Big Data MCQ Links:

https://www.interviewbit.com/big-data-mcq/

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

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

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

Answer: b
Explanation: Hadoop batch processes data distributed over a number of computers ranging in 100s and 1000s.

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

Answer: a
Explanation: Data warehousing integrated with Hadoop would give a better understanding of data.

This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on "Analyzing Data with Hadoop".

1. Mapper implementations are passed the JobConf for the job via the _____ method.
a) JobConfigure.configure
b) JobConfigurable.configure
c) JobConfigurable.configurable
d) None of the mentioned
View Answer

Answer: b
Explanation: JobConfigurable.configure method is overridden to initialize themselves.

2. Point out the correct statement.
a) Applications can use the Reporter to report progress
b) The Hadoop MapReduce framework spawns one map task for each InputSplit generated by the InputFormat for the job
c) The intermediate, sorted outputs are always stored in a simple (key-len, key, value-len, value) format
d) All of the mentioned
View Answer

Answer: d
Explanation: Reporters can be used to set application-level status messages and update Counters.

3. Input to the _____ is the sorted output of the mappers.
a) Reducer
b) Mapper
c) Shuffle
d) All of the mentioned

Answer: a
Explanation: In the Shuffle phase the framework fetches the relevant partition of the output of all the mappers, via HTTP.

4. The right number of reduces seems to be _____
a) 0.90
b) 0.80
c) 0.36
d) 0.95
View Answer

Answer: d
Explanation: The right number of reduces seems to be 0.95 or 1.75.

5. Point out the wrong statement.
a) Reducer has 2 primary phases
b) Increasing the number of reduces increases the framework overhead, but increases load

balancing and lowers the cost of failures
c) It is legal to set the number of reduce-tasks to zero if no reduction is desired
d) The framework groups Reducer inputs by keys (since different mappers may have output the same key) in the sort stage
View Answer

Answer: a
Explanation: Reducer has 3 primary phases: shuffle, sort and reduce.

6. The output of the _____ is not sorted in the Mapreduce framework for Hadoop.
a) Mapper
b) Cascader
c) Scalding
d) None of the mentioned
View Answer

Answer: d
Explanation: The output of the reduce task is typically written to the FileSystem. The output of the Reducer is not sorted.

7. Which of the following phases occur simultaneously?
a) Shuffle and Sort
b) Reduce and Sort
c) Shuffle and Map
d) All of the mentioned
View Answer

Answer: a
Explanation: The shuffle and sort phases occur simultaneously; while map-outputs are being fetched they are merged.

8. 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
View Answer

Answer: c
Explanation: Reporter is a facility for MapReduce applications to report progress, set application-level status messages and update Counters.

9. _____ is a generalization of the facility provided by the MapReduce framework to collect data output by the Mapper or the Reducer.
a) Partitioner

b) OutputCollector
c) Reporter
d) All of the mentioned
View Answer

Answer: b
Explanation: Hadoop MapReduce comes bundled with a library of generally useful mappers, reducers, and partitioners.

10. _____ is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution.
a) Map Parameters
b) JobConf
c) MemoryConf
d) None of the mentioned
View Answer

Answer: b
Explanation: JobConf represents a MapReduce job configuration.

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

Answer: a
Explanation: To use Hive with HBase you'll typically want to launch two clusters, one to run HBase and the other to run Hive.

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

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

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

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

7. All of the following accurately describe Hadoop, EXCEPT _____
a) Open-source
b) Real-time
c) Java-based
d) Distributed computing approach
View Answer

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

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

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

9. _____ 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.

10. Facebook Tackles Big Data With _____ based on Hadoop.
a) 'Project Prism'
b) 'Prism'
c) 'Project Big'
d) 'Project Data'

Answer: a
Explanation: Prism automatically replicates and moves data wherever it's needed across a vast network of computing facilities.

This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on "Data Flow".

1. _____ is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.
a) Hive
b) MapReduce
c) Pig
d) Lucene
View Answer

Answer: b
Explanation: MapReduce is the heart of hadoop.

2. Point out the correct statement.
a) Data locality means movement of the algorithm to the data instead of data to algorithm
b) When the processing is done on the data algorithm is moved across the Action Nodes rather than data to the algorithm
c) Moving Computation is expensive than Moving Data
d) None of the mentioned

Answer: a
Explanation: Data flow framework possesses the feature of data locality.

3. The daemons associated with the MapReduce phase are _____ and task-trackers.
a) job-tracker
b) map-tracker
c) reduce-tracker
d) all of the mentioned
View Answer

Answer: a
Explanation: Map-Reduce jobs are submitted on job-tracker.

