1) What is Hadoop Map Reduce ?

For processing large data sets in parallel across a hadoop cluster, Hadoop MapReduce framework is used. Data analysis uses a two-step map and reduce process.

2) How Hadoop MapReduce works?

In MapReduce, during the map phase it counts the words in each document, while in the reduce phase it aggregates the data as per the document spanning the entire collection. During the map phase the input data is divided into splits for analysis by map tasks running in parallel across Hadoop framework.

3) Explain what is shuffling in MapReduce ?

The process by which the system performs the sort and transfers the map outputs to the reducer as inputs is known as the shuffle

4) Explain what is distributed Cache in MapReduce Framework ?

Distributed Cache is an important feature provided by map reduce framework. When you want to share some files across all nodes in Hadoop Cluster, DistributedCache is used. The files could be an executable jar files or simple properties file.

5) Explain what is NameNode in Hadoop?

NameNode in Hadoop is the node, where Hadoop stores all the file location information in HDFS (Hadoop Distributed File System). In other words, NameNode is the centrepiece of an HDFS file system. It keeps the record of all the files in the file system, and tracks the file data across the cluster or multiple machines

6) Explain what is JobTracker in Hadoop? What are the actions followed by Hadoop?

In Hadoop for submitting and tracking MapReduce jobs, JobTracker is used. Job tracker run on its own JVM process

Hadoop performs following actions in Hadoop

- Client application submit jobs to the job tracker
- JobTracker communicates to the Namemode to determine data location
- Near the data or with available slots JobTracker locates TaskTracker nodes
- On chosen TaskTracker Nodes, it submits the work
- When a task fails, Job tracker notify and decides what to do then.
- The TaskTracker nodes are monitored by JobTracker

7) Explain what is heartbeat in HDFS?

Heartbeat is referred to a signal used between a data node and Name node, and between task tracker and job tracker, if the Name node or job tracker does not respond to the signal, then it is considered there is some issues with data node or task tracker

8) Explain what combiners is and when you should use a combiner in a MapReduce Job?

To increase the efficiency of MapReduce Program, Combiners are used. The amount of data can be reduced with the help of combiner's that need to be transferred across to the reducers. If the operation performed is commutative and associative you can use your reducer code as a combiner. The execution of combiner is not guaranteed in Hadoop

9) What happens when a datanode fails ?

When a datanode fails

- Jobtracker and namenode detect the failure
- On the failed node all tasks are re-scheduled
- Namenode replicates the users data to another node

10) Explain what is Speculative Execution?

In Hadoop during Speculative Execution a certain number of duplicate tasks are launched. On different slave node, multiple copies of same map or reduce task can be executed using Speculative Execution. In simple words, if a particular drive is taking long time to complete a task, Hadoop will create a duplicate task on another disk. Disk that finish the task first are retained and disks that do not finish first are killed.

11) Explain what are the basic parameters of a Mapper?

The basic parameters of a Mapper are

- LongWritable and Text
- Text and IntWritable

12) Explain what is the function of MapReducer partitioner?

The function of MapReducer partitioner is to make sure that all the value of a single key goes to the same reducer, eventually which helps evenly distribution of the map output over the reducers

13) Explain what is difference between an Input Split and HDFS Block?

Logical division of data is known as Split while physical division of data is known as HDFS Block

14) Explain what happens in textinformat ?

In textinputformat, each line in the text file is a record. Value is the content of the line while Key is the byte offset of the line. For instance, Key: longWritable, Value: text

15) Mention what are the main configuration parameters that user need to specify to run Mapreduce Job ?

The user of Mapreduce framework needs to specify

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

16) Explain what is WebDAV in Hadoop?

To support editing and updating files WebDAV is a set of extensions to HTTP. On most operating system WebDAV shares can be mounted as filesystems , so it is possible to access HDFS as a standard filesystem by exposing HDFS over WebDAV.

17) Explain what is sqoop in Hadoop ?

