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

==========

Question Number	Correct Option (Enter either 'A' or 'B' or 'C' or 'D')
Q1.	D
Q2.	B
Q3.	C
Q4	C
Q5	A
Q6	C
Q7	C
Q8.	A
Q9.	B
Q10.	A

Q 2

A. Explain different phases of Map Reduce with word common example?

Ans: Explanation of Map Phase, Reduce Phase, Shuffle Phase, Sort Phase and Combiner/Partition with wordcount example ……..5 marks

B. What are the NOSQL business drivers?

Ans: Explanation business driver volume, variety, variability, agility and diagram ….5marks

C. Explain Datar Gionis Indynk Motwani (DGIM) algorithm.
Ans: …explanation of DGIM algorithm for counting number of 1's in stream, explain different cases of it. ……..5marks

D. Give difference between Traditional data management and analytics approach Versus Big data Approach

Comparing Traditional & Big Data

Feature	Traditional Data	Big Data
Structure	Lake / Pool	Flowing Stream / river
Primary Purpose	Manage business activities	Communicate, Monitor, Insights
Source of data	Business transactions, documents	Social media, Web logs, Machine Generated
Volume of data	Gigabytes, Terabytes	Petabytes, Exabytes
Velocity of data	Ingest level is controlled	Real-time unpredictable ingest
Variety of data	Alphanumeric	Audio, Video, Graphs, Text
Veracity of data	Clean, more trustworthy	Varies depending on source
Structure of data	Well-Structured	Semi- or Un-structured
Physical Storage	In a Storage Area Network	Clusters of commodity computers (in cloud)
Data organization	Relational databases	NoSQL databases
Acess Languages	SQL	NoSQL such as Pig
Data Manipulation	Conventional data processing	Parallel processing
Data Visualization	Variety of tools	Dynamic dashboards with simple measures
Database Tools	Commercial systems	Open Source – Apache Hadoop, Spark, etc
[illegible]	Medium to High	High

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

* Cosine distance is a distance measure.

1. $d(x, y) \geq 0$
2. $d(x, x) = 0$
3. $d(x, y) = d(y, x)$
4. $d(x, y) \leq d(x, z) + d(z, y)$ ..... 2 marks

DOC-1	ABC	cares	me	more	than	XYZ
Term Freq.	1	2	2	1	1	1

DOC-2	RMM	helps	me	more	than	ABC	cares
Term Freq	1	1	2	1	1	1	1

$D_1 = [1, 2, 2, 1, 1, 1, 0]$

$D_2 = [1, 1, 2, 1, 1, 1, 1]$

$D_1 \cdot D_2 = 1 \cdot 1 + 2 \cdot 1 + 2 \cdot 2 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 0 \cdot 1$

$= 10$

$\|D_1\| = \sqrt{1^2 + 2^2 + 2^2 + 1^2 + 1^2 + 1^2 + 0^2} = \sqrt{12} = 3.464$

$\|D_2\| = \sqrt{1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{10} = 3.162$

$\cos(D_1 \cdot D_2) = \frac{D_1 \cdot D_2}{\|D_1\| \cdot \|D_2\|} = \frac{10}{3.464 \times 3.162} = 0.91$

$\cos\theta = 0.91$ OR

$\theta = 24.09^\circ$

Q3

A. Write a short note on Bloom Filter.

Ans: Expalnation about bloom filter with steps 5marks

Analysis of bloom filter ……5mark

B. What is a recommendation system? Explain the collaborative and content based filtering.
Ans: Recommendation system definition ------- 2 marks
Explanation of Collaborative filtering--------4 marks
Explanation of Content Based Filtering--------4 marks

C. What is a community in a Social Network Graph? Explain how the Girvan Newman algorithm finds the different Communities in the graph.

Ans: Community definition …….. 3 marks

Explanation of Girvan Newman algorithm------ 3 marks

Algorithms steps and explanation ……..4 marks

Q 4.

A. Explain Hadoop ecosystem with core components. Explain physical architecture of hadoop. State its limitations

Ans- Draw and explain diagram of Hadoop ecosystem- 4 marks, drawand explain architecture of hadoop- 4 marks, limitation of hadoop-2 marks

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

Ans- Explain the algorithm with pseudo code for natural join-----5 marks and pseudo code for intersection of two sets—5 marks

C. What is clustering? What are applications of cluster analysis? Explain BFR algorithm.
Ans- *Explanation of clustering- 2 Marks*
Applications-2 marks
BFR algorithm with diagram, advantages, disadvantages-6 marks