4. The JobTracker pushes work out to available _____ nodes in the cluster, striving to keep the work as close to the data as possible.
a) DataNodes
b) TaskTracker
c) ActionNodes
d) All of the mentioned
View Answer

Answer: b
Explanation: A heartbeat is sent from the TaskTracker to the JobTracker every few minutes to check its status whether the node is dead or alive.

5. Point out the wrong statement.
a) The map function in Hadoop MapReduce have the following general form:map:(K1, V1) -> list(K2, V2)
b) The reduce function in Hadoop MapReduce have the following general form: reduce: (K2, list(V2)) -> list(K3, V3)
c) MapReduce has a complex model of data processing: inputs and outputs for the map and reduce functions are key-value pairs
d) None of the mentioned
View Answer

Answer: c
Explanation: MapReduce is relatively simple model to implement in Hadoop.

6. InputFormat class calls the _____ function and computes splits for each file and then sends them to the jobtracker.
a) puts
b) gets
c) getSplits
d) all of the mentioned
View Answer

Answer: c
Explanation: InputFormat uses their storage locations to schedule map tasks to process them on the tasktrackers.

7. On a tasktracker, the map task passes the split to the createRecordReader() method on InputFormat to obtain a _____ for that split.
a) InputReader

b) RecordReader
c) OutputReader
d) None of the mentioned
View Answer

Answer: b
Explanation: The RecordReader loads data from its source and converts into key-value pairs suitable for reading by mapper.

8. The default InputFormat is _____ which treats each value of input a new value and the associated key is byte offset.
a) TextFormat
b) TextInputFormat
c) InputFormat
d) All of the mentioned
View Answer

Answer: b
Explanation: A RecordReader is little more than an iterator over records, and the map task uses one to generate record key-value pairs.

9. _____ controls the partitioning of the keys of the intermediate map-outputs.
a) Collector
b) Partitioner
c) InputFormat
d) None of the mentioned
View Answer

Answer: b
Explanation: The output of the mapper is sent to the partitioner.

10. Output of the mapper is first written on the local disk for sorting and _____ process.
a) shuffling
b) secondary sorting
c) forking
d) reducing
View Answer

Answer: a
Explanation: All values corresponding to the same key will go the same reducer.

1. _____ is a platform for developing data flows for the extraction, transformation, and loading (ETL) of huge datasets, as well as for data analysis.

A. Spark
B. HBase
C. Hive
D. Pig

Answer: D) Pig

Explanation:

Pig is a high-level platform or tool that is used to process massive amounts of data at a high level. When processing via the MapReduce framework, it provides a high level of abstraction for the user. It includes a high-level scripting language, known as Pig Latin that is used to construct the data analysis scripts that are employed in the system.

2. In contrast to relational databases, Hive is a query engine that supports the elements of SQL that are specifically designed for querying data.

A. True
B. False

Answer: A) True

Explanation:

Apache Hadoop is a distributed computing platform that allows for simple data summarization, ad hoc searches, and the analysis of big datasets stored in a variety of databases and file systems that connect with Hadoop. It is an easy approach to provide structure to massive volumes of unstructured data and then conduct batch SQL-like queries on that data using Hive as a data warehouse.

---

3. Custom extensions built in the ____ programming language are also supported by Hive.

A. Java
B. C#
C. C
D. C++

Answer: A) Java

Explanation:

Custom extensions built in the Java programming language is also supported by Hive. Apache Hive is built on top of Apache Hadoop and is used to provide data query and analysis capabilities to users. In order to query data stored in multiple databases and file systems that are integrated with Hadoop, Hive provides a SQL-like interface.

---

4. Amongst which of the following is / are correct,

A. Hive is a relational database that supports SQL queries.
B. Pig is a relational database that supports SQL queries.
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation:

Hive and Pig are the relational database that supports SQL queries. There is a special language similar to SQL known as HiveQL that converts the queries into MapReduce programmes that can be executed on datasets in HDFS. Pig provides the use of nested data types- Tuples, Maps, Bags, etc. and supports data operations like Joins, Filters, and Ordering.

---

5. In order to analyze all of this Big Data, Hive is a tool that has been developed.

A. True
B. False

Answer: A) True

Explanation:

In the context of big data analytics, Apache Hive is a distributed, fault-tolerant data warehousing system that can handle huge amounts of data. A data warehouse is a centralized repository of information that can be easily evaluated in order to make data-driven decisions that are informed. Hive lets users to read, write, and manage petabytes of data using SQL, which makes it a powerful tool for data scientists.

---

6. ____ general-purpose model and runtime framework for distributed data analytics.

A. Mapreduce
B. Spark

C. Hive
D. All of the mentioned above

Answer: A) Mapreduce

Explanation:

MapReduce is a programming model which enables massive scalability across hundreds or thousands of servers in a Hadoop cluster. MapReduce is known as the heart and soul of Apache Hadoop because of its processing and computing power.

