Q1. Load an employee detail csv file in Spark RDD, convert the details in DataFrame with proper schema,

1. **EMPLOYEE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Employee ID | First Name | Last Name | Department ID | Salary |
|  |  |  |  |  |
|  |  |  |  |  |

1. **DEPARTMENT**

|  |  |
| --- | --- |
| Department ID | Department Name |
|  |  |
|  |  |

Q2. Add a new column as “Fullname”, populate the fullname as concatenated first name and last name using UDF

Q3. Write a program to get the rank on Salary of the Employees using dataframe API.

Q4. Write a program to get the name of the Employee with maximum salary in each department, using dataframe API

Q5. Write a program which takes 2 strings as input and returns 1 if they are anagrams else returns 0. (**Anagram**: A word, phrase, or name formed by rearranging the letters, such as spar, formed from rasp.)

Q6. A text file contains data like this:

**North Carolina State University, Raleigh**

**Texas A&M University, Houston**

**University of California, San Diego**

**University of Virginia, Charlottesville**

The problem is to count the number of words before comma. The below given code has no errors, but doesn’t give the desired output.

Correct the code to get the right output.

**val file = sc.textFile(“Universities.txt”)**

**file.foreach { x =>**

**var list = x.split(“,”).toList**

**val checkRDD = sc.parallelize(list)**

**print(checkRDD.map { x => x.split(“ “)}.count)**

Q7. How to change one row into multiple row while doing operation on one column

1. Eg:- original row
2. ID Name Class Start\_Date End\_Date

ii. 1 John Xii 20170909 20210909

1. Final rows looks like, After applying logics
2. ID Name Class Start\_Date End\_Date

ii. 1 J Xii 2017-09-09 2021-09-09

iii. 1 o Xii 2017-09-09 2021-09-09

iv. 1 h Xii 2017-09-09 2021-09-09

v. 1 n Xii 2017-09-09 2021-09-09