Sample Questions

Department of Information Technology

Subject Name: Big Data Analytics **Semester: VIII**

Multiple Choice Questions

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	Type of consistency in BASE for NOSQL is
Option A:	Eventual Consistency
Option B:	Strong Consistency
Option C:	Partition Consistency
Option D:	Weak Consistency
2.	An algorithm that divides the entire file of baskets into segments small enough so that all frequent itemset for the segment can be found in main memory is:
Option A:	PCY Algorithm
Option B:	Randomized Algorithm
Option C:	DGIM Algorithm
Option D:	SON Algorithm
3.	Which of the following factors have an impact on the Google PageRank?
Option A:	The total number of inbound links to a page of a web site
Option B:	The subject matter of the website
Option C:	The count of number of times a word repeats on a website
Option D:	The number of outbound links from the page
4.	Map function takes which of the following as input:
Option A:	File on the desktop
Option B:	HDFS block on Data Node
Option C:	File on the server
Option D:	Block on the server
5.	Two k-cliques are adjacent when they share
Option A:	2*k nodes
Option B:	k+1 nodes
Option C:	k-1 nodes
Option D:	k nodes
6.	Identify 3V’s of Big Data
Option A:	Volume, Velocity & Variety
Option B:	Volume, Velocity & Variability
Option C:	Volume, Velocity & Veracity
Option D:	Visualization, Velocity & Value

7.	PCY algorithm is used in the field of big data analytics for
Option A:	Filtering the data stream with large data
Option B:	Hierarchical clustering for large data
Option C:	Frequent itemset mining when the dataset is very large.
Option D:	Counting triangles in social networks
8.	Stream Queries are basically questions asked about the current state of the stream or streams is called as
Option A:	Continuous Queries
Option B:	Adhoc Queries
Option C:	One-time Queries
Option D:	Predefined Queries
9.	Heartbeat is used to communicate between
Option A:	Job Tracker & Task Tracker
Option B:	Name node & Secondary Name Node
Option C:	Job Tracker & Name Node
Option D:	Data Node & Name Node
10.	How Bloom's Filter is different than other filtering algorithms in Data Stream Mining?
Option A:	Bloom's Filter does not use a hash function, whereas other filtering algorithms use hash values.
Option B:	Bloom's Filter uses probabilistic data structure whereas other algorithms do not use probabilistic data structure.
Option C:	Bloom's Filter uses fix structures of data as compared to other.
Option D:	Bloom's Filter is not a filtering algorithm.

11.	Which is an important feature of Big Data Analytics?
Option A:	Portability
Option B:	Scalability
Option C:	Reliability
Option D:	Durability
12.	A sparse matrix system that uses a row and a column as keys is called as
Option A:	Advanced Store
Option B:	Data structures
Option C:	Key-value store
Option D:	Column family store
13.	What do you always have to specify for a MapReduce job?
Option A:	The classes for the mapper and reducer
Option B:	The classes for the mapper, reducer, and combiner
Option C:	The classes for the mapper, reducer, partitioner, and combiner
Option D:	You need not specify anything as all classes have default implementations
14.	The only security feature that exists in Hadoop is

Option A:	Name Node and Data Node Permissions
Option B:	HDFS file- and directory-level ownership and permissions
Option C:	Map Reduce Permissions
Option D:	Zookeeper
15.	In which of the relational algebra operations, the reduce function is identity?
Option A:	Intersection
Option B:	Projection
Option C:	Union
Option D:	Selection
16.	Assume that a text file contains following text. This is a test. Yes it is In a map-reduce logic of finding frequency of occurrence of each word in this file, what is the output of map function?
Option A:	(This,1), (is, 1), (a, 1), (a,1)
Option B:	(This,1), (is, 1), (a, 1), (test., 1), (Yes, 1), (it, 1), (is, 1)
Option C:	(This,1), (is, 2), (a, 1), (test., 1), (Yes, 1), (it, 1), (is, 1)
Option D:	(This,1), (is, 2), (a, 1), (test., 1), (Yes, 1), (it, 1)
17.	Flajolet-Martin Algorithm depends upon
Option A:	Linear function and Binary Equivalent trailing zeros
Option B:	Hash function and Binary Equivalent trailing once
Option C:	Hash function and Binary Equivalent trailing zeros
Option D:	Hash function and Decimal Equivalent trailing zeros
18.	In Decaying window algorithm, we assign
Option A:	more weight to newer elements
Option B:	less weight to newer elements
Option C:	more weight to older elements
Option D:	less weight to older elements
19.	In DGIM algorithm,
Option A:	If a bucket contains a frequent pair, then the bucket is surely frequent
Option B:	If a bucket contains a frequent pair, then the bucket is surely not frequent
Option C:	If a bucket not contains a frequent pair, then the bucket is surely frequent
Option D:	If a bucket not contains a frequent pair, then the bucket is surely not frequent
20.	In FM algorithm , For each stream element a, r(a) be the number of _____ in h(a)
Option A:	trailing 0's
Option B:	trailing 1's
Option C:	all 0's
Option D:	all 1's

