1. **How will you empty the trash in HDFS?**

Empty the HDFS Trash by Typing

hdfs dfs –expunge

You can force hdfs to immediately delete the file and skip the trash bin by giving the rm command a -skipTrash flag.

hdfs dfs -rm -skipTrash /path/to/file/you/want/to/remove/permanently

Trash file location by default is:

/user/hdfs/.Trash

1. **Which tool you used to connect to Hadoop cluster?**

Putty, Winscp, <https://intellipaat.com/tutorial/hadoop-tutorial/hdfs-operations/>

1. **What is Seek time and Transfer rate in Hadoop?**
2. **What is Hadoop Archive Files (HAR)?**

HAR doesn’t support compression, and they are immutable

1. **distcp: Used to parallel copy data in large amount HDFS to another HDFS.**
2. **Which hadoop command will you run to check if Different Hadoop daemons are running?**

JPS command (JVM Process Status Tool).

1. **What is fsck?**

fsck stands for File System Check. It is a command used by HDFS. This command is used to check inconsistencies and if there is any problem in the file. For example, if there are any missing blocks for a file, HDFS gets notified through this command.

1. **What are the main differences between NAS (Network-attached storage) and HDFS?**

The main differences between NAS (Network-attached storage) and HDFS –

* HDFS runs on a cluster of machines while NAS runs on an individual machine. Hence, data redundancy is a common issue in HDFS. On the contrary, the replication protocol is different in case of NAS. Thus the chances of data redundancy are much less.
* Data is stored as data blocks in local drives in case of HDFS. In case of NAS, it is stored in dedicated hardware.

1. **What is the Command to format the NameNode?**

$ hdfs namenode –format

If we format the namenode then the metadata information's are deleted from namenode directory which is specified in the hdfs-site.xml as dfs.namenode.name.dir. But we still have the datas on the hadoop but not namenode meta data

1. **What are the common input formats in Hadoop?**

**Answer:**Below are the common input formats in Hadoop –

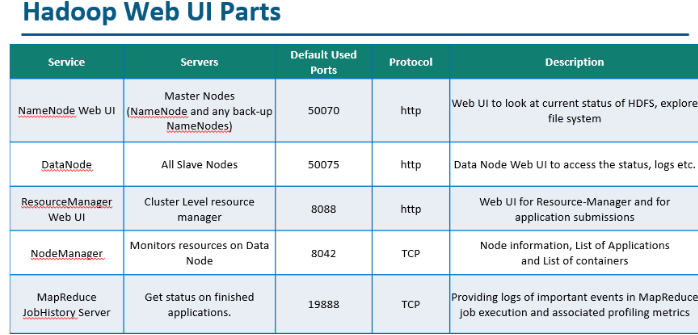
* **Text Input Format** – The default input format defined in Hadoop is the Text Input Format.
* **Sequence File Input Format** – To read files in a sequence, Sequence File Input Format is used.
* **Key Value Input Format** – The input format used for plain text files (files broken into lines) is the Key Value Input Format.

1. **What is Distributed Cache in Hadoop?**

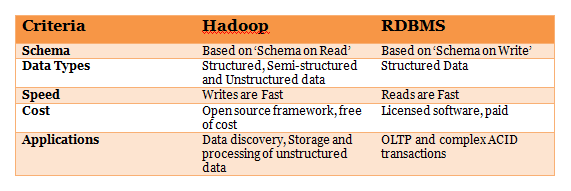
Distributed Cache is a facility provided by the Hadoop MapReduce framework. It cache files when needed by the applications. It can cache read only text files, archives, jar files etc. Once we have cached a file for our job, Hadoop will make it available on each DNs where map/reduce tasks are running.

Thus, we can access files from all the DNs in our map and reduce job**.**

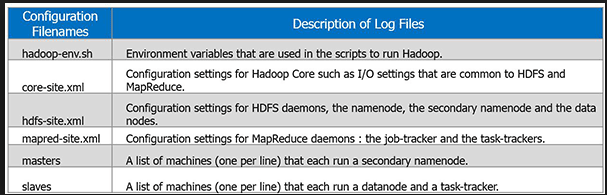
1. **What are different Hadoop web UI?**



1. **Explain the difference between Hadoop and RDBMS.**

**[](https://www.whizlabs.com/wp-content/uploads/2017/11/Hadoop-vs-RDBMS.png)**

1. **What are different configuraiton files in Hadoop environment?**



1. **What are the different configuration files in Hadoop?**

**Answer:**The different configuration files in Hadoop are –

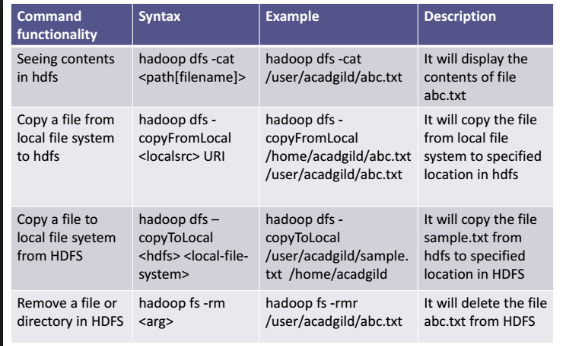
**core-site.xml –** This configuration file contains Hadoop core configuration settings, for example, I/O settings, very common for MapReduce and HDFS. It uses hostname a port.

**mapred-site.xml –**This configuration file specifies a framework name for MapReduce by setting mapreduce.framework.name

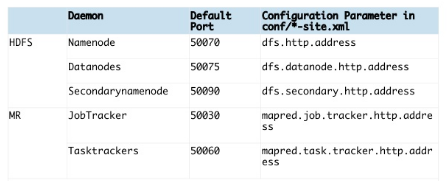
**hdfs-site.xml –** This configuration file contains HDFS daemons configuration settings. It also specifies default block permission and replication checking on HDFS.

**yarn-site.xml –** This configuration file specifies configuration settings for ResourceManager and NodeManager.

1. **What are some important Hadoop commands you are familiar with?**



1. **Hadoop components and default ports?**



<https://ambari.apache.org/1.2.3/installing-hadoop-using-ambari/content/reference_chap2_1.html>

1. **What if you restart NN?**

The Fsimage file will be lost, and editlog will be used to create new fsimage.

