IMPLEMENTATION OF
CLOUD TECHNOLOGY

MODULE 12

Cloud technologies

- Web Services

- Virtualization

- Multi-tenancy – multiple clients can be served

Web Services

- **Web services** are self-contained and self-describing. **Web services** can be discovered using UDDI. **Web services** can be used by other applications. HTTP and XML are the basis for **Web services**.

A complete web service is, therefore, any service that:

- Is available over the Internet or private (intranet) networks

- Uses a standardized XML messaging system

- Is not tied to any one operating system or programming language

- Is self-describing via a common XML grammar

- Is discoverable via a simple find mechanism

VIRTUALIZATION

VIRTUALIZATION

- It is creation of virtual version of something, such as a hardware platform, operating system, a storage device or network resources.

- Why virtualization??
    - Consolidates workload of several under-utilized servers.
    - Virtual machines can be used to provide secure, isolated sandboxes for running untrusted applications.
    - Virtual machines can be used to create operating systems or execution environments with resource limits.

- Virtual machines can provide the illusion of hardware, or hardware configuration.
- Virtual machines allows for powerful debugging and performance monitoring.
- Virtual machines are great tools for research and experiments, since they provide isolation and are safer to work with.
- Virtual machines can be used to create arbitrary test scenarios and can lead to some very imaginative, effective quality assurance.

TYPES OF VIRTUALIZATION

- Hardware Virtualization – refers to the creation of a virtual machine that acts like a real computer with an operating system.

- For example, a computer that is running MS Windows may host a VM that looks like a computer with UBUNTU Linux OS.

- The software or firmware that creates a virtual machine on the host hardware is called a hypervisor or Virtual Machine Monitor.

- Different types of hardware virtualization includes:
  - Full Virtualization: Almost complete simulation of the actual hardware.
  - Partial Virtualization: some but not all the target environment is stimulated.
  - Paravirtualization: hardware environment is not simulated. Guests programs need to be modified to run in this environment.

- Desktop Virtualization – concept of separating logical desktop from the physical machine.
    - For users, it means that they can access their desktop from any location without being tied to a single device.
- Software Virtualization – hosting of several applications in an environment.
- Memory Virtualization – aggregating RAM resources from networked systems into a single memory pool.

Virtual Machine

- A virtual machine (VM) is a software implementation of a machine that executes programs like a physical machine.

- Two major categories:
  - System Virtual Machine – Multiple OS environments can co-exist on the same computer, in isolation from each other.
  - Process Virtual Machine – a process VM runs as a normal application inside the host OS and supports a single process. It is created when process starts and destroyed when process exits.
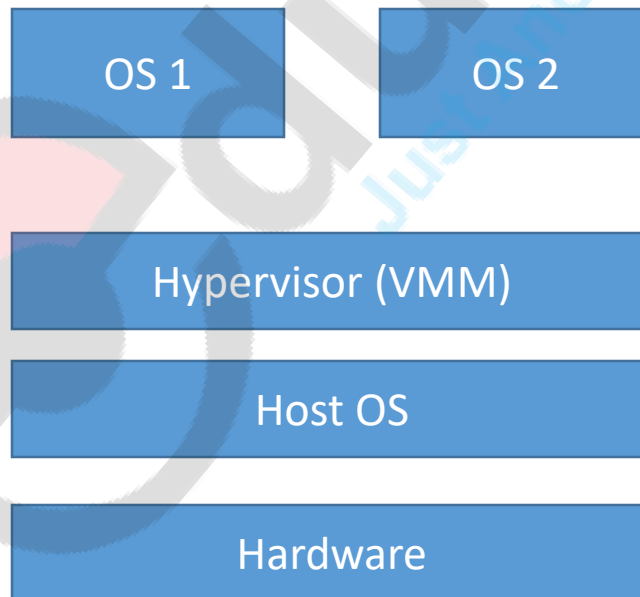
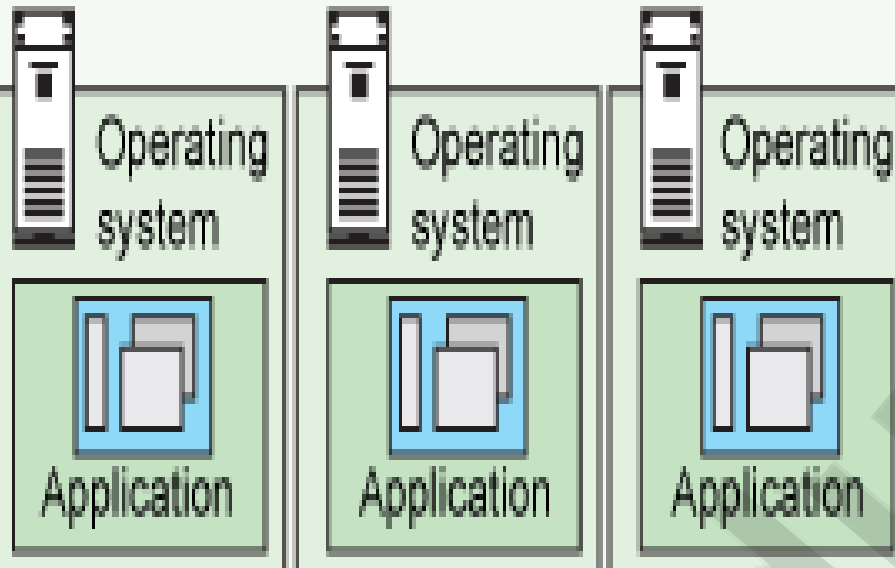# MARKET LEADERS

- Vmware
- Citrix
- Oracle
- Red Hat

Hypervisors

- Also called as Virtual Machine Manager (VMM).
- Hypervisors only task is to run and manage the execution of guest operating system.

