Q.1. Explain hadoop in layman's term:

Hadoop is an open source, Java-based programming framework thatsupports the processing and storage of extremely large data sets in a

distributed computing environment. It is designed to scale up from single servers to thousands of machines, each offering local computation

and storage. MapReduce and Hadoop distributed file system (HDFS) are the main components of Hadoop.

Q.2. Explain the components of Hadoop framework

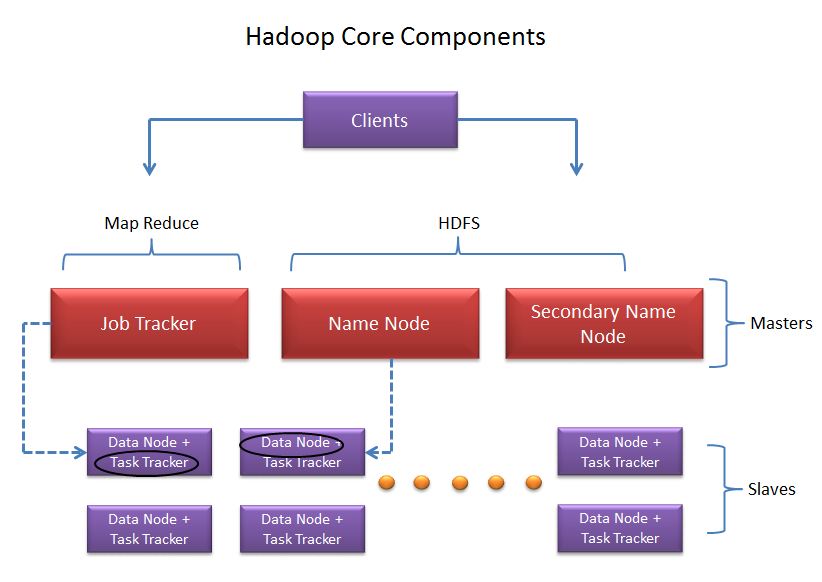
Hadoop cluster has 3 components:

1. Client

2. Master

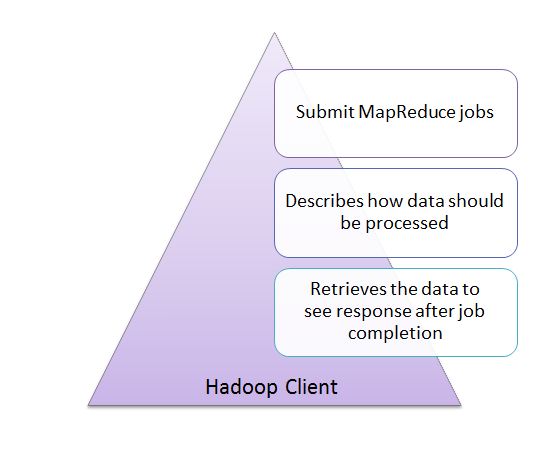
3. Slave

The role of each components are shown in the below image.



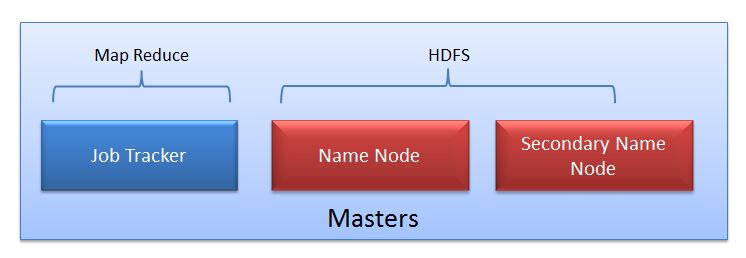
Client:

It is neither master nor slave, rather play a role of loading the data into cluster, submit MapReduce jobs describing how the data should be processed and then retrieve the data to see the response after job completion.



Masters:

The Masters consists of 3 components NameNode, Secondary Node name and JobTracker.

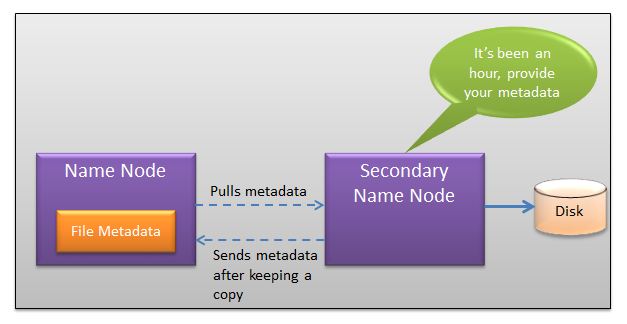


**NameNode:**  
NameNode does NOT store the files but only the file's metadata. In later section we will see it is actually the DataNode which stores the files.   
  
