Map-Reduce

January 31st, 2024

The assignment this week is on map-reduce, which is a distributed and scalable way of extracting/mining required information from multiple datasets stored on multiple servers. Follow the tutorial to understand how you can design mapper and reducer for specific queries/operations.

Tutorial on map-reduce

You can start with a simple word count problem. Say, we have a text file and we want to count the frequency of occurrence of each word. The tutorial below explains how to solve this problem using a map-reduce algorithm.

Tutorial References :

http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

Next you can also look into the following tutorial for slightly harder query (tf-idf scores)

https://www.tutorialspoint.com/map_reduce/map_reduce_tutorial.pdf

For video explanation, please refer to the tutorial.:

https://www.youtube.com/watch?v=30RaNpaupj0&list=PLtzRLOcrx9SS1Ir6_viv-yLJd0PX3GR5O&index=1

How to run and test your code:

After going through the tutorial, we create our own codes and test our codes on a Linux system as follows:

```
cat network.txt | python mapper.py | sort | python reducer.py > result.txt
```

Or just

```
python mapper.py | sort | python reducer.py > result.txt
```

Explanation on the above commands:

1. “cat” is a linux command to print the contents of a file on the console.
2. The pipe operator (|) directs the previous command’s output to the next command.
3. “Sort -n” is a linux command to sort the input numerically.
4. “>” can be used to save the standard output in a file.
5. “Network.txt” is the input file passed to the mapper function.

Makefile:

In the Assignment, each query will be written in a folder and that folder must have its own Makefile to execute the routines. You can find a relevant tutorial here (https://opensource.com/article/18/8/what-how-makefile).

Also in some queries we are to use cp command to copy files from one folder to another.

https://www.geeksforgeeks.org/cp-command-linux-examples/

Tutorial for File Handling:

https://www.w3schools.com/python/python_file_handling.asp

https://www.geeksforgeeks.org/file-handling-python/

Tutorial for Argument Handling :

https://www.geeksforgeeks.org/command-line-arguments-in-python/

Suppose you have to access the Parent folder and use/store a file say red.txt, write ‘../red.txt’ as and when required. ‘../’ takes you back to the parent directory.