OS 1	OS 2

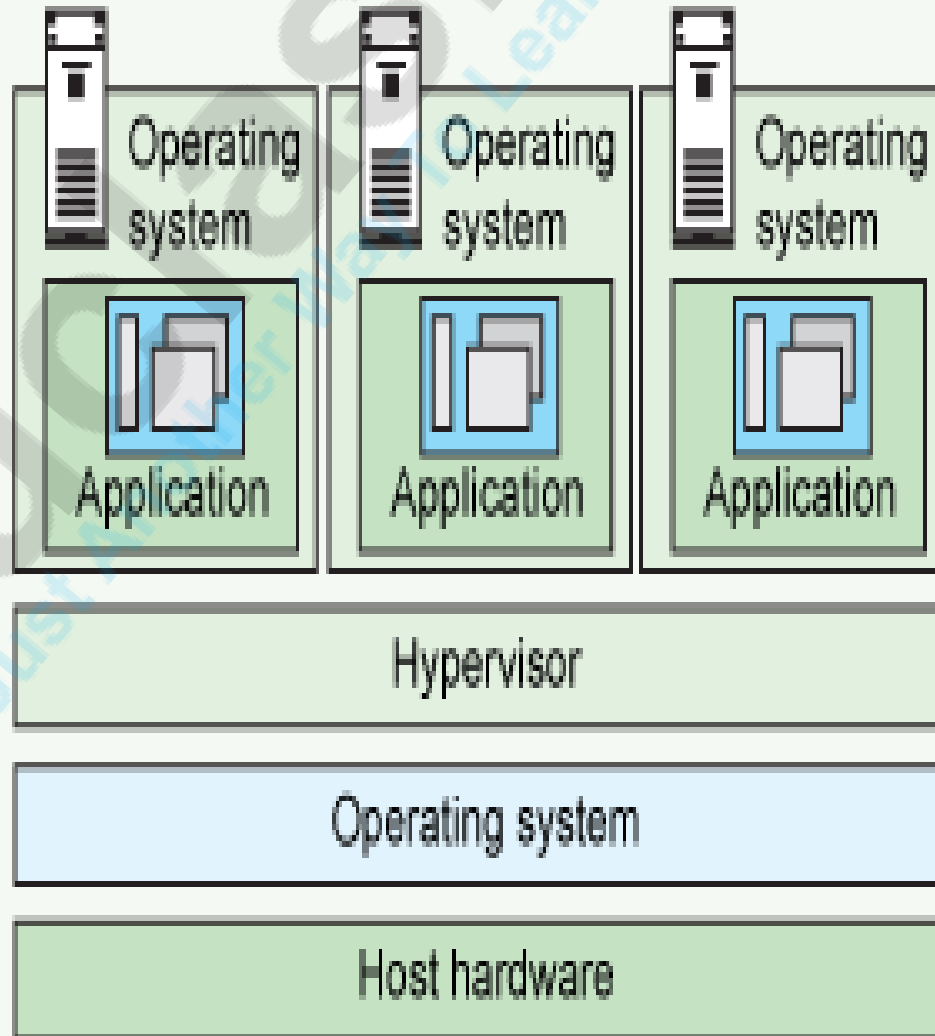| Hypervisor (VMM) |
|------------------|
| Host OS |
| Hardware |

- A **hypervisor** or virtual machine monitor (VMM) is a piece of computer software, firmware or hardware that creates and runs virtual machines. A computer on which a **hypervisor** is running one or more virtual machines is defined as a host machine. Each virtual machine is called a guest machine.

Type 1 Hypervisor

Operating system — Application

Operating system — Application

Operating system — Application

Hypervisor

Host hardware

Type 2 Hypervisor

Operating system — Application

Operating system — Application

Operating system — Application

Hypervisor

Operating system

Host hardware

CLASSIFICATION

- Type 1 (or native, bare metal) – runs directly on the host's hardware
- Type 2 (or hosted) – runs within a conventional operating system environment.

Current players in hypervisor market

- Vmware
- Xen
- Microsoft Virtual Server

AJAX

- It is asynchronous Javascript and XML.

- AJAX is a group of interrelated web development techniques used on the client-side to create asynchronous web applications.

- With Ajax, Web applications can send data to, and retrieve data from, a server asynchronously without interfering with the display and behavior of the existing page.

- Data is stored at receiver's side and is retrieved using object.

