18343

BRAINWARE UNIVERSITY

Term End Examination 2025-2026
Programme – M.Tech.(CSE)-AIML-2025
Course Name – Big Data Analytics
Course Code - MTA10405
(Semester I)

Full Marks : 60 **Time : 2:30 Hours**

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A
(Multiple Choice Type Question) 1 x 15=15

1. *Choose the correct alternative from the following :*

(i) Interpret the role of Volume in Big Data systems.
a) Manages data types b) Handles data speed
c) Measures data size d) Filters useful data

(ii) Data arriving from social media platforms in real time mainly relates to:
a) Volume b) Velocity
c) Variety d) Veracity

(iii) Investigate a scenario where predictive analytics fails.
a) Past trends do not align with future b) Too much training data
c) Lack of volume d) Small database

(iv) Apply a suitable model for analyzing clickstream data
a) Linear regression b) Time series forecasting
c) Classification tree d) Clustering

(v) Implement an approach to scale data storage horizontally
a) Increase CPU cores b) Add more RAM
c) Use a distributed file system d) Deploy GPU computation

(vi) Identify that is not a design factor for Big Data databases
a) Fault tolerance b) Latency reduction
c) Low scalability d) Data security

(vii) Explain the need for distributed processing in Hadoop.
a) To reduce server costs b) To increase latency
c) To support scalability and parallelism d) To increase disk usage

(viii) Recognize a component of the Hadoop Ecosystem.
a) MySQL b) Pig
c) MongoDB d) Redis

(ix) Summarize the purpose of YARN in Hadoop.

a) To store files
b) To query data
c) To manage cluster resources
d) To clean data

(x) Design a data flow between HDFS and MapReduce.

a) Use Sqoop
b) Illustrate HDFS read/write
c) Use command-line utilities
d) Write queries in Hive

(xi) Execute data exploration using .describe() in Pandas.

a) Returns statistical summary
b) Returns nulls
c) Filters columns
d) Deletes duplicates

(xii) Choose method to detect missing values.

a) isnull()
b) dropna()
c) fillna()
d) remove()

(xiii) Assess the usefulness of Pandas for root cause detection.

a) Excellent for patterns
b) Poor at querying
c) Only visualization tool
d) Too complex

(xiv) Classify supervised learning models.

a) Linear regression, K-means
b) Decision trees, Random forest
c) PCA, ICA
d) Naive Bayes, DBSCAN

(xv) Choose the model least prone to overfitting.

a) Decision tree
b) KNN
c) Random forest
d) Linear regression with many features

Group-B

(Short Answer Type Questions) 3 x 5=15

2. Explain the concept of Volume in Big Data. (3)
3. Classify the different types of analytics used in Big Data systems. (3)
4. Design a simplified pipeline using Hadoop Streaming for analyzing log files. (3)
5. Demonstrate correlation matrix for a sample dataset in Python. (3)
6. Differentiate between machine learning and computational models in analytics. (3)

OR

Analyze the limitations of traditional storage in supporting real-time analytics. (3)

Group-C

(Long Answer Type Questions) 5 x 6=30

7. Explain the components of supervised learning and outline a typical workflow. (5)
8. Evaluate the pros and cons of Decision Tree Algorithm for predictive analytics. (5)
9. Describe the design considerations required to build a Big Data storage system. (5)
10. Create a project plan to analyze customer sentiment using Hadoop. (5)
11. Evaluate the role of Exploratory Data Analysis (EDA) in guiding diagnostic analytics workflows. (5)
12. Analyze the importance of the Big Data Analytics Lifecycle in data-driven projects. (5)

OR

Examine the limitations of traditional data processing systems in handling Big Data. (5)

****