**Big Data and Hadoop practice question**

Q.1 Which of the following are the characteristics of Big Data?

A: Volume B: Velocity C: Variety D: All of the above

Q.2 \_\_\_\_deal with the growing number of data sources.

A: Business developer B: Data scientist

C: Sales executive D: Web designer

Q.3 Which of the following are the traditional database technology?

A: RDBMS B: DBMS C: Flat files D: All of the above

Q.5 Half structured data can be called as\_\_\_

A: Structured data B: Unstructured data

C: Semi- Structured data D: Mixed data

Q.6 Which of the following happens In distributed computing?

A: The computing task is divided among several computers

B: Additional high capacity disks are added to the system

C: The results are shared among several users of the network

D: The computing task is moved to the cloud

Q.7 Why are big data applications susceptible to latency?

A: Big data may reside in a different location from the application

B: The volume of big data is too large to be analyzed rapidly

C: Big data cannot use in-memory computing

D: Big data applications are still in the early stages of development.

Q.8 What are the five V’s of Big Data?

A. Volume

B. Velocity

C. Variety

D. All the above

Q.9 What are the challenges with Big Data?

A. Storage

B. Quality

C. Security

D. All the above

Q.10 The data of GPS satellite can be categorized as:

A: Structured data B: Unstructured data

C: Both Structured and unstructured data D: Semi- Structured data

Q.11 According to analysts, for what can traditional IT systems provide a foundation when they’re integrated with big data technologies like Hadoop?

A. Big data management and data mining

B. Data warehousing and business intelligence

C. Management of Hadoop clusters

D. Collecting and storing unstructured data

Q.12 What are the different features of Big Data Analytics?

A. Open-Source

B. Scalability

C. Data Recovery

D. All the above

Q.13 All of the following accurately describe Hadoop, EXCEPT:

A. Open source

B. Real-time

C. Java-based

D. Distributed computing approach

Q.14 Which of the following is not an example of NoSQL database?

A: HBase B: MongoDB

C: Allegro Graph D: Oracle

Q.15 Which of the following is an example of a relational database?

A: SQL B: Oracle

C: SQL SERVER 2012 D: All of the above

Q.16 Which of the following is the daemon of Hadoop?

A: Namenode B: Node manager

C: Datanode D: All of the above

Q.17 Which one of the following is false about Hadoop?

A: It is a distributed framework

B: The main algorithm used in Hadoop is Map Reduce

C: Hadoop can work with commodity hardware

D: All are true

Q.18 Hadoop can work with :

A: Structured B: Unstructured

C: Semi-structured D: All of the above

Q.19 Which of the following characteristic of big data deals with types of data?

A: Volume B: Variety

C: Velocity D: Variable

Q.20 What are the daemons that are required to start the HDFS

A: Resource Manager, Name Node

B: Name Node, Secondary Name Node, Data Node

C: Data Node, Node Manager

D: Data Node, Node Manager, Secondary Name Node

**21**. Which of the following is component of Hadoop?

A: Yarn

B: HDFS

C: Map Reduce

D: All of the above

Q.22 Who is master node in HDFS?

A: Namenode B: Datanode

C: Blocks D: Namespace

Q.23 Programming model of hadoop is\_\_\_\_\_\_?

A: Beep B: Heartbeat

C: Analog pulse D: Map reduce

Q.24 Which of the following defines metadata?

A: Data about data B: data from web logs

C: data from government sources D: data from market surveys

Q.25 Map Reduce can handle\_\_\_\_\_\_\_\_\_\_\_

A: web logs B: images

C: structured data D: unstructured data

Q.26 Which of the following statement explains the use of computing resources by Hadoop architecture?

A: software is distributed to computing resources.

B: data and computing tasks is distributed to computing resources

C: shared memory is created for computing resources.

D: data is distributed to computing resources

Q.27 Hadoop works in

A: master-slave fashion

B: Centralized processing fashion

C: worker/slave fashion

D: All of the mentioned

Q.28 Which of the following is true about Hadoop?

A: Hadoop is schema-less – we don't need to define a schema before writing the data

B: Hadoop can't handle structured data

C: Hadoop MapReduce processes the data at very low latency

D: None of the mentioned

Q.29 Which of the following options most aptly explains the reason behind the creation of Map Reduce?

A: Need to increase the processing power of new hardware

B: Need to perform complex analysis of structured data

C: Need to increase the number of web users

D: Need to spread distributed computing

Q.30 Which of the following describes the Map function?

A: It processes data to create a list of key-value pairs.

B: It indexes the data to list all the words occurring in it

C: It converts a relational database to key value pairs.

D: It tracks data across multiple tables and clusters in Hadoop

Q.31 Which of the following is incorrect big data technology?

A: hadoop

B: kafka

C: pytorch

D: spark

Q.32 The output of a mapper task is

A: The Key-value pair of all the records of the dataset.

