What is -

* 1. Big Data?

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. It can be analysed for insights that lead to better decisions and strategic business moves.

3 V’s of big data:

* Volume
* Velocity
* Variety

==================================================================================

* 1. Hadoop?

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. It modernizes your IT infrastructure and keeps your data secure—in the cloud or on-premises—while helping you drive new revenue streams, improve customer experience, and control costs. It is a key component of a modern data architecture for data at rest.

==================================================================================

* 1. HDFS?

HDFS is a distributed file system that provides high throughput access to application data. It is the primary data storage system used by [Hadoop](http://searchcloudcomputing.techtarget.com/definition/Hadoop) applications. It employs a NameNode and DataNode architecture to implement a [distributed file system](http://searchwindowsserver.techtarget.com/definition/distributed-file-system-DFS) that provides high-performance access to data across highly scalable Hadoop clusters.

==================================================================================

* 1. a framework?

It is a software for developing software applications. It may also include code libraries & compiler and other programs. A framework is a collection of programs which do something useful which is used for developing own applications.

==================================================================================

* 1. a process?

A process is an instance of a computer program that is being executed.

==================================================================================

* 1. a CLI?

Command line interface (CLI) is a text-based interface that is used to operate software and operating systems while allowing the user to respond to visual prompts by typing single commands into the interface and receiving a reply in the same way.

==================================================================================

* 1. 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 occurance of a specific event or condition.

* Namenode – It runs on master node for HDFS.
* Datanode – It runs on slave nodes for HDFS.
* ResourceManager – It runs on master node for Yarn.
* NodeManager – It runs on slave node for Yarn.

==================================================================================

* 1. NameNode?

It is also known as Master node. NameNode does not store actual data or dataset.

* Makes all the decisions regarding distribution and replication
* Stores the metadata of the entire file system
* It is always 1 per cluster
* Regulates access to files by clients

==================================================================================

* 1. DataNode?

It is also known as Slave.

* Stores the data blocks of a file
* They are responsible for serving read/write requests to the clients
* They are usually 1 per node(except namenode machine)
* They send regular heartbeats to th NameNode(periodic heartbeats are block reports

==================================================================================

* 1. SecondaryNameNode?
* It work as a helper daemon
* It is not a hot backup for the namenode
* It is also called the check point node where it regularly does check pointing of the metadata files

==================================================================================

* 1. fault tolerance?

Fault tolerance is the property that enables a [system](https://en.wikipedia.org/wiki/System) to continue operating properly in the event of the failure of some (one or more faults within) of its components. If its operating quality decreases at all, the decrease is proportional to the severity of the failure, as compared to a native designed system in which even a small failure can cause total breakdown. Fault tolerance is particularly sought after in [high-availability](https://en.wikipedia.org/wiki/High-availability) or [life-critical systems](https://en.wikipedia.org/wiki/Life-critical_system). A fault-tolerant design enables a system to continue its intended operation, possibly at a reduced level, rather than failing completely, when some part of the system [fails](https://en.wikipedia.org/wiki/Failure).

==================================================================================

* 1. distributed computing?

Distributed computing is a computing concept that, in its most general sense, refers to multiple computer systems working on a single problem. In distributed computing, a single problem is divided into many parts, and each part is solved by different computers.

==================================================================================

* 1. a node?

A node is a point of intersection/connection within a network. Any system or device connected to a network is also called a node

==================================================================================

* 1. a cluster?

Cluster means Many Computers are worked together as one system. Hadoop Cluster means Computer Cluster used at Hadoop. Hadoop Cluster Mainly designed for storing large amount of unstructured data in Distributed file systems.

==================================================================================

* 1. pseudo distributed mode setup?

The pseudo-distributed mode is also known as a single-node cluster where both NameNode and DataNode will reside on the same machine. In pseudo-distributed mode, all the Hadoop daemons will be running on a single node.

==================================================================================

* 1. fully distributed mode setup?

This mode runs on cluster of machines. This runs on actual hadoop cluster.it is always difficult to debug a mapreduce program as you have mappers running on different machine with different piece of input.

==================================================================================

* 1. RDBMS?

A relational database management system (RDBMS) is a collection of programs and capabilities that enable IT teams and others to create, update, administer and otherwise interact with a relational database. Most commercial RDBMSes use Structured Query Language (SQL) to access the database.

==================================================================================

* 1. OLTP?

OLTP (online transaction processing) is a class of software programs capable of supporting transaction-oriented applications on the Internet.OLTP systems are used for order entry, financial transactions, customer relationship management (CRM) and retail sales. Such systems have a large number of users who conduct short transactions.

==================================================================================

* 1. OLAP?

OLAP (online analytical processing) is a computing method that enables users to easily and selectively extract and query data in order to analyze it from different points of view. OLAP business intelligence queries often aid in trends analysis, financial reporting, sales forecasting, budgeting and other planning purposes.

==================================================================================

* 1. Big Data Analytics?

Big data analytics is the often complex process of examining large and varied data sets -- or big data -- to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.

==================================================================================

* 1. Apache Software Foundation?

The Apache Software Foundation is a decentralized open source community of developers. The software they produce is distributed under the terms of the Apache License and is free and open-source software (FOSS). The Apache projects are characterized by a collaborative, consensus-based development process and an open and pragmatic software license. Each project is managed by a self-selected team of technical experts who are active contributors to the project.

==================================================================================

* 1. Linear Scalability?

it is ability to increase production inputs like labor by a certain percentage and get an equal percentage increase in output ,

vertical scale-increase,the memory (storage)eg;RDBMS.

==================================================================================