**Configuration files in hadoop cluster**

**hadoop-env.sh- and jvm configuration**

**core\_site.sh- name node ip , port , where to store FS image,disk locations where data has to be stored**

**hdfs\_site.sh- block size, replication factor, compression technique**

**mapred\_site.sh- heap size for map task and reducer task**

**yarn\_site.sh- log of applications running location**

**Master**

**Slave**

**(1) SecondaryNamenode, a poorly named component of hadoop.**

Yes

**(2)What are the Side Data DistributionTechniques.**

Side data refers to extra static small data required by map reduce to perform job. Main challenge is the availability of side data on the node where the map would be executed. Hadoop provides two side data distribution techniques.

**(5)Can we change the file cached by Distributed Cache**

The cached files are copied to HDFS at the time of the submission of the job and then later copied locally to the local node by the different task trackers before they spawn M/R tasks. So, the files in the distributed cache can't be changed while the job is running

**(6)What if job tracker machine is down?**

The JobTracker is single point of failure for the Hadoop MapReduce service. If it goes down, all running jobs are halted.

**(7)Can we deploy job tracker other than name node?**

>yes we can

**(8)What are the four modules that make up the Apache Hadoop framework?**

>HDFS

>YARN

>MAP REDUCE

>HADOOP COMMON

**(9)Which modes can Hadoop be run in? List a few features for each mode.**

>STAND ALONE - No demon processses for stand allone

>PSUDO DISTRIBUTED NODE-It woeks on single machine but on different virtual machines

>FULLY DISTRIBUTED MODE- this is used in production and it isgeographically ditributed

**(10)Where are Hadoop’s configuration files located?**

> /usr/lib/hadoop/etc/hadoop

**(11)List Hadoop’s three configuration files.**

>HDFS-SITE.XML

>MAPPERED-SITE.XML

>YARN-SITE.XML

>CORE-SITE.XML

>HADOOP-ENV.SH

**(12)What are “slaves” and “masters” in Hadoop?**

>SLAVE- data node, application manager,node manager

>MASTER- name node,secondary name node,high availability node,resource manager

**(13)How many datanodes can run on a single Hadoop cluster?**

No limit

**(14)What is job tracker in Hadoop?**

JobTracker is the daemon service for submitting and tracking MapReduce jobs in Hadoop. There is only One Job Tracker process run on any hadoop cluster. Job Tracker runs on its own JVM process. In a typical production cluster its run on a separate machine. Each slave node is configured with job tracker node location

* Client applications submit jobs to the Job tracker.
* The JobTracker talks to the NameNode to determine the location of the data
* The JobTracker locates TaskTracker nodes with available slots at or near the data
* The JobTracker submits the work to the chosen TaskTracker nodes.
* The TaskTracker nodes are monitored. If they do not submit heartbeat signals often enough, they are deemed to have failed and the work is scheduled on a different TaskTracker.
* A TaskTracker will notify the JobTracker when a task fails. The JobTracker decides what to do then: it may resubmit the job elsewhere, it may mark that specific record as something to avoid, and it may may even blacklist the TaskTracker as unreliable.
* When the work is completed, the JobTracker updates its status.

JOB TRACKER>

1. It communicates with name node

2.Gets the meta data

3. Allocated the task traker for each map reduce task

4. Maintain the job completetion status

**(15)How many job tracker processes can run on a single Hadoop cluster?**

one job tracker a cluster

**(17)What is “PID”?**

>Process id

**(18)What is “jps”?**

>java process status

**(19)Is there another way to check whether Namenode is working?**

jps

using cloudera manager status

using ping ip

**(20)What is “fsck”?**

file system status check Command for this is - hdfs fsck /

**(23)List the network requirements for using Hadoop.**

**(24)Which port does SSH work on?**

> 22

**(25)What is streaming in Hadoop?**

Hadoop streaming can be performed using languages like Python, Java, PHP, Scala, Perl, UNIX, and many more. The utility allows us to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

**(28)Name the most common Input Formats defined in Hadoop? Which one is default?**

>TextInputFormat

**(32)Consider case scenario: In M/R system,**

- HDFS block size is 64 MB

- Input format is FileInputFormat

- We have 3 files of size 64K, 65Mb and 127Mb

**(33)How many input splits will be made by Hadoop framework?**

Blocks - f1 - 1

f2 - 2

f3 - 2

**(35)What is JobTracker?**

JobTracker is the daemon service for submitting and tracking MapReduce jobs in Hadoop.

