b. Recite how distributed and parallel computing is most suited for handing big data and discuss about the merits and demerits of this system. 10 1 2 3

27. a. With neat sketch illustrate the architecture of YARN and discuss about the workflow and components of hadoop yarn in detail. 10 1 2 1

(OR)

b. The database company maintains the log file of the users creating the tables in the following format. 10 3 3 5

<USER NAME, USER ID, VISITING DATA, TABLE NAME>

It is decided to find out the total number of times each user logged into the server to use the database. Execute the mapper and reducer logic for the a fore mentioned scenario.

Sample file:

David	1001	20-10-13	Emp
John	1002	23-10-17	Student
David	1001	20-11-13	Transaction
Ravi	1003	19-10-14	Employee
John	1002	23-11-17	Enterprise

28. a.i. An analytics company wants to import the large amount of data from SQL (Relational database). Justify the architecture of the framework required to load the data into HDFS environment. 6 4 3 5

ii. Compare the features flume with sqoop. 4 1 1 0

(OR)

b.i. List out the common services provided by zookeeper. 4 2 2 3

ii. Recognize the different types of znodes available in zookeeper and recall the benefits of zookeeper. 6 2 2 3

29. a. With a neat sketch identify the core components of apache spark and discuss in detail about its architecture. 10 2 5 4

(OR)

b.i. Identify the three types of job common in oozie and recall its features. 5 3 2 5

ii. Memorize and brief the main architecture component of apache flink. 5 2 5 5

30. a. Discuss in detail about the data science solutions in the enterprise data science. 10 2 6 3

(OR)

b.i. Write the code snippet in python and R to visualize the data. Assume your own input in printing the output. 6 3 6 5

ii. Mention about the big data visualization tools. 4 2 6 5

* * * * *

Reg. No. | | | | | | | | | | | | | | |

B.Tech. DEGREE EXAMINATION, NOVEMBER 2022
Sixth/ Seventh Semester

18CSE391T – BIG DATA TOOLS AND TECHNIQUES
(For the candidates admitted from the academic year 2018-2019 to 2019-2020)

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** should be answered in answer booklet.

Time: 2½ Hours Max. Marks: 75

PART – A (25 × 1 = 25 Marks)
Answer **ALL** Questions

Marks BL CO PO

1. Pick out the storage component of hadoop architecture. 1 1 1 1
(A) Resource manager (B) Yet another resource negotiator (YARN)
(C) Hadoop distributed file system (HDFS) (D) Map reduce

2. The function that gets an intermediate key and set of values corresponding to the key 1 2 2 1
(A) Map (B) Reduce
(C) Shuffle (D) Sort

3. The initial step in preparing any application for parallel processing is ______. 1 1 1 1
(A) Select the source of input (B) Copy data into HDFS
(C) Identity task that could run concurrently (D) Configure CPUS on different machines

4. The hadoop ecosystem tool used to schedule hadoop jobs in the form of directed a cyclic graph is _______. 1 1 1 1
(A) Oozie (B) Hbase
(C) Hive (D) Solr

5. The type of cloud that remains entirely in the ownership of the organization using it is called as 1 1 1 1
(A) Public cloud (B) Private cloud
(C) Hybrid cloud (D) Community cloud

6. The following hadoop command "hadoop fs – copytoLocal abc.txt abc1.txt" 1 2 1 2
(A) Copies the file abc.txt from HDFS as abc1.txt to local file systems (B) Copies the file abc1.txt from HDFS as abc1.txt
(C) Copies the file abc.txt and the file abc1.txt from HDFS to Ubuntu system (D) Copies the file abc1.txt from local file system as abc.txt to HDFS

7. Which one of the following is not true about check sums in hadoop data integrity? 1 2 1 2

(A) Checksums are computed first time when the file enters the application

(B) Common method used in HDFS is CRC 32 cycle redundancy check

(C) Checksum fixes the data corruption at the network receiving end

(D) Checksum storage is usually less than 1 percent of the file size

8. Choose the command used to check whether all nodes were up in hadoop. 1 2 1 2

(A) SPS (B) CPS

(C) DFS (D) JPS

9. The default input format of map reduce in hadoop is _____. 1 2 1 2

(A) Text input format (B) Key value text input format

(C) Sequence file input format (D) Sequence file as text input format

10. The hadoop scheduler attempt to allocate resources so that all running applications get the same share of resources 1 2 1 2

(A) Capacity scheduler (B) FIFO scheduler

(C) HDFS scheduler (D) Fair scheduler

11. The component of pig that creates directed acyclic graph (DAG) which represent the pig Latin statements and logic operator is ______. 1 2 1 2

(A) Parser (B) Optimizer

(C) Compiler (D) Execution engine

12. ______ type of znode in zookeeper is alive even after the client which created that particular znode is disconnected 1 2 2 5

(A) Ephemeral znode (B) Persistence znode

(C) Sequential znode (D) Non-sequential znode

13. ______ is the hadoop tool designed to support bulk export and import of data into hdfs from the relational databases 1 1 2 5

(A) Oozie (B) Zookeeper

(C) Flume (D) Sqoop

14. The events generated by external source (web server) are consumed by ______ component of flume. 1 1 2 5

(A) Flume data sink (B) Flume data source

(C) Flume data channel (D) HDFS

15. Which pig command is used to display results in the screen? 1 1 2 5

(A) Dump (B) Load

(C) Group (D) Execute

16. Apache spark ______ library is faster than mahout. 1 2 5 4

(A) Spark streaming (B) Spark SQL

(C) Spark MLIB (D) Spark core

17. Hbase and Cassandra are ______ database 1 2 5 5

(A) Column oriented (B) Row oriented

(C) Column and row oriented (D) XML oriented

18. Apache flink manages stateful computations as effectively as stateless one using ______. 1 1 5 5

(A) Protocols (B) Functions

(C) Data store (D) Disk store

19. Which type of database stores the record in JSON format? 1 1 5 5

(A) Cassandra (B) Hbase

(C) MongoDB (D) Hive QL

20. _____ is used by a company wants to show its yearly performance of three products for an interval of years? 1 1 6 3

(A) Heat map (B) Pie chart

(C) Bar chart (D) Histogram

21. The ______ is a popular enterprise level data ware house tool that has been in industry around many years and is highly reliable 1 1 6 3

(A) Exalytics (B) Informatics

(C) Teradata (D) Hbase

22. The ______ chart tries to show the intensity with the depth of colours in different colour palette. 1 1 6 5

(A) Heat map (B) Horizontal bar

(C) Pie (D) Historgram

23. The tool used to create attractive web interface to depict analytics output is ______. 1 2 6 5

(A) Rplot (B) Matplot lib

(C) Rshiny (D) Seaborn

24. The data points [12, 10, 8.5 and so on] are classified as ______. 1 2 6 5

(A) Discrete (B) Continuous

(C) Binned (D) Clustered

25. The best chart best for plotting two or more discrete quantities to compare them against a same base axis 1 1 6 5

(A) Bar chart (B) Pie chart

(C) Line (D) Stacked bar

PART – B (5 × 10 = 50 Marks) Marks BL CO PO

Answer **ALL** Questions

26. a. Draw the map reduce flow chart and explain every step in executing the map reduce paradigm. 10 3 1 1

(OR)