LAB 8 (Python)

Import notebook

. Anuvind M P

. AM.EN.U4AIE22010

```
from pyspark.sql import SparkSession
spark = SparkSession.builder \
.appName("Employee Salary Analysis") \
.getOrCreate()
```

Create an RDD with employee data.

```
data = [
("Aadan", 25, "Engineering", 95000),
("Christo", 30, "Marketing", 85000),
("Denny", 35, "Engineering", 120000),
("Booboo", 28, "HR", 70000),
("Glub Glub", 39, "Marketing", 10000),
("Chu Chu Chu", 45, "Engineering", 105000),
("Cha Cha Cha", 31, "HR", 80000)
]
columns = ["name", "age", "department", "salary"]
df = spark.createDataFrame(data, schema=columns)
df.show()
```

df: pyspark.sql.connect.dataframe.DataFrame = [name: string, age: long ... 2 more fields]

```
+-----------+---+-----------+------+
|       name|age| department|salary|
+-----------+---+-----------+------+
|      Aadan| 25|Engineering| 95000|
|    Christo| 30|  Marketing| 85000|
|      Denny| 35|Engineering|120000|
|     Booboo| 28|         HR| 70000|
|  Glub Glub| 39|  Marketing| 10000|
|Chu Chu Chu| 45|Engineering|105000|
|Cha Cha Cha| 31|         HR| 80000|
+-----------+---+-----------+------+
```

Find the average salary for employees in the company

```
from pyspark.sql.functions import avg
avg_salary = df.agg(avg("salary")).first()[0]
print(f"Average Salary: {avg_salary}")
```

```
Average Salary: 80714.28571428571
```

Identify employees with salaries higher than the average salary.

;

```
    df.filter(df.salary > avg_salary).show()
```

```
+-----------+---+-----------+------+
|       name|age| department|salary|
+-----------+---+-----------+------+
|      Aadan| 25|Engineering| 95000|
|    Christo| 30|  Marketing| 85000|
|      Denny| 35|Engineering|120000|
|Chu Chu Chu| 45|Engineering|105000|
+-----------+---+-----------+------+
```

Sort the employees by salary in descending order and show the top 3 highest-paid employees.

```
    top = df.orderBy(df.salary.desc()).show(3)
```

```
+-----------+---+-----------+------+
|       name|age| department|salary|
+-----------+---+-----------+------+
|      Denny| 35|Engineering|120000|
|Chu Chu Chu| 45|Engineering|105000|
|      Aadan| 25|Engineering| 95000|
+-----------+---+-----------+------+
only showing top 3 rows
```

Group employees by department and compute the average salary per department.

```
+-----------+------------------+
| department|       avg(salary)|
+-----------+------------------+
|Engineering|106666.66666666667|
|  Marketing|           47500.0|
|         HR|           75000.0|
+-----------+------------------+
```