1. **What is Rack Awareness?**

Rack Awareness: No more than one replica is placed on one node. And no more than two replicas are placed on the same rack. This has a constraint that the number of racks used for block replication should be less than the total number of block replicas”.

In [**Big data**](http://data-flair.training/blogs/big-data-history-use-cases/)Hadoop, rack awareness is required for below reasons:

* To improve data[**high availability**](http://data-flair.training/blogs/hadoop-high-availability-tutorial/)and reliability.
* [**Improve the performance**](http://data-flair.training/blogs/hadoop-mapreduce-performance-tuning-best-practices/) of the cluster.
* To improve network bandwidth.
* Avoid losing data if entire rack fails though the chance of the rack failure is far less than that of node failure.
* To keep bulk data in the rack when possible.
* An assumption that in-rack id’s higher bandwidth, lower latency.

1. **Fault tolerance vs high availability?**

High availability is commonly thought of as a way to ensure a resource stays available, but the resource may suffer from some minor downtime. For example, with Hyper-V you have high availability because in the event that a host fails, the guest OSs just stop—there's no time to migrate the active state to another host, so you have some downtime. This is the same with VMware HA. vMotion can't be used because the host just stops and there's no live memory to move. You therefore typically lose the in-memory application state with high availability.

Fault tolerance mean you don't lose the in-memory application state in the event of a failure such as a host crash. Fault Tolerance is much harder than high availability in a virtual environment because you have to maintain two copies of a virtual machine, each on separate hosts. As memory and device state change on the primary, these changes have to be recorded and replayed on the secondary copy at the same time. Secondary name node creates checkpoint and create a ondisk fsimage at the same time it will keep truncating edit log. Now the edit log is small so time to read is short.

1. **How to change the replication of all blocks in HDFS?**

We can change the replication factor of a file using command:

hdfs dfs –setrep –w 3 /user/hdfs/file.txt

We can also change the replication factor of a directory using command:

hdfs dfs -setrep -R 2 /user/hdfs/test

But changing the replication factor for a directory will only affect the existing files and the new files under the directory will get created with the default replication factor (dfs.replication from hdfs-site.xml) of the cluster.

But you can temporarily override and turn off the HDFS default replication factor by passing:

-D dfs.replication=1

If we want to change the replication factor of the existing content in HDFS, which in our case is set to 4.

we can change the dfs.replication value to 4 in $HADOOP\_HOME/conf/hadoop-site.xml file. Which will start replicating to the factor of 4 for any new content that comes in. If we are looking to change for a specific file or a Directory, you can use the below commands to do that. To set replication of an individual file to 4:

$HADOOP\_HOME/bin/hadoop dfs -setrep -w 4 /path of the file

we can also do this on a Directory, which will change for all the files under it recursively.  
To change replication of entire directory under HDFS to 4:

./bin/hadoop dfs -setrep -R -w 4 /Directory path

– this is specific to a directory which we mention and if we give / (root)then it would do for all the files under it.

1. **What exactly is a Namespace, EditLog, FSImage and Metadata in** [**Hadoop**](https://www.quora.com/What-exactly-is-a-Namespace-EditLog-FSImage-and-Metadata-in-hadoop)**?**

**Namespace** in uniquely identifiable names in Hadoop.

Persistence of HDFS metadata broadly breaks down into 2 categories of files:

**fsimage** – An fsimage file contains the complete state of the file system at a point in time. Every file system modification is assigned a unique, monotonically increasing transaction ID. An fsimage file represents the file system state after all modifications up to a specific transaction ID.

**edits** – An edits file is a log that lists each file system change (file creation, deletion or modification) that was made after the most recent fsimage. If you like to see your edit logs and fsimage file location. Just open the hdfs-site.xml file.

In the hdfs-site.xml file check the [http://dfs.namenode.name](http://dfs.namenode.name/).dir value tag to get the location.

1. **What is default Block Size of HDFS?**

128mb in case of Cloudera Hadoop, and 64mb in case of Apache Hadoop.

1. **How to change block size in hdfs?**

We can change the block size using the property named **dfs.block.size** in the **hdfs-site.xml** file. Note: We should mention the size in bits. For example : 134217728 bits = 128 MB.

hadoop fs -D dfs.block.size=134217728 -put local\_name remote\_location

We can also modify your block size in your programs like this

Configuration conf = new Configuration() ;

conf.set( "dfs.block.size", 128\*1024\*1024) ;

1. **What is difference between latency and throughput?** Which one is desirable?

Latency is the time spent to move X from one point to the other.

Throughput is the no of Xs moved from one point to the other in a given amount of time.

1. **Vertical** **scaling vs Horizontal scaling?**

**Vertical scaling** refers to adding more resources (CPU/RAM/DISK) to your server (database or application server is still remains one) as on demand.

**Horizontal Scaling** involves adding more processing units or physical machines to your server

1. **How to install Hadoop MultiNode cluster? Explain in high level?**

Follow given url. <https://www.edureka.co/blog/setting-up-a-multi-node-cluster-in-hadoop-2.X>

**STEP 1:** Check the IP address of all machines.

**STEP 2:** Disable the firewall restrictions.

**STEP 1:** Open hosts file to add master and data node with their respective IP addresses. sudo nano /etc/hosts

**STEP 4:** Restart the sshd service. service sshd restart

**STEP 5:** Create the SSH Key in the master node. ssh-keygen -t rsa -P “”

**STEP 6:** Copy the generated ssh key to master node’s authorized keys.

cat $HOME/.ssh/id\_rsa.pub >> $HOME/.ssh/authorized\_keys

**STEP 7:**Copy the master node’s ssh key to slave’s authorized keys.

ssh-copy-id -i $HOME/.ssh/id\_rsa.pub edureka@slave

**STEP 8:** Extract the Java Tar File on all nodes.

**STEP 9:**Download the Hadoop 2.7.3 Package on all nodes.

**STEP 10:**Extract the Hadoop tar File on all nodes.

**STEP 11:**Add the Hadoop and Java paths in the bash file (.bashrc) on all nodes.

1. **How often data node send heartbeat to name node?**

In [HDFS](http://data-flair.training/blogs/comprehensive-hdfs-guide-introduction-architecture-data-read-write-tutorial/) DataNodes sends the heartbeat to NameNode in every **"3 sec"** (Default value of heartbeat.interval is 3 seconds.)

