

D128771(022)

**B. Tech. (Hon's) (Seventh Semester) Examination,
Nov.-Dec. 2024**

(Data Science)

BIG DATA ANALYTICS

Time Allowed : Three hours

Maximum Marks : 100

Minimum Pass Marks : 35

Note : All questions are compulsory. Part (a) of each unit is compulsory and carries 4 marks. Attempt any two parts from (b), (c) and (d) and carries 8 marks each.

Unit-I

1. (a) Explain the architecture of HDFS. 4
- (b) Represent Hadoop ecosystem and explain each component. 8

- (c) Explain the ethical and legal considerations for big data analytics. 8
- (d) Write a short note on NoSQL database. Establish a comparison between NoSQL and traditional SQL. 8

Unit-II

2. (a) What is the importance of Feature selection? 4
- (b) Differentiate Apache Pig and Hive. Describe the advantages of using Apache Pig and Apache Hive for scalable data pre-processing in big data environments. 8
- (c) What are the techniques to identify and handle outliers in big data? Provide examples of when removing outliers may lead to loss of important information. 8
- (d) Elaborate Map Reduce. Write the classic Map Reduce word count algorithm to count the frequency of words in a large text corpus stored in HDFS. 8

Unit-III

3. (a) Write the difference between Hadoop MapReduce and Spark. 4
- (b) Compare and contrast Resilient Distributed Datasets

(RDDs) and DataFrames in Apache Spark. Provide an example where DataFrames offer performance advantages over RDDs. Also, explain how Spark SQL integrates with DataFrames for querying large datasets.

8

(c) Explain how Apache Flink handles real-time stream processing and state management. Discuss the concept of time windows (event time, processing time) and give an example where Flink's stream processing can be used to monitor real-time sensor data for anomaly detection.

8

(d) Define clustering in machine learning. Explain how it differs from supervised learning algorithms like regression and classification.

8

Unit-IV

4. (a) Explain the architecture and working principles of artificial neural networks (ANNs).

4

(b) Discuss the key techniques used in NLP such as tokenization, stemming, and word embeddings. Provide an example of how NLP can be applied to a large dataset for sentiment analysis.

8

(c) Explain the concept of ensemble learning in machine

learning. Compare and contrast different ensemble techniques like bagging, boosting, and stacking. 8

- (d) Discuss model evaluation techniques, focusing on metrics like accuracy, precision, recall, and the F1 score. 8

Unit-V

5. (a) Define cloud-based big data analytics and briefly describe the role of platforms like Amazon Web Services (AWS) in handling large datasets. 4

- (b) How do Docker and Kubernetes together improve the scalability and portability of big data applications? Provide examples to illustrate their use. 8

- (c) Discuss deployment and scaling strategies in cloud-based big data analytics. Explain horizontal and vertical scaling, and how cloud providers facilitate these scaling options for handling fluctuating workloads. 8

- (d) Compare the big data analytics services offered by Microsoft Azure, and Google Cloud Platform (GCP). Discuss key features, such as data storage, processing, and real-time analytics, that make each platform suitable for large-scale data analysis. 8

D128772(022)

**B. Tech. (Hon's) (Seventh Semester) Examination,
Nov.-Dec. 2024**

(Data Science Branch)

DATA WRANGLING

Time Allowed : Three hours

Maximum Marks : 100

Minimum Pass Marks : 35

Note : Attempt all questions. Part (a) of each question is compulsory, each of 4 marks. Attempt any **two** parts from (b), (c) and (d) each of 8 marks.

Unit-I

1. (a) What is data wrangling, and why is it important in data analysis?

4

- (b) What is Exploratory Data Analysis (EDA), and how does it help in identifying data quality issues? 8
- (c) Compare the advantages and limitations of file I/O and database access for data acquisition. 8
- (d) How do missing data, outliers, and duplicates affect the results of a machine learning model? 8

Unit-II

2. (a) What is one-hot encoding? Provide an example. 4
- (b) Define the operations of stack and unstack in pandas. Provide an example. 8
- (c) Design a step-by-step approach for transforming a raw dataset with mixed types of data (numerical and categorical) into a structured format suitable for analysis. 8
- (d) What are the key techniques for feature extraction from text data? 8

Unit-III

3. (a) Explain the difference between outer join and inner join with suitable illustrations. 4

- (b) Discuss the implications of ignoring duplicate keys in the merging process. What potential issues could arise? 8
- (c) Explain the potential pitfalls of using fuzzy matching in data merging. How can these be addressed? 8
- (d) Using Python's Pandas library, demonstrate how to perform each type of join with code example. 8