---

7. Scalability is prioritized over latency in jobs such as _____.

A. HBase
B. HDFS
C. Hive
D. Mapreduce

Answer: C) Hive

Explanation:

Scalability is prioritized over latency in Hive. The performance of queries is influenced by the size of the cluster and the volume of data. In most cases, increasing cluster capacity alleviates problems caused by memory limits or disc performance limitations. Larger clusters, on the other hand, are more prone to experience various types of scalability challenges, such as a single slow node that causes query performance concerns.

---

8. _____ node serves as the Slave and is responsible for carrying out the Tasks that have been assigned to it by the JobTracker.

A. TaskReduce
B. Mapreduce
C. TaskTracker
D. JobTracker

Answer: C) TaskTracker

Explanation:

TaskTracker node serves as the Slave and is responsible for carrying out the Tasks that have been assigned to it by the JobTracker.

---

9. Apache Hive is data storage and _____ that stores and organizes data for study and querying.

A. Querying tool
B. Mapper
C. MapReduce
D. All of the mentioned above

Answer: A) Querying tool

Explanation:

Apache Hive is data storage and Querying tool that stores and organizes data for study and querying. A hive is a data transformation tool. It gathers information from a variety of sources, primarily HDFS. Hive is a good storage tool for the Hadoop Framework and is included with the framework. Hive is a replica of relational management tables that is used for data storage. It is capable of storing both organized and unstructured data. Hive, on the other hand, is capable of storing data. Hive imports unstructured data from HDFS first, and then develops a structure around it before loading the data.

---

10. The MapReduce framework is responsible for processing one or more pieces of data and producing the output results as _____.

A. Maptask
B. Task execution
C. Mapper
D. All of the mentioned above

Answer: A) Maptask

Explanation:

The MapReduce framework is responsible for processing one or more pieces of data and producing the output results as Maptask. A Map Task is a single instance of the MapReduce software that runs in the background. These jobs are responsible for determining which records from a data block should be processed. The input data is split and examined in parallel using computer resources that have been assigned to it in a Hadoop cluster.

---

11. Apache Hive is a data _____ infrastructure that is built on top of the Hadoop platform.

A. Warehouse
B. Map
C. Reduce
D. None of the mentioned above

Answer: A) Warehouse

Explanation:

Apache Hive is a data warehouse infrastructure built on the Hadoop framework that is ideal for data summarization, analysis, and querying. It is available as a free download. It makes use of a SQL-like language known as HQL (Hive Query Language).

---

12. The Hadoop framework is built in Java, which means that MapReduce applications do not need to be written in _____.

A. C#
B. C
C. Java
D. None of the mentioned above

Answer: C) Java

Explanation:

Hadoop is an Apache open-source framework developed in Java that enables for the distributed processing of massive datasets across clusters of computers by employing simple programming concepts. Hadoop is available for free from the Apache Software Foundation. A distributed storage and computation environment is provided by the Hadoop framework application, which works in conjunction with clusters of computers to deliver its services.

---

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

A. Reducer
B. Mapper
C. File system
D. All of these

Answer: B) Mapper

Explanation:

Mapper maps input key/value pairs to a set of intermediate key/value pairs. Map-Reduce is a programming methodology that is primarily divided into two parts, which are referred to as the Map Phase and the Reduce Phase, respectively. It is intended for the processing of data in parallel, which is distributed across a number of nodes.

---

14. HQL is a query language that is used to construct the custom map-reduce framework in Hive, which is written in _____.

A. Java
B. PHP
C. C#
D. None of the mentioned above

Answer: A) Java

Explanation:

HQL is a query language that is used to construct the custom map-reduce framework in Hive, which is written in Java. Hive supports a SQL parlance known as Hive Query Language (HQL) to retrieve or modify the data. Which is stored in the Hadoop.

---

15. The _____ is the default partitioned in Hadoop, and it offers a method called getPartition that allows us to partition data.

A. HashPartitioner
B. Map function
C. Reduce function
D. All of the mentioned above

Answer: A) HashPartitioner

Explanation:

The HashPartitioner is the default partitioned in Hadoop, and it offers a method called getPartition that allows us to partition data. When a MapReduce job is being executed, the partitioner does the partitioning of the keys of the intermediate map-outputs. The partition is created with the help of the hash function and the key (or a subset of the key). When all divisions are added together, it equals the total number of decrease tasks.

---

16. Hadoop is a framework that can be used in conjunction with a number of related products. Among the most common cohorts are _____.