Some uses of AJAX

- Real-time form data validation

- Autocompletion

- Partial Submit
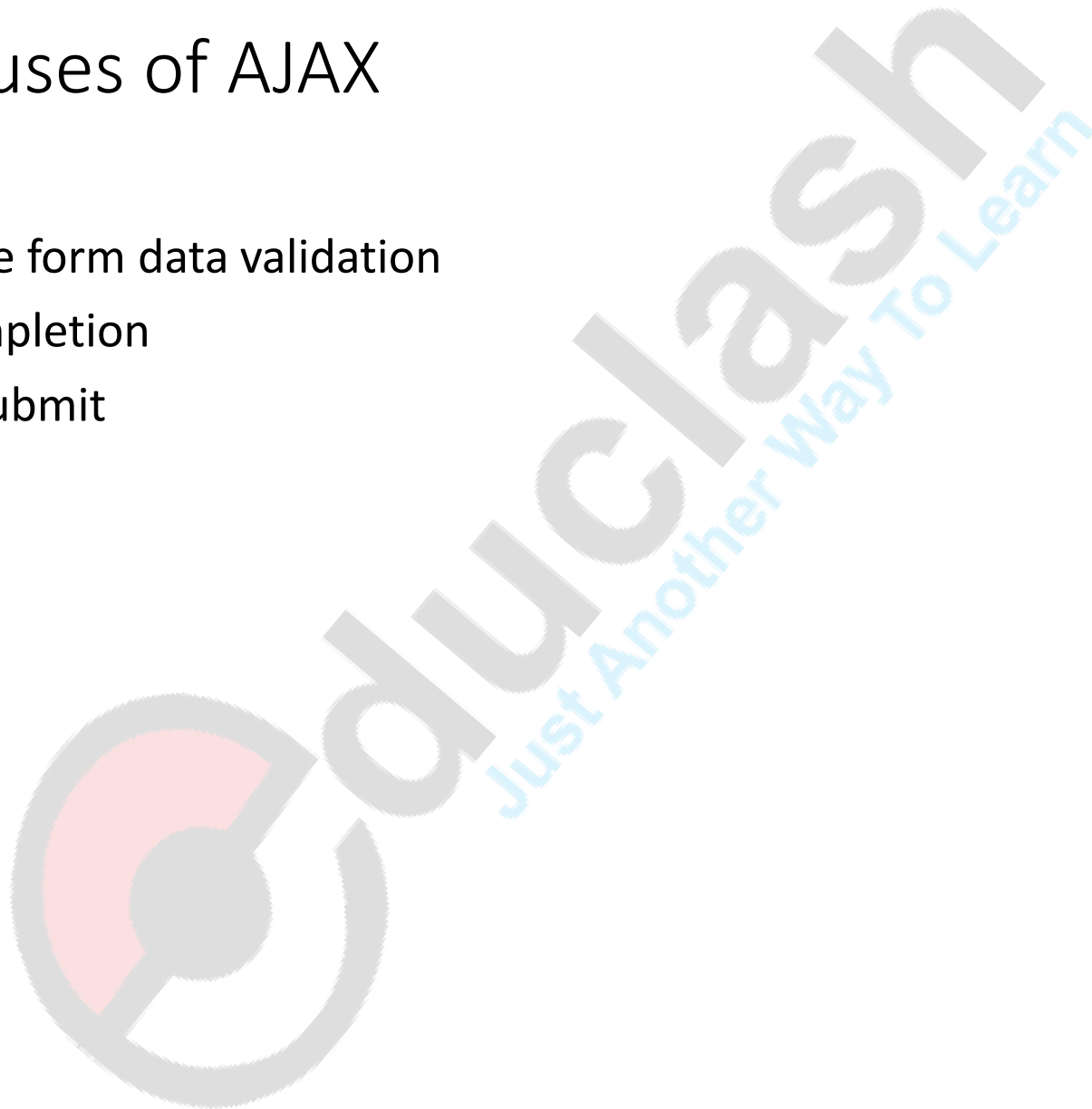
MASHUPS

- It is a Web page or application that uses and combines data, presentation or functionality from two or more sources to create new services.

- Types of Mashups:
  - Business (or enterprise) Mashups –
    - Defines applications that combine their own resources, applications and data with other external web services.
    - Allows for collaboration among business and developers.
  - Consumer Mashups – combines different data from multiple public sources in the browser and organizes it using simple browser interface.
  - Data Mashups – combines similar types of media and information from multiple sources into a single representation. The combination creates a new and distinct web service .

HADOOP
AND
MapReduce

Large-Scale Data Analytics

- MapReduce computing paradigm (E.g., Hadoop) vs. Traditional database systems


vs.

Database
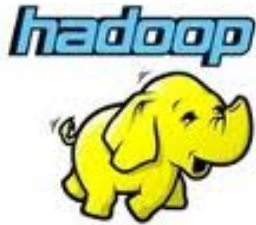
- **Many enterprises are turning to Hadoop**

  - Especially applications generating *big data*

  - Web applications, social networks, scientific applications

What is Hadoop

- Hadoop is a software framework for *distributed processing* of *large datasets* across *large clusters* of computers
  - ***Large datasets*** → Terabytes or petabytes of data
  - ***Large clusters*** → hundreds or thousands of nodes
- Hadoop is open-source implementation for Google ***MapReduce***
- Hadoop is based on a simple programming model called *MapReduce*
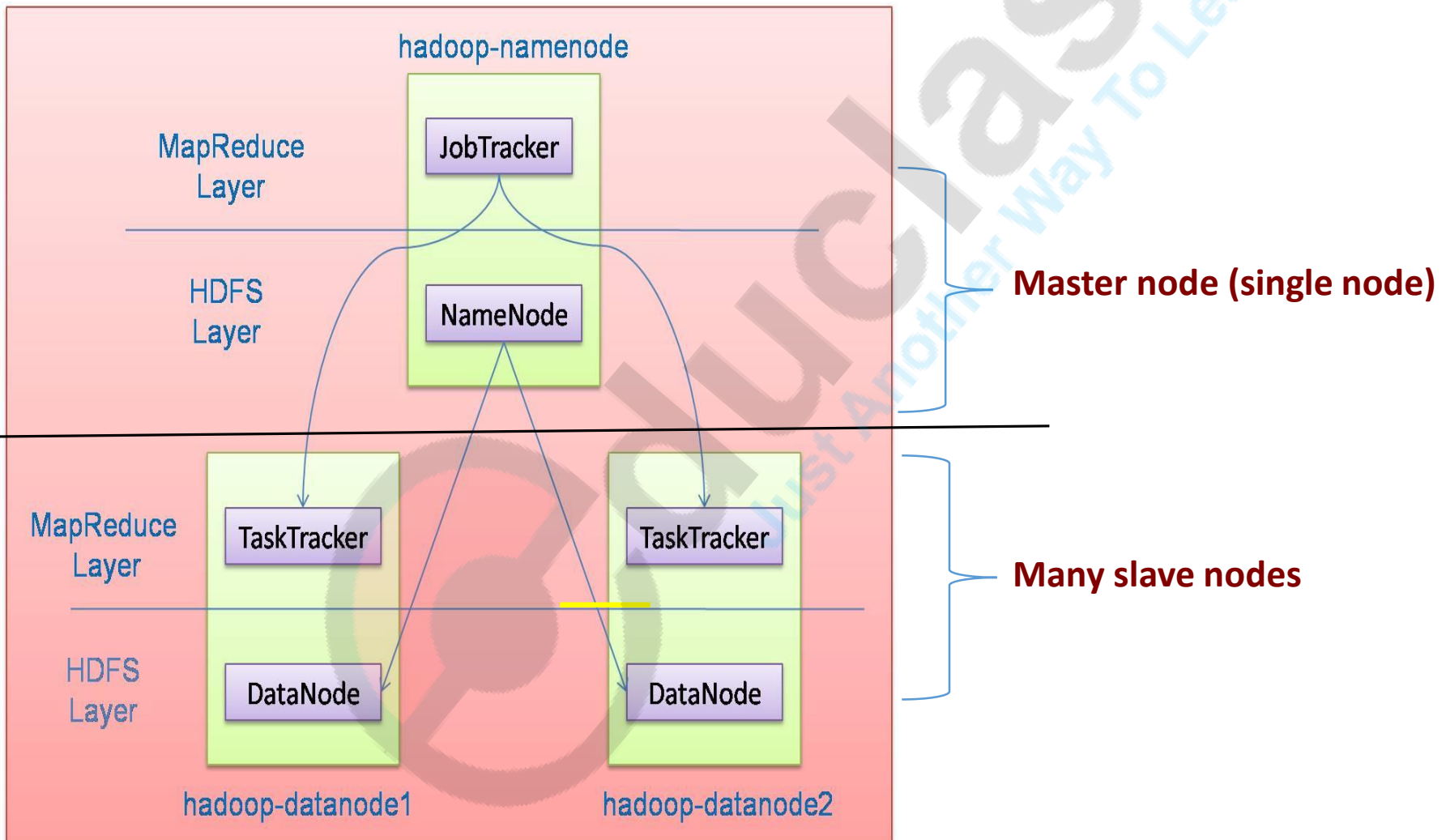- Hadoop is based on a simple data model, *any data will fit*

