## ****Hadoop MapReduce Questions****

* ***Definition:***

MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment.

* ***MapReduce 2.0 or YARN Architecture:***
  + MapReduce framework also follows Master/Slave Topology where the master node (Resource Manager) manages and tracks various MapReduce jobs being executed on the slave nodes (Node Mangers).
  + Resource Manager consists of two main components:
    - **Application Manager:** It accepts job-submissions, negotiates the container for ApplicationMaster and handles failures while executing MapReduce jobs.
    - **Scheduler:** Scheduler allocates resources that is required by various MapReduce application running on the Hadoop cluster.
* ***How MapReduce job works:***
  + As the name MapReduce suggests, reducer phase takes place after mapper phase has been completed.
  + So, the first is the map job, where a block of data is read and processed to produce key-value pairs as intermediate outputs.
  + The reducer receives the key-value pair from multiple map jobs.
  + Then, the reducer aggregates those intermediate data tuples (intermediate key-value pair) into a smaller set of tuples or key-value pairs which is the final output.

## ****1. What do you mean by data locality?****

[***Data locality***](http://www.edureka.co/blog/mapreduce-tutorial/?utm_source=blog&utm_medium=content-link&utm_campaign=hadoop-interview-questions-mapreduce/#data_locality) talks about moving computation unit to data rather data to the computation unit. MapReduce framework achieves data locality by processing data locally i.e. processing of the data happens in the very node by Node Manager where data blocks are present.

## ****2. Is it mandatory to set input and output type/format in MapReduce?****

No, it is not mandatory to set the input and output type/format in MapReduce. By default, the cluster takes the input and the output type as ‘text’.

## ****3. Can we rename the output file?****

Yes, we can rename the output file by implementing multiple format output class.

## ****4. What do you mean by shuffling and sorting in MapReduce?****

Shuffling and sorting takes place after the completion of map task where the input to the every reducer is sorted according to the keys. Basically, the process by which the system sorts the key-value output of the map tasks and transfer it to the reducer is called shuffle.

## ****5. Explain the process of spilling in MapReduce?****

The output of a map task is written into a circular memory buffer (RAM). The default size of buffer is set to 100 MB  which can be tuned by using mapreduce.task.io.sort.mb property. Now, spilling is a process of copying the data from memory buffer to disc when the content of the buffer reaches a certain threshold size. By default, a background thread starts spilling the contents from memory to disc after 80% of the buffer size is filled. Therefore, for a 100 MB size buffer the spilling will start after the content of the buffer reach a size of 80 MB.

**Note:** One can change this spilling threshold using mapreduce.map.sort.spill.percent which is set to 0.8 or 80% by default.

## ****6. What is a distributed cache in MapReduce Framework?****

Distributed Cache can be explained as, a facility provided by the MapReduce framework to cache files needed by applications. Once you have cached a file for your job, Hadoop framework will make it available on each and every data nodes where you map/reduce tasks are running. Therefore, one can access the cache file as a local file in your Mapper or Reducer job.

## ****7. What is a combiner and where you should use it?****

Combiner is like a mini reducer function that allow us to perform a local aggregation of map output before it is transferred to reducer phase. Basically, it is used to optimize the network bandwidth usage during a MapReduce task by cutting down the amount of data that is transferred from a mapper to the reducer.

## ****8. Why the output of map tasks are stored (spilled ) into local disc and not in HDFS?****

The outputs of map task are the intermediate key-value pairs which is then processed by reducer to produce the final aggregated result. Once a MapReduce job is completed, there is no need of the intermediate output produced by map tasks. Therefore, storing these intermediate output into HDFS and replicate it will create unnecessary overhead.

## ****9. What happens when the node running the map task fails before the map output has been sent to the reducer?****

In this case, map task will be assigned a new node and whole task will be run again to re-create the map output.

## ****10. Define Speculative Execution****

If a node appears to be executing a task slower than expected, the master node can redundantly execute another instance of the same task on another node. Then, the task which finishes first will be accepted whereas other task will be killed. This process is called speculative execution.

## ****11. What is the role of a MapReduce Partitioner?****

A partitioner divides the intermediate key-value pairs produced by map tasks into partition. The total number of partition is equal to the number of reducers where each partition is processed by the corresponding reducer. The partitioning is done using the hash function based on a single key or group of keys. The default partitioner available in Hadoop is HashPartitioner.

## ****12. How can we assure that the values regarding a particular key goes to the same reducer?****

By using a partitioner we can control that a particular key – value goes to the same reducer for processing.

## ****13. What is the difference between Input Split and HDFS block?****

HDFS block defines how the data is physically divided in HDFS whereas input split defines the logical boundary of the records required for processing it.

## ****14. What do you mean by InputFormat?****

InputFormat describes the input-specification for a MapReduce job.The MapReduce framework relies on the InputFormat of the job to:

* Validate the input-specification of the job.
* Split-up the input file(s) into logical InputSplit instances, each of which is then assigned to an individual Mapper.
* Provide the RecordReader implementation used to read records from the logical InputSplit for processing by the Mapper.

## ****15. What is the purpose of TextInputFormat?****

TextInputFormat is the default input format present in the MapReduce framework. In TextInputFormat, an input file is produced as keys of type LongWritable (byte offset of the beginning of the line in the file) and values of type Text (content of the line).

## ****16. What is the role of RecordReader in Hadoop MapReduce?****

InputSplit defines a slice of work, but does not describe how to access it. The “RecordReader” class loads the data from its source and converts it into (key, value) pairs suitable for reading by the “Mapper” task. The “RecordReader” instance is defined by the “Input Format”.

## ****17. What are the various configuration parameters required to run a MapReduce job?****

The main [***configuration parameters***](http://www.edureka.co/blog/mapreduce-tutorial/?utm_source=blog&utm_medium=content-link&utm_campaign=hadoop-interview-questions-mapreduce/#explanation_of_mapreduce_program) which users need to specify in “MapReduce” framework are:

* Job’s input locations in the distributed file system
* Job’s output location in the distributed file system
* Input format of data
* Output format of data
* Class containing the map function
* Class containing the reduce function
* JAR file containing the mapper, reducer and driver classes

## ****18. When should you use SequenceFileInputFormat?****

SequenceFileInputFormat is an input format for reading within sequence files. It is a specific compressed binary file format which is optimized for passing the data between the outputs of one “MapReduce” job to the input of some other “MapReduce” job.

Sequence files can be generated as the output of other MapReduce tasks and are an efficient intermediate representation for data that is passing from one MapReduce job to another.

## ****19. What is an identity Mapper and Identity Reducer?****

Identity mapper is the default mapper provided by the Hadoop framework. It runs when no mapper class has been defined in the MapReduce program where it simply passes the input key – value pair for the reducer phase.

Like Identity Mapper, Identity Reducer is also the default reducer class provided by the Hadoop, which is automatically executed if no reducer class has been defined. It also performs no computation or process, rather it just simply write the input key – value pair into the specified output directory.

## ****20. What is a map side join?****

Map side join is a process where two data sets are joined by the mapper.

## ****21. What are the advantages of using map side join in MapReduce?****

The advantages of using map side join in MapReduce are as follows:

* Map-side join helps in minimizing the cost that is incurred for sorting and merging in the shuffle and reduce stages.
* Map-side join also helps in improving the performance of the task by decreasing the time to finish the task.

## ****22. What is reduce side join in MapReduce?****

As the name suggests, in the reduce side join, the reducer is responsible for performing the join operation. It is comparatively simple and easier to implement than the map side join as the sorting and shuffling phase sends the values having identical keys to the same reducer and therefore, by default, the data is organized for us.

♣**Tip**: I would suggest you to go through a dedicated blog on [***reduce side join***](http://www.edureka.co/blog/mapreduce-example-reduce-side-join/?utm_source=blog&utm_medium=content-link&utm_term=hadoop-interview-questions-mapreduce) in MapReduce where the whole process of reduce side join is explained in detail with an example.

## 23. What do you know about NLineInputFormat?

NLineInputFormat splits ‘n’ lines of input as one split.

## ****24. Is it legal to set the number of reducer task to zero? Where the output will be stored in this case?****

Yes, It is legal to set the number of reduce-tasks to zero if there is no need for a reducer. In this case the outputs of the map task is directly stored into the HDFS which is specified in the setOutputPath(Path).

## ****25. Is it necessary to write a MapReduce job in Java?****

No, MapReduce framework supports multiple languages like Python, Ruby etc.

## ****26. How do you stop a running job gracefully?****

One can gracefully stop a MapReduce job by using the command: hadoop job -kill JOBID

## ****27. How will you submit extra files or data ( like jars, static files, etc. ) for a MapReduce job during runtime?****

The distributed cache is used to distribute large read-only files that are needed by map/reduce jobs to the cluster. The framework will copy the necessary files from a URL on to the slave node before any tasks for the job are executed on that node. The files are only copied once per job and so should not be modified by the application.

## ****28. How does inputsplit in MapReduce determines the record boundaries correctly?****

RecordReader is responsible for providing the information regarding record boundaries in an input split.

## ****29. How do reducers communicate with each other?****

This is a tricky question. The “MapReduce” programming model does not allow “reducers” to communicate with each other. “Reducers” run in isolation.

