* + - * [1. What is Big Data ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/" \l "1_What_is_Big_Data)
      * [2. What is Hadoop ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#2_What_is_Hadoop)
      * [3. Who uses Hadoop ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#3_Who_uses_Hadoop)
      * [4. What is commodity hardware?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#4_What_is_commodity_hardware)
      * [5. What is the basic difference between traditional RDBMS and Hadoop?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#5_What_is_the_basic_difference_between_traditional_RDBMS_and_Hadoop)
      * [6. What are the modes in which Hadoop can run ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#6_What_are_the_modes_in_which_Hadoop_can_run)
      * [7. What are main components/projects in Hadoop architecture ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#7_What_are_main_componentsprojects_in_Hadoop_architecture)
      * [8. List important default configuration files in Hadoop cluster ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#8_List_important_default_configuration_files_in_Hadoop_cluster)
      * [9. List important site-specific configuration files in Hadoop cluster ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#9_List_important_site-specific_configuration_files_in_Hadoop_cluster)
      * [10. How many hadoop daemon processes run on a Hadoop System ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#10_How_many_hadoop_daemon_processes_run_on_a_Hadoop_System)
      * [11. How to start all hadoop daemons at a time ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#11_How_to_start_all_hadoop_daemons_at_a_time)
      * [12. If some hadoop daemons are already running and if we need to start any one remaining daemon process then what are the commands to use ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#12_If_some_hadoop_daemons_are_already_running_and_if_we_need_to_start_any_one_remaining_daemon_process_then_what_are_the_commands_to_use)
      * [13. How to stop all the three hadoop daemons at a time ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#13_How_to_stop_all_the_threehadoopdaemons_at_a_time)
      * [14. What are commands that need to be used to bring down a single hadoop daemon?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#14_What_are_commands_that_need_to_be_used_to_bring_down_a_single_hadoop_daemon)
      * [15. How many YARN daemon processes run on a cluster ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#15_How_many_YARN_daemon_processes_run_on_a_cluster)
      * [16. How to start Yarn daemon processes on a hadoop cluster ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#16_How_to_start_Yarn_daemon_processes_on_a_hadoop_cluster)
      * [17. How to verify whether the daemon processes are running or not ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#17_How_to_verify_whether_the_daemon_processes_are_running_or_not)
      * [18. How to bring down the Yarn daemon processes ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#18_How_to_bring_down_the_Yarn_daemon_processes)
      * [19. Can we start both Hadoop daemon processes and Yarn daemon processes with a single command?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#19_Can_we_start_both_Hadoop_daemon_processes_and_Yarn_daemon_processes_with_a_single_command)
      * [20. Can we stop all the above five daemon processes with a single command ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#20_Can_we_stop_all_the_above_five_daemon_processes_with_a_single_command)
      * [21. Which operating systems are supported for Hadoop deployment ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#21_Which_operating_systems_are_supported_for_Hadoop_deployment)
      * [22. How could be the various components of Hadoop cluster deployed in production?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#22_How_could_be_the_various_components_of_Hadoop_cluster_deployed_in_production)
      * [23. What is structured and unstructured data?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#23What_is_structured_and_unstructured_data)
      * [24. Is Namenode also a commodity?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#24_Is_Namenode_also_a_commodity)
      * [25. What is the difference between jps and jps -lm commands ?](http://hadooptutorial.info/hadoop-interview-questions-part-1/#25_What_is_the_difference_between_jps_and_jps_-lm_commands)

**Hadoop Interview Questions and Answers  
1. What is Big Data ?**

Big data is vast amount of data (generally in GBs or TBs of size) that exceeds the regular processing capacity of the traditional computing servers and requires special parallel processing mechanism. This data is too big and its rate of increase gets accelerated. This data can be either structural or unstructured data which may not be able to process by legacy databases.

**2. What is Hadoop ?**

Hadoop is an open source frame work from Apache Software Foundation for storing & processing large-scale data usually called Big Data using clusters of commodity hardware.

**3. Who uses Hadoop ?**

Big organizations in which data grows exponentially day by day and must require Hadoop platform to process such huge data. For example , Facebook, Google, Amazon, Twitter, IBM, LinkedIn etc… companies uses hadoop technology to solve their big data processing problems.

**4. What is commodity hardware?**

Commodity hardware is a non-expensive system which is not of high quality or high-availability.

Hadoop can be installed on any commodity hardware. We don’t need super computers or high-end hardware to work on Hadoop. Commodity hardware includes RAM because there will be some services which will be running on RAM.

**5. What is the basic difference between traditional RDBMS and Hadoop?**

Traditional RDBMS is used for transactional systems to report and archive the data, whereas Hadoop is an approach to store huge amount of data in the distributed file system and process it.

RDBMS will be useful when we want to seek one record from Big data, whereas, Hadoop will be useful when we want Big data in one shot and perform analysis on that later.

**6. What are the modes in which Hadoop can run ?**

Hadoop can run in three modes.

* Stand alone or Local mode – No daemons will be running in this mode and everything runs in a single JVM.
* Pseudo distributed mode – All the Hadoop daemons run on a local machine, simulating cluster on a small scale.
* Fully distributed mode – A cluster of machines will be setup in master/slaves architecture to distribute and process the data across various nodes of commodity hardware.

**7. What are main components/projects in Hadoop architecture ?**

* Hadoop Common: The common utilities that support the other Hadoop modules.
* HDFS: Hadoop distributed file system that provides high-throughput access to application data.
* Hadoop YARN: A framework for job scheduling and cluster resource management.
* Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

**8. List important default configuration files in Hadoop cluster ?**

The default configuration files in hadoop cluster are:

* core-default.xml
* hdfs-default.xml
* yarn-default.xml
* mapred-default.xml

**9. List important site-specific configuration files in Hadoop cluster ?**

In order to override any hadoop configuration property’s default values, we need to provide configuration values in site-specific configuration files. Below are the four site-specific .xml configuration files and environment variable setup file.

