UNIVERSITY OF MUMBAI
SAMPLE MCQ QUESTION BANK (100 questions)

Course Code and Name: BDA ITC801 /R16
Class: BE
Semester:8

Module No	Weightage in Hrs	Q NO	QUESTION (2 marks per question)	OPTIONS				Correct Answer
				A	B	C	D	
1	3	1	What are the 3v's of Big Data?	volume, velocity	Volume, Velocity	volume,	Variety,	B
		2	Apache Hadoop is an _____________ for storage	passive-source	active-source	closed-source	open-source	D
		3	The composition of the data with the	sensitivity of	availability of	structure of	state of data	C
		4	Human generated data __________	Financial data	Network log	Input data	Gaming data	A
		5	What is true about Variety in bigdata?	high in size	speed of data	data from	data in certain	D
		6	which is structured data examples	CSV but XML	social media posts	tab delimited	Medical device	D

Module No	Weighatge in Hrs	Q NO	QUESTION (2 marks per question)	OPTIONS				Correct Answer
				A	B	C	D	
2	10	1	Which type of data storage system cassandra is	distributed	centralized	parallel	dumb	A
		2	The default replication factor in Hadoop is	4	3	5	2	B
		3	What was Hadoop named after?	Creator Doug	Cutting high	The toy	A sound	C
		4	NoSQL CAP theorem -------------	Consistency,Ava	Consistency,Acce	Consistency,A	confidentiality	D
		5	MongoDB provides horizontal scaling through___	Replication	Partitioning	Sharding	Document	C
		6	Point out the correct statement.	DataNode	DataNode is the	DataNode is	Hadoop	B
		7	Which of the following is a wide-column store?	Cassandra	Riak	MongoDB	Redis	A
		8	Procedural language for developing parallel	Pig Latin	Hive	Pig	Oozie	C
		9	Cassendra is a popularly known as _________ data	distributed	centralized	parallel	dumb	A
		10	When a backup node is used in a cluster there is no	Check point	Secondary data	Secondary	Rack awareness	C
		11	Hadoop developed by _______________	Larry Page	Doug Cutting	Mark	Bill Gates	B
		12	No SQL systems are also referred to as	"Not-On-SQL“	"N-Only-SQL“	"No-only-	"Not-Only-	D
		13	One of classified NoSQL databases is	Key-value	value	key	DataNode	A
		14	A ________ store is a simple database that when	document	key-value	graph	simple	B
		15	Social connectionsstores are used to store in	Stack	Tree	Graph	Documents	C
		16	MongoDB scales horizontally using Sharding for	data balancing	load distribution	memory	load balancing	D
		17	HDFS inherited from ------------- file system.	Yahoo	FTFS	Google	Rediff	C
		18	Which is the column-oriented distributed database.	HBase	NOSQL_IBM	MSSQL	MySQL	A

		19	Cost factor in Hadoop cluster setup.	inexpensive	only i7 machine	graphics card	cloud required	A
		20	MongoDB database has been used by number of	backend	proprietary	GUI design	front end	A
Module No	**Hours**	**Q NO**	**QUESTION (2 marks per question)**	**OPTIONS**				**Correct Answer**
				A	**B**	**C**	**D**	
3	9	1	________ node acts as the Slave and is responsible	MapReduce	Mapper	TaskTracker	JobTracker	c
		2	Which function is responsible for consolidating the	Reduce	Map	Reducer	Mapper	a
		3	Which function maps input key/value pairs to a set	Mapper	Reducer	Combiner	Execute	a
		4	_____ is the slave/worker node and holds the user	Data block	NameNode	DataNode	Replication	c
		5	Interface _____________ reduces a set of	Mapper	Reducer	Writable	Readable	b
		6	The MapReduce algorithm contains two important	mapped, reduce	mapping,	Map,	Map, Reduce	d
		7	Which of the following is used to schedules jobs	SlaveNode	MasterNode	JobTracker	Task Tracker	c
		8	HDFS works in a __________ fashion.	worker-master	master-slave	master-worker	slave-master	b
		9	The default block size in hadoop is ______.	16MB	32MB	64MB	128MB	c
		10	HDFS is implemented in _____________	C	Perl	Python	Java	d
		11	Which of the following is a wide-column store?	Cassandra	Riak	MongoDB	Redis	a
		12	Most NoSQL databases support automatic	processing	scalability	replication	reducing	c
		13	mapper and reducer classes extends classes from	org.apache.hadd	apache.hadoop	org.mapreduce	hadoop.mapred	a
		14	What license is Apache Hadoop distributed under?	Apache License	Shareware	Mozilla Public	Commercial	a
		15	A resource used for sharing data globally by all	Distributed	Centralised Cache	secondry	primary memory	a
		16	...........is the master that which manages the jobs	heart beat	Job tracker	Task Tracker	Job history	b
		17	The MapReduce algorithm contains two important	mapped, reduce	mapping,	Map,	Map, Reduce	d
		18	____ can best be described as a programming	MapReduce	Mahout	Oozie	Hbase	a
Module No	**Hours**	**Q NO**	**QUESTION (2 marks per question)**	**OPTIONS**				**Correct Answer**
				A	**B**	**C**	**D**	
4		1	In Flajolet-Martin algorithm if the stream contains n elements with m of them unique, this algorithm runs in	O(n) time	constant time	O(2n) time	O(3n)time	a
		2	which algorithm we will implement to know how many distinct users visited the website till now or in last 2 hours.	DGIM	SVM	FM	Clustering	c
		3	In FM algorithm we shall use estimate...............for the number of distinct elements seen in the stream.	2 to the power R	3 to the power R	2R	3R	a

