

Big data is a field that treats ways to analyze systematically extract information from or otherwise deal with data sets. Data can be large or complex to be dealt with by traditional data processing applications software

A large amount of data

It is a popular term used to express the exponential growth of data.

Big data is difficult to store, collect, maintain, analyze and visualize.

Distributed file system: Distributed file system is a file system in which data is stored on a server. The data is accessed and processed as if it were stored on the local client machine.

Characteristics of distributed file system:

- ❖ Transparency
- ❖ user mobility
- ❖ Performance
- ❖ simplicity and ease of use
- ❖ Scalability
- ❖ high availability
- ❖ high reliability

Big data tools:

- ❖ Apache Hadoop,
- ❖ Apache Storm,
- ❖ Cassandra,
- ❖ Mongo DB
- ❖ Neo4j.

Big data sources:

- ❖ Amazon
- ❖ Redshift
- ❖ Mongo DB

Challenges of big data:

- ❖ Uncertainty of data management
- ❖ The talent gap in big data
- ❖ Getting data into a big data structure
- ❖ Synchronizing across data sources
- ❖ Integration

Benefits of big data:

- ❖ Cost
- ❖ Time reduction
- ❖ Speeding up decision making
- ❖ Analyze in real-time
- ❖ Model and Test variation

Characteristics of big data:

- ❖ Volume
- ❖ Velocity
- ❖ Variety

Types of big data:

- ❖ Structured
- ❖ unstructured
- ❖ Semi-structured
- ❖ hybrid

Use cases of big data

- ❖ Recommendation engine
- ❖ Analyzing call detail records
- ❖ Fraud detection
- ❖ Market basket analysis
- ❖ sentiment analysis

1. What are the main components of big data?

- A. HDFS
- B. MapReduce
- C. YARN
- D. All of the above

Answer - D) All of the above are the main components of big data.

2. On which of the following platforms does Hadoop run?

- A. Debian
- B. Cross-platform
- C. Bare metal
- D. Unix-like

Answer - B) Hadoop runs on cross-platform.

3. Data in ____ bytes size is called big data

- A. Meta
- B. Giga
- C. Tera
- D. Peta

Answer - D) Data in petabyte size is known as big data.

4. Transaction of data of the bank is a type of.

- A. Unstructured data
- B. Structured data
- C. Both a and b
- D. None of the above

Answer - B) A transaction of data of the bank is structured data.

5. The total forms of big data is ____

- A. 1
- B. 2
- C. 3
- D. 4

Answer - C) There are three forms of big data- unstructured, structured, and semi-structured.

6. Identify the incorrect big data Technologies.

- A. Apache Pytorch
- B. Apache Kafka
- C. Apache Hadoop
- D. Apache Spark

Answer - A) Apache PyTorch is incorrect.

7. In which language is Hadoop written?

- A. C++
- B. Java
- C. Rust
- D. Python

Answer - B) Hadoop is required in Java.

8. _____ is a collection of data that is used in volume, yet growing exponentially with time

- A. Big Database
- B. Big DBMS
- C. Big Datafile
- D. Big Data

Answer - D) Big data is a collection of data that is used in volume, yet growing exponentially with time.

9. Identify among the options below which is general-purpose computing model and runtime system for Distributed Data Analytics.

- A. HDFS
- B. MapReduce
- C. Oozie
- D. All of the above

Answer - B) MapReduce Is a general-purpose computing model and runtime system for Distributed Data Analytics.

10. Choose the primary characteristics of big data among the following

- A. Value
- B. Variety
- C. Volume
- D. All of the above

Answer - D) All of the above are primary characteristics of big data.

11. Identify whether true or false: Qubole Is a big data tool.

- A. True
- B. False

Answer - A) True. Qubole Is a big data tool.

12. Choose the languages which are used in data science.

- A. C++
- B. C
- C. R
- D. Ruby

Answer - C) R is used in data science.

13. Which of the following is not a part of the data science process.

- A. Communication building
- B. Discovery
- C. Operationalize
- D. Model planning

Answer - A) Communication building Is not a part of the data science process

14. Identify the different features of Big Data Analytics.

- A. Open-source
- B. Data recovery
- C. Scalability
- D. All of the above

Answer - D) All of the above are features of Big Data Analytics.

15. Total V's of big data is ____

- A. 3
- B. 4
- C. 5
- D. 6

Answer - C) There are a total of 5 V's of big data.

16. Among the following options choose the one which depicts the correct reason why big data analysis is difficult to optimize.

- A. The technology to mine data
- B. Both data and cost-effective ways to mine data to make business sense out of it
- C. Big data is not difficult to optimize
- D. None of the above

Answer - B) Both data and cost-effective ways to mine data to make business sense out of it.

17. All of the following accurately describe Hadoop, except

- A. Open source
- B. Java-based
- C. Real-time
- D. Distributed computing approach

Answer - C) Hadoop is not a real-time platform.

18. Which of the following are the Benefits of Big Data Processing?

- A. Businesses can utilize outside intelligence while taking decisions.
- B. Better operational efficiency
- C. Improve customer service
- D. All of the above

Answer - D) All of the above are benefits of big data processing.

19. Big data analysis does the following except?

- A. Spreads data
- B. Analyze data
- C. Organizes data
- D. Collect data

Answer - b) Big data analysis doesn't analyze data.

20. Which of the following is true about big data?

- A. Big data can be processed using traditional techniques
- B. Big data refers to data sets that are at least a petabyte in size
- C. Big data analysis does not involve reporting and data mining techniques
- D. Big data has low velocity meaning that it is generated slowly

Answer - B) Big data refers to data sets that are at least of petabyte in size is true.

21. Which of the following can be generally used to clean and prepare big data.

- A. Pandas
- B. Data lake
- C. U-SQL
- D. Data warehouse

Answer - D) Data warehouse is generally used to clean and prepare big data.

22. Identify the operation which can be performed in the data warehouse.

- A. Alter
- B. Modify
- C. Scan
- D. Read/write

Answer - C) Scan operation can be performed on the data warehouse.

23. Among the following options which component deals with ingesting streaming data into Hadoop?

- A. Oozie
- B. Hive
- C. Kafka
- D. Flume

Answer - D) Flume deals with ingesting streaming data into Hadoop.

24. Among the following option which of the following property gets configured on mapred-site.xml

- A. Java environment variables
- B. Replication factor
- C. Directory names to store hdfs files
- D. Host and port where MapReduce task runs.

Answer - D) Host and Port get configured on mapred-site.xml.

25. Mapper class is

- A. Static type
- B. Generic type
- C. Abstract type
- D. Final

Answer - B) Mapper class is a generic type.

26. Among the following which does the Job control in Hadoop?

- A. Task class
- B. Mapper class
- C. Job class
- D. Reducer class

Answer - C) Job control is handled by the Job class.

27. Identify the term used to define the multidimensional model of the data warehouse.

- A. Table
- B. Data cube
- C. Tree
- D. Data structure

Answer - B) The multidimensional model of a data warehouse is known as a data cube.

28. Fixed-size pieces of MapReduce job is known as _____

- A. Splits
- B. Tasks
- C. Maps
- D. Records

Answer - A) Fixed size pieces of MapReduce job is known as splits.

29. The output of map tasks is written in?

- A. Local disk
- B. File system
- C. HDFS
- D. Secondary storage

30. What is the time horizon in the data warehouse?

- A. 3-4 years
- B. 5-6 years
- C. 5-10 years
- D. 1-2 years

Answer - C) The time horizon in the data warehouse is 5-10 years.

31. Where can the data be updated?

- A. Informational environment
- B. Data warehouse environment
- C. Operational environment
- D. Data mining environment

Answer - C) Data can be updated in an operational environment

32. Hadoop Common Package contains?

- A. msi files
- B. war files
- C. exe files
- D. jar files

Answer - D) Hadoop Common Package contains jar files.

33. Small logical units where data warehouses hold large amounts of data is known as ____.

- A. Access layers
- B. Data marts
- C. Data storage
- D. Data miners

34. Choose the incorrect property of the data warehouse.

- A. Collection from heterogeneous sources
- B. Subject oriented
- C. Time variant
- D. Volatile

Answer - D) Volatile is not a property of the data warehouse.

35. Identify the slave node among the following.

- A. Job node
- B. Data node
- C. Task node
- D. Name node

Answer - B) Data node is known as the slave node.

36. What is the source of all data warehouse data known as?

- A. Formal environment
- B. Data warehouse environment
- C. Operational environment
- D. Technology environment

Answer - C) The source of all data warehouse data is known as the Operational environment.

37. Fact tables are _____

- A. HDFS
- B. MapReduce
- C. YARN
- D. All of the above

Answer - D) All of the above are the main components of big data.

38. Identify the correct definition of Reconciled data.

- A. Reconcile data is data stored in one operational system in the organization.
- B. Reconcile data is the data that has been selected and formatted for end-user support applications.
- C. Reconcile data is the current data intended to be the single source for all decision support systems

Answer - C) Reconcile data is the current data intended to be the single source for all decision support systems.

39. Identify the node which acts as a checkpoint node in HDFS.

- A. Secondary Name node
- B. Secondary data node
- C. Name node
- D. Data node

40. Identify the most common source of change data in refreshing a data warehouse.

- A. Logged change data
- B. Cooperative change data
- C. Queryable change data
- D. Snapshot change data

Answer - C) Queryable change data is the most common source of change data in accessing a data warehouse.

41. DSS in data warehouse stands for _____

- A. Decision single system
- B. Decision support system
- C. Data support system
- D. Data storable system

Answer - B) DSS stands for a Decision support system.

42. _____ is data about data.

- A. HDFS
- B. MapReduce
- C. YARN
- D. All of the above

Answer - D) All of the above are the main components of big data.

43. How many approaches are there in data warehousing to integrate heterogeneous databases?

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

Answer - A) There are two different approaches to integrating heterogeneous databases - a) query driver approach b) update driven approach.

44. Identify the correct options which are considered before investing in data mining

- A. Vendor consideration
- B. Functionality
- C. Compatibility
- D. All of the above

Answer - D) All of the above are considered before investing in data mining.

45. Efficiency and scalability of data mining algorithms" issues come under?

- A. Mining Methodology and User Interaction Issues
- B. Performance Issues
- C. Diverse Data Types Issues
- D. None of the above

Answer - B) Efficiency and scalability of data mining algorithms" issues come under performance issues.

46. Identify among the following for which system of data warehousing is mostly used.

- A. Data mining and data storage
- B. Data integration and data storage
- C. Reporting and data analysis
- D. Data cleaning and data storage

Answer - C) System of data warehousing is mostly used for Reporting and data analysis.

47. What is the use of data cleaning?

- A. To remove the noisy data
- B. Transformations to correct the wrong data.
- C. Correct the inconsistencies in data
- D. All of the above

Answer - D) All of the above are uses of data mining.

48. What is the minimum amount of data that a disk can read or write in HDFS?

- A. Byte size
- B. Block size
- C. Heap
- D. None of the above

Answer - B) The minimum amount of data that a disk can read or write in HDFS is of block size.