* core-site.xml : Common properties are configured in this file.
* hdfs-site.xml : Site specific hdfs properties are configured in this file
* yarn-site.xml : Yarn specific properties can be provided in this file.
* mapred-site.xml : Mapreduce framework specific properties will defined here.
* hadoop-env.sh : Hadoop environment variables are setup in this file.

All these configuration files should be placed in hadoop’s configuration directory etc/hadoop from hadoop’s home directory.

**10. How many hadoop daemon processes run on a Hadoop System ?**

As of hadoop-2.5.0 release, three hadoop daemon processes run on a hadoop cluster.

* NameNode daemon – Only one daemon runs for entire hadoop cluster.
* Secondary NameNode daemon – Only one daemon runs for entire hadoop cluster.
* DataNode daemon – One datanode daemon per each datanode in hadoop cluster

**11. How to start all hadoop daemons at a time ?**

$ start-dfs.sh command can be used to start all hadoop daemons from terminal at a time.

**12. If some hadoop daemons are already running and if we need to start any one remaining daemon process then what are the commands to use ?**

Instead of start-dfs.sh which will trigger all the hadoop three daemons at a time, we can also start running each daemon separately by the below commands.

$ hadoop-daemon.sh start namenode

$ hadoop-daemon.sh start secondarynamenode

$ hadoop-daemon.sh start datanode

**13. How to stop all the three hadoop daemons at a time ?**

By using stop-dfs.sh command, we can stop the above three daemon processes with a single command.

**14. What are commands that need to be used to bring down a single hadoop daemon?**

Below hadoop-daemon.sh commands can be used to bring down each hadoop daemon separately.

$ hadoop-daemon.sh stop namenode

$ hadoop-daemon.sh stop secondarynamenode

$ hadoop-daemon.sh stop datanode

**15. How many YARN daemon processes run on a cluster ?**

Two types of Yarn daemons will be running on hadoop cluster in master/slave fashion.

* ResourceManager – Master daemon process
* NodeManager – One Slave daemon process per node in a cluster.

**16. How to start Yarn daemon processes on a hadoop cluster ?**

By running $ start-yarn.sh command from terminal on each machine on hadoop cluster, Yarn daemons can be started.

**17. How to verify whether the daemon processes are running or not ?**

By using java’s processes command $ jps to check what are all the java processes running on a machine. This command lists down all the daemon processes running on a machine along with their process ids.

**18. How to bring down the Yarn daemon processes ?**

Using $ stop-yarn.sh command, we can bring down both the Yarn daemon processes running on a machine.

**19. Can we start both Hadoop daemon processes and Yarn daemon processes with a single command?**

Yes, we can start all the above mentioned five daemon processes (3 hadoop + 2 Yarn) with a single command $ start-all.sh

**20. Can we stop all the above five daemon processes with a single command ?**

Yes, by using $ stop-all.sh command all the above five daemon processes can be bring down in a single shot.

**21. Which operating systems are supported for Hadoop deployment ?**

The only supported operating system for hadoop’s production deployment is Linux. However, with some additional software Hadoop can be deployed on Windows for test environments.

**22. How could be the various components of Hadoop cluster deployed in production?**

Both Name Node and Resource Manager can be deployed on a Master Node, and Data nodes and node managers can be deployed on multiple slave nodes.

There is a need for only one master node for namenode and Resource Manager on the system. The number of slave nodes for datanodes & node managers depends on the size of the cluster.

One more node with hardware specifications same as master node will be needed for secondary namenode.

**23. What is structured and unstructured data?**

**Structured data** is the data that is easily identifiable as it is organized in a structure. The most common form of structured data is a database where specific information is stored in tables, that is, rows and columns.

**Unstructured data** refers to any data that cannot be identified easily. It could be in the form of images, videos, documents, email, logs and random text. It is not in the form of rows and columns.

**24. Is Namenode also a commodity?**

No. Namenode can never be a commodity hardware because the entire HDFS rely on it. It is the single point of failure in HDFS. Namenode has to be a high-availability machine.

**25. What is the difference between**jps**and**jps -lm**commands ?**

jps command returns the process id and short names for running processes. But jps -lm returns long messages along with process id and short names as shown below.