To transfer the data between Relational database management (RDBMS) and Hadoop HDFS a tool is used known as Sqoop. Using Sqoop data can be transferred from RDMS like MySQL or Oracle into HDFS as well as exporting data from HDFS file to RDBMS

18) Explain how JobTracker schedules a task ?

The task tracker send out heartbeat messages to Jobtracker usually every few minutes to make sure that JobTracker is active and functioning. The message also informs JobTracker about the number of available slots, so the JobTracker can stay upto date with where in the cluster work can be delegated

19) Explain what is Sequencefileinputformat?

Sequencefileinputformat is used for reading files in sequence. It is a specific compressed binary file format which is optimized for passing data between the output of one MapReduce job to the input of some other MapReduce job.

20) Explain what does the conf.setMapper Class do ?

Conf.setMapperclass sets the mapper class and all the stuff related to map job such as reading data and generating a key-value pair out of the mapper

21) Explain what is Hadoop?

It is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides enormous processing power and massive storage for any type of data.

22) Mention what is the difference between an RDBMS and Hadoop?

RDBMS	Hadoop
RDBMS is relational database management system	Hadoop is node based flat structure
It used for OLTP processing whereas Hadoop	It is currently used for analytical and for BIG DATA processing
In RDBMS, the database cluster uses the same data files stored in shared storage	In Hadoop, the storage data can be stored independently in each processing node.
You need to preprocess data before storing it	you don't need to preprocess data before storing it

23) Mention Hadoop core components?

Hadoop core components include,

- HDFS
- MapReduce

24) What is NameNode in Hadoop?

NameNode in Hadoop is where Hadoop stores all the file location information in HDFS. It is the master node on which job tracker runs and consists of metadata.

25) Mention what are the data components used by Hadoop?

Data components used by Hadoop are

- Pig
- Hive

26) Mention what is the data storage component used by Hadoop?

The data storage component used by Hadoop is HBase.

27) Mention what are the most common input formats defined in Hadoop?

The most common input formats defined in Hadoop are;

- TextInputFormat
- KeyValueInputFormat
- SequenceFileInputFormat

28) In Hadoop what is InputSplit?

It splits input files into chunks and assign each split to a mapper for processing.

29) For a Hadoop job, how will you write a custom partitioner?

You write a custom partitioner for a Hadoop job, you follow the following path

- Create a new class that extends Partitioner Class
- Override method getPartition
- In the wrapper that runs the MapReduce
- Add the custom partitioner to the job by using method set Partitioner Class or – add the custom partitioner to the job as a config file

30) For a job in Hadoop, is it possible to change the number of mappers to be created?

No, it is not possible to change the number of mappers to be created. The number of mappers is determined by the number of input splits.

31) Explain what is a sequence file in Hadoop?

To store binary key/value pairs, sequence file is used. Unlike regular compressed file, sequence file support splitting even when the data inside the file is compressed.

32) When Namenode is down what happens to job tracker?

Namenode is the single point of failure in HDFS so when Namenode is down your cluster will set off.

33) Explain how indexing in HDFS is done?

Hadoop has a unique way of indexing. Once the data is stored as per the block size, the HDFS will keep on storing the last part of the data which say where the next part of the data will be.

34) Explain is it possible to search for files using wildcards?

Yes, it is possible to search for files using wildcards.

35) List out Hadoop's three configuration files?

The three configuration files are

- core-site.xml
- mapred-site.xml
- hdfs-site.xml

36) Explain how can you check whether Namenode is working beside using the jps command?

Beside using the jps command, to check whether Namenode are working you can also use

/etc/init.d/hadoop-0.20-namenode status.

37) Explain what is "map" and what is "reducer" 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.

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

38) In Hadoop, which file controls reporting in Hadoop?

In Hadoop, the hadoop-metrics.properties file controls reporting.

39) For using Hadoop list the network requirements?

