DATABASE DESIGN II – 1DL400 – 2022

Assignment 4 (Lab 4)
MapReduce Paradigm

Examination

This assignment aims to familiarize you with the **MapReduce** paradigm by letting you implement simple tasks. In order to complete this assignment, you may need to do research beyond the material given in the class or provided on Studium. The assignment must be submitted before the indicated deadline on **Studium**.

In this assignment, you need to complete a template Python script provided in "mapreduce.py". There, you need to create your own implementations of the classes "mapper" and "reducer" and your own functions for each of the questions. The template file will be mostly complete. The areas where you need to write your solution will be marked with the comment "#fill_in".

To use the template script you will need to have installed some languages and packages depending on your operating system. Follow the section **"Installation Instructions"** at the end of this document.

Submission Instructions

The assignment must be submitted before the indicated deadline on **Studium**. Please notice that you are not expected to finish this assignment during the lab: it may take longer. You must submit the result of this assignment on **Studium** as one single PDF file and one single script file per group. Please indicate group name, group participants with ID (personnummer, if you have one).

Your submission should include:

1. The modified file "mapreduce.py" that includes the code to run the answers to all the questions.
2. A pdf file that includes
    a. Instructions on how to run you programs (e.g. if user input or arguments are required)
    b. The output of your results (or sample of the result if it is too large for a particular question)

Warmup (You do not need to submit this):

From the assignment page on Studium, download the file "lorem.txt" on the same directory as "mapreduce.py". Using the MapReduce paradigm, write a program that implements Word-Count and execute it using the downloaded file as input. The template file already has implementations of the Mapper and Reducer classes that you can use to perform a Word-Count. Make sure it executes correctly and you understand how it works.

Question 1:

By using the MapReduce paradigm write a program that implements the K-mer Counter. Use the file "sequence.txt" as input.

***K*-mer Counter** is a utility designed for counting *k*-mers (sequences of consecutive *k* symbols) in a set of reads from genome sequencing projects. *K*-mer counting is important for many bioinformatics applications. **Apply the k-mer counter utility for k=3 and the genome: ACACACAGT**. For instance, for k=3 and the genome: ACACACAGT, we get the following 3-mers: ACA, CAC, ACA, CAC, ACA, CAG, AGT. K-mer counting counts the frequency of 3-mers. E.g. ACA appears 3 times.

For the following questions (Questions 2,3) download the files "City.dat" on the same directory as "mapreduce.py" from the assignment page on Studium . Using the downloaded file as input write programs that implement the following SQL queries.

Question 2:

```
SELECT name, country_code, province, population
FROM City
WHERE population < 100000
```

Question 3:

```
SELECT country_code, province, SUM(population)
FROM City
GROUP BY country_code, province
ORDER BY country_code
```

Installation Instructions

Irrespective of the OS that you are using you will need to have Java (>=8) and Python (3) installed. If you haven't installed these you can look for instructions on the web.

Depending on previous installations or configurations on your OS you might need additional fixes to properly install pyspark using the following instructions. If this takes up more time than you can allocate, we also provide an Ubuntu virtual machine with pyspark preinstalled and ready to run "mapreduce.py". You can download it from the following link. Bear in mind the download is ~5GB.
https://uppsala.box.com/s/frndih8j731ucdvhgd2s30hb5u7rbf10
Defaults (you can change these yourself)

- Password: 123456
- Keyboard Layout:English-US

Windows:

1. Make sure you have Java(>=8) and Python(3) installed.
2. Open a powershell window and navigate to an easily accessible directory (e.g. C:\Users\<user>\Documents\db2_lab4\)
3. Check that you can call "python". If not, search online for how to add python to your Windows "Path".
4. Check if "pip" is installed alongside your python installation by executing "pip help"
    a. If not installed, download this file and place it in the directory that your powershell is currently in.
    b. Execute "python get-pip.py" and the installation of "pip" will start. When finished:
5. Execute "pip install pyspark" to install the pyspark package.
6. Download winutils.exe and place it in the following directory: "C:\winutils\bin\"
7. Type "Edit the system and environment variables" in the windows search bar and click the first result. Click "Environment Variables" and then "New" system variable (lower box). Enter variable name "HADOOP_HOME", variable value "C:\winutils" and click ok.
8. From Studium download the template script: "mapreduce.py" and place it in the directory where your powershell is. Execute it and see that it terminates without errors.
    a. You may need to edit the and uncomment the paths under "#Windows" to reflect your actual installation paths for Java,Python. The example should give you an idea for where to look.
9. Edit the template file to complete your tasks.

Linux:

1. Make sure you have Java (>=8) and Python (3) installed.
2. Open a terminal in an accessible directory and execute “pip install pyspark”
3. From Studium download the template script: “mapreduce.py” and place it in the directory where your terminal is. Execute it and see that it terminates without errors.
    a. You may need to edit the and uncomment the paths under “#Linux” to reflect your actual installation paths for Java, Python. The example should give you an idea for where to look.
4. Edit the template file to complete your tasks.

Mac:

1. Make sure Python 3.7.x, Java (>=8) are installed. Check pip is correctly installed and linked using "pip3 --help".
2. Execute the following: pip3 install pyspark
3. Download mapreduce.py from Studium and navigate to the downloaded folder through the terminal. Execute it to see whether it outputs errors.
    a. Note: You may need to uncomment some of the paths under #Mac and edit them to reflect the paths your binaries are saved under. To find where a binary is stored on your system, run "which <binary>" on a terminal (e.g. for Python 3, run "which python3").
    b. Note 2: If the error "No module named pyspark" appears, try running "pip3 install findspark" in a terminal, and include the line "import findspark" at the start of mapreduce.py.
4. Edit your template file to complete the tasks.