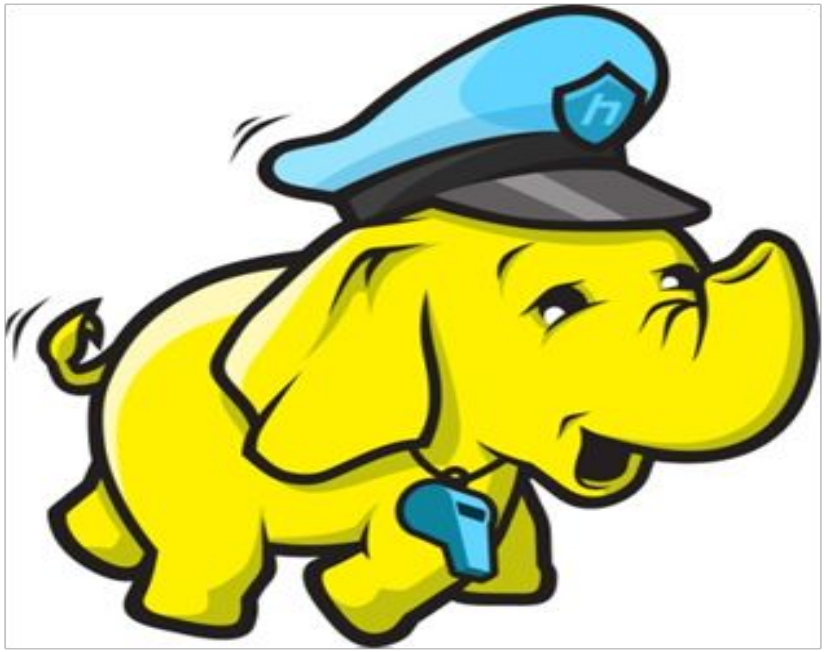
Problem with Traditional approach
Vs
Requirement of new approach
(Hadoop)



Agenda

Data, Data, Everywhere
Traditional large-scale computation
Store/Process data in TB/PB/ZB
Structured Vs Un-structured data
Processing Structured/Un-structured data
Partial failures of the system
Scalability
Economical factor

Data, Data, Everywhere

Social networking. Like Facebook/Linkedin/Twitter.

Online transactions/Online shopping.

Purchases at departmental stores.

Traditional Large-Scale Computation

Traditionally, computation was processor-bound.

Relatively small amounts of data (MB/GB).

For decades, primary push was to increase the computing power of a single machine.

Faster processor, more RAM.

But this can increase performance till some extent.

Store/Process data in TB/PB/ZB

Traditional system can’t store data in TB/PB/ZB.

It’s not possible to process large amount of data on a single machine.

Even good processor can’t process billion of rows.

On the other hand, Hadoop can process large amount of data very easily.

History data is analysed for business intelligence.

Structured data

Emp Id	F Name	L Name	Department	Contact No.
123	Srini	Reddy	Accounts	9876789876
124	Venkat	Eluri	Networking	8876896545

Un-structured data

INFO : Server started successfully at 2:14:52 pm
WARN : Not enough memory
FATAL Error : Server crashed unexpectedly

Process Structured/Un-structured data

RDBMS needs a schema but Hadoop doesn't.

Data must be structured in traditional approach.

You must know the structure of the table before inserting the data.

Hadoop can process structured/un-structured data easily.

Hadoop can process text files directly, without loading them into database.

Partial failures of the system

There is no partial failure in traditional approach.

Either the query is 100% completed or 100% failed.

If some processing is failed, you need to start it all over again.

In Hadoop, failure of 1 machine doesn't let the complete job failed.

If a job is partially failed, it is rescheduled on alternate machine automatically.

Scalability

Traditional systems are not scalable.

In traditional approach, adding load to the system decreases the performance.

On the other hand, Hadoop is scalable.

You can add/remove machines from Hadoop cluster at runtime.

You can increase the cluster size if data size increases.

Economical Factor

Traditional approach is more expensive.

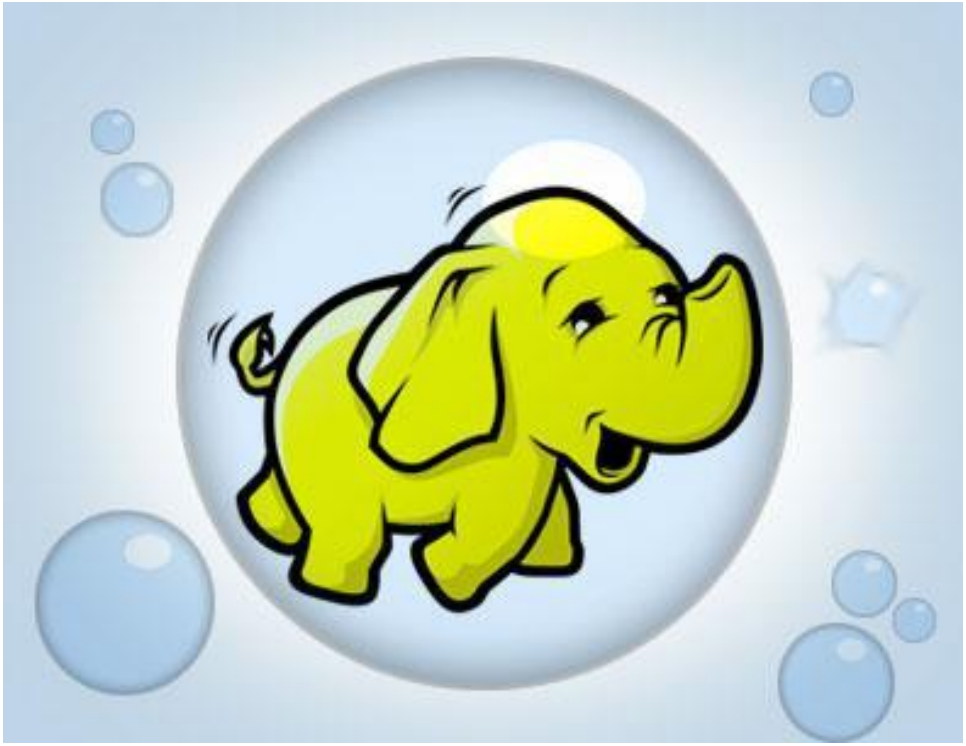
Instead of commodity hardware, you need high performance servers.

You need to pay high for RDBMS Softwares.

On the other hand, Hadoop is open source/free software.

You can use Hadoop for personal or enterprise use free of cost.

…Thanks…



For online Hadoop training, send mail to neeraj.ymca.2k6@gmail.com