Printed Page:-04 **Subject Code:- ACSDS0601**

Roll. No:

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech

SEM: VI - THEORY EXAMINATION (2024 - 2025)

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-a. State which component in the Hadoop ecosystem provides a distributed coordination service for distributed applications? [CO1,K1] 1

(a) ZooKeeper
(b) Spark
(c) Mahout
(d) Flume

1-b. State what is the role of the Data Node in HDFS? [CO1,K1] 1

(a) To manage the metadata of the files stored in the HDFS
(b) To store the data in the HDFS
(c) To process and analyze the data stored in the HDFS
(d) To provide a SQL-like interface to query and analyze data stored in Hadoop

1-c. State whether Fact tables are _____ ? [CO2,K1] 1

(a) HDFS
(b) MapReduce
(c) YARN
(d) All of the Above

1-d. Identify the correct definition of Reconciled data? [CO2,K1] 1

(a) Reconcile data is a data scored in one operational system in the organization
(b) Reconcile data is the data that has been selected and formatted for end-user support

.

applications.

(c) Reconcile data is the current data intented to be the single source for all decision support systems

(d) None

1-e. State which of the following is not a benefit of file-based data structures in Hadoop? [CO3,K1] 1

(a) Efficient data compression

(b) High data compression ratio

(c) Fast query processing

(d) High data redundancy

1-f. State which file-based data structure supports nested data types such as maps, arrays, and structs? [CO3,K1] 1

(a) Avro

(b) Parquet

(c) ORC

(d) JSON

1-g. State If the database contains some tables then it can be forced to drop without dropping the tables by using the keyword? [CO4,K1] 1

(a) RESTRICT

(b) OVERWRITE

(c) F DROP

(d) CASCADE

1-h. Recall whether the thrift service component in hive is used for? [CO4,K1] 1

(a) moving hive data files between different servers

(b) use multiple hive versions

(c) submit hive queries from a remote client

(d) Installing hive

1-i. State whether Sqoop written in? [CO5,K1] 1

(a) C

(b) C++

(c) Java

(d) hadoop

1-j. List how Data processed by Scoop can be used for? [CO5,K1] 1

(a) Hbase

(b) HDFS

(c) Mapreduce

(d) MahOut

2. Attempt all parts:-

.

2.a.	Discuss the importance of Big Data in analytics and state how Machine learning can be implemented on Big Data Hadoop ? [CO1,K2]	2
2.b.	State and explain which file format is better for your data Avro vs Parquet? [CO2,K1]	2
2.c.	Discuss in detail about Hadoop Distributed File System (HDFS) and list all its commands? [CO3, K2]	2
2.d.	Discuss what is the purpose of the "SELECT" command in HiveQL? [CO4,K2]	2
2.e.	Discuss Apache Pig all execution mechanisms in detail ? [CO5, K2]	2

SECTION-B 30

3. Answer any five of the following:-

3-a.	Describe Big Data architecture with neat and clean diagram and explain all ecosystems in detail? [CO1, K2]	6
3-b.	Discuss Big Data in Healthcare, Transportation & Medicine? [CO1, K2]	6
3-c.	Explain the following (i) HDFS replication (ii) Rack awareness (iii) Locality reference ? [CO2,K2]	6
3-d.	Explain the following (i) HDFS safe mode (ii) Name Node high availability [CO2,K2]	6
3.e.	Explain what happens if during the PUT operation, HDFS block is assigned a replication factor 1 instead of the default value 3 ? [CO3, K2]	6
3.f.	Explain with Neat sketch explain in detail Apache Hive architecture? [CO4, K2]	6
3.g.	Discuss and write a program in Spark Core Processing RDD to run Word Count program? [CO5, K2]	6

SECTION-C 50

4. Answer any one of the following:-

4-a.	Discuss the Pig Latin data types and explain with examples? [CO1,K2]	10
4-b.	Explain Concept of blocks in details, Let incoming data be 300 files and per file is 200 MB. IF block size is 64 MB and RF 03, then how many blocks will be created? [CO1, K1]	10

5. Answer any one of the following:-

5-a.	Discuss and write a short note on Serialization, Bytes Writable? [CO2, K1]	10
5-b.	Discuss and Write short notes on? [CO2, K1] (i) Driver code (ii) Mapper code (iii) Reducer code (iv) Combiner	10

6. Answer any one of the following:-

6-a.	Explain Hadoop "Mapred-Site.xml" conf file in detail? [CO3,K2]	10
6-b.	Differentiate between compression and serialization in detail? [CO3,K4]	10

7. Answer any one of the following:-

7-a.	Explain Hive Integration & Work Flow steps involved with a diagram? [CO4, K2]	10
7-b.	Discuss what are the important components of the Spark ecosystem, Explain how Spark runs applications with the help of its architecture? [CO4,K2]	10

8. Answer any one of the following:-

8-a.	Explain what is the role of a shuffle operation in Hadoop's Spark framework, and how does it contribute to sorting and aggregating data? [CO5, K1]	10
8-b.	Describe what is Apache ZooKeeper and Apache Oozie, How do you configure an “Oozie” job in Hadoop, Explain in Detail? [CO5,K2]	10

.