Heartbeat interval can be set in hdfs-site.xml as follows:

<property>

<name>heartbeat.recheck.interval</name>

<value>15</value>

<description>Determines data node heartbeat interval in seconds.</description>

</property>

**Timeout** = 2 \* heartbeat.recheck.interval + 10 \* heartbeat.interval.  
After which namenode will consider the datanode as a dead node and namenode stops sending I/O requests to this datanode. NameNode then schedules the creation of replicas of those [Data Blocks](http://data-flair.training/blogs/data-blocks-hdfs-hadoop-distributed-file-system/) on another datanodes.

1. **What happens when two users try to access the same file in the HDFS?**

HDFS NameNode supports exclusive write only. Hence, only the first user will receive the grant for file access and the second user will be rejected.

1. **How to restart all the daemons in Hadoop?**

To restart all the daemons, it is required to stop all the daemons first. The Hadoop directory contains sbin directory that stores the script files to stop and start daemons in Hadoop.

Use stop daemons command /sbin/stop-all.sh to stop all the daemons and then use /sin/start-all.sh command to start all the daemons again.

1. **What is the use of jps command in Hadoop?**

The jps command is used to check if the Hadoop daemons are running properly or not. This command shows all the daemons running on a machine i.e. Datanode, Namenode, NodeManager, ResourceManager etc.

1. **Explain the process that overwrites the replication factors in HDFS.**

There are two methods to overwrite the replication factors in HDFS –

**Method 1: On File Basis**

In this method, the replication factor is changed on the basis of file using Hadoop FS shell. The command used for this is:

$hadoop fs – setrep –w2/my/test\_file

Here, test\_file is the filename that’s replication factor will be set to 2.

**Method 2: On Directory Basis**

In this method, the replication factor is changed on directory basis i.e. the replication factor for all the files under a given directory is modified.

$hadoop fs –setrep –w5/my/test\_dir

Here, test\_dir is the name of the directory, the replication factor for the directory and all the files in it will be set to 5.

1. **What will happen with a NameNode that doesn’t have any data?**

**Answer:** A NameNode without any data doesn’t exist in Hadoop. If there is a NameNode, it will contain some data in it or it won’t exist.

1. **Explain NameNode recovery process.**

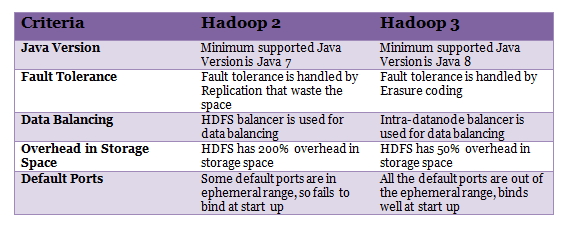
The NameNode recovery process involves the below-mentioned steps to make Hadoop cluster running:

* In the first step in the recovery process, file system metadata replica (FsImage) starts a new NameNode.
* The next step is to configure DataNodes and Clients. These DataNodes and Clients will then acknowledge new NameNode.
* During the final step, the new NameNode starts serving the client on the completion of last checkpoint FsImage loading and receiving block reports from the DataNodes.

**Note:** Don’t forget to mention, this NameNode recovery process consumes a lot of time on large Hadoop clusters. Thus, it makes routine maintenance difficult. For this reason, HDFS high availability architecture is recommended to use.

1. **What are the differences between Hadoop 2 and Hadoop 3?**

Following are the differences between Hadoop 2 and Hadoop 3 –

**[](https://www.whizlabs.com/wp-content/uploads/2017/11/Hadoop2-and-Hadop3.png)**

1. **What are the Port Numbers for NameNode, Task Tracker, and Job Tracker?**

* **NameNode** – Port 50070
* **Task Tracker** – Port 50060
* **Job Tracker** – Port 50030

1. **DFS can handle a large volume of data then why do we need Hadoop framework?**

Hadoop is not only for storing large data but also to process those big data. Though DFS(Distributed File System) too can store the data, but it lacks below features-

* It is not fault tolerant
* Data movement over a network depends on bandwidth.

1. **What is Sequencefileinputformat?**

Hadoop uses a specific file format which is known as Sequence file. The sequence file stores data in a serialized key-value pair. Sequencefileinputformat is an input format to read sequence files.

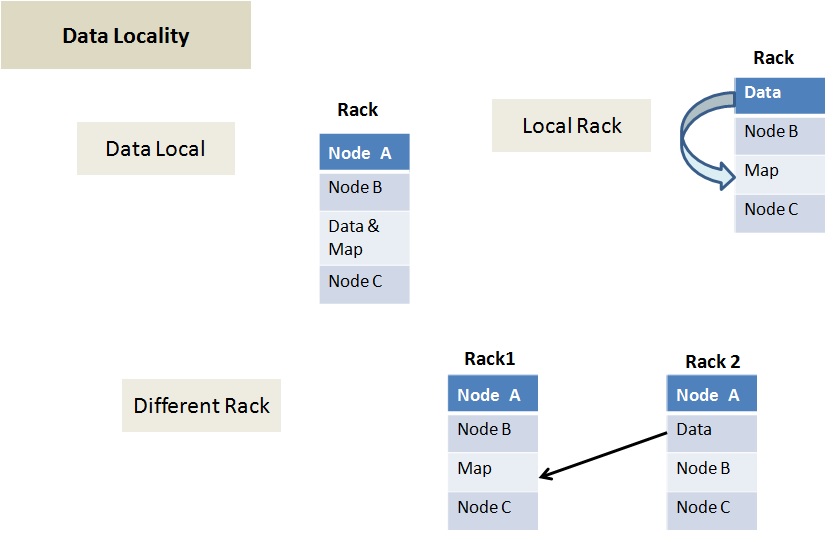
1. **Why is HDFS only suitable for large data sets and not the correct tool to use for many small files?**

This is due to the performance issue of NameNode. Usually, NameNode is allocated with huge space to store metadata for the large-scale file. The metadata is supposed to be a from a single file for optimum space utilization and cost benefit. In case of small size files, NameNode does not utilize the entire space which is a performance optimization issue.