**(36)What are some typical functions of Job Tracker?**

* Client applications submit jobs to the Job tracker.
* The JobTracker talks to the NameNode to determine the location of the data
* The JobTracker locates TaskTracker nodes with available slots at or near the data
* The JobTracker submits the work to the chosen TaskTracker nodes.
* The TaskTracker nodes are monitored. If they do not submit heartbeat signals often enough, they are deemed to have failed and the work is scheduled on a different TaskTracker.
* A TaskTracker will notify the JobTracker when a task fails. The JobTracker decides what to do then: it may resubmit the job elsewhere, it may mark that specific record as something to avoid, and it may may even blacklist the TaskTracker as unreliable.
* When the work is completed, the JobTracker updates its status.

**(37)What is TaskTracker?**

A TaskTracker is a slave node daemon in the cluster that accepts tasks (Map, Reduce and Shuffle operations) from a JobTracker. There is only One Task Tracker process run on any hadoop slave node.

**(38)What is the relationship between Jobs and Tasks in Hadoop?**

>job is divided into tasks which will be assigned to each data node

**(39)Suppose Hadoop spawned 100 tasks for a job and one of the task failed. What will Hadoop do?**

task speculation

**(40)How does speculative execution work in Hadoop?**

> Process of launching a new task only after old task runs for the significant amount of time and framework detects it is running slow as compared to other tasks, for the same job

**(41)Using command line in Linux, how will you**

- See all jobs running in the Hadoop cluster

> job –list

**(42)What is BIG DATA?**

Data whihc satisfy five properties 1. vloume 2. velocity 3. variety 4.veracity 5. value

**(43)Can you give some examples of Big Data?**

>Social media data

>Health care domain data

>Trading data

>Financial data

**(44)How analysis of Big Data is useful for organizations?**

>Analysis of big data helps organization to get the insights of data which helps organization to take business decisions

**(45)What are some of the characteristics of Hadoop framework?**

>Distribute data storage

>Schema on Read

>Data Locality

>quick analysis on Big data

>Horizontal Scalability

**(46)Give a brief overview of Hadoop history.**

>Google file system paper is reference for hadoop

**(47)Give examples of some companies that are using Hadoop structure?**

>Amazon,yahoo,facebook,twitter,city bank,barclay

**(48)What is the basic difference between traditional RDBMS and Hadoop?**

**Architecture**  
• RDBMS have ACID properties.  
• [**Hadoop**](http://data-flair.training/blogs/hadoop-introduction-tutorial-quick-guide/) is distributed computing framework having two main components: Distributed file system ([**HDFS**](http://data-flair.training/blogs/comprehensive-hdfs-guide-introduction-architecture-data-read-write-tutorial/)) and **[MapReduce](http://data-flair.training/blogs/hadoop-mapreduce-introduction-tutorial-comprehensive-guide/)**.  
**Data Size**  
• RDMS: Giga bytes of data  
• Hadoop: petabytes of data  
**Updates**  
• RDMS: we can able to read and write many times  
• Hadoop: we can read many times and writeis limited

**Data acceptance**  
• RDBMS has reasonable data sets.  
• Hadoop has inexpensive data storage,(structured, semi-structured). So we can able store everything in our database and there will be no data loss.  
**Scalability**  
• RDBMS is provided vertical scalability. So if the data increases for storing then we have to increase particular system configuration.  
• Hadoop provides horizontal scalability. So we just have to add one or more node to the cluster if there is any requirement for an increase in data.  
**OLTP**  
• RDMS support OLTP (Real-time data processing).  
• OLTP is not supported in Apache Hadoop.  
• Hadoop supports large scale Batch Processing workloads (OLAP).  
**Cost**  
• RDMS -Licensed software, therefore we have to pay for the software.  
• Hadoop is open source framework, so we don’t need to pay for software.