For using Hadoop the list of network requirements are:

- Password-less SSH connection
- Secure Shell (SSH) for launching server processes

40) Mention what is rack awareness?

Rack awareness is the way in which the namenode determines on how to place blocks based on the rack definitions.

41) Explain what is a Task Tracker in Hadoop?

A Task Tracker in Hadoop is a slave node daemon in the cluster that accepts tasks from a JobTracker. It also sends out the heartbeat messages to the JobTracker, every few minutes, to confirm that the JobTracker is still alive.

42) Mention what daemons run on a master node and slave nodes?

- Daemons run on Master node is "NameNode"
- Daemons run on each Slave nodes are “Task Tracker” and "Data"

43) Explain how can you debug Hadoop code?

The popular methods for debugging Hadoop code are:

- By using web interface provided by Hadoop framework
- By using Counters

44) Explain what is storage and compute nodes?

- The storage node is the machine or computer where your file system resides to store the processing data
- The compute node is the computer or machine where your actual business logic will be executed.

45) Mention what is the use of Context Object?

The Context Object enables the mapper to interact with the rest of the Hadoop

system. It includes configuration data for the job, as well as interfaces which allow it to emit output.

46) Mention what is the next step after Mapper or MapTask?

The next step after Mapper or MapTask is that the output of the Mapper are sorted, and partitions will be created for the output.

47) Mention what is the number of default partitioner in Hadoop?

In Hadoop, the default partitioner is a “Hash” Partitioner.

48) Explain what is the purpose of RecordReader in Hadoop?

In Hadoop, the RecordReader loads the data from its source and converts it into (key, value) pairs suitable for reading by the Mapper.

49) Explain how is data partitioned before it is sent to the reducer if no custom partitioner is defined in Hadoop?

If no custom partitioner is defined in Hadoop, then a default partitioner computes a hash value for the key and assigns the partition based on the result.

50) Explain what happens when Hadoop spawned 50 tasks for a job and one of the task failed?

It will restart the task again on some other TaskTracker if the task fails more than the defined limit.

51) Mention what is the best way to copy files between HDFS clusters?

The best way to copy files between HDFS clusters is by using multiple nodes and the distcp command, so the workload is shared.

52) Mention what is the difference between HDFS and NAS?

HDFS data blocks are distributed across local drives of all machines in a cluster while NAS data is stored on dedicated hardware.

53) Mention how Hadoop is different from other data processing tools?

In Hadoop, you can increase or decrease the number of mappers without worrying about the volume of data to be processed.

54) Mention what job does the conf class do?

Job conf class separate different jobs running on the same cluster. It does the job level settings such as declaring a job in a real environment.

55) Mention what is the Hadoop MapReduce APIs contract for a key and value class?

For a key and value class, there are two Hadoop MapReduce APIs contract

- The value must be defining the org.apache.hadoop.io.Writable interface
- The key must be defining the org.apache.hadoop.io.WritableComparable interface

56) Mention what are the three modes in which Hadoop can be run?

The three modes in which Hadoop can be run are

- Pseudo distributed mode
- Standalone (local) mode
- Fully distributed mode

57) Mention what does the text input format do?

The text input format will create a line object that is an hexadecimal number. The value is considered as a whole line text while the key is considered as a line object. The mapper will receive the value as 'text' parameter while key as 'longwriteable' parameter.

58) Mention how many InputSplits is made by a Hadoop Framework?

Hadoop will make 5 splits

- 1 split for 64K files
- 2 split for 65mb files
- 2 splits for 127mb files

59) Mention what is distributed cache in Hadoop?

Distributed cache in Hadoop is a facility provided by MapReduce framework. At the time of execution of the job, it is used to cache file. The Framework copies the necessary files to the slave node before the execution of any task at that node.

60) Explain how does Hadoop Classpath plays a vital role in stopping or starting in Hadoop daemons?

Classpath will consist of a list of directories containing jar files to stop or start daemons.

Selenium	CCNA	AngularJS	NodeJS	PLSQL