KUMARAGURU
college of technology
character is life

Register Number:……………..

M.TECH DEGREE EXAMINATIONS: APRIL / MAY 2023

(Regulation 2018)

Second Semester

DATA SCIENCE

P18ITI2207: Big Data Technologies

COURSE OUTCOMES

CO1: Identify applications require big data technologies.

CO2: Explain Hadoop Architecture - HDFS, YARN and Map Reduce.

CO3: Perform administration and configuration of Hadoop Ecosystem.

CO4: Write basic queries and scripts in Hive and Pig.

CO5: Write advanced queries and scripts using hive and pig - aggregation, joins, sorting.

CO6: Discuss the need of HBase and write queries to use HBase as data source for Big Data.

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-

PART A (10 x 1 = 10 Marks)

1. **Assertion (A):** Hadoop do need specialized hardware to process the data. CO3 [K_2]

   Reason (R): Hadoop 2.0 allows live stream processing of real time data.

   a) Both A and R are Individually true and R is the correct explanation of A
   b) Both A and R are Individually true but R is not the correct explanation of A
   c) A is true but R is false
   d) A is false but R is true

2. Which of the following is used by Facebook to Tackle Big Data based on Hadoop. CO1 [K_1]

   a) Project Prism
   b) Prism
   c) Project Big
   d) Project Data

3. Match List I with List II CO4 [K_2]

List I	List II
A) hdfs fsck / -files -blocks	i) Periodically merge the namespace image with the edit log.
B) Secondary namenode	ii) Blocks that make up each file in the filesystem.
C) YARN	iii) Local
D) PigUnit	iv) Replication

	A	B	C	D
a)	ii	i	iii	iv
b)	iii	iv	i	ii
c)	iv	i	ii	iii
d)	iii	i	ii	iv

4. Match List I with List II CO2 [K_2]

List I	List II
A. HBase	i) Data summarization and ad hoc querying
B. Hive	ii) Parallel computation
C. Pig	iii)Performance coordination service
D. Zookeeper	iv) Structured data storage for large tables

	A	B	C	D
a)	ii	i	iii	iv
b)	iii	iv	i	ii
c)	iv	i	ii	iii
d)	iii	i	ii	iv

5. **Assertion (A):** RAID is turned off by default CO3 [K_2]

Reason (R): Hadoop is designed to be a highly redundant distributed system

a) Both A and R are Individually true and R is the correct explanation of A

b) Both A and R are Individually true but R is not the correct explanation of A

c) A is true but R is false

d) A is false but R is true

6. Which of the following server is a machine that keeps a copy of the state of the entire system and persists this information in local log files. CO2 [K_1]

a) Master

b) Region

c) Zookeeper

d) MapReduce

7. Point out the correct statement : CO5 [K_2]

i) Hive Commands are non-SQL statement such as setting a property or adding a resource

ii) Set -v prints a list of configuration variables that are overridden by the user or Hive

iii) Set sets a list of variables that are overridden by the user or Hive

iv) HiveServer2 has a new JDBC driver

a) ii, iii

b) i

c) ii

d) iii

8. Which of the following statement will create column with varchar datatype? CO5 [K_2]

i) CREATE TABLE foo (bar CHAR(10))

ii) CREATE TABLE foo (bar VARCHAR(10))

iii) CREATE TABLE foo (bar CHARVARYING(10))

a) i b) ii, iii

c) iii d) i, ii, iii

9. Order the execution of map reduce. CO2 [K_2]
1) Shuffling 2) Reducer 3) Mapping 4) Inputsplits

5) Input 6) Output

a) 1-2-4-3-5-6 b) 3-4-5-1-2-3

c) 5-4-3-1-2-6 d) 2-5-6-1-3-4

10. Which of the following most popular high-level Java API in Hadoop Ecosystem? CO3 [K_1]

a) Scalding b) HCatalog

c) Cascalog d) Cascading

PART B (10 x 2 = 20 Marks)

11. What is Big Data Analytics? List the tools for Big Data Analytics. CO1 [K_2]

12. Explain the problems with traditional methods of approach for Big Data Analytics? CO1 [K_2]

13. List the features of HDFS. CO2 [K_1]

14. What is role of Rack Awareness Algorithm in HDFS? CO2 [K_2]

15. List the difference between NameNode and DataNode. CO3 [K_2]

16. What is a heartbeat in HDFS? CO3 [K_2]

17. Suppose there is file of size 514 MB stored in HDFS (Hadoop 2.x) using default block size configuration and default replication factor. Then, how many blocks will be created in total and what will be the size of each block? CO4 [K_2]

18. What is a distributed cache in MapReduce Framework? CO4 [K_1]

19. What are the different data types in Pig Latin? CO5 [K_1]

20. When should we use SORT BY instead of ORDER BY? CO5 [K_2]

PART C (10 x 5 = 50 Marks)

21. Describe about the various Hadoop daemons and their roles in a Hadoop cluster. CO1 [K_2]

22. Explain “Big Data” and what are five V’s of Big Data? CO1 [K_2]

23. State the reason why we can’t perform “aggregation” (addition) in mapper? Why do we need the “reducer” for this? CO2 [K_2]

24. Differentiate between Hadoop 1.x and Hadoop 2.x architecture. CO2 [K_2]

25. Explain briefly about Apache Hadoop HDFS Architecture. CO3 [K_2]

26. Enumerate how does replication management takes place in HDFS with neat diagram? CO3 [K_2]

27. What is YARN and explain the main components of YARN. CO4 [K_2]

28. Describe about Hadoop ecosystem with neat diagram. CO4 [K_2]

29. How does Apache Pig provide abstraction over MapReduce and list the features? CO5 [K_2]

30. Explain the features of HBase and Zookeeper. CO5 [K_2]

Answer any TWO Questions

PART D (2 x 10 = 20 Marks)

31. Illustrate the steps involved in setting up a single node hadoop cluster with all necessary configuration changes needed. CO1 [K_2]

32. Explain about MapReduce framework to perform distributed and parallel processing on large data sets in a distributed environment. Consider the text input Dear, Bear, River, Car, Car, River, Deer, Car and Bear, use the input to explain MapReduce. CO3 [K_3]

33. Describe about components of Hive and explain about Hive DDL and DML commands for Online Analytical Processing. CO5 [K_3]

****