**(49)What are the core components of Hadoop?**

>YARN,HDFS,MR

**(50)What is HDFS?**

>Hadoop Distributed File system

**(51)What are the key features of HDFS?**

>Distributed file system

>Schema on read

>Suitable for Big Data

**(52)What is Fault Tolerance?**

if any node of cluster is down and then also cluster is working with remaining nodes without having effect on the cluster of that node

**(53)Replication causes data redundancy then why is is pursued in HDFS?**

>To increase avalibility of data

**(54)Since the data is replicated thrice in HDFS, does it mean that any calculation done on one node will also be replicated on the other two?**

>yes

**(55)What is a commodity hardware? Does commodity hardware include RAM?**

>Commodity hardware 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

**(56)What is a metadata?**

>Metadata is stored in name node which holds Ip address of data node,where Replica stored ,Information about each file blocks stored ,

**(57)Why do we use HDFS for applications having large data sets and not when there are lot of small files?**

>Because the HDFS gives good performance for large files and less performance for small files

**(58)What is a daemon?**

>It is the process running in the background .

**(59)Is Namenode machine same as datanode machine as in terms of hardware?**

>No

**(60)What is a heartbeat in HDFS?**

>Hearbeat is the signal sent by data node to name node to convey the detailed information about the data availabe with it ,it sends this signal after evry 3 sec.

**(61)Are Namenode and job tracker on the same host?**

>may or may not be

**(62)Are job tracker and task trackers present in separate machines?**

>yes

**(63)On what basis Namenode will decide which datanode to write on?**

RAA

**(64)Doesn’t Google have its very own version of DFS?**

It has GFS

**(65)Who is a ‘user’ in HDFS?**

**(66)Is client the end user in HDFS?**

No

No it is the node which fires the query and computation jobs

**(67)What is the communication channel between client and namenode/datanode?**

secure shell

**(68)What is a rack?**

A rack is a collection of 30 or 40 nodes that are physically stored close together and are all connected to the same network switch

