b. Examine the different types of NoSQL databases. Compare the components of traditional databases and MongoDB. 8 1 5 5

25. a. With code snippets, state how can we create line, histogram, pie and bar chart in Python. 8 1 6 5

(OR)

b. List and detail on the on-premises and cloud computing enterprise infrastructure solutions. 8 1 6 4

PART – C (1 × 15 = 15 Marks)

Answer **ANY ONE** Question

Marks BL CO PO

26. Describe on applying map reduce programming paradigm for the given scenario. The requirement is to find the total number of clicks to every link from the website's log file. The log file format is 15 2 3 5

<Link ID, Link Name, Click_Count>

Example entries in the file:

L123, Contact Us, 3
L321, Feedback, 5
L123, Contact Us, 4
L114, Locate Us, 2
L321, Feedback, 1
L891, MBank, 3
L114, Locate Us, 1

The expected output format is

Contact Us, 7
Feedback 6
Locate Us, 3
MBank, 4

Discuss on inputs and output of each phase of MR and the logic.

27. Discuss the HDFS command used for the below purposes with code snippets. 15 2 3 5

(i) Recursively remove a directory and its contents including the contents of its subdirectories
(ii) Move files between different folders of HDFS
(iii) Create a new file in HDFS and display its content
(iv) Copy of file from local file system to HDFS and delete the local copy

* * * * *

Reg. No.

B.Tech/ M.Tech (Integrated) DEGREE EXAMINATION, MAY 2023

Fourth Semester

21CSE222T – BIG DATA TOOLS AND TECHNIQUES

(For the candidates admitted from the academic year 2022-2023 onwards)

Note:

(i) **Part - A** should be answered in OMR sheet within first 40 minutes and OMR sheet should be handed over to hall invigilator at the end of 40th minute.

(ii) **Part - B** and **Part - C** should be answered in answer booklet.

Time: 3 Hours Max. Marks: 75

PART – A (20 × 1 = 20Marks)

Answer **ALL** Questions

Marks BL CO PO

1. ______ is the component responsible for storage in hadoop. 1 1 2 5
(A) Map reduce (B) Hadoop Distributed File System (HDFS)
(C) Yet Another Resource Negotiator (YARD) (D) Resource manager

2. The voice transcriptions of a media file has to be processed to answer few questions. This is _______ data. 1 1 1 5
(A) Quasi-structured (B) Semi-structured
(C) Unstructured (D) Structured

3. The first step in preparing an application for parallel processing is ______. 1 1 1 5
(A) Creating destination for writing output (B) Clean the data for missing values
(C) Configure node manager on different machines (D) Identity tasks that could run concurrently

4. _______ is not a NoSQL database. 1 1 2 5
(A) Cassandra (B) Google Big Table
(C) Microsoft Access (D) Redis

5. Which one of the below has the smallest size? 1 1 3 5
(A) Hard disk block (B) Operating system file system block
(C) Map reduce input file split (D) HDFS block

6. The map reduce phase first segments the input files into splits, which have the default size of ______. 1 1 3 5
(A) 128 Mb (B) User defined
(C) 64 Mb (D) 1 GB

7. Hadoop ______ is a utility that comes with Hadoop, in which mapper and reducer will be scripts or programs in any language. 1 1 3 5
(A) Streaming (B) Lucene
(C) Yarn (D) Solr

8. When there is a single key, the type of architecture that best fits map reduce application is _______. 1 2 3 5

(A) Parallel reduce (B) Zero reduce
(C) Single reduce (D) Multiple reduce

9. Redundant Array of Independent Disks (RAID) is not used by Hadoop because _______. 1 1 4 5

(A) An entire array cannot be used if one disk fails (B) It is not redundant
(C) Uses available hard disks without special configuration (D) It is faster

10. In configuration of a Hadoop cluster, where 'n' denotes the node. 'r' denotes rack and 'd' represents data center, which one of the below have least distance from d2/r2/n3? 1 2 4 5

(A) d2/r1/n1 (B) d2/r3/n3
(C) d2/r2/n50 (D) d1/r2/n4

11. An application processing event data, wants to log the events in persistent medium for later usage. The component that can be used for this is ______. 1 2 4 5

(A) Yarn (B) Flume
(C) Zookeeper (D) Sqoop

12. The _______ type of table in Hive manages both data and meta data 1 1 4 5

(A) External (B) Imported
(C) Linked (D) Managed

13. All the communications between processes in Apache flink is done through ______. 1 1 5 5

(A) Rest API (B) Metastore
(C) Stored procedures (D) Remote procedures

14. Which of the following statements is incorrect about Mongo DB? 1 1 5 5

(A) Older records can be dropped based on size limit (B) Schema and tables should be created before data inserts
(C) Multiple key indexing possible (D) Need not define a schema beforehand

15. Stateful computations are effectively managed in apache flink similar to stateless computations using ______. 1 1 5 5

(A) Metastore (B) Dictionaries
(C) Keystore (D) Functions

16. ______ object in Apache spark is the entry point of a spark program. 1 1 5 5

(A) Text context (B) Spark context
(C) Main spark (D) Root context

17. The data point [3.67, 8.19, 4.33, 12.12……] represent ______ data. 1 1 6 5

(A) Continuous (B) Clustered
(C) Binned (D) Discrete

18. The png command in R is used to 1 1 6 5

(A) Convert text to images (B) Print image on printer
(C) Give file name to the plot (D) Display plot on a separate window

19. A car company wants to show the mileage of its different car models in a chart. Which one is more suitable 1 1 6 5

(A) Line chart (B) Bar chart
(C) Heat map (D) Pie chart

20. The _______ tool is a standard data warehousing tool used in industry for many years in a reliable manner. 1 1 6 5

(A) Hbase (B) Hive
(C) Teradata (D) kdB+

PART – B (5 × 8 = 40 Marks)

Answer **ALL** Questions

Marks BL CO PO

21. a. Indicate on the significance of using cloud computing for big data, its advantages, limitations and different architectures. 8 2 1 5

(OR)

b. Summarize the steps in building a corporate big data strategy as a part of big data mining. 8 2 1 5

22. a. Relate on how integrity and compression is done on data stored in HDFS and Hadoop. 8 2 3 5

(OR)

b. Discuss on the different types of schedules available in Yarn and when to use each of them. 8 2 3 5

23. a. State and detail the purpose of the below pig Latin commands. 8 1 4 5

(i) STORE
(ii) LOAD
(iii) DUMP
(iv) DESCRIBE

(OR)

b. Using Hive, describe how we execute basic create table, load data and select data from the below table, with code snippets. 8 1 4 5

Link click
viewTime : Int
userID : Big Int
subLinks : Array
properties : map of key value pairs

24. a. State the reasons for using spark and detail on its core data structures. 8 1 5 5

(OR)