**Q1. Name the most common Input Formats defined in Hadoop? Which one is default?**

 The two most common Input Formats defined in Hadoop are:

 – TextInputFormat

- KeyValueInputF2ormat

- SequenceFileInputFormat

 TextInputFormat is the Hadoop default.

**Q2. What is the difference between TextInputFormat and KeyValueInputFormat class?**

**TextInputFormat:** 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.

**KeyValueInputFormat:** 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.

**Q3. What is InputSplit in Hadoop?1**

 When a Hadoop job is run, it splits input files into chunks and assign each split to a mapper to process. This is called InputSplit.

**Q4. How is the splitting of file invoked in Hadoop framework?**

 It is invoked by the Hadoop framework by running getInputSplit()method of the Input format class (like FileInputFormat) defined by the user.

**Q5. Consider case scenario: In M/R system, - HDFS block size is 64 MB**

**- Input format is FileInputFormat**

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

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

Hadoop will make 5 splits as follows:

- 1 split for 64K files

- 2 splits for 65MB files

- 2 splits for 127MB files

**Q6. What is the purpose of RecordReader in Hadoop?**

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

**Q7. After the Map phase finishes, the Hadoop framework does “Partitioning, Shuffle and sort”. Explain what happens in this phase?**

**Partitioning:** It is the process of determining which reducer instance will receive which intermediate keys and values. Each mapper must determine for all of its output (key, value) pairs which reducer will receive1 them. It is necessary that for any key, regardless of which mapper instance generated it, the destination partition is the same.

**Shuffle:** After the first map tasks have completed, the nodes may still be performing several more map tasks each. But they also begin exchanging the intermediate outputs from the map tasks to where they are required by the reducers. This process of moving map outputs to the reducers is known as shuffling.

**Sort:** Each reduce task is responsible for reducing the values associated with several intermediate keys. The set of intermediate keys on a single node is automatically sorted by Hadoop before they are presented to the Reducer.

**Q8. If no custom partitioner is defined in Hadoop then how is data partitioned before it is sent to the reducer?**

 The default partitioner computes a hash value for the key and assigns the partition based on this result.

**Q9. What is a Combiner?**

The Combiner is a ‘mini-reduce’ process which operates only on data generated by a mapper. The Combiner will receive as input all data emitted by the Mapper instances on a given node. The output from the Combiner is then sent to the Reducers, instead of the output from the Mappers.

**Q10. What is JobTracker?**

 JobTracker is the service within Hadoop that runs MapReduce jobs on the cluster.

**Q11. What are some typical functions o1f Job Tracker?**

 The following are some typical tasks of JobTracker:-

- Accepts jobs from clients

- It talks to the NameNode to determine the location of the data.

- It locates TaskTracker nodes with available slots at or near the data.

- It submits the work to the chosen TaskTracker nodes and monitors progress of each task by receiving heartbeat signals from Task tracker.

**Q12. What is TaskTracker?**

 TaskTracker is a node in the cluster that accepts tasks like MapReduce and Shuffle operations – from a JobTracker.

**Q13. What is the relationship between Jobs and Tasks in Hadoop?**

 One job is broken down into one or many tasks in Hadoop.

**Q14. Suppose Hadoop spawned 100 tasks for a job and one of the task failed. What will Hadoop do?**

 It will restart the task again on some other TaskTracker and only if the task fails more than four (default setting and can be changed) times will it kill the job.

**Q15. Hadoop achieves parallelism by dividing the tasks across many nodes, it is possible for a few slow nodes to rate-limit the rest of the program and slow down the program. What mechanism Hadoop provides to combat this?**

 Speculative Execution.

**Q16. How does speculative execution work in Hadoop?**

JobTracker makes different TaskTrackers pr2ocess same input. When tasks complete, they announce this fact to the JobTracker. Whichever copy of a task finishes first becomes the definitive copy. If other copies were executing speculatively, Hadoop tells the TaskTrackers to abandon the tasks and discard their outputs. The Reducers then receive their inputs from whichever Mapper completed successfully, first.

**Q17. Using command line in Linux, how will you**

-**See all jobs running in the Hadoop cluster**

**- Kill a job?**

 Hadoop job – list

Hadoop job – kill jobID

**Q18. What is Hadoop Streaming?**

Streaming is a generic API that allows programs written in virtually any language to be used as Hadoop Mapper and Reducer implementations.

**Q19. What is the characteristic of streaming API that makes it flexible run MapReduce jobs in languages like Perl, Ruby, Awk etc.?**

Hadoop Streaming allows to use arbitrary programs for the Mapper and Reducer phases of a MapReduce job by having both Mappers and Reducers receive their input on stdin and emit output (key, value) pairs on stdout.