A. MapReduce, Hive and HBase
B. Hive, Spark and HBase
C. Spark, Hive and ZooKeeper
D. Spark, HBase and Hive

Answer: A) MapReduce, Hive and HBase

Explanation:

Hadoop is a framework that can be used with a number of related products. Among the most common cohorts are MapReduce, Hive and HBase. The Hadoop software library is a framework that enables for the distributed processing of massive data sets across clusters of computers using simple programming models. It is a component of the Apache Hadoop software library. It is intended to grow from a small number of servers to thousands of devices, each of which can do computing and storage on its own.

---

17. _____ is best described as a programming model that is used to construct Hadoop-based applications that can be scaled up and down.

A. Oozie
B. Zookepper
C. MapReduce
D. All of the mentioned above

Answer: C) MapReduce

Explanation:

MapReduce is best described as a programming model that is used to construct Hadoop-based applications that can be scaled up and down. When it comes to the Hadoop framework, MapReduce is a programming paradigm or pattern that is used to retrieve large amounts of data stored in the Hadoop File System.

---

18. Amongst which of the following is/are the Hive function Meta commands.

A. Show functions
B. Describe function
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation:

Show functions and Describe function and are the Hive function Meta commands.

---

19. _____ is a shell utility that can be used to run Hive queries in either interactive or batch mode, depending on the situation.

A. $HIVE_HOME/bin/hive
B. $HIVE/bin/
C. $HIVE_HOME/hive
D. All of the mentioned above

Answer: A) $HIVE_HOME/bin/hive

Explanation:

$HIVE_HOME/bin/hive is a shell utility that can be used to run Hive queries in either interactive or batch mode, depending on the situation.

---

20. The _____ tool has the capability of listing all of the possible database schemas.

A. sqoop-list-databases
B. Hbase-list
C. hive schema
D. sqoop-list-columns

Answer: A) sqoop-list-databases

Explanation:

The sqoop-list-databases tool has the capability of listing all of the possible database schemas. The Sqoop List Databases tool is a database server query execution and parsing tool that runs and parses the "SHOW DATABASES" query. This displays a list of every database that is currently present on the database server. The primary goal of this utility is to compile a list of all of the database schemas that are currently available on the server.

---

21. Amongst which of the following is/are true with reference to User-defined Functions of Hive.

A. function that fetches one or more columns from a row as arguments
B. It returns a single value
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation:

User-defined Functions of Hive are the functions that fetch one or more columns from a row as arguments. It returns a single value. These functions give us the ability to design custom functions that process individual records or groups of related records. Hive comes pre-loaded with a large number of useful functionalities. Although there is some exclusion, there are also some specific circumstances for which UDFs are the best option.

---

22. Amongst which of the following is/are correct.

A. Default location of Hadoop configuration is in $HADOOP /conf/ HOME
B. If $HADOOP HOME is specified, Sqoop will utilise the default installation location
C. default location of Hadoop configuration is in $HADOOP HOME/conf/
D. Sqoop command-line tool serves as a wrapper for the bin/hadoop script that is included with Hadoop as a base.

Answer: D) Sqoop command-line tool serves as a wrapper for the bin/hadoop script that is included with Hadoop as a base

Explanation:

Sqoop command-line tool serves as a wrapper for the bin/hadoop script that is included with Hadoop as a base. Sqoop provides a straightforward command line, allowing us to retrieve data from a variety of databases using sqoop commands. They are written in Java and communicate with other databases through the JDBC interface. In addition to being an open-source technology, it also stands for "SQL to Hadoop" and "Hadoop to SQL."

---

23. A _____ serves as the master, and each cluster has just one NameNode.

A. Data Node
B. Block Size
C. Data block
D. NameNode

Answer: D) NameNode

Explanation:

A NameNode serves as the master, and each cluster has just one NameNode. Managing the File System Namespace and controlling access to files by clients are the responsibilities of the NameNode, which is a very highly available server.

---

24. HDFS always needs to work with large data sets.

A. True
B. False

Answer: A) True

Explanation:

HDFS always needs to work with large data sets. It will not be fruitful if HDFS is deployed to process several small data sets ranging in some MB or GB. The architecture of HDFS is designed in such a way so that it can be best fitted to store and retrieve huge amount of data. What are required are high cumulative data bandwidth and the scalability feature to spread out from a single node cluster to a hundred or a thousand-node cluster. The acid test is that HDFS should be able to manage tens of millions of files in a single occurrence.

---

25. HDFS operates in a ____ manner.

A. Master-slave architecture
B. Master-worker architecture
C. Worker-slave architecture
D. All of the mentioned above

Answer: B) Master-worker architecture

Explanation:

HDFS operates in a Master-worker architecture manner. It indicates that there is a single master node and a number of worker nodes in a given cluster. The Namenode is the node that serves as the master node. The namenode is the master node that runs on a different node in the cluster from the rest of the nodes.