	7	4	In sliding window of size w an element arriving at time t expires at	w	t	t+w	t-w	c
		5	Real-time data stream is ________	sequence of data items that arrive in some order and may be seen only once.	sequence of data items that arrive in some order and may be seen twice.	sequence of data items that arrive in same order	sequence of data items that arrive in different order	a
		6	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.	B
		7	Which of the following statements about standard Bloom filters is correct?	It is possible to delete an element from a Bloom filter.	A Bloom filter always returns the correct result.	It is possible to alter the hash functions of a full Bloom filter to create more space.	A Bloom filter always returns TRUE when testing for a previously added element.	d
		8	What are DGIM’s maximum error boundaries?	DGIM always underestimates the true count; at most by 25%	DGIM either underestimates or overestimates the true count; at most by 50%	DGIM always overestimates the count; at most by 50%	DGIM either underestimates or overestimates the true count; at most by 25%	B
		9	Which of the following statements about the standard DGIM algorithm are false?	DGIM operates on a time-based window.	DGIM reduces memory consumption through a clever way of storing counts.	In DGIM, the size of a bucket is always a power of two.	The maximum number of buckets has to be chosen beforehand.	d

		10	In DGIM,whenever forming a bucket then_____	Every bucket should have at least one 1, else no bucket can be formed	Every bucket should have at least two 1, else no bucket can be formed	Every bucket should have at least three 1, else no bucket can be formed	Every bucket should have at least four 1, else no bucket can be formed	A
		11	Which attribute is not indicative for data streaming?	Limited amount of memory	Limited amount of processing time	Limited amount of input data	Limited amount of processing power	C
		12	In Filtering Streams_____________	Accept those tuples in the stream that meet a criterion.	Accept data in the stream that meet a criterion.	Accept those class in the stream that meet a criterion.	Accept rows in the stream that meet a criterion.	a
		13	A Bloom filter consists of_________	An array of n bits, initially all 0’s.	An array of 1 bits, initially all 0’s.	An array of 2 bits, initially all 0’s.	An array of n bits, initially all 1’s.	a
		14	The purpose of the Bloom filter is to allow_____________	through all stream elements whose keys are in Set	through all stream elements whose keys are in class	through all data elements whose keys are in Set	through all touple elements whose keys are in Set	a

Module No	Weightage in Hrs	Q NO	QUESTION (2 marks per question)	OPTIONS				Correct Answer
				A	B	C	D	
5		1	The phenomenon that occurs because of feature changes or changes in behaviour of the data itself is known as	Concept Drift	Streaming	Sampling	Batch Processing	A
		2	Identify the heirarchical clustering type which calculates the average distance between clusters before merging.	Average Link Clustering	Centroid Link Clustering	Single Link Clustering	Complete Link Clustering	A
		3	Which of the following stream clustering algorithm can be used for counting 1's in a stream	FM Algorithm	PCY Algorithm	BDMO Algorithm	SON Algorithm	C

10	4	Which term indicated the degree of corelation in dataset between X and Y, if the given association rule given is X-->Y	Confidence	Monotonicity	Distinct	Hashing	A
	5	Which technique is used to filter unnecessary itemset in PCY algorithm	Association Rule	Hashing Technique	Data Mining	Market basket	B
	6	In association rule, which of the following indicates the measure of how frequently the items occur in a dataset ?	Support	Confidence	Basket	Itemset	A
	7	Identify the property of frequent itemsets which is defined as follows ' If a set of items in a dataset is frequent , then so are all its subsets'	Support	Confidence	Monotonicity	Distinct	C
	8	Identify the algorithm in which, on the first pass we count the item themselves and then determine which items are frequent. On the second pass we count only the pairs of item both of which are found frequent on first pass	DGIM	CURE	Pagerank	Apriori	D
	9	SON algorithm is also known as	PCY Algorithm	Multistage Algorithm	Multihash Algorithm	Partition Algorithm	D
	10	which of the following clustering technique is used by K- Means Algorithm	Hierarchical Technique	Partitional technique	Divisive	Agglomerative	B
	11	Producing clusters in a determined location based on the high density of data set participants is known as	Single Link	Hierarchical Technique	Partitional technique	Density Based Clustering	D
	12	A version of k-means algorithm used to cluster data that is too large to fit in main memory is........................	BFR Algorithm	FM Algorithm	PCY Algorithm	SON Algorithm	A
	13	Which of the following Hierarchichal approach begins with each observation in a distinct (singleton) cluster, and successively merges clusters together until a stopping criterion is satisfied.	Divisive	Agglomerative	Single Link	Complete Link	B
	14	Identify the large scale clustering algorithm which uses a combination of partition based and hierarchical algorithms	FM Algorithm	PCY Algorithm	SON Algorithm	CURE Algorithm	D
	15	Which of theclassification algorithm uses a hyperplane which separates the data into classes.	SVM Classifier	PCY Algorithm	K-Nearest neighbour	BFR Algorithm	A
	16	Which of the algorithm maps the input data to a specific category	Classifier	Multi Label Classification	Multi Class Classification	Feature	A
	17	A classification model that uses a treelike structure to represent multiple decision paths is..........................	PCY Algorithm	SVM Classifier	Decision tree	K-Nearest neighbour	C

