**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 Questions

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

Chapter 1: Introduction to Big Data

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

a) Velocity

b) Variety

c) Volume

d) Value

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

a) Organizes data

b) Analyzes data

c) Spread data

d) Collects data

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

a) no limit

b) limit

c) serial limit

d) factor dependent limit

4. DFS State for

1. Direct file system
2. Distributed file system
3. Disk file system
4. Data file system

5. Which is NOT an Application of Big data?

1. Banking
2. Education
3. Government
4. RDBMS

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

1. Scalability
2. Open-Source
3. Data Recovery
4. data redundancy

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

1. Data Recovery
2. Stream processing
3. Data mirroring
4. Web Scripting

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

1. Data units
2. Structural database
3. Unstructured database
4. Web data

|  |
| --- |
| 9. Which of the following batch Processing instance is NOT an example of \_\_\_\_\_\_\_\_\_\_\_Big Data Batch Processing? |
| a)Processing 10 GB sales data every 6 hours |
| b)Processing flights sensor data |
| c)Web crawling app |
| 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 |
| b)validation |
| c)transparency |
| d)decryption |

2. \_\_\_\_\_\_ is a framework for performing Machine Learning related task.
a) Drill
b) BigTop
c) Mahout
d) Chukwa

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

a) Grunt

b) FS

c) HDFS

d) CL

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

a) block size

b) average file size in the cluster

c) mininum hard disk size in the cluster

d) number of DataNodes

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

a) Increases

b) Decreases

c) Remains unchanged

d) May increase or decrease

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

a) HDFS.

b) Task Tracker.

c) Job Tracker.

d) Name Node.

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

a) CouchDB

b) MongoDB

c) Neo4j

d) Cloudant

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

1. SQL
2. JSON
3. CODE
4. HTML

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

a) Nodes

 b)Text

 c) Data structures

 d) Documents

10. Column family is Store\_\_\_\_\_\_\_\_.

1. Node
2. Matrix
3. Data
4. Query

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

1. Map Structure
2. Tree Structure
3. New Structure
4. Back Structure

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

1. Key-value
2. Graph
3. Column
4. RDBMS

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

1. Small data
2. Big data
3. Traditional data
4. Unstructured data

14. What type of storage used in MongoDB?

a) Document Oriented

b) Graph Oriented

c) Database Oriented

d) Group Oriented

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

a) Java

b) Phython

c) Android

d) Spark

16. Define Data Locality?

a) Code information platform

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

c) Locating computation logic by moving data to the application space.

d) Data information centres.

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

a) Code

b) Kerberos

c) Delta

d) Spar

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

a) Dynamic

b) Authentic

c) Static

d) Hybrid

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

a) Dynamic

b) Mixed

c) Static

d) Hybrid

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

a) Fully distributed

b) Partially distributed

c) Statically distributed

d) Centralized

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

a) CQL

b) Apache Lucene

c) SQL

d) Scala

**Chapter 3: MapReduce Paradigm**

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

d) HdfsConf

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

a)Reduce task

b)Task tracker

c)Map task

d) Job tracker

3. Input splits created by

a) Driver program

b) Job tracker

c) Map task

d) Reduce task

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.

b) Keys are presented to reducer in sorted order; values for a given key are sorted in ascending order.

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

d) Keys are presented to a reducer in random order; values for a given key are sorted in ascending order.

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

a) The Hadoop framework holds the intermediate data in the Task Tracker’s memory

b) The Mapper transfers the intermediate data to the JobTracker, which then sends it to the Reducers

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

d) The Mapper transfers the intermediate data to the reducers as soon as it is generated by the Map task

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

a) Master Task

b) MapReduce

c) Reduce Task

d) Datanode

7. Mathematical algorithms may does NOT include the following −

1. Sorting
2. Searching
3. Indexing
4. Algorithm

8. The map task is done by means of

1. Object Class
2. Mapper Class
3. Task Class
4. Title Class

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

1. MapReduce Class
2. Shuffle Class
3. Reducer Class
4. MAP Class

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

1. both relations.
2. any one relation
3. left relation
4. right relation

11. MapReduce programming framework uses for which tasks:

1. Map and Reduce
2. Object Class
3. Inheritance
4. HTML and CSS

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

1. different
2. same
3. duplicate
4. Writable

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

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

a) Reducer

b) Input splits

c) Output splits

d) Data splits

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

a) Reducer

b) Input step

c) Output buffer

d) Mapper

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

a) more than one

b) one

c) only two

d) buffered

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

a) Map and Reduce

b) Shuffle

c) Reduce

d) Map

|  |
| --- |
| 17.\_\_\_\_\_\_\_\_ Controls the partitioning of the keys of the intermediate map-outputs. |
| a)Collector |
| b)Partitioner |
| c)InputFormat |
| d)HCatalog |

