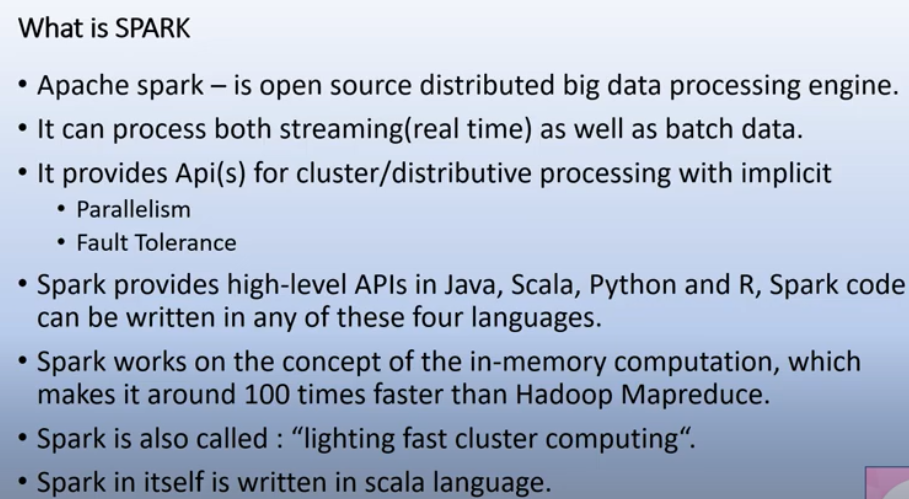
**What is Spark**



It is a distributed big data processing engine which means it follows the architecture of distributing programming where you can process one execution across no. of nodes in a cluster so it works on a cluster topology where you have no. of nodes , nodes are simple hardware commodity hardware machines which are connected with each other and can work in a cluster for execution of one single program.

Big data , in which data means in petabytes and terabytes and processing engine means we can do slicing, dicing and transformations on that particular data, we can extract or read data from diff files system and other source and load them in any target location so when we say that spark is a open source means we don’t need any license, you can download the libraries, jars and start working.

How spark diff from map reduce, one important factor is it support streaming data handling.

The API s provide for cluster for distributing processing is itself or implicitly is parallelism and fault tolerance so you don’t need to write externally any code to handle fault tolerance and parallelism.

Spark works on the concept of in memory computation, the entire processing of data across nodes happen in memory that means all your data is loading in primary memory that’s where the processing happen due to which we cut it down the input output overhead that is the reason why spark is 100 times faster that Hadoop.as you know Hadoop is very input output and disk oriented processing engine, so all the files, output is stored in the disk that’s why the input output overhead in Hadoop is more.

Spark itself written in scala language and that’s the reason why scala preferred more for data processing in spark.

