What is HADOOP ?

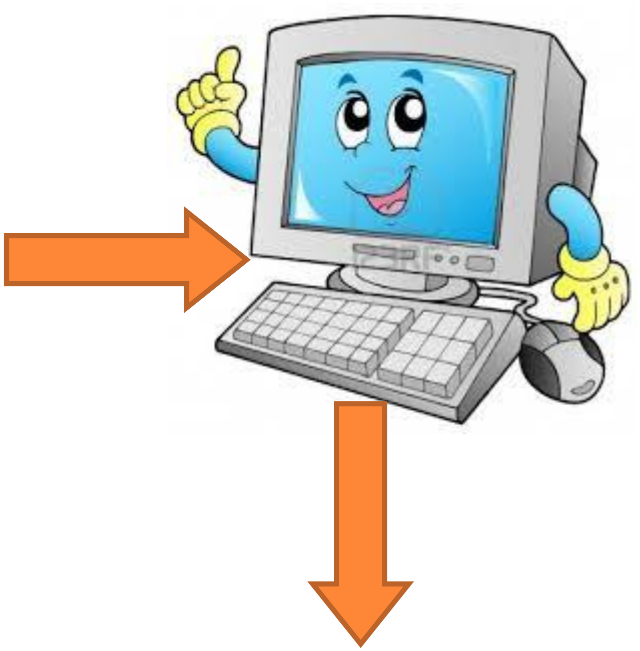
Agenda

- Introduction to Distributed Computing
- An overview of Super Computing and its challenges
- Hadoop as a solution
- History of Hadoop

One man = One day

Adding the Numbers

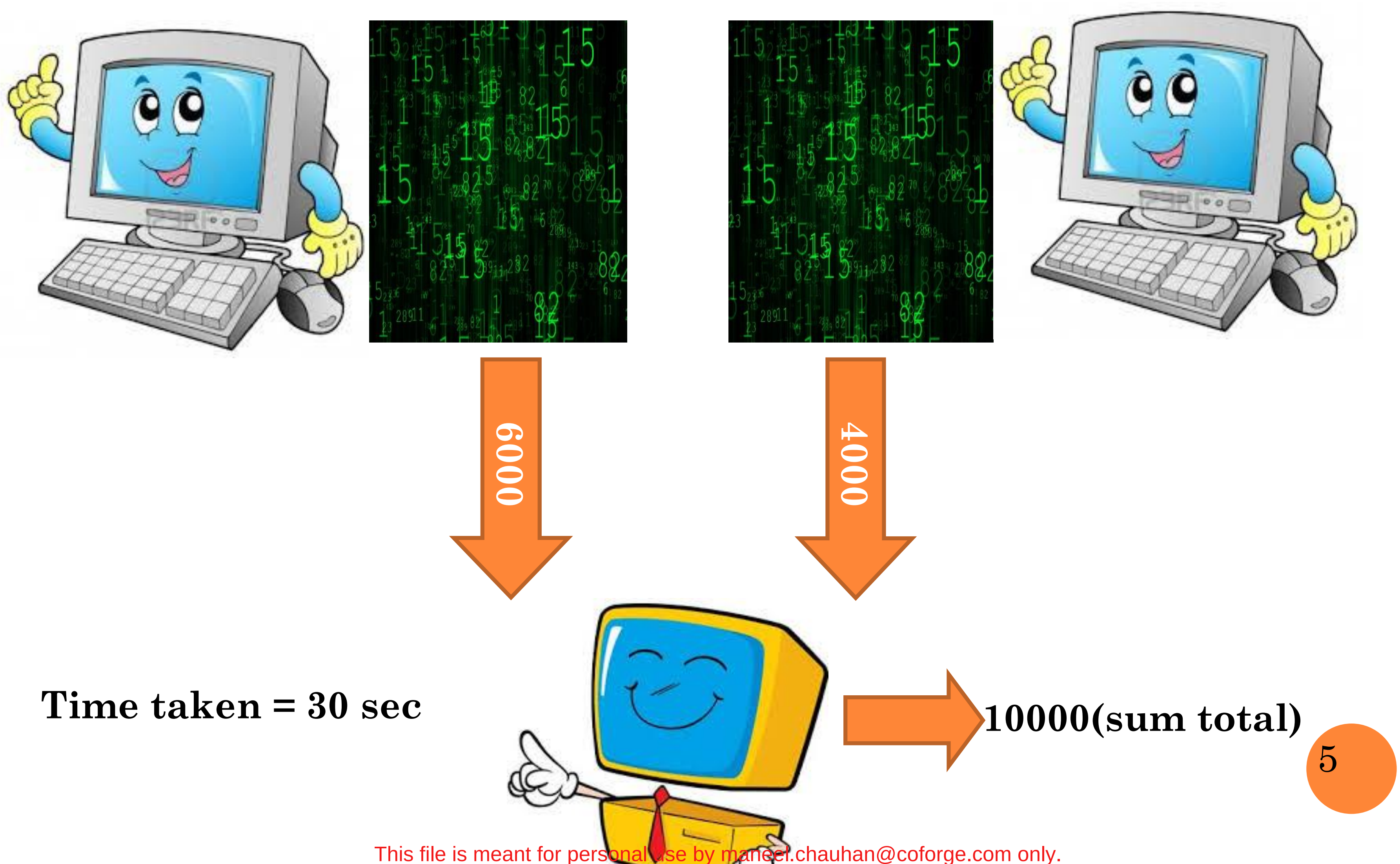
Input file (1 GB)



