German International University of Applied Sciences

Informatics and Computer Science

Distributed Web-based System, Spring 2023

Assignment 1

Due date is April 1, 2024 at 11:59 PM

In this assignment, you will examine the Map Reduce Methodology. Given ***"DataScienceJobSalaries"*** dataset, You will be using the "Data Science Job Salaries"dataset provided in the file ′dssalaries.txt′ Each line inside the input file contains the following information (Year, Job Title, Salary, Company Location), each separated with ", "

You are required to do the following Tasks:

a) Task 1

Find the average salary per Job, then observe the highest paying job from the output file:

1. The Mapper goal is to make a key value pair of the Job and its, Salary Value So the Key is the Job Title and its Value is the Salary <Job, Salary>

2. The Reducer Input as follows < Job, < S1, S2, ...> >, The Key is the Job Title and the Value is a set of all taken salaries within the specified job title, Reducer should compute the average of those salaries and output a key value pair as follows <Job, average salary>

b) Task 2

Find the average salary per Country, then observe the highest paying country from the output File:

1. The Mapper goal is to make a key value pair of the country and its salary, so the Key is the country and its value are the salary <Country, Salary>

2. The Reducer Input as follows < Country, < S1, S2, ...> >, The Key is the country and the value is a set of all taken salaries within the specified country, Reducer should compute the average of those salaries and output a key value pair as follows <country, average salary>

Deliverables:

a) Your code needs to be submitted on The Google form Link **https://forms.office.com/r/HngVn6Hj4u**

(No mails are allowed)

b) You have to submit the notebook that contains both tasks and the obtained output files.

c) You have to submit also a pdf file that contains screenshots of the code, screenshots of the output files, and a brief description of how you solved the tasks.

d) Make sure to comment on every step while coding.

e) Make sure to use only **MRJob** for both tasks, no need to use spark or any other distributed system packages package.

f) Feel free to use any other python packages for example (pandas, numpy, …. etc).

PLAGIARISM IS NOT TOLERATED, AND COPIED WORK WILL BE AWARDED 0 POINTS FOR BOTH PERSONS INVOLVED!