1. What is Big Data?

* Big data is a term applied to datasets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.

2. What is Hadoop?

* Apache Hadoop is a framework that allows for the distributed processing of large data sets across the clusters of commodity computers using a simple programming model.

-> It is an open-source Data Management either scale-out storage or distributed processing.

3. What is OLTP?

* OLTP (online transaction processing) is a class of software programs capable of supporting

Transaction-oriented applications on the Internet. Typically, OLTP systems are used for order

entry, financial transactions, customer relationship management (CRM) and retail sales.

4. What is OLAP?

* OLAP is an acronym for Online Analytical Processing. OLAP performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modelling. It is the foundation for many kinds of business applications for Business Performance Management, Planning, Budgeting, Forecasting, Financial Reporting, Analysis, Simulation Models, Knowledge Discovery, and Data Warehouse Reporting.

-> OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making.

5. What is RDBMS?

* A relational database refers to a database that stores data in a structured format, using rows and columns which makes it easy to locate and access specific values within the database. It is "relational" because the values within each table are related to each other. Tables may also be related to other tables. The relational structure makes it possible to run queries across multiple tables at once.

6. What is HDFS?

* HDFS is part of the Apache Hadoop Core project. In Hadoop environment Data is stored in a file system referred to as a HDFS- the Hadoop Distributes File System. Within HDFS, data is broken down into chunks and distributed across a cluster of machines.

-> HDFS has a master/slave architecture.

7. What is MapReduce?

* Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

-> It is a programming model used by Google.

-> It is a combination of the Map and Reduce models.

8. What is NameNode?

* In Hadoop environment HDFS has a master/slave architecture. A HDFS cluster consist of a single NameNode which is a master server that manages the file system namespace and regulates access to files by clients.

The NameNode has the following characteristics:

-> It is the master of the DataNodes

-> It maintains and executes file system namespace operations such as opening, closing, renaming files and directories.

-> It determines mapping of blocks to DataNodes.

9. What is DataNode?

* In Hadoop environment HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The Data Nodes are responsible for :

-> Handling read and write requests from application clients.

-> Performing block creation, deletion, and replication upon instruction from the NameNode. (The NameNode makes all decisions regarding replication of blocks.)

-> Sending heartbeats to the NameNode.

-> sending the blocks stored on the dataNode in a Blockreport.

10. What is Secondary NameNode?

* In a Hadoop environment Secondary NameNode connects to NameNode every hour. It

Performs Housekeeping, backup of NameNode metadata. This saved metadata can be

Rebuild a failed NameNode.

11. What is ResourceManager?

* ResourceManager (RM) is the master that arbitrates all the available cluster resources and thus helps manage the distributed applications running on the YARN. There will be only one Resource Manager per cluster. It has two components:

-> ApplcationsManager => The ApplicationsManager is responsible for accepting job-submissions, negotiating the first container to start ApplicationsMaster

-> Scheduler => The Scheduler is responsible for allocating resources to the various running jobs.

12. What is NodeManager?

* In a YARN environment the NodeManager is the agent of the ResourceManager, running on DataNode. The NodeManager is responsible for containers, monitoring their resource usage (CPU+RAM) and reporting the same to the ResourceManager (Scheduler). The per-job ApplicationsMaster has the responsibility of negotiating appropriate resource containers from the scheduler, tracking their status and monitoring for progress.

13. What is Distributed File System?

* The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

14. What is Parallel Processing?

* Parallel processing is a method of simultaneously breaking up and running program tasks on multiple microprocessors, thereby reducing processing time. Parallel processing may be accomplished via a computer with two or more processors.

15. What is Commodity Hardware?

* A Commodity Hardware is Computer hardware that is affordable and easy to obtain. Typically it is a low-performance system that is IBM PC-compatible and is capable of running Microsoft Windows, Linux, or MS-DOS without requiring any special devices or equipment.

