# Hadoop:

* It is framework of tools and this FW used to support applications which run on Big Data.
* Hadoop is an open source and is distributed under apache license.
* Big Data is creating challenges and Hadoop is addressing.
* Challenges come in three levels as velocity, volume, variety.
* In traditional Big Data handles by powerful computers.
* Hadoop takes its challenges and broken the Big Data into pieces and send into computation, then combined the result.
* Architecture -> Map reduce, File system(HDFS) and project(set of tools) component.
* Its works on Linux based tools.
* Two nodes as Task tracker and Data Node and combine known’s as slave.
* Job of TT is to process small piece of task to particular node
* Data node job is manage the piece of data
* All computer with TT & DN called as slaves
* Master has four component Job trackers, Name node, TT, DN.
* TT & JT fall under map reduce
* DN & NN fall under HDFS
* Application that runs on Hadoop that contacts the master node in queue by apache processing tools.
* JT runs on master node is to break big task into small piece and send to task tracker in slaves node and process it, send it back to JT.
* NN is responsible to keep the index which data resides on which DN.
* Hadoop always maintains three copies of processing file in different system(replication process)
* In any node dies in DN, JT ask the TT to perform the same operation that saved on system.
* Master dies, then it should be single point of failure
* Hadoop take care of data well. Tables that maintained by NN has index where the data is residing. All those tables are backed up in different systems
* In Enterprise Edition of Hadoop keep’s two Master, one as Master and other as back up Master
* Programmers would focus on writing scaling free programs.
* Scalability cost should be linear.
* If you want ur processing speed is high, then you need double the computer.
* Projects combine HIVE, HBASE, MAHOUT, PIG, OOZIE, FLUME, and SCOOP.
* In Hadoop was created by Doug Cutting and Michael Cafarella of Yahoo 2005.
* He named as Hadoop.
* The Name was came from Doug’s played with his elephant toy and named it as Hadoop.
* In 2006 the Hadoop project was donated to apache.
* Users of Hadoop in Enterprise level divided into two types 1) Administrators 2)Users
* Administrators are responsible for installation, Manage/Monitor system, and tune system. In short, he is overall responsible for software health.
* Any area deals with Big Data are using Hadoop.
* User is responsible for 1) Design Application 2) Import/Export data 3) Work with tools.
* Big Data Size 1) Terabyte=10^12 2) Petabyte= 10^15
* Big Data is unstructured.
* Traditional Hard Disk 1Tb read time =10000 sec(167 min)
* Solid state disk in Hadoop world, 1Tb read time =2000sec(33 min)
* In 2003 Google released the FILE called GFS(Google file system)
* FLUME-🡪Distribute service for collecting, aggregating, moving huge amount of log data.
* HIVE🡪 Data warehouse infra structure built on top of Hadoop for providing data summarization.
* HBASE🡪An open source, non relational distributed database, written in JAVA
* MAHOUT🡪 A distributed and scalable machine learning algorithms on Hadoop.
* OOZIE🡪 Java based application it runs on web server
* PIG-🡪similar to SQL programs.
* SCOOP🡪Tool designed for transfer bulk data between Hadoop and data stores as databases.

# Big Data

* Data which is beyond to storage capacity or processing power is called Big Data.
* It comes from Hospitality, social networks, Airlines, Sensors etc.
* Computation: The program which you are applying on your data to retrieve it.
* Data types
  + - * 1. Structured data -> RDMS data
        2. Semi structured data🡪 XML data, log file
        3. Un Structured data🡪Videos, Audios, Images etc.
* 3v’s( velocity, volume, variety) together called Big data and this definition is given by IBM.
* Hadoop is used for Storing and processing the huge data only.
* In year 2003 Google comes with their FILE system called GFS(Google file system)
* Followed by year 2004 they came out with Map reduce concept too.
* Then in 2005 yahoo people who named as Doug Cutting and Michael Cafarella developed Hadoop and later it donated to Apache.
* Hadoop name came from an elephant toy which is played by Dough cutting son.
* Hadoop is an open source Frame work under Apache license which used to processing large data with cluster’s of commodity hardware’s.
* Cluster means set of machines in single Lan and commodity hardware’s are local hardware’s.
* Hadoop has two core concepts.
  + - * 1. HDFS🡪 used to store the Big data with cluster of commodity hardwares with streaming access pattern.
        2. Streaming access pattern🡪 Write once, read any number of times, but don’t change the content of file once you keep the file into HDFS.
        3. File system🡪 Storing Files on directives.
        4. Mapreduce🡪 used to Process the data.
* HDFS Block size = 64mb (default) and you can make this to 128mb also.
* How HDFS works and it has five important Deamon’s.
* Deamon’s means some back-ground process that not visible.
* Five Deamon’s
  + - * 1. Name Node(NN)
        2. Secondary name Node
        3. Job Tracker
        4. Task Tracker
        5. Data Node
* Top three are Master Deamon’s (NN, Secondary NN, Job Tracker)
* Rest are slave nodes (TT, DN)
* Every top and bottom services talk with each other.
* Data about data is called **Meta data.**
* HDFS given **3 replications** by default.
* To overcome the data lost problem, we go for replication.
* Name Node is master and it has the corresponding slave node as Data Node.
* Data Node send its info to NN for every blog report and heart beat (3 sec). It say I have some Blog files and am alive and process properly.
* Blog report for am having this kind of data and heart beat for is alive.
* Name node has the index where data is stored in Data Node.
* Data Node is used to store your data.
* If DN fails to send its heart beat to NN, then NN think that DN is lost.
* Administrators take to lost DN and perform some action to rejoin in cluster with new storage.
* NN is lost is lost entire cluster will lost. Why NN has meta data where file store in DN. To overcome this NN h/w will be more reliable one.
* NN called as single point of failure.
* After storing your data, you need to process it, here JT come into scene.
* Process is given by client using map reduce and it assign to JT.
* JT doesn’t no where your file exists. JT send a request to NN, bcz NN has its Meta data of where file exist in DN.
* Now JT has the meta data and it work is assigning task to TT.
* Even your data exist in 3 nodes due to replication process. The request from JT not goes to all TT of that file present in DN.
* It goes to nearest TT of your in DN.
* Then TT assign the process to DN is called as **MAP**
* Request from client called as **Input File** and stores into no of files in DN called as **Input splits**.
* **No of input splits is equal to no of mappers**.
* **JT will arrange all these things(mappers).**
* If DN is lost, then TT will let to JT am not alive, then JT assign the same task to other TT of DN having requested file.
* TT is slave for JT in master node and send its info of every heart beat (3 secs) to JT, says am alive.
* If proper heart beat is not come from TT, JT will wait for 10heart beats (30 secs), until if not came JT will decide either TT works slow or it will dead.
* TT also sends available slots size to JT too.
* JT also single point failure, if it lost entire task will lost.
* After all process we need the outputs right, who will do?
* Reducers will do that. **No of output’s equal to no of reducers.**