ii. What are the three modes in which Hadoop can run? Describe with suitable example for each. **5**

iii. Who takes care of replication consistency in a Hadoop cluster and what do under/over replicated blocks mean? **5**

******

Total No. of Questions: 6 *Total No. of Printed Pages:4*

Enrollment No.....

Faculty of Engineering

End Sem (Even) Examination May-2022

CS3ET05 Big Data Analytics

Programme: B.Tech. Branch/Specialisation: CSE

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d.

Q.1 i. What type of variable is a telephone area code? **1**

(a) Numerical, continuous (b) Numerical, discrete

(c) Categorical (d) Categorical, ordinal

ii. Real time processing deals with _____ characteristics of data. **1**

(a) Variety (b) Variability (c) Velocity (d) Volatility

iii. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop? **1**

(a) Big data management and data mining

(b) Data warehousing and business intelligence

(c) Management of Hadoop clusters

(d) Collecting and storing unstructured data

iv. Which of the following platforms does Hadoop run on ___? **1**

(a) Bare metal (b) Debian

(c) Cross-platform (d) Unix-like

v. A file in HDFS that is smaller than a single block size: **1**

(a) Cannot be stored in HDFS.

(b) Occupies the full block's size.

(c) Occupies only the size it needs and not the full block.

(d) Can span over multiple blocks.

vi. HDFS provides a command line interface called _____ used to interact with HDFS. **1**

(a) "HDFS Shell" (b) "FS Shell"

(c) "DFS Shell" (d) hbase

vii. _____ maps input key/value pairs to a set of intermediate key/value pairs. **1**
(a) Mapper (b) Reducer
(c) Both (a) and (b) (d) None of these

viii. _____ is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution. **1**
(a) Map Parameters (b) JobConf
(c) MemoryConf (d) None of these

ix. A _____ node acts as the Slave and is responsible for executing a Task assigned to it by the ResourceManager. **1**
(a) MapReduce (b) Mapper
(c) NodeManager (d) DataNode

x. What is the full form of YARN? **1**
(a) Yet Another Resource Network
(b) Yet Another Relational Negotiator
(c) Yet Another Resource Negotiator
(d) Yet Another Relational Network

Q.2 i. What size of the data can be considered as Big Data? Is size of the data the only attribute of the data that makes it a Big Data? **2**

ii. With proper examples discuss and differentiate structured, unstructured and semi-structured data. **3**

iii. What is specific need of Big Data that cannot be accommodated by RDBMS? What is the solution? **5**

OR iv. What can be the real-life applications of clustering in the Big data like Banking Data, Mega Store transaction data? What inferences can be made using clustering algorithms? **5**

Q.3 i. What are the advantages of Hadoop? **2**

ii. What is the role of Hadoop Cluster Management? Describe following properties of Hadoop Cluster Management with suitable example. **8**
(a) SSI (Single System Image)
(b) Fault tolerance
(c) High availability

OR iii. Define with suitable block diagram the role of following component in Hadoop Architecture: **8**
(a) NameNode (b) Node Manager
(c) Resource Manager (d) Data Node
(e) Secondary NameNode

Q.4 i. Write the use and syntax of following HDFS commands: **3**
(a) mkdir (b) Chown (c) put

ii. In HDFS file system in Hadoop framework distributes the data over different nodes. What are the criteria for the block size of the data? What is the effect of having very small or very large block size? **7**

OR iii. How does read and write operation of perform with a HDFS cluster? Define step –by step operations of read and write with suitable block diagram. **7**

Q.5 i. How is MR-I different than MR-2? What all things are taken care of with this update? **4**

ii. One input file (File-1) contains following text dataset. Design a block diagram to depict execution of algorithm for WordCount using MapReduce programming paradigm. **6**

Java, Java, .NET, Scala
SQL, Java, C#, C++
C#, C++, C++ C#
Java, SQL, Scala, .NET

File 1

OR iii. N dimensional numerical values are written in a row in the Text file. **6**
(a) Write Map- Reduce pseudo code for implementing K Means clustering, clearly specify the key-Value pair.
(b) What are the challenge in doing K Means clustering using Map Reduce?