## Top Answers to Map Reduce Interview Questions

**1. Compare MapReduce and Spark**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **MapReduce** | **Spark** |
| Processing Speeds | Good | Exceptional |
| Standalone mode | Needs Hadoop | Can work independently |
| Ease of use | Needs extensive Java program | APIs for Python, Java, & Scala |
| Versatility | Real-time & machine learning applications | Not optimized for real-time & machine learning applications |

**2. What is MapReduce?**

Referred as the core of Hadoop, MapReduce is a programming framework to process large sets of data or big data across thousands of servers in a Hadoop Cluster. The concept of MapReduce is similar to the cluster scale-out data processing systems. The term MapReduce refers to two important processes of Hadoop program operates.

First is the map() job, which converts a set of data into another breaking down individual elements into key/value pairs (tuples). Then comes reduce() job into play, wherein the output from the map, i.e. the tuples serve as the input and are combined into smaller set of tuples. As the name suggests, the map job every time occurs before the reduce one.

**Learn more about MapReduce in this insightful**[*Hadoop MapReduce – What it Refers To?*](https://intellipaat.com/blog/hadoop-mapreduce-what-it-refers-to/)**now.**

**Learn MapReduce in 18 hrs. Download e-book now**

**GET CERTIFIED**

**3. Illustrate a simple example of the working of MapReduce.**

Let’s take a simple example to understand the functioning of MapReduce. However, in real-time projects and applications, this is going to be elaborate and complex as the data we deal with Hadoop and MapReduce is extensive and massive.

Assume you have five files and each file consists of two key/value pairs as in two columns in each file – a city name and its temperature recorded. Here, name of city is the key and the temperature is value.  
San Francisco, 22  
Los Angeles, 15  
Vancouver, 30  
London, 25  
Los Angeles, 16  
Vancouver, 28  
London,12

It is important to note that each file may consist of the data for same city multiple times. Now, out of this data, we need to calculate the maximum temperature for each city across these five files. As explained, the MapReduce framework will divide it into five map tasks and each map task will perform data functions on one of the five files and returns maxim temperature for each city.

(San Francisco, 22)(Los Angeles, 16)(Vancouver, 30)(London, 25)  
Similarly each mapper performs it for the other four files and produce intermediate results, for instance like below.

(San Francisco, 32)(Los Angeles, 2)(Vancouver, 8)(London, 27)  
(San Francisco, 29)(Los Angeles, 19)(Vancouver, 28)(London, 12)  
(San Francisco, 18)(Los Angeles, 24)(Vancouver, 36)(London, 10)  
(San Francisco, 30)(Los Angeles, 11)(Vancouver, 12)(London, 5)

These tasks are then passed to the reduce job, where the input from all files are combined to output a single value. The final results here would be:

(San Francisco, 32)(Los Angeles, 24)(Vancouver, 36)(London, 27)

These calculations are perform instantly and are extremely efficient to calculate outputs on a large dataset.

Are you interested in learning MapReduce? Well, we have the comprehensive[*big data hadoop training*](https://intellipaat.com/big-data-hadoop-training/) Course to give you a head start in your career.

**4. What are the main components of MapReduce Job?**

Main Driver Class: providing job configuration parameters  
Mapper Class: must extend org.apache.hadoop.mapreduce.Mapper class and performs execution of map() method  
Reducer Class: must extend org.apache.hadoop.mapreduce.Reducer class

**5. What is Shuffling and Sorting in a MapReduce?**

Shuffling and Sorting are two major processes operating simultaneously during the working of mapper and reducer.

The process of transferring data from Mapper to reducer is Shuffling. It is a mandatory operation for reducers to proceed their jobs further as the shuffling process serves as input for the reduce tasks.

In MapReduce, the output key-value pairs between the map and reduce phases (after the mapper) are automatically sorted before moving to the Reducer. This feature is helpful in programs where you need sorting at some stages. It also saves the programmer’s overall time.

This MapReduce Tutorial will give you a [clear understanding of MapReduce](https://intellipaat.com/tutorial/mapreduce-tutorial/)

**6. What is Partitioner and its usage?**

Partitioner is yet another important phase that controls the partitioning of the intermediate map-reduce output keys using a hash function. The process of partitioning determines in what reducer, a key-value pair (of the map output) is sent. The number of partitions is equal to the total number of reduce jobs for the process.

Hash Partitioner is the default class available in Hadoop , which implements the following function.int getPartition(K key, V value, int numReduceTasks)  
The function returns the partition number using the numReduceTasks is the number of fixed reducers.

Learn about MapReduce and its various components in this [java mapreduce tutorial](https://intellipaat.com/blog/java-mapreduce-tutorial/).

[**Wish to Learn Mapreduce? Click Here**](https://intellipaat.com/hadoop-mapreduce-training/#course-content)

**7. What is Identity Mapper and Chain Mapper?**

Identity Mapper is the default Mapper class provided by Hadoop. when no other Mapper class is defined, Identify will be executed. It only writes the input data into output and do not perform and computations and calculations on the input data.

The class name is org.apache.hadoop.mapred.lib.IdentityMapper.

Chain Mapper is the implementation of simple Mapper class through chain operations across a set of Mapper classes, within a single map task. In this, the output from the first mapper becomes the input for second mapper and second mapper’s output the input for third mapper and so on until the last mapper.

The class name is org.apache.hadoop.mapreduce.lib.ChainMapper.

**8. What main configuration parameters are specified in MapReduce?**

The MapReduce programmers need to specify following configuration parameters to perform the map and reduce jobs:

* The input location of the job in HDFs.
* The output location of the job in HDFS.
* The input’s and output’s format.
* The classes containing map and reduce functions, respectively.
* The .jar file for mapper, reducer and driver classes

Learn about [overview, dataflow and performance conditions of MapReduce](https://en.wikipedia.org/wiki/MapReduce)

**9. Name Job control options specified by MapReduce.**

Since this framework supports chained operations wherein an input of one map job serves as the output for other, there is a need for job controls to govern these complex operations.

The various job control options are:

**Job.submit() :** to submit the job to the cluster and immediately return

**Job.waitforCompletion(boolean) :** to submit the job to the cluster and wait for its completion

**Learn more about**[hadoop online course](https://intellipaat.com/big-data-hadoop-training/#curriculum)**to get ahead in your career!**

**10. What is InputFormat in Hadoop?**

Another important feature in MapReduce programming, InputFormat defines the input specifications for a job. It performs the following functions:

* Validates the input-specification of job.
* Split the input file(s) into logical instances called InputSplit. Each of these split files are then assigned to individual Mapper.
* Provides implementation of RecordReader to extract input records from the above instances for further Mapper processing

**11. What is the difference between HDFS block and InputSplit?**

An HDFS block splits data into physical divisions while InputSplit in MapReduce splits input files logically.

While InputSplit is used to control number of mappers, the size of splits is user defined. On the contrary, the HDFS block size is fixed to 64 MB, i.e. for 1GB data , it will be 1GB/64MB = 16 splits/blocks. However, if input split size is not defined by user, it takes the HDFS default block size.

**12. What is Text Input Format?**

It is the default InputFormat for plain text files in a given job having input files with .gz extension. In TextInputFormat, files are broken into lines, wherein key is position in the file and value refers to the line of text. Programmers can write their own InputFormat.  
The hierarchy is:

java.lang.Object

org.apache.hadoop.mapreduce.InputFormat

org.apache.hadoop.mapreduce.lib.input.FileInputFormat

org.apache.hadoop.mapreduce.lib.input.TextInputFormat

**13. What is JobTracker?**

JobTracker is a Hadoop service used for the processing of MapReduce jobs  in the cluster. It submits and tracks the jobs to specific nodes having data. Only one JobTracker runs on single Hadoop cluster on its own JVM process. if JobTracker goes down, all the jobs halt.

**Go for the**[online Hadoop training](https://intellipaat.com/big-data-hadoop-training/)**for mastering MapReduce.**

**14. Explain job scheduling through JobTracker.**

JobTracker communicates with NameNode to identify data location and submits the work to TaskTracker node. The TaskTracker plays a major role as it notifies the JobTracker for any job failure. It actually is referred to the heartbeat reporter reassuring the JobTracker that it is still alive. Later, the JobTracker is responsible for the actions as in it may either resubmit the job or mark a specific record as unreliable or blacklist it.

**15. What is SequenceFileInputFormat?**

A compressed binary output file format to read in sequence files and extends the FileInputFormat.It passes data between output-input (between output of one MapReduce job to input of another MapReduce job)phases of MapReduce jobs.

**16. How to set mappers and reducers for Hadoop jobs?**

Users can configure JobConf variable to set number of mappers and reducers.

job.setNumMaptasks()

job.setNumreduceTasks()

Check this insightful blog to find out what are the [must-have skills of Hadoop professionals](https://intellipaat.com/blog/must-skills-for-hadoop-professionals/)?

**17. Explain JobConf in MapReduce.**

It is a primary interface to define a map-reduce job in the Hadoop for job execution. JobConf specifies mapper, Combiner, partitioner, Reducer,InputFormat , OutputFormat implementations and other advanced job faets liek Comparators.

**18. What is a MapReduce Combiner?**

Also known as semi-reducer, Combiner is an optional class to combine the map out records using the same key. The main function of a combiner is to accept inputs from Map Class and pass those key-value pairs to Reducer class

**19. What is RecordReader in a Map Reduce?**

RecordReader is used to read key/value pairs form the InputSplit by converting the byte-oriented view  and presenting record-oriented view to Mapper.

**20. Define Writable data types in MapReduce.**

Hadoop reads and writes data in a serialized form in writable interface. The Writable interface has several classes like Text (storing String data), IntWritable, LongWriatble, FloatWritable, BooleanWritable. users are free to define their personal Writable classes as well.

