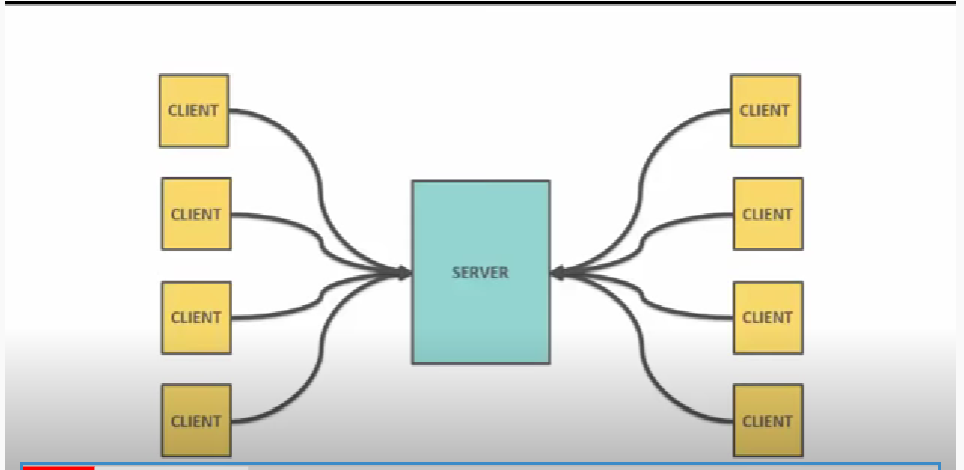
**History of data management and Hadoop**

[**https://www.youtube.com/watch?v=AZovvBgRLIY**](https://www.youtube.com/watch?v=AZovvBgRLIY)

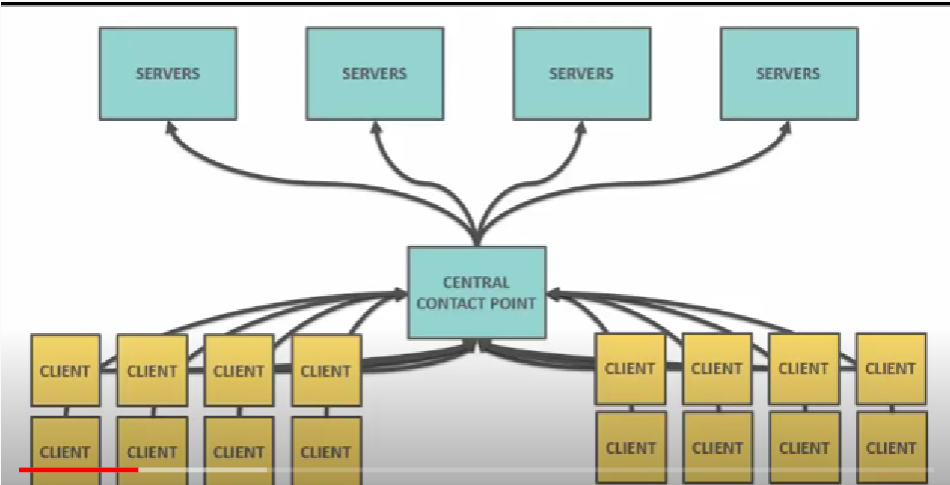
In most of the database storages – it works in client server model (it was exciting because many clients can connect to one server so that they can divide the work load among the client and server owns the data, client would come to the server and server provides data to them…

But If we loaded to much of data. servers will be heavily loaded and would not response to the client request normally and may hung up and cause problems of delay response… because there is limit for the backend server to deal with the data. so, at the end of the day this would leave you unhappy client and overloaded servers.



