Sample Questions

Big Data Analytics (ITC801)

(CBCGS-H) SEM VIII

SET 1

1. Consider the following scenario and list the V's satisfied in the scenario: "Businesses and governments have a large amount of data that needs to be analysed and processed very quickly. If this data is fragmented into smaller chunks and spread over many machines, all those machines process their portion of the data in parallel and the results are obtained extremely fast."
   a. Velocity, Volume
   b. Velocity, Variety
   c. Volume, Variety
   d. Velocity

2. What makes Big Data analysis difficult to optimize?
   (A) Big Data is not difficult to optimize
   (B) Both data and cost-effective ways to mine data to make business sense out of it
   (C) The technology to mine data
   (D) All of the above

3. Which of the following is not a big data open source software framework?
   a. Hadoop
   b. Apache Spark
   c. Apache Storm
   d. Xplenty

4. XML data is:
   a. Structured
   b. Semi-structured
   c. Unstructured
   d. Highly structured

5. Which of the following is the advantage of Big Data approach over Traditional approach?
   a. Pre-processing is required before storing.
   b. There is a limit on how much data to be stored and for how long.
   c. No protection against hardware failure.
   d. Any unstructured data such as text, images and videos can be stored.

6.Which one of the below is not a core component of Hadoop:

1. YARN
2. NameNode
3. HDFS
4. MapReduce

7. _____ is a high-level scripting language which is used to compile MapReduce

1. Pig
2. SQL
3. C
4. Sqoop

8. _____ is the slave node and holds the user data in the form of Data Blocks:
a) DataNode
b) NameNode
c) Data block
d) Replication

9. Which of the following is not a NoSQL Data Architectural Pattern:

1. Key Value Store
2. Row Store
3. Document Store
4. Column Store

10. _____ is an example of Graph Store:

1. Cassandra
2. MongoDB
3. Neo4j
4. Redis

11. The process of horizontal partitioning of a large database which can be stored on different servers is called:

1. Replication
2. Sharding
3. Distribution
4. Hashing

12. CAP stands for:

1. Coordination, Atomicity and Partitioning
2. Consistency, Availability and Partition Tolerance
3. Consistency, Atomicity and Partition
4. Contingency, Atomicity and Partition Tolerance

13. ZooKeeper allows distributed processes to coordinate with each other through a shared hierarchial name space of data registers called _____

1. hnodes
2. znodes
3. ynodes
4. qnodes

14. Hadoop's target is to run on clusters of the order of _____ nodes.

1. 10
2. 100
3. 1000
4. 10,000

15. Hadoop license is distributed under _____
A. Apache License 2.0
B. Mozilla 2.0
C. Shareware
D. Google Chrome

16. Hadoop written in_____
A. Python

B. PHP
C. Java
D. JSP

17. _____ is the most popular high-level Java API in Hadoop Ecosystem
A. Scalding
B. HCatalog
C. Cascalog
D. Cascading

18. Point out the correct statement
A. Hadoop do need specialized hardware to process the data
B. Hadoop 2.0 allows live stream processing of real time data
C. In Hadoop programming framework output files are divided in to lines or records
D. None of the mentioned

19. Which of the following is not an input format in Hadoop ?
A. TextInput Format
B. ByteInput Format
C. SequenceFile Inputf ormat
D. KepInput Format

20. What was Hadoop named after?
A. Creator Doug Cutting favorite circus act
B. Cutting high school rock band
C. The toy elephant of Cutting son
D. A sound Cutting laptop made during Hadoop development

21. What protocol is used by the mapping nodes in order to transfer the output of the map process to the reducer node?

a. TCP
b. SMTP
c. IP
d. UDP

22. The inbound traffic that is faced by reducer node is known as

a. Traffic congestion
b. In-cast
c. inside traffic
d. Transfer traffic

23. What is format of inputs given to Map process?

a. document Format
b. File Format
c. Graph/Tree
d. key-value pair

24. When the Reducer function is both associative & commutative, then some of the tasks are computed at the Map side using _____

a. shuffler
b. sort
c. reducer
d. combiners

25. In projection using Map reduce, what is the task of reducer function?

a. eliminating unwanted attributes
b. To eliminate two identical tuples
c. To transfer intermediate keys
d. To provide key, value pair to combiner

