

Term End Examination January – 2026

Course	: MSCDS	Date	: 25/02/2026
Subject Code	: MSCDS-203	Time	: 11:30am to 01:45pm
Subject Name	: Big Data Analytics using “R”	Duration	: 02.15 Hours
		Max. Marks	: 70

Section A

Answer the following (Attempt any three)

(30)

1. Define Big Data and explain its key characteristics (5 Vs). Discuss why the R programming language is considered a powerful tool for large-scale data analytics.
2. Explain the architecture of the Hadoop Distributed File System (HDFS). How do the NameNode and DataNode components work together to ensure data reliability and fault tolerance?
3. Describe the MapReduce programming model. Illustrate the Map and Reduce phases with a practical example, such as a word count or log analysis program.
4. Discuss Nonlinear Models in R. Explain the significance of Decision Trees and Random Forests when predicting categorical outcomes in complex datasets.
5. Compare and contrast SQL and NoSQL databases. In what scenarios would a Data Scientist choose a NoSQL environment like MongoDB over a traditional relational database?

Section B

Answer the following (Attempt any four)

(20)

1. Explain the different Data Types in R (Vectors, Matrices, and Data Frames) and how they are used for data manipulation and storage.
2. What is Data Ingestion? Describe how tools like Apache Flume or Sqoop facilitate the movement of data from external sources into the Hadoop ecosystem.
3. Explain the purpose of the RHIPE package. How does it enable users to perform big data analysis by integrating R with Hadoop?
4. Discuss Generalized Additive Models (GAMs). What advantages do they offer over simple linear regression when dealing with non-linear relationships?
5. What is Feature Importance in the context of Random Forests? Explain how evaluating feature importance helps in the process of model refinement and selection.
6. Discuss the role of Data Visualization in Big Data analytics and name three essential R packages used for creating high-quality graphical representations.

Section C

Part – A (Multiple Choice Questions)

(10)

- Which of the following is NOT one of the 5 Vs of Big Data?
A Volume
B Velocity
C Validity
D Variety
- In HDFS, what is the default block size in Hadoop 2.x and later?
A 64 MB
B 128 MB
C 256 MB
D 512 MB
- Which R function is used to read a CSV file into a Data Frame?
A load.csv()
B read.csv()
C get.csv()
D import.csv()

- 4 In a MapReduce job, which phase comes immediately after the Map phase?
A Reduce B Shuffle and Sort
C Output D Combine
- 5 Which NoSQL database is known for its "Column-family" storage model?
A MongoDB B Cassandra
C Redis D Neo4j
- 6 Which R package is widely used for creating complex and aesthetic data visualizations?
A ggplot2 B dplyr
C tidyr D shiny
- 7 The process of reducing the number of input variables when developing a predictive model is:
A Data Cleaning B Feature Selection
C Over fitting D Under fitting
- 8 Which Hadoop component is responsible for resource management and job scheduling?
A HDFS B YARN
C Hive D Pig
- 9 A Random Forest is an ensemble of multiple:
A Linear Regressions B Decision Trees
C K-Means Clusters D Neural Networks
- 10 Which operator in R is used for the "Pipe" operation in the tidyverse?
A -> B ==
C %>% D %*%

Part – B (Do as Directed)

(10)

State whether the following statements are true or false

- 1 R is an open-source language primarily used for statistical computing and graphics.
- 2 HDFS is designed to store data on a single high-end expensive server.
- 3 The 'Reducer' function in MapReduce aggregates the intermediate data based on keys.
- 4 MongoDB stores data in flexible, JSON-like documents.
- 5 In R, a Vector can store elements of different data types simultaneously.
- 6 Logistic Regression is used for predicting categorical dependent variables.
- 7 Apache Spark is often faster than MapReduce because it performs in-memory processing.
- 8 Non-linear least squares (NLS) models always converge on the first iteration.
- 9 The 'rm()' function in R is used to remove objects from the workspace.
- 10 Big Data analytics can help organizations identify trends and patterns for better decision-making.