Solution to over come this is obviously big strong server box with lot of hardware, memory and resource but they are quit expensive and as we can keep on adding but down the line this is not the solution…

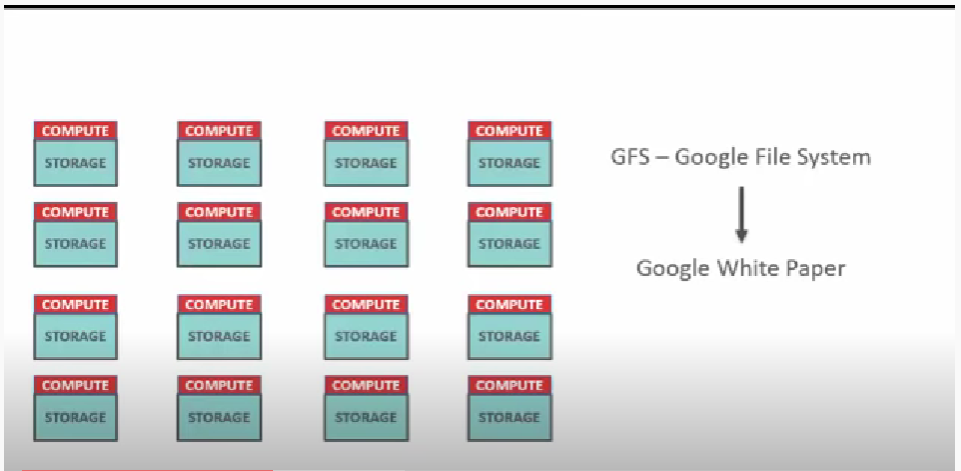
Another idea is instead of one big server machine why don’t we have multiple storage servers and distributed storage system and one central contact point which help to communicate which server the client want to talk to or communicate and use: this approach give you better scalability than just upgrading the machine but here also we still have a trouble if number of clients are more and also and cause our servers again to face heavily loaded and again requirements is to increase server and cause this distributed system crash….



So as the client number increase, they over whelmed the servers and cause the problems….

So to overcome this issue. google has developed their own system called GFS (google file system)

Here again no is no single data storage but data is distributed across the nodes with replica and each node stores the data and as well can compute …so as we increase the nodes both the storage and computing power will add up..



Utilizing this concepts doug and mike developed the Hadoop

Mike this guy refering to his son toys he given elephant as the symbol on Hadoop and doug moves on working on cloudera..

Hadoop core:

Hadoop as distributed storage system called hdfs(hadoop distributed file system): for storage

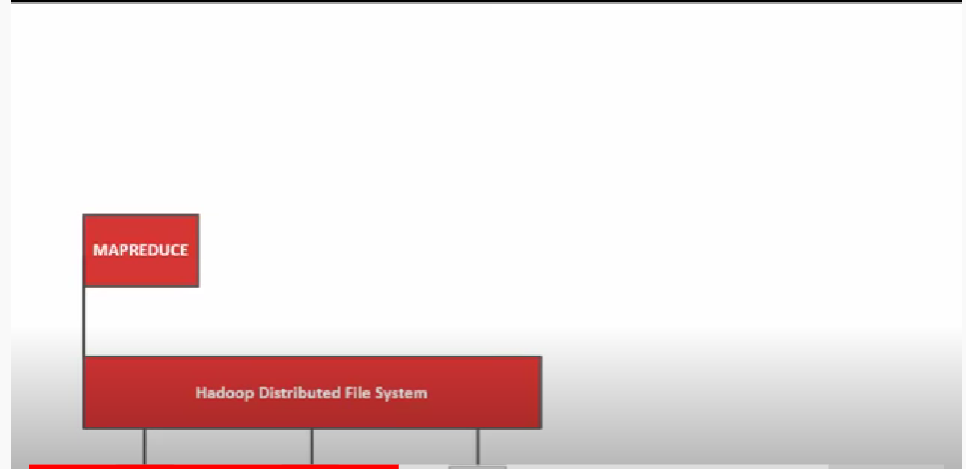
And the processing unit that pairs with hdfs is called : mapreduce

Hdfs file can store in any of the operating system.

Mapreduce is a combination of map procedure and reduce procedure

Map is for filtering and sorting the data.

Reduce is for summarizing operations.