**(69)On what basis data will be stored on a rack?**

Rack Awareness Algorithm

**(70)Do we need to place 2nd and 3rd data in rack 2 only?**

not necessary

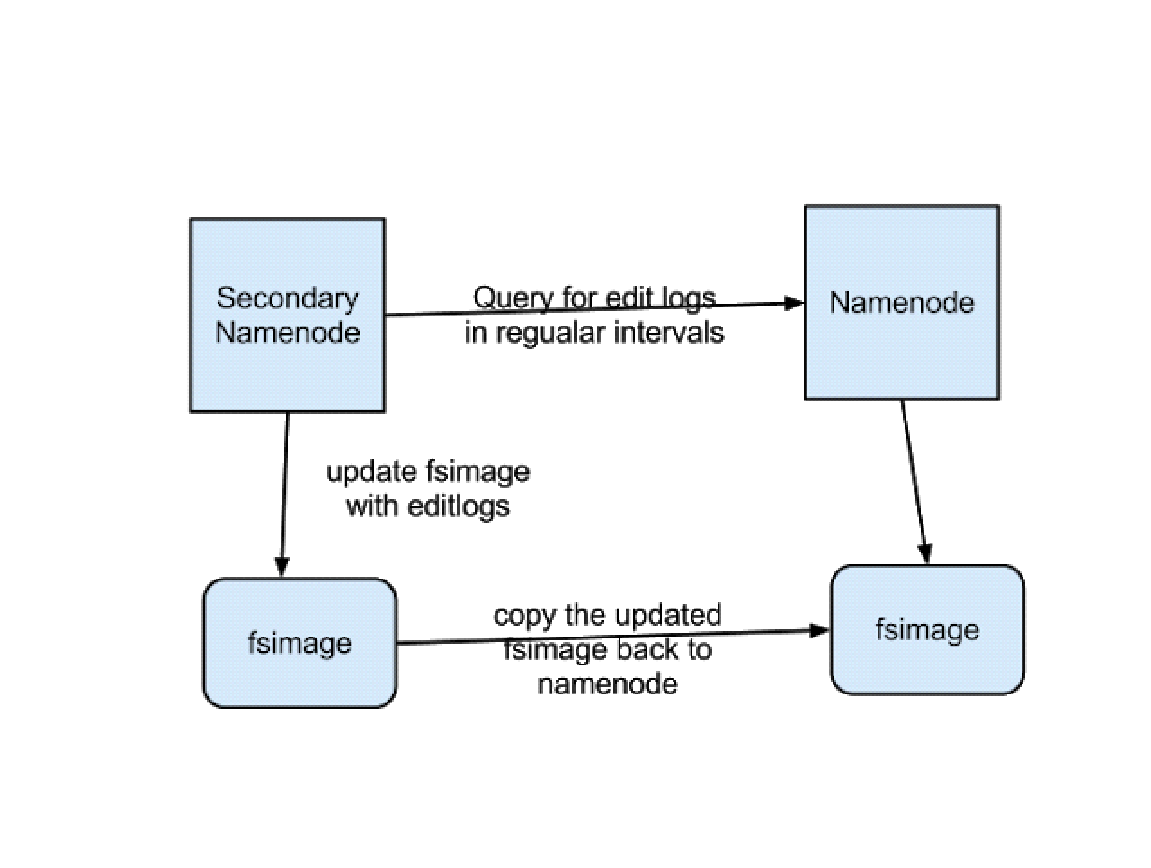
**(71)What is a Secondary Namenode? Is it a substitute to the Namenode?**

Secondary Namenode- It frequently collects back up of edit logs and FS image

Two different files are

fsimage - Its the snapshot of the filesystem when namenode started

Edit logs - Its the sequence of changes made to the filesystem after namenode started



**(72)What is a Task Tracker in Hadoop? How many instances of Task Tracker run on a hadoop cluster**

A TaskTracker is a slave node daemon in the cluster that accepts tasks (Map, Reduce and Shuffle operations) from a JobTracker. There is only One Task Tracker process run on any hadoop slave node. Task Tracker runs on its own JVM process. Every TaskTracker is configured with a set of slots, these indicate the number of tasks that it can accept. The TaskTracker starts a separate JVM processes to do the actual work (called as Task Instance) this is to ensure that process failure does not take down the task tracker. The TaskTracker monitors these task instances, capturing the output and exit codes. When the Task instances finish, successfully or not, the task tracker notifies the JobTracker

**(73)Why ‘Reading‘ is done in parallel and ‘Writing‘ is not in HDFS?**

Writting in parallel will creat inconsistency .

**74)What is a DataNode? How many instances of DataNode run on a Hadoop Cluster?**

A DataNode stores data in the Hadoop File System HDFS. There is only One DataNode process run on any hadoop slave node. DataNode runs on its own JVM process. On startup, a DataNode connects to the NameNode. DataNode instances can talk to each other, this is mostly during replicating data

**(75)How NameNode Handles data node failures?**

NameNode periodically receives a Heartbeat and a Blockreport from each of the DataNodes in the cluster. Receipt of a Heartbeat implies that the DataNode is functioning properly. A Blockreport contains a list of all blocks on a DataNode. When NameNode notices that it has not recieved a hearbeat message from a data node after a certain amount of time, the data node is marked as dead. Since blocks will be under replicated the system begins replicating the blocks that were stored on the dead datanode. The NameNode Orchestrates the replication of data blocks from one datanode to another. The replication data transfer happens directly between datanodes and the data never passes through the namenode.