1. Data in _____ bytes size is called Big Data.

- A. Tera
- B. Giga
- C. Peta
- D. Meta

2. How many V's of Big Data

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

3. Transaction data of the bank is?

- A. structured data
- B. unstructured data
- C. Both A and B
- D. None of the above

4. In how many forms BigData could be found?

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

5. Which of the following are Benefits of Big Data Processing?

- A. Businesses can utilize outside intelligence while taking decisions
- B. Improved customer service
- C. Better operational efficiency
- D. All of the above

6. Which of the following are incorrect Big Data Technologies?

- A. Apache Hadoop
- B. Apache Spark
- C. Apache Kafka
- D. Apache Pytarch

7. The overall percentage of the world's total data has been created just within the past two years is ?

- A. 80%
- B. 85%
- C. 90%
- D. 95%

8. Apache Kafka is an open-source platform that was created by?
- A. LinkedIn
 - B. Facebook
 - C. Google
 - D. IBM
9. What was Hadoop named after?
- A. Creator Doug Cutting's favorite circus act
 - B. Cuttings high school rock band
 - C. The toy elephant of Cutting's son
 - D. A sound Cutting's laptop made during Hadoop development
10. What are the main components of Big Data?
- A. MapReduce
 - B. HDFS
 - C. YARN
 - D. All of the above
11. All of the following accurately describe Hadoop, EXCEPT _____
- A. Open-source
 - B. Real-time
 - C. Java-based
 - D. Distributed computing approach
12. _____ has the world's largest Hadoop cluster.
- A. Apple
 - B. Datamatics
 - C. Facebook
 - D. None of the above
13. Facebook Tackles Big Data With _____ based on Hadoop.
- A. Project Prism
 - B. Prism
 - C. Project Big
 - D. Project Data
14. _____ is general-purpose computing model and runtime system for distributed data analytics.
- A. Mapreduce
 - B. Drill
 - C. Oozie
 - D. None of the above

15. The examination of large amounts of data to see what patterns or other useful information can be found is known as

- A. Data examination
- B. Information analysis
- C. Big data analytics
- D. Data analysis

16. Big data analysis does the following except?

- A. Collects data
- B. Spreads data
- C. Organizes data
- D. Analyzes data

17. 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. None of the above

18. The new source of big data that will trigger a Big Data revolution in the years to come is?

- A. Business transactions
- B. Social media
- C. Transactional data and sensor data
- D. RDBMS

19. The unit of data that flows through a Flume agent is

- A. Log
- B. Row
- C. Record
- D. Event

20. Listed below are the three steps that are followed to deploy a Big Data Solution except

- A. Data Processing
- B. Data dissemination
- C. Data Storage
- D. Data Ingestion

|

1. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including _____

- a) Improved data storage and information retrieval
- b) Improved extract, transform and load features for data integration
- c) Improved data warehousing functionality
- d) Improved security, workload management, and SQL support

2. 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 the Hadoop programming framework output files are divided into lines or records
- d) None of the mentioned

3. 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

4. Hadoop is a framework that works with a variety of related tools. Common cohorts include _____

- a) MapReduce, Hive and HBase
- b) MapReduce, MySQL and Google Apps
- c) MapReduce, Hummer and Iguana
- d) MapReduce, Heron and Trumpet

5. Point out the wrong statement.

- a) Hardtop processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data
- b) Hadoop uses a programming model called "MapReduce", all the programs should conform to this model in order to work on the Hadoop platform
- c) The programming model, MapReduce, used by Hadoop is difficult to write and test
- d) All of the mentioned

6. What was Hadoop named after?

- a) Creator Doug Cutting's favorite circus act
- b) Cutting's high school rock band
- c) The toy elephant of Cutting's son
- d) A sound Cutting's laptop made during Hadoop development

7. All of the following accurately describe Hadoop, EXCEPT _____

- a) Open-source
- b) Real-time
- c) Java-based
- d) Distributed computing approach

8. _____ 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

9. _____ has the world's largest Hadoop cluster.

- a) Apple
- b) Datamatics
- c) Facebook
- d) None of the mentioned

10. Facebook Tackles Big Data With _____ based on Hadoop.

- a) 'Project Prism'
- b) 'Prism'
- c) 'Project Big'
- d) 'Project Data'

According to analysts, for what can traditional IT systems provide a foundation?

- (A) Big data management and data mining
- (B) Data warehousing and business intelligence
- (C) Management of Hadoop clusters
- (D) Collecting and storing unstructured data

Answer -A

2.What are the main components of Big Data?

- (A) MapReduce
- (B) HDFS
- (C) YARN
- (D) All of these

Answer -D

3.What are the different features of Big Data Analytics?

- (A) Open-Source
- (B) Scalability
- (C) Data Recovery
- (D) All the above

Answer -D

4.According to analysts, for what can traditional IT systems provide a foundation?

- (A) Big data management and data mining
- (B) Data warehousing and business intelligence
- (C) Management of Hadoop clusters
- (D) Collecting and storing unstructured data

Answer -A

5.What are the four V's of Big Data?

- (A) Volume
- (B) Velocity
- (C) Variety
- (D) All the above

Answer-D

6.IBM and _____ have announced a major initiative to use Hadoop to support university courses in distributed computer programming

- a) Google Latitude
- b) Android (operating system)
- c) Google Variations
- d) Google

Answer: d

7.Point out the correct statement.

- a) Hadoop is an ideal environment for extracting and transforming small volumes
- b) Hadoop stores data in HDFS and supports data compression/decompression
- c) The Giraph framework is less useful than a MapReduce job to solve graph analysis
- d) None of the mentioned

Answer: b

8. What license is Hadoop distributed under?

- a) Apache License 2.0
- b) Mozilla Public License
- c) Shareware
- d) Commercial

Answer: a

Explanation: Hadoop is Open Source, released under Apache 2 license.

9. Which of the following platforms does Hadoop run on?

- a) Bare metal
- b) Debian
- c) Cross-platform
- d) Unix-like

Answer: c

Explanation: Hadoop has support for cross-platform operating system.

Hadoop achieves reliability by replicating the data across multiple hosts and hence does not require _____ storage on hosts.

- a) RAID
- b) Standard RAID levels
- c) ZFS
- d) Operating system

Answer: a

Explanation: With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack.

Above the file systems comes the _____ engine, which consists of one Job Tracker, to which client applications submit MapReduce jobs.

- a) MapReduce
- b) Google
- c) Functional programming
- d) Facebook

Answer: a

Explanation: MapReduce engine uses to distribute work around a cluster.

The Hadoop list includes the HBase database, the Apache Mahout _____ system, and matrix operations.

- a) Machine learning
- b) Pattern recognition
- c) Statistical classification
- d) Artificial intelligence

Answer: a

Explanation: The Apache Mahout project's goal is to build a scalable machine learning tool.

As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including _____

- a) Improved data storage and information retrieval
- b) Improved extract, transform and load features for data integration
- c) Improved data warehousing functionality
- d) Improved security, workload management, and SQL support

Answer: d

Explanation: Adding security to Hadoop is challenging because all the interactions do not follow the classic client-server pattern.

Point out the wrong statement.

- a) Hardtop processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data
- b) Hadoop uses a programming model called “MapReduce”, all the programs should confirm to this model in order to work on Hadoop platform
- c) The programming model, MapReduce, used by Hadoop is difficult to write and test
- d) All of the mentioned

Answer: c

Explanation: The programming model, MapReduce, used by Hadoop is simple to write and test.

What was Hadoop named after?

- a) Creator Doug Cutting’s favorite circus act
- b) Cutting’s high school rock band
- c) The toy elephant of Cutting’s son
- d) A sound Cutting’s laptop made during Hadoop development

Answer: c

Explanation: Doug Cutting, Hadoop creator, named the framework after his child’s stuffed toy elephant.

All of the following accurately describe Hadoop, EXCEPT _____

- a) Open-source
- b) Real-time
- c) Java-based
- d) Distributed computing approach

Answer: b

Explanation: Apache Hadoop is an open-source software framework for distributed storage and distributed processing of Big Data on clusters.

_____ 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

Answer: a

Explanation: MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel distributed architecture.

_____ has the world’s largest Hadoop cluster.

- a) Apple
- b) Datamatics
- c) Facebook
- d) None of the mentioned

Answer: c

Explanation: Facebook has many Hadoop clusters, the largest among them is the one that is used for Data warehousing.

_____ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.

- a) Scalding
- b) HCatalog
- c) Cascalog
- d) All of the mentioned

Answer: c

Explanation: Cascalog also adds Logic Programming concepts inspired by Datalog. Hence the name “Cascalog” is a contraction of Cascading and Logic.

Hive also support custom extensions written in _____

- a) C#
- b) Java
- c) C
- d) C++

Answer: b

Explanation: Hive also support custom extensions written in Java, including user-defined functions (UDFs) and serializers-deserializers for reading and optionally writing custom formats.

Point out the wrong statement.

- a) Elastic MapReduce (EMR) is Facebook’s packaged Hadoop offering
- b) Amazon Web Service Elastic MapReduce (EMR) is Amazon’s packaged Hadoop offering
- c) Scalding is a Scala API on top of Cascading that removes most Java boilerplate
- d) All of the mentioned

Answer: a

Explanation: Rather than building Hadoop deployments manually on EC2 (Elastic Compute Cloud) clusters, users can spin up fully configured Amazon EMR clusters.

_____ is the most popular high-level Java API in Hadoop Ecosystem

- a) Scalding
- b) HCatalog
- c) Cascalog
- d) Cascading

Answer: d

Explanation: Cascading hides many of the complexities of MapReduce programming behind more intuitive pipes and data flow abstractions.

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

- a) Mapreduce
- b) Drill
- c) Oozie
- d) None of the mentioned

Answer: a

Explanation: Mapreduce provides a flexible and scalable foundation for analytics, from traditional reporting to leading-edge machine learning.

The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to _____

- a) SQL
- b) JSON
- c) XML
- d) All of the mentioned

Answer: a

Explanation: Pig Latin, in essence, is designed to fill the gap between the declarative style of SQL and the low-level procedural style of MapReduce.

_____ jobs are optimized for scalability but not latency.

- a) Mapreduce
- b) Drill
- c) Oozie
- d) Hive

Answer: d

Explanation: Hive Queries are translated to MapReduce jobs to exploit the scalability of MapReduce.

_____ is a framework for performing remote procedure calls and data serialization.

- a) Drill
- b) BigTop
- c) Avro
- d) Chukwa

Answer: c

Explanation: In the context of Hadoop, Avro can be used to pass data from one program or language to another.

Big Data Question 1:

The data node and name node in HADOOP are

1. Worker Node and Master Node respectively
2. Master Node and Worker Node respectively
3. Both Worker Nodes
4. Both Master Nodes

Answer (Detailed Solution Below)

Option 1 : Worker Node and Master Node respectively

Big Data Question 2:

Point out the wrong statement :

1. Non-Relational databases require that schemas be defined before you can add data.
2. NoSQL databases are built to allow the insertion of data without a predefined schema.
3. NewSQL databases are built to allow the insertion of data without a predefined schema.
4. All of the options.

Answer (Detailed Solution Below)

Option 1 : Non-Relational databases require that schemas be defined before you can add data.

Big Data Question 3:

Which of the following is component of Hadoop?

1. YARN
2. HDFS
3. Map reduce
4. All of the options

Answer (Detailed Solution Below)

Option 4 : All of the options

Big Data Question 4:

Big Data is generally characterised by three Vs that stand for _____, _____ and _____.

1. Volume ; Viscosity ; Variety
2. Variety ; Velocity ; Vivid
3. Viscosity ; Volume ; Velocity
4. Volume ; Variety ; Velocity
5. Volume ; Variety ; Viscosity

Answer (Detailed Solution Below)

Option 4 : Volume ; Variety ; Velocity

Big Data Question 5:

Which of the following statement/s is/are true?

- (i) Facebook has the world's largest Hadoop cluster.
- (ii) Hadoop 2.0 allows live stream processing of real time data

1. Neither (i) nor (ii)
2. Both (i) and (ii)
3. (i) only
4. (ii) only

Answer ([Detailed Solution Below](#))

Option 2 : Both (i) and (ii)

Big Data Question 6

[Download Solution PDF](#)

The data node and name node in HADOOP are

1. Worker Node and Master Node respectively
2. Master Node and Worker Node respectively
3. Both Worker Nodes
4. Both Master Nodes

Answer ([Detailed Solution Below](#))

Option 1 : Worker Node and Master Node respectively

Big Data Question 7

[Download Solution PDF](#)

Which of the following is component of Hadoop?

1. YARN
2. HDFS
3. Map reduce
4. All of the options

Answer ([Detailed Solution Below](#))

Option 4 : All of the options

Big Data Question 8

[Download Solution PDF](#)

Point out the wrong statement :

1. Non-Relational databases require that schemas be defined before you can add data.
2. NoSQL databases are built to allow the insertion of data without a predefined schema.
3. NewSQL databases are built to allow the insertion of data without a predefined schema.
4. All of the options.

Answer ([Detailed Solution Below](#))

Option 1 : Non-Relational databases require that schemas be defined before you can add data.