Lets say we have some number of cards to count those cards it will take a while so instead we can share some set of cards to four different people to count this is called map ( telling each person to count is called map procedure) and these four people counting and handing back to you as a single count is called reduce procedure..

Here each person is a called node which has (cpu and storage disk) and these set of nodes are called cluster.

Now on these to get data in to Hadoop, read data from Hadoop storage , manage resources and schedule jobs etc now Hadoop has broader ecosystem…

**First component** to discuss is called : YARN ( yet another resource negotiator): this compute resource and clusters and schedule the users application….

**Second component**: sqoop : this is used for pulling RDBMS data in to hdfs

**Third component** : Flume : Now when it comes to web data , stream event data to pull this data in to hdfs we use flume.

**Fourth component**: Now to write code and access data and work with data we have many languages but to avoid big code we can achieve by writing simple single line queries and this can be done by hive. But here hive sql will convert in to mapreduce java code at the backend to communicate with the hdfs…

**Fifth component**: impala: this bypass MapReduce backend code and it is low latency sql engine which bypass the MR..

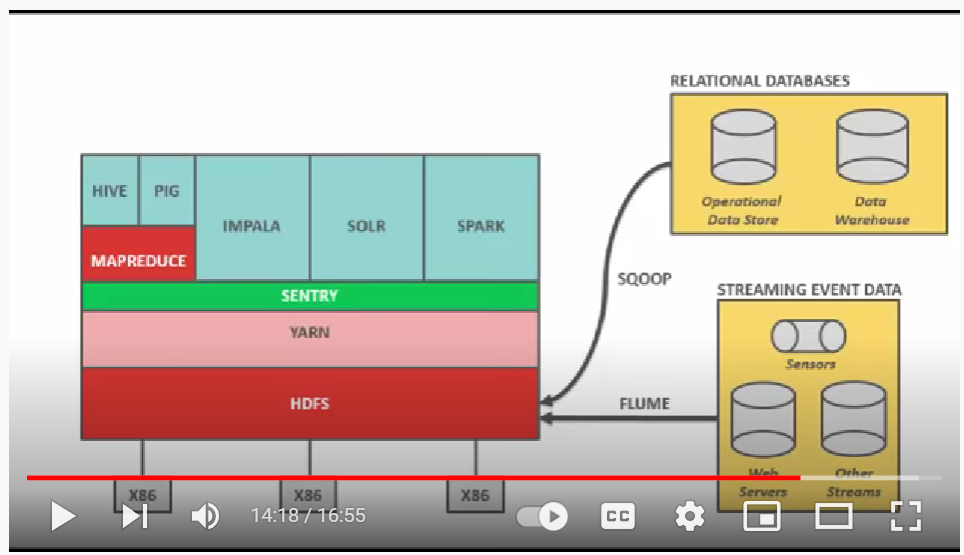
So query to return 10 rows from million count table hive takes 10-20 seconds where as Impala takes only milliseconds,,

**Sixth component**: search component: indexing on the hadoop data is by SOLR ( its an API)

It also index data in batches..

Now comes raising star: SPARK: part of ecosystem: in memory computation rather than writing in to disk…like MR..

Now for all these components to communicate with hdfs system utilizing yarn architecture for this there as to be some security and authentication right and this can be done by one more components called SENTRY( they call it has kerbose authentication)…



**Now lets find HOW cloudera fits in to this picture ?**

Now we know Hadoop it is open-source framework from Apache foundation…

But it is bit difficult to setup and maintain this open source. commercial vendors come in to picture with user friendly package tools with components for easy of working and with there distributed system called CDH and this by cloudera but this use 100% opensource Hadoop developed by Apache foundation…

Addition to packaging software, cloudera also provides Value add enterprise ready tool to drive and manages the data called cloudera manager

Cloudera manager manages entire hadoop infrastructure for you.: it’s a administrative tool for fast and easy and secure deployment, also helps in alerting, monitoring and management , configurations..

