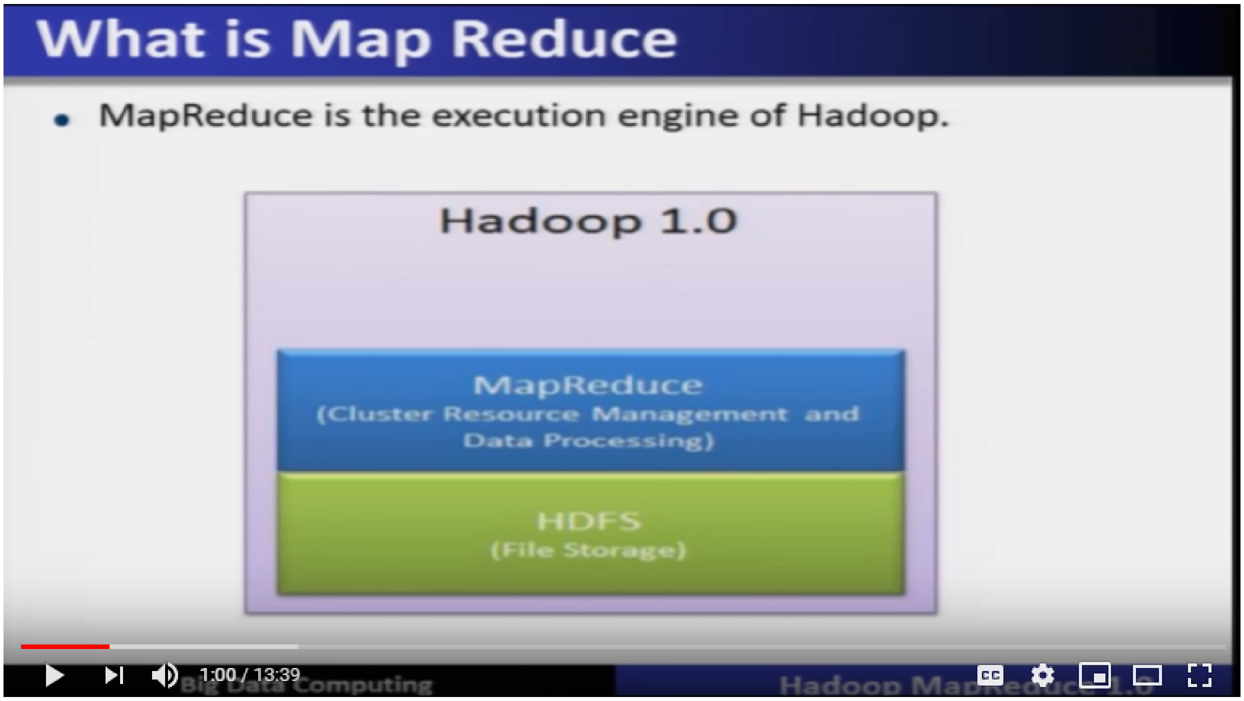
Lecture 05
Hadoop MapReduce 1.0

Hadoop MapReduce 1.0 version

Refer slide time :(0:18)

