Reg. No															

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

Fourth Semester

21CSE222T - BIGDATA TOOLS AND TECHNIQUES

(For the candidates admitted during 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 = 20 Marks)

Answer **all** Questions

		Marks	BL	CO
1.	Which one of the following is the storage component of Hadoop? (A) Yet Another Resource Negotiator (B) Resource Manager (C) Hadoop Distributed File System (HDFS) (D) Map Reduce	1	1	1
2.	The voice transcriptions of multimedia content is a type of ____. (A) Semi-structured data (B) Structured data (C) Unstructured data (D) Quasi-structured data	1	1	2
3.	The first step in preparing an application for parallel processing is _____. (A) Select the destination for writing output (B) Copy data into HDFS (C) Configure CPUs on different machines (D) Identify tasks that could run concurrently	1	1	1
4.	Which one of the below services is owned by cloud service provider in Infrastructure as a Service (IaaS) (A) Applications (B) Data (C) Storage (D) Runtime	1	1	2
5.	In Hadoop, _____ is the concept of having more than one name node when the cluster size is huge. (A) HDFS Blocking (B) HDFS Federation (C) Name space imaging (D) Name node pooling	1	1	3
6.	Common method used in HDFS fpr data integrity check is _____. (A) CRC-32C Cyclic redundancy check (B) MD5 Standard (C) Token based authentication (D) Hashmap convergence	1	1	3
7.	The _____ object in a map reduce program is used as an interface by mappers and reducers to emit their output to and read their input from. (A) RecordReader (B) Context (C) OutputFormatClass (D) InputValueClass	1	1	3
8.	Which of the following is NOT TRUE about a map reduce file split? (A) A map reduce split is always of the size 128 MB (B) A good map reduce split should be of the size of the HDSFS block (C) Managing too small splits are difficult (D) Map function is executed once for each line or record of a split	1	2	3

9.	The _____ service in Zookeeper takes care of managing the leaving and joining of a node in the cluster. (A) elector (B) leader (C) cluster management (D) ensemble	1	1	4
10.	Which of the below is NOT a part of Pig architecture? (A) Parser (B) Combiner (C) Optimizer (D) Execution Engine	1	2	4
11.	In Pig Latin data model, _____ is a Single value irrespective of data type. (A) field (B) atom (C) row (D) bag	1	1	4
12.	Hive architecture is _____. (A) store-based (B) update-based (C) write-based (D) read based	1	1	4
13.	Apache flink manages stateful computations as effectively as stateless ones using _____. (A) protocols (B) functions (C) datastore (D) diskstore	1	2	5
14.	In Apache Spark, the _____ operation is one that computes a new RDD from existing RDDs. (A) count (B) collect (C) take (D) join	1	1	5
15.	In MongoDB, _____ model, we refer the sub documents in the original document, using references. (A) relational (B) normalized (C) embedded (D) external	1	2	5
16.	The _____ is the batch processing API of Flink (A) DataSetAPI (B) DataStreamAPI (C) FlinkAPI (D) DataAPI	1	1	5
17.	The _____ chart is best for plotting two or more discrete quantities against categorical variables to compare them against a same base axis. (A) bar (B) pie (C) heatmap (D) stacked bar	1	2	6
18.	The _____ is the top most challenge in enterprise data science and big data. (A) selection of tools for data analysis (B) storage and delivery of analysis output (C) speeding of data processing (D) data cleansing	1	2	6
19.	The _____ from HP is a column-oriented, massively parallel processing database system with features such as support for in-database machine learning, and native integration to open source systems. (A) Cassandra (B) Vertica (C) GreenPlum (D) Hana	1	1	6
20.	The _____ is the library used for image recognition. (A) OpenCV (B) Cognos (C) seaborn (D) pandas	1	2	6

PART - B (5 × 8 = 40 Marks) — Marks | BL | CO

Answer **all** Questions

21.	(a) Identify and list the different types of data with an example for each type.	8	1	2
	(OR)			
	(b) Explain the uses of enterprise big data mining and the steps involved in the same			

22.	(a) Detail on the purpose of Application Master in YARN. With a clear diagram, explain the process of YARN application job scheduling.	8	2	3
	(OR)			
	(b) Describe on how data integrity and compression is performed in Hadoop.			
23.	(a) Compare and identify the purpose of usage of RAID and JBOD type of hard disks. Hadoop uses which type of hard disks for HDFS storage and why.	8	3	4
	(OR)			
	(b) Summarize with code snippets on the process involved in creating a Pig relation out of a file and running diagnostic operators on the relation. Assume sample data of your choice.			
24.	(a) Summarize on the data model used in MongoDB, as compared with the traditional databases	8	2	5
	(OR)			
	(b) Compare and contrast on the purpose and usage of the different data structures or abstractions in Spark			
25.	(a) Compare and summarize on any two Enterprise and open source NoSQL Databases used for Enterprise data warehouse and data mining.	8	2	6
	(OR)			
	(b) With sample code snippets, detail on the usage of bar charts and heat maps.			

PART - C (1 × 15 = 15 Marks)	**Marks**	**BL**	**CO**
Answer **any 1** Questions			
26. Discuss the HDFS commands used for the below purposes with sample code snippets. i. Recursively remove a directory and its contents including the contents of its sub directories ii. Copy a file from local file system to HDFS and then delete the local copy iii. Create a new file in HDFS iv. Recursive listing of the contents of a folder including the contents of the sub directories.	15	2	2
27. Given a web server log file having entries of the format <Date, Time, UserID, PageName, ActionPerformedInPage>. The company has decided to use map reduce programming paradigm to get frequently visited page for each user. Present the logic for mapper and reducer along with the input to each phase. Sample log file entries: 11-07-2021,11:10:26,saru80,Deposit,Submit 12-07-2021,12:40:18,nisha_1980,Transactions,Download 12-07-2021,15:36:52,roha@bank,ManageBeneficiary,AddPayee 13-07-2021,18:22:18,deepak04,Bill,PayEB 13-07-2021,11:10:32,nisha_1980,Account,ViewBalance 14-07-2021,12:11:40,nisha_1980,Account,ViewBalance	15	2	2

* * * * *