first_name	count(*)
Alice	1
Eve	1
Jane	1

```
SELECT *
FROM your_table
GROUP BY column1, column2, ..., columnN
HAVING COUNT(*) > 1;
```

PYTHON

FOR_LOOP

In [5]:

```python
def remove_duplicates(my_list):
    Empty_List = []
    for i in my_list:
        if i not in Empty_List:
            Empty_List.append(i)
    return Empty_List

# Example usage:
my_list = [1, 2, 2, 3, 4, 4, 5]
result = remove_duplicates(my_list)
print(result)
```

```
[1, 2, 3, 4, 5]
```

ORDEREDDICT

In [6]:

```python
my_list = [1, 2, 2, 3, 4, 4, 5]
from collections import OrderedDict
my_list_no_duplicates = list(OrderedDict.fromkeys(my_list))
print(my_list_no_duplicates)
```

```
[1, 2, 3, 4, 5]
```

SET

In [4]:

```python
my_list = [1, 2, 2, 3, 4, 4, 5]
my_list_no_duplicates = list(set(my_list))
print(my_list_no_duplicates)
```

```
[1, 2, 3, 4, 5]
```

PANDAS

In [8]:

```
import pandas as pd

# Create a dictionary of data
data = {'Name': ['Alice','Alice', 'Bob', 'Charlie', 'David','David'],
        'Age': [25,20, 30, 30, 40, 40],
        'City': ['New York','New York', 'San Francisco', 'Los Angeles'

# Create a DataFrame from the dictionary
df = pd.DataFrame(data)

# Display the DataFrame
print(df)
```

```
      Name  Age           City
0    Alice   25       New York
1    Alice   20       New York
2      Bob   30  San Francisco
3  Charlie   30    Los Angeles
4    David   40        Chicago
5    David   40        Chicago
```

In [13]:

```
data_no_duplicates=df.drop_duplicates(data)
print(data_no_duplicates)
```

```
      Name  Age           City
0    Alice   25       New York
1    Alice   20       New York
2      Bob   30  San Francisco
3  Charlie   30    Los Angeles
4    David   40        Chicago
```

PYSPARK

In [17]:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("RemoveDuplicates").getOrCreate()
data = [("Alice", 25), ("Bob", 30), ("Alice", 25), ("Charlie", 35), ("
columns = ["Name", "Age"]
df = spark.createDataFrame(data, columns)
df.show()
```

```
+-------+---+
|   Name|Age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|  Alice| 25|
|Charlie| 35|
|  David| 40|
|    Bob| 30|
|    Bob|  3|
+-------+---+
```

In [19]:

```
df_no_duplicates = df.dropDuplicates()
df_no_duplicates.show()
```

```
+-------+---+
|   Name|Age|
+-------+---+
|  Alice| 25|
|    Bob| 30|
|Charlie| 35|
|  David| 40|
|    Bob|  3|
+-------+---+
```

In [ ]: