Question: You are tasked with performing various data processing tasks using Apache Spark's Resilient Distributed Datasets (RDDs) on a dataset containing employee information. Use the provided dataset named employee\_data which has the following structure:

employee\_data Schema: employee\_id:int, employee\_name:string, department:string, salary:double, hire\_date:string

Write Spark applications using RDDs to accomplish the following tasks:

1. Load the employee\_data dataset into an RDD.

Solution: employee\_rdd = sc.textFile("employee\_data.csv")

1. Filter the employee\_data RDD to retain records where the salary is greater than 50000.

Solution: filtered\_rdd = employee\_rdd.filter(lambda x: float(x.split(",")[3]) > 50000)

1. Map the employee\_data RDD to extract only the employee\_name and department fields.

Solution: name\_department\_rdd = employee\_rdd.map(lambda x: (x.split(",")[1], x.split(",")[2]))

1. Apply a flatMap transformation to split the employee\_name by space and return individual words.

Solution: words\_rdd = name\_department\_rdd.flatMap(lambda x: x[0].split(" "))

1. Calculate the total salary for each department by reducing the employee\_data RDD by key.

Solution: department\_total\_salary\_rdd = employee\_rdd.map(lambda x: (x.split(",")[2], float(x.split(",")[3]))).reduceByKey(lambda a, b: a + b)

1. Sort the employee\_data RDD by the salary field in descending order.

Solution: sorted\_rdd = employee\_rdd.map(lambda x: (x, float(x.split(",")[3]))).sortBy (lambda x: x[1], ascending=False)

1. Group the employee\_data RDD by the hire\_date field.

Solution: grouped\_by\_date\_rdd = employee\_rdd.groupBy(lambda x: x.split(",")[4])

1. Join the employee\_data RDD with another RDD named department\_data on the department field.

Solution: department\_rdd = sc.textFile("department\_data.csv")joined\_rdd = employee\_rdd.map(lambda x: (x.split(",")[2], x)).join(department\_rdd.map(lambda x: (x.split(",")[0], x)))

1. Collect the first 5 records of the employee\_data RDD into the driver program.

Solution: first\_5\_records = employee\_rdd.take(5)