Q=QUESTIO	question_description	question_explanation	question_type	question_difficulty
A=ANSWER	answer_description	answer_explanation	answer_isright	answer_position
Q	The unit of data that flows through a Flume agent is		M	1
A	Log		0	1
A	Row		0	2
A	Event		1	3
A	Record		0	4
Q	Refers to how accurate and correct the data is for its intended to use		M	1
A	Varacity		0	1
A	Validity		1	2
A	Varience		0	3
A	Value		0	4
Q	Which of the following is not an input format in Hadoop ?		M	1
A	TextInputFormat		0	1
A	ByteInputFormat		1	2
A	SequenceFileInputformat		0	3
A	KepInputFormat		0	4
Q	A Combiner, also known as		M	1
A	a semi-reducer		1	1
A	a semi-mapper		0	2
A	a full reducer		0	3
A	Hive		0	4
Q	Mapper and Reducer classes, the user can specify type information during the class declarant Id t		M	1
A	< VALUEIN, KEYOUT,VALUEOUT>		0	1
A	<KEYIN, VALUEIN, KEYOUT>		0	2
A	<KEYIN, VALUEIN>		0	3
A	<KEYIN, VALUEIN, KEYOUT,VALUEOUT>		1	4
Q	Grouping and aggregation can be performed in ……….. MapReduce job.		M	1
A	1		1	1
A	2		0	2
A	4		0	3
A	3		0	4
Q	What is getJobState()		M	1

A	Checks if the job is finished or not.		0	1
A	User-specified job name.		0	2
A	Returns the current state of the Job		1	3
A	Sets the Mapper for the job		0	4
Q	The partition phase takes place		M	1
A	after the Map phase and before the Reduce phase.		1	1
A	Before the Map phase and After the Reduce phase.		0	2
A	After Combiner		0	3
A	Before Shuffle		0	4
Q	Which is not a Distance Measure		M	1
A	Levenshtein distance		0	1
A	Cosine Distance		0	2
A	Jaccard Distance		0	3
A	Read Distance		1	4
Q	Select the false statement		M	1
A	Edit Distance is a measure of the similarity between two strings.		0	1
A	The edit distance between s and t is the number of deletions required to transform s into t.		1	2
A	Jaccard distance is only applicable to set data.		0	3
A	Edit distance(xs, ys) is defined as len(xs) + len(ys) - 2 * lcs(xs, ys)		0	4
Q	Algorithm to estimate number of distinct elements seen in the stream.		M	1
A	FM Algorithm		1	1
A	DGIM algorithm		0	2
A	HITS Algorithm		0	3
A	Bloom Filter		0	4
Q	In Bloom Filter		M	1
A	fale positive are not possible ,but false negative is		0	1
A	false positives matches and false negative both are not possible		0	2
A	false positives matches are possible,but false negative is not		1	3
A	always true		0	4
Q	A page is a good hub page with respect to a given query		M	1
A	if it points to many good hub page with respect to the query		0	1
A	if it points to many good authoritive page with respect to the query		1	2
A	it points to large number of pages		0	3
A	it points to small numnber of pages		0	4

Q	pages that provide information about a topic are called		M	1
A	authorities		1	1
A	hubs		0	2
A	page rank		0	3
A	tendril		0	4
Q	Technique to handle dead ends is:		M	1
A	Remove all pages with no out outgoing links and remove their in links too		1	1
A	Remove all pages with no out outgoing links and dont remove their in links		0	2
A	Avoid it		0	3
A	dead end can't be handled		0	4
Q	Identify the property of frequent itemsets which is defined as follows ' If a set of items is frequen		M	1
A	Support		0	1
A	Confidence		0	2
A	Monotonicity		1	3
A	Distinct		0	4
Q	Identify the algorithm where in first pass the entire file of baskets is divided into small segments,		M	1
A	SON Algorithm		1	1
A	Pagerank Algorithm		0	2
A	PCY Algorithm		0	3
A	Blooms Filter		0	4
Q	Identify the algorithm, which is an extension of apriori algorithm where a hash table is created on		M	1
A	DGIM		0	1
A	Pagerank Algorithm		0	2
A	PCY Algorithm		1	3
A	Blooms Filter		0	4
Q	Hierarchical clustering type where distance between two clusters are taken as the shortest distanc		M	1
A	Centroid Link Clustering		0	1
A	Single Link Clustering		1	2
A	Complete Link Clustering		0	3
A	Average Link Clustering		0	4
Q	Identify the large scale clustering algorithm which uses a combination of partition based and hier		M	1
A	FM Algorithm		0	1
A	PCY Algorithm		0	2
A	SON Algorithm		0	3

A	CURE Algorithm		1	4
Q	Which of the following algorithm is most sensitive to outliers?		M	1
A	K-means clustering algorithm		1	1
A	K-medians clustering algorithm		0	2
A	K-modes clustering algorithm		0	3
A	K-medoids clustering algorithm		0	4
Q	Point out the wrong statement.		M	1
A	k-means clustering is a method of vector quantization		0	1
A	k-means clustering aims to partition n observations into k clusters		0	2
A	k-nearest neighbor is same as k-means		1	3
A	k-means clustering is a method of quantization		0	4
Q	Hadoop is a framework that works with a variety of related tools. Common cohorts include _____		M	1
A	MapReduce, Hive and HBase		1	1
A	MapReduce, MySQL and Google Apps		0	2
A	MapReduce, Hummer and Iguana		0	3
A	MapReduce, Heron and Trumpet		0	4
Q	What was Hadoop named after?		M	1
A	Creator Doug Cutting’s favorite circus act		0	1
A	Cutting’s high school rock band		0	2
A	The toy elephant of Cutting’s son		1	3
A	A sound Cutting’s laptop made during Hadoop development		0	4
Q	What makes Big Data analysis difficult to optimize?		M	1
A	Big Data is not difficult to optimize		0	1
A	Both data and cost effective ways to mine data to make business sense out of it		1	2
A	The technology to mine data		0	3
A	cost effective		0	4