---

26. HDFS follows the write-once, read-many.

A. True
B. False

Answer: A) True

Explanation:

HDFS follows the write-once, read-many approach for its files and applications. It assumes that a file in HDFS once written will not be modified, though it can be access 'n' number of times. At present, in HDFS strictly has one writer at any time. This assumption enables high throughput data access and also simplifies data coherency issues.

---

27. Amongst which of the following is not aligns as a characteristic of HDFS?

A. HDFS file system is well suited for storing data associated with applications that require low latency data access.
B. HDFS is well-suited for storing data connected to applications that require low-latency data access to be performed.
C. HDFS is not suited for instances in which multiple/simultaneous writes to the same file are required.
D. None of the mentioned above

Answer: C) HDFS is not suited for instances in which multiple/simultaneous writes to the same file are required.

Explanation:

HDFS is extremely fault-tolerant and to be implemented on low-cost hardware, this feature makes this best choice to cloud storage. A high-throughput access to application data is provided by HDFS, which is well-suited for applications that deal with massive amounts of data.

---

28. In order to interact with HDFS, a command line interface named _____ is provided.

A. HDFS Shell
B. DFS Shell
C. K Shell
D. FS Shell

Answer: D) FS Shell

Explanation:

It is possible to view the contents of a directory indicated in the path provided by the user by using the Hadoop FS shell command ls. All FS shell commands accept URIs for path names as inputs. Scheme:/authority/path is the URI format used in this example. In the case of HDFS, the scheme is hdfs, and in the case of the Local File System, the scheme is file. The scheme as well as the authority is entirely optional. Otherwise, the default scheme supplied in the configuration will be utilized if it is not explicitly specified.

---

29. HDFS stores data in a distributed manner, the data can be processed in parallel on a _____ of nodes.

A. Cluster

A.
B. Data Node
C. Master Node
D. None of the mentioned above

Answer: A) Cluster

Explanation:

HDFS stores data in a distributed manner, the data can be processed in parallel on a cluster of nodes. This, plus data locality, cut the processing time and enable high throughput. With HDFS, computation happens on the DataNodes where the data resides, rather than having the data move to where the computational unit is. By minimizing the distance between the data and the computing process, this approach decreases network congestion and boosts a system's overall throughput.

---

30. With reference to HDFS, Name Node is the prime node which contains metadata.

A. True
B. False

Answer: A) True

Explanation:

Name Node is the prime node which contains metadata (data about data) requiring comparatively fewer resources than the data nodes that stores the actual data. These data nodes are commodity hardware in the distributed environment. HDFS maintains all the coordination between the clusters and hardware, thus working at the heart of the system.

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

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

Answer: C) Peta

Explanation:

Big Data refers to data that is larger than a petabyte in size. The estimated volume of data that will be processed by Big Data solutions is significant and expected to continue to grow. In addition to increased storage and processing requirements, large data volumes necessitate the implementation of extra data preparation, curation, and management activities.

---

2. How many V's of Big Data?

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

Answer: D) 5

Explanation:

There are five V's of big data; these are Volume, Velocity, Variety, Value and Veracity. Knowing the 5 V's enables data scientists to extract more value from their data while also enabling the scientists' organizations to become more customer-centric as a result of their knowledge.

---

3. Unprocessed data or processed data are observations or measurements that can be expressed as text, numbers, or other types of media.

A. True
B. False

Answer: A) True

Explanation:

Unprocessed data or processed data are observations or measurements that can be expressed as text, numbers, or other types of media. In statistics, a data point, or observation, is a collection of one or more measurements taken on a single member of the observation unit (or unit of observation). Example: If the unit of observation is an individual and the research question is the determinants of money demand, a data point can be the values of income, wealth, the individual's age, and the number of dependents.

---

4. In computers, a ____ is a symbolic representation of facts or concepts from which information may be obtained with a reasonable degree of confidence.

A. Data
B. Knowledge
C. Program
D. Algorithm

Answer: A) Data

Explanation:

Information can be derived from data in computing if the data provides a symbolic representation of facts or concepts from which some probability can be calculated. While the summarizing of very large data sets might result in smaller data sets that are primarily composed of symbolic data, symbolic data are distinct in their own right on any sized data set, no matter how large or tiny it is.

---

5. In Big Data environments, Velocity refers –

A. Data can arrive at fast speed
B. Enormous datasets can accumulate within very short periods of time
C. Velocity of data translates into the amount of time it takes for the data to be processed
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

Large datasets can accumulate in a short amount of time in Big Data environments, since data arrives at lightning speed and accumulates in massive quantities. From the perspective of an enterprise, the velocity of data can be defined as the amount of time it takes for data to be processed once it enters the business's perimeter. In order to keep up with the rapid input of data, businesses must develop highly elastic and available data processing solutions, as well as the associated data storage capacities.

