## Group – E

8. (a) What is 'Map-Reduce' (MR)? Which part of MR is responsible for processing one or more chunks of data and producing the output results? Is it essential that MR applications must be written in Java?

   (b) Mention two advantages and disadvantages each of using MR.

   (c) Provide the outline and algorithm of an MR-based solution to the following problem:
   A Web-site user log contains user ids and lengths of each session (in time unit) by the user. Compute the average session length for each user, given the following key-value pairs.
   Input:
   - key : user_id
   - value : session_length

   Output:
   - key : user_id
   - value : average_session_length

**(1 + 1 + 1) + 4 + 5 = 12**

9. Suppose we have an n×n matrix M, whose element in row i and column j will be denoted $m_{ij}$. Suppose we also have a vector v of length n, whose jth element is $v_j$. In the said context answer the following:-
   i. What would be the matrix vector product?
   ii. What would be the Map function?
   iii. What would be the Reduce function?

**3 + 4.5 + 4.5 = 12**





# WEB INTELLIGENCE AND BIG DATA
# (CSEN 4165)

**Time Allotted : 3 hrs** **Full Marks : 70**

***Figures out of the right margin indicate full marks.***

***Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.***

***Candidates are required to give answer in their own words as far as practicable.***

## Group – A
## (Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

   (i) The HDFS client software implements _____ checking on the contents of HDFS files.
   (a) checksum (b) even parity
   (c) odd parity (d) all of the above mentioned.

   (ii) Which of the following is finally produced by hierarchical clustering ?
   (a) final estimate of cluster centroids
   (b) tree showing how close things are to each other
   (c) assignment of each point to clusters
   (d) all of the above mentioned.

   (iii) Which of the following platforms does Hadoop run on?
   (a) Windows (b) Debian
   (c) Cross-platform (d) Linux.

   (iv) The HDFS file system is temporarily unavailable whenever the _____ is down.
   (a) ActionNode (b) DataNode
   (c) MapNode (d) NameNode.

   (v) Which type of clustering method does the agglomerative clustering fall under?
   (a) Hierarchical (b) Partitioned
   (c) Probabilistic (d) Heuristic.

   (vi) Work out the approximate retrieval time for a 500-TB dataset distributed across a 2000-node cluster, assuming an average data transfer rate of 1 Gbps.
   (a) 67 minutes (b) 34 minutes
   (c) 17 minutes (d) 7 minutes.

(vii) Which of the following combination is incorrect ?
(a) Continuous – euclidean distance
(b) Continuous – correlation similarity
(c) Binary – manhattan distance
(d) None of the above mentioned.

(viii) Point out the wrong statement:
(a) k-means clustering is a method of vector quantization
(b) k-means clustering aims to partition n observations into k clusters
(c) k-nearest neighbor is same as k-means
(d) None of the above mentioned.

(ix) _____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.
(a) Map-Reduce (b) Mahout
(c) Hive (d) HDFS.

(x) The daemons associated with the Map-Reduce phase are _____ and task-tracker.
(a) map-tracker (b) job-tracker
(b) reduce-tracker (d) shuffle-tracker.

## Group – B

2. (a) Draw the architecture schematic for a typical 'Collective Intelligence' (CI) enabled *event-driven* Web-based application with *asynchronous services*. What changes will be needed to make it a *non-event-driven* one?

(b) Is a typical Google Search an example of a 'Synchronous Service' or an 'Asynchronous Service' in the context of CI? Explain in brief.

(c) Mention <u>one</u> real-life example <u>each</u> of Web-sites exploiting 'Explicit Intelligence', 'Implicit Intelligence', and 'Derived Intelligence'.

(d) Mention <u>one</u> real-life example <u>each</u> for the following types of user profile metadata attributes – Numeric, Nominal Ordinal, and Nominal Categorical.

**(3 + 1) + 2 + 3 + 3 = 12**

3. (a) What is 'Tagging' and how do tags help? Name the different types of tags, with <u>one</u> suitable example for <u>each</u>. What is a 'Tag Cloud'?

(b) Draw a data persistence model for tagging done by users for an online news portal (like say *yahoo.com*).





(c) Farmer-A has tagged Article-1 for tags {apple, banana, fruit}, and has tagged Article-2 for tags {fruit, mango, orange}; Farmer-B has tagged Article-3 for tags {cherry, fruit, orange}. Derive the vocabulary and use that to create a blank tabular format for collection of raw data about tagging these fruit-related articles by different farmers.

**(1 + 3 + 1) + 4 + 3 = 12**

## Group – C

4. (a) What is a 'Recommendation Engine' (RE)? Name the <u>two</u> basic types of RE.

(b) What is 'Jaccard Similarity Index'? Explain in brief.

(c) Prove that the cosine similarity between the two given points (6, 10, 4) and (4, 9, 6) is 0.97.

(d) What is 'Minkowski Distance'? Provide an example when the 'Triangle Inequality' property is violated for distance between three points on a two-dimensional plane.

**(1 + 2) + 3 + 2 + (2 + 2) = 12**

5. (a) Compare and contrast 'Classification' and 'Clustering'.

(b) "Classification is a two-step process" – explain in brief.

(c) Provide the outline of the basic algorithm for classification, and the conditions for stopping partitioning.

**3 + 4 + 5 = 12**

## Group - D

6. (a) Discuss in detail about Hadoop Distributed File System.

(b) What is set and map and explain their operation.

(c) Explain streaming mechanism with Hadoop.

**4 + 4 + 4 = 12**

7. (a) What are the <u>three</u> modes in which Hadoop can be run? Explain briefly the utility of these modes and when to use them.

(b ) Explain, in brief, what the following are: HBase, Hive, and Pig.

**(3 + 3) + (2 + 2 + 2) = 12**