**(3)What is shuffleing in mapreduce?**

Once map tasks started to complete, A communication from reducers is started. where map output sent to reducer, which is looking for the output data to process. at same time data nodes are still process multiple other tasks. The data transfer of mappers output to reducer known as shuffling.

**(4)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 identifies the reducer as an recipient of mapper output. All the key, no matter which mapper has generated this, must lie with same reducer.

**(16)What does the mapred.job.tracker command do?**

**(21)What is a “map” in Hadoop?**

>Mapper task processes each input record and it generates a new <key, value> pairs.

>output-(Key,value )

**(22)What is a “reducer” in Hadoop?**

>Reducer takes a set of an intermediate key-value pair produced by the mapper as the input. After that, it runs a reduce function on each of them to generate the output. The output of the reducer is the final output, which it store in HDFS

>Output- (key value List)

**(26)What is the difference between Input Split and an HDFS Block?**

Logical division- Input Split dependant on size of block

Phisical division- Block-default (128mb)

**(29)What is the difference between TextInputFormat and KeyValueInputFormat class?**

TextInputFormat:

1) It reads lines of text files and provides the offset of the line as key to the Mapper and actual line as Value to the mapper.

2) TextInputFormat is the default file format in Hadoop .

KeyValueInputFormat:

1) Reads text file and parses lines into key, Val pairs. Everything up to the first tab character is sent as key to the Mapper and the remainder of the line is sent as value to the mapper.

**(30)What is InputSplit in Hadoop?**

It is the input to the mapper task

**(31)How is the splitting of file invoked in Hadoop framework**

based on block size

**(34)After the Map phase finishes, the Hadoop framework does “Partitioning, Shuffle and sort”. Explain what happens in this phase?**

**Shuffle Phase-**Once the first map tasks are completed, the nodes continue to perform several other map tasks and also exchange the intermediate outputs with the reducers as required. This process of moving the intermediate outputs of map tasks to the reducer is referred to as Shuffling.

**Sort Phase-** Hadoop MapReduce automatically sorts the set of intermediate keys on a single node before they are given as input to the reducer.

**Partitioning Phase-**The process that determines which intermediate keys and value will be received by each reducer instance is referred to as partitioning. The destination partition is same for any key irrespective of the mapper instance that generated it.

**(79)Can I set the number of reducers to zero?**

yes

**(83)If reducers do not start before all mappers finish then why does the progress on MapReduce job shows something like Map(50%) Reduce(10%)? Why reducers progress percentage is displayed when mapper is not finished yet?**

Reducers start copying intermediate key-value pairs from the mappers as soon as they are available. The progress calculation also takes in account the processing of data transfer which is done by reduce process, therefore the reduce progress starts showing up as soon as any intermediate key-value pair for a mapper is available to be transferred to reducer. Though the reducer progress is updated still the programmer defined reduce method is called only after all the mappers have finished.

**(84)What is next step after Mapper or MapTask?**

Combiner or reducer

**(86)Does MapReduce programming model provide a way for reducers to communicate with each other? In a MapReduce job can a reducer communicate with another reducer?**

Not required

**(90)Where is the Mapper Output (intermediate kay-value data) stored ?**

**On local file system of slave node**

**(91)What is identity mapper and reducer? In which cases can we use them?**

* org.apache.hadoop.mapred.lib.IdentityMapper Implements the identity function, mapping inputs directly to outputs. If MapReduce programmer do not set the Mapper Class using JobConf.setMapperClass then IdentityMapper.class is used as a default value.
* org.apache.hadoop.mapred.lib.IdentityReducer Performs no reduction, writing all input values directly to the output. If MapReduce programmer do not set the Reducer Class using JobConf.setReducerClass then IdentityReducer.class is used as a default value.

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

**Yes, It is possible to use multiple inputs in hadoop. There are various ways in which this can be done -**

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

2. Use FileInputFormat.addInputPaths() method, that can take a comma separated list of multiple inputs ex -

FileInputFormat.addInputPaths(“user1/file0.gz,user2.file.gz…………”).

3. Use Multiple mappers.

4. Use input file in Distributed Cache.

5. Use MultipleInputs.addInputPath() method tospecify different input files like -

MultipleInputs.addInputPath(job, ClouderaPath, TextInputFormat.class, JoinclouderaMapper.class);

MultipleInputs.addInputPath(job, HdpPath, TextInputFormat.class, HdpMapper.class)

**(107)What is ‘Key value pair’ in HDFS?**

**(109)Is map like a pointer?**

**(111)Why are the number of splits equal to the number of maps?**

**(118)What are the four basic parameters of a mapper?**

LongWritable, text, text and IntWritable. The first two represent input parameters and the second two represent intermediate output parameters.