hadoop1@ubuntu-1:~$ jps

5314 SecondaryNameNode

5121 DataNode

5458 Jps

4995 NameNode

hadoop1@ubuntu-1:~$ jps -lm

5314 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode

5121 org.apache.hadoop.hdfs.server.datanode.DataNode

5473 sun.tools.jps.Jps -lm

4995 org.apache.hadoop.hdfs.server.namenode.NameNode

* + - * [1. What is HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#1_What_is_HDFS)
      * [2. What are the objectives of HDFS file system?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#2_What_are_the_objectives_of_HDFS_file_system)
      * [3. What are the limitations of HDFS file systems?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#3_What_are_the_limitations_of_HDFS_file_systems)
      * [4. What is a block in HDFS and what is its size?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#4_What_is_a_block_in_HDFS_and_what_is_its_size)
      * [5. What are the core components in HDFS Cluster ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#5_What_are_the_core_components_in_HDFS_Cluster)
      * [6. What is a NameNode ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#6_What_is_a_NameNode)
      * [7. What is a DataNode ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#7_What_is_a_DataNode)
      * [8. Is Namenode machine same as datanode machine as in terms of hardware?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#8_Is_Namenode_machine_same_as_datanode_machine_as_in_terms_of_hardware)
      * [9. What is a Secondary Namenode?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#9_What_is_a_Secondary_Namenode)
      * [10. What is a Checkpoint Node?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#10_What_is_a_Checkpoint_Node)
      * [11. What is a checkpoint ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#11_What_is_a_checkpoint)
      * [12. What is a daemon?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#12_What_is_a_daemon)
      * [13. What is a heartbeat in HDFS?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#13_What_is_a_heartbeat_in_HDFS)
      * [14. Are Namenode and Resource Manager run on the same host?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#14_Are_Namenode_and_Resource_Manager_run_on_the_same_host)
      * [15. What is the communication mode between namenode and datanode?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#15_What_is_the_communication_mode_between_namenode_and_datanode)
      * [16. If we want to copy 20 blocks from one machine to another, but another machine can copy only 18.5 blocks, can the blocks be broken at the time of replication?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#16_If_we_want_to_copy_20_blocks_from_one_machine_to_another_but_anothermachine_can_copy_only_185_blocks_can_the_blocks_be_broken_at_the_time_ofreplication)
      * [17. How indexing is done in HDFS?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#17_How_indexing_is_done_in_HDFS)
      * [18. If a data Node is full, then how is it identified?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#18_If_a_data_Node_is_full_then_how_is_it_identified)
      * [19. If datanodes increase, then do we need to upgrade Namenode?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#19_If_datanodes_increase_then_do_we_need_to_upgrade_Namenode)
      * [20. Why Reading is done in parallel and Writing is not in HDFS?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#20Why_Reading_is_done_in_parallel_and_Writing_is_not_in_HDFS)
      * [21. What is fsimage and edit log in hadoop?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-2/#21_What_is_fsimage_and_edit_log_in_hadoop)

### Hadoop Interview Questions and Answers

###### 1. What is HDFS ?

HDFS is a distributed file system implemented on Hadoop’s framework. It is a block-structured distributed file system designed to store vast amount of data on low cost commodity hardware and ensuring high speed process on data.

HDFS stores files across multiple machines and maintains reliability and fault tolerance. HDFS support parallel processing of data by Mapreduce framework.

###### 2. What are the objectives of HDFS file system?

* Easily Store large amount of data across multiple machines
* Data reliability and fault tolerance by maintaining multiple copies of each block of a file.
* Capacity to move computation to data instead of moving data to computation server. I.e. processing data locally.
* Able to provide parallel processing of data by Mapreduce framework.

###### 3. What are the limitations of HDFS file systems?

* HDFS supports file operations reads, writes, appends and deletes efficiently but it doesn’t support file updates.
* HDFS is not suitable for large number of small sized files but best suits for large sized files. Because file system namespace maintained by Namenode is limited by it’s main memory capacity as namespace is stored in namenode’s main memory and large number of files will result in big **fsimage** file.

###### 4. What is a block in HDFS and what is its size?

It is a fixed size chunk of data usually of size **128 MB**. It is the minimum of size of data that HDFS can read/write.

HDFS files are broken into these fixed size chunks of data across multiple machines on a cluster.

Thus, blocks are building bricks of a HDFS file. Each block is maintained in at least 3 copies as mentioned by replication factor in Hadoop configuration to provide data redundancy and maintain fault-tolerance.

###### 5. What are the core components in HDFS Cluster ?

* Name Node
* Secondary Name Node
* Data Nodes
* Checkpoint Nodes
* Backup Node

###### 6. What is a NameNode ?

Namenode is a dedicated machine in HDFS cluster which acts as a master serve that maintains file system namespace in its main memory and serves the file access requests by users. File system namespace mainly contains fsimage and edits files. Fsimage is a file which contains file names, file permissions and block locations of each file.

Usually only one active namenode is allowed in HDFS default architecture.

###### 7. What is a DataNode ?

DataNodes are slave nodes of HDFS architecture which store the blocks of HDFS files and sends blocks information to namenode periodically through heart beat messages.

Data Nodes serve read and write requests of clients on HDFS files and also perform block creation, replication and deletions.

###### 8. Is Namenode machine same as datanode machine as in terms of hardware?

It depends upon the cluster we are trying to create. The Hadoop VM can be there on the same machine or on another machine. For instance, in a single node cluster, there is only one machine,whereas in the development or in a testing environment, Namenode and datanodes are on different machines.

###### 9. What is a Secondary Namenode?

The Secondary NameNode is a helper to the primary NameNode. Secondary NameNode is a specially dedicated node in HDFS cluster whose main function is to take checkpoints of the file system metadata present on namenode. It is not a backup namenode and doesn’t act as a namenode in case of primary namenode’s failures. It just checkpoints namenode’s file system namespace.

For in depth details please refer the [post here](http://hadooptutorial.info/what-is-secondary-namenode/)**.**

###### 10. What is a Checkpoint Node?

It is an enhanced secondary namenode whose main functionality is to take checkpoints of namenode’s file system metadata periodically. It replaces the role of secondary namenode. Advantage of Checkpoint node over the secondary namenode is that it can upload the result of merge operation of fsimage and edits log files while checkpointing.

For indepth details please refer the [post here.](http://hadooptutorial.info/what-is-checkpoint-node/)

###### 11. What is a checkpoint ?

During Checkpointing process, fsimage file is merged with edits log file and a new fsimage file will be created which is usually called as a checkpoint.

###### 12. What is a daemon?

Daemon is a process or service that runs in background. In general, we use this word in UNIX environment. Hadoop or Yarn daemons are Java processes which can be verified with **jps** command.

###### 13. What is a heartbeat in HDFS?

A heartbeat is a **signal** indicating that it is alive. A datanode sends heartbeat to Namenode and task tracker will send its heart beat to job tracker. If the Namenode or job tracker does not receive heart beat then they will decide that there is some problem in datanode or task tracker is unable to perform the assigned task.

