**University of Mumbai**

**Examination 2020 under cluster 5 (APSIT)**

Program: BE Information Technology

Curriculum Scheme: Revised 2016

Examination: Final Year Semester VIII

Course Code: BEITC801 and Course Name: Big Data Analytics

Sample University Multiple Choice MCQ ANS. KEY

 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 **Chapter 1: Introduction to Big Data**

1. Speed of storing and processing data represented as\_\_\_\_\_\_\_\_

a) Velocity

2. Big data analysis does the following except\_\_\_\_\_\_\_\_\_\_

c) Spread data

3. There is \_\_\_\_\_\_ to how much data needs to be stored and for how long in Hadoop.

a) no limit

4. DFS State for

b) Distributed file system

5. Which is NOT an Application of Big data?

 d) RDBMS

6. Which is NOT a feature of Big Data Analytics?

d ) data redundancy

7. Apache Spark is capable of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

b) Stream processing

8. Traditional data management\_\_\_\_\_\_\_\_.

b) Structural database

9. Which of the following Batch Processing instance is NOT an example of \_\_\_\_\_\_\_\_\_\_\_Big Data Batch Processing?

d)Trending topic analysis of tweets for last 15 minutes

|  |
| --- |
|  |
|  |
|  |
|   |
|  |

**Chapter 2: Introduction to Big Data Frameworks: Hadoop, NoSQL**

|  |
| --- |
| 1. Hadoop does not provide \_\_\_\_\_\_\_\_\_ at the storage and network level. |
| a) encryption |

2. \_\_\_\_\_\_ is a framework for performing Machine Learning related task.
c) Mahout

3. You can run Pig in interactive mode using the \_\_\_\_\_\_ shell.

a) Grunt

4. In Hadoop, the optimal input split size is the same as the\_\_\_\_\_\_\_\_

a) block size

5. When you increase the number of files stored in HDFS, The memory required by namenode

a) Increases

6. Which demon is responsible for replication of data in Hadoop?

d) Name Node.

7. \_\_\_\_\_\_\_\_\_\_is a Graph Based NoSql Database.

c) Neo4j

8. A document database in NoSQL is a type of database that is designed to store and query data in \_\_\_\_\_\_ format.

b) JSON

9. Graph is used to store \_\_\_\_\_\_ in NoSQL

c)Data structures

10. Column family is Store

b) Matrix

11. Document Stored in NoSQL is of\_\_\_\_\_\_\_\_.

b) Tree Structure

12. In NoSQL Data Architected Pattern is NOT represent \_\_\_\_\_\_\_.

d) RDBMS

13. Hadoop is not fit for\_\_\_\_\_\_\_\_.

1. Small data

14. What type of storage used in MongoDB.

a) Document Oriented

15.Oozie is a \_\_\_\_\_\_\_\_\_ web application.

a) Java

16. Define Data Locality?

b) Locating computation logic near to data, instead of moving data to the computation logic or application space.

17. Hadoop supports \_\_\_\_\_\_\_\_\_\_\_\_ authentication.

b) Kerberos

18.. Traditional data structure consist\_\_\_\_\_\_\_\_\_\_\_\_ schema.

c) Static

19. Big data data structure consist\_\_\_\_\_\_\_\_\_\_\_\_ schema.

a) Dynamic

20. In Big data, data sources are \_\_\_\_\_\_\_\_\_\_\_\_\_\_.

a) Fully distributed

21. Neo4j supports Indexes by using \_\_\_\_\_\_\_\_\_\_\_.

b) Apache Lucene

**Chapter 3: MapReduce Paradigm**

1. \_\_\_\_\_\_\_\_\_ is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution.
b) JobConf

2. Task scheduling is handled by\_\_\_\_\_\_\_\_\_\_\_\_\_

d) Job tracker

3. Input splits created by

a) Driver program

4. How are keys and values presented and passed to the reducers during a standard sort and shuffle phase of MapReduce?

a) Keys are presented to reducer in sorted order; values for a given key are not sorted.

5. In the execution of a MapReduce job, where does the Mapper place the intermediate data of each Map task?

c) The Mapper stores the intermediate data on the underlying filesystem of the local disk of the machine which ran Map task

6. What do you call the processing technique and program model for distributed computing based on java in Hadoop?

b) MapReduce

7. Mathematical algorithms may does NOT include the following −

d) Algorithm

8. The map task is done by means of

b) Mapper Class

9. The reduce task is done by means of \_\_\_\_\_\_\_\_\_\_\_.

c) Reducer Class

10. In the Intersection operation in MapReduce, the reduce function must produce a tuple only if \_\_\_\_\_\_\_have the tuple.

1. both relations.

11. MapReduce programming framework uses for which tasks

1. Map and Reduce

12. For union operation in MapReduce, both the relation needs to have the \_\_\_\_\_\_\_\_ schema.

b)same

13. An input to a MapReduce is been divided into the fixed size of pieces named as

the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

b) Input splits

14. Shuffled and sorted data is passed as input to the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

a) Reducer

15. Map can emit \_\_\_\_\_\_\_\_\_ intermediate key-value pair.

a) more than one

16. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_function processes a key/value pair to generate a set of intermediate key/value pairs.

d) Map

.

|  |
| --- |
| 17.\_\_\_\_\_\_\_\_ Controls the partitioning of the keys of the intermediate map-outputs. |
|  |
| b)Partitioner |

18. Input splits,\_\_\_\_\_\_\_\_\_\_\_, Shuffling, Reducer are the phases in MapReduce operation.

b) Mapping

**Chapter 4: Mining Big Data Streams**

1. The timestamp of new bucket is the timestamp of the \_\_\_\_\_\_\_(later in time) of the two buckets

a) rightmost

2. \_\_\_\_\_\_\_\_\_\_\_\_ uses O(log2 N) bits to represent a window of N bits.