16. What is JRE?

* The Java Runtime Environment (JRE) is a set of software tools for development of Java applications. It combines the Java Virtual Machine (JVM), platform core classes and supporting libraries. JRE is part of the Java Development Kit (JDK), but can be downloaded separately.

17. What is JDK?

* The Java Development Kit (JDK) is a software development environment used for developing Java applications and applets. It includes the Java Runtime Environment (JRE), an interpreter/loader (java), a compiler (javac), an archiver (jar), a documentation generator (javadoc) and other tools needed in Java development.

18. What is a process?

* A MapReduce program under execution, is called a JOB. It is the YARN's responsibility to handle the job execution

19. What is a daemon?

* A daemon is a type of program on Unix-like operating systems that runs unobtrusively in the

background, rather than under the direct control of a user, waiting to be activated by the

occurrence of a specific event or condition.

20. What is a service?

* A service is anything that starts on boot and provides critical system functioning that starts

on each boot. So a "service" would be a special type of process or a way of referring to a

group of related processes that collectively fulfil that role.

21. What is a Mapper?

* Hadoop Mapper task processes each input record and it generates a new <key, value> pairs. The <key, value> pairs can be completely different from the input pair. In mapper task, the output is the full collection of all these <key, value> pairs. Before writing the output for each mapper task, partitioning of output take place on the basis of the key and then sorting is done. This partitioning specifies that all the values for each key are grouped together.

22. What is a Reducer?

* The Reducer process the output of the mapper. After processing the data, it produces a new set of output. HDFS stores this output data. Hadoop Reducer takes a set of an intermediate key-value pair produced by the mapper as the input and runs a Reducer function on each of them. One can aggregate, filter, and combine this data (key, value) in a number of ways for a wide range of processing. Reducer first processes the intermediate values for particular key generated by the map function and then generates the output (zero or more key-value pair).

23. What is map?

* A map is an initial phase of MapReduce program. This is the phase where the raw data can be read, extracted and transformed. The results from this map phase is written out to HDFS or moved on to reducer for aggregate processing, such as file count, sum, min, max, etc. The map phase can also be thought as the ETL or projection step of MapReduce.

24. What is reduce?

* This is the final phase where data is stored on a user-defined key and grouped by that same key. The Reducer has the option to perform an aggregate function on the data. The Reducer phase can be thought as the aggregation step.

25. What is Input Split?

* InputSplit in Hadoop MapReduce is the logical representation of data. It describes a unit of work that contains a single map task in a MapReduce program.InputSplit in Hadoop is user defined . User can control split size according to the size of data in MapReduce program. Thus the number of map tasks is equal to the number of InputSplits.

26. What are map intermediate results?

* When the map tasks are complete, the intermediate results are gathered in the partition and a shuffling occurs, sorting the output for optimal processing by reduce.

27. What is shuffle and sort?

* The output from the mapper is an input to the shuffle and sort. Shuffle phase in Hadoop transfers the map output from Mapper to a Reducer in MapReduce whereas the Sort phase in MapReduce covers the merging and sorting of map outputs. Data from the mapper are grouped by the key, split among reducers and sorted by the key. Every reducer obtains all values associated with the same key. Shuffle and sort phase in Hadoop occur simultaneously and are done by the MapReduce framework.

28. What is Application Master?

* In a Hadoop environment, the Application Master is a first container created by the scheduler to perform a job. The Application Master initializes the job by creating a number of bookkeeping objects to keep track of the jobs progress as it will receive progress and completion records from the tasks.

29. What is a container?

* A container is a collection of resources to run an application. It is basically a combination of CPU + RAM.

30. What is a task?

* MapReduce job consists of several tasks (they could be either map or reduce tasks). If a task fails, it is launched again on another node. Those are task attempts.

31. What is a job in YARN?

* In terms of YARN, the programs that are being run on a cluster are called applications. In terms of MapReduce they are called jobs. So, if you are running MapReduce on YARN, job and application are the same thing.

32. What is monolithic computing?

* Monolithic computing is the simplest form of computing. A single computer such as personal computer is used for computing. The computer that does monolithic