Sum total = 10000

Time taken = 50 seconds

Parallel Processing – Faster computing !



6000

4000

Time taken = 30 sec

10000(sum total)

SUPER COMPUTER – cluster of computers

IBM

FUJITSU

CRAY
THE SUPERCOMPUTER COMPANY

intel

https://en.wikipedia.org/wiki/History_of_supercomputing

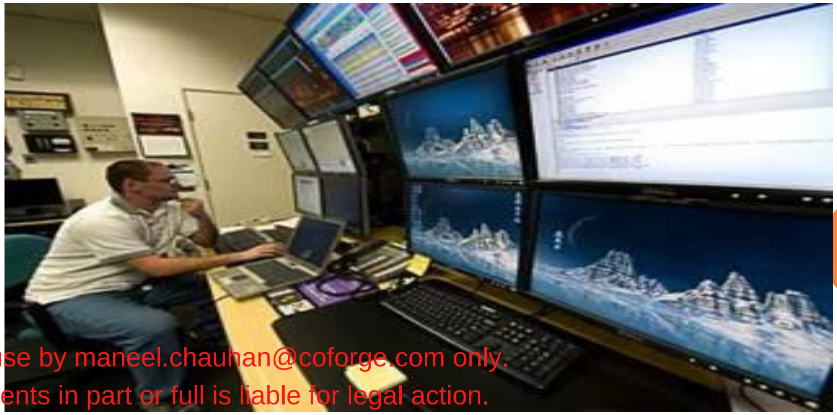
Uses of Super Computers

Used in weather forecasting, seismic data analysis etc.

University research labs for studies related to computational fluid dynamics, bioinformatics etc.

Department of defense for nuclear research.

