Middle East Technical University

Department of Computer Engineering

CENG 495

Cloud Computing

Spring 2024

HW - 3: Hadoop - MapReduce

Due date: May 30, 2024, Thursday, 23:59

1 Introduction

In this homework, you will use Apache Hadoop's MapReduce to get insights from the Machine Learning Engineer Salary in 2024 dataset. You will use the Java language for this homework.

2 Setup

Setup Hadoop in Pseudo-Distributed operation mode, using the single node cluster approach. You can follow the Hadoop tutorial here and the MapReduce tutorial here.

3 Task

Download the dataset and extract the `.csv` file. This is the only input file you will need for this assignment. You might want to preprocess the dataset to make your job easier for the later tasks (e.g. convert the `csv` to a `tsv`). If your Java program expects a preprocessed dataset, make sure to include a script that takes the original dataset and converts it to the one your program expects.

I recommend using visidata to inspect the dataset.

3.1 Tasks

Report on the following;

- Total amount of machine learning engineer's salaries (Salary in USD) (**total**)
- List the average salary (Salary in USD) for each job title (**jobtitle**)
- List the average salary (Salary in USD) for each job title for SE, MI, and EX experience levels (**titleexperince**)
- Separate the employee residences according to whether they are *US* or *non-US*, then list the average salary (Salary in USD) for the salaries in those two separate categories. Output US employee residence salaries in `part-r-00000` and non US employee residence salaries in `part-r-00001` (**employeeresidence**)

- Partition the salaries by work year; first partition is the salaries in the 2024 (work year = 2024), the second partition is for salaries in the 2023 (work year = 2023), and final partition is for salaries before 2023 (work year < 2023). Report the average salary (Salary in USD) for these 3 partitions (`part-r-00000` to `part-r-00002`) (**averageyear**).

4 Submission

- Use Java programming language using the Apache Hadoop library with MapReduce programming model.
- **Your submission must include your Java source code:** Java source code should be in a separate folder in the submission because I will compile it as mentioned below.
- **Your submission must include input and output files:** Since input file may be updated within the homework duration, put your final version of input file in the submission. Also, put your final version of output files that your code created.
- **Your submission must include a README file:** The README file should include the following information:
  - Operating system used for developing the homework
  - Hadoop release used for the homework
  - JDK (Java Development Kit) version used for the homework
  - IDE (Integrated Development Environment) used for the homework
  - If you want to deviate from the commands given below within a reason, explain how to build & run your project, and I will use that during grading.
- Archive your submission as a `.zip` file and name it as "firstname_lastname.zip" and submit to odtuclass.
- This is an individual assignment. You can discuss your ideas with your peers but using implementation specific code that is not your own is strictly forbidden and constitutes as cheating. This includes but not limited to friends, any previous homework, CENG homework repositories on GitHub, or the Internet in general. The violators will get no grade from this assignment and will be punished according to the department regulations.

Your code will be evaluated using the Pseudo-Distributed local mode of Hadoop. Your submission will be extracted, and the following commands will be executed on the top level of your submission:

```
# compilation
hadoop com.sun.tools.javac.Main *.java
jar cf Hw3.jar *.class

# running
hadoop jar Hw3.jar Hw3 total <input.csv> output_total
hadoop jar Hw3.jar Hw3 jobtitle <input.csv> output_jobtitle
hadoop jar Hw3.jar Hw3 titleexperince <input.csv> output_titleexperince
hadoop jar Hw3.jar Hw3 employeeresidence <input.csv> output_employeeresidence
hadoop jar Hw3.jar Hw3 averageyear <input.csv> output_averageyear
```