Read this blog to see how the processor speeds of Hadoop are increasing and Learn more about [the Mapper and Reducer parts of MapReduce](https://intellipaat.com/blog/hadoop-processor-rising-speed-of-big-data-technologies/)

**21. What is OutputCommitter?**

OutPutCommitter describes the commit of MapReduce task. FileOutputCommitter is the default available class available for OutputCommitter in MapReduce. It performs the following operations:

* Create temporary output directory for the job during initialization.
* Then, it cleans the job as in removes temporary output directory post job completion.
* Sets up the task temporary output.
* Identifies whether a task needs commit. The commit is applied if required.
* JobSetup, JobCleanup and TaskCleanup are important tasks during output commit.

**22. What is a “map” in Hadoop?**

In Hadoop, a map is a phase in HDFS query solving. A map reads data from an input location, and outputs a key value pair according to the input type.

**23. What is a “reducer” in Hadoop?**

In Hadoop, a reducer collects the output generated by the mapper, processes it, and creates a final output of its own.

**Get the**[*hadoop developer training course*](https://intellipaat.com/hadoop-developer-training/)**for an unbelievable price now!**

**24. What are the parameters of mappers and reducers?**

The four parameters for mappers are:

* LongWritable (input)
* text (input)
* text (intermediate output)
* IntWritable (intermediate output)

The four parameters for reducers are:

* Text (intermediate output)
* IntWritable (intermediate output)
* Text (final output)
* IntWritable (final output)

**25. What are the key differences between Pig vs MapReduce?**

PIG is a data flow language, the key focus of Pig is manage the flow of data from input source to output store. As part of managing this data flow it moves data feeding it to

**process 1.** taking output and feeding it to

**process2.** The core features are preventing execution of subsequent stages if previous stage fails, manages temporary storage of data and most importantly compresses and rearranges processing steps for faster processing. While this can be done for any kind of processing tasks Pig is written specifically for managing data flow of Map reduce type of jobs. Most if not all jobs in a Pig are map reduce jobs or data movement jobs. Pig allows for custom functions to be added which can be used for processing in Pig, some default ones are like ordering, grouping, distinct, count etc.

Mapreduce on the other hand is a data processing paradigm, it is a framework for application developers to write code in so that its easily scaled to PB of tasks, this creates a separation between the developer that writes the application vs the developer that scales the application. Not all applications can be migrated to Map reduce but good few can be including complex ones like k-means to simple ones like counting uniques in a dataset.

**26. What is partitioning?**

Partitioning is a process to identify the reducer instance which would be used to supply the mappers output. Before mapper emits the data (Key Value) pair to reducer, mapper identify the reducer as an recipient of mapper output. All the key, no matter which mapper has generated this, must lie with same reducer.

**27. How to set which framework would be used to run mapreduce program?**

mapreduce.framework.name. it can be

1. Local
2. classic
3. Yarn

**28. What platform and Java version is required to run Hadoop?**

Java 1.6.x or higher version are good for Hadoop, preferably from Sun. Linux and Windows are the supported operating system for Hadoop, but BSD, Mac OS/X and Solaris are more famous to work.

**29. Can MapReduce program be written in any language other than Java?**

Yes, Mapreduce can be written in many programming languages Java, R, C++, scripting Languages (Python, PHP). Any language able to read from stadin and write to stdout and parse tab and newline characters should work . Hadoop streaming (A Hadoop Utility) allows you to create and run Map/Reduce jobs with any executable or scripts as the mapper and/or the reducer.

**What is partition and combiner in MapReduce?**

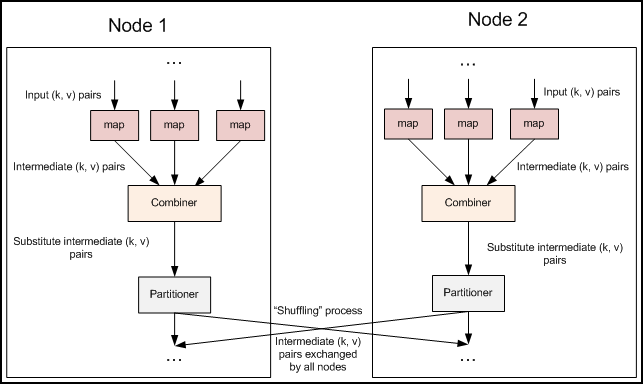
Partition and combiner are the two phase of a MapReduce operation those are executed before the reduce phase and after the map phase. Here are the details of partition and combiner in MapReduce.

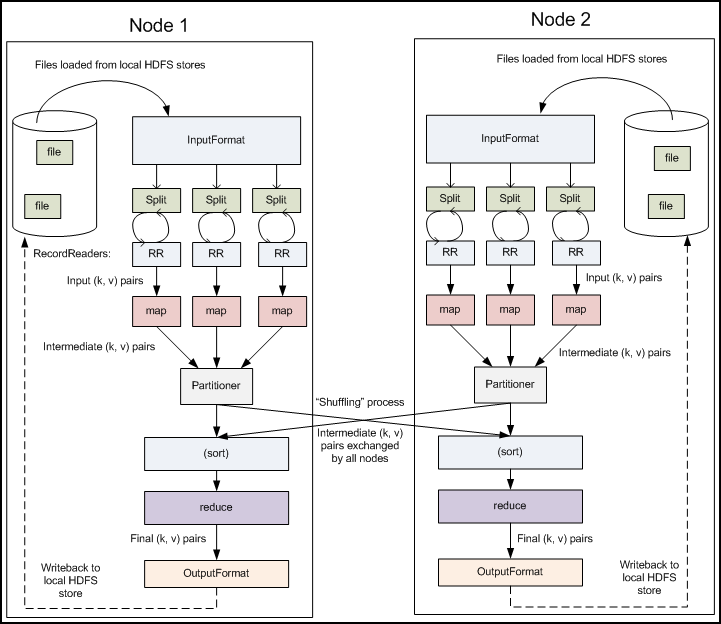
**Combiner:** Combiner works like a mini reducer in Map phase which takes the input from map phase. It performs local reduce function on mapper result before they are distributed further. Once combiner functionality is executed (if required) then the output is passed to the reducer phase.

**Partition:**Partition comes into picture when you are using more than one reducer. Partition decides which reducer is responsible for a particular key.

It takes the input from mapper phase or Combiner phase (if used) and then sends it across the responsible reducer based on the key. The number of partitions is equal to the number of reducers.

So in partition and combiner, combiner comes first and then partition. The below image from Yahoo depicts the operation beautifully.

**When Combiner is used**  


**When Combiner is not being used**  


**2. What is HCatalog?**

HCatalog enables reading and writing of data in any format for which we use SerDe in Hive. By default, HCatalog supports [RC File, CSV, JSON, and Sequence File formats](http://www.hdfstutorial.com/input-file-formats-in-hadoop/). But for custom formats, the user needs to provide InputFormat, OutputFormat, and SerDe information.

It is built on the top of Hive metastore and incorporates components from the Hive DDL. HCatalog also provides the read and write an interface for Pig and MapReduce and uses Hive CLI for issuing commands.

So in short, HCatalog opens up the hive metastore to the other MapReduce tools. As we know every MapReduce tool has its own perception about the HDFS data. PIG consider the data as a set of file while Hive considers it as a set of tables. HCatalog simply simplifies the process.

**Example:**  
Let’s say we have a table (employee) in Hive with the following details-

|  |  |
| --- | --- |
| ID | Name |
| 1 | John |
| 2 | Chris |
| 3 | Peter |
| 4 | Lisa |

Now if you want to use this table to load data in Pig, then you will have to use **pig –useHCatalog** command.

Let’s see how to use the commands to export the data in PIG-

A = LOAD ’employee’ USING org.apache.hcatalog.pig.HCatLoader();

HCatalog can be accessed through webchat which makes use of [rest API](http://www.restapitutorial.com/).

**3. What are Hadoop and Hadoop ecosystems?**

Well, this question can be simply answered by anyone. I am just writing few lines for it.

Hadoop is an open source java-based programming framework which is used to process and store large data sets in distributed environment.

Hadoop is one of the top projects of [Apache Software Foundation](https://www.apache.org/). Also, Hadoop makes use of commodity hardware for its nodes ([DataNodes](http://www.hdfstutorial.com/important-hadoop-terminology/)) and so maintaining a low-cost system.

Here are some of the Hadoop ecosystems which are frequently being used-

* **HDFS:** To store the large sets of data
* **Hive:** To process structured data
* **Pig:** To process unstructured data
* **Oozie:** Create workflow jobs
* **Flume:** Get real-time data from other sources
* **MapReduce:** Data analysis
* **HBase:** NoSQL database used for record level operation
* **Sqoop:** Import/Export data to and from RDBMS to Hadoop system
* **Kafka:** For messaging

There are many another component in Hadoop ecosystem, but the above are important and mostly used.

**4. What is view? At first, we have created a view on top of the table (two rows- empid, empname). Then added 100 Rows on the table whether the newly added rows will see in view.**

A view is a virtual table and is created based on the result from one or more real table.

Here is the syntax to create view on the table-

CREATE VIEW view\_name AS SELECT col1, col2 FROM tablename;

Then you can view the data of the view using the below command-

Select \* from view\_name;

If you later want to drop the view, just use the below statement-

Drop view view\_name;

Now as you know what view is and how to create a view, let’s jump to the second statement of the question.

As the view is created with the contents of the table itself and so updating table will update the view as well. So if you will add 100 new rows to the table then while checking the view, it will also have those newly added 100 rows.

