K24N 0223

Reg. No. : .....

Name : .....

Third Semester M.Sc. Degree (CBSS – Regular) Examination, October 2023
(2022 Admission)
STATISTICS WITH DATA ANALYTICS
MST3C11 : Big Data Analytics

Time : 3 Hours Max. Marks : 80

PART – A

Answer **all** questions. **Each** question carries **2** marks. **(8×2=16)**

1. Explain the concept of Big Data characteristics.
2. Define Big Data sources with examples.
3. Discuss the importance of Data Ownership in the context of Big Data analytics. What are the legal and ethical considerations associated with it ?
4. Describe the Hadoop framework and its role in Big Data processing.
5. Explain the concept of Data Privacy.
6. Types of files in HDFS.
7. Define Hive architecture and its components.
8. Describe the role of Pig in Big Data analytics and its advantages.

PART – B

Answer **any four** questions. **Each** question carries **4** marks. **(4×4=16)**

9. Explain the design goals of HDFS. How does HDFS ensure fault tolerance and high availability ?
10. Discuss the Master-slave architecture of Hadoop. What are the responsibilities of the NameNode and DataNode in this architecture ?

P.T.O.

11. Describe the process of installing HDFS. What are the steps involved in setting up a Hadoop cluster ?

12. Compare and contrast Spark with traditional MapReduce. How does Spark overcome the limitations of MapReduce ?

13. Discuss the Spark architecture, including Resilient Distributed Dataset (RDD) and Directed Acyclic Graph (DAG). How do these components contribute to Spark's performance ?

14. Explain the concept of stream processing models and tools. How does Spark Streaming enable real-time data processing ?

PART – C

Answer **any four** questions. **Each** question carries **12** marks. **(4×12=48)**

15. Discuss the role of Apache Spark in Big Data analytics. How do Spark's ecosystem components, such as MLlib, Spark GraphX, and SparkSQL, contribute to data analysis ?

16. Explain the architecture of the Hadoop Distributed File System (HDFS). How does HDFS manage data storage and replication across a distributed environment ?

17. Describe the process of reading and writing data in HDFS. What are the advantages of using HDFS for large-scale data storage ?

18. Discuss the components and functionality of Apache Pig. How does Pig Latin facilitate data processing tasks in Big Data environments ?

19. Compare and contrast Apache Hive with traditional relational database systems. What are the benefits of using Hive for data warehousing and analysis ?

20. Explain the concept of partitioning in Hive. How does partitioning improve query performance in Hive ?