1. **Why do we need Data Locality in Hadoop? Explain.**

Datasets in HDFS store as blocks in DataNodes the Hadoop cluster. During the execution of a MapReduce job the individual Mapper processes the blocks (Input Splits). If the data does not reside in the same node where the Mapper is executing the job, the data needs to be copied from the DataNode over the network to the mapper DataNode.

Now if a MapReduce job has more than 100 Mapper and each Mapper tries to copy the data from other DataNode in the cluster simultaneously, it would cause serious network congestion which is a big performance issue of the overall system. Hence, data proximity to the computation is an effective and cost-effective solution which is technically termed as Data locality in Hadoop. It helps to increase the overall throughput of the system.



Data locality can be of three types:

* **Data local –** In this type data and the mapper resides on the same node. This is the closest proximity of data and the most preferred scenario.
* **Rack Local –** In this scenarios mapper and data reside on the same rack but on the different data nodes.
* **Different Rack –** In this scenario mapper and data reside on the different racks.

1. **How Is Hadoop CLASSPATH essential to start or stop Hadoop daemons?**

CLASSPATH includes necessary directories that contain jar files to start or stop Hadoop daemons. Hence, setting CLASSPATH is essential to start or stop Hadoop daemons.

However, setting up CLASSPATH every time is not the standard that we follow. Usually CLASSPATH is written inside /etc/hadoop/hadoop-env.sh file. Hence, once we run Hadoop, it will load the CLASSPATH automatically.

1. **What are the different file permissions in HDFS for files or directory levels?**

Hadoop distributed file system (HDFS) uses a specific permissions model for files and directories. Following user levels are used in HDFS –

* Owner
* Group
* Others.

For each of the user mentioned above following permissions are applicable –

* read (r)
* write (w)
* execute(x).

Above mentioned permissions work differently for files and directories.

For files –

* The **r** permission is for reading a file
* The **w** permission is for writing a file.

For directories –

* The **r** permission lists the contents of a specific directory.
* The **w** permission creates or deletes a directory.
* The**X** permission is for accessing a child directory.

1. **What are the three running modes of Hadoop?**

The three running modes of Hadoop are as follows:

**i. Standalone or local**: This is the default mode and does not need any configuration. In this mode, all the following components of Hadoop uses local file system and runs on a single JVM –

* NameNode
* DataNode
* ResourceManager
* NodeManager

**ii. Pseudo-distributed**: In this mode, all the master and slave Hadoop services are deployed and executed on a single node.

**iii. Fully distributed**: In this mode, Hadoop master and slave services are deployed and executed on separate nodes.

1. **Explain JobTracker in Hadoop**

JobTracker is a JVM process in Hadoop to submit and track MapReduce jobs.

JobTracker performs the following activities in Hadoop in a sequence –

* JobTracker receives jobs that a client application submits to the job tracker
* JobTracker notifies NameNode to determine data node
* JobTracker allocates TaskTracker nodes based on available slots.
* it submits the work on allocated TaskTracker Nodes,
* JobTracker monitors the TaskTracker nodes.
* When a task fails, JobTracker is notified and decides how to reallocate the task.

1. **How can you achieve security in Hadoop?**

**Answer:**Kerberos are used to achieve security in Hadoop. There are 3 steps to access a service while using Kerberos, at a high level. Each step involves a message exchange with a server.

1. **Authentication –** The first step involves authentication of the client to the authentication server, and then provides a time-stamped TGT (Ticket-Granting Ticket) to the client.
2. **Authorization –** In this step, the client uses received TGT to request a service ticket from the TGS (Ticket Granting Server).
3. **Service Request –** It is the final step to achieve security in Hadoop. Then the client uses service ticket to authenticate himself to the server.
4. **How to recover a NameNode when it is down?**

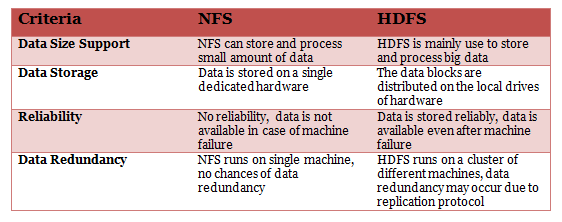
The following steps need to execute to make the Hadoop cluster up and running:

1. Use the FsImage which is file system metadata replica to start a new NameNode.
2. Configure the DataNodes and also the clients to make them acknowledge the newly started NameNode.
3. Once the new NameNode completes loading the last checkpoint FsImage which has received enough block reports from the DataNodes, it will start to serve the client.

In case of large Hadoop clusters, the NameNode recovery process consumes a lot of time which turns out to be a more significant challenge in case of routine maintenance.

1. **How is NFS different from HDFS?**

**Answer:**There are a number of distributed file systems that work in their own way. NFS (Network File System) is one of the oldest and popular distributed file storage systems whereas HDFS (Hadoop Distributed File System) is the recently used and popular one to handle big data.The main differences between NFS and HDFS are as follows –

**[](https://www.whizlabs.com/wp-content/uploads/2017/11/NFS-vs-HDFS.png)**

1. **What do you understand by Rack Awareness in Hadoop?**

It is an algorithm applied to the NameNode to decide how blocks and its replicas are placed. Depending on rack definitions network traffic is minimized between DataNodes within the same rack. For example, if we consider replication factor as 3, two copies will be placed on one rack whereas the third copy in a separate rack.

1. **What is the difference between “HDFS Block” and “Input Split”?**

The HDFS divides the input data physically into blocks for processing which is known as HDFS Block.

Input Split is a logical division of data by mapper for mapping operation.