What is Hadoop (Cont'd)

- **Hadoop framework consists of two main layers**
  - Distributed file system (HDFS)
  - Execution engine (MapReduce)
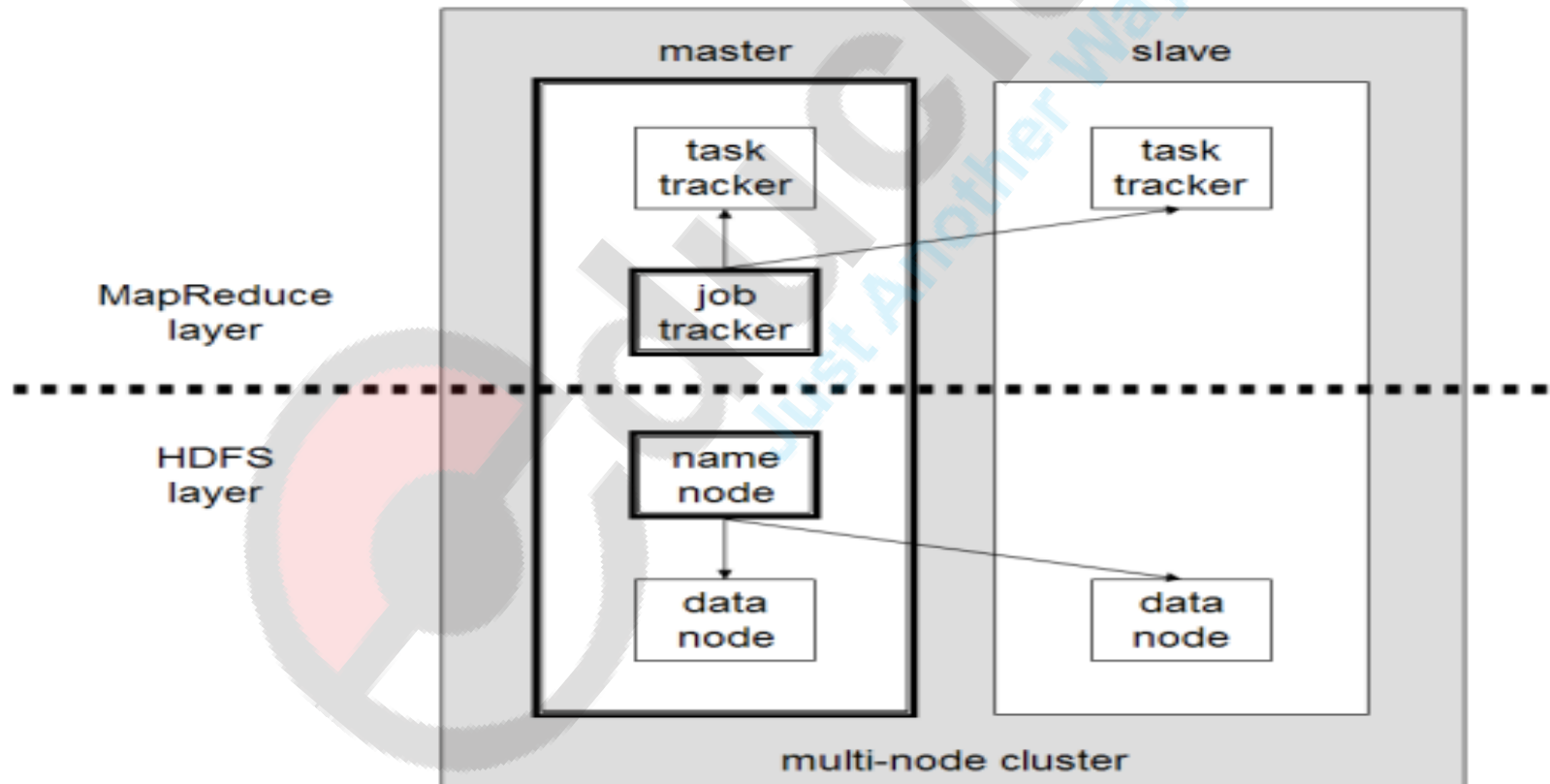
Hadoop Master/Slave Architecture