Big Data Question 9:

The data node and name node in HADOOP are

1. Worker Node and Master Node respectively
2. Master Node and Worker Node respectively
3. Both Worker Nodes
4. Both Master Nodes

Answer ([Detailed Solution Below](#))

Option 1 : Worker Node and Master Node respectively

Big Data Question 10:

Hadoop (a big data tool) works with number of related tools. Choose from the following, the common tools included into Hadoop:

1. MySQL, Google API and Map reduce
2. Map reduce, Scala and hummer
3. Map reduce, H base and Hive
4. Map reduce, hummer and Heron

Answer ([Detailed Solution Below](#))

Option 3 : Map reduce, H base and Hive

Big Data Question 11:

Which of the following is component of Hadoop?

1. YARN
2. HDFS
3. Map reduce
4. All of the options

Answer ([Detailed Solution Below](#))

Option 4 : All of the options

Big Data Question 12:

Which of the following statement/s is/are true?

- (i) Facebook has the world's largest Hadoop cluster.
- (ii) Hadoop 2.0 allows live stream processing of real time data

1. Neither (i) nor (ii)
2. Both (i) and (ii)
3. (i) only
4. (ii) only

Answer ([Detailed Solution Below](#))

Option 2 : Both (i) and (ii)

Big Data Question 13:

In Jan 2021, All India Council for Technical Education (AICTE) joined hands with which of the following to train 5 lakh students and faculty on cybersecurity?

1. Everdata Technologies
2. Quick Heal Technologies Ltd
3. eRaksha Foundation
4. Cyber Peace Foundation

Answer ([Detailed Solution Below](#))

Option 4 : Cyber Peace Foundation

Big Data Question 14:

Big Data is generally characterised by three Vs that stand for _____, _____ and _____.

1. Volume ; Viscosity ; Variety
2. Variety ; Velocity ; Vivid
3. Viscosity ; Volume ; Velocity
4. Volume ; Variety ; Velocity
5. Volume ; Variety ; Viscosity

Answer ([Detailed Solution Below](#))

Option 4 : Volume ; Variety ; Velocity

Big Data Question 15:

Point out the wrong statement :

1. Non-Relational databases require that schemas be defined before you can add data.
2. NoSQL databases are built to allow the insertion of data without a predefined schema.
3. NewSQL databases are built to allow the insertion of data without a predefined schema.
4. All of the options.

Answer ([Detailed Solution Below](#))

Option 1 : Non-Relational databases require that schemas be defined before you can add data.

- 1. What are the main components of Big Data?**
 - (a) MapReduce
 - (b) HDFS
 - (c) YARN
 - (d) All of these
- 2. What are the different features of Big Data Analytics?**
 - (a) Open-Source
 - (b) Scalability
 - (c) Data Recovery
 - (d) All the above
- 3. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with Big Data?**
 - (a) Big data management and data mining
 - (b) Data warehousing and business intelligence
 - (c) Management of Hadoop clusters
 - (d) Collecting and storing unstructured data
- 4. What are the four V's of Big Data?**
 - (a) Volume
 - (b) Velocity
 - (c) Variety
 - (d) All the above
- 5. All of the following accurately describe Hadoop, EXCEPT:**
 - (a) Open-source
 - (b) Real-time
 - (c) Java-based
 - (d) Distributed computing approach
- 6. _____ is general-purpose computing model and runtime system for distributed data analytics.**
 - (a) Mapreduce
 - (b) Drill
 - (c) Oozie
 - (d) None of the above
- 7. The examination of large amounts of data to see what patterns or other useful information can be found is called:**
 - (a) Data examination
 - (b) Information analysis
 - (c) Big data analytics
 - (d) Data analysis
- 8. Big data analysis does the following except**
 - (a) Collects data
 - (b) Spreads data
 - (c) Organizes data
 - (d) Analyzes data

- 9. 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
- 10. The new source of big data that will trigger a Big Data revolution in the years to come is**
- (a) Business transactions
 - (b) Social media
 - (c) Transactional data and sensor data
 - (d) RDBMS
- 11. The unit of data that flows through a Flume agent is**
- (a) Log
 - (b) Row
 - (c) Event
 - (d) Record
- 12. Listed below are the three steps that are followed to deploy a Big Data Solution except**
- (a) Data Ingestion
 - (b) Data Processing
 - (c) Data dissemination
 - (d) Data Storage
- 13. Check below the best answer to "which industries employ the use of so-called "Big Data" in their day to d**
- (a) Weather forecasting
 - (b) Marketing
 - (c) Healthcare
 - (d) All of the above
- 14. There are almost as many bits of information in the digital universe as there are stars in the actual univers**
- (a) True
 - (b) False
- 15. The word 'Big data' was coined by**
- (a) Roger Mougallas
 - (b) John Philips
 - (c) Simon Woods
 - (d) Martin Green
- 16. The word 'Big Data' was coined in the year**
- (a) 2000
 - (b) 1970
 - (c) 1998
 - (d) 2005

- 17. Concerning the Forms of Big Data, which one of these is odd?**
- (a) Structured
 - (b) Unstructured
 - (c) Processed
 - (d) Semi-Structured
- 18. Big Data applications benefit the media and entertainment industry by**
- (a) Predicting what the audience wants
 - (b) Ad targeting
 - (c) Scheduling optimization
 - (d) All of the above
- 19. The feature of big data that refers to the quality of the stored data is _____**
- (a) Variety
 - (b) Volume
 - (c) Variability
 - (d) Veracity

Unit1: Big data MCQ

1. How many V's are present in the Big Data?

- a. 3
- b. 4
- c. 5
- d. 6

Answer: c

2. Data in a Relational Database is:

- a. Structured
- b. Un-Structured
- c. Semi Structured
- d. Meta Data

Answer: a

3. Data is found in the big data, in how many forms?

- a. 2
- b. 3
- c. 4
- d. 5

Answer: b

4. What kind of data is in Log files?

- a. Structured
- b. Un-Structured
- c. Semi Structured
- d. Meta Data

Answer: c

5. What is the overall percentage of the world's total data created within the past two years is:

- a. 80%
- b. 85%
- c. 90%
- d. 95%

Answer: c

6. What are the main components present in the Big Data Analytics?

- a. MapReduce
- b. HDFS
- c. YARN
- d. All of the above

Answer: d

6. What are the major benefits of Big Data Processing?

- a. Businesses can utilize outside intelligence while taking decisions
- b. Improved customer service
- c. Better operational efficiency
- d. All of the above

Answer: d

7. The Hadoop is written in which programming language?

- a. C
- b. C++
- c. Java
- d. Python

Answer: c

8. Which of the following option given are NOT related to the big data problem(s)?

- a. Parsing 5 MB XML file every 2 minutes
- b. Processing the twitter data
- c. Processing online banking transactions
- d. both (a) and (c)

Answer: d

9. What does the characteristics “Velocity” in Big Data represents?

- a. Speed of input data generation
- b. Speed of individual machine processors
- c. Speed of ONLY storing data
- d. Speed of storing and processing data

Answer: d

10. Which of the following are example(s) of Real Time Big Data Processing?

- a. Complex Event Processing (CEP) platforms
- b. Stock market data analysis
- c. Bank fraud transactions detection
- d. both (a) and (c)

Answer: d

11. Hadoop is open source.

- a. ALWAYS True
- b. True only for Apache Hadoop
- c. True only for Apache and Cloudera Hadoop
- d. ALWAYS False

Answer: b

12. Which of the following is not an example of Social Media?

- a. Twitter
- b. Google
- c. Insta
- d. Youtube

Answer: b

13. By 2027, the volume of data produced digitally will reach to

- a. TB
- b. YB
- c. ZB
- d. EB

Answer: c

14. For Drawing insights for Business what are need?

- a. Collecting the data
- b. Storing the data
- c. Analysing the data
- d. All the above

Answer: d

15. Does Facebook uses "Big Data " to determine the behavior of its users? Is this True or False.

- a. TRUE
- b. FALSE

Answer: a

16. The Process of describing the data that is huge and complex to store and process is known as

- a. Analytics
- b. Data mining
- c. Big Data
- d. Data Warehouse

Answer: c

17. Data generated from online transactions is one of the example for volume of big data. Is this true or False.

- a. TRUE
- b. FALSE

Answer: a

18. Velocity is the speed at which the data is processed

- a. TRUE
- b. FALSE

Answer: b

19. have a structure but cannot be stored in a database.

- a. Structured
- b. Semi-Structured
- c. Unstructured
- d. None of these

Answer: b

20. _____ refers to the ability to turn your data useful for business.

- a. Velocity
- b. Variety
- c. Value
- d. Volume

Answer: c

21. Value tells the trustworthiness of data in terms of quality and accuracy.

- a. TRUE
- b. FALSE

Answer: b

22. Files are divided into sized Chunks.

- a. Static
- b. Dynamic
- c. Fixed
- d. Variable

Answer: c

23. _____ is an open source framework for storing data and running application on clusters of commodity hardware.

- a. HDFS
- b. Hadoop
- c. MapReduce
- d. Cloud

Answer: b

24. Hadoop MapReduce allows you to perform distributed parallel processing on large volumes of data quickly and efficiently: statement is True or False

- a. TRUE
- b. FALSE

Answer: a

25. In Relational database Management System the property of Scaling is applicable.

- a. TRUE
- b. FALSE

Answer: b

26. Which of the following options is not the example of NoSql ?

- a. Google
- b. Netflix
- c. Amazon
- d. CERN

Answer: c

27. Scalability and better performance of NoSQL is Achieved by sacrificing ACID Compatibility Is it TRUE?

- a. TRUE
- b. FALSE

Answer: a

28. For Scalability and better performance of No SQL is attained by compromising ACID Compatibility Is it TRUE?

- a. TRUE
- b. FALSE

Answer: a

29. _____ is a programming model for writing applications that can process Big Data in parallel on multiple nodes.

- a. HDFS
- b. MAP REDUCE
- c. HADOOP
- d. HIVE

Answer: b

30. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

- a. Decision Tree
- b. Regression
- c. Classification
- d. Random Forest

Answer: d

31. Data Set is the:

- a. Tweets stored in a flat file
- b. A collection of image files in a directory
- c. An extract of rows from a database table stored in a CSV formatted file
- d. All the above

Answer: d

32. Data analysis is the process of:

- a. Examining data to find facts
- b. Relationships,
- c. Patterns, insights and/or trends.
- d. All the above

Answer: d

33. What are the general categories of analytics that are distinguished by the results they produce:

- a. Descriptive analytics
- b. Diagnostic analytics
- c. Predictive analytics
- d. All the above

Answer: d

34. BI enables an organization to gain insight into the performance of an enterprise

- a. By analyzing data generated by its business processes and information systems.
- b. By examining data to find facts
- c. From relationships,
- d. All the above

Answer: a

35. Data variety refers to;

- a. Multiple schemas
- b. Multiple formats and types of data
- c. Multiple Data Models
- d. None of above

Answer: b

36. Unstructured Data Consists of:

- a. Text file, Audio Files.
- b. Video files, Text data
- c. Tagged Data
- d. a) and b)

Answer: d

37. Multiple internal and external data in the big data comes from the multiple sources as :

- a. Sensors, Social network sites
- b. Email, Xml, Multimedia
- c. a) and b)
- d. None of the above

Answer: c

38. Ingestion Layer Should have the capability to:

- a. validate, cleanse, transform, reduce
- b. integrate
- c. Preprocess the data
- d. a) and b)

Answer: d

39. 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
- c. Management of Hadoop clusters
- d. Collecting and storing unstructured data

Answer: a

40. What are the main components of Big Data?

- a. MapReduce
- b. HDFS
- c. YARN
- d. All of these

Answer: (d)

41. What are the different features of Big Data Analytics?

- a. Open-Source
- b. Scalability
- c. Data Recovery
- d. All the above

Answer: (d)

42. What are the four V's of Big Data?

- a. Volume
- b. Velocity
- c. Variety
- d. All the above

Answer: (d)

43. All of the following accurately describe Hadoop, EXCEPT:

- a. Open-source
- b. Real-time
- c. Java-based
- d. Distributed computing approach

Answer: (b)

44. _____ is general-purpose computing model and runtime system for distributed dataanalytics.

