Master PySpark: From Zero to Big Data Hero!!

Joins Part 3

- **Coding Question:**
- Write a PySpark query to find employees whose location matches the location of their department. Display emp_id, emp_name, emp_location, dept_name, and dept_location for matching records.
- Modify the code to find departments that have no employees assigned to them. Display dept_id, dept_name, and dept_head.
- Write a PySpark query to get the average salary of employees in each department, displaying dept_name and the calculated average_salary.
- List the employees who earn more than the average salary of their department. Display emp_id, emp_name, emp_salary, dept_name, and dept_location.

Example → for joins with emp and dept data

```
from pyspark.sql import SparkSession
from pyspark.sql import Row

# Sample DataFrames
emp_data = [
    Row(emp_id=1, emp_name="Alice", emp_salary=50000,
emp_dept_id=101, emp_location="New York"),
    Row(emp_id=2, emp_name="Bob", emp_salary=60000,
emp_dept_id=102, emp_location="Los Angeles"),
    Row(emp_id=3, emp_name="Charlie", emp_salary=55000,
emp_dept_id=101, emp_location="Chicago"),
    Row(emp_id=4, emp_name="David", emp_salary=70000,
emp_dept_id=103, emp_location="San Francisco"),
    Row(emp_id=5, emp_name="Eve", emp_salary=48000,
emp_dept_id=102, emp_location="Houston")
]

dept_data = [
    Row(dept_id=101, dept_name="Engineering", dept_head="John",
dept_location="New York"),
    Row(dept_id=102, dept_name="Marketing", dept_head="Mary",
dept_location="Los Angeles"),
```

```
    Row(dept_id=103, dept_name="Finance", dept_head="Frank",
dept_location="Chicago")
]

emp_columns = ["emp_id", "emp_name", "emp_salary", "emp_dept_id",
"emp_location"]
dept_columns = ["dept_id", "dept_name", "dept_head",
"dept_location"]

emp_df = spark.createDataFrame(emp_data, emp_columns)
dept_df = spark.createDataFrame(dept_data, dept_columns)

# Display emp data
print("emp_data:")
emp_df.show()

# Display dept data
print("dept_data:")
dept_df.show()
```

```
emp_data:
+------+--------+----------+-----------+-------------+
|emp_id|emp_name|emp_salary|emp_dept_id| emp_location|
+------+--------+----------+-----------+-------------+
|     1|   Alice|     50000|        101|     New York|
|     2|     Bob|     60000|        102|  Los Angeles|
|     3| Charlie|     55000|        101|      Chicago|
|     4|   David|     70000|        103|San Francisco|
|     5|     Eve|     48000|        102|      Houston|
+------+--------+----------+-----------+-------------+

dept_data:
+-------+-----------+---------+-------------+
|dept_id|  dept_name|dept_head|dept_location|
+-------+-----------+---------+-------------+
|    101|Engineering|     John|     New York|
|    102|  Marketing|     Mary|  Los Angeles|
|    103|    Finance|    Frank|      Chicago|
+-------+-----------+---------+-------------+
```

```
# Inner Join on emp_dept_id and dept_id
inner_join = emp_df.join(dept_df, emp_df["emp_dept_id"] ==
dept_df["dept_id"], "inner")
```

```python
# Display the result
print("Inner Join Result:")
inner_join.show()

# Inner Join with Filtering Columns and WHERE Condition
inner_join = emp_df.join(dept_df, emp_df["emp_dept_id"] ==
dept_df["dept_id"], "inner")\
    .select("emp_id", "emp_name", "emp_salary", "dept_name",
"dept_location")\
    .filter("emp_salary > 55000")  # Add a WHERE condition

# Display the result
print("Inner Join with Filter and WHERE Condition:")
inner_join.show()
```

```
Inner Join Result:
+------+--------+----------+-----------+-------------+-------+-----------+---------+-------------+
|emp_id|emp_name|emp_salary|emp_dept_id| emp_location|dept_id|  dept_name|dept_head|dept_location|
+------+--------+----------+-----------+-------------+-------+-----------+---------+-------------+
|     1|   Alice|     50000|        101|     New York|    101|Engineering|     John|     New York|
|     3| Charlie|     55000|        101|      Chicago|    101|Engineering|     John|     New York|
|     2|     Bob|     60000|        102|  Los Angeles|    102|  Marketing|     Mary|  Los Angeles|
|     5|     Eve|     48000|        102|      Houston|    102|  Marketing|     Mary|  Los Angeles|
|     4|   David|     70000|        103|San Francisco|    103|    Finance|    Frank|      Chicago|
+------+--------+----------+-----------+-------------+-------+-----------+---------+-------------+

Inner Join with Filter and WHERE Condition:
+------+--------+----------+---------+-------------+
|emp_id|emp_name|emp_salary|dept_name|dept_location|
+------+--------+----------+---------+-------------+
|     2|     Bob|     60000|Marketing|  Los Angeles|
|     4|   David|     70000|  Finance|      Chicago|
+------+--------+----------+---------+-------------+
```

```python
# Left Join with Filtering Columns and WHERE Condition
left_join_filtered = emp_df.join(dept_df, emp_df["emp_dept_id"] ==
dept_df["dept_id"], "left")\
    .select("emp_id", "emp_name", "dept_name", "dept_location")\
    .filter("emp_salary > 55000")  # Add a WHERE condition

# Display the result
print("Left Join with Filter and WHERE Condition:")
left_join_filtered.show()
```

```python
# Left Anti Join
left_anti_join = emp_df.join(dept_df, emp_df["emp_dept_id"] ==
dept_df["dept_id"], "left_anti")

# Display the result
print("Left Anti Join Result:")
left_anti_join.show()
```

```
Left Join with Filter and WHERE Condition:
+------+--------+---------+-------------+
|emp_id|emp_name|dept_name|dept_location|
+------+--------+---------+-------------+
|     2|     Bob|Marketing|  Los Angeles|
|     4|   David|  Finance|      Chicago|
+------+--------+---------+-------------+

Left Anti Join Result:
+------+--------+----------+-----------+------------+
|emp_id|emp_name|emp_salary|emp_dept_id|emp_location|
+------+--------+----------+-----------+------------+
+------+--------+----------+-----------+------------+
```

```
Left Join Result without filter:
+------+--------+----------+-----------+-------------+-------+-----------+---------+-------------+
|emp_id|emp_name|emp_salary|emp_dept_id| emp_location|dept_id|  dept_name|dept_head|dept_location|
+------+--------+----------+-----------+-------------+-------+-----------+---------+-------------+
|     1|   Alice|     50000|        101|     New York|    101|Engineering|     John|     New York|
|     2|     Bob|     60000|        102|  Los Angeles|    102|  Marketing|     Mary|  Los Angeles|
|     3| Charlie|     55000|        101|      Chicago|    101|Engineering|     John|     New York|
|     4|   David|     70000|        103|San Francisco|    103|    Finance|    Frank|      Chicago|
|     5|     Eve|     48000|        102|      Houston|    102|  Marketing|     Mary|  Los Angeles|
+------+--------+----------+-----------+-------------+-------+-----------+---------+-------------+

Left Anti Join Result without filter:
+------+--------+----------+-----------+------------+
|emp_id|emp_name|emp_salary|emp_dept_id|emp_location|
+------+--------+----------+-----------+------------+
+------+--------+----------+-----------+------------+
```