21.	Euclidean Distance between Age 21 and 24 and Income 500 and 504 is
Option A:	5
Option B:	25
Option C:	7
Option D:	678
22.	Jaccard Distance between Set1 = {1,0,1,1,1} and Set2 = {1,0,0,1,1} is
Option A:	3/4
Option B:	1/4
Option C:	2/4
Option D:	1
23.	A Bloom filter consists of an array of n bits, initially all :
Option A:	Garbage Value
Option B:	1's
Option C:	0’s.
Option D:	Combination of 0's and 1's
24.	Algorithm to estimate number of distinct elements seen in the stream.
Option A:	FM Algorithm
Option B:	DGIM algorithm
Option C:	HITS Algorithm
Option D:	Bloom Filter
25.	The right end of a bucket in DGIM algorithm is always a position with a
Option A:	even number
Option B:	combination 0 's and 1's
Option C:	0
Option D:	1
26.	A collection of pages whose purpose is to increase the PageRank of a certain page or pages is called a

Option A:	page rank
Option B:	spam farm.
Option C:	dead end
Option D:	spider trap
27.	To compute page rank we need to know the
Option A:	probablity that a random surfer will land at the page
Option B:	size of the page in bytes
Option C:	sequence of the page
Option D:	web servers name
28.	In PCY Algorithm which technique is used to filter unnecessary itemset
Option A:	Association Rule
Option B:	Hashing Technique
Option C:	Data Mining
Option D:	Market basket
29.	Euclidean Distance between Age 21 and 24 and Income 500 and 504 is
Option A:	5
Option B:	25
Option C:	7
Option D:	678
30.	Jaccard Distance between Set1 = {1,0,1,1,1} and Set2 = {1,0,0,1,1} is
Option A:	3/4
Option B:	1/4
Option C:	2/4
Option D:	1

Descriptive Questions

Q No	10 marks each
1	Explain the types of NoSQL data stores and their typical usage.
2	Explain working of all phases of MapReduce with one common example.
3	Explain how Hadoop goals are covered in Hadoop distributed file system.
4	Explain Page rank algorithm with an example. State the problems occurred in the algorithm and ways to solve them.
5	Explain Park-Chen-Yu algorithm. How memory mapping is done in PCY.
6	How is recommendation done based on properties of product? Explain with suitable example.
7	Explain CURE algorithm with Initialization and Completion phases.
8	Explain PageRank algorithm with a suitable example.
9	Explain Girvan Newman method for community detection in social network.
10	Explain NOSQL design patterns with its benefits and example.
11	Discuss 2 step Matrix-Matrix Multiplication algorithm using MapReduce with example.
12	What is Hadoop? Describe HDFS architecture in detail. Give advantages and limitations of Hadoop.
13	What is HDFS? List features of HDFS?
14	Define PageRank? Illustrate PageRank calculation?
15	Define Jaccard Distance? Find Jaccard distances between the following pair of vectors? [1, 2, 3, 4, 5, 6] and [3, 4, 5, 6, 7, 8]

Q No	5 marks each
1	What are three V's of Big Data? Give two example of big data case studies. Indicate which V's are satisfied by these case studies.
2	For following operations write the Map Reduce pseudo code: 1. Matrix Vector multiplication 2. Selection 3. Union
3	List the different issues and challenges in data stream query processing.
4	Explain how failures are handled in MapReduce job?
5	What is DGIM? State the rules used in DGIM Algorithm.
6	Explain CURE algorithm, clearly stating its advantages over traditional clustering

	algorithm.
7	Give problems in Flajolet-Martin (FM) algorithm to count distinct elements in a stream.
8	Explain the nearest neighbor problem. What similarity measure can be used in an application to find plagiarism in documents.
9	Explain the importance of counting triangles in social networks.
10	Give importance of “Shuffle and Sort” phase of Hadoop.
11	Differentiate between SQL and NoSQL.
12	Define Blooms filter and list its application.
13	Explain FM algorithm with example.
14	Explain HITS algorithm.
15	List and comment different models of Recommendation System.

—X—X—