.

Printed Page:-3 **Subject Code:- BCSAI0603**

Roll. No:

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA
(An Autonomous Institute Affiliated to AKTU, Lucknow)
B.Tech
SEM: VI - THEORY EXAMINATION (2025 - 2026)
Subject: Big Data Analytics

Time: 3 Hours **Max. Marks: 100**

General Instructions:

IMP: *Verify that you have received the question paper with the correct course, code, branch etc.*

1. *This Question paper comprises of* ***three Sections -A, B, & C.*** *It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.*

2. *Maximum marks for each question are indicated on right -hand side of each question.*

3. *Illustrate your answers with neat sketches wherever necessary.*

4. *Assume suitable data if necessary.*

5. *Preferably, write the answers in sequential order.*

6. *No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.*

SECTION-A 20

1. Attempt all parts:-

(1.1) Big Data is primarily characterized by: (CO1, K1) 1

(a) Volume, Variety, Velocity, Veracity, Value
(b) Size, Shape, Speed, Security, Storage
(c) Data, Drivers, Design, Deployment, Distribution
(d) Accuracy, Access, Availability, Architecture, Auditing

(1.2) Data replication in HDFS ensures: (CO1, K1) 1

(a) Faster computation only
(b) Fault tolerance and reliability
(c) Reduced storage requirements
(d) Elimination of metadata

(1.3) The master node in HDFS is: (CO2, K1) 1

(a) DataNode
(b) NameNode
(c) NodeManager
(d) Reducer

(1.4) YARN is responsible for: (CO2, K1) 1

(a) Storage
(b) Resource management
(c) Querying
(d) Visualization

(1.5) Default execution mode of Pig (CO3, K1) 1

(a) Local
(b) MapReduce
(c) Distributed
(d) Hybrid

(1.6) HiveQL is similar to (CO3, K1) 1

(a) C
(b) SQL
(c) Java
(d) Python

(1.7) Sqoop job helps in: (CO4, K1) 1

(a) Creating reusable import/export commands
(b) Creating Hive partitions
(c) Monitoring Flume channels
(d) Serializing Kafka messages

(1.8) One of the supported Sqoop file formats is: (CO4, K1) 1

(a) Avro
(b) YAML
(c) DOCX
(d) HTML

(1.9) The keyword used to define a mutable variable in Scala is (CO5, K1) 1

(a) val
(b) const
(c) var
(d) mut

(1.10) A class in Scala is mainly used to (CO5, K1) 1

(a) Define a blueprint for objects
(b) Store only numbers
(c) Execute shell commands
(d) Replace loops

2. Attempt all parts:-

(2.1) Explain how data replication is handled in HDFS. (CO1, K2) 2

(2.2) Mention two advantages associated with fault tolerance in Hadoop. (CO2, K2) 2

(2.3) List different ways in which Pig programs can be executed. (CO3, K2) 2

(2.4) Mention the use of --hive-import option. (CO4, K2) 2

(2.5) Illustrate arithmetic and relational operators in Scala with examples. (CO5, K2) 2

SECTION-B 30

Answer any one of the following:-

(3.1) Compare data replication vs block storage in Hadoop. (CO1, K3) 6

(3.2) Examine the role of Hadoop I/O compression. (CO1, K4) 6

Answer any one of the following:-

(3.3)	Explain the concept of replication in HDFS and discuss how it ensures fault tolerance. (CO2, K3)	6
(3.4)	Describe Hadoop daemons and explain their roles. (CO2, K2)	6

Answer any one of the following:-

(3.5)	Analyze the characteristics of Pig Latin that support efficient data transformation (CO3, K3)	6
(3.6)	Differentiate the approach adopted by Hive from that of conventional database systems. (CO3, K3)	6

Answer any one of the following:-

(3.7)	Describe incremental import in Sqoop using append and lastmodified modes. (CO4, K3)	6
(3.8)	Explain multiplexing in Flume using event headers and routing logic. (CO4, K3)	6

Answer any one of the following:-

(3.9)	Explain Scala basic types and operators with suitable code snippets. (CO5, K2)	6
(3.10)	Explain the use of MLlib for machine learning tasks in Spark. (CO5, K3)	6

SECTION-C 50

4. Answer any one of the following:-

(4.1)	Critically analyze the history of Hadoop and its evolution. (CO1, K4)	10
(4.2)	Critically analyze the importance of Big Data in social media expansion. (CO1, K4)	10

5. Answer any one of the following:-

(5.1)	A company faces high processing time despite powerful nodes. Examine how data locality impacts performance and suggest improvements. (CO2, K4)	10
(5.2)	Investigate task execution in MapReduce and challenges in distributed systems. (CO2, K4)	10

6. Answer any one of the following:-

(6.1)	Analyze both the structural and execution aspects involved in Pig data processing. (CO3, K4)	10
(6.2)	Describe important considerations involved in designing schemas within HBase. (CO3, K4)	10

7. Answer any one of the following:-

(7.1)	Design a Flume-based solution for collecting logs from multiple servers into centralized storage. (CO4, K4)	10
(7.2)	Explain Kafka producer and consumer workflow with suitable examples. (CO4, K3)	10

8. Answer any one of the following:-

(8.1)	Explain RDDs in detail including transformations, actions, persistence, lineage, and fault tolerance. (CO5, K4)	10
(8.2)	Develop a detailed case-oriented solution for enterprise-scale Big Data analytics using Spark and related tools. (CO5, K4)	10