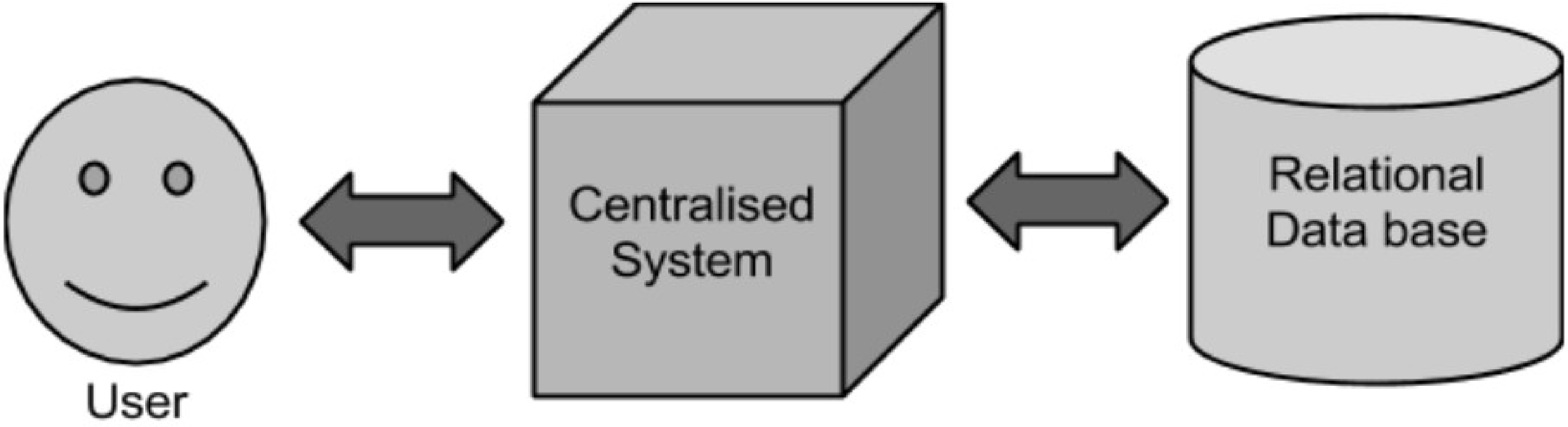
Big Data Solutions

Tushar B. Kute,
http://tusharkute.com

Traditional Enterprise Approach



Traditional Enterprise Approach

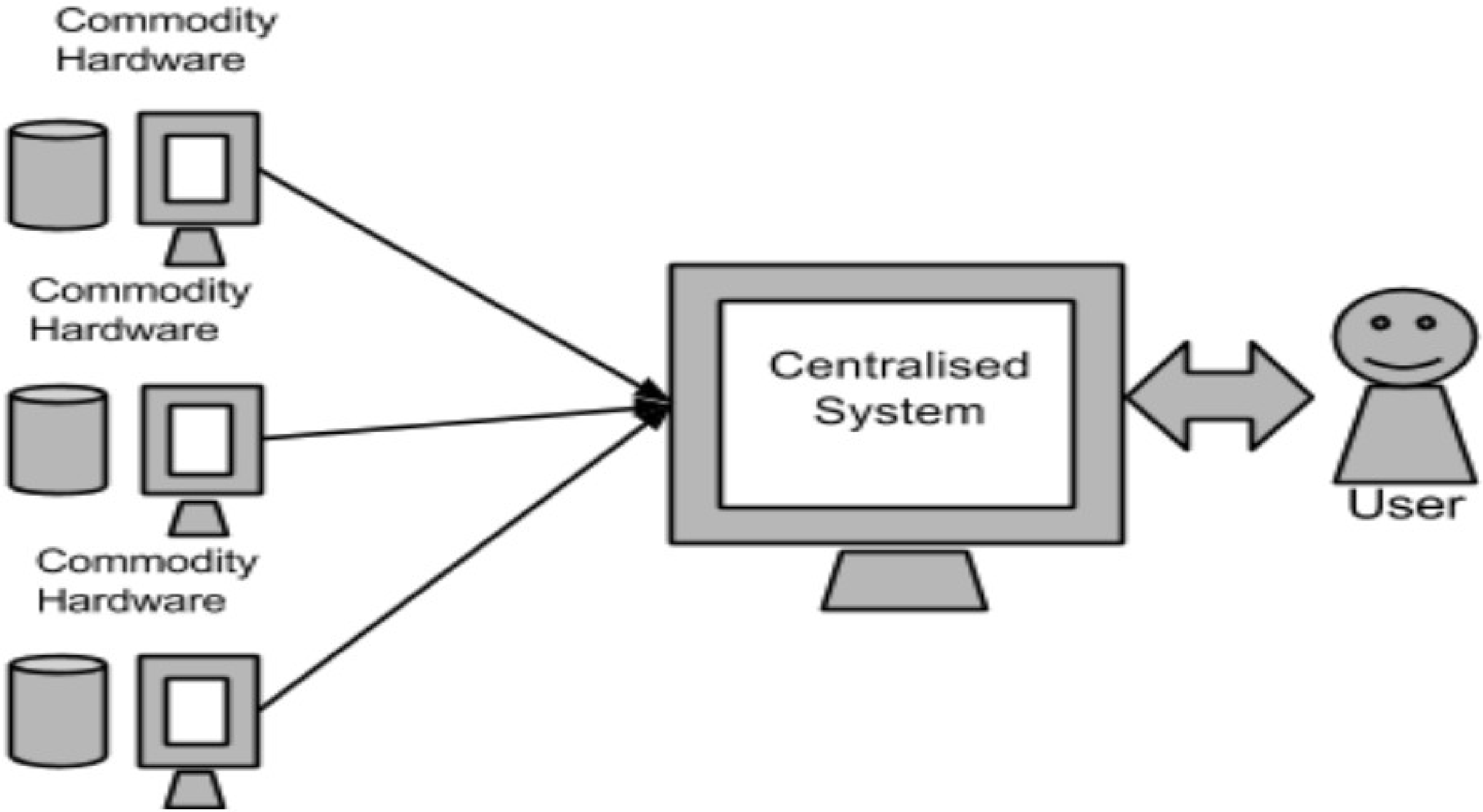
Limitation

- This approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data.
- But when it comes to dealing with huge amounts of scalable data, it is a hectic task to process such data through a single database bottleneck.

Google's Solution

- Google solved this problem using an algorithm called MapReduce.
- This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.