Q.6 Attempt any two:

i. What is eraser coding? How does it help to overcome storage issues of Hadoop -3.x over Hadoop 2.x? **5**

Marking Scheme
CS3ET05 Big Data Analytics

Q.1	i.	What type of variable is a telephone area code? (c) Categorical		**1**
	ii.	Real time processing deals with _____ characteristics of data. (c) Velocity		**1**
	iii.	According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop? (a) Big data management and data mining		**1**
	iv.	Which of the following platforms does Hadoop run on ___? (c) Cross-platform		**1**
	v.	A file in HDFS that is smaller than a single block size: (c) Occupies only the size it needs and not the full block.		**1**
	vi.	HDFS provides a command line interface called _____ used to interact with HDFS. (b) "FS Shell"		**1**
	vii.	_____ maps input key/value pairs to a set of intermediate key/value pairs. (a) Mapper		**1**
	viii.	_____ is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution. (b) JobConf		**1**
	ix.	A _____ node acts as the Slave and is responsible for executing a Task assigned to it by the ResourceManager. (c) NodeManager		**1**
	x.	What is the full form of YARN? (c) Yet Another Resource Negotiator		**1**
Q.2	i.	Size of the data can be considered as Big Data	1 mark	**2**
		Size of the data the only attribute of the data	1 mark	
	ii.	Structured data	1 mark	**3**
		Unstructured data	1 mark	
		Semi-structured data	1 mark	
	iii.	Specific need of Big Data that cannot be accommodated by RDBMS	3 marks	**5**
		Solution	2 marks	
OR	iv.	Identify a real word bid data applications	3 marks	**5**
		Define clustering problem for selected big data	2 marks	
Q.3	i.	Define Hadoop	1 mark	**2**
		Advantages of Hadoop	1 mark	
	ii.	Role of Hadoop Cluster Management	2 marks	**8**
		(a) SSI (Single System Image)	2 marks	
		(b) Fault tolerance	2 marks	
		(c) High availability	2 marks	
OR	iii.	Block diagram of Hadoop cluster and explanation	3 marks	**8**
		(a) NameNode	1 mark	
		(b) Node Manager	1 mark	
		(c) Resource Manager	1 mark	
		(d) Data Node	1 mark	
		(e) Secondary NameNode	1 mark	
Q.4	i.	Write the use and syntax of following HDFS commands: 1 mark for each	(1 mark * 3)	**3**
	ii.	Criteria for the block size of the data	4 marks	**7**
		Effect of having very small or very large block size	3 marks	
OR	iii.	Define step –by step operations of read and write 2 marks for each (2 marks * 2)	4 marks	**7**
		Block diagram for read and write operation (one for each) 1.5 marks for each (1.5 marks * 2)	3 marks	
Q.5	i.	New things of MR-2 introduced over MR-I	2 marks	**4**
		Define advantages for MR-2 over MR-I	2 marks	
	ii.	Draw a flow diagram for work count program using MapReduce framework	3 marks	**6**
		Explanation of execution flow of above wordcount program	3 marks	
	iii.	(a) Map- Reduce pseudo code for implementing K Means clustering, clearly specify the key-Value pair.	3 marks	**6**
		(b) Challenge in doing K Means clustering using Map Reduce	3 marks	

Q.6		Attempt any two:		
	i.	Eraser coding	2 marks	**5**
		Implementation of eraser coding in Hadoop -3.x	1 mark	
		Advantages of eraser coding with Hadoop -3.x over Hadoop 2.x	2 marks	
	ii.	Three modes in which Hadoop	3 marks	**5**
		Example for each	2 marsk	
	iii.	Explanation of replication consistency in a Hadoop cluster	3 marks	**5**
		Under/over replicated blocks mean	2 marks	

******