• Hadoop is designed as a *master-slave shared-nothing* architecture

WHO Uses MapReduce/Hadoop

- Google
- Yahoo
- IBM, Microsoft, Oracle
- Facebook, Amazon, AOL, NetFlex
- Many others + universities and research labs
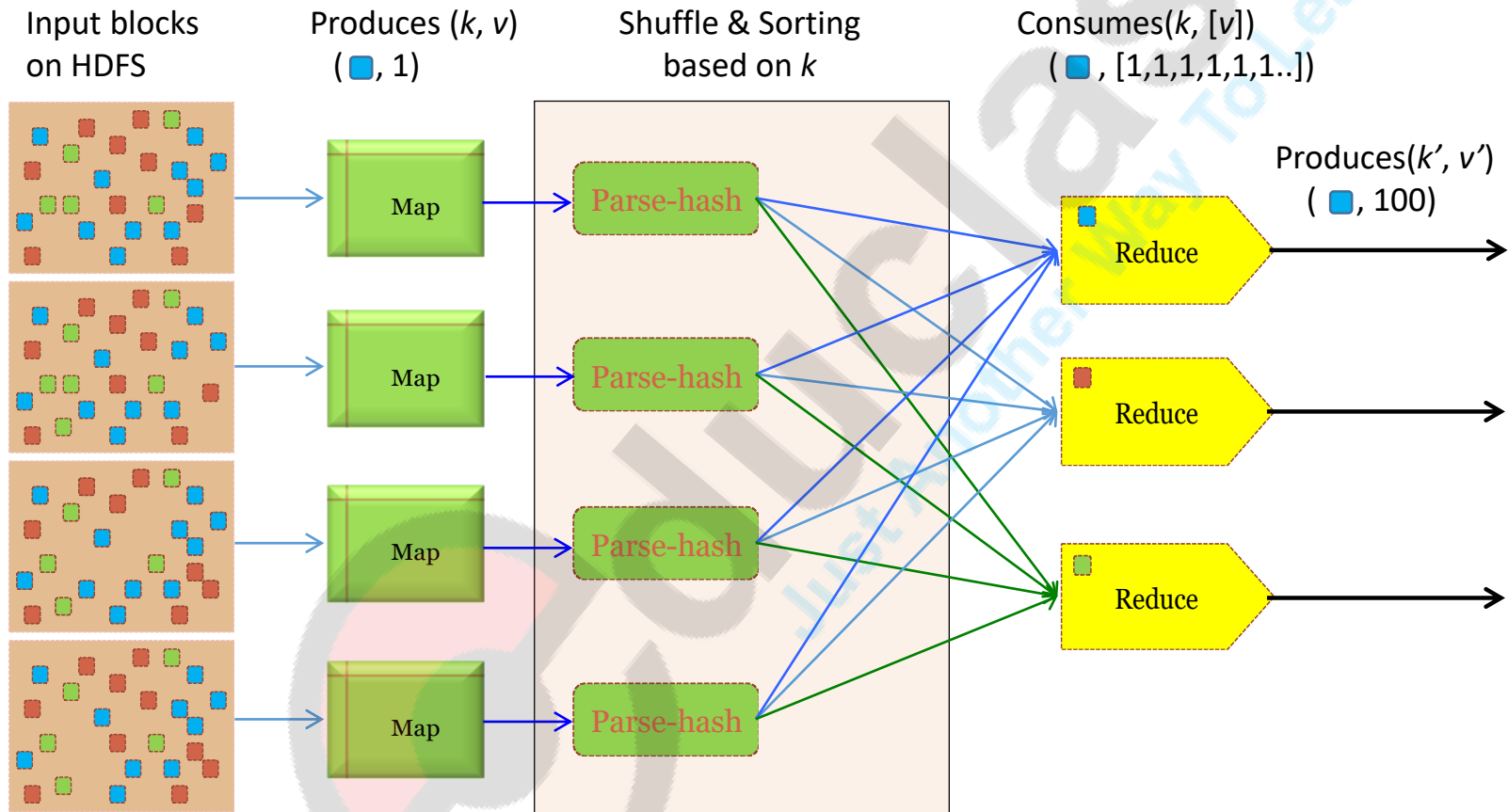
MapReduce

- **MapReduce** is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

- **MapReduce** is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster.

Functions in the Model

- Map
  - Process a key/value pair to generate intermediate key/value pairs
- Reduce
  - Merge all intermediate values associated with the same key
- Partition
  - By default : `hash(key) mod R`
  - Well balanced

Map-Reduce Execution Engine
(Example: Color Count)

Input blocks on HDFS

Produces (*k*, *v*) (🟦, 1)

Shuffle & Sorting based on *k*

Consumes(*k*, [*v*]) (🟦, [1,1,1,1,1,1..])

Produces(*k'*, *v'*) (🟦, 100)