Enhance your Big Data skills with the experts. Here is the [Complete List of Big Data Blogs](https://www.whizlabs.com/blog/a-complete-list-of-big-data-blogs/) where you can find latest news, trends, updates, and concepts of Big Data.

1. **What is a block in HDFS and what is its default size in Hadoop 1 and Hadoop 2? Can we change the block size?**

Blocks are smallest continuous data storage in a hard drive. For HDFS, blocks are stored across Hadoop cluster.

* The default block size in Hadoop 1 is: 64 MB
* The default block size in Hadoop 2 is: 128 MB

Yes, we can change block size by using the parameter – **dfs.block.size** located in the hdfs-site.xml file.

We can specify block size on per file basis.

1. **What is Distcp tool and when to use it**

DistCp (distributed copy) is a tool used for large inter/intra-cluster copying. It uses MapReduce to effect its distribution, error handling and recovery, and reporting

The most common invocation of DistCp is an inter-cluster copy:

bash$ hadoop distcp hdfs://nn1:8020/foo/bar \   
                    hdfs://nn2:8020/bar/foo

We can also specify multiple source directories on the command line:

bash$ hadoop distcp hdfs://nn1:8020/foo/a \   
                    hdfs://nn1:8020/foo/b \   
                    hdfs://nn2:8020/bar/foo

1. **How to start Ambari Server? 8080**

To start the Ambari Server: ambari-server start

To check the Ambari Server processes: ps -ef | grep Ambari

To stop the Ambari Server: ambari-server stop

Once you have started the Ambari service, you can access the Ambari Install Wizard through your browser.

1. Point your browser to http://{main.install.hostname}:8080.
2. Log in to the Ambari Server using the default username/password: admin/admin. You can change this later to whatever you wish.
3. **Is it possible to run MapReduce locally, without HDFS and Hadoop cluster?**In Windows to execute and test the Map Reduce code we need to take the help of cygwin (A windows to Unix simulator).  
   Steps to follow: 1.First install cywin. 2.set the cygpath environmental variable  
   now we can give local files as input to MR program.
4. **What does WORM stand for?**

Write Once Read Many

1. **YARN (MRv2) and MapReduce (MRv1) Schedulers**

A scheduler determines which jobs run, where and when they run, and the resources allocated to the jobs. The YARN (MRv2) and MapReduce (MRv1) computation frameworks support the following schedulers:

**FIFO** - Allocates resources based on arrival time.

**Fair** - Allocates resources to weighted pools, with fair sharing within each pool. When configuring the scheduling policy of a pool, Domain Resource Fairness (DRF) is a type of fair scheduler.

* CDH 4 Fair Scheduler
* CDH 5 Fair Scheduler

**Capacity** - Allocates resources to pools, with FIFO scheduling within each pool.

* CDH 4 Capacity Scheduler
* CDH 5 Capacity Scheduler

1. **When to Use Each Scheduler in Hadoop?**

The capacity scheduler is the right choice while we want to ensure guaranteed access with the potential in order to reuse unused capacity as well as prioritize jobs within queues, while we are running a large Hadoop cluster, along with the multiple clients.  
Whereas, when we use both small and large clusters for the same organization with a limited number of workloads, the fair scheduler works well. Also, in a simpler and less configurable way, it offers the means to non-uniformly distribute capacity to pools (of jobs). Furthermore, it can offer fast response times for small jobs mixed with larger jobs (supporting more interactive use models). Hence, it is useful in the presence of diverse jobs.

1. **How to enable Hadoop Fair Scheduler?**

In addition, it is a “contrib” module. Though, by copying it from Hadoop’s control/fair scheduler directory to the lib directory, place its JAR file on Hadoop’s classpath, to enable it.  
Furthermore, just set the **mapred.jobtracker.taskScheduler** property to:  
org.apache.hadoop.mapred.FairScheduler

1. **What is a daemon?**

It can be also called as a process. Windows equivalent of daemon is a service.

1. **What is CAP Theorem?**

Consistency, Partition Tolerance, Availability.

1. **RPC Means?**

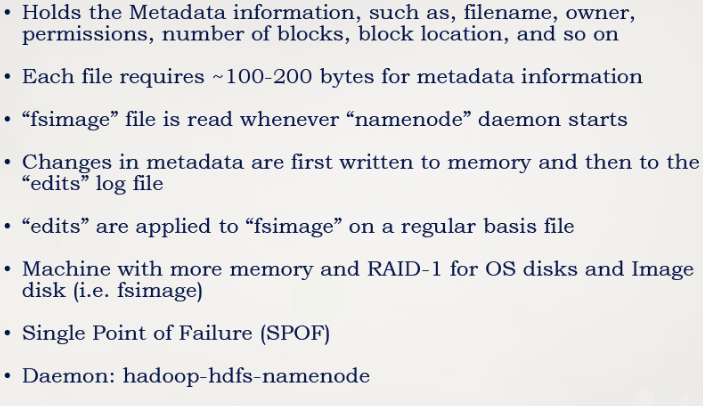
Remote procedure call

1. **Secondary NameNode vs Passive NameNode?**
2. **Block Replication happens in sequence but data is written in parallel?**
3. **A 400mb file is being copied to HDFS. System has finished copying 250 mb, what happens if client tries to access it?**

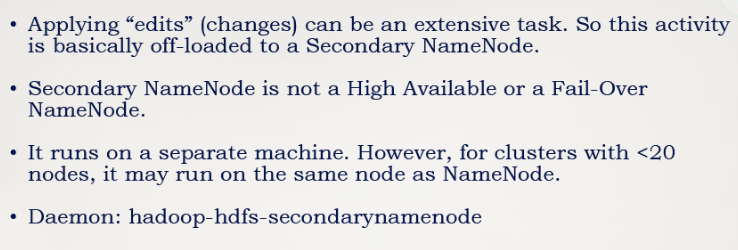
Client can read upto block that’s successfully written.