Unit-IV

4. (a) What are spatial joins? How do they differ from regular joins in databases? 4
- (b) Describe the difference between upsampling and downsampling in time series resampling. When would each be used? 8
- (c) Describe a scenario where geocoding would be applied. What are the benefits of using geocoded data. 8
- (d) Describe the challenges faced when merging different datasets in GIS. How can these challenges be overcome? 8

Unit-V

5. (a) Explain the basic workflow of using Git for a new project. 4
- (b) Discuss the importance of documentation in data wrangling and how it impacts project outcomes. 8
- (c) Discuss how reproducibility affects results and collaboration. 8
- (d) Discuss the essential components of a well-organized data wrangling project. 8

D127773(022)

**B. Tech. (Hon's) (Seventh Semester) Examination,
Nov.-Dec. 2024**

(New Scheme)

(Specialization : Artificial Intelligence)

Time Allowed : Three hours

Maximum Marks : 100

Minimum Pass Marks : 35

Note : Attempt all questions. Part (a) of each question is compulsory, each of 4 marks. Attempt any **two** parts from (b), (c) and (d) each of 8 marks.

Unit-I

- | | |
|---|---|
| 1. (a) What are the uses of SRS document? | 4 |
| (b) Describe requirements engineering. | 8 |

- (c) Explain the main phases in the software development life cycle. 8
- (d) What are the principles and practices of agile software development model. 8

Unit-II

2. (a) What is modularity principle of software design? 4
- (b) Define different levels of abstraction. Explain various types of attributes in entity-relationship diagram with example. 8
- (c) What is unified modeling language? Explain various types of structural diagrams in unified modeling language. 8
- (d) Write short notes on : 8
- (i) Client-Server architecture
- (ii) Microservices architecture

Unit-III

3. (a) What do you understand by conventions in software engineering? What are the various coding conventions? 2+2=4

- (b) Explain different programming paradigms in software development. Give two example of documentation tools. 6+2=8
- (c) What is unit testing? Explain test driven development. 8
- (d) What is "big bang" non-incremental integration? Explain the types of integration testing. 2+6=8

Unit-IV

4. (a) Explain version control systems. 4
- (b) Draw the work flow of software configuration management. What approaches are used for modernizing the legacy systems? 2+6=8
- (c) Define software maintenance. Explain the types of maintenance. 8
- (d) Describe various code refactoring techniques. Give example of encapsulate-field refactoring technique. 6+2=8

Unit-V

5. (a) Explain work breakdown structure. 4

(b) What are the roles and events of scrum framework?

Explain the components of Kanban framework. 4+4=8

(c) What are the steps involved in risk management?

Discuss various mitigation strategies in software projects. 4+4=8

(d) (i) Describe two effort estimation models. 4+4=8

(ii) Suppose we have the following counts for different elements, with their respective weights :

Element type	Count	Weight
External inputs	15	4
External outputs	6	5
Internal logic	5	7

Calculate the following :

(a) Unadjusted function points

(b) Adjusted function points based on adjustment factor (CAF) = 1.2

D127733(022)

**B. Tech. (Hon's) (Seventh Semester) Examination,
Nov.-Dec. 2024**

(AICTE Scheme)

(Data Science/Artificial Intelligence)

IMAGE PROCESSING

Time Allowed : Three hours

Maximum Marks : 100

Minimum Pass Marks : 35

Note : All questions are compulsory. Part (a) of each unit is compulsory and carries 4 marks. Attempt any **two** parts from (b), (c) and (d) and carries 8 marks each.

Unit-I

1. (a) Write specialized image processing hardware? 4
- (b) Explain the concept of image processing? 8

D127733(022)

PTO

- (c) Explain Image Acquisition and Pre-processing Techniques. 8
- (d) Discuss the various Image Processing Steps. 8

Unit-II

2. (a) What is Brightness Adoption? 4
- (b) Explain the different types of gray level transformation. 8
- (c) Write short notes on : 8
- (i) Image Denoising
- (ii) Image deblurring.
- (d) Explain Frequency domain Enhancement Techniques. 8

Unit-III

3. (a) What is image segmentation. 4
- (b) Explain various feature extraction techniques. 8
- (c) Write short notes on : 8

- (i) Thresholding
- (ii) Region based segmentation.
- (d) Describe Edge Detection Algorithms. 8

Unit-IV

4. (a) Define Mathematical Morphology. 4
- (b) Explain various mathematical morphology operation. 8
- (c) Explain application of morphological operation in image processing. 8
- (d) Discuss Morphological image Processing for Noise removal and feature extraction. 8

Unit-V

5. (a) Define wavelet based compression. 4
- (b) Discuss Image Compression Techniques. 8

- (c) Write short notes on : 8
- (i) Image Registration
 - (ii) Fusion Method.
- (d) Write case study on Medical imaging. 8