###### 14. Are Namenode and Resource Manager run on the same host?

**No**, in practical environment, Namenode runs on a separate host and Resource Manager runs on a separate host.

###### 15. What is the communication mode between namenode and datanode?

The mode of communication is SSH.

###### 16. If we want to copy 20 blocks from one machine to another, but another machine can copy only 18.5 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.

###### 17. How indexing is done in HDFS?

Hadoop has its own way of indexing. Depending upon the block size, once the data is stored, HDFS will keep on storing the last part of the data which will say where the next part of the data will be. In fact, this is the base of HDFS.

###### 18. If a data Node is full, then how is it 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.

###### 19. 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.

###### 20. Why Reading is done in parallel and Writing is not in HDFS?

Reading is done in parallel because by doing so we can access the data fast. But we do not perform the write operation in parallel because it might result in data written by one node can be overwritten by other.

For example, we have a file and two nodes are trying to write data into the file in parallel, then the first node does not know what the second node has written and vice-versa. So, this makes it confusing which data to be stored and accessed.

###### 21. What is fsimage and edit log in hadoop?

Fsimage is a file which contains file names, file permissions and block locations of each file, and this file is maintained by Namenode for indexing of files in HDFS. We can call it as metadata about HDFS files. The fsimage file contains a serialized form of all the directory and file inodes in the filesystem.

EditLog is a transaction log which contains records for every change that occurs to file system metadata.

**Note: Whenever a NameNode is restarted, the latest status of FsImage is built by applying edits records on last saved copy of FsImage.**

* + - * [1. What is a Backup Node?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#1_What_is_a_Backup_Node)
      * [2. What are the differences between Backup Node and Checkpoint node or Secondary Namenode ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#2_What_are_the_differences_between_Backup_Node_and_Checkpoint_node_or_SecondaryNamenode)
      * [3. What is Safe Mode in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#3_What_is_Safe_Mode_in_HDFS)
      * [4. What is Data Locality in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#4_What_is_Data_Locality_in_HDFS)
      * [5. What is a rack?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#5_What_is_a_rack)
      * [6. What is Rack Awareness ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#6_What_is_Rack_Awareness)
      * [7. How does HDFS File Deletes or Undeletes work?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#7_How_does_HDFS_File_Deletes_or_Undeletes_work)
      * [8. What is a Rebalancer in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#8_What_is_a_Rebalancer_in_HDFS)
      * [9. What is the need for Rebalancer in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-3/#9_What_is_the_need_for_Rebalancer_in_HDFS)

##### Hadoop Interview questions and answers

###### 1. What is a Backup Node?

It is an extended checkpoint node that performs checkpointing and also supports online streaming of file system edits.

It maintains an in memory, up-to-date copy of file system namespace and accepts a real time online stream of file system edits and applies these edits on its own copy of namespace in its main memory.

Thus, it maintains always a latest backup of current file system namespace.

###### 2. What are the differences between Backup Node and Checkpoint node or Secondary Namenode ?

* **Multiple checkpoint nodes** can be registered with namenode but only a **single backup node** is allowed to register with namenode at any point of time.
* To create a checkpoint, checkpoint node or secondary namenode needs to **download fsimage and edits files** from active namenode and apply edits to fsimage and saves a copy of new fsimage as a checkpoint.

But in backup node, **no need to download** fsimage and edits files from active namenode because, it already has an up-to-date copy of fsimage in its main memory and accesses online streaming of edits which are provided by namenode. So, applying these edits into fsimage in its own main memory and saving a copy in local FS.

So, checkpoint creation in backup node is faster than that of checkpoint node or secondary namenode.

* The diff between **checkpoint node** and secondary namenode is that checkpoint node **can upload** the new copy of **fsimage file back to namenode** after checkpoint creation where as a secondary namenode can’t upload but can only store in its local FS.
* Backup node provides the option of running **namenode with no persistent storage**but a checkpoint node or secondary namenode doesn’t provide such option.
* In case of namenode failures, **data loss** in checkpoint node or secondary namenode **is certain** at least to a minimum amount of data due to time gap between two checkpoints.

But in backup node, data loss is not certain and it maintains namespace which is in sync with namenode at any point of time.

###### 3. What is Safe Mode in HDFS ?

Safe Mode is a maintenance state of NameNode during which Name Node doesn’t allow any changes to the file system.

During Safe Mode, HDFS cluster is ready-only and doesn’t replicate or delete blocks.

Name Node automatically enters safe mode during its start up and maintain blocks replication value within minimum and maximum allowable replication limit.

###### 4. What is Data Locality in HDFS ?

One of the HDFS design idea is that **“Moving Computation is cheaper than Moving data”.**

If data sets are huge, running applications on nodes where the actual data resides will give efficient results than moving data to nodes where applications are running.

This concept of moving applications to data, **is called Data Locality.**

This **reduces network traffic** and **increases speed** of data processing and accuracy of data since there is **no chance of data loss** during data transfer through network channels because there is no need to move data.

###### 5. What is a rack?

Rack is a **storage area with all the datanodes** put together. These datanodes can be physically located at different places. Rack is a physical collection of datanodes which are stored at a single location. There can be multiple racks in a single location.

###### 6. What is Rack Awareness ?

The concept of maintaining Rack Id information by NameNode and using these rack ids for choosing closest data nodes for HDFS file read or writes requests **is called Rack Awareness.**

By choosing closest data nodes for read/writes request through rack awareness policy, **minimizes the write cost and maximizing read speed.**

###### 7. How does HDFS File Deletes or Undeletes work?

When a file is deleted from HDFS, it will not be removed immediately from HDFS, but HDFS moves the file into **/trash**directory. After certain period of time interval, the NameNode deletes the file from the HDFS **/trash** directory. The deletion of a file releases the blocks associated with the file.