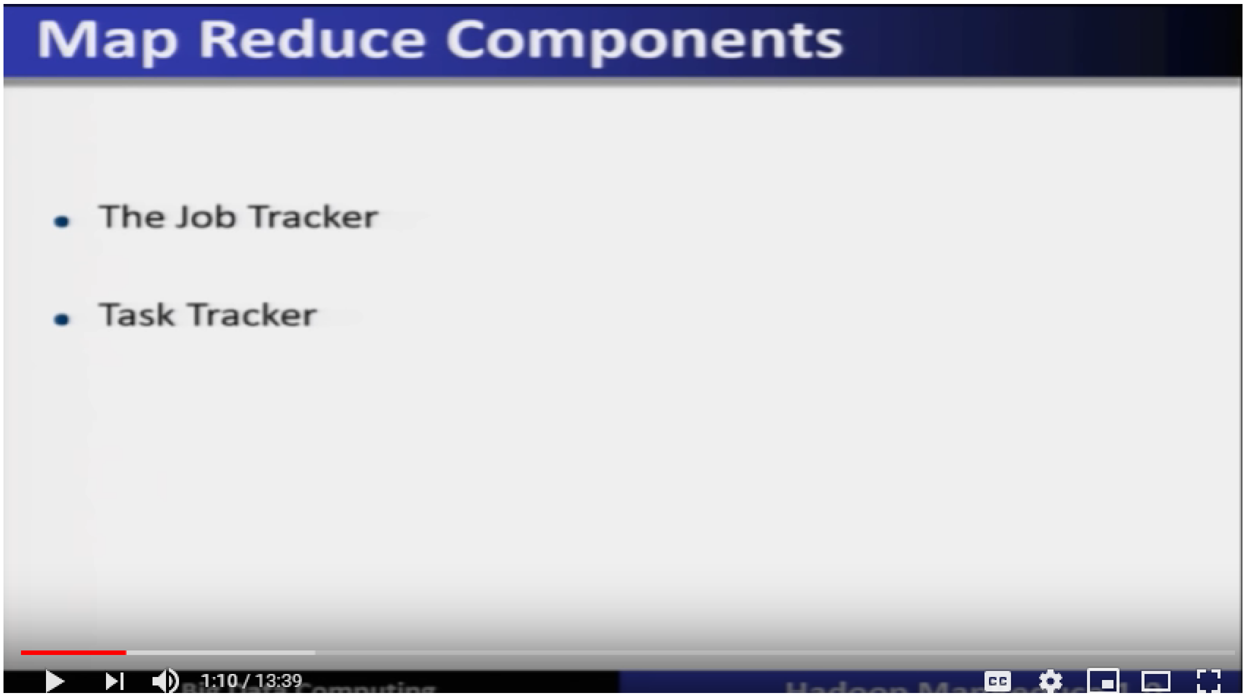


What is Map Reduce

- MapReduce is the execution engine of Hadoop.

So, MapReduce is an execution engine of Hadoop and we are going to briefly describe Hadoop 1.0, its components, that is Mapreduce which errands or HDFS. So, MapReduce is the programming paradigm, of the Hadoop system, for big data computing and it also performs, in MapReduce version 1.0 the resource management and the data processing, aspects. Also which runs over HDFS 1.0. So, we are going to cover this Hadoop or a MapReduce 1.0 version, in a brief.

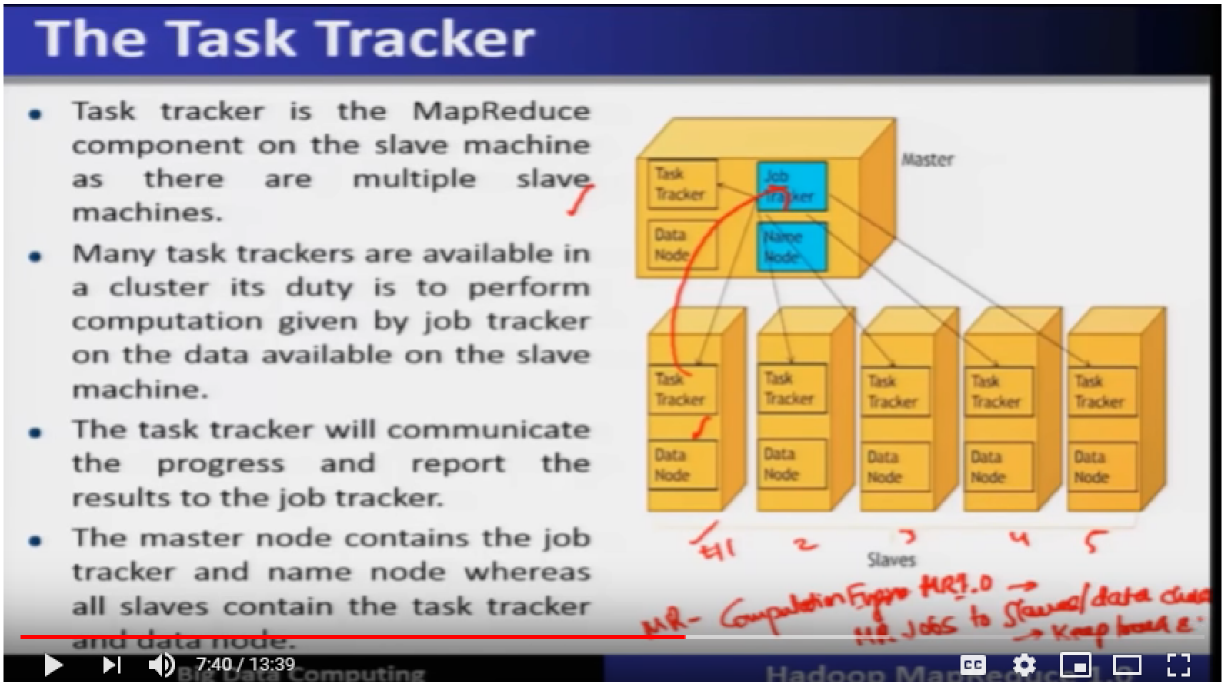
Refer slide time :(01:04)