Map

Parse-hash

Reduce

Users only provide the "Map" and "Reduce" functions

Architecture overview

Master node

user

Job tracker

Slave node 1

Task tracker

Workers

Slave node 2

Task tracker

Workers

Slave node N

Task tracker

Workers
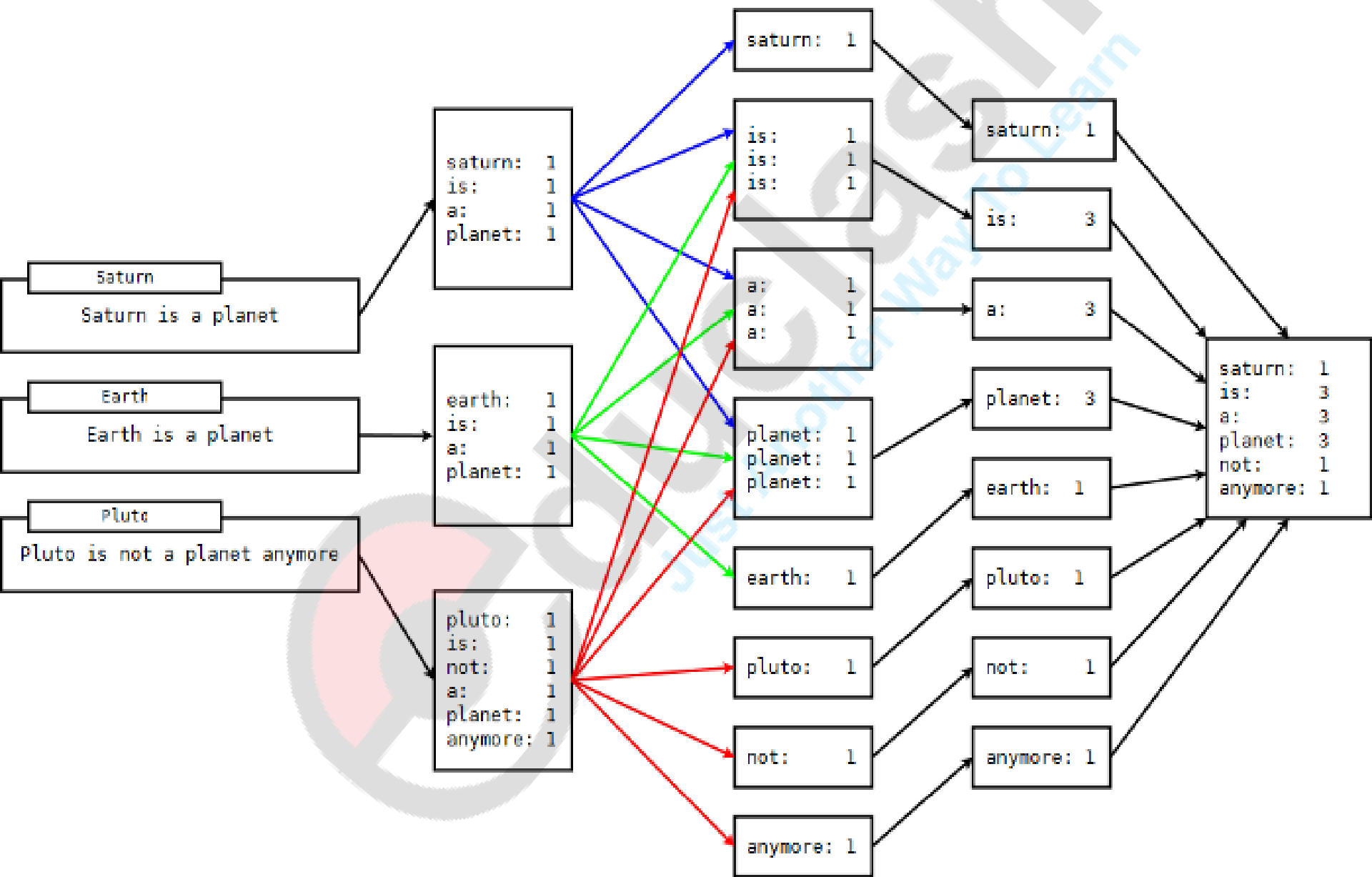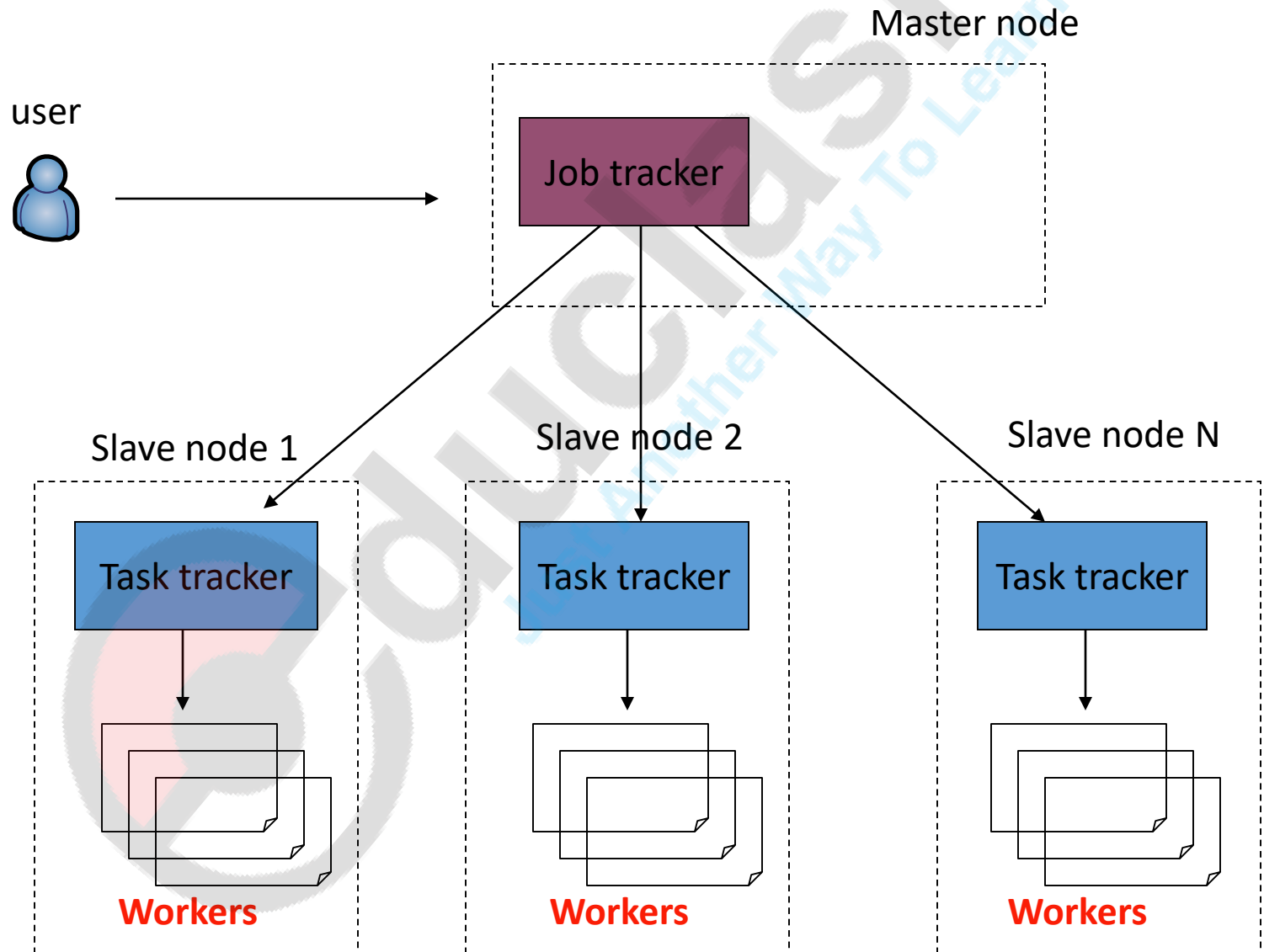
Properties of MapReduce Engine