- a. Mapreduce
- b. Drill
- c. Oozie
- d. None of the above

Answer: (a)

45. The examination of large amounts of data to see what patterns or other useful information can be found is known as

- a. Data examination
- b. Information analysis
- c. Big data analytics
- d. Data analysis

Answer: (c)

46. Big data analysis does the following except

- a. Collects data
- b. Spreads data
- c. Organizes data
- d. Analyzes data

Answer: (b)

47. 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

Answer: (b)

48. The new source of big data that will trigger a Big Data revolution in the years to come is

- a. Business transactions
- b. Social media
- c. Transactional data and sensor data
- d. RDBMS

Answer: (c)

49. The unit of data that flows through a Flume agent is

- a. Log
- b. Row
- c. Event
- d. Record

Answer: (c)

50. Listed below are the three steps that are followed to deploy a Big Data Solution except

- a. Data Ingestion
- b. Data Processing
- c. Data dissemination
- d. Data Storage

Answer: (c)

51. Check below the best answer to “which industries employ the use of so-called “Big Data” in their day to day operations?

- a. Weather forecasting
- b. Marketing
- c. Healthcare
- d. All of the above

Answer: (d)

52. There are almost as many bits of information in the digital universe as there are stars in the actual universe?

- a. True
- b. False

Answer: (a)

53. The word ‘Big data’ was coined by

- a. Roger Mougallas
- b. John Philips
- c. Simon Woods
- d. Martin Green

Answer: (a)

54. The word ‘Big Data’ was coined in the year

- a. 2000
- b. 1970
- c. 1998
- d. 2005

Answer: (c)

55. Concerning the Forms of Big Data, which one of these is odd?

- a. Structured
- b. Unstructured
- c. Processed
- d. Semi-Structured

Answer: (c)

56. Big Data applications benefit the media and entertainment industry by

- a. Predicting what the audience wants
- b. Ad targeting
- c. Scheduling optimization
- d. All of the above

Answer: (d)

57. The feature of big data that refers to the quality of the stored data is _____

- a. Variety
- b. Volume
- c. Variability
- d. Veracity

Answer: (d)

58. _____ is a framework for performing remote procedure calls and data serialization.

- a. Drill
- b. BigTop
- c. Avro
- d. Chukwa

Answer: c

59. Which of the following is a characteristic of Big Data?

- a. Huge volume of data
- b. Complexity of data types and structures
- c. Speed of data creation and growth
- d. All of the mentioned

Answer: d

60. Concurrent access to shared data may result in_____

- a. Data consistency
- b. Data insecurity
- c. Data inconsistency
- d. None of the mentioned

Answer: c

61. Mutual exclusion implies that :

- a. If a process is executing in its critical section, then no other process must be executing in their critical sections
- b. If a process is executing in its critical section, then other processes must be executing in their critical sections
- c. If a process is executing in its critical section, then all the resources of the system must be blocked until it finishes execution
- d. None of the mentioned

Answer: a

62. In the memory hierarchy, as the speed of operation increases the memory size also increases.

- a. True
- b. False

Answer: b

63. To use a _____ network service, the service user first establishes a connection, uses the connection, and terminates the connection.

- a. Connection-oriented
- b. Connection-less
- c. Service-oriented
- d. Service-less

Answer: a

64. Which layer is responsible for the process-to process delivery ?

- a. Network
- b. Transport
- c. Application
- d. Physical

Answer: b

65. _____ refers to the biases, noise and abnormality in data, trustworthiness of data.

- a. Value
- b. Veracity
- c. Velocity
- d. Volume

Answer: b

66. _____ refers to the connectedness of big data.

- 1. Value
- 2. Veracity
- 3. Velocity
- 4. Valence

Answer: d

Unit2: Big data MCQ

1. Which one of the following is false about Hadoop?

- a. It is a distributed framework
- b. The main algorithm used in it is Map Reduce
- c. It runs with commodity hardware
- d. All are true

Answer: (d)

2. What license is Apache Hadoop distributed under?

- a. Apache License 2.0
- b. Shareware
- c. Mozilla Public License
- d. Commercial

Answer: (a)

3. Which of the following platforms does Apache Hadoop run on ?

- a. Bare metal
- b. Unix-like
- c. Cross-platform
- d. Debian

Answer: (c)

4. Apache Hadoop achieves reliability by replicating the data across multiple hosts and hence does not require _____ storage on hosts.

- a. Standard RAID levels
- b. RAID
- c. ZFS
- d. Operating system

Answer: Option (b)

5. Hadoop works in

- a. master-worker fashion
- b. master – slave fashion
- c. worker/slave fashion
- d. All of the mentioned

Answer: (b)

6. Which type of data Hadoop can deal with is

- a. Structured
- b. Semi-structured
- c. Unstructured
- d. All of the above

Answer: (d)

7. Which statement is false about Hadoop

- a. It runs with commodity hardware
- b. It is a part of the Apache project sponsored by the ASF
- c. It is best for live streaming of data
- d. None of the above

Answer: (c)

8. As compared to RDBMS, Apache 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.

Answer: (d)

9. Hadoop can be used to create distributed clusters, based on commodity servers, that provide low-cost processing and storage for unstructured data

- a. True
- b. False

Answer: (a)

10. _____ is a framework for performing remote procedure calls and data serialization.

- a. Drill
- b. BigTop
- c. Avro
- d. Chukwa

Answer: (c)

11. IBM and _____ have announced a major initiative to use Hadoop to support university courses in distributed computer programming.

- a. Google Latitude
- b. Android (operating system)
- c. Google Variations
- d. Google

Answer: (d)

12. What was Hadoop written in?

- a. Java (software platform)
- b. Perl
- c. Java (programming language)
- d. Lua (programming language)

Answer: (c)

13. Apache _____ is a serialization framework that produces data in a compact binary format.

- a. Oozie
- b. Impala
- c. Kafka
- d. Avro

Answer: (d)

14. Avro schemas describe the format of the message and are defined using _____

- a. JSON
- b. XML
- c. JS
- d. All of the mentioned

Answer: (a)

15. In which all languages you can code in Hadoop

- a. Java
- b. Python
- c. C++
- d. All of the above

Answer: (d)

16. All of the following accurately describe Hadoop, EXCEPT

- a. Open source
- b. Real-time
- c. Java-based
- d. Distributed computing approach

Answer: (b)

17. _____ has the world's largest Hadoop cluster.

- a. Apple
- b. Datamatics
- c. Facebook
- d. None of the mentioned

Answer: (c)

18. Which among the following is the default OutputFormat?

- a. SequenceFileOutputFormat
- b. LazyOutputFormat
- c. DBOutputFormat
- d. TextOutputFormat

Answer: (d)

19. Which of the following is not an input format in Hadoop?

- a. ByteInputFormat
- b. TextInputFormat
- c. SequenceFileInputFormat
- d. KeyValueInputFormat

Answer: (a)

20. What is the correct sequence of data flow in MapReduce?

- a. InputFormat
- b. Mapper
- c. Combiner
- d. Reducer
- e. Partitioner
- f. OutputFormat

- a. abcdfe
- b. abcedf
- c. acdefb
- d. abcdef

Answer: (b)

21. In which InputFormat tab character ('\t') is used

- a. KeyValueTextInputFormat
- b. TextInputFormat
- c. FileInputFormat
- d. SequenceFileInputFormat

Answer: (a)

Which among the following is true about SequenceFileInputFormat

- a. Key- byte offset. Value- It is the contents of the line
- b. Key- Everything up to tab character. Value- Remaining part of the line after tab character
- c. Key and value- Both are user defined
- d. None of the above

Answer:(c)

22. Which is key and value in TextInputFormat

- a. Key- byte offset Value- It is the contents of the line
- b. Key- Everything up to tab character Value- Remaining part of the line after tab character
- c. Key and value- Both are user defined
- d. None of the above

Answer: (a)

23. Which of the following are Built-In Counters in Hadoop?

- a. FileSystem Counters
- b. FileInputFormat Counters
- c. FileOutputFormat counters
- d. All of the above

Answer: (d)

24. Which of the following is not an output format in Hadoop?

- a. TextoutputFormat
- b. ByteoutputFormat
- c. SequenceFileOutputFormat
- d. DBOutputFormat

Answer: (b)

25. Is it mandatory to set input and output type/format in Hadoop MapReduce?

- a. Yes
- b. No

Answer: (b)

26. The parameters for Mappers are:

- a. text (input)
- b. LongWritable(input)
- c. text (intermediate output)
- d. All of the above

Answer: (d)

27. For 514 MB file how many InputSplit will be created

- a. 4
- b. 5
- c. 6
- d. 10

Answer: (b)

28. Which among the following is used to provide multiple inputs to Hadoop?

- a. MultipleInputs class
- b. MultipleInputFormat
- c. FileInputFormat
- d. DBInputFormat

Answer: (a)

29. The Mapper implementation processes one line at a time via _____ method.

- a. map
- b. reduce
- c. mapper
- d. reducer

Answer: (a)

30. The Hadoop MapReduce framework spawns one map task for each _____ generated by the InputFormat for the job.

- a. OutputSplit
- b. InputSplit
- c. InputSplitStream
- d. All of the mentioned

Answer: (b)

31. _____ 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

Answer: (a)

32. _____ part of the MapReduce is responsible for processing one or more chunks of data and producing the output results.

- a. Maptask
- b. Mapper
- c. Task execution
- d. All of the mentioned

Answer: (a)

33. _____ function is responsible for consolidating the results produced by each of the Map() functions/tasks.

- a. Map
- b. Reduce
- c. Reducer
- d. Reduced

Answer: (b)

34. The number of maps is usually driven by the total size of

- a. task
- b. output
- c. input
- d. none

Answer: (c)

35. The right number of reduces seems to be :

- a. 0.65
- b. 0.55
- c. 0.95
- d. 0.68

Answer: (c)

36. Mapper and Reducer implementations can use the _____ to report progress or just indicate that they are alive.

- a. Partitioner
- b. OutputCollector
- c. Reporter
- d. All of the mentioned

Answer: (c)

37. The major components in the Hadoop 2.0 are:

- a. 2
- b. 3
- c. 4
- d. 5

Answer: (b)

38. Which of the statement is true about PIG.

- a. Pig is also a data ware house system used for analysing the Big Data Stored in the HDFS
- b. .It uses the Data Flow Language for analysing the data
- c. a and b
- d. Relational Database Management System

Answer: (c)

39. Which of the following platforms does Hadoop run on?

- a. Bare metal
- b. Debian
- c. Cross-platform
- d. Unix-like

Answer: (c)

40. The Hadoop list includes the HBase database, the Apache Mahout _____ system, and matrix operations.

- a. Machine learning
- b. Pattern recognition
- c. Statistical classification
- d. Artificial intelligence

Answer: (a)

41. Which of the Node serves as the master and there is only one NameNode per cluster.

- a. Data Node
- b. NameNode
- c. Data block
- d. Replication

Answer: (b)

42. HDFS consists as the

- a. master-worker
- b. master node and slave node
- c. worker/slave
- d. all of the mentioned

Answer: (b)

43. The name node used, when the secondary node get failed is .

- a. Rack
- b. Data node
- c. Secondary node
- d. None of the mentioned

Answer: (c)

44. Which of the following scenario may not be a good fit for HDFS?

- a. HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file
- b. HDFS is suitable for storing data related to applications requiring low latency data access
- c. HDFS is suitable for storing data related to applications requiring low latency data access
- d. None of the mentioned

Answer: (a)

45. The need for data replication occurs:

- a. Replication Factor is changed
- b. DataNode goes down
- c. Data Blocks get corrupted
- d. All of the mentioned

Answer: (d)

46. HDFS uses only one language for implementation:

- a. C++
- b. Java
- c. Scala
- d. None of the Above

Answer: (d)

47. In YARN which node is responsible for managing the resources

- a. Data Node
- b. NameNode
- c. Resource Manager
- d. Replication

Answer: (c)

48. As Hadoop framework is implemented in Java, MapReduce applications are required to be written in Java Language

- a. True
- b. False

Answer: (b)

49. _____ maps input key/value pairs to a set of intermediate key/value pairs.