**(76)How many Daemon processes run on a Hadoop system?**

Hadoop is comprised of five separate daemons. Each of these daemon run in its own JVM. Following 3 Daemons run on Master nodes NameNode - This daemon stores and maintains the metadata for HDFS. Secondary NameNode - Performs housekeeping functions for the NameNode. JobTracker - Manages MapReduce jobs, distributes individual tasks to machines running the Task Tracker. Following 2 Daemons run on each Slave nodes DataNode – Stores actual HDFS data blocks. TaskTracker - Responsible for instantiating and monitoring individual Map and Reduce tasks.

**(77) How many JVMs run on a slave node?**

Single instance of a Task Tracker is run on each Slave node. Task tracker is run as a separate JVM process.

Single instance of a DataNode daemon is run on each Slave node. DataNode daemon is run as a separate JVM process.

One or Multiple instances of Task Instance is run on each slave node. Each task instance is run as a separate JVM process. The number of Task instances can be controlled by configuration. Typically a high end machine is configured to run more task instances.

**(80)How the HDFS Blocks are replicated?**

In pipeline of data node - one data node write on another data node and ack to previous node in pipe line

**(81)What is a Task instance in Hadoop? Where does it run?**

task instance is a mapper task assigned to a data node which will be monitored by task tracker

**(82)what is meaning Replication factor?**

Number of times that block will be replicated on data node in a cluster

**(85)What are the default configuration files that are used in Hadoop?**

core-site.xml

mapred-site.xml

hdfs-site.xml

hadoop-env.sh

yarn-site.xml

**(87)What is HDFS Block size? How is it different from traditional file system block size?**

128mb block size traditional block size is 4kb

**(88)what is SPF?**

single point of failure - If name node fails in version one

**(89)How NameNode Handles data node failures?**

RAA

**(47)Hadoop achieves parallelism by dividing the tasks across many nodes, it is possible for a few slow nodes to rate-limit the rest of the program and slow down the program. What mechanism Hadoop provides to combat this?**

Speculative execution

**(78)What is throughput? How does HDFS get a good throughput?**

In hadoop, the task is divided among different blocks, the processing is done parallel and independent to each other. so because of parallel processing, HDFS has high throughput.

The HDFS is based on Write Once and Read Many Model, it simplifies the data coherency issues as the data is written once can’t be modified and therefore, provides high throughput data access.

Apache Hadoop works on Data Locality principle which states that move computation to data instead of data to computation, this reduces network congestion and therefore, enhances the overall system throughput.

**(89)If we want to copy 10 blocks from one machine to another, but another machine can copy only 8.5 blocks, can the blocks be broken at the time of replication?**

In HDFS, blocks cannot be broken down. Before copying the blocks from one machine to another, the Master node will figure out what is the actual amount of space required, how many block are being used, how much space is available, and it will allocate the blocks accordingly.

**(91)If a data Node is full how it’s identified?**

When data is stored in datanode, then the metadata of that data will be stored in the Namenode. So Namenode will identify if the data node is full.

**(92)If datanodes increase, then do we need to upgrade Namenode?**

While installing the Hadoop system, Namenode is determined based on the

size of the clusters. Most of the time, we do not need to upgrade the Namenode because it does not store the actual data, but just the metadata, so such a requirement rarely arise.

**(104)What if rack 2 and datanode fails?**

Due to replication no data lose will be there as the replica storage is done by following rack awareness algorithm

**(106)What is the difference between Gen1 and Gen2 Hadoop with regards to the Namenode?**

In Gen 1 Hadoop, Namenode is the single point of failure. In Gen 2 Hadoop, we have what is known as Active and Passive Namenodes kind of a structure. If the active Namenode fails, passive Namenode takes over the charge.