- **Job Tracker is the master node (runs with the namenode)**
  - Receives the user's job
  - Decides on how many tasks will run (number of mappers)
  - Decides on where to run each mapper (concept of locality)

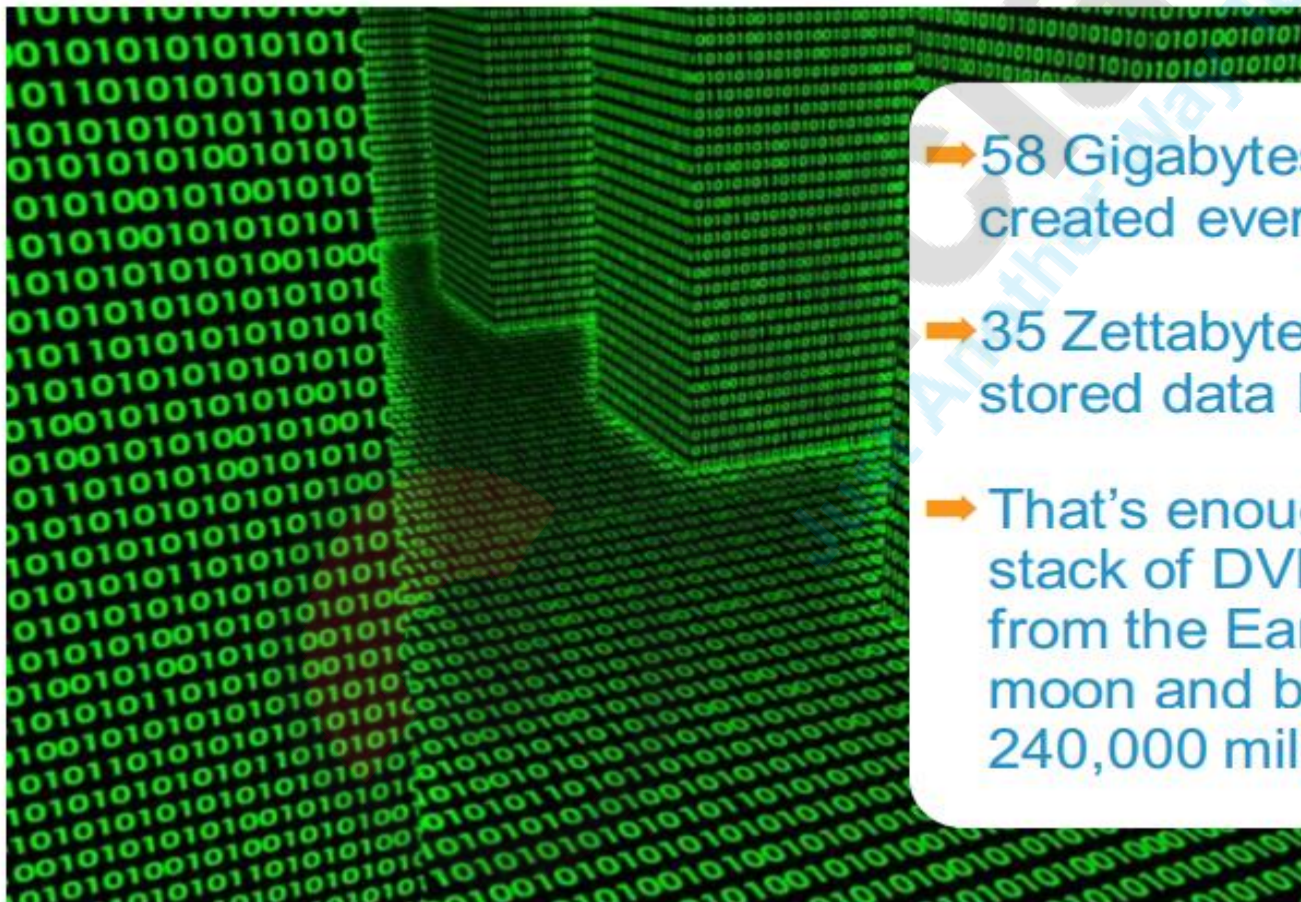# Properties of MapReduce Engine (Cont'd)