1. **Will secondary NameNode become active in case Active NameNode fails? No Explain**

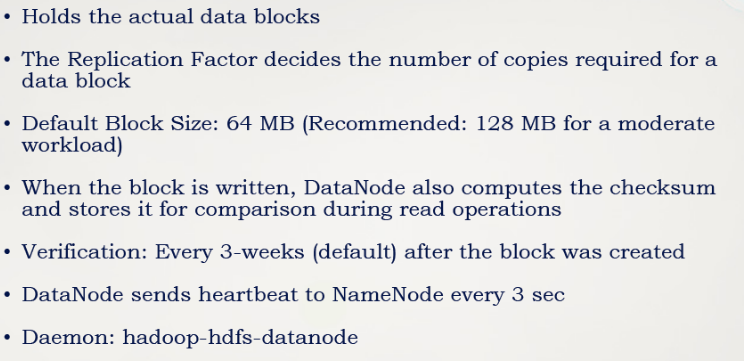
**Name node**:



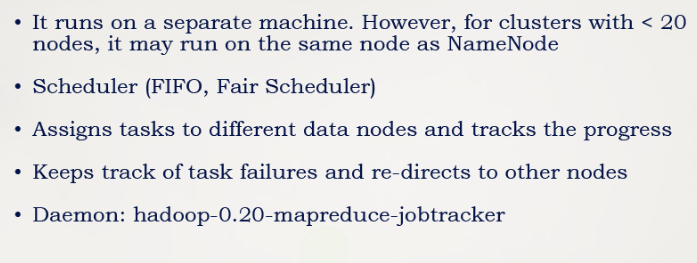
**Secondary Name Node:**



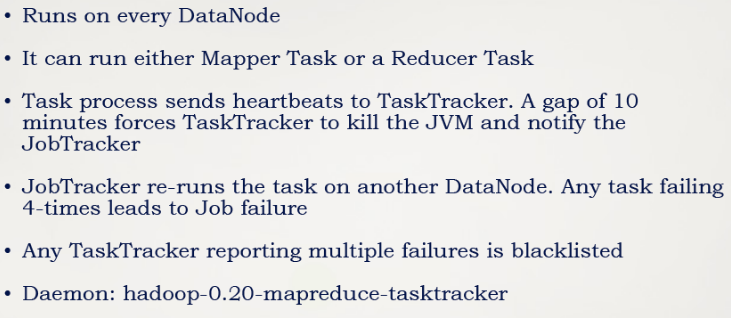
**Data Node**



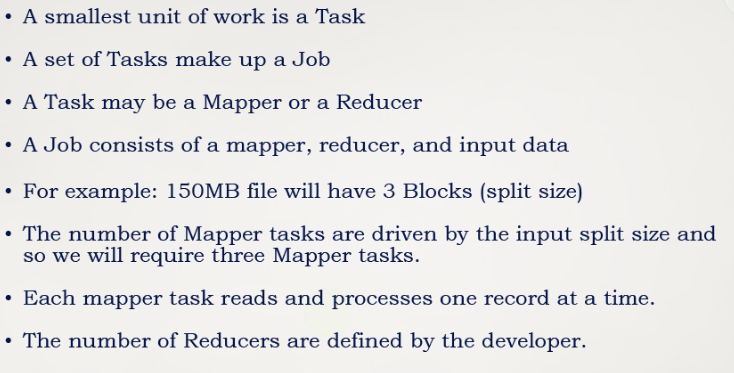
**Job Tracker**



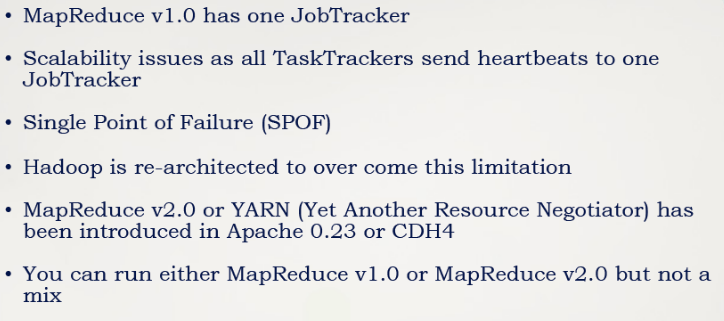
**Task Tracker**



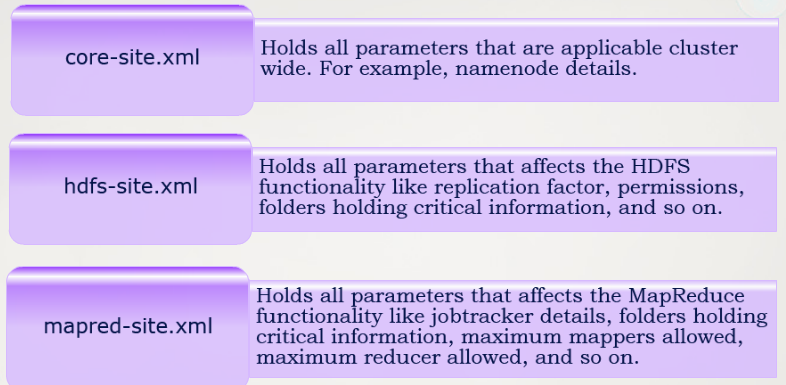
**Job Vs Task**



**MR1 vs MR2**



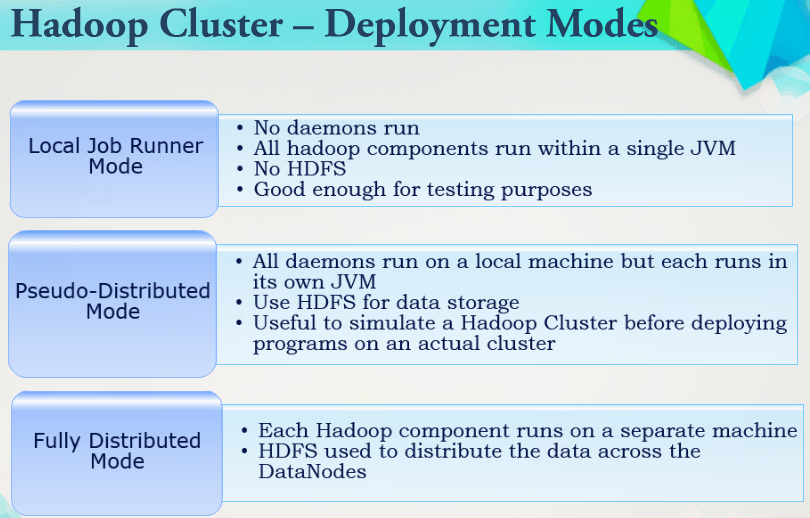
**Important Files**



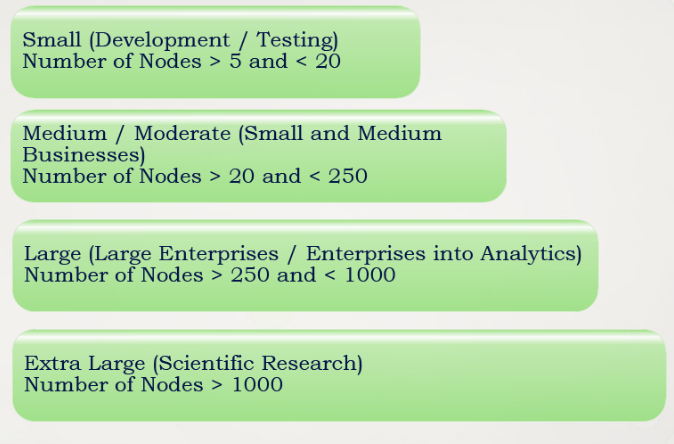
1. **What is the location of above 3 configuration files?**

/etc/hadoop/conf

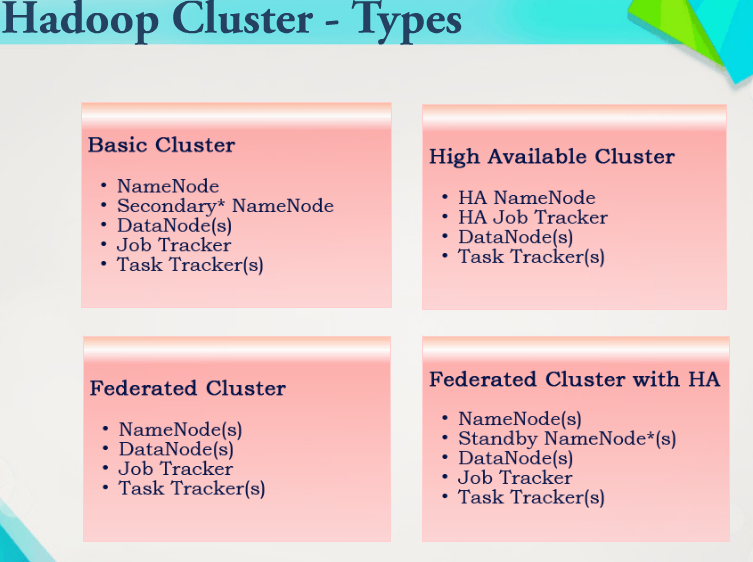
1. **What are different Hadoop modes?**



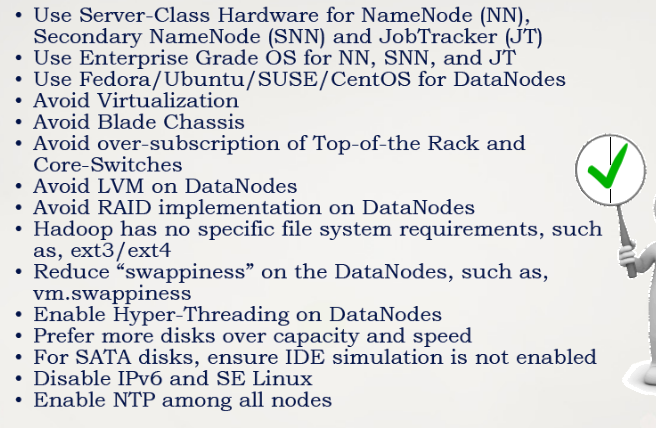
1. **What are cluster size for different kind of requirement?**



1. **What are different Hadoop cluster types?**



1. **Points to keep in mind for Hadoop environment?**



# Basic Hadoop Questions

1. **Explain some important features of Hadoop.**

Hadoop supports the storage and processing of big data. It is the best solution for handling big data challenges. Some important features of Hadoop are –

* **Open Source –** Hadoop is an open source framework which means it is available free of cost. Also, the users are allowed to change the source code as per their requirements.
* **Distributed Processing –** Hadoop supports distributed processing of data i.e. faster processing. The data in Hadoop HDFS is stored in a distributed manner and MapReduce is responsible for the parallel processing of data.
