1) Why are big data applications libel to instancy?

- BD cannot use in-memory computing
- BD applications are still in the early stages of development
- The volume of BD is too large to be analyzed rapidly
- BD may reside in a different in a location from the application

2) Data locality means:

- Moving data to computation
- Moving computation to data
- Moving data to computation and Moving computation to data
- None

3) Which of the following is fundamental datastruture of Spark?

- RDD
- Data frame
- Dataset
- None

4) Typically, a HBase region server is collocated with

- HDFS namenode
- HDFS datanode
- As a client to hdfs server
- Resource manager

5) which command is used to check the status of all daemons running in hdfs

- Fsck
- Distcp
- Jps
- None

6) module for structured data processing

- GraphX
- MLlib
- Spark SQl
- Spark R

7) what message is generated by a datanode to indicate its connectivity with namenode ?

- Beep
- Heartbeat
- Analog pulse
- Map

8) _____________is not a component of spark

- SparkR
- Mllib
- GraphX
- Sqoop

9) As compared to RDBMS, Hadoop

A - Has higher data Integrity.

B - Does ACID transactions

C - IS suitable for read and write many times

D - Works better on unstructured and semi-structured data.

10) why do we use ssh in hadoop cluster

- To perform the password less authentication
- To establish the communication between Master and worker
- Both
- None

11) which of the following services is provided by yarn

- Global resource management
- Record reader
- MapReduce engine
- Data mining

12) hive provides a sql like language called

- SQL Hive
- Hive QL
- DB QL
- Hive Data

13) Which of the following statements about Hadoop are false?
a) Hadoop is a distributed framework
b) The main algorithm used in Hadoop is MapReduce
c) Hadoop runs with commodity hardware
d) All true

14) The main advantage of creating table partition is

A - Effective storage memory utilization

B - faster query performance

C - Less RAM required by namenode

D - simpler query syntax

15) What can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.

(a) MapReduce
(b) Mahout
(c) Oozie
(d) All of the mentioned

16) Which of the below property gets configured on core-site.xml ?

A - Replication factor

B - Directory names to store hdfs files.

C - Host and port where MapReduce task runs.

D - Java Environment variables.

17) on dropping a managed/internal table

- The schema dropped without dropping data
- The data dropped without dropping schema
- An error is throw
- Both the schema and the data is dropped

18) to retrieve all rows of a table in Hbase we use

- get
- scan
- put
- select

19) In a word count query using MapReduce, what does the map function do?

A) It sorts the words alphabetically and returns a list of the most frequently used words.

B) It returns a list with each document as a key and the number of words in it as the value.The master JobTracker sends map and reduce functions to the same machines or nodes in a cluster.

C) It creates a list with each word as a key and every occurrence as value 1.

D) It creates a list with each word as a key and the number of occurrences as the value.

20) You use the hadoop fs –put command to write a 300 MB file using and HDFS block size of 64 MB. Just after this command has finished writing 200 MB of this file, what would another user see when trying to access this file?

a.) They would see Hadoop throw an Concurrent File Access Exception when they try to access this file.
b.) They would see the current state of the file, up to the last bit written by the command.
c.) They would see the current of the file through the last completed block.
d.) They would see no content until the whole file written and closed.

21) Which of the following deal with small files issue?

> - A. Hadoop archives
> - HBase
> - Sequence files
> - All of the above

22) Apache HBase was modeled after Google's

- Foundation DB
- Big Top
- Big Table
- None

23) if a big data analyst were to analyze data from a database of call logs provided by a telecom service provider which element of big data would be dealing with?

- Volume
- Variety
- Velocity
- Variable

24) udf stands for

- Universal Defined Function
- Unique Defined Function
- Universal Disk format
- Unique definition of function

25) Can you provide multiple input paths to a map-reduce jobs?

A. Yes, but only in Hadoop 0.22+.

B. No, Hadoop always operates on one input directory.

C. Yes, developers can add any number of input paths.

D. Yes, but the limit is currently capped at 10 input paths

26) the create statement in hive is related to

- DDL statement
- DML
- Session control
- Embedded sql

27) one of your HBase region server (in a well configured in a sized HBase and Hadoop cluster) is reporting bad performance (slow response) what can be the possible reason?

- Small rows and Colum names
- Uneven key space distribution
- Small Colum family name
- None

28) Hadoop framework is written in:

- Java
- Python
- Scala
- C++

29) the parameter fs.default.name is set in _____________ configuration file?

- Hadoop-env.sh
- Mapred-site.xml
- Core-site.xml
- Hdfs-site.xml

30) which of the following statements are correct?

- spark can run on the top of Hadoop\
- Spark can process data stored in HDFS
- Spark can use Yarn as resource management layer
- All

31) Common feature of RDD and DataFrame?

- Immutability
- In-memory
- Resilient
- All

32) which characteristics does not belong to big data

- Volume
- Variety
- Velocity
- Variable

33) hive uses __________ to store metadata

- Derby database
- HiveQL
- NoSQL
- SQl

34) HBase is defined as ___

- Row oriented
- Colum oriented
- Tuple oriented
- None

35) which of the following is true about Apache airflow?

- Open source
- Workflow management platform
- Data transformation pipeline
- All