- **Task Tracker is the slave node (runs on each datanode)**
  - Receives the task from Job Tracker
  - Runs the task until completion (either map or reduce task)
  - Always in communication with the Job Tracker reporting progress
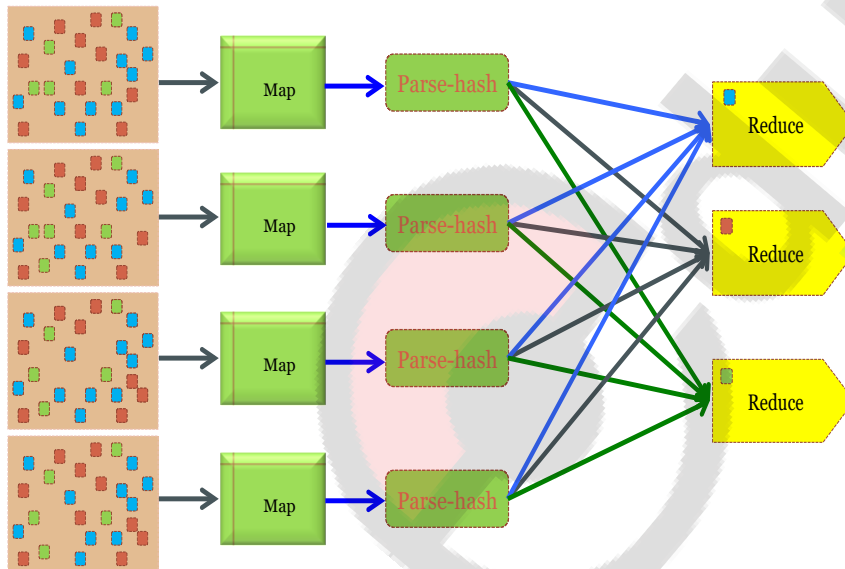
In this example, 1 map-reduce job consists of 4 map tasks and 3 reduce tasks
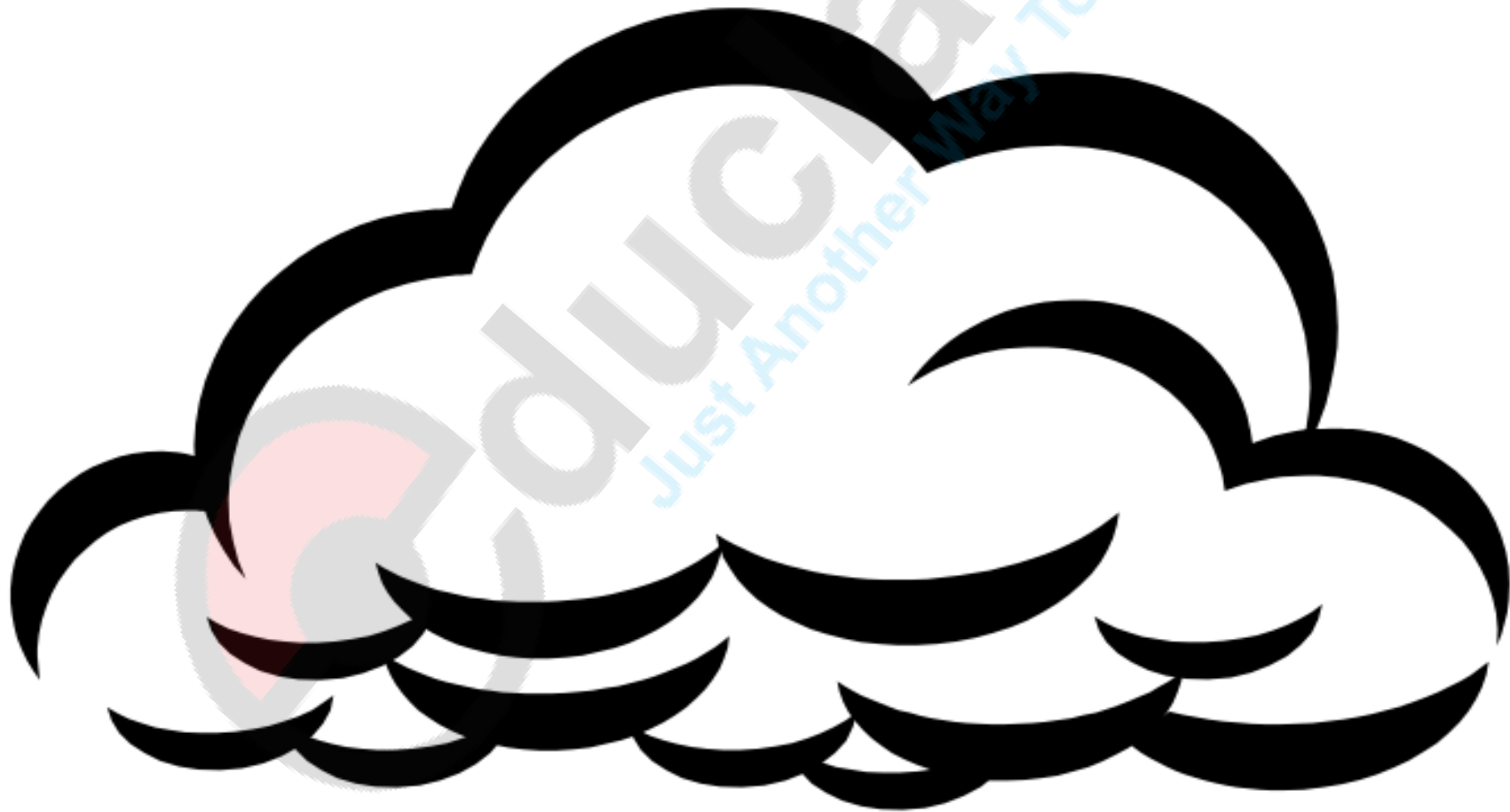
Cloud Data Centre

It's all about the data

→ 58 Gigabytes of data created every second

→ 35 Zettabytes of digitally stored data by 2020

→ That's enough data to fill a stack of DVDs reaching from the Earth to the moon and back — about 240,000 miles each way.

The 'Cloud' Will Save Us!

Only one problem

Cloud is powered by massive centralized infrastructure

Today's Cloud: Centralized Data Centers

Apple

Amazon

Google

Case Studies

- Amazon
- Microsoft
- Google
- Refer 'Cloud Computing By Deshpande/Bangare'