from pyspark.sql import SparkSession

from pyspark.sql.functions import col

# Initialize a Spark session

spark = SparkSession.builder \

.appName("Employee Data Analysis") \

.getOrCreate()

# Sample employee data

data = [

(1, 'Arjun', 'IT', 75000),

(2, 'Vijay', 'Finance', 85000),

(3, 'Shalini', 'IT', 90000),

(4, 'Sneha', 'HR', 50000),

(5, 'Rahul', 'Finance', 60000),

(6, 'Amit', 'IT', 55000)

]

# Define schema (columns)

columns = ['EmployeeID', 'EmployeeName', 'Department', 'Salary']

# Create DataFrame

employee\_df = spark.createDataFrame(data, columns)

# Show the DataFrame

employee\_df.show()

#1. \*\*Task 1: Filter Employees by Salary\*\*

filtered\_df = employee\_df.filter(col('Salary') > 60000)

print("Employees with Salary > 60000:")

filtered\_df.show()

#2. \*\*Task 2: Calculate the Average Salary by Department\*\*

from pyspark.sql.functions import avg

avg\_salary\_df = employee\_df.groupBy('Department').agg(avg('Salary').alias('AverageSalary'))

print("Average Salary by Department:")

avg\_salary\_df.show()

#3. \*\*Task 3: Sort Employees by Salary\*\*

sorted\_salary\_df = employee\_df.orderBy(col("Salary").desc())

print("Employees sorted by salary:")

sorted\_salary\_df.show()

#4. \*\*Task 4: Add a Bonus Column\*\*

from pyspark.sql.functions import expr

bonus\_df = employee\_df.withColumn("Bonus", col("Salary") \* 0.10)

print("Employee DataFrame with Bonus Column:")

bonus\_df.show()