* **Fault Tolerance –** Hadoop is highly fault-tolerant. It creates three replicas for each block at different nodes, by default. This number can be changed according to the requirement. So, we can recover the data from another node if one node fails. The detection of node failure and recovery of data is done automatically.
* **Reliability –**Hadoop stores data on the cluster in a reliable manner that is independent of machine. So, the data stored in Hadoop environment is not affected by the failure of the machine.
* **Scalability –** Another important feature of Hadoop is the scalability. It is compatible with the other hardware and we can easily ass the new hardware to the nodes.
* **High Availability –** The data stored in Hadoop is available to access even after the hardware failure. In case of hardware failure, the data can be accessed from another path.

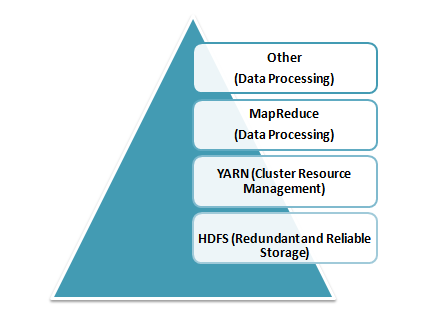
1. **Explain the different modes in which Hadoop run.**

**Answer:**Apache Hadoop runs in the following three modes –

* **Standalone (Local) Mode –** By default, Hadoop runs in a local mode i.e. on a non-distributed, single node. This mode uses the local file system to perform input and output operation. This mode does not support the use of HDFS, so it is used for debugging. No custom configuration is needed for configuration files in this mode.
* **Pseudo-Distributed Mode –** In the pseudo-distributed mode, Hadoop runs on a single node just like the Standalone mode. In this mode, each daemon runs in a separate Java process. As all the daemons run on a single node, there is the same node for both the Master and Slave nodes.
* **Fully – Distributed Mode –** In the fully-distributed mode, all the daemons run on separate individual nodes and thus forms a multi-node cluster. There are different nodes for Master and Slave nodes.