Time interval for which a file remains in /trash directory can be configured with **fs.trash.interval** property stored in **core-site.xml.**

**As long as a file remains in /trash directory, the file can be undeleted** by moving the file from /trash directory into required location in HDFS. **Default trash interval is set to 0**. So, HDFS Deletes file without storing in trash.

###### 8. What is a Rebalancer in HDFS ?

Rebalancer is a administration tool in HDFS, to balance the distribution of blocks uniformly across all the data nodes in the cluster.

Rebalancing will be done on demand only. It will not get triggered automatically.

HDFS administrator issues this command on request to balance the cluster

$ hdfs balancer

If a Rebalancer is triggered, NameNode will scan entire data node list and when

* Under-utilized data node is found, it moves blocks from over-utilized data nodes or not-under-utilized data nodes to this current data node
* If Over-utilized data node is found, it moves blocks from this data node to other under-utilized or not-over-utilized data nodes.

###### 9. What is the need for Rebalancer in HDFS ?

Whenever a new data node is added to the existing HDFS cluster or a data node is removed from the cluster then some of the data nodes in the cluster will have more/less blocks compared to other data nodes.

In this unbalanced cluster, data read/write requests become very busy on some data nodes and some data nodes are under utilized.

In such cases, to make all the data nodes space is uniformly utilized for blocks distribution, rebalancing will be done by Hadoop Administrator.

* + - * [1. What is the default block size in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#1_What_is_the_default_block_size_in_HDFS)
      * [2. What is the benefit of large block size in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#2_What_is_the_benefit_of_large_block_size_in_HDFS)
      * [3. What are the overheads of maintaining too large Block size ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#3_What_are_the_overheads_of_maintaining_too_large_Block_size)
      * [4. If a file of size 10 MB is copied on to HDFS of block size 256 MB, then how much storage will be allocated to the file on HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#4_If_a_file_of_size_10_MB_is_copied_on_to_HDFS_of_block_size_256_MB_then_how_much_storage_will_be_allocated_to_the_file_on_HDFS)
      * [5. What are the benefits of block structure concept in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#5_What_are_the_benefits_of_block_structure_concept_in_HDFS)
      * [6. What if we upgrade our Hadoop version in which, default block size is higher than the current Hadoop version’s default block size. Suppose 128 MB (Hadoop 0.20.2) to 256 MB (Hadoop 2.4.0).](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#6_What_if_we_upgrade_our_Hadoop_version_in_which_default_block_size_is_higher_than_the_current_Hadoop_versions_default_block_size_Suppose_128_MB_Hadoop_0202_to_256_MB_Hadoop_240)
      * [7. What is Block replication ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#7_What_is_Block_replication)
      * [8. What is default replication factor and how to configure it ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#8_What_is_default_replication_factor_and_how_to_configure_it)
      * [9. What is HDFS distributed copy (distcp) ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#9_What_is_HDFS_distributed_copy_distcp)
      * [10. What is the use of fsck command in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-and-answers-part-4/#10_What_is_the_use_offsckcommand_in_HDFS)

###### 1. What is the default block size in HDFS ?

As of Hadoop-2.4.0 release, the default block size in HDFS is 256 MB and prior to that it was 128 MB.

###### 2. What is the benefit of large block size in HDFS ?

Main benefit of large block size is Quick Seek Time. The time to transfer a large file of multiple blocks operates at the disk transfer rate instead of depending much on seek time.

###### 3. What are the overheads of maintaining too large Block size ?

Usually in Mapreduce framework, each map task operate on one block at a time. So, having too few blocks result in too few map tasks running in parallel for longer time which finally results in overall slow down of job performance.

###### 4. If a file of size 10 MB is copied on to HDFS of block size 256 MB, then how much storage will be allocated to the file on HDFS ?

Even though the default HDFS block size is 256 MB, a file which is smaller than a single block doesn’t occupy full block size. So, in this case, the file will occupy just 10 MB but not 256 MB.

###### 5. What are the benefits of block structure concept in HDFS ?

* Main benefit is that the ability to store very large files which can be even larger than the size of single disk (node) as the file is broken into blocks and distributed across various nodes on cluster.
* Another important advantage is simplicity of storage management as the blocks are fixed size, it is easy to calculate how many can be stored on a given disk.
* Blocks replication feature is useful in fault tolerance.

###### 6. What if we upgrade our Hadoop version in which, default block size is higher than the current Hadoop version’s default block size. Suppose 128 MB (Hadoop 0.20.2) to 256 MB (Hadoop 2.4.0).

All the existing files are maintained at block size of 128 MB but any new files copied on to upgraded hadoop are broken into blocks of size 256 MB.

###### 7. What is Block replication ?

Block replication is a way of maintaining multiple copies of same block across various nodes on cluster to achieve fault tolerance. In this, though one of the data node containing the block becomes dead, the block data can be obtained from other live data nodes which contain the same copy of the block data.

###### 8. What is default replication factor and how to configure it ?

The default replication factor in fully distributed HDFS is 3.

This can be configured with **dfs.replication**in**hdfs-site.xml**file at site level.

Replication factor can be setup at file level with below FS command.

$ hadoop fs -setrep N /filename

In above command ‘N’ is the new replication factor for the file “/filename".

###### 9. What is HDFS distributed copy (distcp) ?

distcp is an utility for launching MapReduce jobs to copy large amounts of HDFS files within or in between HDFS clusters.

