Department of Computer Systems Engineering
University of Engineering & Technology Peshawar
Khyber Pukhtunkhwa, Pakistan, *Phone No # +92-922-560576*

EXAM: FINAL-TERM
SUBJECT: DATA ANALYTICS

PART- B
TOTAL MARKS: 20

SEMESTER: Fall-2023
TIME: 95 MINS

Question 2: Briefly answer the following questions.

1. You have a large dataset of log files from a web server. The dataset is so large that it cannot be processed on a single machine. You want to use Hadoop to process the dataset and generate a report on the most popular pages on the website. [5]

2. You are working as a Hadoop admin in some company, and you lost the NameNode at some stage during the big data processing, will you need to reinitiate the whole process, or do you have an alternate way to achieve the high availability? Keep in mind the constraint that the company can't afford an expensive solution. [5]

3. To assess the significance of possible variation in performance in a certain test between the government schools of a city, a common test was given to several students taken at random from the fifth class of the 3 schools concerned. The results are given below [5]

A	B	C
9	13	14
11	12	13
13	10	17
9	15	7
8	5	9

Make the analysis of variance for the given data.

Question 3: Short questions. **[1x5]**

1. What is the purpose of a Dockerfile?

2. How does Docker facilitate microservices architecture?

3. Name a few components of metadata in NameNode.

4. What if we decrease the replication factor?

5. If we have 4GB of data? How many blocks will be created in Hadoop version 2.X with RF = 5?