1. **Explain the core components of Hadoop.**

**Answer:**Hadoop is an open source framework that is meant for storage and processing of big data in a distributed manner. The core components of Hadoop are –

[](https://www.whizlabs.com/wp-content/uploads/2017/11/Core-Components-of-Hadoop.png)

* **HDFS (Hadoop Distributed File System) –** HDFS is the basic storage system of Hadoop. The large data files running on a cluster of commodity hardware are stored in HDFS. It can store data in a reliable manner even when hardware fails.
* **Hadoop MapReduce –** MapReduce is the Hadoop layer that is responsible for data processing. It writes an application to process unstructured and structured data stored in HDFS. It is responsible for the parallel processing of high volume of data by dividing data into independent tasks. The processing is done in two phases Map and Reduce. The Map is the first phase of processing that specifies complex logic code and the Reduce is the second phase of processing that specifies light-weight operations.
* **YARN –** The processing framework in Hadoop is YARN. It is used for resource management and provides multiple data processing engines i.e. data science, real-time streaming, and batch processing.

1. **Do you prefer good data or good models? Why?**

**How to Approach:**This is a tricky question but generally asked in the big data interview. It asks you to choose between good data or good models. As a candidate, you should try to answer it from your experience. Many companies want to follow a strict process of evaluating data; means they have already selected data models. In this case, having good data can be game-changing. The other way around also works as a model is chosen based on good data.

As we already mentioned, answer it from your experience. However, don’t say that having both good data and good models is important as it is hard to have both in real life projects.

1. **Will you optimize algorithms or code to make them run faster?**

**How to Approach:**The answer to this question should always be “Yes.” Real world performance matters and it doesn’t depend on the data or model you are using in your project.

The interviewer might also be interested to know if you have had any previous experience in code or algorithm optimization. For a beginner, it obviously depends on which projects he worked on in the past. Experienced candidates can share their experience accordingly as well. However, be honest about your work, and it is fine if you haven’t optimized code in the past. Just let the interviewer know your real experience and you will be able to crack the big data interview.

1. **How do you approach data preparation?**

**How to Approach:**Data preparation is one of the crucial steps in big data projects. A big data interview may involve at least one question based on data preparation. When the interviewer asks you this question, he wants to know what steps or precautions you take during data preparation.

As you already know, data preparation is required to get necessary data which can then further be used for modeling purposes. You should convey this message to the interviewer. You should also emphasize the type of model you are going to use and reasons behind choosing that particular model. Last, but not the least, you should also discuss important data preparation terms such as transforming variables, outlier values, unstructured data, identifying gaps, and others.

1. **How would you transform unstructured data into structured data?**

**How to Approach:**Unstructured data is very common in big data. The unstructured data should be transformed into structured data to ensure proper data analysis. You can start answering the question by briefly differentiating between the two. Once done, you can now discuss the methods you use to transform one form to another. You might also share the real-world situation where you did it. If you have recently been graduated, then you can share information related to your academic projects.

By answering this question correctly, you are signaling that you understand the types of data, both structured and unstructured, and also have the practical experience to work with these. If you give an answer to this question specifically, you will definitely be able to crack the big data interview.

1. **Which hardware configuration is most beneficial for Hadoop jobs?**

Dual processors or core machines with a configuration of 4 / 8 GB RAM and ECC memory is ideal for running Hadoop operations. However, the hardware configuration varies based on the project-specific workflow and process flow and need customization accordingly.

Secondary NameNode vs Stand By Name Node?

Cloudera vs Hortonworks

**Node Manager Vs Task Tracker?**

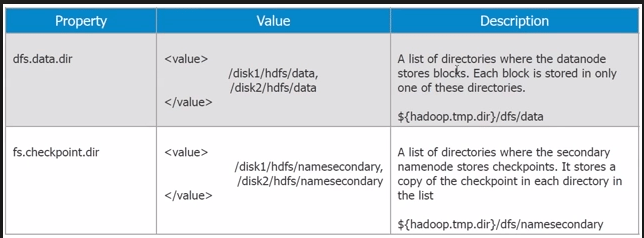
**Client-Side Buffering:**

**How Fault Tolerance is achieved in Hadoop?**

**Resource Manager Vs Node Manager.**

**Task Tracker Vs Job Tracker**

**Splitting of file into Data Blocks.**







<https://data-flair.training/blogs/rack-awareness-hadoop-hdfs/>

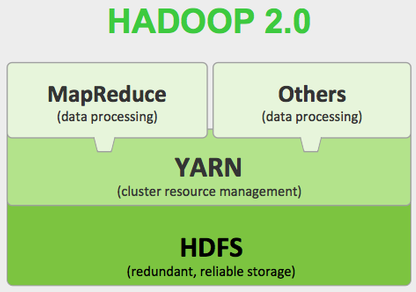
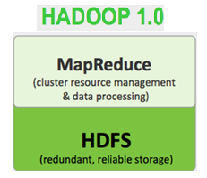
<https://www.youtube.com/channel/UC8OU1Tc1kxiI37uXBAbTX7A>

[https://data-.training/blogs/hadoop-tutorials-home/#tutorials](https://data-flair.training/blogs/hadoop-tutorials-home/#tutorials)

[**https://data-flair.training/blogs/hadoop-schedulers/**](https://data-flair.training/blogs/hadoop-schedulers/)

[**https://data-flair.training/blogs/hadoop-vs-cassandra/**](https://data-flair.training/blogs/hadoop-vs-cassandra/)

**What are differences between Hadoop 1x and 2x?**

****