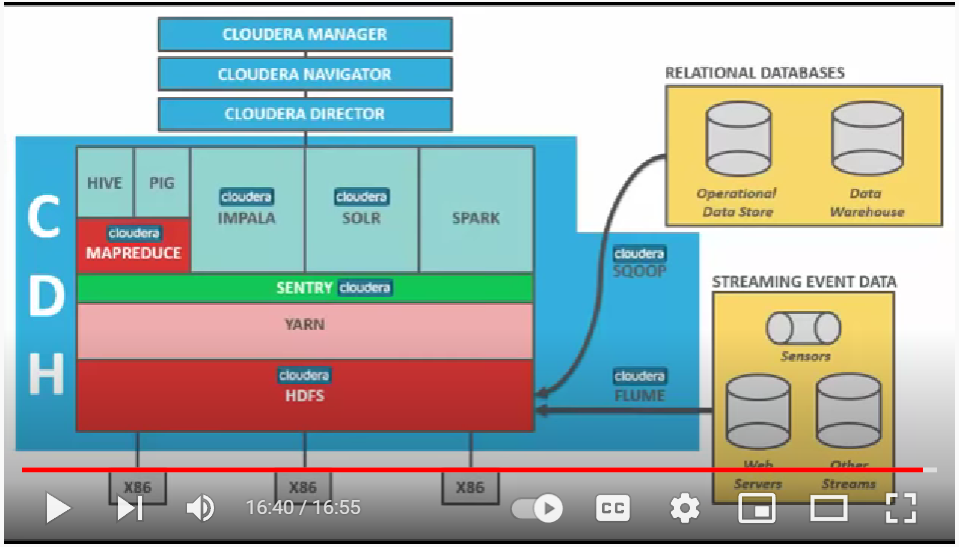
Cloudera also provides single logs in one place where you can see error and rectify them…

Cloudera navigator: provides data security

Cloudera director : provides to cloud

So cloudera manager( include navigator ) is a self service tool for deploying cluster hadoop component and mange’s them

So cloudera manager( include navigator and director) is a self service tool for deploying cluster hadoop in the cloud



Cloudera hive and impala runs on kubernities to pause , resume and up and down the services..

<https://www.youtube.com/watch?v=HloGuAzP_H8>

**Different distributions are**

Cloudera(CDH)

MAPR

Hortonworks

Actually core hadoop/base hadoop is apache hadoop with MAPR distributions

But other vendors like cloudera took the concepts of base hadoop and added the some features to UI security, monitoring and released there own distribution CDH and called them as cloudera hadoop.. apart from this they provide the support on there distributions

Now what maked cloudera is the leader in hadoop distributions..

1. First vendor to provide hadoop as package and leader in industry
2. First to provide sql for hadoop with its impala query engine..
3. It also includes open source distribution..

It provides support on whether you run on CDH on own server or hosted infrasture service like

Amazon vc2, rackspace, vcloud etc..

Because of all these benefits cloudera manger handles double the market share of other hadoop distributions..

Once you open cloudera VM machine, you will get view

Go to desktop

Launch cloudera express in order to launch cloudera manager…..

Now open terminal and see all the components

Cd /usr/lib/

Cd /opt/cloudera

Here we will see parcels: will discuss what is parcels

Now launch cloudera express

Go to terminal and enter command to start cloudera manager

Sudo service cloudera-scm-server start

practice all the components:

<https://www.edureka.co/blog/cloudera-hadoop-tutorial/>

cloudera vm machine:

https://www.youtube.com/watch?v=PLEt8FuDnjk

impala:

https://app.pluralsight.com/library/courses/hdfs-getting-started/table-of-contents

impala:

<https://www.youtube.com/watch?v=rdegSAMzXhY>

hive:

<https://app.pluralsight.com/library/courses/hive-relational-database-developers-getting-started/table-of-contents>

<https://www.youtube.com/watch?v=rr17cbPGWGA>

hbase:

<https://app.pluralsight.com/library/courses/hbase-hadoop-getting-started/table-of-contents>