6. In Big Data environments, Variety of data includes –

A. Includes multiple formats and types of data
B. Includes structured data in the form of financial transactions,
C. Includes semi-structured data in the form of emails and unstructured data in the form of images
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

A wide range of data formats and kinds are required to be supported by Big Data systems, and this is referred to as data diversity. Enterprises have a number of issues when it comes to data integration, transformation, processing, and storage because of the variety of data. For example, financial transactions may contain structured data, while emails and photos may contain semi-structured data and unstructured data, respectively.

---

7. In Big Data environment, Veracity of data refers -

A. Quality or fidelity of data
B. Large size of the data that cannot be process
C. Small size of the data that can easily process
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

The quality or fidelity of data is referred to as veracity. Data entering Big Data environments must be evaluated for quality, which may necessitate data processing activities in order to resolve erroneous data and remove noise from the data stream. When it comes to authenticity, data can be either part of the signal or part of the noise of a dataset.

---

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

A. Cost Reduction
B. Time Reductions
C. Smarter Business Decisions

D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

Cost reduction, time reductions and smarter business decisions are the benefits of big data processing.

---

9. Structured data conforms to a data model or schema and is often stored in tabular form.

A. True
B. False

Answer: A) True

Explanation:

Structured data is data that has been organized according to a data model or schema and is frequently kept in tabular format. Due to the fact that it is used to record relationships between distinct things, it is most typically kept in a relational database. Enterprise applications and information systems, such as ERP and CRM systems, are frequently responsible for the generation of structured data.

---

10. Data that does not conform to a data model or data schema is known as _____.

A. Structured data
B. Unstructured data
C. Semi-structured data
D. All of the mentioned above

Answer: B) Unstructured data

Explanation:

It is referred to as unstructured data when the data does not comply to a data model or a data schema. Unstructured data is believed to account for 80 percent of all data within a given organization, according to some estimates. The growth rate of unstructured data is faster than that of structured data. SQL cannot be used to process or query unstructured data since it is not structured. As a binary large object, it is saved in a table in a relational database if it is necessary to be stored within the database (BLOB).

11. Amongst which of the following is/are not Big Data Technologies?

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

Answer: D) Apache Pytarch

Explanation:

Apache Pytarch is not a Big Data technology in the traditional sense. As part of a big data solution, Apache Hadoop, Apache Spark, and Apache Kafka are utilized. The Hadoop Distributed File System (HDFS) and a data processing engine that executes the MapReduce program to filter and sort data are the two primary components of Apache Hadoop. HDFS is a distributed file system that stores and distributes data across several computers. Apache Spark can also be used in conjunction with HDFS or another distributed file system. Hadoop MapReduce is capable of processing significantly larger data sets than Spark, particularly when the total size of the data collection exceeds the amount of memory that is available.

---

12. _____ involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task.

A. Parallel data processing
B. Single channel processing
C. Multi data processing
D. None of the mentioned above

Answer: A) Parallel data processing

Explanation:

Parallelism, which is defined in computing as the simultaneous execution of multiple processes, is the fundamental notion behind parallel data analysis. Parallel data processing is the simultaneous execution of several sub-tasks that together represent a bigger task in the context of a larger task. Parallel data analysis is a technique for analyzing data by running parallel processes on numerous computers at the same time.

---

13. Amongst which of the following can be considered as the main source of unstructured data.

A. Twitter
B. Facebook
C. Webpages
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

Unstructured data is primarily derived from social media platforms such as Twitter, Facebook, and the Internet. In the context of data storage, unstructured data refers to information that has not been organized according to a pre-determined data model or schema, and hence cannot be stored in a standard relational database management system (RDBMS). Text and multimedia are two types of unstructured content that are frequently encountered. Many business documents, as well as email messages, videos, images, webpages, and audio files, are unstructured in their content.

---

14. Amongst which of the following shows an example of unstructured data,

A. Students roll number, age
B. Videos
C. Audio files
D. Both B and C

Answer: D) Both B and C

Explanation:

Unstructured data includes files such as audio and video files, to name a few examples. In contrast to structured data, unstructured data does not fit well into a spreadsheet or database. It might be either textual or non-textual in nature. It can be created by a human or by a machine. Audio and video files, photos, text files - Word docs, PowerPoint presentations, email, chat logs, and other types of unstructured data are examples of unstructured data. Based on information gathered from social networking sites such as Facebook, Twitter, and LinkedIn Text messages, geolocation, chat, and call records are all examples of mobile data.