- a. Mapper
- b. Reducer
- c. Both Mapper and Reducer
- d. None of the mentioned

Answer: (d)

50. The number of maps is usually driven by the total size of _____

- a. Inputs
- b. Outputs
- c. Tasks
- d. None of the mentioned

Answer: (a)

51. which of the File system is used by HBase

- a. Hive
- b. Imphala
- c. Hadoop
- d. Scala

Answer: (c)

52. The information mapping data blocks with their corresponding files is stored in

- a. Namenode
- b. Datanode
- c. Job Tracker
- d. Task Tracker

Answer: (a)

53. In HDFS the files cannot be

- a. read
- b. deleted
- c. excuted
- d. archived

Answer: (d)

54. The datanode and namenode are, respectivley, which of the following?

- a. Slave and Master nodes
- b. Master and Worker nodes
- c. Both worker nodes
- d. both master nodes

Answer: (a)

55. Hadoop is a framework that works with a variety of related tools. Common cohorts include

- a. MapReduce, Hive and HBase
- b. MapReduce, MySQL and Google Apps
- c. MapReduce, Hummer and Iguana
- d. MapReduce, Heron and Trumpet

Answer: (a)

56. Hadoop was named after?

- a. Creator Doug Cuttings favorite circus act
- b. The toy elephant of Cuttings son
- c. Cuttings high school rock band
- d. A sound Cuttings laptop made during Hadoops development

Answer: (b)

57. All of the following accurately describe Hadoop, EXCEPT:

- a. Open source
- b. Java-based
- c. Distributed computing approach
- d. Real-time

Answer: (d)

58. Hive also support custom extensions written in:

- a. C
- b. C#
- c. C++
- d. Java

Answer: (d)

59. The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to :

- a. JSON
- b. XML
- c. SQL
- d. JQuery

Answer: (c)

60. In comparison to Rational DBMS, Hadoop

- a. A – Has higher data In
- b. B – Does ACID transactions
- c. C – IS suitable for read and write many times
- d. D – Works better on unstructured and semi-structured data.

Answer: (d)

61. The Files in HDFS are ment for

- a. Low latency data access
- b. Multiple writers and modifications at arbitrary offsets.
- c. Only append at the end of file
- d. Writing into a file only once.

Answer: (b)

62. The main role of the secondary namenode is to

- a. Copy the filesystem metadata from primary namenode.
- b. Copy the filesystem metadata from NFS stored by primary namenode
- c. Monitor if the primary namenode is up and running.
- d. Periodically merge the namespace image with the edit log.

Answer: (b)

63. The MapReduce algorithm contains three important tasks, namely _____.

- a. Splitting, mapping, reducing
- b. scanning, mapping, Reduction
- c. Map, Reduction, decluttering
- d. Cleaning, Map, Reduce

Answer: (a)

64. In how many stages the MapReduce program executes?

- a. 2
- b. 3
- c. 4
- d. 5

Answer: (d)

65. What is the function of Mapper in the MapReduce?

- a. Splitting the Data File
- b. Job
- c. Scanning the subblock of files
- d. PayLoad

Answer: (c)

66. Although the Hadoop framework is implemented in Java, MapReduce applications need be written in _____

- a. C
- b. C#
- c. Java
- d. None of the above

Answer: (d)

67. What is the meaning of commodity Hardware in Hadoop

- a. Very cheap hardware
- b. Industry standard hardware
- c. Discarded hardware
- d. Low specifications Industry grade hardware

Answer: (d)

68. Which of the following are true for Hadoop?

- a. It's a tool for Big Data analysis
- b. It supports structured and unstructured data analysis
- c. It aims for vertical scaling out/in scenarios
- d. Both (a) and (b)

Answer: (d)

69. Which of the following are the core components of Hadoop 2.0?

- a. HDFS
- b. Map Reduce
- c. YARN
- d. all the above

Answer: (d)

70. Pig Programming Language is used for real time queries.

- a. TRUE
- b. FALSE

Answer: (b)

71. What is the default HDFS block size for Hadoop 2.0?

- a. 32 MB
- b. 128 MB
- c. 128 KB
- d. 64 MB

Answer: (b)

72. Which of the following phases occur simultaneously ?

- a. Shuffle and Sort
- b. Reduce and Sort
- c. Shuffle and Map
- d. All of the mentioned

Answer: (a)

73. Major Components of Hadoop 1.0 are:

- a. HDFS and MapReduce
- b. Map Reduce, HDFS and YARN
- c. YARN and HDFS
- d. None of Above

Answer: (a)

Unit 3 Big data

1. A _____ serves as the master and there is only one NameNode per cluster.

- a. Data Node
- b. NameNode
- c. Data block
- d. Replication

Answer: b

2. Point out the correct statement.

- a. DataNode is the slave/worker node and holds the user data in the form of Data Blocks
- b. Each incoming file is broken into 32 MB by default
- c. Data blocks are replicated across different nodes in the cluster to ensure a low degree of fault tolerance
- d. None of the mentioned

Answer: a

3. HDFS works in a _____ fashion.

- a. master-worker
- b. master-slave
- c. worker/slave
- d. all of the mentioned

Answer: a

4. _____ NameNode is used when the Primary NameNode goes down.

- a. Rack
- b. Data
- c. Secondary
- d. None of the mentioned

Answer: c

5. Point out the wrong statement.

- a. Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level
- b. Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode
- c. User data is stored on the local file system of DataNodes
- d. DataNode is aware of the files to which the blocks stored on it belong to

Answer: d

6. Which of the following scenario may not be a good fit for HDFS?

- a. HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file
- b. HDFS is suitable for storing data related to applications requiring low latency data access
- c. HDFS is suitable for storing data related to applications requiring low latency data access
- d. None of the mentioned

Answer: a

7. The need for data replication can arise in various scenarios like _____

- a. Replication Factor is changed
- b. DataNode goes down
- c. Data Blocks get corrupted
- d. All of the mentioned

Answer: d

8. _____ is the slave/worker node and holds the user data in the form of Data Blocks.

- a. DataNode
- b. NameNode
- c. Data block

-21

20

- d. Replication

Answer: a

9. HDFS provides a command line interface called _____ used to interact with HDFS.

- a. "HDFS Shell"
- b. "FS Shell"
- c. "DFS Shell"
- d. None of the mentioned

Answer: b

10. HDFS is implemented in _____ programming language.

- a. C++
- b. Java
- c. Scala
- d. None of the mentioned

Answer: b

11. For YARN, the _____ Manager UI provides host and port information.

- a. Data Node
- b. NameNode
- c. Resource
- d. Replication

Answer: c

12. Point out the correct statement.

- a. The Hadoop framework publishes the job flow status to an internally running web server on the master nodes of the Hadoop cluster
- b. Each incoming file is broken into 32 MB by default
- c. Data blocks are replicated across different nodes in the cluster to ensure a low degree of fault tolerance
- d. None of the mentioned

Answer: a

13. For _____ the HBase Master UI provides information about the HBase Master uptime.

- a. HBase
- b. Oozie
- c. Kafka
- d. All of the mentioned

Answer: a

14. During start up, the _____ loads the file system state from the fsimage and the edits log file.

- a. DataNode
- b. NameNode
- c. ActionNode
- d. None of the mentioned

Answer: b

15. What is the utility of the HBase ?

- a. It is the tool for Random and Fast Read/Write operations in Hadoop
- b. Acts as Faster Read only query engine in Hadoop
- c. It is MapReduce alternative in Hadoop
- d. It is Fast MapReduce layer in Hadoop

Answer: a

16. What is Hive used as?

- a. Hadoop query engine
- b. MapReduce wrapper
- c. Hadoop SQL interface
- d. All of the above

Answer: d

17. What is the default size of the HDFS block ?

- a. 32 MB
- b. 64 KB
- c. 128 KB
- d. 64 MB

Answer: d

18. In the HDFS what is the default replication factor of the Data Node?

- a. 4
- b. 1
- c. 3
- d. 2

Answer: c

19. What is the protocol name that is used to create replica in HDFS?

- a. Forward protocol
- b. Sliding Window Protocol
- c. HDFS protocol
- d. Store and Forward protocol

Answer: c

20. HDFS data blocks can be read in parallel.

- a. True
- b. False

Answer: a

21. Which of the following is fact about combiners in HDFS?

- a. Combiners can be used for mapper only job
- b. Combiners can be used for any Map Reduce operation
- c. Mappers can be used as a combiner class
- d. Combiners are primarily aimed to improve Map Reduce performance
- e. Combiners can't be applied for associative operations

Answer: d

22. In HDFS the Distributed Cache is used in which of the following

- a. Mapper phase only
- b. Reducer phase only
- c. In either phase, but not on both sides simultaneously
- d. In either phase

Answer: d

23. Which of the following type of joins can be performed in Reduce side join operation?

- a. Equi Join
- b. Left Outer Join
- c. Right Outer Join
- d. Full Outer Join
- e. All of the above

Answer: e

24. A Map reduce function can be written:

- a. Java
- b. Ruby
- c. Python
- d. Any Language which can read from input stream

Answer: d

25. In the map is there any input format?

- a. Yes, but only in Hadoop 0.22+.
- b. Yes, there is a special format for map files.
- c. No, but sequence file input format can read map files.
- d. Both 2 and 3 are correct answers

Answer: c

26. Which MapReduce phase is theoretically able to utilize features of the underlying file system in order to optimize parallel execution?

- a. Split
- b. Map
- c. Combine
- d. Reduce

Answer: a

27. Which method of the FileSystem object is used for reading a file in HDFS

- a. open()
- b. access()
- c. select()
- d. None of the above

Answer: a

28. The world's largest Hadoop cluster.

- a. Apple
- b. Facebook
- c. Datamatics
- d. None of the mentioned

Answer: b

29. The Big Data Tackles Facebook are based on _____ Hadoop.

- a. 'Project Data
- b. 'Prism'
- c. 'Project Big'
- d. 'Project Prism'

Answer: d

30. Which SequenceFile are present in Hadoop I/O

?

- a. 2
- b. 8
- c. 9
- d. 3

Answer: c

31. slowest compression technique is _____

- a. Bzip2
- b. LZO
- c. Gzip
- d. All of the mentioned

Answer: c

32. Which of the following is a typically compresses files which are best available techniques. 10% to 15 %.

- a. Bzip2
- b. LZO
- c. Gzip
- d. both Dand C

Answer: a

33. Which of the following is provides search technology? and Java-based indexing

- a. Solr
- b. Lucy
- c. Lucene Core
- d. None of these

Answer: c

34. Are defined with Avro schemas _____

- a. JAVA
- b. XML
- c. All of the mentioned
- d. JSON

Answer: d

35. _____ of the field is used to Thrift resolves possible conflicts.

- a. Name
- b. UID
- c. Static number
- d. All of the mentioned

Answer: c

36. _____ layer of is said to be the future Hadoop.

Avro.

