

In who we trust

Cloud Computing

Spring 2019

Dr. Farshad Khunjush

Homework 1: Map-Reduce

Due date: April 30th, 23:55

The given Dataset that is the log of a server and has 10787 rows and 4 columns. Columns are IP, Time, URL, and Response Status.

Using MapReduce, You Should find:

- 1- N most visited URLs with the number of visits for each URL. (N is input) (40 points)
- 2- N most visited URLs in M most Popular hours in all times. (N and M are inputs) (60 points)
- 3- Most visited URLs with distinct visitors (Optional) (20 points)

Reminders:

- Each HW has to be done individually
- The programs should be developed in JAVA.
- You should submit a report and explain the problem-solving method and show your results. (50% of your points belong to the report for each part)
- Your codes and the report should be submitted before the due date through email to TA.
- In-person delivery will be announced.