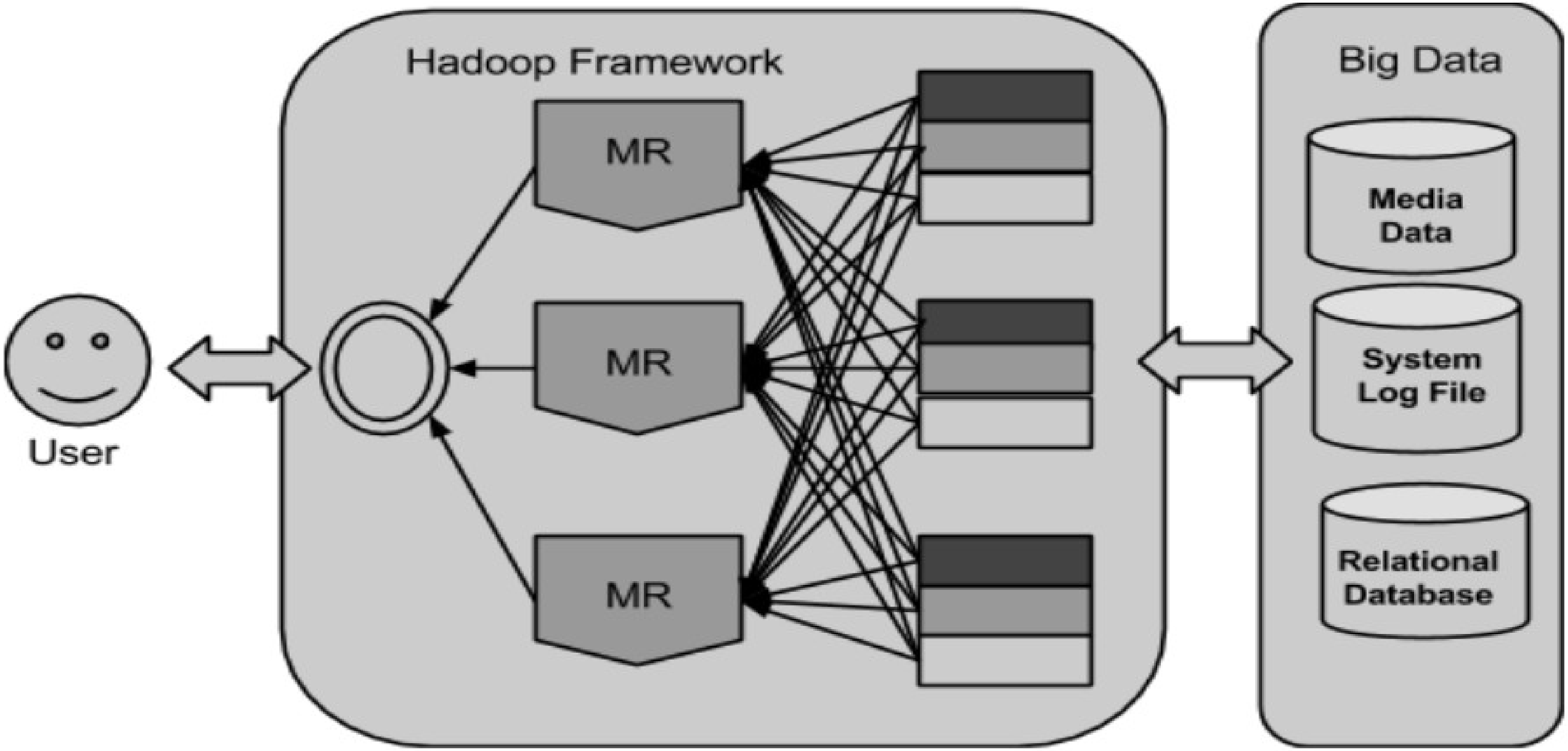
Google's Solution – MapReduce



Hadoop

- Using the solution provided by Google, **Doug Cutting** and his team developed an Open Source Project called HADOOP.
- Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others.
- In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