**5. At first, we have created a view on top of the table (two rows empid, empname). Then again added a third row (i.e. address ). Whether can we see the newly added row in the view?**

This can also be derived from the above Capgemini Hadoop interview questions and answers. Again if the table is getting updated, the view will get updated.

Here you just need to ensure that while creating the view no such condition has been applied which will prevent the view from updating.

For example, if you have just used a couple of column names (col1 and col2) while creating the view then if you will add the third column in the table also, it won’t affect the view.

Hope that clarifies. Please comment for any further questions.

**6. In Hive is the delimiter is mandatory? What is the default delimiter?**

While you create a table in Hive, you specify the delimiter to let Hive know the format of the data you have in the input file.

CREATE TABLE emp(  
Id int, name string)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘\t’;

In the above Hive create table example, ‘\t’ is the delimiter.

The default record delimiter in Hive is − \n

And the field delimiters in Hive are − \001,\002,\003

ROW FORMAT should have delimiters used to terminate the fields and lines as shown above.

**Hey! I am first heading line feel free to change me**

**7. Is the mapper is mandatory or is the reducer is mandatory?**

Mapper class is mandatory in Hadoop. If you do not use the mapper class in the driver, then **IdentityMapper** will be used.

Reducer is not the mandatory one. Also, if you are dealing the import operation through Sqoop, the only mapper run and not the reducer.

**8. Without loading the data from HDFS is it possible to load the data to hive?**

Yes, you can use the files placed on the local system. Another option is to copy the other tables data t create a new table. But in that case also, indirectly you are using the HDFS data.

So, the best option to load data in Hive without copying in HDFS is to use local inpath command while loading data in Hive table as below.

Load data local inpath ‘localfilepath’ into table tbl\_name

**9. Without inserting or without loading the data to the hive is it possible to load the data to hive?**

Yes, we can do so by creating a table from anther existing table. Here is the syntax for that-

CREATE TABLE tbl1 row format delimited fields terminated by ‘\t’ AS select \* from tbl2

You can also create an external table in Hive using HBase as below-

CREATE EXTERNAL TABLE hbase\_hive\_names(fields\_names) STORED BY ‘org.apache.hadoop.hive.hbase.HBaseStorageHandler’ WITH SERDEPROPERTIES (“hbase.columns.mapping” = “:key,id:id,name:fn,name:ln,age:age”) TBLPROPERTIES(“hbase.table.name” = “hbase\_2\_hive\_names”);

**10. What is high definition cluster?**

High definition cluster is similar to High Availability cluster. You can check our details article [Design Hadoop cluster](http://www.hdfstutorial.com/blog/design-hadoop-cluster/) for the details.

High availability (HA) Hadoop cluster is not but the group of the system which acts like a single system and provides high uptime.

An HA cluster is usually used for load balancing, backup, failover and disaster recovery (DR) purpose.

**11. What is Flume? Do you have any knowledge on Flume?**

Explain a bit about Flume and why it is used.

Flume is a reliable distributed system to aggregate a large amount of real-time streaming data in HDFS. Flume is used in Hadoop ecosystem to get the data from the sources like Twitter, Facebook, and other sites to Hadoop systems.

The primary use of Flume is to gather the social media data to analyze further. We can call for sentiment data analysis; we take help of Flume to gather the data.

**12. Where will the data store in HDFS?**

The data will be stored in HDFS in the directory specified in dfs**.datanode.data.dir** and **/**dfs**/data** suffix that you see in the default value will not be appended.

You can find this information in hdfs**-site.xml** easily. But remember if you edit hdfs**-site.xml**, you need to restart the DataNodes services for the changes to take place.

You should also check [HDFS file processing](http://www.hdfstutorial.com/hdfs-file-processing/) for the detailed operation.

**13. In HDFS the data will store why again we need to load the data to HIVE?**

Actually, Hive does not store anything. It is the HDFS which store the data which we put into Hive table.

The main reason to load the data again in Hive table is for further computation. For example, you have partitioned table, and so the system will make directories based on the partition and will keep file related to that partition.

Now if you will manually place the file, that won’t be accessible for any operation. So to avoid such conditions, we load the data in Hive table even we have files in HDFS location.

**14. What is Hive Partition and have you used hive partition?**

Hive partition is an important concept in Hive and is one of the major factors for [Hive performance tuning](http://www.hdfstutorial.com/blog/hive-performance-tuning/).

Let’s consider a scenario where you have a table having the population of India. It will have billions of records…right?

Now if you have to find some records from the table how much time will it take? A lot more than the usual which is not a good practice.

To avoid this, Partition has been allowed on the Hive table.

**15. What is Hive Partition?**

It is a way to divide the Hive table into different related parts based on the partitioned columns such as city, date, country, etc.

It is very easy to query the partitioned table as the data is already structured. Here is the syntax to create partitioned table-

create table population\_partition(  
pID int,  
Name string,  
Address string  
)partitioned by (state string)  
row format delimited  
fields terminated by ‘\t’  
stored as textfile

Remember partitioned column should not be written while creating the columns.

Now if you want to load the data in the partitioned column, follow the below steps-

load data inpath ‘user/hive/population/state.csv’  
overwrite into table population\_partition partition (state)

There are two types of Hive partitions-

* Static Partition
* Dynamic Partition

Static partition is the default partition in Hive, but you need to set some parameters to enable the dynamic partition. Here is the set of properties which need to be configured to allow the dynamic partition in Hive.

set hive.exec.dynamic.partition=true;  
set hive.exec.dynamic.partition.mode=nonstrict;

**16. What is the purpose of using the Hadoop in your project?**

Here you can explain as per your project. A simple answer can be like below-

We are working for an ecommerce company, and we receive the log file based on the sessions from our client. Now explain how the log file comes, and then you parse the file.

Again use any tool like Hive/Pig to analyze those file and get the required data like the popular products which use clicked most.

Then say you are using some tools like Tableau to present the data in graphical form to share with the client.

Remember this is just a typical case, and you may explain as per your project. Here are few Hadoop use cases which you can consider while explaining- [Big Data ECommerce case studies](http://www.hdfstutorial.com/blog/big-data-ecommerce-case-studies/), [Hadoop use cases in Education](http://www.hdfstutorial.com/blog/hadoop-use-cases-in-education/), [Big Data use cases in banking and financial services](http://www.hdfstutorial.com/blog/big-data-use-cases-in-banking-and-financial-services/).

**17. Where is the production environment for your project?**

Again the answer will depend on your project. You can say the location of your production server and then share the configurations etc.

**18. What is the default path where the hive data will store?**

The default path where hive stores data in HDFS is **/usr/hive/warehouse**.

But you can change/configure the default location as per your need from the Hive configuration file **hive-site.xml**.

Inside this directory, you will find the directories depending on the hive tables you have created.

You will find a section in **hive-site.xml** like below-

**<property>**  
**<name>hive.metastore.warehouse.dir</name>**  
**<value>Your\_Path\_HERE</value>**  
**<description>location of default database for the warehouse</description>**  
**</property>**

Now put the path you want in place of “your\_path\_HERE”. One thing you should note here is, the path you are providing should have Read/Write/Execute permissions.

For that you can execute the below-

sudo chown -R user  
sudo chmod -R 777

And then you should restart the Hadoop services using stop and start commands-

stop-all.sh  
start-all.sh

**19. Explain syntax to create hive table and load data into the hive. Also, share its mandatory items?**

This is relatively a simple question, and you can refer below syntax.  
**Hive Create table syntax (internal table)**

Create table tblname (  
Id int, name string..)  
Row format delimited fields terminated by ‘\t’

**Hive load data syntax (from local system)**

Load data local inpath ‘local\_pth’ into table tblname;

**Hive load data syntax (from HDFS location)**

Load data inpath ‘local\_pth’ into table tblname;

**Hive Create table syntax (internal table)**

Create external table Etblname (  
Id int, name string..)  
Row format delimited fields terminated by ‘\t’  
Location ‘hdfs\_external\_location’

**What are the different tables in Hive?**

There are mainly two types of Hive tables-

• External table  
• Internal table

Below are the syntaxes to create both the table.

**Hive Create table syntax (internal table)**

Create table tblname (  
Id int, name string..)  
Row format delimited fields terminated by ‘\t’

**Hive Create table syntax (internal table)**

Create external table Etblname (  
Id int, name string..)  
Row format delimited fields terminated by ‘\t’  
Location ‘hdfs\_external\_location’

**Difference between Hive External table and Hive Internal table**

For the difference between Hive external and internal table, please check Q9 of [Hadoop scenario based questions](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/).

There is one more table that we use in the hive, and that is Skewed Table.

**Hive Skewed Table**

A skewed table is a special type of table where the values that appear very often (heavy skew) are split out into separate files and rest of the values go to some other file.

**Create Hive Skewed Table Syntax**

create table T (c1 string, c2 string) skewed by (c1) on (‘x1’)

The skewed table is again one of the important Hive query optimization techniques.

**20. Explain Hive Internal table vs. Hive External table. If we have two tables (i.e. internal table & external table) and we drop both the tables. Then if we run (select \* from external/internal table ) query on both external and internal table. Whether can we get the output. Then what will be the output.**

Please refer the above question for the difference between an external table and internal table.

Now as we know if we drop the internal table, both metadata and data will be lost. While if we will delete the external table, only metadata will be lost and not the actual data as it will be in some other directory.

Now if you will drop both the table and then do a query, it will show you the error message- “table not found”.