**(112)Is a job split into maps?**

job is slit into input split and processed by map

**(120)Can we do online transactions(OLTP) using Hadoop? SRVMTrainings**

No

**(236)What is Schema on Read and Schema on Write?**

schema on read is defining the schema at read time and this is followed in hadoop

**(238)What is Internal and External table in Hive?**

Internal tables are called managed tables and external tables are only pointer to the tables and if we delete that tables then

**(239)what is Small File Problem in Hadoop**

The Hadoop Distributed File System- [HDFS](http://data-flair.training/blogs/comprehensive-hdfs-guide-introduction-architecture-data-read-write-tutorial/) is a distributed file system. [Hadoop](http://data-flair.training/blogs/hadoop-introduction-tutorial-quick-guide/) is mainly designed for batch processing of large volume of data. The default [Data Block](http://data-flair.training/blogs/data-blocks-hdfs-hadoop-distributed-file-system/) size of HDFS is 128 MB. When file size is significantly smaller than the block size the efficiency degrades.  
Mainly there are two reasons for producing small files:

* Files could be the piece of a larger logical file. Since HDFS has only recently supported appends, these unbounded files are saved by writing them in chunks into HDFS.
* Another reason is some files cannot be combined together into one larger file and are essentially small. e.g. – A large corpus of images where each image is a distinct file.

**(137)Whats the default port that jobtrackers listens ?**

**50030**

**(241)What should be the ideal replication factor in Hadoop?**

To decide what would be the ideal replication factor for your Hadoop cluster, analyse these three parameters.  
1. Find out how many nodes are there in the cluster, make sure you cannot have more replicas than nodes in your cluster.  
2. How much data you plan to store ? Your expected usage.  
3. Your cluster's storage capacity

**(242)What is the optimal block size in HDFS?**

128mb

**(243)explain Metadata in Namenode**

In [Hadoop](http://data-flair.training/blogs/hadoop-introduction-tutorial-quick-guide/), Namenode has 2 types of metadata:

1) File System Metadata: It his having all the information about file i.e what are the permission to the particular file, time-stamp, size, replication factor, [Data Blocks](http://data-flair.training/blogs/data-blocks-hdfs-hadoop-distributed-file-system/) of that file. It further categorized into two categories:

a) fsimage: It is having the snapshot of the file system with which Namenode will start for very first time. It always persisted to our disk due to which we don't loose data even after shut down of Namenode.

b) edits(Journal): Whatever operations i.e write, delete, update, etc, it all stores in Journal and hence it keeps on growing during the life of cluster.  
Journal will be committed to fsimage after either Rebooting or using Checkpoints.

Checkpoints: It commits the Journal to fsimage.

2) Bitmap: It is having information about all the Mapping between block and data node i.e which block on which node. This metadata will never persist on dis

What is the port number for NameNode, Task Tracker and Job Tracker?   
  
 NameNode 50070   
Job Tracker 50030   
Task Tracker 50060

How can you debug Hadoop code?   
  
 First, check the list of MapReduce jobs currently running. Next, we need to see that there are no orphaned jobs running; if yes, you need to determine the location of RM logs.  
1. Run: “ps –ef | grep –I ResourceManager”  
and look for log directory in the displayed result. Find out the job-id from the displayed list and check if there is any error message associated with that job. 2. On the basis of RM logs, identify the worker node that was involved in execution of the task.  
3. Now, login to that node and run – “ps –ef | grep –iNodeManager”  
4. Examine the Node Manager log. The majority of errors come from user level logs for each map-reduce job.

32. Elaborate Hadoop Process?

RPC Port:

* 8020- NameNode
* 8021: Job Tracker

Http Port:

* 50070 – name Node
* 50075 – DataNode
* 50090 – Secondary namenode
* 50030 – JobTracker
* 50060 – TaskTracker

66. What are the types of schedulers?

FIFO: Default scheduler it is. It schedule the Jobs in First In First Out format.

FAIR: Give priority dynamically

CAPACITY: Give priority in % to process a job. Highly recommendable in 2.x

I have 20000 records in a table. I want copy them to two separate files( records equally distributed) into HDFS (using Sqoop).   
How do we achieve this, if table does not have primary key or unique key?

go ahead with your sqoop import query with two mappers (-m 2).

