UNIVERSITY OF MUMBAI

SAMPLE MCQ QUESTION BANK

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

Q NO	QUESTION (2 marks per question)	OPTIONS			
		A	B	C	D
1	What are the 3v's of Big Data?	volume, velocity and value	Volume, Velocity and Variety	volume, varacity and variety	Variety, Variability and Volume
2	Apache Hadoop is an _____ for storage and large-scale processing of data-sets on clusters of commodity hardware.	passive-source software framework	active-source software framework	closed-source software framework	open-source software framework
3	The composition of the data with the	sensitivity of data	availability of data	structure of data	state of data
4	Human generated data _____	Financial data	Network log	Input data	Gaming data
5	What is true about Variety in bigdata?	high in size	speed of data	data from many sources	data in certain format
6	which is structured data examples	CSV but XML	social media posts	tab delimited files	Medical device data
7	Which type of data storage system cassandra is used?	distributed	centralized	parallel	dumb
8	The default replication factor in Hadoop is	4	3	5	2
9	What was Hadoop named after?	Creator Doug Cutting favorite circus act	Cutting high school rock band	The toy elephant of Cutting son	A sound Cutting laptop made during Hadoop development
10	NoSQL CAP theorem ----------	Consistency, Availability and Performance	Consistency, Acceptibility and Parttition Tolerance	Consistency, Availability and Parttition Tolerance	confidentiality , Availability and Parttition Tolerance
11	MongoDB provides horizontal scaling through___	Replication	Partitioning	Sharding	Document
12	Point out the correct statement.	DataNode resposible for replication	DataNode is the slave/worker node and holds the user data in the form of Data Blocks	DataNode is also called as YARN	Hadoop designed in Python
13	Which of the following is a wide-column store?	Cassandra	Riak	MongoDB	Redis
14	Procedural language for developing parallel processing applications for large data sets in the Hadoop environment is	Pig Latin	Hive	Pig	Oozie
15	Cassendra is a popularly known as _____ data storage system.	distributed	centralized	parallel	dumb
16	When a backup node is used in a cluster there is no need of _____	Check point node	Secondary data node	Secondary name node	Rack awareness
17	Hadoop developed by _____	Larry Page	Doug Cutting	Mark Zuckerberg	Bill Gates
18	No SQL systems are also referred to as _____ to emphasize that they do in fact allow SQL-like query languages to be used.	"Not-On-SQL“	"N-Only-SQL“	"No-only-SQL“	"Not-Only-SQL“
19	One of classified NoSQL databases is _____	Key-value	value	key	DataNode

UNIVERSITY OF MUMBAI
SAMPLE MCQ QUESTION BANK

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

Q NO	QUESTION (2 marks per question)	OPTIONS			
		A	B	C	D
20	A _____ store is a simple database that when presented with a simple string (the key) returns an arbitrary large BLOB of data (the value).	document	key-value	graph	simple
21	Social connectionsstores are used to store in _____	Stack	Tree	Graph	Documents
22	MongoDB scales horizontally using Sharding for _____ purpose.	data balancing	load distribution	memory balancing	load balancing
23	HDFS inherited from ---------- file system.	Yahoo	FTFS	Google	Rediff
24	Which is the column-oriented distributed database.	HBase	NOSQL_IBM	MSSQL	MySQL
25	Cost factor in Hadoop cluster setup.	inexpensive commodity hardware	only i7 machine	graphics card required	cloud required
26	MongoDB database has been used by number of software,major websites and services as _____	backend	proprietary	GUI design	front end
27	_____ node acts as the Slave and is responsible for executing a Task assigned to it by the JobTracker.	MapReduce	Mapper	TaskTracker	JobTracker
28	Which function is responsible for consolidating the results produced by each of the Map() functions/tasks.	Reduce	Map	Reducer	Mapper
29	Which function maps input key/value pairs to a set of intermediate key/value pairs.	Mapper	Reducer	Combiner	Execute
30	_____ is the slave/worker node and holds the user data in the form of Data Blocks.	Data block	NameNode	DataNode	Replication
31	Interface _____ reduces a set of intermediate values which share a key to a smaller set of values.	Mapper	Reducer	Writable	Readable
32	The MapReduce algorithm contains two important tasks, namely _____.	mapped, reduce	mapping, Reduction	Map, Reduction	Map, Reduce
33	Which of the following is used to schedules jobs and tracks the assign jobs to Task tracker?	SlaveNode	MasterNode	JobTracker	Task Tracker
34	HDFS works in a _____ fashion.	worker-master fashion	master-slave fashion	master-worker fashion	slave-master fashion
35	The default block size in hadoop is _____.	16MB	32MB	64MB	128MB
36	HDFS is implemented in _____ language.	C	Perl	Python	Java
37	Which of the following is a wide-column store?	Cassandra	Riak	MongoDB	Redis
38	Most NoSQL databases support automatic _____ meaning that you get high availability and disaster recovery.	processing	scalability	replication	reducing
39	mapper and reducer classes extends classes from the package	org.apache. haddop. mapreduce	apache.hadoop	org. mapreduce	hadoop. mapreduce
40	What license is Apache Hadoop distributed under?	Apache License 2.0	Shareware	Mozilla Public License	Commercial
41	A resource used for sharing data globally by all nodes is	Distributed Cache	Centralised Cache	secondry memory	primary memory
42	.....is the master that which manages the jobs and res ources in a cluster	heart beat	Job tracker	Task Tracker	Job history server
43	The MapReduce algorithm contains two important tasks, namely _____.	mapped, reduce	mapping, Reduction	Map, Reduction	Map, Reduce