Syntax for using this tool.

hadoop distcp hdfs://namenodeX/src hdfs://namenodeY/dest

###### 10. What is the use of fsck command in HDFS ?

HDFS **fsck** command is a useful to get the files and blocks details of the file system. It’s syntax is:

$ hadoop fsck<path> [-move|-delete|-openforwrite][-files[-blocks[-locations|-racks]]]

|  |  |
| --- | --- |
|  |  |

below are the command options and their purpose.

|  |  |
| --- | --- |
| -move | Move corrupted files to /lost+found |
| -delete | Delete corrupted files. |
| -openforwrite | Print out files opened for write. |
| -files | Print out files being checked. |
| -blocks | Print out block report. |
| -locations | Print out locations for every block. |
| -racks | Print out network topology for data-node locations. |

* [1. As 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?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#1_As_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)
* [2. What happens if you get a ‘connection refused java exception’ when you type hadoop fs -ls /?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#2_What_happens_if_you_get_a_connection_refused_java_exception_when_you_type_hadoop_fs_-ls)
* [3. Can we have multiple entries in the master files?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#3_Can_we_have_multiple_entries_in_the_master_files)
* [4. Why do we need a password-less SSH in Fully Distributed environment?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#4_Why_do_we_need_a_password-less_SSH_in_Fully_Distributed_environment)
* [5. Does this lead to security issues?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#5_Does_this_lead_to_security_issues)
* [6. which port does SSH work on?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#6_which_port_does_SSH_work_on)
* [7. If we only had 32 MB of memory how can we sort one terabyte of data?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#7_If_we_only_had_32_MB_of_memory_how_can_we_sort_one_terabyte_of_data)
* [8. How can we create an empty file in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#8_How_can_we_create_an_empty_file_in_HDFS)
* [9. How can we see the contents of a Snappy compressed file or SequenceFile via command line?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#9_How_can_we_see_the_contents_of_a_Snappy_compressed_file_or_SequenceFile_via_command_line)
* [10. How can we check the existence of a file in HDFS ?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#10_How_can_we_check_the_existence_of_a_file_in_HDFS)
* [11. How can we set the replication factor of directory via command line?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#11_How_can_we_set_the_replication_factor_of_directory_via_command_line)
* [12. How can we apply the permission to all the files under a HDFS directory recursively?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#12_How_can_we_apply_the_permission_to_all_the_files_under_a_HDFS_directory_recursively)
* [13. What is hadoop fs -stat command used for?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#13_What_is_hadoop_fs_-stat_command_used_for)
* [14. What is expunge command in HDFS and why is it used for ?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#14_What_is_expunge_command_in_HDFS_and_why_is_it_used_for)
* [15. How can we see the output of a MR job as a single file if the reducer might have created multiple part-r-0000\* files ?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#15_How_can_we_see_the_output_of_a_MR_job_as_a_single_file_if_the_reducer_might_have_created_multiple_part-r-0000_files)
* [16. Which of the following is most important when selecting hardware for our new Hadoop cluster?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#16_Which_of_the_following_is_most_important_when_selecting_hardware_for_our_newHadoop_cluster)
* [17. Why would you likely not want to use network storage in your cluster?](http://hadooptutorial.info/hadoop-interview-questions-answers-part-5/#17_Why_would_you_likely_not_want_to_use_network_storage_in_your_cluster)

###### 1. As 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?

Since there are 3 nodes, when we send the MapReduce programs, calculations will be done only on the original data. The master node will know which node exactly has that particular data. In case, if one of the nodes is not responding, it is assumed to be failed. Only then, the required calculation will be done on the second replica.

###### 2. What happens if you get a ‘connection refused java exception’ when you type hadoop fs -ls /?

It could mean that the Namenode is not working on our hadoop cluster.

###### 3. Can we have multiple entries in the master files?

**Yes**, we can have multiple entries in the Master files.

###### 4. Why do we need a password-less SSH in Fully Distributed environment?

We need a password-less SSH in a Fully-Distributed environment because when the cluster is LIVE and running in Fully Distributed environment, the communication is too frequent. The Resource Manager/Namenode should be able to send a task to Node manager /Datanodes quickly.

###### 5. Does this lead to security issues?

**No**, not at all. Hadoop cluster is an isolated cluster. And generally it has nothing to do with an internet. It has a different kind of a configuration. We needn’t worry about that kind of a security breach, for instance, someone hacking through the internet, and so on. Hadoop has implemented Kerberose for strong security way to connect to other machines to fetch and to process data.

###### 6. which port does SSH work on?

SSH works on Port **22**, though it can be configured. 22 is the default Port number.

###### 7. If we only had 32 MB of memory how can we sort one terabyte of data?

Take a smaller chunk of data and sort it in memory, so we partition it in lots of little data sets. Then merge those sorted lists into one big list before writing the results back to disk.

###### 8. How can we create an empty file in HDFS ?

We can create empty file with **hadoop fs -touchz** command. It creates a zero byte file.

$ hadoop fs -touchz /user/hadoop/filename

###### 9. How can we see the contents of a Snappy compressed file or SequenceFile via command line?

With the help of **$ hadoop fs -text /user/hadoop/filename** command we can see the contents of sequencefile or any compressed format file in text format.

###### 10. How can we check the existence of a file in HDFS ?

The hadoop test is used for file test operations. The syntax is shown below:

$ hadoop fs -test -[ezd] URI

Here “e” for checking the existence of a file, “z” for checking the file is zero length or not, “d” for checking the path is a directory or not. On success, the test command returns 1 else 0.

###### 11. How can we set the replication factor of directory via command line?

Hadoop **setrep** is used to change the replication factor of a file. Use the **-R** option for recursively changing the replication factor.