**Q20. What is Distributed Cache in Hadoop?**

Distributed Cache is a facility provided by the MapReduce framework to cache files (text, archives, jars and so on) needed by applications during execution of the job. The framework will copy the necessary files to the slave node before any tasks for the job are executed on that node.

**Q21. What is the benefit of Distributed cache? Why can we just have the file in HDFS and have the application read it?**

This is because distributed cache is much faster. It copies the file to all trackers at the start of the job. Now if the task tracker runs 10 or 100 Mappers or Reducer, it will use the same copy of distributed cache. On the other hand, if you put code in file to read it from HDFS in the MR Job then every Mapper will try to access it from HDFS hence if a TaskTracker run 100 map jobs then it will try to read this file 100 times from HDFS. Also HDFS is not very efficient when used like this.

**Q.22 What mechanism does Hadoop framework provide to synchronise changes made in Distribution Cache during runtime of the application?**

This is a tricky question. There is no such mechanism. Distributed Cache by design is read only during the time of Job execution.

**Q23. Have you ever used Counters in Hadoop. Give us an example scenario?**

Anybody who claims to have worked on a Hadoop project is expected to use counters.

**Q24. 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, the input format class provides methods to add multiple directories as input to a Hadoop job.

**Q25. Is it possible to have Hadoop job output in multiple directories? If yes, how?**

Yes, by using Multiple Outputs class.

**Q26. What will a Hadoop job do if you try to run it with an output directory that is already present? Will it**

**- Overwrite it**

**- Warn you and continue**

**- Throw an exception and exit**

The Hadoop job will throw an exception and exit.

**Q27. How can you set an arbitrary number of mappers to be created for a job in Hadoop?**

You cannot set it.

**Q28. How can you set an arbitrary number of Reducers to be created for a job in Hadoop?**

You can either do it programmatically by using method setNumReduceTasks in the Jobconf Class or set it up as a configuration setting.

**Q29. How will you write a custom partitioner for a Hadoop job?**

 To have Hadoop use a custom partitioner you will have to do minimum the following three:

- Create a new class that extends Partitioner Class

- Override method getPartition

- In the wrapper that runs the Mapreduce, either

- Add the custom partitioner to the job programmatically using method set Partitioner Class or – add the custom partitioner to the job as a config file (if your wrapper reads from config file or oozie)

**Q30. How did you debug your Hadoop code?**

 There can be several ways of doing this but most common ways are:-

- By using counters.

- The web interface provided by Hadoop framework.

**Q31. Did you ever built a production process in Hadoop? If yes, what was the process when your Hadoop job fails due to any reason?**

It is an open-ended question but most candidates if they have written a production job, should talk about some type of alert mechanism like email is sent or there monitoring system sends an alert. Since [Hadoop works on unstructured data,](http://wiziqlmp.wpengine.com/decoding-big-data-analytics-hadoop/) it is very important to have a good alerting system for errors since unexpected data can very easily break the job.

**1.What is BIG DATA?**

Ans: Big Data is nothing but an assortment of such a huge and complex data that it becomes very tedious to capture, store, process, retrieve and analyze it with the help of on-hand database management tools or traditional data processing techniques.

**2. Can you give some examples of Big Data?**

Ans: There are many real life examples of Big Data! Facebook is generating 50 of data per day, NYSE (New York Stock Exchange) generates about 1 terabyte 11Zl data per day, a jet airline collects 10 terabytes of censor data for every 30 time. All these are day to day examples of Big Data!

**3. Can you give a detailed overview about the by Facebook? •**

tes

ying

ta bLgenerated

Ans: As of December 31, 2012, there are 1.06 billion million mobile users. On an average, 3.2 billion like Facebook. 72% of web audience is on Facebook. on facebook from wall posts, sharing images, fact, Facebook started using Hadoop in mid

sers on Facebook and 680 s are posted every day on re are so many activities going omments and liking posts, etc. In e of the initial users of Hadoop.

4. **According to IBM, what a aracteristics of Big Data?**

Ans: According to IBM, the 500+ terabytes of data per reason for losses. Varie i

**5. How Big is Bi**

• Ans: With time e is growing exponentially. Earlier we used to talk about Megabytes or Gigabytes s arrived when we talk about data volume in terms of terabytes, petabytes ettabytes! Global data volume was around 1.8ZB in 2011 and is expected to be 7.9ZB also known that the global information doubles in every two years!

s of Big Data are: Volume: Facebook generating a yzing 2 million records each day to identify the ideo, sensor data, log files, etc.

**6. Ho lysis of Big Data is useful for organizations?**

Ans: Effective analysis of Big Data provides a lot of business advantage as organiz.ations will learn which areas to focus on and which areas are less important. Big data analysis provides some early key indicators that can prevent the company from a huge loss or help in grasping a great opportunity with open hands! A precise analysis of Big Data helps in decision making! For instance, nowadays people rely so much on Facebook and Twitter before buying any product or service. All thanks to the Big Data explosion.