UNIVERSITY OF MUMBAI
SAMPLE MCQ QUESTION BANK

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

Q NO	QUESTION (2 marks per question)	OPTIONS			
		A	B	C	D
44	____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.	MapReduce	Mahout	Oozie	Hbase
45	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
46	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
47	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
48	In sliding window of size w an element arriving at time t expires at	w	t	t+w	t-w
49	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
50	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.
51	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.
52	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%
53	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.
54	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

UNIVERSITY OF MUMBAI
SAMPLE MCQ QUESTION BANK

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

Q NO	QUESTION (2 marks per question)	OPTIONS			
		A	B	C	D
55	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
56	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.
57	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.
58	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
59	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
60	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
61	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
62	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
63	Which technique is used to filter unnecessary itemset in PCY algorithm	Association Rule	Hashing Technique	Data Mining	Market basket
64	In association rule, which of the following indicates the measure of how frequently the items occur in a dataset ?	Support	Confidence	Basket	Itemset
65	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
66	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
67	SON algorithm is also known as	PCY Algorithm	Multistage Algorithm	Multihash Algorithm	Partition Algorithm
68	which of the following clustering technique is used by K- Means Algorithm	Hierarchical Technique	Partitional technique	Divisive	Agglomerative
69	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

UNIVERSITY OF MUMBAI
SAMPLE MCQ QUESTION BANK

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

Q NO	QUESTION (2 marks per question)	OPTIONS			
		A	B	C	D
70	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
71	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
72	Identify the large scale clustering algorithm which uses a combination of partition based and hierarchical algorithms	FM Algorithm	PCY Algorithm	SON Algorithm	CURE Algorithm
73	Which of theclassification algorithm uses a hyperplane which separates the data into classes.	SVM Classifier	PCY Algorithm	K-Nearest neighbour	BFR Algorithm
74	Which of the algorithm maps the input data to a specific category	Classifier	Multi Label Classification	Multi Class Classificatio n	Feature
75	A classification model that uses a treelike structure to represent multiple decision paths is.....	PCY Algorithm	SVM Classifier	Decision tree	K-Nearest neighbour
76	The distance between two mean points of a cluster is known as	Density	Average	Centroid	Divisive
77	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 Classificatio n	Feature
78	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
79	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
80	Page Rank Helps in measuring _____of a web page within a set of similar entities.	Interconnectio ns	relative importance	Incomming Links	Outgoging Links
81	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
82	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
83	---------- are theset of pages whose outlinks reach to the pages only from that set	Dead Ends	Hubs	Spider Traps	Tubes
84	One lagre portion of web which is more or less strongly connected Componet also called as _____	Tubes	Core	Tendrils	InComponets
85	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
86	In PageRank computation highest eigen value of a Markov matrix is	0	1	-1	2
87	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

UNIVERSITY OF MUMBAI

SAMPLE MCQ QUESTION BANK

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

Q NO	QUESTION (2 marks per question)	OPTIONS			
		A	B	C	D
88	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
89	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
90	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
91	….....allows us to discover groups of interacting objects and relationship between them	Node	Community	Map reduce	Combiners
92	The process of identifying similar users and recommending what similar users like is called	collaborative filtering	Content-Based systems	Page rank	stream filtering
93	The concept which explains the advantage of on-line vendors over conventional, brick-and mortar vendors is called	Short tail	Tailing	Long-tail	ZeroTail
94	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
95	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
96	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
97	A and B have an intersection of size 1 and a union of size 5. then their Jaccard distance is	5	43835	43926	1
98	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
99	finding maximal cliques is a	not a NP-complete problem	NP-complete problem	easy task	moderate problem
100	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