a) DGIM Algorithm

3. For Filtering Stream \_\_\_\_\_\_\_\_\_is used.

c) Bloom Filter

4. Park, Chen, Yu algorithm is useful for \_\_\_\_\_\_\_\_\_\_in Big Data Application.

b) Find Frequent Itemset

5. Distinct Element can be find using\_\_\_\_\_\_\_\_\_.

c) FM Algorithm

6. Which queries can be written by retaining a simple summary from the maximum of all stream elements, and not needed to record the entire stream?

a) Ad-hoc

7. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_is example of stream source?

c.) Internet and Web Traffic

8. Which of the following streaming windows show valid bucket representations according to the DGIM rules?

|  |
| --- |
| d) 1 0 1 1 0 0 0 1 0 1 1 1 0 1 1 0 0 1 0 1 1 |

9. To initialize the bit array in bloom filter, it always begins with all bits as \_\_\_\_\_\_\_.

a) ZERO

10.Bloom filter is a \_\_\_\_\_\_\_\_\_\_data structure.

a)Space efficient probabilistic

11. Which one is NOT a data mining tasks

d) Process of Data flow

12. Frequent Patterns and Association Rules

1. Support confidence

13. Algorithm FM is

1. Flajolet-Martin

14. Data mining is

a) The actual discovery phase of a knowledge

|  |
| --- |
| 15.Match the following a) Bloom filter i) Frequent Pattern Mining b) FM Algorithm ii) Filtering Stream c) PCY Algorithm iii) Distinct Element Findd) DGIM Algorithm iv) Counting 1’s in window |
|  |
| b)-ii), b-iii), c-i), d-iv) |

**Chapter 5: Big Data Mining Algorithms**

1. Clustering Using Representatives algorithm means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

b) CURE Algorithm

2. \_\_\_\_\_\_\_\_Clustering detects cluster with Spherical shape with Variable size.

a) CURE

3. To find Frequent Itemset \_\_\_\_\_\_\_\_algorithm used.

a) SON

4. Data Items which is common in all subsets is called as\_ \_\_\_\_\_\_\_in SON Algorithm

c) Candidate Key

5. \_\_\_\_\_\_\_\_\_\_ has 2 Map Reduce Phase.

a) SON Algorithm

6. Classification of Objects into groups is called as \_\_\_\_\_\_\_\_\_\_

b) Clustering

7. Consider a point that is correctly classified and distant from the decision boundary. Which of the following methods will be unaffected by this point?

b) SVM

8. The goal of clustering a set of data is to

a)divide them into groups of data that are near each other

 9. One type of Hierarchical Clustering is:

d) Bottom-Top and Top Down Clustering

10. Which command is used for Copy file or directories recursively?

b) Distcp

11. Strategic value of data mining is

c)time-sensitive

12. Frequent-Pattern Mining Does NOT represents \_\_\_\_\_\_\_\_\_\_\_\_.

d) Text format

13. Complete graph and then applying a search PageRank method is to \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

a) ranking all the nodes

14. SVD state for

b) Singular Value Decomposition

15. The canopy clustering algorithm is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_algorithm.

a) unsupervised pre-clustering

16. PCY algorithm used for \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_when the dataset is very large.

d) frequent itemset mining

17. PCY algorithm uses the technique of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_to filter out unnecessary item sets for next candidate itemset generation.

b) hashing

18. The \_\_\_\_\_\_\_\_\_\_\_\_\_ used to derive a classification from the K-nearest neighbors.

a) decision rule

19. In \_\_\_\_\_\_\_\_\_ main memory is divided into Chunks of memory.

a) SON

20.SON can be implemented in \_\_\_\_\_\_\_\_ sets of MapReduce.

a) Two

**Chapter 6: Big Data Analytics Applications**

1. The techniques for artificially increasing the \_\_\_\_\_\_\_\_\_\_\_of a page are collectively called Link Spam.

c)Page Rank

2. \_\_\_\_\_\_\_ is a Link analysis algorithm that rates the web pages.

b) HITS

3. v=(A^T\*u) is a \_\_\_\_\_\_\_\_\_\_\_\_\_.

a) Authority Weight Vector

4.\_\_\_\_\_\_\_\_ in the Social graph are Nodes.

a) Entities.

5. Simrank is used to analyzing \_\_\_\_\_\_\_\_\_\_\_\_\_.

a) Social Network graphs

6.\_\_\_\_\_\_\_\_\_\_\_\_\_ can be defined as Systems that evaluate quality based on the preferences of others with a similar point of view

a) Recommender systems

7. In-Degree and Out-Degree method is used in \_\_\_\_\_\_Algorithm.

c) HITS

8. Big Data With Facebook Tackles is based on what?

d) ProjectPrism.

9. \_\_\_\_\_\_\_\_is driven by the total size and number of maps usually

a) Outputs

10. On how many nodes mainly in HDFS Pig operates?

c) 2

11. Pig in batch mode uses which command?

a) Pig scripts

12. NoSQL database Apache Cassandra is use for

b) Facebook

13. \_\_\_\_\_\_\_\_\_\_\_\_\_is a measure of correlation between two variables in peer-based collaborative filtering.

a)Pearson Correlation Coefficient

14. How many type of document similarity exist?

b) Two

15. In \_\_\_\_\_\_Similarity it is considered as documents are similar if they contain large, identical sequences of character.

 a) Lexical

16. Decision tree is a type of \_\_\_\_\_\_\_\_\_\_\_, which consist collection of node, arranged as a binary tree.

a) Classifier

17. What Utility matrix offers?

d) Known information about the degree to which a user likes an item and predicts the values of the unknown entries based on the values of the known entries.