- a. RMC
- b. RPC
- c. RDC
- d. All of the mentioned

Answer: b

37. High storage density Which of the following has high storage density?

- a. RAM_DISK
- b. ARCHIVE
- c. ROM_DISK
- d. All of the mentioned

Answer: b

38. HDFS provides a command line interface called _____ used to interact with HDFS.

- a. "HDFS Shell"
- b. "FS Shell"
- c. "DFS Shell"
- d. None of the mentioned

Answer: b

39. Which format from the given format is more compression-aggressive?

- a. Partition Compressed
- b. Record Compressed
- c. Block-Compressed
- d. Uncompressed

Answer: c

40. Avro schemas describe the format of the message and are defined using ____

- a. JSON
- b. XML
- c. JS
- d. All of the mentioned

Answer: b

41. Which editor is used for editing files in HDFS

- a. Vi Editor
- b. Python editor
- c. DOS editor
- d. DEV C++ Editor

Answer: a

42. Command to view the directories and files in specific directory:

- a. Ls
- b. Fs -ls
- c. Hadoop fs -ls
- d. Hadoop fs

Answer: a

43. Which among the following is correct?

S1: MapReduce is a programming model for data processing

S2: Hadoop can run MapReduce programs written in various languages

S3: MapReduce programs are inherently parallel

- a. S1 and S2
- b. S2 and S3
- c. S1 and S3
- d. S1, S2 and S3

Answer: d

44. Mapper class is

- a. generic type
- b. abstract type
- c. static type
- d. final

Answer: a

45. Which package provides the basic types of Hadoop?

- a. org.apache.hadoop.io
- b. org.apache.hadoop.util
- c. org.apache.hadoop.type
- d. org.apache.hadoop.lang

Answer: a

46. Which among the following does the Job control in Hadoop?

- a. Mapper class
- b. Reducer class
- c. Task class
- d. Job class

Answer: d

47. Hadoop runs the jobs by dividing them into

- a. maps
- b. tasks
- c. individual files
- d. None of these

Answer: b

48. Which are the two nodes that control the job execution process of Hadoop?

- a. Job Tracker and Task Tracker
- b. Map Tracker and Reduce Tracker
- c. Map Tracker and Job Tracker
- d. Map Tracker and Task Tracker

Answer: a

49. Which among the following schedules tasks to be run?

- a. Job Tracker
- b. Task Tracker
- c. Job Scheduler
- d. Task Controller

Answer: A

50. What are fixed size pieces of MapReduce job called?

- a. records
- b. splits
- c. tasks
- d. maps

Answer: b

51. Where is the output of map tasks written?

- a. local disk
- b. HDFS
- c. File System
- d. secondary storage

Answer: a

52. Which among the following is responsible for processing one or more chunks of data and producing the output results.

- a. Maptask
- b. jobtask
- c. Mapper class
- d. Reducetask

Answer: a

53. Which acts as an interface between Hadoop and the program written?

- a. Hadoop Cluster
- b. Hadoop Streams
- c. Hadoop Sequencing
- d. Hadoop Streaming

Answer: d

54. What are Hadoop Pipes?

- a. Java interface to Hadoop MapReduce
- b. C++ interface to Hadoop MapReduce
- c. Ruby interface to Hadoop MapReduce
- d. Python interface to Hadoop MapReduce

Answer: b

55. What does Hadoop Common Package contain?

- a. war files
- b. msi files
- c. jar files
- d. exe files

Answer: c

56. Which among the following is the master node?

- a. Name Node
- b. Data Node
- c. Job Node
- d. Task Node

Answer: a

57. Which among the following is the slave node?

- a. Name Node
- b. Data Node
- c. Job Node
- d. Task Node

Answer: b

58. Which acts as a checkpoint node in HDFS?

- a. Name Node
- b. Data Node
- c. Secondary Name Node
- d. Secondary Data Node

Answer: c

59. Which among the following holds the location of data?

- a. Name Node
- b. Data Node
- c. Job Tracker
- d. Task Tracker

Answer: a

60. What is the process of applying the code received by the JobTracker on the file called?

- a. Naming
- b. Tracker
- c. Mapper
- d. Reducer

Answer: a

61. In which mode should Hadoop run in order to run pipes job?

- a. distributed mode
- b. centralized mode
- c. pseudo distributed mode
- d. parallel mode

Answer: b

62. Which of the following are correct? S1:

Namespace volumes are independent of each other S2: Namespace volumes are managed by namenode

- a. S1 only
- b. S2 only
- c. Both S1 and S2
- d. Neither S1 nor S2

Answer: c

63. Which among the following architectural changes need to attain High availability in HDFS?

- a. Clients must be configured to handle namenode failover
- b. Datanodes must send block reports to both namenodes since the block mappings are stored in a namenode's memory, and not on disk
- c. namenodes must use highly-available shared storage to share the edit log
- d. All of the above

Answer: d

64. Which controller in HDFS manages the transition from the active namenode to the standby?

- a. failover controller
- b. recovery controller
- c. failsafe controller
- d. fencing controller

Answer: a

65. Which among the following is not an fencing mechanism employed by system in HDFS?

- a. killing the namenode's process
- b. disabling namenode's network port via a remote management command
- c. revoking namenode's access to the shared storage directory
- d. None of the above

Answer: d

66. What is the value of the property dfs.replication et in case of pseudo distributed mode?

- a. 0
- b. 1
- c. null
- d. yes

Answer: b

67. What is the minimum amount of data that a disk can read or write in HDFS?

- a. block size
- b. byte size
- c. heap
- d. None

Answer: a

68. Which HDFS command checks file system and lists the blocks?

- a. hfsck
- b. fcsk
- c. fblock
- d. fsck

Answer: d

69. What is an administered group used to manage cache permissions and resource usage?

- a. Cache pools
- b. block pool
- c. Namenodes
- d. HDFS Cluster

Answer: a

70. Which object encapsulates a client or server's configuration?

- a. File Object
- b. Configuration object
- c. Path Object
- d. Stream Object

Answer: b

71. Which interface permits seeking to a position in the file and provides a query method for the current offset from the start of the file?

DataStream

- a. Seekable
- b. PositionedReadable
- c. Progressable

Answer: b

72. Which method is used to list the contents of a directory?

- a. listFiles
- b. listContents
- c. listStatus
- d. listPaths

Answer: C

73. What is the operation that use wildcard characters to match multiple files with a single expression called?

- a. globbing
- b. pattern matching
- c. regex
- d. regexfilter

74. What does the globStatus() methods return?

- a. an array of FileStatus objects
- b. an array of ListStatus objects
- c. an array of PathStatus objects
- d. an array of FilterStatus objects

Answer: a

75. What does the glob question mark(?) matches?

- a. zero or more characters
- b. one or more characters
- c. a single character
- d. metacharacter

Answer: c

76. Which method on FileSystem is used to permanently remove files or directories?

- a. remove()
- b. rm()
- c. del()
- d. delete()

Answer: d

77. Which streams the packets to the first datanode in the pipeline?

- a. DataStreamer
- b. FileStreamer
- c. InputStreamer
- d. PathStreamer

Answer: a

78. Which queue is responsible for asking the namenode to allocate new blocks by picking a list of suitable datanodes to store the replicas?

- a. ack queue
- b. data queue
- c. path queue
- d. stream queue

Answer: b

79. Which command is used to copy files/directories?

- a. distcp
- b. hcp
- c. copy
- d. cp

Answer: a

80. Which flag is used with distcp to delete any files or directories from the destination?

- a. -remove
- b. -rm
- c. -del
- d. -delete

Answer: d

Unit4: Big data MCQ

1. Which among the following is Hadoop's cluster resource management system?

- a. GLOB
- b. YARN
- c. ARM
- d. SPARK

Answer: b

2. Which of the following processing framework interacts with YARN directly?

- a. Pig
- b. Hive
- c. Crunch
- d. None of these

Answer: D

3. Which of the following processing frameworks run on MapReduce?

- a. Pig
- b. Hive
- c. Crunch
- d. All of the above

Answer: d

4. Which among the following are the core services of YARN?

- a. resource manager and node manager
- b. namenode and datanode
- c. data manager and resource manager
- d. data manager and application manager

Answer: a

5. Which constraints can be used to request a container on a specific node or rack, or anywhere on the cluster in YARN?

- a. Container constraints
- b. Space constraints
- c. Locality constraints
- d. Resource constraints

Answer: c

6. Which among the following can be used to model YARN applications?

- a. one application per user job
- b. run one application per workflow
- c. long-running application that is shared by different users
- d. All of the above

Answer: d

7. Which follows one application per user job model?

- a. MapReduce
- b. Spark
- c. Apache Slider
- d. Samza

Answer: a

8. Which application runs per user session?

- a. MapReduce
- b. Spark
- c. Apache Slider
- d. None of the above

Answer: b

9. Which among the following has a long-running application master for launching other applications on the cluster?

- a. MapReduce
- b. Spark
- c. Apache Slider
- d. None of the above

Answer: c

10. Which among the following can be used for stream processing?

- a. Spark
- b. Samza
- c. Storm
- d. All of the above

Answer: d

11. Which provides a simple programming model for developing distributed applications on YARN?

- a. Apache Slider
- b. Apache Twill
- c. Spark
- d. Tez

Answer: b

12. Which among the following statements are true with respect to Apache Twill? S1: Twill supports real-time logging S2: Allows the usage of a Java Runnable interface

- a. S1 only
- b. S2 only
- c. Both S1 and S2
- d. Neither S1 nor S2

Answer: c

13. Which daemon control the job execution process in MapReduce 1?

- a. jobtracker
- b. tasktrackers
- c. Both jobtracker and tasktrackers
- d. Name node and data node

Answer: c

14. Which among the following coordinates all the jobs run on the system by scheduling tasks in MapReduce 1?

- a. jobtracker
- b. tasktrackers
- c. data node
- d. Name node

Answer: a

15. Which of the following which keeps a record of the overall progress of each job in MapReduce 1?

- a. jobtracker
- b. tasktrackers
- c. data node
- d. Name node

Answer: a

16. Which among the following run tasks and send progress reports in MapReduce 1?

- a. jobtracker
- b. tasktrackers
- c. data node
- d. Name node

Answer: b

17. Choose the tasks of jobtracker in MapReduce 1?

- a. job scheduling
- b. task progress monitoring
- c. task bookkeeping
- d. All of the above

Answer: d

18. Which is responsible for storing job history in MapReduce 1?

- a. jobtracker
- b. tasktrackers
- c. data node
- d. Name node

Answer: a

19. In YARN, the responsibility of jobtracker is handled by

- a. Resource manager
- b. application master
- c. timeline server
- d. All of the above

Answer: d

20. In YARN, the responsibility of tasktracker is handled by

- a. Resource manager
- b. application master
- c. timeline server
- d. Node manager

Answer: d

21. Which stores the application history in YARN?

- a. Resource manager
- b. application master
- c. timeline server
- d. Node manager

Answer: c

22. Which among the following are the features of YARN?

- a. Scalability
- b. Multitenancy
- c. Availabilit
- d. All of the above

Answer: d

23. Which among the following schedulers available in YARN?

- a. FIFO
- b. Shortest Job First
- c. Round Robin
- d. Shortest Remaining Time

Answer: a

24. Which are/is the schedulers available in YARN?

- a. FIFO
- b. Capacity
- c. Fair Schedulers
- d. All of the above

Answer: d

25. Which among the following schedulers attempts to allocate resources so that all running applications get the same share of resources in YARN

- a. FIFO
- b. Capacity
- c. Fair Schedulers
- d. Round Robin

Answer: c

26. Which among the following schedulers provides queue elasticity in YARN?

- a. FIFO
- b. Capacity
- c. Fair Schedulers
- d. Round Robin

Answer: b

27. Which among the following schedulers in YARN is used by default?

- FIFO
- Capacity
- Fair Schedulers
- Round Robin

Answer: b

28. In which xml, is the default configuration of schedulers to be changed?

- a. yarn-site.xml
- b. config.xml
- c. scheduler.xml
- d. yarn-scheduler.xml

Answer: a

29. Which among the following queue scheduling policies are/is supported by Fair Schedulers in YARN?

- a. FIFO
- b. Dominant Resource Fairness
- c. preemption
- d. All of the above

Answer: d

30. Which holds the list of rules for queue placement in Fair Scheduling?

- a. queuePlacementPolicy
- b. rulePlacementolicy
- c. scheduleQueuePolicy
- d. schedulingPolicy

Answer: a

31. Which of the setting is used to set preemption globally?

- a. yarn.scheduler.fair.preemption = true
- b. yarn.scheduler.preemption = true
- c. yarn.scheduler.global.preemption = true
- d. yarn.scheduler.enable.preemption = true

Answer: a

32. Which among the following supports delay scheduling?

- a. FIFO
- b. Capacity Scheduler
- c. Fair Scheduler
- d. Both Capacity and Fair Scheduler

Answer: d

33. What is the default period of heartbeat request sent by node manager?

- a. one per millisecond
- b. one per second
- c. one per minute
- d. one per nanosecond

Answer: b

34. Which error detection code is used in HDFS?

- a. CRC-32
- b. CRC-32C
- c. SHA
- d. SHA-1

Answer: b

35. CRC-32C has the storage overhead

- a. less than 1%
- b. less than 5%
- c. less than 10%
- d. less than 2.5%

Answer: a

36. The heartbeat signal are sent from

- a. Jobtracker to Tasktracker
- b. Tasktracker to Job tracker
- c. Jobtracker to namenode
- d. Tasktracker to namenode

Answer: b

37. Spark was initially started by _____ at UC Berkeley AMPLab in 2009.

- a. Mahek Zaharia
- b. Matei Zaharia
- c. Doug Cutting
- d. Stonebraker

Answer: (b)

38. _____ is a component on top of Spark Core.

- a. Spark Streaming
- b. Spark SQL
- c. RDDs
- d. All of the mentioned

Answer: (b)

39. Spark SQL provides a domain-specific language to manipulate _____ in Scala, Java, or Python.

- a. Spark Streaming
- b. Spark SQL
- c. RDDs
- d. All of the mentioned

Answer: (c)

40. _____ leverages Spark Core fast scheduling capability to perform streaming analytics.