**(119)What is the input type/format in MapReduce by default?**

Text InputFormat- It is the default InputFormat of MapReduce

**(128)On What concept the Hadoop framework works?**

**Map reduce**

**(153)Where do you specify the Mapper Implementation?**

**(160)How would you tackle calculating the number of unique visitors for each hour by mining a huge**

**Apache log? You can use post processing on the output of the MapReduce job.**

**(161)How did you debug your Hadoop code ?**

**(162)How will you write a custom partitioner for a Hadoop job?**

**(163)How can you add the arbitrary key-value pairs in your mapper?**

**(165)What are combiners? When should I use a combiner in my MapReduce Job? (166)How Mapper is instantiated in a running job?**

**(167)Which interface needs to be implemented to create Mapper and Reducer for the Hadoop?**

**171)Explain the WordCount implementation via Hadoop framework ?**

**(172)What Mapper does?**

**(173)what is MAP REDUCE?**

**(174)Explain the Reducer?s Sort phase?**

**(175)What are the primary phases of the Reducer?**

**(176)Explain the Reducer's reduce phase?**

**(177)Explain the shuffle?**

**(178)What happens if number of reducers are 0?**

**(179)How many Reducers should be configured?**

**(180)What is Writable & WritableComparable interface?**

**(181)What is the Hadoop MapReduce API contract for a key and value Class?**

**(182)Where is the Mapper Output (intermediate kay-value data) stored ?**

**(186)can we write map reduce program in other than java programming language. how.**

**(188)What is the use of Context object?**

**(189)What is the Reducer used for?**

**(190)What is the use of Combiner?**

**(194)How does Mappers run() method works?**

**(196)It can be possible that a Job has 0 reducers? (197)How many maps are there in a particular Job?**

**(199)How can we control particular key should go in a specific reducer?**

**(203)Describe what happens to a MapReduce job from submission to output?**

**Cache during runtime of the application**

**(206)Did you ever ran into a lop sided job that resulted in out of memory error, if yes then how did you handled it**

**215)Does MapReduce programming model provide a way for reducers to communicate with each other? In a MapReduce job can a reducer communicate with another reducer?**

**(217)What are combiners? When should I use a combiner in my MapReduce Job?**

**(219)When is the reducers are started in a MapReduce job?**

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)

4. Tell me some Map implementations?

HashMap (unsorted)

TreeMap (Sorted values)

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

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

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

6. What Happens in Shuffle Phase in Map Reduce?

All the part files will be exchanged between reduce tasks

part files will be generated by partitioners

map output will be transferred over network…

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

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

locations…Usually Buckets are array….

9. What are the parameters in Mapper class

map(key, value, context)

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

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

37. What is SetUp() and CleanUp()?

These Mapreduce methods include at the start and end of the each split.

SetUp for initialize the resources.

Map and reduce is processing the data.

Cleanup is close the resources.

Comparison also trigger. Setup and cleanup is trigger in both map and reducer.

Map reduce can give record level control, but these two can give block level control. file level also allows in Input Format.

**(245)what is difference between int and intwritable**

**(247)In Map Reduce why map write output to Local Disk instead of HDFS?**

**(255) what are mapreduce new and old apis while writing map reduce program . explain how it works**

**(256)How to write a Custom Key Class?**

**(257)What is the utility of using Writable Comparable (Custom Class) in Map Reduce code?**

**(258)What are Input Format, Input Split & Record Reader and what they do?**

**(259)Why we use IntWritable instead of Int? Why we use LongWritable instead of Long?**

**(277)Why Mapreduce output written in local disk**

**(289)What is a Record Reader in hadoop?**

**(311)What is difference between reducer and combiner?**

The main use of combiner is to reduce the number of key value pairs which is passed from mapper to reducer so that network traffic can be reduced.

A [**Combiner**](http://data-flair.training/blogs/combiner-in-hadoop-mapreduce-advantages-disadvantages/) in **[Hadoop](http://data-flair.training/blogs/hadoop-introduction-tutorial-quick-guide/)** is a **mini**[**reducer**](http://data-flair.training/blogs/reducer-in-hadoop-mapreduce/) that performs the local reduce task. Many **[MapReduce](http://data-flair.training/blogs/hadoop-mapreduce-introduction-tutorial-comprehensive-guide/)** jobs are limited by the network bandwidth available on the cluster, so the combiner minimizes the data transferred between map and reduce tasks.

A [**Combiner**](http://data-flair.training/blogs/combiner-in-hadoop-mapreduce-advantages-disadvantages/) is executed (optionally) after the **[Mapper](http://data-flair.training/blogs/mapper-in-hadoop-mapreduce/)** phase in the same Node which runs the Mapper. So there is no Network I/O involved. Thus is it also know as a **Local Reducer**. It does similar things as a Reducer (i.e.) group data in order to reduce Network traffic between Mapper and Reducer Nodes.

A [**Reducer**](http://data-flair.training/blogs/reducer-in-hadoop-mapreduce/), on the other hand, is executed after data from multiple Mappers are partitioned and based on an algorithm they are Shuffled into various Nodes across the network. When the data is of [**Key/Value pair**](http://data-flair.training/blogs/key-value-pairs-hadoop-mapreduce/) type, Values with the same Key always lands at the same Reducer

**(56)How can you set an arbitrary number of mappers to be created for a job in Hadoop?**

**3-Explain the usage of Context Object.**

Context Object is used to help the mapper interact with other Hadoop systems. Context Object can be used for updating counters, to report the progress and to provide any application level status updates. ContextObject has the configuration details for the job and also interfaces, that helps it to generating the output.

**(58)How will you write a custom partitioner for a Hadoop job?**

**35.What is Combiner? Where does it fit and give an example? Preferably from your project.**

**36.What is Partitioner? Why do you need it and give an example? Preferably from your project.**