**21. A Table has two rows (emp\_name and emp\_salary). How to get the Max salary without using the Max keyword or without using UDFs?**

There are many methods to do so like below-

**Using MIN keyword:**

select MIN(-1 \* col)\*-1 as col from tableName;

**Using self-join**

select A.col, B.col  
from tableName as A, tableName as B  
where A.col < B.col

**Using Limit**

SELECT column  
FROM YOUR\_TABLE  
ORDER BY column DESC  
LIMIT 1

**22. How to create a Hive table in AVRO format?**

Here is the syntax to create Hive table in Avro format-

CREATE EXTERNAL TABLE avro\_hive\_table  
ROW FORMAT SERDE ‘org.apache.hadoop.hive.serde2.avro.AvroSerDe’TBLPROPERTIES  
(‘avro.schema.url’=’hdfs://localdomain/user/avro/schemas/activity.avsc’)  
STORED AS INPUTFORMAT ‘org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat’  
OUTPUTFORMAT ‘org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat’  
LOCATION ‘/user/avro/applog\_avro’;

**Contents**[[hide](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/)]

* [1 1. What are the differences between -copyFromLocal and -put command](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#1_What_are_the_differences_between_-copyFromLocal_and_-put_command)
* [2 2. What are the differences between -copyToLocal and -put command](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#2_What_are_the_differences_between_-copyToLocal_and_-put_command)
* [3 3. What is the default block size in Hadoop and can it be increased?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#3_What_is_the_default_block_size_in_Hadoop_and_can_it_be_increased)
* [4 4. How to import RDBMS table in Hadoop using Sqoop when the table doesn’t have a primary key column?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#4_How_to_import_RDBMS_table_in_Hadoop_using_Sqoop_when_the_table_doesnt_have_a_primary_key_column)
* [5 5. What is CBO in Hive?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#5_What_is_CBO_in_Hive)
* [6 6. Can we use LIKE operator in Hive?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#6_Can_we_use_LIKE_operator_in_Hive)
* [7 7. Can you use IN/EXIST operator in Hive?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#7_Can_you_use_INEXIST_operator_in_Hive)
* [8 8. What are the differences between INNER JOIN and LEFT SEMI JOIN?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#8_What_are_the_differences_between_INNER_JOIN_and_LEFT_SEMI_JOIN)
* [9 9. What are the differences between External and Internal Tables in Hive](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#9_What_are_the_differences_between_External_and_Internal_Tables_in_Hive)
* [10 10. When to use external and internal tables in Hive?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#10_When_to_use_external_and_internal_tables_in_Hive)
* [11 11. We have a Hive partitioned table with partition column as country. We have 10 partition and data for now is jut for one country, If we will copy the data manually for other 9 partitions, whether those will be reflected if we will run a command.](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#11_We_have_a_Hive_partitioned_table_with_partition_column_as_country_We_have_10_partition_and_data_for_now_is_jut_for_one_country_If_we_will_copy_the_data_manually_for_other_9_partitions_whether_those_will_be_reflected_if_we_will_run_a_command)
* [12 12. Where the Mapper’s Intermediate data will be stored?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#12_Where_the_Mappers_Intermediate_data_will_be_stored)
* [13 13. What is Partition and Combiner in MapReduce?](http://www.hdfstutorial.com/blog/hadoop-scenario-based-interview-questions/#13_What_is_Partition_and_Combiner_in_MapReduce)

**1. What are the differences between -copyFromLocal and -put command**

**Ans:** Basically, both put and copyFromLocal fulfill similar purposes, but there are some differences. First, see what both the command does-

**-put:** it can copy the file from source to destination

**– copyFromLocal:** It copies the file from local file system to Hadoop system

As you saw, put can do what copyFromLocal is doing but the reverse is not true. So the main difference between -copyFromLocal and -put commands is, in -copyFromLocal, the source has to be the local file system which is not mandatory for –put command.

Uses of these commands-

hadoop fs -copyFromLocal <localsrc> URI

hadoop fs -put <localsrc> … <destination>

**2. What are the differences between -copyToLocal and -put command**

**Ans:** The answer will be similar to what I explained in the above question. The only difference is, there it was –copyFromLocal and here it is –copyToLocal.

So in –copyToLocal command, the destination has to be the local file system.

**3. What is the default block size in Hadoop and can it be increased?**

Ans: The default block size in Hadoop 1 is 64 MB while in Hadoop 2, it is 128MB.

It can be increased as per your requirements. You can check [Hadoop Terminology](http://hdfstutorial.com/important-hadoop-terminology/) for more details.

In fact changing the block size is very easy and you can do it by setting fs.local.block.size in the configuration file easily. Use the below command to change the default block size in Hadoop.

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

Just put the size you want of a block in KB in place of “sizeinKB” variable.

**4. How to import RDBMS table in Hadoop using Sqoop when the table doesn’t have a primary key column?**

Ans: Usually, we import an RDBMS table in Hadoop using [Sqoop Import](http://hdfstutorial.com/sqoop-import-function/) when it has a primary key column. If it doesn’t have the primary key column, it will give you the below error-

ERROR tool.ImportTool: Error during import: No primary key could be found for table <table\_name>. Please specify one with –split-by or perform a sequential import with ‘-m 1’

Here is the solution of what to do when you don’t have a primary key column in RDBMS, and you want to import using Sqoop.

If your table doesn’t have the primary key column, you need to specify *-m 1* option for importing the data, or you have to provide ***–split-by*** argument with some column name.

Here are the scripts which you can use to import an RDBMS table in Hadoop using Sqoop when you don’t have a primary key column.

sqoop import \  
–connect jdbc:mysql://localhost/dbname \  
–username root \  
–password root \  
–table user \  
–target-dir /user/root/user\_data \  
–columns “first\_name, last\_name, created\_date”  
-m 1

or

sqoop import \  
–connect jdbc:mysql://localhost/ dbname\  
–username root \  
–password root \  
–table user \  
–target-dir /user/root/user\_data \  
–columns “first\_name, last\_name, created\_date”  
–split-by created\_date

**5. What is CBO in Hive?**

Ans: CBO is cost-based optimization and applies to any database or any tool where optimization can be used.

So it is similar to what you call Hive Query optimization. Here are the few parameters, you need to take care while dealing with CBO in Hive.

* Parse and validate query
* Generate possible execution plans
* For each logically equivalent plan, assign a cost

You can also check [Hortonworks technical sheet](http://hortonworks.com/blog/hive-0-14-cost-based-optimizer-cbo-technical-overview/) on this for more details.

**6. Can we use LIKE operator in Hive?**

Yes, Hive supports LIKE operator, but it doesn’t support multi-value LIKE queries like below-

SELECT \* FROM user\_table WHERE first\_name LIKE ANY ( ‘root~%’ , ‘user~%’ );

So you can easily use LIKE operator in Hive as and when you require. Also, when you have to use a multi-like operator, break it so that it can work in Hive.  
E.g.:

WHERE table2.product LIKE concat(‘%’, table1.brand, ‘%’)

**7. Can you use IN/EXIST operator in Hive?**

No, Hive doesn’t support IN or EXIST operators. Instead, you can use left semi join here. Left Semi Join performs the same operation IN do in SQL.

So if you have the below query in SQL-

SELECT a.key, a.value  
FROM a  
WHERE a.key in  
(SELECT b.key  
FROM B);

Then the suitable query for the same in Hive can be-

SELECT a.key, a.val  
FROM a LEFT SEMI JOIN b on (a.key = b.key)

Both will fulfill the same purpose.

**8. What are the differences between INNER JOIN and LEFT SEMI JOIN?**

Ans: Left semi-join in Hive is used instead of IN operator (as IN is not supported in Hive). Now coming to the differences, inner join returns the common data from both the table depending on condition applied while left semi joins only returns the records from the left-hand table.

**9. What are the differences between External and Internal Tables in Hive**

Ans: As we know there are a couple of kinds of tables in Hive- Internal and External (Managed) table. In the internal table (default), data will be stored at the default Hive location while in the external table; you can specify the location.

The major difference between the internal and external tables are-

| **External Table** | **Internal Table** |
| --- | --- |
| External table stores files on the HDFS | Stored in a directory based on settings in hive.metastore.warehouse.dir, by default internal tables are stored in the following directory “/user/hive/warehouse” you can change it by updating the location in the config file. |
| If you delete an external table the file still remains on the HDFS server. As an example if you create an external table called “table\_test” in HIVE using HIVE-QL and link the table to file “file”, then deleting “table\_test” from HIVE will not delete “file” from HDFS. | Deleting the table deletes the metadata & data from master-node and HDFS respectively. |
| External table files are accessible to anyone who has access to HDFS file structure and therefore security needs to be managed at the HDFS file/folder level. | Deleting the table deletes the metadata & data from master-node and HDFS respectively. • Internal table file security is controlled solely via HIVE. Security needs to be managed within HIVE, probably at the schema level (depends on organisation to organisation). |
| Meta data is maintained on master node and deleting an external table from HIVE, only deletes the metadata not the data/file. | It is the default table in Hive. |

Showing 1 to 4 of 4 entries

Hive may have internal or external tables this is a choice that affects how data is loaded, controlled, and managed.

**10. When to use external and internal tables in Hive?**

**Use EXTERNAL tables when:**

* The data is also used outside of Hive. For example, the data files are read and processed by an existing program that doesn’t lock the files.
* Data needs to remain in the underlying location even after a DROP TABLE. This can apply if you are pointing multiple schemas (tables or views) at a single data set or if you are iterating through various possible schemas.
* Hive should not own data and control settings, dirs, etc., you may have another program or process that will do those things.
* You are not creating a table based on existing table (AS SELECT).