- a. MLlib
- b. Spark Streaming
- c. GraphX
- d. RDDs

Answer: (b)

41. _____ is a distributed machine learning framework on top of Spark.

- a. MLlib
- b. Spark Streaming
- c. GraphX
- d. RDDs

Answer: (a)

42. Users can easily run Spark on top of Amazon's _____

- a. Infosphere
- b. EC2
- c. EMR
- d. None of the mentioned

Answer: (b)

43. Which of the following can be used to launch Spark jobs inside MapReduce?

- a. SIM
- b. SIMR
- c. SIR
- d. RIS

Answer: (b)

44. Which of the following language is not supported by Spark?

- a. Java
- b. Pascal
- c. Scala
- d. Python

Answer: (b)

45. Spark is packaged with higher level libraries, including support for _____ queries.

- a. SQL
- b. C
- c. C++
- d. None of the mentioned

Answer: (a)

46. Spark includes a collection over _____ operators for transforming data and familiar data frame APIs for manipulating semi structured data.

- a. 50
- b. 60
- c. 70
- d. 80

Answer: (d)

47. Spark is engineered from the bottom-up for performance, running _____ faster than Hadoop by exploiting in memory computing and other optimizations.

- a. 100x
- b. 150x
- c. 200x
- d. None of the mentioned

Answer: (a)

48. Spark powers a stack of high-level tools including Spark SQL, MLlib for _____

- a. regression models
- b. statistics
- c. machine learning
- d. reproductive research

Answer: (c)

49. For Multiclass classification problem which algorithm is not the solution?

- a. Naive Bayes
- b. Random Forests
- c. Logistic Regression
- d. Decision Trees

Answer: (d)

50. Which of the following is a tool of Machine Learning Library?

- a. Persistence
- b. Utilities like linear algebra, statistics
- c. Pipelines
- d. All of the above

Answer: (d)

51. Which of the following is true for Spark core?

- a. It is the kernel of Spark
- b. It enables users to run SQL / HQL queries on the top of Spark.
- c. It is the scalable machine learning library which delivers efficiencies
- d. Improves the performance of iterative algorithm drastically.

Answer: (a)

52. Which of the following is true for Spark MLlib?

- a. Provides an execution platform for all the Spark applications
- b. It is the scalable machine learning library which delivers efficiencies
- c. enables powerful interactive and data analytics application across live streaming data
- d. All of the above

Answer: (b)

53. Which of the following is true for RDD?

- a. We can operate Spark RDDs in parallel with a low-level API
- b. RDDs are similar to the table in a relational database
- c. It allows processing of a large amount of structured data
- d. It has built-in optimization engine

Answer: (a)

54. RDD is fault-tolerant and immutable

- a. True
- b. False

Answer: (a)

55. The read operation on RDD is

- a. Fine-grained
- b. Coarse-grained
- c. Either fine-grained or coarse-grained
- d. Neither fine-grained nor coarse-grained

Answer: (c)

56. The write operation on RDD is

- a. Fine-grained
- b. Coarse-grained
- c. Either fine-grained or coarse-grained
- d. Neither fine-grained nor coarse-grained

Answer: (b)

57. Is it possible to mitigate stragglers in RDD?

- a. Yes
- b. No

Answer: (a)

58. Fault Tolerance in RDD is achieved using

- a. Immutable nature of RDD
- b. DAG (Directed Acyclic Graph)
- c. Lazy-evaluation
- d. None of the above

Answer: (b)

59. What is action in Spark RDD?

- a. The ways to send result from executors to the driver
- b. Takes RDD as input and produces one or more RDD as output.
- c. Creates one or many new RDDs
- d. All of the above

Answer: (a)

60. The shortcomings of Hadoop MapReduce was overcome by Spark RDD by

- a. Lazy-evaluation
- b. DAG
- c. In-memory processing
- d. All of the above

Answer: (d)

61. Spark is developed in which language

- a. Java
- b. Scala
- c. Python
- d. R

Answer: (b)

62. Which of the following is not a component of the Spark Ecosystem?

- (a) Sqoop
- (b) GraphX
- (c) MLlib
- (d) BlinkDB

Answer: (a)

63. Which of the following algorithm is not present in MLlib?

- a. Streaming Linear Regression
- b. Streaming KMeans
- c. Tanimoto distance
- d. None of the above

Answer: (c)

64. Which of the following is not the feature of Spark?

- a. Supports in-memory computation
- b. Fault-tolerance
- c. It is cost-efficient
- d. Compatible with other file storage system

Answer: (c)

65. Which of the following is the reason for Spark being Speedy than MapReduce?

- a. DAG execution engine and in-memory computation
- b. Support for different language APIs like Scala, Java, Python and R
- c. RDDs are immutable and fault-tolerant
- d. None of the above

Answer: (a)

66. Which of the following is true for RDD?

- a. RDD is a programming paradigm
- b. RDD in Apache Spark is an immutable collection of objects
- c. It is a database
- d. None of the above

Answer: (b)

67. Which of the following is a tool of the Machine Learning Library?

- a. Persistence
- b. Utilities like linear algebra, statistics
- c. Pipelines
- d. All of the above

Answer: (d)

68. _____ is a online NoSQL developed by Cloudera.

- a. HCatalog
- b. Hbase
- c. Imphala
- d. Oozie

Answer: (b)

69. Which of the following is not a NoSQL database?

- a. SQL Server
- b. MongoDB
- c. Cassandra
- d. None of the mentioned

Answer: (a)

70. Which of the following is a NoSQL Database Type?

- a. SQL
- b. Document databases
- c. JSON
- d. All of the mentioned

Answer: (b)

71. Which of the following is a wide-column store?

- a. Cassandra
- b. Riak
- c. MongoDB
- d. Redis

Answer: (a)

72. "Sharding" a database across many server instances can be achieved with _

- a. LAN
- b. SAN
- c. MAN
- d. All of the mentioned

Answer: (b)

73. Most NoSQL databases support automatic _____ meaning that you get high availability and disaster recovery.

- a. processing
- b. scalability
- c. replication
- d. all of the mentioned

Answer: (c)

74. Which of the following are the simplest NoSQL databases?

- a. Key-value
- b. Wide-column
- c. Document
- d. All of the mentioned

Answer: (a)

75. _____ stores are used to store information about networks, such as social connections.

- a. Key-value
- b. Wide-column
- c. Document
- d. Graph

Answer: (d)

76. NoSQL databases is used mainly for handling large volumes of _____ data.

- a. unstructured
- b. structured
- c. semi-structured
- d. all of the mentioned

Answer: (a)

77. Which of the following language is MongoDB written in?

- a. Javascript
- b. C
- c. C++
- d. All of the mentioned

Answer: (d)

78. Point out the correct statement.

- a. MongoDB is classified as a NoSQL database
- b. MongoDB favors XML format more than JSON
- c. MongoDB is column-oriented database store
- d. All of the mentioned

Answer: (a)

79. Which of the following format is supported by MongoDB?

- a. SQL
- b. XML
- c. BSON
- d. All of the mentioned

Answer: (c)

80. NoSQL was designed with security in mind, so developers or security teams don't need to worry about implementing a security layer. Is it true or false?

- a. True
- b. False

Answer: (b)

81. Which of the following is not a reason NoSQL has become a popular solution for some organizations?

- a. Better scalability
- b. Improved ability to keep data consistent
- c. Faster access to data than relational database management systems (RDBMS)
- d. More easily allows for data to be held across multiple servers

Answer: (b)

82. NoSQL prohibits structured query language (SQL). Is it True or False?

- a. True
- b. False

Answer: (b)

83. When is it best to use a NoSQL database?

- a. When providing confidentiality, integrity, and availability is crucial
- b. When the data is predictable
- c. When the retrieval of large quantities of data is needed
- d. When the retrieval speed of data is not critical

Answer: (c)

84. Which of the following companies developed NoSQL database Apache Cassandra?

- a. LinkedIn
- b. Twitter
- c. MySpace
- d. Facebook

Answer: (d)

85. NoSQL databases are most often referred to as:

- a. Relational
- b. Distributed
- c. Object-oriented
- d. Network

Answer: (b)

86. SQL databases are:

- a. Horizontally scalable
- b. Vertically scalable
- c. Either horizontally or vertically scalable
- d. They don't scale

Answer: (b)

87. Which of the following is not an example of a NoSQL database?

- a. CouchDB
- b. MongoDB
- c. HBase
- d. PostgreSQL

Answer: (d)

88. SQL command types include data manipulation language (DML) and data definition language (DDL).

- a. True
- b. False

Answer: (a)

89. _____ systems are scale-out file-based (HDD) systems moving to more uses of memory in the nodes.

- a. NoSQL
- b. NewSQL
- c. SQL
- d. All of the mentioned

Answer: (a)

90. Point out the correct statement.

- a. Hadoop is ideal for the analytical, post operational, data-warehouse-ish type of workload
- b. HDFS runs on a small cluster of commodity class nodes
- c. NEWSQL is frequently the collection point for big data
- d. None of the mentioned

Answer: (a)

91. Which is an advantage of NewSQL ?

- a. Less complex applications, greater consistency.
- b. Convenient standard tooling.
- c. SQL influenced extensions.
- d. All of the mentioned

Answer: (d)

92. Following represent column in NoSQL

_____.

- a. Database
- b. Field
- c. Document
- d. Collection

Answer:(b)

93. What is the aim of NoSQL?

- a. NoSQL provides an alternative to SQL databases to store textual data.
- b. NoSQL databases allow storing non structured data.
- c. NoSQL is not suitable for storing structured data.
- d. NoSQL is a new data format to store large datasets.

Answer: (d)

94. Which of the following is not a feature for NoSQL databases?

- a. Data can be easily held across multiple servers
- b. Relational Data
- c. Scalability
- d. Faster data access than SQL databases

Ans : b

95. Which of the following statement is correct with respect to mongoDB?

- a. MongoDB is a NoSQL Database
- b. MongoDB used XML over JSON for data exchange
- c. MongoDB is not scalable
- d. All of the above

Ans : a

96. Which of the following represent column in mongoDB?

- a. document
- b. database
- c. collection
- d. field

Ans : d

97. The system generated _id field is?

- a. A 12 byte hexadecimal value
- b. A 16 byte octal value
- c. A 12 byte decimal value
- d. A 10 bytes binary value

Ans : a

98. Which of the following true about mongoDB?

- a. MongoDB is a cross-platform
- b. MongoDB is a document oriented database
- c. MongoDB provides high performance
- d. All of the above

Ans : d

99. Collection is a group of MongoDB __?

- a.Database
- b. Document
- c.Field
- d. None of the above

Ans : b

100. A developer want to develop a database for LFC system where the data stored is mostly in similar manner. Which database should use?

- a. Relational
- b. NoSQL
- c. Both A and B can be used
- d. None of the above

Ans : b

101. Documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data is known as ?

- a. dynamic schema
- b. mongod
- c. mongo
- d. Embedded Documents

Ans : a

102. Instead of Primary Key mongoDB use?

- a. Embedded Documents
- b. Default key _id
- c. mongod
- d. mongo

Ans : B

Unit5: Big data MCQ

1. A _____ serves as the master and there is only one NameNode per cluster.

- a. Data Node
- b. NameNode
- c. Data block
- d. Replication

Answer: (b)

2. Point out the correct statement.

- a. DataNode is the slave/worker node and holds the user data in the form of Data Blocks
- b. Each incoming file is broken into 32 MB by default
- c. Data blocks are replicated across different nodes in the cluster to ensure a low degree of fault tolerance
- d. None of the mentioned

Answer: (a)

3. HDFS works in a _____ fashion.

- a. master-worker
- b. master-slave
- c. worker/slave
- d. all of the mentioned

Answer: (a)

4. _____ NameNode is used when the Primary NameNode goes down.

- a. Rack
- b. Data
- c. Secondary
- d. None of the mentioned

Answer: (c)

5. Which of the following scenario may not be a good fit for HDFS?

- a. HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file
- b. HDFS is suitable for storing data related to applications requiring low latency data access
- c. HDFS is suitable for storing data related to applications requiring low latency data access
- d. None of the mentioned

Answer: (a)

6. _____ is the slave/worker node and holds the user data in the form of Data Blocks.