---

15. Scalability, elasticity, resource pooling, self-service, low cost and fault tolerance are the features of,

A. Cloud computing
B. Power BI
C. System development
D. None of the mentioned above

Answer: A) Cloud computing

Explanation:

Cloud computing is characterized by its scalability, elasticity, resource pooling, self-service, cheap cost, and fault tolerance, among other characteristics. While "Big Data" refers to massive volumes of data that have been collected, cloud computing refers to the technology that remotely receives this data and conducts any actions that have been specified on that information.

---

16. Amongst which of the following is/are the cloud deployment models,

A. Public Cloud
B. Private Cloud
C. Hybrid Cloud
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

Public Cloud, Private Cloud and Hybrid Cloud are the cloud deployment models. A cloud deployment model is characterized by the location of the infrastructure that will be used for the deployment and the authority that will be exercised over that infrastructure.

---

17. Virtualization separates resources and services from the underlying physical delivery environment.

A. True
B. False

Answer: A) True

Explanation:

Virtualization is the process of isolating resources and services from the physical delivery environment that they are delivered in. Big data virtualization is a method that focuses on the creation of virtual structures for large-scale data storage and processing environments. The usage of big data virtualization can be beneficial to businesses and other organizations because it helps them to make use of all of the data assets they have collected in order to achieve a variety of goals and objectives.

18. What is a Virtual Machine (VM)?

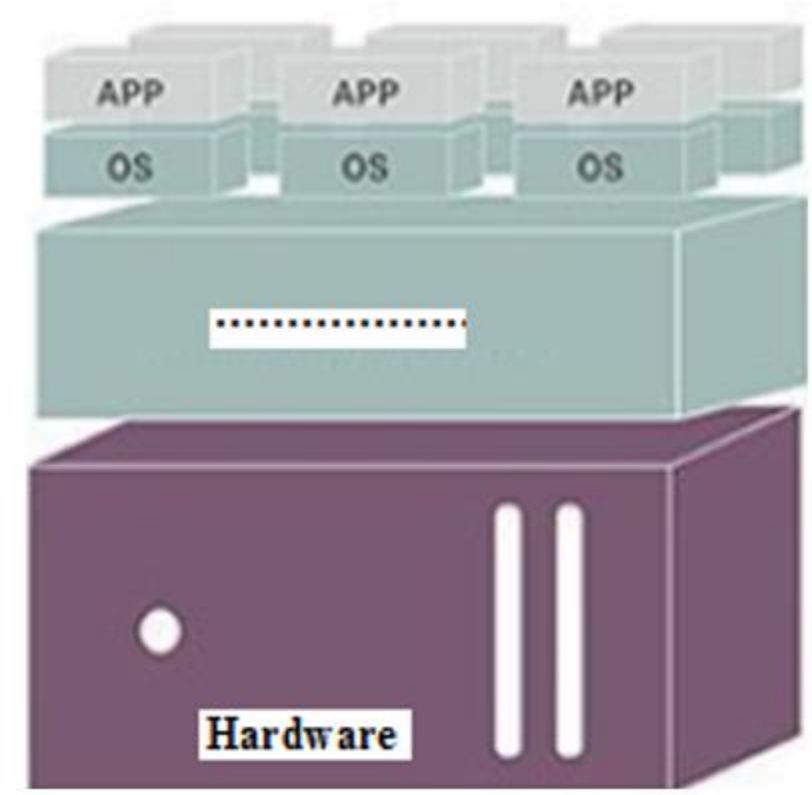
A. Virtual representation of a physical computer
B. Virtual representation of a logical computer
C. Virtual System Integration
D. All of the mentioned above

Answer: A) Virtual representation of a physical computer

Explanation:

A virtual machine (VM) is a representation of a physical computer that exists only in virtual space. In addition to a CPU, memory, and discs to store your stuff, it is capable of connecting to the internet if necessary. In contrast to the real and tangible components of your computer (referred to as hardware), virtual machines (VMs) are typically conceived of as virtual computers or software-defined computers that run on physical servers and exist solely as code.

19. In the given Virtual Architecture, name the missing layer,



A. Virtualization layer
B. Storage layer
C. Abstract layer
D. None of the mentioned above

Answer: A) Virtualization layer

Explanation:

It serves as an additional abstraction layer between network and storage hardware, processing, and the application executing on that hardware.... A machine that has a virtualization layer can create other (virtual) machines, which can then be used to install alternative operating systems on top of the main machine.

---

20. MongoDB is a ____ database.

A. SQL
B. DBMS
C. NoSQL
D. RDBMS

Answer: C) NoSQL

Explanation:

MongoDB is a NoSQL (non-relational) database. Databases that are not tabular in nature, such as NoSQL databases, store data in a different way than relational tables. NoSQL databases are classified into a number of categories based on the data model they use. Document, key-value, wide-column, and graph are the four most common types. They are capable of supporting variable schemas and scaling quickly when dealing with big amounts of data and significant user traffic.