Map Reduce Components

- The Job Tracker
- Task Tracker

So, MapReduce has two different major components, one is called the,' Job Tracker,' the other one is called the,' Trust Tracker'.
Refer slide time :(01:15)



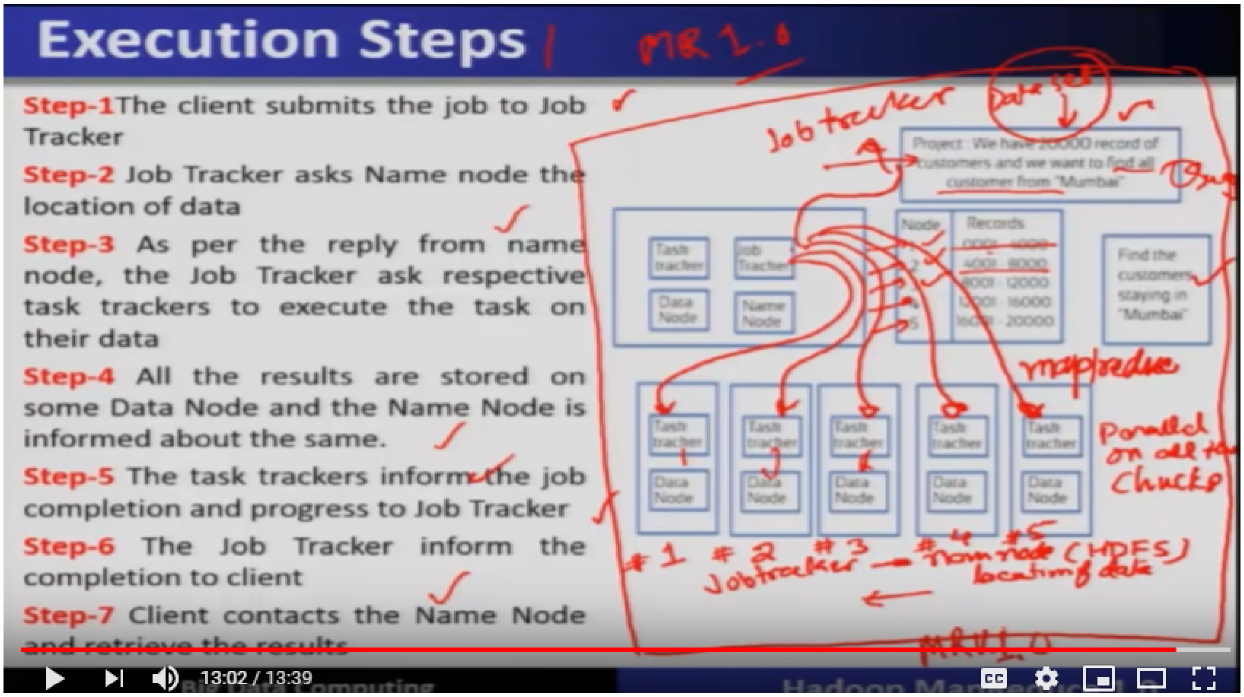
The Task Tracker

- Task tracker is the MapReduce component on the slave machine as there are multiple slave machines.
- Many task trackers are available in a cluster its duty is to perform computation given by job tracker on the data available on the slave machine.
- The task tracker will communicate the progress and report the results to the job tracker.
- The master node contains the job tracker and name node whereas all slaves contain the task tracker and data node.

And the scenario of the job tracker is, you can see here, in this particular diagram. So, this is the job tracker, which is a part of MapReduce version one. Now this particular job tracker runs, on the master or this is the master node. So, this master node and this name node, is a part of HDFS, HDFS 1.0 version. So, both of them may resides on the same machine or may not be in the same machine, but for the sake of simplicity we assume that the name node and the job tracker resides, on the same node which is called a,' Master', over here in this particular scenario, since it is having a master and several slaves. So, hence this is basically a client-server architecture, which is followed in MapReduce, version 1.0 and also in the HDFS version 1.0. So, in this particular diagram, we can see here, the job tracker resides on a particular node which is basically the, the master node. And on the same master node, let us assume that as DFS name node is there. So, we are not going to refer in this part of the discussion why because, we are focusing only on the MapReduce.