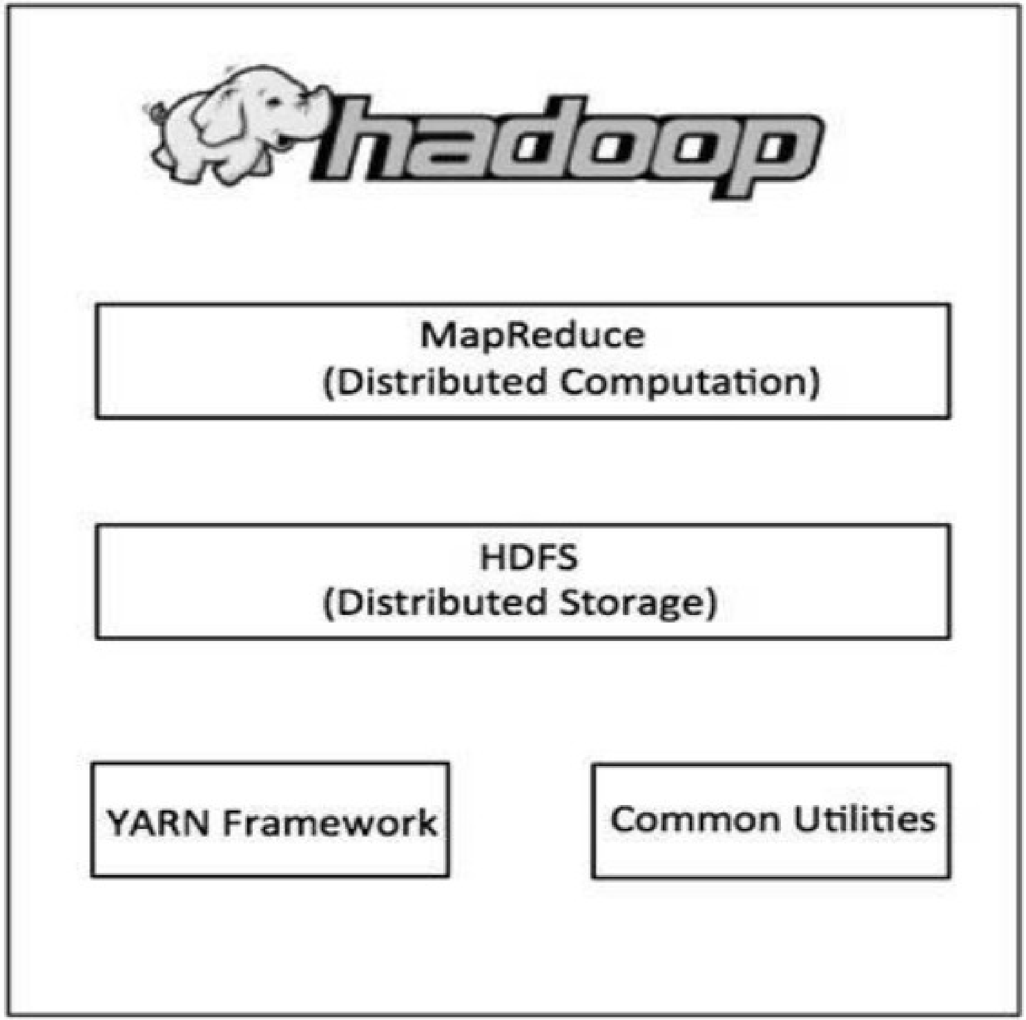
Hadoop


Hadoop Framework
MR
MR
MR
User
Big Data
Media Data
System Log File
Relational Database

What is Hadoop ?

- Hadoop is an Apache open source framework written in Java that allows distributed processing of large datasets across clusters of computers using simple programming models.
- The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers.
- Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture



What is MapReduce?

- MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.
- The MapReduce program runs on Hadoop which is an Apache open-source framework.

Hadoop Distributed File System

- The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware.
- It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant.
- It is highly fault-tolerant and is designed to be deployed on low-cost hardware.
- It provides high throughput access to application data and is suitable for applications having large datasets.

Hadoop Distributed File System

- Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules:
  - Hadoop Common: These are Java libraries and utilities required by other Hadoop modules.
  - Hadoop YARN: This is a framework for job scheduling and cluster resource management.

What is Hadoop ?

- Hadoop is an Apache open source framework written in Java that allows distributed processing of large datasets across clusters of computers using simple programming models.
- The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers.
- Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

How does Hadoop work?

- It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput.
- Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

How does Hadoop work?

- Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs:
  - Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).
  - These files are then distributed across various cluster nodes for further processing.
  - HDFS, being on top of the local file system, supervises the processing.
  - Blocks are replicated for handling hardware failure.
  - Checking that the code was executed successfully.
  - Performing the sort that takes place between the map and reduce stages.
  - Sending the sorted data to a certain computer.
  - Writing the debugging logs for each job.

Advantages of Hadoop

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

Thank you

Web Resources
http://mitu.co.in
http://tusharkute.com

Blogs
http://digitallocha.blogspot.in
http://kyamputar.blogspot.in

tushar@tusharkute.com