B: The Key-value pair of all the records from the input split processed by the mapper

C: Only the sorted Keys from the input split

D: The number of rows processed by the mapper task.

Q.33 What is the role of map function in a word count query?

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

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

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

D: It returns a list with each document as a key and the number of words in it as the value.

Q.34 Which of the following absorbs the huge inflow of data and sort it out in different categories?

A: Data sources B: Ingestion

C: Security D: Visualization

Q.35 \_\_\_\_\_\_is an open source framework that enables you to store large volumes of data in a distributed manner across multiple machines.

A: Hadoop B: Hive

C: Pig D: Zookeeper

Q.36 In Hadoop 2.x release HDFS federation means

A: Allowing namenodes to communicate with each other.

B: Allow a cluster to scale by adding more datanodes under one namenode.

C: Allow a cluster to scale by adding more namenodes.

D: Adding more physical memory to both namenode and datanode.

Q.37 Which of the following is the correct statement:

A: Data locality means moving computation to data instead of data to computation B: Data locality means moving data to computation instead of computation to data

C: Both A and B

D: None of the above

Q.38 Which of the following is managed by the Map Reduce environment?

A: web logs B: images

C: structured data D: unstructured data

Q.39 Datanode and namenode are\_\_\_\_\_ and \_\_\_\_\_\_

A: worker, master

B: master, worker

C: worker, worker

D: master, master

Q.40 Which of the following is not the aim of HDFS?

A: to handle volume of data

B: prevent deletion of data

C: fault detection and recovery

D: Provide high network bandwidth

Q.41 Which one of the following statements is true regarding <key,value> pairs of a MapReduce job?

A: A value class must extend WritableComparable.

B: A key class must implement WritableComparable.

C: A value class must implement WritableComparable.

D: A key class must implement Writable.

Q.42 The MapReduce algorithm contains two important tasks, namely \_\_\_\_\_\_\_\_\_\_.

A. mapped, reduce  
B. mapping, Reduction  
C. Map, Reduction  
D. Map, Reduce

Q.43 \_\_\_\_\_ takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

A. Map  
B. Reduce  
C. Both A and B  
D. Node

Q.44 \_\_\_\_\_\_ task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

A. Map  
B. Reduce  
C. Node  
D. Both A and B

Q.45 In how many stages the MapReduce program executes?

A. 2  
B. 3  
C. 4  
D. 5

46. Which of the following is/are INCORRECT with respect to Hive?

A. Hive provides SQL interface to process large amount of data  
B. Hive needs a relational database like oracle to perform query operations and store data.  
C. Hive works well on all files stored in HDFS  
D. Both A and B

47. Which of the following is not a Features of HiveQL?

A. Supports joins  
B. Supports indexes  
C. Support views  
D. Support Transactions

Q.48 Renaming an existing table name with a new name

A: ALTER TABLE old\_table\_name RENAME TO new\_table\_name;

B: ALTER TABLE RENAME TO new\_table\_name;

C: ALTER TABLE old\_table\_name new\_table\_name;

D: None of the above

Q.49 Derby database is used by \_\_\_\_\_\_ to store metadata:

A: Hive B: Pig

C: NoSQL D: SQL

Q.50 In Hive, Directories are created in:

A: Partitioning

B: Bucketing

C: Both a and b

D: none

Q.51 Which is not the component of hive architecture?

A: Driver B: MAP C: Execution engine D: shell

52.  What is Hive?

[**A.**](javascript:%20void(0)) An open source data warehouse system

[**B.**](javascript:%20void(0)) relational database

[**C.**](javascript:%20void(0)) OLTP

[**D.**](javascript:%20void(0)) An langauge

Q.53 What is the command to print a list of default tables in hive?

A: show tables

B: describe tables

C: list tables

D: None of the above

Q.54 What is the command to drop a hive table?

A: Drop table tablename

B: drop –e table

C: exit table

D: None

Q.55 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

Q.56 While loading data into managed tables, If the LOCAL clause is mentioned, it

A: Moves the data from local filesystem to the target files system

B: Copies the data from local filesystem to target final file system

C: Overwrites the data in the target file system

D: Merges with the data in the target file system

#### 

Q.57 The command SET in Hive is useful when :

A: You want to set an explicit property in ‘hive-site.xml’.

B: You want to set an internal property in ‘hive-site.xml’.

C: You are not permitted to use any other command.

D: You cannot add external properties like SET.