---

21. MongoDB support cross platform and is written in _____ language.

A. Python
B. C++
C. R
D. Java

Answer: B) C++

Explanation:

MongoDB is a cross-platform database that is created in the C++ programming language. MongoDB stores data as flexible, JSON-like documents, which means that fields can differ from document to document and that the data structure can be altered as the database matures. Because MongoDB is a distributed database at its heart, it comes with built-in features such as high availability, horizontal scaling, and geographic distribution that are simple to utilize.

---

22. Amongst which of the following is / are true to run MongoDB?

A. High availability through built-in replication and failover
B. Management tooling for automation, monitoring, and backup
C. Fully elastic database as a service with built-in best practices
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

It is necessary to have high availability with built-in replication and failover as well as management tooling for automation, monitoring, and backup. It is also necessary to have MongoDB running as a fully elastic database as a service with built-in best practices.

---

23. Big data deals with high-volume, high-velocity and high-variety information assets,

A. True
B. False

Answer: A) True

Explanation:

Big data is defined as information assets with a high volume, a high velocity, and a high variety of information assets that necessitate the use of cost-effective, creative types of information processing in order to get greater insight and make better decisions.

---

24. _____ hypervisor runs directly on the underlying host system. It is also known as "Native Hypervisor" or "Bare metal hypervisor".

A. TYPE-1 Hypervisor
B. TYPE- 2 Hypervisor
C. Both A and B
D. None of the mentioned above

Answer: A) TYPE-1 Hypervisor

Explanation:

TYPE-1 The hypervisor is a virtual machine that runs on top of the underlying host system. It is also referred to as a "Native Hypervisor" or a "Bare metal hypervisor" in some circles.

---

25. ____ is also known as "Hosted Hypervisor".

A. TYPE-1 Hypervisor
B. TYPE- 2 Hypervisor

C. Both A and B
D. None of the mentioned above

Answer: B) TYPE- 2 Hypervisor

Explanation:

The term "Hosted Hypervisor" refers to a TYPE-2 hypervisor that is hosted on a third-party server. Type 2 hypervisors are often encountered in setups with a modest number of servers, as the name suggests. The fact that we do not have to install a management console on another machine in order to set up and administer virtual machines is what makes them so convenient. All of this can be accomplished on the server where the hypervisor is installed. Hosted hypervisors are mostly used as management consoles for virtual machines, and we may do any activity with the help of the built-in functions of the hypervisor.

---

26. In the layered architecture of Big Data Stack, Interfaces and feeds,

A. Internally managed data
B. Data feeds from external sources.
C. It provides access to each and every layer & components of big data stack
D. All of the mentioned above

Answer: D) All of the mentioned above

Explanation:

Interfaces and feeds internally maintained data; data feeds from external sources; and offers access to each and every layer and component of the Big Data Stack are all included in the tiered design of the Big Data Stack. In order to create the Data Pipeline in accordance with the diverse requirements of either the Batch Processing System or the Stream Processing System, the Big Data Architecture is used. This architecture is made up of six layers, each of which is responsible for ensuring the secure transit of data.

---

27. _____ is the supporting physical infrastructure is fundamental to the operation and scalability of big data architecture.

A. Redundant physical infrastructure
B. Integrated System
C. Integrated Database
D. All of the mentioned above

Answer: A) Redundant physical infrastructure

Explanation:

It is critical for the functioning and scalability of big data architecture to have redundant physical infrastructure as a backbone of the infrastructure. Ideally, networks should be redundant and have sufficient capacity to handle the projected amount and velocity of inbound and outbound data, in addition to the "regular" network traffic encountered by the organization.

---

28. The physical infrastructure of a big data is based on a distributed computing model.

A. True
B. False

Answer: A) True

Explanation:

It is a distributed computing model that underpins the physical infrastructure of a big data system. The physical infrastructure will "make or break" the deployment of big data since it is concerned with high-velocity, high-volume, and high-variety data streams. The majority of big data deployments require high availability, which necessitates the use of resilient and redundant networks, servers, and physical storage systems. The concepts of resilience and redundancy are intertwined.

---

29. Security infrastructure refers the data about your constituents needs to be protected to ____.

A. Meet compliance requirements
B. Protect the privacy
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation:

Security infrastructure refers to the data about your constituents that needs to be protected in order to comply with regulatory obligations and protect their personal information.

---

30. Reporting and visualization enables.

A. Processing of data
B. User friendly representation
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

Explanation:

Reporting and visualization make it possible to process data and present it in a user-friendly manner. Putting data into a chart, graph, or other visual format that may be used to inform analysis and interpretation is known as data visualization (or data presentation). Different stakeholders can engage with and learn from data visualizations since they make the examined data easily understandable.