**7. Who are Data Scientists?**

Ans: Data scientists are soon replacing business analysts or data analysts. Data scientists are experts who find solutions to analyze data. Just as web analysis, we have data scienti ho have good business insight as to how to handle a business challenge. Sharp data sc' e not only involved in dealing business problems, but also choosing the relevant issues t b ng value-addition to the organization.

**8. What is Hadoop?**

Ans. Hadoop is a framework that allows for distributed pros g of la lata sets across clusters of commodity computers using a simple progra el. 9. Why the name Hadoop? Ans. Hadoop doesn't have any expanding r s. The charming yellow elephant

you see is basically named after Dougs 's

hant,

**10. Why do we need Had Ans. Everyday a large a k... ctured data is getting dumped into our machines. The major challenge is of for 1 e data sets in our systems but to retrieve and analyze the big data in the** orga t too data present in different machines at different locations. In this situation ec ity for Hadoop arises. Hadoop has the ability to analyze the data present in d erent chines at different locations very quickly and in a very cost effective way. It us colkept of MapReduce which enables it to divide the query into small parts an em in parallel. This is also known as parallel computing.

**11. Wh e of the characteristics of Hadoop framework?**

Ans. ramework is written in Java. It is designed to solve problems that involve analyzing rge data (e.g. petabytes). The programming model is based on Google's MapReduce. The infrastructure is based on Google's Big Data and Distributed File System. Hadoop handles large files/data throughput and supports data intensive distributed applications. Hadoop is scalable as more nodes can be easily added to it.

**12. Give a brief overview of Hadoop history.**

Ans. In 2002, Doug Cutting created an open source, web crawler project. In 2004, Google published MapReduce, GFS papers. In 2006, Doug Cutting developed thenpen source, Mapreduce and HDFS project In 2008, Yahoo ran 4,000 node Hadoop cluster and Hadoop won terabyte sort benchmark. In 2009, Facebook launched SQL support for Hadoop.

**13. Give examples of some companies that are using Hadoop str ture?**

Ans. A lot of companies are using the Hadoop structure such as Cloudera, EMC, M Hortonworks, Amazon, Facebook, eBay, Twitter, Google and so on.

**14. What is the basic difference between traditio 1 RDrkSjind Hadoop? •**

Ans. Traditional RDBMS is used for transactional syst whereas Hadoop is an approach to store huge am and process it. RDBMS will be useful when you w whereas, Hadoop will be useful when you wa B1 that later.

**15. What is structured and u ata?**

Ans. Structured data is the da most common form of stTuc tables, that is, rows and co easily. It could be in th in the form of ms's and

d archive the data, distributed file system record from Big data, shot and perform analysis on

'd ntifiable as it is organized in a structure. The base where specific information is stored in ured data refers to any data that cannot be identified ens, documents, email, logs and random text. It is not

**16. What are tlreJmponents of Hadoop?**

Ans. Core co large da

f Hadoop are HDFS and Map Reduce. HDFS is basically used to store ap Reduce is used to process such large data sets.

**17. HDFS?**

Ans. HDFS is a file system designed for storing very large files with streaming data access patterns, running clusters on commodity hardware. 18.What are the key features of HDFS? Ans. HDFS is highly fault-tolerant, with high throughput, suitable for applications with large data sets, streaming access to file system data and can be built out of commodity hardware.

**19. What is Fault Tolerance?**

Ans: Suppose you have a file stored in a system, and due to some technical problem that file gets destroyed. Then there is no chance of getting the data back present in that file. To avoid such situations, Hadoop has introduced the feature of fault tolerance in HDFS. In Hadoop, when we store a file, it automatically gets replicated at two other locations also. So even if one or two of the systems collapse, the file is still available on the third system.

**20.What is the difference between a Hadoop database and Rel Database?**

Ans: Hadoop is not a database, it is an architecture with a filesystem c3r,(iii4The data is stored in HDFS which does not have any predefined containers. elati al d se stores data in predefined containers.

•

**21.What is MAP REDUCE?**

Ans: Map Reduce is a set of programs used to ac man ulate large data sets over a Hadoop cluster.

**22.What is the InputSplit in m**

**software?**

Ans: An inputsplit is the slice of ssed by a single Mapper. It generally is of the block size which is stored on the

**23.what is meaning factor?**

Ans. Replication fact r de number of times a given data block is stored in the cluster. The defau r licati factor is 3. This also means that you need to have 3times the amount of stora • e &store the data. Each file is split into data blocks and spread across the clus

**24.what ult replication factor in HDFS?**

Ans. It hadoop comes with 3 replication factor. You can set the replication level individua for each file in HDFS. In addition to fault tolerance having replicas allow jobs that consume the same data to be run in parallel. Also if there are replicas of the data hadoop can attempt to run multiple copies of the same task and take which ever finishes first. This is useful if for some reason a box is being slow. Most Hadoop administrators set the default replication factor for their files to be three. The main assumption here is that if you keep three copies of the data, your data is safe. this to be true in the big clusters that we manage and operate. In addition to fault tolerance having replicas allow jobs that consume the same data to be run in parallel. Also if there are replicas of the data hadoop can attempt to run multiple copies of the same task and take which ever finishes first. This is useful if for some reason a box is being slow.