- a. DataNode
- b. NameNode
- c. Data block
- d. Replication

Answer: (a)

7. HDFS provides a command line interface called _____ used to interact with HDFS.

- a. "HDFS Shell"
- b. "FS Shell"
- c. "DFS Shell"
- d. None of the mentioned

Answer: (b)

8. For YARN, the _____ Manager UI provides host and port information.

- a. Data Node
- b. NameNode
- c. Resource
- d. Replication

Answer: (c)

9. During start up, the _____ loads the file system state from the fsimage and the edits log file.

- a. DataNode
- b. NameNode
- c. ActionNode
- d. None of the mentioned

Answer: (b)

10. In HDFS the files cannot be

- a. read
- b. deleted
- c. executed
- d. Archived

Answer: (c)

11. Which of the following command sets the value of a particular configuration variable (key)?

- a. set -v
- b. set =
- c. set
- d. reset

Answer: (b)

12. Which of the following operator executes a shell command from the Hive shell?

- a. |
- b. !
- c. ^
- d. +

Answer: (b)

13. Hive specific commands can be run from Beeline, when the Hive _____ driver is used.

- a. ODBC
- b. JDBC
- c. ODBC-JDBC
- d. All of the Mentioned

Answer: Option (b)

14. Which of the following data type is supported by Hive?

- a. map
- b. record
- c. string
- d. enum

Answer: (d)

15. Avro-backed tables can simply be created by using _____ in a DDL statement.

- a. "STORED AS AVRO"
- b. "STORED AS HIVE"
- c. "STORED AS AVROHIVE"
- d. "STORED AS SERDE"

Answer: (a)

16. Types that may be null must be defined as

a _____ of that type and Null within Avro.

- a. Union
- b. Intersection
- c. Set
- d. All of the mentioned

Answer: (a)

17. _____ is interpolated into the quotes to correctly handle spaces within the schema.

- a. \$SCHEMA
- b. \$ROW
- c. \$SCHEMASPACES
- d. \$NAMESPACES

Answer: (a)

18. _____ was designed to overcome the limitations of the other Hive file formats.

- a. ORC
- b. OPC
- c. ODC
- d. None of the mentioned

Answer: (a)

19. An ORC file contains groups of row data called _____

- a. postscript
- b. stripes
- c. script
- d. none of the mentioned

Answer: (b)

20. HBase is a distributed _____ database built on top of the Hadoop file system.

- a. Column-oriented
- b. Row-oriented
- c. Tuple-oriented
- d. None of the mentioned

Answer: (a)

21. HBase is _____ defines only column families.

- a. Row Oriented
- b. Schema-less
- c. Fixed Schema
- d. All of the mentioned

Answer: (b)

22. The _____ Server assigns regions to the region servers and takes the help of Apache ZooKeeper for this task.

- a. Region
- b. Master
- c. Zookeeper
- d. All of the mentioned

Answer: (b)

23. Which of the following command provides information about the user?

- a. status
- b. version
- c. whoami
- d. user

Answer: (c)

24. _____ command fetches the contents of a row or a cell.

- a. select
- b. get
- c. put
- d. none of the mentioned

Answer: (b)

25. HBaseAdmin and _____ are the two important classes in this package that provide DDL functionalities.

- a. HTableDescriptor
- b. HDescriptor
- c. HTable
- d. HTabDescriptor

Answer: (a)

26. The minimum number of row versions to keep is configured per column family via _____

- a. HBaseDecriptor
- b. HTabDescriptor
- c. HColumnDescriptor
- d. All of the mentioned

Answer: (c)

27. HBase supports a _____ interface via Put and Result.

- a. "bytes-in/bytes-out"
- b. "bytes-in"
- c. "bytes-out"
- d. none of the mentioned

Answer: (a)

28. One supported data type that deserves special mention are _____

- a. money
- b. counters
- c. smallint
- d. tinyint

Answer: (b)

29. _____ does re-write data and pack rows into columns for certain time-periods.

- a. OpenTS
- b. OpenTSDB
- c. OpenTSD
- d. OpenDB

Answer: (b)

30. _____ command disables drops and recreates a table.

- a. drop
- b. truncate
- c. delete
- d. none of the mentioned

Answer: (b)

34. When a _____ is triggered the client receives a packet saying that the znode has changed.

- a. event
- b. watch
- c. row
- d. value

Answer: (b)

35. The underlying client-server protocol has changed in version _____ of ZooKeeper.

- a. 2.0.0
- b. 3.0.0
- c. 4.0.0
- d. 6.0.0

Answer: (b)

36. A number of constants used in the client ZooKeeper API were renamed in order to reduce _____ collision.

- a. value
- b. namespace
- c. counter
- d. none of the mentioned

Answer: (b)

37. ZooKeeper allows distributed processes to coordinate with each other through registers, known as _____

- a. znodes
- b. hnodes
- c. vnodes
- d. rnodes

Answer: (a)

38. Zookeeper essentially mirrors the _____ functionality exposed in the Linux kernel.

- a. iread
- b. inotify
- c. iwrite
- d. icount

Answer: (b)

39. ZooKeeper's architecture supports high _____ through redundant services.

- a. flexibility
- b. scalability
- c. availability
- d. interactivity

Answer: (c)

40. You need to have _____ installed before running ZooKeeper.

- a. Java
- b. C
- c. C++
- d. SQLGUI

Answer: (a)

41. To register a “watch” on a znode data, you need to use the _____ commands to access the current content or metadata.

- a. stat
- b. put
- c. receive
- d. gets

Answer: (a)

42. _____ has a design policy of using ZooKeeper only for transient data.

- a. Hive
- b. Imphala
- c. Hbase
- d. Oozie

Answer: (c)

43. The _____ master will register its own address in this znode at startup, making this znode the source of truth for identifying which server is the Master.

- a. active
- b. passive
- c. region
- d. all of the mentioned

Answer: (a)

44. Pig operates in mainly how many nodes?

- a. Two
- b. Three
- c. Four
- d. Five

Answer: (a)

45. You can run Pig in batch mode using _____

- a. Pig shell command
- b. Pig scripts
- c. Pig options
- d. All of the mentioned

Answer: (b)

46. Which of the following function is used to read data in PIG?

- a. WRITE
- b. READ
- c. LOAD
- d. None of the mentioned

Answer:(c)

47. You can run Pig in interactive mode using the _____ shell.

- a. Grunt
- b. FS
- c. HDFS
- d. None of the mentioned

Answer: (a)

48. Which of the following is the default mode?

- a. Mapreduce
- b. Tez
- c. Local
- d. All of the mentioned

Answer: (a)

49. _____ is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.

- a. Pig Latin
- b. Oozie
- c. Pig
- d. Hive

Answer: (c)

50. Hive also support custom extensions written in :

- a. C
- b. C++
- c. C#
- d. Java

Answer: (d)

51. Which of the following is not true about Pig?

- a. Apache Pig is an abstraction over MapReduce
- b. Pig can not perform all the data manipulation operations in Hadoop.
- c. Pig is a tool/platform which is used to analyze larger sets of data representing them as data flows.
- d. None of the above

Ans : b

52. Which of the following is/are a feature of Pig?

- a. Rich set of operators
- b. Ease of programming
- c. Extensibility
- d. All of the above

Ans : d

53. In which year apache Pig was released?

- a. 2005
- b. 2006
- c. 2007
- d. 2008

Ans : b

54. Pig operates in mainly how many nodes?

- a. 2
- b. 3
- c. 4
- d. 5

Ans : a

55. Which of the following company has developed PIG?

- a. Google
- b. Yahoo
- c. Microsoft
- d. Apple

Ans : b

56. Which of the following function is used to read data in PIG?

- a. Write
- b. Read
- c. Perform
- d. Load

Ans : d

57. _____ is a framework for collecting and storing script-level statistics for Pig Latin.

- a. Pig Stats
- b. PStatistics
- c. Pig Statistics
- d. All of the above

Ans : c

58. Which of the following is true statement?

- a. Pig is a high level language.
- b. Performing a Join operation in Apache Pig is pretty simple.
- c. Apache Pig is a data flow language.
- d. All of the above

Ans : d

59. Which of the following will compile the Pigunit?

- a. \$pig_trunk ant pigunit-jar
- b. \$pig_tr ant pigunit-jar
- c. \$pig_ ant pigunit-jar
- d. \$pigtr_ ant pigunit-jar

Ans : a

60. Point out the wrong statement.

- a. Pig can invoke code in language like Java Only
- b. Pig enables data workers to write complex data transformations without knowing Java
- c. Pig's simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL
- d. Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig

Ans : a

61. You can run Pig in interactive mode using the_____ shell

- a. Grunt
- b. FS
- c. HDFS
- d. None of the mentioned

Ans : a

62. Which of the following is the default mode?

- a. Mapreduce
- b. Tez
- c. Local
- d. All of the mentioned

Ans : d

63. Use the _____ command to run a Pig script that can interact with the Grunt shell (interactive mode)

- a. fetch
- b. declare
- c. run
- d. all of the mentioned

Ans : c

64. What are the different complex data types in

PIG

- a. Maps
- b. Tuples
- c. Bags
- d. All of these

Answer: d

65. What are the various diagnostic operators available in Apache Pig?

- a. Dump Operator
- b. Describe Operator
- c. Explain Operator
- d. All of these

66. If data has less elements than the specified schema elements in pig, then?

- a. Pig will not do any thing
- b. It will pad the end of the record columns with nulls
- c. Pig will through error
- d. Pig will warn you before it throws error

Answer: b

67. Which of the following command sets the value of a particular configuration variable (key)?

- a. set -v
- b. set =
- c. set
- d. reset

Answer: b

68. Point out the correct statement.

- a. Hive Commands are non-SQL statement such as setting a property or adding a resource
- b. Set -v prints a list of configuration variables that are overridden by the user or Hive
- c. Set sets a list of variables that are overridden by the user or Hive
- d. None of the mentioned

Answer: a

69. Which of the following will remove the resource(s) from the distributed cache?

- a. delete FILE[S] *
- b. delete JAR[S] *
- c. delete ARCHIVE[S] *
- d. all of the mentioned

Answer: d

70. _____ is a shell utility which can be used to run Hive queries in either interactive or batch mode.

- a. \$HIVE/bin/hive
- b. \$HIVE_HOME/hive
- c. \$HIVE_HOME/bin/hive
- d. All of the mentioned

Answer: c

71. HiveServer2 introduced in Hive 0.11 has a new CLI called _____

- a. Beeline
- b. SqlLine
- c. HiveLine
- d. CLILine

Answer: a

72. Variable Substitution is disabled by using _____

- a. set hive.variable.substitute=false;
- b. set hive.variable.substitutevalues=false;
- c. set hive.variable.substitute=true;
- d. all of the mentioned

Answer: a

73. _____ supports a new command shell Beeline that works with HiveServer2.

- a. HiveServer2
- b. HiveServer3
- c. HiveServer4
- d. None of the mentioned

Answer: a

74. In _____ mode HiveServer2 only accepts valid Thrift calls.

- a. Remote
- b. HTTP
- c. Embedded
- d. Interactive

Answer: a

75. The Hbase tables are

- a. Made read only by setting the read-only option
- b. Always writeable
- c. Always read-only
- d. Are made read only using the query to the

Answer: a

76. Every row in a Hbase table has

- a. Same number of columns
- b. Same number of column families
- c. Different number of columns
- d. Different number of column families

Answer: d

77. Hbase creates a new version of a record during

- a. Creation of a record
- b. Modification of a record
- c. Deletion of a record
- d. All the above

Answer: d

78. HBaseAdmin and _____ are the two important classes in this package that provide DDL functionalities.

- a. HTableDescriptor
- b. HDescriptor
- c. HTable
- d. HTabDescriptor

Answer: a

79. Mention how many operational commands in Hbase?

- a. Get
- b. Put
- c. Delete
- d. All of the mentioned

Answer: d

80. The _____ Server assigns regions to the region servers and takes the help of Apache ZooKeeper for this task.

- a. Region
- b. Master
- c. Zookeeper
- d. All of the mentioned