**Use INTERNAL tables when:**

* The data is temporary
* You want Hive to completely manage the lifecycle of the table and data

**11. We have a Hive partitioned table with partition column as country. We have 10 partition and data for now is jut for one country, If we will copy the data manually for other 9 partitions, whether those will be reflected if we will run a command.**

Ans: This is really a good question. As the data has been kept manually in all the other file directory and so directly it won’t be available.

Data will be available directly for all partition when you will put it through command and not manually.

**12. Where the Mapper’s Intermediate data will be stored?**

**Ans:**The mapper output (which is intermediate data) is stored on the Local file system (not in HDFS) of each mapper nodes. This is a temporary directory location which can be setup in the configuration file by the Hadoop administrator. The intermediate data is cleaned up after the Hadoop Job completes.

## Top Answers to Hive Questions

**1. Compare Pig and Hive**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Pig** | **Hive** |
| Architecture | Procedural data flow language | SQL type declarative language |
| Application | Programming purposes | Report creation |
| Operational field | Client side | Server side |
| Support for avro files | Yes | No |

**2. What is the definition of Hive? What is the present version of Hive and explain about ACID transactions in Hive?**

Hive is an open source data warehouse system. We can use Hive for analyzing and querying in large data sets of Hadoop files. It’s similar to SQL. The present version of hive is 0.13.1. Hive supports ACID transactions: The full form of ACID is Atomicity, Consistency, Isolation, and Durability. ACID transactions are provided at the row levels, there are Insert, Delete, and Update options so that Hive supports ACID transaction.

* Insert
* Delete
* Update

Get a better understanding of Hive by going through this [Hive Tutorial](https://intellipaat.com/tutorial/hadoop-tutorial/apache-hive/) now.

**3. Explain what is a Hive variable? What do we use it for?**

Hive variable is basically created in the Hive environment that is referenced by Hive scripting languages. It provides to pass some values to the hive queries when the query starts executing. It uses the source command.

**4. What kind of data warehouse application is suitable for Hive? What are the types of tables in Hive?**

Hive is not considered as a full database. The design rules and regulations of Hadoop and HDFS put restrictions on what Hive can do.Hive is most suitable for data warehouse applications.  
Where :

* Analyzing the relatively static data.
* Less Responsive time.
* No rapid changes in data.Hive doesn’t provide fundamental features required for OLTP, Online Transaction Processing.Hive is suitable for data warehouse applications in large data sets.**Two types of tables in Hive**

1. Managed table.
2. External table.

**5. Can We Change settings within Hive Session? If Yes, How?**

Yes we can change the settings within Hive session, using the SET command. It helps to change Hive job settings for an exact query.  
Example: The following commands shows buckets are occupied according to the table definition.

hive> SET hive.enforce.bucketing=true;

We can see the current value of any property by using SET with the property name. SET will list all the properties with their values set by Hive.

hive> SET hive.enforce.bucketing;

hive.enforce.bucketing=true

And this list will not include defaults of Hadoop. So we should use the below like

SET -v

It will list all the properties including the Hadoop defaults in the system.

**Interested in learning Hive? Well, we have the comprehensive**[*Hive Training Course*](https://intellipaat.com/big-data-hadoop-training/#curriculum)**to give you a head start in your career.**

**6. Is it possible to add 100 nodes when we have 100 nodes already in Hive? How.**

**Yes, we can add the nodes by following the below steps.**

1. Take a new system create a new username and password.
2. Install the SSH and with master node setup ssh connections.
3. Add ssh public\_rsa id key to the authorized keys file.
4. Add the new data node host name, IP address and other details in /etc/hosts slaves file  
   192.168.1.102 slave3.in slave3.
5. Start the Data Node on New Node.
6. Login to the new node like suhadoop or ssh -X hadoop@192.168.1.103.
7. Start HDFS of a newly added slave node by using the following command  
   ./bin/hadoop-daemon.sh start data node.
8. Check the output of jps command on a new node

**7. Explain the concatenation function in Hive with an example ?**

Concatenate function will join the input strings.We can specify the  
‘N’ number of strings separated by a comma.  
Example:

CONCAT ('Intellipaat','-','is','-','a','-','eLearning',’-’,’provider’);

Output:

Intellipaat-is-a-eLearning-provider

So, every time we set the limits of the strings by ‘-‘. If it is common for every strings, then Hive provides another command

CONCAT\_WS. In this case,we have to specify the set limits of operator first.

CONCAT\_WS ('-',’Intellipaat’,’is’,’a’,’eLearning’,‘provider’);

Output: Intellipaat-is-a-eLearning-provider.

Learn more about the Apache Hive [features, architecture, and HiveQL](https://en.wikipedia.org/wiki/Apache_Hive), in this article now.

[**Wish to Learn Hive? Click Here**](https://intellipaat.com/big-data-hadoop-training/#course-content)

**8. Trim and Reverse function in Hive with examples?**

Trim function will delete the spaces associated with a string.  
Example:

TRIM(‘ INTELLIPAAT ‘);

Output:

INTELLIPAAT

To remove the Leading space

LTRIM(‘ INTELLIPAAT’);

To remove the trailing space

RTRIM(‘INTELLIPAAT ‘);

In Reverse function, characters are reversed in the string.

Example:

REVERSE(‘INTELLIPAAT’);

Output:

TAAPILLETNI

**9. How to change the column data type in Hive? Explain RLIKEin Hive.**

We can change the column data type by using ALTER and CHANGE.  
The syntax is :

ALTER TABLE table\_name CHANGE column\_namecolumn\_namenew\_datatype;

Example: If we want to change the data type of the salary column from integer to bigint in the employee table.  
ALTER TABLE employee CHANGE salary salary BIGINT;RLIKE: Its full form is Right-Like and it is a special function in the Hive. It helps to examine the two substrings. i.e, if the substring of A matches with B then it evaluates to true.  
Example:

‘Intellipaat’ RLIKE ‘tell’ ◊ True

‘Intellipaat’ RLIKE ‘^I.\*’ ◊ True (this is a regular expression)

[Learn what is Hadoop Hive](https://intellipaat.com/blog/what-is-hadoop-hive/) in this detailed blog post now.

**Download Hive Interview questions asked by top MNCs in 2017 ?**

Top of Form



Bottom of Form

**10. What are the components used in Hive query processor?**

The components of a Hive query processor include

* Logical Plan of Generation.
* Physical Plan of Generation.
* Execution Engine.
* Operators.
* UDF’s and UDAF’s.
* Optimizer.
* Parser.
* Semantic Analyzer.
* Type Checking

**11. What is Buckets in Hive?**

The data which is currently present are partitioned and divided further into Buckets  
The division is performed based on Hash of particular columns that is selected in the table.

**12. Explain process to access sub directories recursively in Hive queries?**

By using below commands we can access sub directories recursively in Hive

hive> Set mapred.input.dir.recursive=true;

hive> Set hive.mapred.supports.subdirectories=true;

Hive tables can be pointed to the higher level directory and this is suitable for the directory structure which is like /data/country/state/city/

**13. What are the components used in Hive query processor?**

**The components of a Hive query processor include**

* Logical Plan of Generation
* Physical Plan of Generation
* Execution Engine
* Operators
* UDF’s and UDAF’s
* Optimizer
* Parser
* Semantic Analyzer
* Type Checking

**14. How to skip header rows from a table in Hive?**

Header records in log files  
System=….  
Version=…  
Sub-version=….  
In the above three lines of headers that we do not want to include in our Hive query. To skip header lines from our tables in the Hive,set a table property that will allow us to skip the header lines.

CREATE EXTERNAL TABLE employee (

name STRING,

job STRING,

dob STRING,

id INT,

salary INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘ ‘ STORED AS TEXTFILE

LOCATION ‘/user/data’

TBLPROPERTIES("skip.header.line.count"="2”);

**15. The maximum size of string data type supported by hive? Mention the Hive support binary formats?**

The maximum size of string data type supported by hive is 2 GB.  
Hive supports the text file format by default and it supports the binary format Sequence files, ORC files, Avro Data files, Parquet files.  
Sequence files: Splittable, compressible and row oriented are the general binary format.  
ORC files: Full form of ORC is optimized row columnar format files. It is a Record columnar file and column oriented storage file. It divides the table in row split. In each split stores that value of the first row in the first column and followed sub subsequently.  
AVRO data files: It is same as a sequence file splittable, compressible and row oriented, but except the support of schema evolution and multilingual binding support.

**Learn more about Hive in this**[*Training Course*](https://intellipaat.com/big-data-hadoop-training/#sample-video)**to get ahead in your career!**

**16. What is the precedence order of HIVE configuration?**

We are using a precedence hierarchy for setting the properties

1. SET Command in HIVE
2. The command line –hiveconf option
3. Hive-site.XML
4. Hive-default.xml
5. Hadoop-site.xml
6. Hadoop-default.xml

**17. If you run a select \* query in Hive, Why does it not run MapReduce?**

It’s an optimization technique. Hive, fetch, task conversion property can minimize the latency of map-reduce overhead. When queried like SELECT, FILTER LIMIT queries, this property skips map reduce and using FETCH task. As a result, Hive can execute queries without running MapReduce task.

By default its value is minimal. Which optimize- SELECT \*, FILTER on partition columns, LIMIT queries only, Where another value is more which optimize- SELECT, FILTER, LIMIT.

**18. How Hive can improve performance with ORC format tables?**

We can store the hive data in highly efficient manner in the Optimized Row Columnar file format. It can simplify many Hive file format limitations. We can improve the performance by using ORC files while reading, writing and processing the data.