58. Hive can load the data from:

A. Local File system

B. HDFS File system

C. Output of a Pig Job

D. All of the above

59. The partitioning of a table in Hive creates more

[**A.**](javascript:%20void(0)) subdirectories under the database name

[**B.**](javascript:%20void(0)) subdirectories under the table name

[**C.**](javascript:%20void(0)) files under database name

[**D.**](javascript:%20void(0)) files under the table name

60. The clause used to limit the number of rows returned by a query is

[**A.**](javascript:%20void(0)) Rownum

[**B.**](javascript:%20void(0)) Restrict

[**C.**](javascript:%20void(0)) Maxrow

[**D.**](javascript:%20void(0)) Limit

61. Which of the following scenarios are not prevented by enabling strict mode in Hive?

[**A.**](javascript:%20void(0)) Scanning all the partitions

[**B.**](javascript:%20void(0)) Generating random sample of data

[**C.**](javascript:%20void(0)) Running a order by clause without a LIMIT clause

[**D.**](javascript:%20void(0)) Cartesian product

62. Which of the following is foundation of Hbase?

a. Foundation DB

b. Big top

c. Big Tables

d. None of the above

63. Which of the following is the possible cause of bad performance of HBase region server ?

a. uneven key space distribution

b. small rows and column names

c. small column family name

d. None of the above

#### Q.64 We can use Spark over Hadoop in-

1.Standalone

2.YARN

3.SIMR (Spark in MapReduce)

4. All of the above

#### Q.65 Which of the following is fundamental datastructure of Spark?

1.RDD

2. DataFrame

3. Dataset

4. None of the above

#### Q.66 DataFrame in Apache Spark prevails over RDD and doesnot contain any feature of RDD.

1. TRUE

2. FALSE

#### Q.67 Which of the following is not true for DataFrame?

1. DataFrame in Apache Spark is behind RDD

2. We can build DataFrame from different data sources. structured data file, tables in Hive

3. The Application Programming Interface (APIs) of DataFrame is available in various languages.

4. Both in Scala and Java, we represent DataFrame as Dataset of rows.

#### Q.68 We can create DataFrame using-

1.Tables in Hive

2. Structured data files

3. External databases

4. All of the above

#### Q.69 RDD allows Java serialization.

1. TRUE

2. FALSE

#### Q.70 Which of the following is not true for Apache Spark Execution?

1.To simplify working with structured data it provides DataFrame abstraction in Python, Java, and Scala.

2. The data can be read and written in a variety of structured formats. For example, JSON, Hive Tables, and Parquet.

3. Using SQL we can query data,only from inside a Spark program and not from external tools.

4. The best way to use Spark SQL is inside a Spark application. This empowers us to load data and query it with SQL.

#### Q.71 If we launch the application through Spark submit, application jar doesnot gets distributed to all worker nodes.

1. TRUE

2. FALSE

#### Q.72 When SQL run from the other programming language the result will be-

1.DataFrame

2. DataSet

3. Either DataFrame or Dataset

4. Neither DataFrame nor Dataset

#### Q.73 Which of the following are uses of Apache Spark SQL?

1. It executes SQL queries.

2. We can read data from existing Hive installation using SparkSQL.

3. When we run SQL within another programming language we will get the result as Dataset/DataFrame.

4. All of the above

#### Q.74 With the help of Spark SQL, we can query structured data as a distributed dataset (RDD).

1. TRUE

2. FALSE

#### Q.75 Which of the following is true for Spark SQL?

1.Hive transactions are not supported by Spark SQL.

2. No support for time-stamp in Avro table.

3. Even if the inserted value exceeds the size limit, no error will occur.

4. All of the above

#### Q.76 Which of the followimg is true for Catalyst optimizer?

1.The optimizer helps us to run queries much faster than their counter RDD part.

2.The optimizer helps us to run queries little faster than their counter RDD part.

3.The optimizer helps us to run queries in the same speed as their counter RDD part.

4. None of the above

#### Q.77 The Dataset API is accessible in-

1. Java and Scala

2. Java, Scala and python

3. Scala and Python

4. Scala and R

#### Q.78 Does Dataset API support Python and R.

1. yes

2. no

#### Q.79 which of the following provide object-oriented programming interface?

1.RDD

2.DataFrame

3.Dataset

4.All of the above

#### Q.80 Which of the following the data is organized into named column?

1.RDD

2.DataFrame

3.Dataset

4.a and b

#### Q.81 Spark SQL can connect through JDBC or ODBC.

1.TRUE

2.FALSE

#### Q.82 Which type of processing Apache Spark can handle?

1. Batch Processing

2. Interactive Processing

3. Stream Processing

4. Graph Processing

5. All of the above

#### Q.83 Spark is developed in

1. Scala

2. Java

#### Q.84 Which of the following statements are correct?

1. Spark can run on the top of Hadoop

2. Spark can process data stored in HDFS

3. Spark can use Yarn as resource management layer

4. All of the above

#### Q.85 Spark's core is a batch engine

1. TRUE

2. FALSE

#### Q.86 Spark is 100x faster than Map Reduce due to

1. In-memory computing

2. Development in Scala