24BDAH201 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. BIG DATA ANALYTICS SECOND SEMESTER DEGREE EXAMINATION

MAY 2025

Big Data Architechure & Hadoop Ecosystem

Duration:3 Hours **Max Marks:70**

PART A

I. Answer any FOUR of the following **(4×5= 20)**

a) Explain the importance of Big data analytics with a suitable example.

b) With a suitable example demonstrate why Shuffle and Sort are intermediary steps in MapReduce execution flow.

c) Consider a CSV file containing movie reviews with information about the Reviewer, the Movie title and the Rating. Create a table in Hive to define the schema for this data. Demonstrate the operation of loading the data and display the average rating for the movie titled "Dune".

d) Describe the three fundamental configuration units in Apache Puppet using a suitable example.

e) Explain the authentication procedure followed by Sentry in Hadoop.

PART B

II. Answer any FIVE questions selecting at least one question from each unit:

(5×10= 50)

UNIT-I

2. 1. Explain five significant open source projects that are part of the Hadoop Ecosystem. (5M)
   2. Explain the steps required to deploy a Big Data Solution. (5M)

3. 1. Highlight the advantages of a Hadoop Cluster. (5M)
   2. Discuss how cluster computing is an effective approach for big data solutions. (5M)

UNIT-II

4. Compare and contrats the different characteristics of NoSQL and relational databases.

5. With an illustration explain the YARN model and its services.

UNIT-III

6. Explain the two approaches to indexing for data management in hadoop.

7. In marketing, big data comprises of gathering and analyzing massive amounts of digital information to improve business operations. Citing 4 examples, analyse how Big Data is transforming marketing and sales.

UNIT-IV

8. Consider a CSV file with details of bills paid towards Utility expenditures such as Gas, Electricity and Telephone. Using Pig Latin Script,
   1. Display the total expenditure for the year.
   2. Display the total expenditure per month
   3. Display the total expenditure for each of the three utilities during the year.

9. Describe the basic primitives that Apache Storm provides for performing stream transformations.

****

24BDAH202 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. BIG DATA ANALYTICS SECOND SEMESTER DEGREE EXAMINATION MAY 2025

Machine Learning and Deep Learning

Duration:3 Hours Max Marks:70

PART A

I. Answer any FOUR of the following (4×5= 20)

a) Describe the concept of supervised machine learning. Elaborate on any two examples of supervised machine learning tasks.

b) Explain metrics for multiclass Classification.

c) Expalin covariant shift in Distribution Shift.

d) Explain in detail the concept of custom block.

e) Distinguish between PCA and NMF as a dimensionality reduction models.

PART B

II. Answer any FIVE questions selecting at least one question from each unit: (5×10= 50)

UNIT-I

2. Apply and analyse k-Nearest Neighbors regression model on make_wave dataset available in mglearn.

3. Discuss on the concept of generalization and underfitting in supervised machine learning.

UNIT-II

4. Elaborate on Hierarchical clustering and dendrograms .

5. Explain RobustScaler and Normalizer with an illustration.

UNIT-III

6. Explain the implementation of Softmax Regression from Scratch using python code.

7. Elaborate on the below activation functions:
   1) ReLu function
   2) Sigmoid function

UNIT-IV

8. Explain in detail the concept of CNN.

9. Explain with an illustration, Cross-correlation computation with Multiple Input Channels.

****

24BDAH203

Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. BIG DATA ANALYTICS SECOND SEMESTER DEGREE EXAMINATION

MAY 2026

Multivariate Techniques for Data Analysis

Duration:3 Hours **Max Marks:70**

PART A

Answer any FOUR of the following **(4×5= 20)**

a) Enumerate any five relationship between Multivariate Dependence Methods.

b) Describe the application of factor analysis in the field of Buisness and Machine Learning.

c) Explain the working of hierarchical procedure with an example.

d) Differentiate between ANOVA and MANOVA

e) Explain why measurement scale is important in data analysis and also provide the impact of choice of Measurement Scale.

PART B

II. Answer any FIVE questions selecting at least one question from each unit: **(5×18= 50)**

UNIT-I

2. Explain 5 real-world cases where logistic regression was effectively used.

3. Describe ANOVA table computation in terms of fitting of multiple linear regression equation on one dependent variable and 'n' independent variable.

UNIT-II

4. Using PCA convert the following dataset and find the first four components

X	2.5	0.5	2.2	1.9
Y	2.4	0.7	2.9	2.2

5. The production manager of a company maintains that flow type in days(Y) depends on the number of operations (X) to be performed. The following data gives the necessary information.Plot a scatter diagram. Calculate the value of Karl Pearson's coefficient of correlation.

X	2	2	3	4	4	5	6	6	7	7
Y	8	13	14	11	20	10	22	26	22	25

UNIT-III

6. Describe first canonical variable, second canonical variable and kth canonical variable.

7. Explain the designing of conjoint analysis.

UNIT-IV

8. Elaborate the decision process of MANCOVA

9. Differentiate between Analysis of Variance and Analysis of Covariance.

****

24BOAS207 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. BIG DATA ANALYTICS SECOND SEMESTER DEGREE EXAMINATION

MAY 2025

Natural Language Processing

Duration:3 Hours Max Marks:70

PART A

I. Answer any FOUR of the following (4×5= 20)

a) Explain the Edit-distance mechanism for spell correction with a suitable example.

b) Develop a python program to extract some of the most insightful named entities from a given text file

c) Explain any five applications where Language Detection and Optical Character Recognition is used.

d) Develop python code to demonstrate briefly the application of Stochastic Gradient Descent algorithm to generate the classification report when performing text classification.

e) Develop a python program using Genism library to remove stop words.

PART B

II. Answer any FIVE questions selecting at least one question from each unit: (5×10= 50)

UNIT-I

2. Compare and contrast the various N-gram taggers. Develop a python program to demonstrate the working of N-gram tagger.

3. Discuss the applications of NLP in
   a. Routing Support Tickets.
   b. Sentiment Analysis.
   c. Business Data Analysis
   d. Customer Service Chatbots
   e. Interactive Voice Response (IVR) Systems.

UNIT-II

4 With an example, explain in detail how the parse tree is generated using the Shift-reduce parser.

5. Explain the need for parsing text in NLP applications. For the sentence "The clever fox jumped over the wall", explain the generation of the parse tree with a suitable illustration.

UNIT-III

6. Compare and contrast between Question Answering Systems with Dialog Systems.

7. Explain the underlying concept used by most Information Retrieval systems. Discuss how itis applied to 1. Boolean Retrieval Model. 2. Vector Space Model.

UNIT-IV

8. Develop Web Crawlers using Python Programming Language
   1. To find the current temperature in a given city and display it.
   2. To print three levels of headers for a given website.

9. Consider a CSV file containing SMS messages, where each message contains two attributes, SMS_ID, SMS_DATA. Develop a python program to perform tokenization, removing stop words and words less than three letters, lemmatize and convert all text to lowercase.