		18	The distance between two mean points of a cluster is known as	Density	Average	Centroid	Divisive	C
		19	An individual measurable property of a phenomenon used in classification algorithm that is being observed is known as	Multi Label Classification	Multi Class Classification	Binary Classification	Feature	D
		20	classification of a sample is dependent on the target values of the neighboring points falls under which of the following classification algorithm type	Multi Label Classification	K-Nearest neighbour	PCY Algorithm	SVM Classifier	B

Module	Weighatge in Hrs	Q NO	QUESTION (2 marks per question)	OPTIONS				Correct Answer
				A	B	C	D	
6		1	In a web graph ______ is consider as nodes and edges connecting nodes are ______to the pages	Web page & links	links & Web page	Hubs & Authorities	Authorities & Hubs	A
		2	Page Rank Helps in measuring __________of a web page within a set of similar entities.	Interconnections	relative importance	Incomming Links	Outgoging Links	B
		3	In page Rank compuation in a web a Dead Ends are the pages with no ______ in the web graph.	Trust Rank	In links	out links	Hub Score	C
		4	In Structure of web some pages that reach from in-components to the out-componets withtout linking it to any pages in SCC(Strongly connected Componets), are called as	Dead Ends	Hubs	Spider Traps	Tubes	D
		5	------------- are theset of pages whose outlinks reach to the pages only from that set	Dead Ends	Hubs	Spider Traps	Tubes	C
		6	One lagre portion of web which is more or less strongly connected Componet also called as ______	Tubes	Core	Tendrils	InComponets	B
		7	Technique used to attempts to measure what fraction of PageRank value could be due to spam is called as	Spam Mass	Trust Rank	Page Rank	Hub Score	A
		8	In PageRank computation highest eigen value of a Markov matrix is	0	1	-1	2	B
		9	Which statement is true about Page Rank	PageRank is Query Dependent	PageRank is Query Independent , works on lagre portion of web	PageRank is Query Dependent , works on small portion of web	PageRank works on small portion of web	C
		10	Which Statement is true about HITS algorithm	HITS work on entire Web graph	HITS work on small subgraph from the web garph	HITS assign pageRank to webpages	It use idea of random surfer	B

11	11	Which of the following factors have an impact on the Google PageRank?	The total number of inbound links to a page of a web site	The subject matter of the site providing the inbound link to a page of a web site	The text used to describe the inbound link to a page of a web site	The number of outbound links on the page that contains the inbound link to a page of a web site	A
	12	An alogrithm which visits each node X once and computes the number of shortest paths from X to each of the other nodes that go through each of the edges is:	DGIM Algorithm	Girvan-Newman Algorithm	Page Rank Algorithm	FM Algorithm	B
	13	……………………allows us to discover groups of interacting objects and relationship between them	Node	Community	Map reduce	Combiners	B
	14	The process of identifying similar users and recommending what similar users like is called	collaborative filtering	Content-Based systems	Page rank	stream filtering	A
	15	The concept which explains the advantage of on-line vendors over conventional, brick-and mortar vendors is called	Short tail	Tailing	Long-tail	ZeroTail	C
	16	For an edge e in a graph, edge betweeness of e is defined as the number of ………………path between all nodes paira(Vi,Vj)in the graph such that the shortest path between Vi and Vj passes through e	shortest	farthest	equal	zero length	A
	17	Girwan and Newman proposed a hierachical divisive clustering technique for social graphs that use the:	Edge Betweeness as a distance measure	Centrality as a distance measure	Jaccard distance as a distance measure	Euclidean distance as a distance measure	A
	18	A measure that says "two objects are considered to be similar if they are refrenced by similar objects" is:	Page Rank	Trust Rank	Graph Rank	Sim Rank	D
	19	A and B have an intersection of size 1 and a union of size 5. then their Jaccard distance is	5	43835	43926	1	C
	20	We can enumerate or count the triangles in a graph with m edges in	O(m to the power 3/2)time	O(m cube)time	O(m)time	O(m square)time	A
	21	finding maximal cliques is a	not a NP-complete problem	NP-complete problem	easy task	moderate problem	B
	22	The number of triangles per node in a social network graph is an important measure of the ……………………………… of a community	page rank	authority	TrustRank	closeness	D