Set hive.compute.query.using.stats-true;

Set hive.stats.dbclass-fs;

CREATE TABLE orc\_table (

idint,

name string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\:’

LINES TERMINATED BY ‘\n’

STORES AS ORC;

Need a reason to learn Apache Hadoop and Hive? Well, go through this blog post to find out why [Hadoop is the new black](https://intellipaat.com/blog/hadoop-is-the-new-black/).

**19. Explain the functionality of Object-Inspector?**

It helps to analyze the internal structure of row object and individual structure of columns in HIVE. It also provides a uniform way to access complex objects that can be stored in multiple formats in the memory.  
Instance of Java class  
A standard Java object  
A lazily initialized object  
The Object-Inspector tells structure of the object and also ways to access the internal fields inside the object.

**20. Explain the reason for whenever we run hive query, new metastore\_db is created. Why?**

Local metastore is created when we run Hive in embedded mode. And before creating it checks whether the metastore exists or not and this metastore property is defined in the configuration file hive-site.xml. Property is“javax.jdo.option.ConnectionURL” with default value “jdbc:derby:;databaseName=metastore\_db;create=true”.So to change the behavior of the location to an absolute path, so that from that location meta-store will be used.

**Give your career a big boost by going through our**[*Hive Training Course*](https://intellipaat.com/big-data-hadoop-training/#certification)**now!**

**21. Differentiate between Hive and Hbase**

|  |  |
| --- | --- |
| **Hive** | **HBase** |
| Enables most of the SQL queries | This doesn’t allow SQL queries |
| Doesn’t support record level insert, update, and delete operations on table | It supports |
| It is a data warehouse framework | It is NoSQL database |
| Hive run on the top of MapReduce | HBase runs on the top of HDFS |

**22. Differentiate between Hive and Hbase**

By using below commands we can access sub directories recursively in Hive

hive> Set mapred.input.dir.recursive=true;

hive> Set hive.mapred.supports.subdirectories=true;

Hive tables can be pointed to the higher level directory and this is suitable for the directory structure which is like /data/country/state/city/

**23. What are the uses of explode Hive?**

Hadoop developers consider the array as their inputs and convert them into a separate table row. To convert complicate data types into desired table formats Hive is essentially using explode.

**24. What is available mechanism for connecting from applications, when we run hive as a server?**

1. Thrift Client: Using thrift you can call hive commands from various programming languages. Example: C++, PHP,Java, Python and Ruby.
2. JDBC Driver: JDBC Driver supports the Type 4 (pure Java) JDBC Driver
3. ODBC Driver: ODBC Driver supports the ODBC protocol.

**25. How do we write our own custom SerDe?**

End users want to read their own data format instead of writing, so the user wants to write a Deserializer than SerDe.  
Example: The RegexDeserializer will deserialize the data using the configuration parameter ‘regex’, and a list of column names.  
If our SerDe supports DDL, we probably want to implement a protocol based on DynamicSerDe. It’s non-trivial to write a “thrift DDL” parser.

**26. Mention the date data type in Hive? Name the Hive data type collection.**

The TIMESTAMP data type stores date in java.sql.timestamp format.

**Three collection data types in Hive**

1. ARRAY
2. MAP
3. STRUCTGo through this IBM article and learn how [Hive works perfectly with Big SQL](https://www-01.ibm.com/software/data/infosphere/hadoop/hive/)

**27. Can we run UNIX shell commands from Hive? Can Hive queries be executed from script files? How? Give an example.**

Yes, we can run UNIX shell commands from Hive using the! Mark before the command .For example: !pwd at hive prompt will list the current directory.  
We can execute Hive queries from the script files by using the source command.  
Example −

Hive> source /path/to/file/file\_with\_query.hql

**1. Compare Hadoop & Spark**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Hadoop** | **Spark** |
| Dedicated storage | HDFS | None |
| Speed of processing | average | excellent |
| Libraries | Separate tools available | Spark Core, SQL, Streaming, MLlib, GraphX |

**2. What are real-time industry applications of Hadoop?**

Hadoop, well known as Apache Hadoop, is an open-source software platform for scalable and distributed computing of large volumes of data. It provides rapid, high performance and cost-effective analysis of structured and unstructured data generated on digital platforms and within the enterprise. It is used in almost all departments and sectors today.Some of the instances where Hadoop is used:

* Managing traffic on streets.
* Streaming processing.
* Content Management and Archiving Emails.
* Processing Rat Brain Neuronal Signals using a Hadoop Computing Cluster.
* Fraud detection and Prevention.
* Advertisements Targeting Platforms are using Hadoop to capture and analyze click stream, transaction, video and social media data.
* Managing content, posts, images and videos on social media platforms.
* Analyzing customer data in real-time for improving business performance.
* Public sector fields such as intelligence, defense, cyber security and scientific research.
* Financial agencies are using Big Data Hadoop to reduce risk, analyze fraud patterns, identify rogue traders, more precisely target their marketing campaigns based on customer segmentation, and improve customer satisfaction.
* Getting access to unstructured data like output from medical devices, doctor’s notes, lab results, imaging reports, medical correspondence, clinical data, and financial data.

**3. How is Hadoop different from other parallel computing systems?**

Hadoop is a distributed file system, which lets you store and handle massive amount of data on a cloud of machines, handling data redundancy. Go through this [HDFS content to know how the distributed file system works](https://intellipaat.com/tutorial/hadoop-tutorial/hdfs-overview/). The primary benefit is that since data is stored in several nodes, it is better to process it in distributed manner. Each node can process the data stored on it instead of spending time in moving it over the network.

On the contrary, in Relational database computing system, you can query data in real-time, but it is not efficient to store data in tables, records and columns when the data is huge.

[Learn about Oracle DBA](https://intellipaat.com/tutorial/oracle-dba-tutorial/) now.

Hadoop also provides a scheme to build a Column Database with Hadoop HBase, for runtime queries on rows.

Learn more in this [HBase Tutorial](https://intellipaat.com/tutorial/hbase-tutorial/)**.**

**4. What all modes Hadoop can be run in?**

Hadoop can run in three modes:

* **Standalone Mode:** Default mode of Hadoop, it uses local file stystem for input and output operations. This mode is mainly used for debugging purpose, and it does not support the use of HDFS. Further, in this mode, there is no custom configuration required for mapred-site.xml, core-site.xml, hdfs-site.xml files. Much faster when compared to other modes.
* **Pseudo-Distributed Mode (Single Node Cluster):** In this case, you need configuration for all the three files mentioned above. In this case, all daemons are running on one node and thus, both Master and Slave node are the same.
* **Fully Distributed Mode (Multiple Cluster Node):** This is the production phase of Hadoop (what Hadoop is known for) where data is used and distributed across several nodes on a Hadoop cluster. Separate nodes are allotted as Master and Slave.

**5. Explain the major difference between HDFS block and InputSplit.**

In simple terms, block is the physical representation of data while split is the logical representation of data present in the block. Split acts a s an intermediary between block and mapper.  
Suppose we have two blocks:  
**Block 1: ii nntteell  
Block 2: Ii ppaatt**  
Now, considering the map, it will read first block from ii till ll, but does not know how to process the second block at the same time. Here comes Split into play, which will form a logical group of Block1 and Block 2 as a single block.

It then forms key-value pair using inputformat and records reader and sends map for further processing With inputsplit, if you have limited resources, you can increase the split size to limit the number of maps. For instance, if there are 10 blocks of 640MB (64MB each) and there are limited resources, you can assign ‘split size’ as 128MB. This will form a logical group of 128MB, with only 5 maps executing at a time.

However, if the ‘split size’ property is set to false, whole file will form one inputsplit and is processed by single map, consuming more time when the file is bigger.

**6. What is distributed cache and what are its benefits?**

Distributed Cache, in Hadoop, is a service by MapReduce framework to cache files when needed. Learn more in this [MapReduce Tutorial](https://intellipaat.com/tutorial/mapreduce-tutorial/) now. Once a file is cached for a specific job, hadoop will make it available on each data node both in system and in memory, where map and reduce tasks are executing.Later, you can easily access and read the cache file and populate any collection (like array, hashmap) in your code.

**Benefits of using distributed cache are:**

* It distributes simple, read only text/data files and/or complex types like jars, archives and others. These archives are then un-archived at the slave node.
* Distributed cache tracks the modification timestamps of cache files, which notifies that the files should not be modified until a job is executing currently.

**7. Explain the difference between NameNode, Checkpoint NameNode and BackupNode.**

* **NameNode** is the core of HDFS that manages the metadata – the information of what file maps to what block locations and what blocks are stored on what datanode. In simple terms, it’s the data about the data being stored. NameNode supports a directory tree-like structure consisting of all the files present in HDFS on a Hadoop cluster. It uses following files for namespace:  
  fsimage file- It keeps track of the latest checkpoint of the namespace.  
  edits file-It is a log of changes that have been made to the namespace since checkpoint.
* **Checkpoint NameNode** has the same directory structure as NameNode, and creates checkpoints for namespace at regular intervals by downloading the fsimage and edits file and margining them within the local directory. The new image after merging is then uploaded to NameNode.  
  There is a similar node like Checkpoint, commonly known as Secondary Node, but it does not support the ‘upload to NameNode’ functionality.
* **Backup Node** provides similar functionality as Checkpoint, enforcing synchronization with NameNode. It maintains an up-to-date in-memory copy of file system namespace and doesn’t require getting hold of changes after regular intervals. The backup node needs to save the current state in-memory to an image file to create a new checkpoint.