26 Which of the following maps input key/value pairs to a set of intermediate key value pair?

a. Reducer
b. Partitioner
c. Mapper
d. Combiner

27. What task is performed simultaneously multiple times?

a. Shuffle & map
b. Shuffle & sort
c. Reduce & sort
d. Map & sort

28. Arrange the following into correct order of MapReduce process

i. Output format
ii. shuffle & sort
iii. Mapper,
iv. input data
v. reducer

a. iv->iii->ii->v->i
b. iii->iv->ii->i-> v
c. i->ii->iii->iv->v
d. v->iv->iii->ii->i

29. A task tracker node acts as the Slave and is responsible for executing a Task assigned to it by the_____.

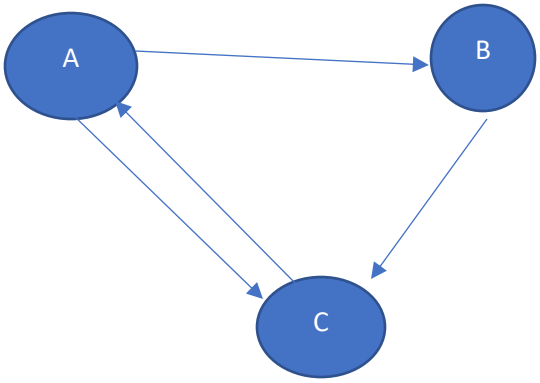
A. MapReduce

B. Mapper

C. Task

D. JobTracker

30. Find the PageRank for the node A:



a) 1
b) 0.575
c) 1.06375
d) 1.0541875

Sample Questions

Big Data Analytics (ITC801)

(CBCGS-H) SEM VIII

SET 2

1. Which algorithm is used for finding frequent item sets/pairs?
   a. Apriori algorithm
   b. Regression
   c. Support vector machine (SVM)
   d. Naïve Bayesian

2. 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

3. A Bloom filter guarantees no

   a. false positives
   b. false negatives
   c. false positives and false negatives
   d. false positives or false negatives, depending on the Bloom filter type

4. Flajolet-Martin algorithm runs in _____ time and needs _____ memory

   a. O(n) , O(log(m))
   b. O(n) , O(log(n))
   c. O(log(n)) ,O(n)
   d. O(log(m)) ,O(n)

5. Mapper maps ____ key/value pairs to a set of intermediate key/value pairs.

A. input

B. output

C. internal

D. external

6. Point out the correct statement

A. Map Task in MapReduce is performed using the Mapper() function
B. Reduce Task in MapReduce is performed using the Map() function
C. All of the mentioned
D. MapReduce tries to place the data and the compute as close as possible

6. In which approach is structured data stored in data marts and data warehouses?
   a. Big Data Approach
   b. Traditional Approach
   c. A/B Testing
   d. Data Mining
7. Which of the following is a column-oriented database in Hadoop Ecosystem:
   1. Sqoop
   2. Hive
   3. HBase
   4. Pig

9. Which of the following is a collection of pre-existing data mining algorithms:

   1. Mahout
   2. Oozie
   3. HBase
   4. Hive

10. CAP theorem is also known as:

   a. Brewer's Theorem

   b. Brew Theorem

   c. Bayes Theorem

   d. Consistency Theorem

11. Semi structured data separate data elements using _____

a. Tags
b. Fixed fields
c. Random fields
d. User generated field

12. Which one given below is not an example of Key Value Store:

1. Redis

2. Amazon Dynamo

3. MongoDB

4. Azure Table Storage

13. What are DGIM's maximum error boundaries?

a.DGIM always underestimates the true count; at most by 25%

b.DGIM always overestimates the count; at most by 50%

c.DGIM either underestimates or overestimates the true count; at most by 25%

d.DGIM either underestimates or overestimates the true count; at most by 50%

14. Which algorithm works by selecting representative points for each cluster?

a. BFR Algorithm
b. CURE Algorithm
c. PCY Algorithm
d. SON Algorithm

15. The organization which has the world's largest Hadoop cluster is :

A. Apple

B. Google

C. Facebook

D. IBM

16. “A stock price checker alerts the user when a stock price crosses a particular price point”. This is an example of _____

a) Continuous queries

b) One-time queries

c) Join queries

d) Ad-hoc queries

17. 10. Which of the following is correct format for Mapper function?
a.map (InputKeyType inputKey, InputValueType inputValue)
b. map (intermediatetype intermediate key, iterator)
c.map (InputValueType inputValue, InputKeyType inputKey)

d.map (shuffler, sorter)

18. Falsifying the origin of an internet communication (emails, webpages) in order to mislead the recipient is called:

a) Spoofing
b) Spamdexing
c) Web spam
d) Search spam

19. How many key- value pair will be generated for a multiplication of Matrix A (2,3) and Matrix B (3,4) using MapReduce after completing mapping?

a. 6
b. 4
c. 8
d. 10

15. _____ is general-purpose computing model and runtime system for distributed data analytics.

