## COMSATS University Islamabad, Lahore Campus

### Final Examination – Semester Fall 2021

| | | | | Course | CSC465 | Credit | 3 |
|---|---|---|---|---|---|---|---|
| Course Title: | Introduction to Data Science | | | Program Name: | BS(CS) | | |
| Course | Mr. Shahid Bhatti | | Section: | | Date | 15-November-2021 | |
| Semester: | 6th | Batch: | | Maximum | | 25 | |
| Time Allowed: | 1.5 Hours | | | Reg. No. | FA18 BCS | | |
| Student's Name: | | | | | | | |

Important Instructions / Guidelines:
- All parts are compulsory.

### Section I

[Marks 10]

**Question 1:**

A. In a range selection on a range-partitioned attribute, it is possible that only one disk may need to be accessed. Describe the benefits and drawbacks of this property. [2]

B. What is a parallel join? In what situation fragment-and-replicate join is preferred. [3]

C. Write the names of map reduces phases from data input to produce the final results. [3]

D. What are the responsibilities of the Name Node in the Hadoop HDFS? [2]

*[handwritten: Master Node: manage data access • data can be replicated • to send data to reducers after map is done — chunks]*

### Section-II

[Marks 15]

**Question 2:**

A. Consider a dataset with one variable x and output y. Recall your linear regression concepts and. write down the following equations for linear regression. [7]
- Hypothesis
- Cost Function
- Minimization using Gradient Descent

*[handwritten: $J(\theta, \theta) = \hat{y} - 2l$ ; $m, \text{theta}$ ; $W \approx 0$ ]*

B. Which of the following is the cause of overfitting in logistic regression.? Justify your answer? [4]

- High variance
- High Bias

*[handwritten: if $Y \geq 0.5$ predict 1 $\theta_j - \frac{1}{2m}$ ]*

C. Categorize the following as regression or classification problems. Give reason. [4]

I. Predicting Car speed from camera image installed on top of the camera.

II. To determine whether a transaction of Rs. 10,000,456 is a fraud or not.

III. Predicting company's profit at the end of the fiscal year based on performance in first two quarters.

IV. To determine whether a product review of 500 words is positive or negative based on the content of the review.

*[handwritten: $\frac{1}{1+e^{-z}}$ ; $\uparrow x$ $0$ ; $G(z) = \frac{1}{1+e^{-\theta x}}$ ]*