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

a) Merging

b) Mapping

c) Sorting

d) Mini Combiner

**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

b) leftmost

c) side most

d) equal

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

a) DGIM Algorithm

b) PCY Algorithm

c) FM Algorithm

d) Bloom filter

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

a) PCY Filter

b) FM Filter

c) Bloom Filter

d) Block Filter

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

a) Filtering Stream

b) Find Frequent Itemset

c) Find Distinct element

d) Counting Window

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

a) DGIM Algorithm

b) PCY Algorithm

c) FM Algorithm

d) Bloom filter

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

b) Standard based

c.) Standing

d) Standing and Ad-hoc

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

a) Cain and Able tool

b) Query engine

c.) Internet and Web Traffic

d) Relational database

|  |
| --- |
| 8. Which of the following streaming windows show valid bucket representations according to the DGIM rules? |
| a)1 0 1 1 1 0 1 0 1 1 1 1 0 1 0 1 |
| b)1 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 1 0 0 1 |
| c)1 1 1 1 0 0 1 1 1 0 1 0 1 |
| 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

b) ONE

c.) SIMILAR

d) ZERO AND ONE

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

a)Space efficient probabilistic

b)Simple

c)Complex

d) Structured

11. Which one is NOT a data mining tasks

1. Association,
2. correlation,
3. causality analysis
4. Process of Data flow

12. Frequent Patterns and Association Rules

1. Support confidence
2. Set connection
3. Super confidence
4. Simple condition

13. Algorithm FM is

1. Flajolet-Martin
2. Flow- Math
3. Function-Matrix
4. time variant non-volatile collection of data

14. Data mining is

1. The actual discovery phase of a knowledge
2. The actual discovery post of a knowledge
3. task of assigning a classification
4. time variant non-volatile collection of data

|  |
| --- |
| 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 |
| a)-ii), b-i), c-iii), d-iv) |
| b)-ii), b-iii), c-i), d-iv) |
| c)-iv), b-iii), c-ii), d-i) |
| d)-i), b-ii), c-iii), d-iv) |

**Chapter 5: Big Data Mining Algorithms**

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

a) SON Algorithm

b) CURE Algorithm

c) K-means Algorithm

d) K-Medoid Algorithm

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

a) CURE

b) k-means

c) Hierarchical

d) k-medoids

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

a) SON

b) CURE

c) FM

d) K-Medoid

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

a) Primary Key

b) Foreign Key

c) Candidate Key

d) Unique Key

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

a) SON Algorithm

b) CURE Algorithm

c) FM Algorithm

d) K-Medoid Algorithm

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

a) Filtering

b) Clustering

c) Frequent Itemset

d) Streaming

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?

a) Nearest neighbor

b) SVM

c) Logistic regression

d) Linear regression

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

1. divide them into groups of data that are near each other
2. determine the nearest neighbors of
3. each of the data
4. choose the best data from the set

9. One type of Hierarchical Clustering is:

1. Bottom-Top Clustering
2. Top-Down Clustering (Divisive)
3. Right- Down
4. Bottom-Top and Top Down Clustering

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

1. Dtcp
2. Distcp
3. Dcp
4. distc

11. Strategic value of data mining is

* 1. cost-sensitive
	2. work-sensitive
	3. time-sensitive
	4. technical-sensitive

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

1. Apriori (Candidate generation & test)
2. Projection-based (FPgrowth, CLOSET+, ...)
3. Vertical format approach (CHARM, ...)
4. Text format

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

 a) Ranking all the nodes

b) Arrange of nodes

c) Deleted nodes

d) Sorting a node

14. SVD state for

1. SignValue Decomposition
2. Singular Value Decomposition
3. SameValue Decomposition
4. System Value Decomposition

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

a) unsupervised pre-clustering

b) unsupervised response

c) Clustering supervised

d) supervised

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

a) closed unsupervised mining

b) unsupervised response mining

c) frequent unsupervised representatives

d) frequent itemset mining

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

a) mirroring

b) hashing

c) bloom filtering

d) page ranking

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

a) decision rule

b) hashing

c) filtering

d) page ranking

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

a) SON

b) FM

c) DGIM

d) CURE

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

a) Two

b) One

c) Five

d) Three

**Chapter 6: Big Data Analytics Applications**

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

a)Hub

b)Authority

c)Page Rank

d) Trust Rank

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

a) In-Degree

b) HITS

c) page rank

d)Out-Degree

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

a) Authority Weight Vector

b) Hub Weight Vector

c) Page Rank

d) web graph

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

a) Entities.

b) Computer

c) Network

d)People

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

a)Social Network graphs

b)Web graphs

c) Page graphs

d) Link 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

b) Content-Based system

c) collaborative filtering

d) text mining

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

a) SON

b) CURE

c) HITS

d) PCY

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

1. Prism
2. ProjectBid
3. ProjectData
4. ProjectPrism

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

1. Outputs
2. Next
3. Inputs
4. Tasks

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

1. 3
2. 4
3. 2
4. 1

11. Pig in batch mode uses which command ?

1. Pig scripts
2. Pig options
3. Pig function
4. Pig shell command

12. NoSQL database Apache Cassandra is use for

* 1. MySpace
	2. Facebook
	3. LinkedIn
	4. Twit

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

* 1. Pearson Correlation Coefficient
	2. Correlation Coefficient
	3. Pearls Coefficient
	4. Point Correlation Coefficient

14. How many type of document similarity exist?

* 1. Three
	2. Two
	3. Four
	4. Five

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

a) Lexical

b) Semantic

c) Paralytics

d )Frequency

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

a) Classifier

b) Content Analysis

c) Collaborative filter

d )Decision

17. What Utility matrix offers?

a) Unknown 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.

1. Known information about the values to which a user likes an item and predicts the values of the unknown entries.
2. Known information about the degree to which a user dislikes an item and predicts the values of the known entries based on the values of the unknown entries.
3. 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.