A. Mapreduce
B. Drill
C. Oozie
D. Hive

17. Mapper implementations are passed the JobConf for the job via the _____ method

A. JobConfigure.configure
B. JobConfigurable.configure
C. JobConfigurable.configureable
D. None of the mentioned

18. Which of the following phases occur simultaneously ?

A. Reduce and Sort
B. Shuffle and Sort
C. Shuffle and Map
D. All of the mentioned

19. HDFS works in a _____ fashion

A. master-worker
B. master-slave

C. worker/slave.

D. All of the mentioned

20. _____ is the process of extracting useful knowledge from continuous, rapid data streams.

a) Data Stream Mining
b) Pattern mining
c) Changes Mining
d) Concept drift

1. Which of the following is not an example of data stream applications?

a) Sensor Networks
b) Network traffic analysis
c) Human behavior
d) Transaction Log analysis

20. Which of the following is not a data stream query processing issue?

a) Aproximate query answering
b) Sliding windows
c) Bounded memory requirements
d) Blocking Operators

21. What relates to observing an infinite stream of data, looking at each of its items and quantifying whether the item is of interest and should be stored for further evaluation?

a) Sampling
b) Stream Query
c) Decaying windows
d) Filtering

22. Bloom filter has a recall rate of 100% because_____

a) False negative matches are possible but false positives are not.
b) False positive matches are possible but false negatives are not.
c) There is nothing called as false positive or false negative.
d) False positive matches and negatives both are always possible.

23. Which algorithm is used to estimate the number of distinct elements in the stream?

a) Flajolet-Martin (FM) Algorithm
b) Datar-Gionis-Indyk-Motwani (DGIM) Algorithm
c) Bloom Filter
d) Randomized Sampling Algorithm

24. What is the complexity of the Datar-Gionis-Indyk-Motwani (DGIM) Algorithm?

a) O(n)
b) $O(\log_2 N)$
c) O(nlogN)
d) $O(n^2)$

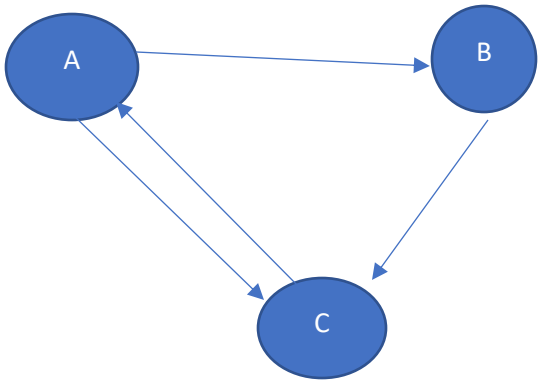
25. Which of the following is not a constraint that must be satisfied in DGIM Algorithm?

a) The right end of a bucket always starts with a position with a 1.
b) Number of 1s must be a power of 2.
c) Buckets overlap in timestamps.
d) Buckets are sorted by size. Earlier buckets are not smaller than later buckets

27. Which of the following statements about the standard DGIM algorithm are false?

a. DGIM not operates on a time-based window.
b. In DGIM, the size of a bucket is always a power of two.
c. The maximum number of buckets has to be chosen beforehand.
d. The buckets contain the count of 1's and each 1's specific position in the stream.

28. 30. Find the PageRank for the node C:



a) 1
b) 0.575
c) 1.06375
d) 1.0541875

29. Recommendation based on similarity measures between users and/or items is an example of_

a) Content-based systems
b) Collaborative filtering systems
c) Nearest neighbor technique
d) User Profiles

30. Which of the following means selecting a subset of data to be analysed?

a) Sampling
b) Filtering
c) Counting
d) Querying