$ hadoop fs -setrep -w 4 -R /user/hadoop/dir

###### 12. How can we apply the permission to all the files under a HDFS directory recursively?

Using **$ hadoop fs -chmod -R 755 /user/hadoop/dir** command we can set the permissions to a directory recursively.

###### 13. What is hadoop fs -stat command used for?

Hadoop stat returns the stats information of a file. It returns the last updated date and time. The syntax of stat is shown below:

$ hadoop fs -stat /filename

2014-09-28 18:34:06

###### 14. What is expunge command in HDFS and why is it used for ?

Hadoop fs expunge command is used to empty the trash directory in HDFS. Syntax is:

$ hadoop fs –expunge

###### 15. How can we see the output of a MR job as a single file if the reducer might have created multiple part-r-0000\* files ?

We can use hadoop fs -getmerge command to combine all the part-r-0000\* files into single file and this file can be browsed to view the entire results of the MR job at a time. Syntax is:

$ hadoop fs -getmerge &lt;hdfssrc&gt; &lt;localdestination&gt; [addnl]

The addnl option is for adding new line character at the end of each file.

**16. Which of the following is most important when selecting hardware for our new Hadoop cluster?**

1. The number of CPU cores and their speed.

2. The amount of physical memory.

3. The amount of storage.

4. The speed of the storage.

5. It depends on the most likely workload.

Answer 5 – Though some general guidelines are possible and we may need to generalize whether our cluster will be running a variety of jobs, the best fit depends on the anticipated workload.

17. Why would you likely not want to use network storage in your cluster?

1. Because it may introduce a new single point of failure.

2. Because it most likely has approaches to redundancy and fault-tolerance that may

be unnecessary given Hadoop’s fault tolerance.

3. Because such a single device may have inferior performance to Hadoop’s use of

multiple local disks simultaneously.

4. All of the above.

Answer 4: Network storage comes in many flavors but in many cases we may find a large Hadoop cluster

of hundreds of hosts reliant on a single (or usually a pair of) storage devices. This adds a new failure scenario to the cluster and one with a less uncommon likelihood than many others. Where storage technology does look to address failure mitigation it is usually through disk-level redundancy.

**Real Time questions:**

1. Hive – Where do you use Internal or Managed table? What scenarios?
2. In your resume, what do you mean by, “monitoring & managing MapReduce jobs"? Explain?
3. Interviewer’s Project: How to modify the RDBMs’ Nested SQL queries into Hadoop framework using Pig.
4. Sqoop: Need to know very well. Some of the current projects are importing data from other RDBMs sources into HDFS.
5. Can you join or transform tables/columns when importing using Sqoop?
6. Can you do the above with different RDBMs (not clear)?
7. How do you transfer flat files from Unix systems?
8. What is your Pig/Hive programming level (1- 10)? (Almost all interviewers asked this.)
9. Learn Scala! – Interviewer repeatedly told me.

Other Interview Questions:

1. Hive – Interval vs External How do you save your files in Hive
2. Sqoop – Incremental vs hast modified relate to your project
3. Sqoop – How to check if RDBMS Table Columns added/removed and how to incorporate these changes into the import job.
4. What are the challenges you’ve faced in your project? Give 2 examples.
5. How do you check Data Integrity (log files)
6. How to improve performance in your script (PIG)?
7. Tell me about your project? work.
8. How do you use Partitioning/Bucketing in your project? (Examples from your project)
9. Where do you look for answers? (user groups, Apache Web, stack overflow)
10. NOSQL- HBase – Unstructured data storage?
11. How to debug Production issue? Give example. (logs, script counters, JVM)
12. Data Ingestion
13. What is the file size you’ve used?

Dev. Environmet

Production Environmet

1. Does Hive support indexing? (How does this relate to Partition and Bucketing)
2. Pig support Conditional Loop?
3. Hive – What type of data stored?
4. Recruiter: In your experience, what is the jump from DB developer to Hadoop without Java experience?

More Technical type Interview Questions:

1. What functions did you use in PIG?
2. Filter – What did you filter out?
3. Join – What did you join?
4. What is your cluster size?
5. What is the file size for production environment?
6. How long does it take to run your script in Production cluster?
7. Are you planning for anything to improve the performance?
8. What size of file do you use for Development?
9. What did you work on HBase?
10. Why Hadoop? compare to RDBMS.
11. Hive – What did you do to increase the performance.
12. PIG – what did you do to increase the performance
13. What Java UDF did you write?
14. What scenario do you think you can use Java for?
15. You can process log files in RDBMS too. Why Hadoop?
16. Hive partitioning – your project example? Why?

 Hive – What file format do you use in your work? (Avro, Parquet, Sequence file)

1. Hadoop – What is the challenge or difficulty you’ve faced?
2. PIG – What is the challenge or difficulty you’ve faced?
3. Flume – What is the challenge or difficulty you’ve faced?
4. Sqoop – What is the challenge or difficulty you’ve faced? (he didn’t ask this question)
5. How experienced are you in Linux?
6. What shell type do you use?
7. How about your experience in Cloudera Manager?
8. How about your experience in Cloudera Manager?
9. Do you use Impala? (I compared it with Hive and explained in more details)
10. How do you select the Eco system tools for your project?

InfoSys – Interview Questions:

As you can see, questions are mostly based on theory.

