Top 20 Hadoop & MapReduce Interview Question

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 ?

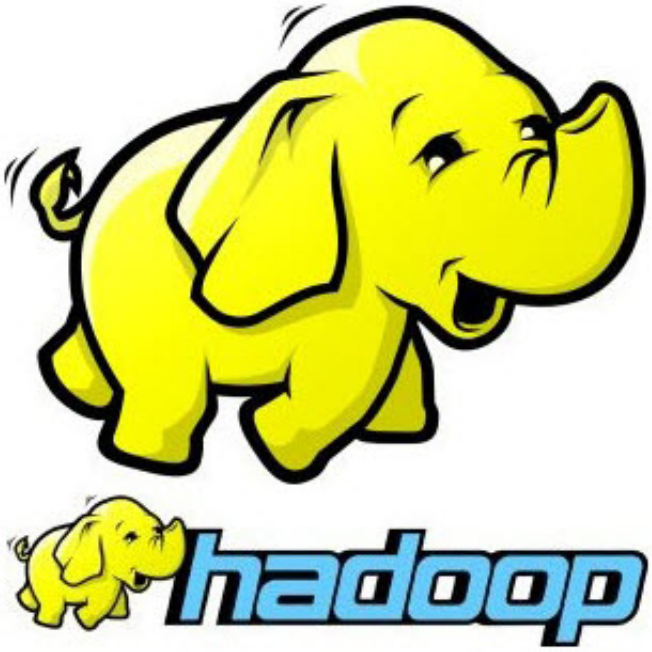
The process by which the system performs the sort and transfers the map outputs to the reducer as inputs is known as the shuffle

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

---

Accounting	Ethical Hacking	Cloud Computing	Jmeter
Manual Testing	QTP	Selenium	Test Management
Load Runner	Quality Center	Mobile Testing	Live Selenium Project
Enterprise Testing	Live Testing Project	Sap & All Modules