University of Mumbai
Examination 2020

Examinations Commencing from 23rd December 2020 to 6th January 2021 and from 7th January 2021 to 20th January 2021

Program: **Computer Engineering**

Curriculum Scheme: Rev 2016

Examination: BE Semester VII

Course Code: CSDLO7032 and Course Name: Big Data Analytics

Time: 2 hour Max. Marks: 80

==========

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	Real Time Analysis Processing can be achieved by _____.
Option A:	DBMSs
Option B:	RDBMS
Option C:	Big Data Architecture & Technology
Option D:	OLAP data warehousing
2.	In Hadoop ecosystem projects, _____ provides a method to import data from tables in relational database into HDFS
Option A:	Hue
Option B:	Oozie
Option C:	Sqoop
Option D:	Mahout
3.	_____ is responsible for storing data on the cluster.
Option A:	Mapreduce
Option B:	HDFS
Option C:	Job Tracker
Option D:	Task Tracker
4.	HDFS uses _____ architecture
Option A:	Pipeline
Option B:	Master / Slave
Option C:	Peer to Peer
Option D:	tree
5.	σ C(R) in Relation Algebra is used to denote _____ Operation in MapReduce.
Option A:	Natural Join
Option B:	Selection
Option C:	Projection
Option D:	Aggregation

6.	In MapReduce, ---------Specifies how to combine the maps for local aggregation
Option A:	Combiner class
Option B:	Mapper Class
Option C:	Reducer Class
Option D:	Shuffle Class
7.	CAP properties of NOSQL are not available in the form
Option A:	CP tolerant
Option B:	AP tolerant
Option C:	CA tolerant
Option D:	PA tolerant
8.	Amazon DynamoDB is example of _____ store.
Option A:	Column family
Option B:	Key-value
Option C:	Document
Option D:	Graph
9.	The Jaccard similarity (Jaccard coefficient) of two sets S1, S2 is the size of their ----------divided by the size of their union.
Option A:	Union all
Option B:	Intersection
Option C:	Difference
Option D:	Minus
10.	In Cosine Similarity If the vectors are aligned (correlated) angle is ------ degrees and cos(X,Y)=1
Option A:	90
Option B:	180
Option C:	0
Option D:	360
11.	Find Hamming distance between two vectors of categorical attributes x =(married, low income, cheat), y = (single, low income, not cheat)
Option A:	3
Option B:	2
Option C:	1
Option D:	3
12.	--------is the average of all (data) points in the cluster also called as “artificial” point.
Option A:	Data Point
Option B:	Clustroid

Option C:	Outlier
Option D:	Centroid
13.	-------- is Isolated points waiting to be assigned to a compression set
Option A:	Extended Compression set (CS):
Option B:	Discard set (DS):
Option C:	Retained set (RS):
Option D:	Outlier set:
14.	Select the Category for recommendation of Top 10, Most Popular videos on YouTube
Option A:	Editorial and hand curated
Option B:	Simple aggregates
Option C:	Tailored to individual users
Option D:	False suggession
15.	Extrapolate unknown ratings from by learning ratings from user actions is known as ----way
Option A:	Implicit
Option B:	Explicit
Option C:	Active
Option D:	Passive
16.	In --------- system Cannot predict ratings for new item till some similar users have rated it
Option A:	Item Cold-Start problem
Option B:	Scalability
Option C:	Sparsity
Option D:	User Cold-Start problem
17.	Following algorithm is used for Counting Distinct Elements in a Stream
Option A:	Flajolet-Martin Algorithm
Option B:	SON Algorithm
Option C:	CURE Algorithm
Option D:	DGIM Algorithm
18.	Which of the following is not the block of data stream management system
Option A:	Stream Processor
Option B:	working Storage
Option C:	Archival Storage
Option D:	Query Processor
19.	A Bloom filter guarantees no ------
Option A:	false positives
Option B:	false positives and false negatives

Option C:	false negatives
Option D:	false positives or false negatives, depending on the Bloom filter type
20.	The Girvan-Newman (GN) algorithm begins by performing a ------- of the graph, starting at the node X.
Option A:	advanced-first search(AFS)
Option B:	breadth-first search(BFS)
Option C:	depth-first search(DFS)
Option D:	random-first search(RFS)

Q2. **(20 Marks Each)**	**Solve any Four out of Six** **5 marks each**
A	Give difference between Traditional data management and analytics approach Versus Big data Approach
B	Write Map Reduce pseudo code for “Group By” “Aggregation” in database
C	Explain Jaccard Similarity with example
D	Describe Characteristics of NoSQL database
E	List down the steps in Page Rank Algorithm.
F	Explain Data stream management Architecture?

Q3. **(20 Marks Each)**	**Solve any Two Questions out of Three** **10 marks each**
A	What is Hadoop? Describe HDFS architecture with diagram
B	Explain Girvan-Newman algorithm to mine Social Graphs.
C	What are the different data architecture pattern in NoSQL . Explain graph store and column family store patterns with relevant examples