**(305)Safe-mode execeptions**

cannot create or modify data in safe mode

**(306)What is the meaning of the term "non-DFS used" in Hadoop web-console?**

The configurations which will be kept for local machine use

**(307)What is AMI**

An Amazon Machine Image (AMI) is a special type of virtual appliance that is used to instantiate (create) a virtual machine within the Amazon Elastic Compute Cloud (""EC2"")

**(308)Can we submit the mapreduce job from slave node?**

No

**(313)How to change replication factor of files already stored in HDFS**

>hdfs dfs -setrep -w 3 /user/dataflair/dir1

**(315)How can one set space quota in Hadoop (HDFS) directory**

The space quota is a hard limit on the number of bytes used by files in the tree rooted at that directory. Block allocations fail if the quota would not allow a full block to be written. Each replica of a block counts against the quota

**(316)How can one increase replication factor to a desired value in Hadoop?**

<property>

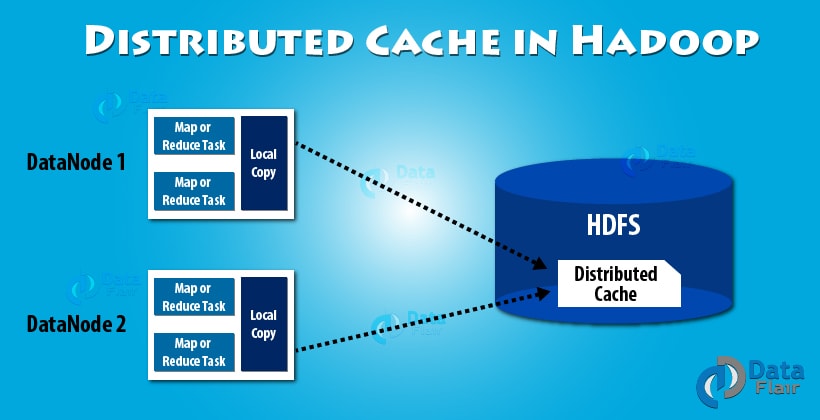
<name>dfs.replication</name>

<value>(desired value)</value>

</property>

>hdfs dfs -setrep -w 3 /user/dataflair/dir1

**(52)What is Distributed Cache in Hadoop?**



**DistributedCache** is a facility provided by the Map-Reduce framework to **cache** files needed by applications. Once you **cache** a file for your job,**hadoop** framework will make it available on each and every data nodes (in file system, not in memory) where you map/reduce tasks are running.

**(54)Is it possible to have Hadoop job output in multiple directories? If yes, how?**

**- Overwrite it**

**- Warn you and continue**

**- Throw an exception and exit**

Yes, it is possible to have the output of Hadoop MapReduce Job written to multiple directories.

In **[Hadoop MapReduce](http://data-flair.training/blogs/hadoop-mapreduce-introduction-tutorial-comprehensive-guide/)**, the output of Reducer is the final output of a Job, and thus its written in to the Hadoop Local File System(HDFS).

The task of writing the output to the HDFS file is done by RecordWriter with the help of OutputFormat.

[**OutputFormat**](http://data-flair.training/blogs/outputformat-recordwriter-hadoop-mapreduce/) is an Interface that defines the way in which the (Key, Value) pair produced by Reducer written to the Output Files. TextOutputFormat is the default OutputFormat.

There is an abstract class called FileOutputFormat which is the base class for all file-based OutputFormats.  
And by default this class writes the output to files based on the number of reducers within the same directory.  
FileOutputFormat and its subclasses generate a set of output files in the output directory.  
There is one file per reducer, and files are named by the partition number: part-00000, part-00001, etc.

But to write output to multiple files we use the class called MultipleFileOutputFormat.  
This abstract class extends the FileOutputFormat, allowing to write the output data to different output files.

