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
   a) Direct file system
   b) Distributed file system
   c) Disk file system
   d) Data file system

5. Which is NOT an Application of Big data?

a) Banking
b) Education
c) Government
d) RDBMS

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

a) Scalability
b) Open-Source
c) Data Recovery
d) data redundancy

7. Apache Spark is capable of _____.

a) Data Recovery
b) Stream processing
c) Data mirroring
d) Web Scripting

8. Traditional data management_____.

a) Data units
b) Structural database
c) Unstructured database
d) 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.

a) SQL
b) JSON
c) CODE
d) HTML

9. Graph is used to store _____ in NoSQL
a) Nodes
b)Text
c) Data structures
d) Documents

10. Column family is Store_____.
a) Node
b) Matrix
c) Data
d) Query

11. Document Stored in NoSQL is of_____.
a) Map Structure
b) Tree Structure
c) New Structure
d) Back Structure

12. In NoSQL Data Architected Pattern is NOT representing _____.
a) Key-value
b) Graph
c) Column
d) RDBMS

13. Hadoop is not fit for_____.

a) Small data
b) Big data
c) Traditional data
d) 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 –

a) Sorting
b) Searching
c) Indexing
d) Algorithm

8. The map task is done by means of

a) Object Class
b) Mapper Class
c) Task Class
d) Title Class

9. The reduce task is done by means of _____.

a) MapReduce Class
b) Shuffle Class
c) Reducer Class
d) MAP Class

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

a) both relations.
b) any one relation
c) left relation
d) right relation

11. MapReduce programming framework uses for which tasks:

a) Map and Reduce
b) Object Class
c) Inheritance
d) HTML and CSS

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

a) different
b) same
c) duplicate
d) 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

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
a) Association,
b) correlation,
c) causality analysis
d) Process of Data flow

12. Frequent Patterns and Association Rules
a) Support confidence
b) Set connection
c) Super confidence
d) Simple condition

13. Algorithm FM is
a) Flajolet-Martin
b) Flow- Math
c) Function-Matrix
d) time variant non-volatile collection of data

14. Data mining is
a) The actual discovery phase of a knowledge
b) The actual discovery post of a knowledge
c) task of assigning a classification
d) 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 Find
d) 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
   a) divide them into groups of data that are near each other
   b) determine the nearest neighbors of
   c) each of the data
   d) choose the best data from the set

9. One type of Hierarchical Clustering is:
   a) Bottom-Top Clustering
   b) Top-Down Clustering (Divisive)
   c) Right- Down
   d) Bottom-Top and Top Down Clustering

10. Which command is used for Copy file or directories recursively?
   a) Dtcp
   b) Distcp
   c) Dcp
   d) distc

11. Strategic value of data mining is
   a) cost-sensitive
   b) work-sensitive
   c) time-sensitive
   d) technical-sensitive

12. Frequent-Pattern Mining Does NOT represents _____.

   a) Apriori (Candidate generation & test)
   b) Projection-based (FPgrowth, CLOSET+, ...)

c) Vertical format approach (CHARM, ...)
d) 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
a) SignValue Decomposition
b) Singular Value Decomposition
c) SameValue Decomposition
d) 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?
a) Prism
b) ProjectBid
c) ProjectData
d) ProjectPrism

9. _____ is driven by the total size and number of maps usually.
a) Outputs
b) Next
c) Inputs
d) Tasks

10. On how many nodes mainly in HDFS Pig operates?
a) 3
b) 4
c) 2
d) 1

11. Pig in batch mode uses which command ?
a) Pig scripts
b) Pig options
c) Pig function
d) Pig shell command

12. NoSQL database Apache Cassandra is use for

a) MySpace
b) Facebook
c) LinkedIn
d) Twit

13. _____is a measure of correlation between two variables in peer-based collaborative filtering.

a) Pearson Correlation Coefficient
b) Correlation Coefficient
c) Pearls Coefficient
d) Point Correlation Coefficient

14. How many type of document similarity exist?

a) Three
b) Two
c) Four
d) 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.

b) Known information about the values to which a user likes an item and predicts the values of the unknown entries.
c) 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.
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.