Scheme	**R2012**
Semester	VIII
Course Code	CPE8035
Course Name	Big Data Analytics

Question No.	Question	a	b	c	d
1	Find the L1 and L2 distances between the points (5, 6, 7) and (8, 2, 4).	L1 =10 , L2 = 5.83	L1 =10 , L2 = 5	L1 =11 , L2 = 4.9	L1 =9 , L2 = 5.83
2	The hardware term used to describe Hadoop hardware requirements is	Commodity firmware	Commodity software	Commodity hardware	Cluster hardware
3	CSV and JSON can be described as	Structured data	Unstructured data	Semi-structured data	Multi-structured data
4	Which of the following Operation can be implemented with Combiners?	Selection	Projection	Natural Join	Union
5	Which of the following is not a Hadoop Distributions?	MAPR	Cloudera	Hortonworks	RMAP
6	If size of file is 4 GB and block size is 64 MB then number of mappers required for MapReduce task is	8	16	32	64
7	A Reduce task receives	one or more keys and their associated value list	key value pair	list of keys and their associated values	list of key value pairs
8	NOSQL is	Not only SQL	Not SQL	Not Over SQL	No SQL
9	Neo4j is an example of which of the following NoSQL architectural pattern?	Key-value store	Graph Store	Document Store	Column-based Store
10	_____is a batch-based, distributed computing framework modeled after Google's paper.	MapCompute	MapReuse	MapCluster	MapReduce
11	"Sharding" a database across many server instances can be achieved with _____	MAN	LAN	WAN	SAN
12	ETL stands for _____	Extraction transformation and loading	Extract Taken Lend	Enterprise Transfer Load	Entertainment Transference Load
13	Hadoop is the solution for:	Database software	Big Data Software	Data Mining software	Distribution software
14	The time between elements of one stream	need not be uniform	need to be uniform	must be 1ms.	must be 1ns
15	In Bloom filter an array of n bits is initialized with	all 0s	all 1s	half 0s and half 1s	all -1
16	Which of the following is not the default daemon of Hadoop?	Namenode	Datanode	Job Tracker	Job history server
17	What do you mean by sampling of stream data?	Sampling reduces the amount of data fed to a subsequent data mining algorithm.	Sampling reduces the diversity of the data stream	Sampling aims to keep statistical properties of the data intact.	Sampling algorithms often doesn't need multiple passes over the data
18	Which of the following statements about data streaming is true?	Stream data is always unstructured data.	Stream data often has a high velocity.	Stream elements cannot be stored on disk.	Stream data is always structured data.

19	The DGIM algorithm was developed to estimate the counts of 1's occur within the last k bits of a stream window N. Which of the following statements is true about the estimate of the number of 0's based on DGIM?	The number of 0's cannot be estimated at all.	The number of 0's can be estimated with a maximum guaranteed error	To estimate the number of 0s and 1s with a guaranteed maximum error, DGIM has to be employed twice, one creating buckets based on 1's, and once created buckets based on 0's.	Determine whether an element has already occurred in previous stream data.
20	if Distance measure d(x, y)= d(y, x) then it is called	Symmetric	identical	positiveness	triangle inequality
21	_____ stores are used to store information about networks, such as social connections.	Key-value	Wide-column	Document	graph
22	What is the edit distance between A=father and B=feather ?	5	1	4	2
23	Sliding window operations typically fall in the category	OLTP Transactions	Big Data Batch Processing	Big Data Real Time Processing	Small Batch Processing
24	_____ systems focus on the relationship between users and items for recommendation.	DGIM	Collaborative-Filtering	Content Based and Collaborative Filtering	Content Based
25	Find Hamming Distance for vectors A=100101011 B=100010010	2	4	3	1