University of Mumbai

Program: **Computer Engineering**
Curriculum Scheme: 2016
Examination: BE Semester VII
Course Code: CSDLO7032 and Course Name: BIG DATA ANALYTICS
Time: 2.30 hour Max. Marks: 80

==========

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	Stream management is important when the input rate is controlled --------
Option A:	Internally
Option B:	constant
Option C:	Static
Option D:	externally
2.	What are DGIM's maximum error boundaries?
Option A:	DGIM always underestimates the true count; at most by 25%
Option B:	DGIM either underestimates or overestimates the true count; at most by 50%
Option C:	DGIM always overestimates the count; at most by 50%
Option D:	DGIM either underestimates or overestimates the true count; at most by 25%
3.	Which software tool allows real time data processing in big data?
Option A:	Hive
Option B:	Sqoop
Option C:	Flume
Option D:	PIG
4.	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
5.	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
6.	There is a need for storing transactional data generated by a Bank's ATM. The data is to be stored in a tabular format. According to CAP theorem, which type of datastore is to be used for this?
Option A:	CP
Option B:	AP
Option C:	CA

Option D:	CAP
7.	_____ system recommend items based on similarity measures between users and/or items.
Option A:	Content-based filtering
Option B:	General filtering
Option C:	Collaborative Filtering
Option D:	User-based filtering
8.	A Bloom filter guarantees no
Option A:	False negatives
Option B:	False positives
Option C:	False positives and false negatives
Option D:	False positives or false negatives, depending on the Bloom filter type
9.	Which of the following term can be used to describe nodes that contain the maximum amount of information about a network?
Option A:	Social Networks
Option B:	Degree Centrality
Option C:	BetweenessCentrality
Option D:	Broadcasters
10.	Flajolet-Martin(FM) algorithm is used to_____
Option A:	Count distinct elements in the stream
Option B:	Count frequent items in the stream
Option C:	Count ones in the streams
Option D:	Check item in the stream

Q2 (20 Marks)	**Solve any Four out of Six** **5 marks each**
A	Explain different phases of Map Reduce with word common example?
B	What are the NOSQL business drivers?
C	Explain Datar Gionis Indynk Motwani (DGIM) algorithm.
D	Give difference between Traditional data management and analytics approach Versus Big data Approach
E	Why Cosine Distance is a Distance Measure? Find the Cosine Similarity between two documents DOC_1: ABC cares me more than XYZ cares me DOC_2: RMM helps me more than ABC cares me

Q3	**Solve any Two Questions out of Three** **10 marks each**

(20 Marks)	
A	Write a short note on Bloom Filter
B	What is a recommendation system? Explain the collaborative and content based filtering.
C	What is a community in a Social Network Graph? Explain how the Girvan Newman algorithm finds the different Communities in the graph

Q4. **(20 Marks)**	**Solve any Two Questions out of Three** **10 marks each**
A	Explain Hadoop ecosystem with core components. Explain physical architecture of hadoop. State its limitations
B	Show map reduce implementation with the help of pseudo code i. algorithm for Natural Join of two relations ii. algorithm to perform intersection of two sets
C	What is clustering? What are applications of cluster analysis? Explain BFR algorithm