***Learn about the various Hadoop components in this***[*Big Data Hadoop Video Tutorial*](https://intellipaat.com/big-data-hadoop-video-tutorial/)***.***

**8. What are the most common Input Formats in Hadoop?**

There are three most common input formats in Hadoop:

* **Text Input Format:** Default input format in Hadoop.
* **Key Value Input Format:** used for plain text files where the files are broken into lines
* **Sequence File Input Format:** used for reading files in sequence

**9. Define DataNode and how does NameNode tackle DataNode failures?**

DataNode stores data in HDFS; it is a node where actual data resides in the file system. Each datanode sends a heartbeat message to notify that it is alive. If the namenode does noit receive a message from datanode for 10 minutes, it considers it to be dead or out of place, and starts replication of blocks that were hosted on that data node such that they are hosted on some other data node.A BlockReport contains list of all blocks on a DataNode. Now, the system starts to replicate what were stored in dead DataNode.

The NameNode manages the replication of data blocksfrom one DataNode to other. In this process, the replication data transfers directly between DataNode such that the data never passes the NameNode.

**Download Hadoop Interview questions asked by top MNCs in 2017 ?**

Top of Form



Bottom of Form

**10. What are the core methods of a Reducer?**

The three core methods of a Reducer are:

1. **setup():** this method is used for configuring various parameters like input data size, distributed cache.  
   public void setup (context)
2. **reduce():** heart of the reducer always called once per key with the associated reduced task  
   public void reduce(Key, Value, context)
3. **cleanup():** this method is called to clean temporary files, only once at the end of the task  
   public void cleanup (context)

**11. What is SequenceFile in Hadoop?**

Extensively used in MapReduce I/O formats, SequenceFile is a flat file containing binary key/value pairs. The map outputs are stored as SequenceFile internally. It provides Reader, Writer and Sorter classes. The three SequenceFile formats are:

1. Uncompressed key/value records.
2. Record compressed key/value records – only ‘values’ are compressed here.
3. Block compressed key/value records – both keys and values are collected in ‘blocks’ separately and compressed. The size of the ‘block’ is configurable.

**12. What is Job Tracker role in Hadoop?**

Job Tracker’s primary function is resource management (managing the task trackers), tracking resource availability and task life cycle management (tracking the taks progress and fault tolerance).

* It is a process that runs on a separate node, not on a DataNode often.
* Job Tracker communicates with the NameNode to identify data location.
* Finds the best Task Tracker Nodes to execute tasks on given nodes.
* Monitors individual Task Trackers and submits the overall job back to the client.
* It tracks the execution of MapReduce workloads local to the slave node.

**13. What is the use of RecordReader in Hadoop?**

Since Hadoop splits data into various blocks, RecordReader is used to read the slit data into single record. For instance, if our input data is split like:  
**Row1: Welcome to**

**Row2: Intellipaat**  
It will be read as “Welcome to Intellipaat” using RecordReader.

**14. What is Speculative Execution in Hadoop?**

One limitation of Hadoop is that by distributing the tasks on several nodes, there are chances that few slow nodes limit the rest of the program. Tehre are various reasons for the tasks to be slow, which are sometimes not easy to detect. Instead of identifying and fixing the slow-running tasks, Hadoop tries to detect when the task runs slower than expected and then launches other equivalent task as backup. This backup mechanism in Hadoop is Speculative Execution.

It creates a duplicate task on another disk. The same input can be processed multiple times in parallel. When most tasks in a job comes to completion, the speculative execution mechanism schedules duplicate copies of remaining tasks (which are slower) across the nodes that are free currently. When these tasks finish, it is intimated to the JobTracker. If other copies are executing speculatively, Hadoop notifies the TaskTrackers to quit those tasks and reject their output.

Speculative execution is by default true in Hadoop. To disable, set mapred.map.tasks.speculative.execution and mapred.reduce.tasks.speculative.execution  
JobConf options to false.

**15. What happens if you try to run a Hadoop job with an output directory that is already present?**

It will throw an exception saying that the output file directory already exists.

**To run** the MapReduce job, you need to ensure that the output directory does not exist before in the HDFS.

**To delete** the directory before running the job, you can use shell:Hadoop fs –rmr /path/to/your/output/Or via the Java API: FileSystem.getlocal(conf).delete(outputDir, true);

Prepare yourself for the [MapReduce Interview questions and answers](https://intellipaat.com/interview-question/map-reduce-interview-questions/)Now

**16. 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.

**17. How to configure Replication Factor in HDFS?**

**hdfs-site.xml** is used to configure HDFS. Changing the dfs.replication property in hdfs-site.xml will change the default replication for all files placed in HDFS.  
You can also modify the replication factor on a per-file basis using the

Hadoop FS Shell:[training@localhost ~]$ hadoopfs –setrep –w 3 /my/fileConversely,

you can also change the replication factor of all the files under a directory.

[training@localhost ~]$ hadoopfs –setrep –w 3 -R /my/dir

Go through [Hadoop Administration Training](https://intellipaat.com/hadoop-administration-training/) to learn about Replication Factor In HDFS now!

**18. How to compress mapper output but not the reducer output?**

To achieve this compression, you should set:

conf.set("mapreduce.map.output.compress", true)

conf.set("mapreduce.output.fileoutputformat.compress", false)

**19. What is the difference between Map Side join and Reduce Side Join?**

Map side Join at map side is performed data reaches the map. You need a strict structure for defining map side join. On the other hand, Reduce side Join (Repartitioned Join) is simpler than map side join since the input datasets need not be structured. However, it is less efficient as it will have to go through sort and shuffle phases, coming with network overheads.

**20. How can you transfer data from Hive to HDFS?**

By writing the query:

**hive> insert overwrite directory '/' select \* from emp;**

You can write your query for the data you want to import from Hive to HDFS. The output you receive will be stored in part files in the specified HDFS path.

**21. What companies use Hadoop, any idea?**

[Learn how Big Data and Hadoop have changed the rules of the game](https://intellipaat.com/blog/big-data-changing-disruptive-innovation/) in this blog post. Yahoo! (the biggest contributor to the creation of Hadoop) – Yahoo search engine uses Hadoop, Facebook – Developed Hive for analysis,Amazon,Netflix,Adobe,eBay,Spotify,Twitter,Adobe.

Real Time Hadoop Interview Questions from Various interviews

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 ꮕ�at ꒋ�les 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 ꒋ�les in Hive

2. Sqoop – Incremental vs hast modiꒋ�ed 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 ꒋ�les)

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 overꮕ�ow)

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 ꒋ�le size you’ve used?

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 룊�lter out?

3. Join – What did you join?

4. What is your cluster size?

5. What is the 룊�le 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 룊�le 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 룊�les in RDBMS too. Why Hadoop?

16. Hive partitioning – your project example? Why?

1. Hive – What 룊�le format do you use in your work? (Avro, Parquet, Sequence 룊�le)

2. Hadoop – What is the challenge or dif룊�culty you’ve faced?

3. PIG – What is the challenge or dif룊�culty you’ve faced?

4. Flume – What is the challenge or dif룊�culty you’ve faced?

5. Sqoop – What is the challenge or dif룊�culty you’ve faced? (he didn’t ask this question)

6. How experienced are you in Linux?

7. What shell type do you use?

8. How about your experience in Cloudera Manager?

9. How about your experience in Cloudera Manager?

10. Do you use Impala? (I compared it with Hive and explained in more details)

11. 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” 룊�le?

4. What is SecondaryNameNode?

5. Explain the MapReduce processing framework? (start to end)

6. What is Combiner? Where does it 룊�t 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.

    

3/22/2017 Real Time Hadoop Interview Questions From Different Readers ­ Hadoop Online Tutorials

http://hadooptutorial.info/real­time­hadoop­interview­questions­from­different­readers/ 3/10

8. Oozie – What are the nodes?

9. What are the actions in Action Node?

10. Explain your Pig project?

11. What log 韛�le 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 韛�lter 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 con韛�gurations? (RAM, CPU core, Disk size)

26. What is the NameNode con韛�gurations? (RAM, CPU core, Disk size)

27. How many Map slots & reducer slots con韛�gured in each DataNode? (he didn’t ask this)

28. How do you copy 韛�le 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

{

    

3/22/2017 Real Time Hadoop Interview Questions From Different Readers ­ Hadoop Online Tutorials

http://hadooptutorial.info/real­time­hadoop­interview­questions­from­different­readers/ 4/10

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

    

3/22/2017 Real Time Hadoop Interview Questions From Different Readers ­ Hadoop Online Tutorials

http://hadooptutorial.info/real­time­hadoop­interview­questions­from­different­readers/ 5/10

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 Shuf醜�e Phase in Map Reduce?

All the part 覜les will be exchanged between reduce tasks

part 覜les 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()

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 覜elds into same 覜eld.

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

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

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

Read record by record from each 覜le and compare 覜rst record from 覜rst 覜le with 覜rst record in 2nd

覜le and same way with 2nd record b/w the 覜les….

If 覜rst rec in 1st 覜le < 1st record 2nd 覜le then i will emit 1st record in 1st 覜le and i will move cursor

of 覜rst 覜le to 2nd record in the 覜rst 覜le then check with 1st record in 2nd 覜le and so on…

    

3/22/2017 Real Time Hadoop Interview Questions From Different Readers ­ Hadoop Online Tutorials

http://hadooptutorial.info/real­time­hadoop­interview­questions­from­different­readers/ 6/10

8

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

I will copy records from the remaining ⋈le 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 ⋈les of 1 TB on two disks, you should Merge the ⋈les 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 ⋈les are compared in the Map Reducer Phase?

If one of the ⋈le 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://collabnix.com/archives/458>