So, hence we are not going to discuss this name node part, of HDFS. Now another component, is called the,' Task Tracker', the task record may resides, on the same node also resides on other different slave nodes. So, the job tracker and task tracker they, they run in the form of a client-server model. So, job tracker is basically, running as a must is a server and the task records is run as its client. So, it's a client-server model let us understand, more into the functionality, of job tracker. So, job tracker as, I have already mentioned, is hosted inside the master and receives the job execution requests, from the client. So, the so, the client or the application, which basically is nothing but a MapReduce, program when it is submitted by the client, then the job tracker has to deal with that particular program execution. So, its main duty is to break, down the receive, its main duties are to break down, the received job, that is a that is the big data computation, specified in the form of MapReduce jobs. And this particular MapReduce, is divided into the smaller chunks and that is the small parts and these small parts that is called,' Chunks', are allocated with the map function and map and reduced function and this particular partial, contradictions are happening at this particular slave nodes with the help of the, task tracker. And this is the, the entire unit of execution, of this particular job. So, let us see the more detail of the task tracker. So, the task tracker is the MapReduce component, on the slave machine, as there are multiple slave machines, as we have shown here five of them. So, many task records are available, in the cluster its duties to perform the computations, which are assigned by the by that the job tracker, on the data which is available on the slave machines. So, the task tracker will communicate, the progress and report the result, back to the job tracker, the master node contains the job tracker and the name node whereas all the slave nodes contain the, the task tracker and data nodes. So, in this particular way, the job tracker keeps the track, of the map and reduce jobs, which are being allocated at different nodes and which are executing on, the data set which are assigned which are allocated, to these particular nodes, where the data is there the, the map and reduced function or the configuration will be performed. So, it's a computation engine. So, MapReduce is a computation engine, in version 1.0. so, not only it allocates, the MapReduce jobs to different slave nodes, where the data also resides, in the form of a chunks and it will then connect and so, basically not only it assigns but also it tracks, keep track ,of the progress and the resources, which is being allocated. Okay? Okay? Discovery hmm physically in order the execution steps.

Refer slide time :(08:04)



Execution Steps

Step-1 The client submits the job to Job Tracker

Step-2 Job Tracker asks Name node the location of data

Step-3 As per the reply from name node, the Job Tracker ask respective task trackers to execute the task on their data

Step-4 All the results are stored on some Data Node and the Name Node is informed about the same.

Step-5 The task trackers inform the job completion and progress to Job Tracker

Step-6 The Job Tracker inform the completion to client

Step-7 Client contacts the Name Node and retrieve the results

Project : We have 20000 record of customers and we want to find all customer from "Mumbai"

Node	Records
1	0001 - 4000
2	4001 - 8000
3	8001 - 12000
4	12001 - 16000
5	16001 - 20000

Find the customers staying in "Mumbai"

So, we are going to now trace all the execution steps, for the life cycle of a MapReduce job, till the application is submitted, by the client and to the to the MapReduce and it finishes and we are going to trace the, the execution cycle or the execution steps in the MapReduce, version 1.0. so, the first step is, the client submits the job, to the job tracker for example here, we have to find out we have to 20,000 records, of the customer and we want to find out all the customers from Mumbai. So, that means the query is basically to be executed, on these data set and this particular operation to find all the customer, this is the query, which is to be performed, using the MapReduce program which is being submitted. So, this particular request is being submitted to the job tracker and job tracker will ask the name note, about the location of this particular data set. So, the job tracker will consult the name node, which is a part of HDFS, to find out the location, of the data where it is being installed. So, now I say that this 20,000 records are divided like this, there are five different nodes, which stores all of them this is node number one two three four and five. So, the records first four thousand records stored are installed on this particular node number one and the next 4,000 is stored on the node number two and then next four thousand node number three and furthermore node number 4 and 5 respectively stores the remaining twenty thousand records.

Now this is called the chunks or the splits. The entire, two thousand twenty thousand record is splitted and stored on four different, five different nodes and this information will be given from, this name node back to the job tracker now the job tracker as per the reply by the name node the job tracker ask the respective task tracker to execute the tasks on their data. So, therefore, this particular job tracker, now assigns the, the, the map function or map and reduce function map function, on the stash tracker to execute, on that data chunk. So, with this particular direction the task tracker will perform this execution, at all places there in parallel. So, this particular execution of MapReduce program is done in parallel, on all the chunks. So, after the computation, of a MapReduce so, all the results are stored on the same data node, so whatever is the result it will be stored on the same data node and the name node is informed, about these particular results. So, the task tracker informs the job tracker, about the completion and the progress of this jobs assigned to those job tracker now the job tracker informs this particular completion, to that particular client and the client contacts the name node and retrieve the result back. So, after completing it this particular job tracker will inform, to the client about the completion of this particular job, that means now the result of the query is now ready, which the client will be able to access, with the help of name node and gets back the result. So, this is the entire execution engine, of which is, there in the form of, a MapReduce version 1.0, that we have briefly explained. Thank you.