NameNode oversees the health of DataNode and coordinates access to the data stored in DataNode.   
Name node keeps track of all the file system related information such as to

* + Which section of file is saved in which part of the cluster
  + Last access time for the files
  + User permissions like which user have access to the file

**JobTracker:**  
JobTracker coordinates the parallel processing of data using MapReduce.

**Secondary Name Node:**  
Don't get confused with the name "Secondary". Secondary Node is NOT the backup or high availability node for Name node.

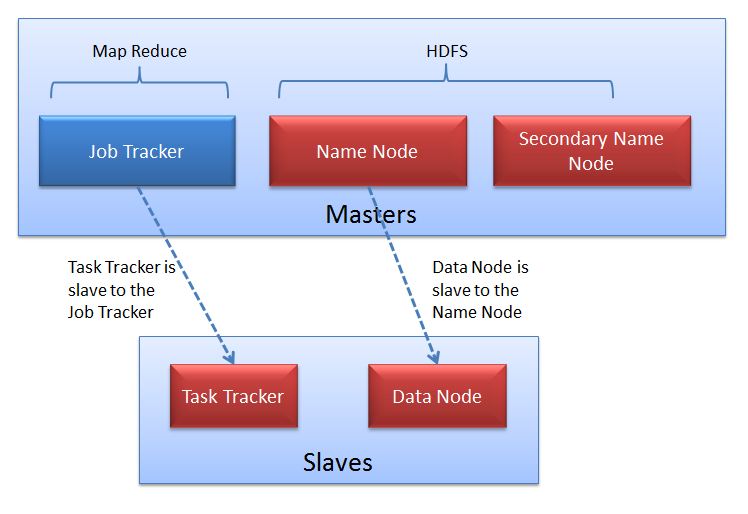


the job of Secondary Node is to contact NameNode in a periodic manner after certain time interval(by default 1 hour).   
NameNode which keeps all filesystem metadata in RAM has no capability to process that metadata on to disk. So if NameNode crashes, you lose everything in RAM itself and you don't have any backup of filesystem. What secondary node does is it contacts NameNode in an hour and pulls copy of metadata information out of NameNode. It shuffle and merge this information into clean file folder and sent to back again to NameNode, while keeping a copy for itself. Hence Secondary Node is not the backup rather it does job of housekeeping.   
In case of NameNode failure, saved metadata can rebuild it easily.

Slaves:

Slave nodes are the majority of machines in Hadoop Cluster and are responsible to

* + Store the data
  + Process the computation



Each slave runs both a DataNode and Task Tracker daemon which communicates to their masters. The Task Tracker daemon is a slave to the JobTracker and the DataNode daemon a slave to the NameNode.

Q.3. Explain the reasons to learn Big data technologies

**1. No signs of slowing down.**What’s fueled the meteoric rise of larger data sets? More people have access to mobile devices that sense and acquire information through the likes of cameras, microphones, etc. Consider this, since 2012, about 2.5 exabytes of data are created daily! That said, the Big Data revolution will continue to grow.

**2. Everyone uses Big Data.** Think Big Data is limited to IT circles? Think again. Big Data is everywhere, from politics to health care to even sports. Healthcare organizations use it to provide more personalized prescriptions, predictive analysis, and many other services. And sports-wise, more teams are using Big Data analysis to scout for athletes who best fit their needs.

**3. Information Managers in Demand.** Someone has to be able to implement, run, and manage the software used to analyze Big Data. So, in conjunction with the rise of Big Data, the demand for information management specialists has increased. Knowing how to use Big Data technology can make you highly desirable in a number of industries, even more so if you are able to break down that data and make it more streamlined. Other desired Big Data skills include data mining, information warehousing.

**4. Software Options Galore.**Just how big is Big Data? Major players such as Microsoft, IBM, Oracle have spent billions investing in data analytics and management software. There are a lot of solutions out there. There are also open source options such as Apache Hadoop, meaning the cost of entry doesn’t have to be outrageous.

**5. You’ll learn other tech.**Want to help companies really leverage Big Data? You’ll need to be able to utilize cloud-based services. They not only can provide the storage needed for large amounts of data, but the power needed to manage and analyze all of the data. Also, because Big Data is more than just numbers (it also involves media files such as audio and video) your data analysis skills are likely to expand.