**(94)When we send a data to a node, do we allow settling in time, before sending another data to that node?**

Yes we do

**(113)Which are the two types of ‘writes’ in HDFS?**

Data can be written to HDFS two ways, using httpfs (port 14000) or webhdfs (port 50070).

**(115)Can Hadoop be compared to NOSQL database like Cassandra?**

Apache **Cassandra** is a **NoSQL database** ideal for high-speed, online transactional data, while **Hadoop** is a **big data** analytics system that focuses on data warehousing and data lake use cases. ... **As** a result, you **can**process very large data sets very quickly

**(138)unable to read options file while i tried to import data from mysql to hdfs.**

$cat > import.txt

import

--connect

jdbc:mysql://localhost/hadoopdb

--username

hadoop

-P

sqoop --options-file /user/cloudera/import.txt --table employee

**(141)What is the difference between a Hadoop and Relational Database and Nosql?**

|  |  |  |
| --- | --- | --- |
| **Hadoop** | **RDBMS** | **Nosql** |
| **Not a database** | **Database** | **NoSQL** is created especially as a **database** framework |
| **Both** | **Structured data** | **Unstructured data** |
| **Any format data** | **Row and column** | **Column format data** |

**(155)Explain the core methods of the Reducer?**

1)setup () – This method of the reducer is used for configuring various parameters like the input data size, distributed cache, heap size, etc.

Function Definition- public void setup (context)

2)reduce () it is heart of the reducer which is called once per key with the associated reduce task.

Function Definition -public void reduce (Key,Value,context)

3)cleanup () - This method is called only once at the end of reduce task for clearing all the temporary files.

**(157)Is it possible to provide multiple input to Hadoop? If yes then how can you give multiple directories as input to the Hadoop job**

Yes, It is possible to use multiple inputs in **[hadoop](http://data-flair.training/blogs/hadoop-tutorial-for-beginners/)**. There are various ways in which this can be done -

**1.** If Multiple input files are present in the same directory - By default hadoop doesnt read the directory recursively. But suppose if multiple input files like data1, data2,etc are present in /folder1, then Set mapreduce.input.fileinputformat.input.dir.recursive to true and then use FileInputFormat.addInputPath to specify the input directory.This can also be done in driver class code by adding FileInputFormat.setInputDirRecursive(job, true); Before FileInputFormat.addInputPath(job, new Path(args[0])); in your Map Reduce Code.

**2.** Use FileInputFormat.addInputPaths() method, that can take a comma separated list of multiple inputs ex -  
FileInputFormat.addInputPaths(“user1/file0.gz,user2.file.gz…………”).

**3.** Use Multiple mappers.

**4.** Use input file in [**Distributed Cache**](http://data-flair.training/blogs/hadoop-distributed-cache/).

**5.** Use MultipleInputs.addInputPath() method tospecify different input files like -  
MultipleInputs.addInputPath(job, ClouderaPath, TextInputFormat.class, JoinclouderaMapper.class);  
MultipleInputs.addInputPath(job, HdpPath, TextInputFormat.class, HdpMapper.class)

**(262)mkdir: org.apache.hadoop.hdfs.server.namenode.SafeModeException: Cannot create directory**

/user/hadoop/inpdata. Name node is in safemode.

**(269)In which location Name Node sores its Metadata and why?**

1) edit logs files  
2) FsImage files  
These files are available on namenode disk(persistent data storage).

**(121)Explain how HDFS communicates with Linux native file system**

**(270)Should we use RAID in Hadoop or not?**

**No**

**(278)When Hadoop Enter in Safe Mode**

During the startup of namenode daemon, the namenode enters safe mode for a certain period of time

To be asked to Viraj

**(205)Did you ever built a production process in Hadoop ? If yes then what was the process when your hadoop job fails due to any reason**

**(309)How to resolve small file problem in hdfs?**

**(304)What is crontab? Explain with suitable example.**

**(301)How to handle bad records during parsing?**

**(92)Which object can be used to get the progress of a particular job**

**(130)What other technologies have you used in hadoop stack?**