**25. what is the typical block size of an HDFS block?** Ans: Default blocksize is 64mb. But 128mb is typical. **26. How does master slave architecture in the Hadoop?**

Ans: Totally 5 daemons run in Hadoop Master-slave architecture . On Master Node : Name Node and Job Tracker and Secondary name node On Slave : Data Node and Task T er But its recommended to run Secondary name node in a separate machine which h node capacity.

**27. What is compute and Storage nodes?**

C94 Ans: I do define Hadoop into 2 ways : Distributed Processing : a cVlistributed Storage : HDFS Name Node holds Meta info and Data holds exact dates 'ts MR p gram.

**28. Explain how input and output data**

e adoop framework?

Ans: Fileinputformat, textinputformat, ke etc rmat, sequencefileinputformat, sequencefileasinputtextformat, wholefilefljn. e t rmats in hadoop framework

29**. How can we control pa hould go in a specific reducer?**

Ans. By using a custom part% 30.What is the Re s or? Ans. Reducer is us d cot, e the multiple outputs of mapper to one. 31. What h imary phases of the Reducer? Ans: r 3 primary phases: shuffle, sort and reduce. **32.Wh appens if number of reducers are 0?**

Ans: It is legal to set the number of reduce-tasks to zero if no reduction is desired. In this case the outputs of the map-tasks go directly to the FileSystem, into the output path set by setOutputPath(Path). The framework does not sort the map-outputs before writing them out to the FileSystem.

**33.How many instances of JobTracker can run on a HadoopCluser?**

Ans: One. There can only be one JobTracker in the cluster. This can be run on the same machine running the NameNode. 34. How NameNode Handles data node failures? Ans. Through checksums. every data has a record followed by a checksum. if the m doesnot match with the original then it reports an data corrupted error.

35. Can I set the number of reducers to zero? Ans. can be given as zero. So, the mapper output is anfinalise utput acthigi res in HDFS. 36. What is a SequenceFile in Hadoop? Ans. A. Sequence File contains a binary encoding bitra number of homogeneous writable objects.

B. ASequenceFilecontains a binary encodi writable objects.

number of heterogeneous

C. ASequenceFilecontains a bina n ar itTary number of WritableComparable objects, in sorted order.

D. ASequenceFilecontains "nary nc g of an arbitrary number key-value pairs. Each key must be the same type. h ch ue e sametype.

**37. How many s tes des Writable interface defines in Hadoop? •**

Ans. Two

38. quence files and why are they important in Hadoop? Ans: S nee files are binary format files that are compressed and are splitable. They are often used in high-performance map-reduce jobs 39. What are map files and why are they important in Hadoop? Ans: Map files are sorted sequence files that also have an index. The index allows fast data look up.

**40.How can you use binary data in MapReduce in Hadoop?**

Ans. Binary data can be used directly by a map-reduce job. Often binary data is added to a sequence file. **41. What is map — side join in Hadoop?** Ans. Map-side join is done in the map phase and done in memory 42.What is reduce — side join in Hadoop?

Ans: Reduce-side join is a technique for merging data from different s icesICIed on a specific key. There are no memory restrictions

**43.How can you disable the reduce step in**

Ans : A developer can always set the number of t disable the reduce step.

**44.Why would a developer cre Hadoop?**

NA\? • ers t ero. That will completely -reduce without the reduce step

Ans: There is a CPU intensive%ccu etween the map and reduce steps. Disabling the reduce step speeds up data p ce in

45. What is the de

ormat in Hadoop?

Ans: The default i mitt s TextInputFormat with byte offset as a key and entire line as a value.

46. Ho overwrite the default input format in Hadoop?

Ans to overwrite default input format, a developer has to net new input format on job con efore submitting the job to a cluster.

47. Is there a map input format in Hadoop?

Ans: No, but **sequence** file input format can read map files.

48.What happens if mapper output does not match reducer input in Hadoop? Ans: A real-time exception will be thrown and map-reduce job will fail. 49. Can you provide multiple input paths to a map-reduce jobs Hadoop? Ans: Yes, developers can add any number of input paths.

50. Since the data is replicated thrice in HDFS, does it mea calculation done on one node will also be replicated on o?

Ans: Since there are 3 nodes, when we send the MapReduce done only on the original data. The master node will kno, particular data. In case, if one of the nodes is not resp then, the required calculation will be done on the second replica.