1. Why Hadoop? (Compare to RDBMS)
2. What would happen if NameNode failed? How do you bring it up?
3. What details are in the “fsimage" file?
4. What is SecondaryNameNode?
5. Explain the MapReduce processing framework? (start to end)
6. What is Combiner? Where does it fit and give an example? Preferably from your project.
7. What is Partitioner? Why do you need it and give an example? Preferably from your project.
8. Oozie – What are the nodes?
9. What are the actions in Action Node?
10. Explain your Pig project?
11. What log file loaders did you use in Pig?
12. Hive Joining? What did you join?
13. Explain Partitioning & Bucketing (based on your project)?
14. Why do we need bucketing?
15. Did you write any Hive UDFs?
16. Filter – What did you filter out?
17. HBase?
18. Flume?
19. Sqoop?
20. Zookeeper?
21. Impala? Explain the use of Impala?
22. Cassandra? What do you know about Cassandra?
23. ClickStream.
24. What is your cluster size?
25. What is the DataNode configurations? (RAM, CPU core, Disk size)
26. What is the NameNode configurations? (RAM, CPU core, Disk size)
27. How many Map slots & reducer slots configured in each DataNode? (he didn’t ask this)
28. How do you copy file from cluster to cluster?
29. What commands do you use to check to check system health, jobs, etc.?
30. Do you use Cloudera Manager to monitor and manage the jobs, cluster, etc.?
31. What is Speculative execution?
32. What do you know about Scala? (interviewer asked about the skills that I listed in my resume)

Java Interview Questions:

Had an array of the follwing elements: [29 12 24 18 -11 -5]  
Need an O/P of sorting of arrays ,== [12 18 24 29 -5 -11]  
Need an O/P of even and odd numbers in array ,==[12 18 24] && [29 -5 -11]//Declaring an araylist  
ArrayList<Integer> arraylist = new ArrayList<Integer>();/\* Sorting of arraylist using Collections.sort\*/Collections.sort(arraylist);

for(int counter: arraylist)  
{  
System.out.println(counter);  
}

/\*Sort array in reverse order\*/

Collections.reverseOrder(arraylist);

System.out.println(“\*\*\*\*\*\* Reverse Sorted String Array \*\*\*\*\*\*\*”);  
for (int i : stringArray)  
{  
System.out.println(i);  
}

/\* sort an array to even numbers and odd numbers\*/  
public class SortNumbers  
{  
private static int[] array = {12 18 24 29 -5 -11};  
private static List<Integer> even = new ArrayList<>();  
private static List<Integer> odd = new ArrayList<>();

public static void even(int[] arr, List even , List odd)  
{  
for(int i = 0 ; i < arr.length ; i++)  
{  
if(arr[i] % 2 == 0)  
even.add(arr[i]);  
else  
odd.add(arr[i]);  
}  
}

//To Display the even and Odd numbers  
public static void display(List<Integer> list)  
{ for(Integer i : list)  
System.out.println(i);  
}

public static void main(String[] args){  
classify(array,even,odd);  
display(even);  
}  
}  
}

2)How to make your class compatible with Java Hash Maps?  
Overriding hashcode() and equals() method.

3)You have two tables Employee and Dept with the below columns.Select Maximum salary by Department.

Employee EMPID NAME SAL DEPTID  
DEPT dept\_id dept\_name

SELECT d.dept\_name, MAX(e.SAL) FROM Employee e,Dept d where (d.dept\_id = e.dept\_id) group by

On 07/28/2015

1. Tell me some List implementations?

ArrayList

Linkedlist

2. In what Purposes you use ArrayList and Linkedlist?

ArrayList for fast searching,

LinkedList,for more insertions/deletes

3. In both Arraylist and Linkedlist, which is faster?

ArrayList is faster as it containis duplicates, no sorting

Linkedlist is slow as it contains adding and removing of elements

4. Tell me some Map implementations?

HashMap (unsorted)

TreeMap (Sorted values)

LinkedHashMap( if you want near-HashMap performance and insertion-order iteration)

5. Which of the Map implementations is faster and why?

Hash map is fast as there is no need of extra burden in sorting values…

6. What Happens in Shuffle Phase in Map Reduce?

All the part files will be exchanged between reduce tasks

part files will be generated by partitioners

map output will be transferred over network…

7. What is the Fundemental Data Structure inside a HashMap?

Integer, For calculating hash value for all keys stored into buckets…Buckets are used as storage

locations…Usually Buckets are array….

8. How do you use Map Reduce methods?

map is method to parse the input records

reduce for aggregating the results reading input from map

9. What are the parameters in Mapper class

map(key, value, context)

10. What is the interface on a Main function on a Mapper?

In Mapper Class you write…..

setup() and map(key, value, context)—-( return type of map method is void…but it writes output to context)

cleanup()

11. Is it possible to get multiple Key,value pairs from the Map phase?

Yes, by concatenating two or more fields into same field.

12. Imagine you have a Server Class Computer, If you have two files of 1 GB each on Hard disk,

These files consists of Integers from smaller to larger, how do you Merge the files into one File

and generate an output of Sorted Order? Tell me the Logic

Read record by record from each file and compare first record from first file with first record in 2nd

file and same way with 2nd record b/w the files….

If first rec in 1st file < 1st record 2nd file then i will emit 1st record in 1st file and i will move cursor

of first file to 2nd record in the first file then check with 1st record in 2nd file and so on…

13. What if there are no records in one of the files in the above Scenario?

I will copy records from the remaining file as it is without comparing

14. What is the execution time of the above Program?

1-2 minutes…in Hadoop

15. If you have two files of 1 TB on two disks, you should Merge the files into one File and generate

an output of Sorted Order? What will you do?

Write all the above logic in map method of map reduce job….or reduce method

16. How the records of the two files are compared in the Map Reducer Phase?

If one of the file is small then i can read that into memory through distributed cache in setup

method of mapper class

17. What Problems you face in the Reducer Phase?

Out of Memory Problem (To overcome this problem increase the heap size mapreduce.child.java.opts)

http://www.hadooptpoint.org/accenture-hadoop-interview-questions-and-answers-for-experienced/