WED	THU	FRI	SAT	SUN	MON	TUE
52	62	62	58	55	52	52
40	48	52	52	45	40	40



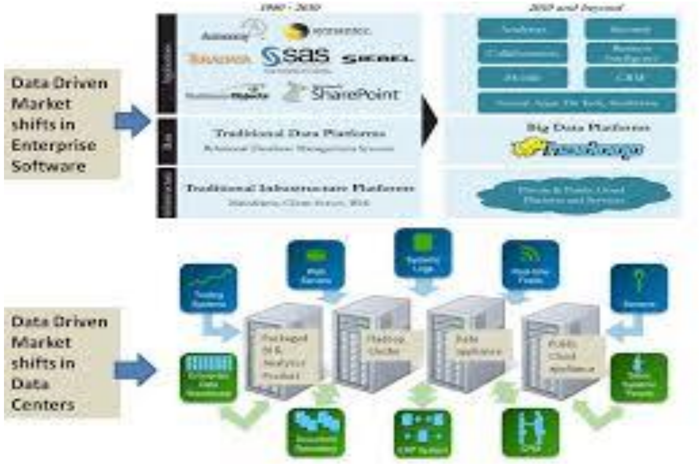
Challenges With Super Computing

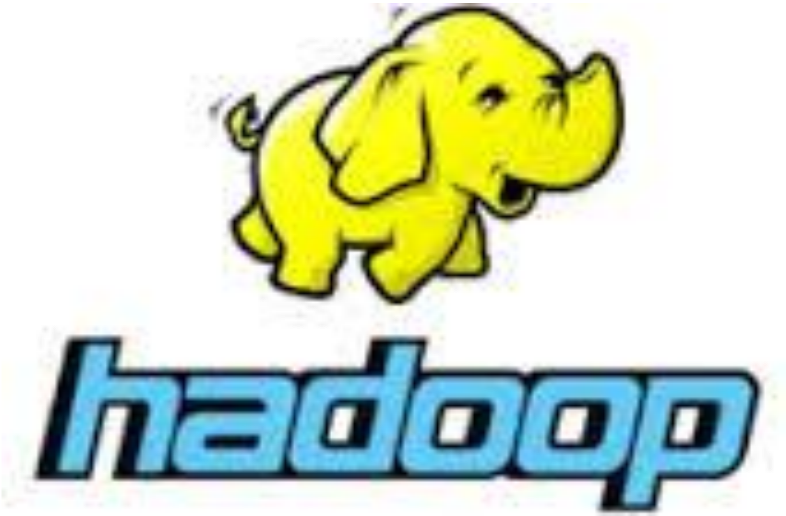
- A general purpose operating system like framework for parallel computing needs did not exist
- Companies procuring super computers were locked to specific vendors for hardware support
- High initial cost of the hardware
- Develop custom software for individual use cases
- High cost of software maintenance and upgrades which had to be taken care in house the organizations using a super computer.
- Not simple to scale horizontally

HADOOP Comes to the Rescue

- A general purpose operating system like framework for parallel comuting needs
- Its free software (open source) with free upgrades
- Has options for upgrading the software and its free !
- Opens up the power of distributed computing to a wider set of audience.
- Mid sized organizations need not be locked to specific vendors for hardware support – Hadoop works on commoditity hardware
- The software challenges of the orgnazation having to write proprietary softwares is no longer the case.

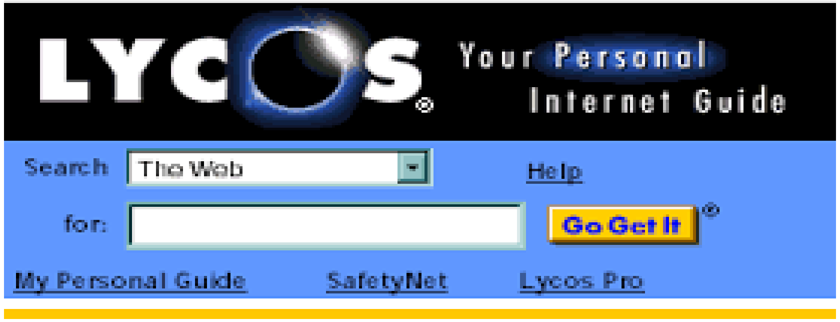
The BUZZZZ…WORD is HADOOOPPPP


1980 - 2010
2010 and beyond
Data Driven Market shifts in Enterprise Software
SAS
SIEBEL
SharePoint
Traditional Data Platforms
Big Data Platforms
hadoop
Traditional Infrastructure Platform
Data Driven Market shifts in Data Centers
Hadoop Cluster



Late 90's to Early 2000's – dot com Bubble !

- This was the age of the INTERNET BOOM
- The need of the hour was a scalable web search engine
- Major players in this business of internet search engines back then were



Evolution of Hadoop From Internet Search Engines

- The need of the hour was scalable search engine for the growing internet
- Internet Archive search director Doug Cutting and University of Washington graduate student Mike Cafarella set out to build a search engine and the project named NUTCH in the year 2001-2002
- Google's distributed file system paper came out in 2003 & first file map-reduce paper came out in 2004
- In 2006 Dough Cutting joined YAHOO and created an open source framework called HADOOP (name of his son's toy elephant) HADOOP traces back its root to NUTCH, Google's distributed file system and map-reduce processing engine.
- It went to become a full fledged Apache project and a stable version of Hadoop was used in Yahoo in the year 2008

Summary

Overview of parallel processing and its challenges

Hadoop as a rescue tool for those challenges

What is hadoop and its evolution?