Faculty of Computing
Sabaragamuwa University of Sri Lanka
SE6103 - Parallel and Distributed Systems
Assignment 03

Time Frame: (2 hours)

Instructions:

- The Assignment contains two questions.
- **Question 1** is a theory question and should be written on a paper and handed over within the given time.
- **Question 2** is a practical question. It should be done using your previous practical knowledge and,
  - prepare a Word document with screenshots of the steps you follow.
  - Create a GitHub repository and upload it.
  - Submit the repository link to the following link.
- **Submit here:** https://forms.gle/FpzmSws6yXuZQF2NA

Question 1

1. Explain the differences between **Docker Containers** and **Virtual machines**.
2. What does the **-d** flag in the docker run command do? Why is it important when running services like Nginx?
3. Explain the difference between the following commands: **docker run -d nginx:latest** and **docker run nginx:latest**. Which one would you use for long-running services?
4. When running the command **docker run -d -p 8080:80 nginx**, what does **-p 8080:80** accomplish? Why is port mapping necessary?

5. What is Hadoop and what is it used for?
6. What are the advantages of using Apache Spark over Hadoop? Explain.

Question 2

1. Use the knowledge you gained during the Hadoop lab sessions to analyze the following example of a text and find the number of occurrences of each word in the text.

   Hadoop is an open-source framework that stores and processes large amounts of data across a network of computers. Hadoop is a very interesting framework to learn about networking.