Copyright Notice

BigData

Prakash Badhe

prakash.badhe@vishwasoft.in

What Happens in an **Internet Minute?**

And **Future Growth** is **Staggering**

Today, the number of networked devices = the global population

By **2015**, the number of networked devices = **2x** the global population

In **2015**, it would take you **5 years** to view all video crossing IP networks each **second**

IP

intel

More and more Data

- Today We Live in the Data Age.
- Due to Internet networks, the speed of of data accumulation is keeps on increasing and increasing.
- And the World is getting more “Hungrier and Hungrier for Data”

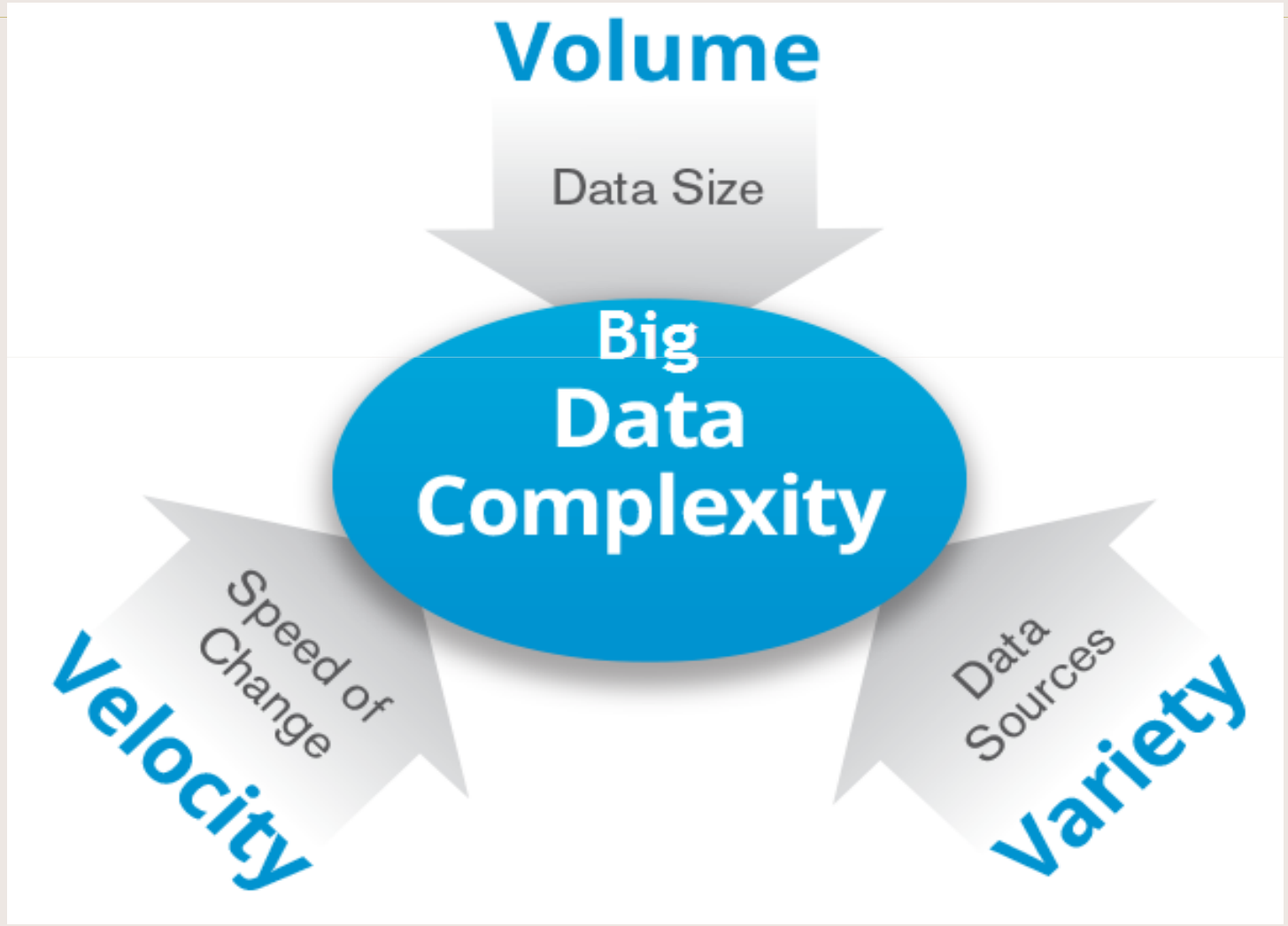
- AADHAR – Government of India's UIDAI project is considered as one of the largest BigData project in the World...!

Feb 14th 2011 – IBM's Super Computer **"WATSON" built using BigData Technology.**

What is Big Data

- BigData is the any amount of data that is **structured and/or unstructured data which is beyond the storage and processing capabilities of a single physical machine and traditional database techniques.**
- Data that has extra large Volume, comes from Variety of sources, Variety of formats and comes at us with a great Velocity is normally refers to as BigData.

The 3 V's of data



Database limitations

- Handles only structured data with known schemas
- High cost of maintenance
- Limited scalability
- Transaction execution limits the performance.
- Data is normally NOT replicated.
- User needs to understand complex SQL Queries.

Big Data at early stage

- Google originated in the year **1998.**
- They faced serious challenge in early 2000 to handle the BigData.
- In 2004 Google related two papers to handle Big data
  - GFS: Google File System
  - MapReduce: A Programming Model

- Apache Hadoop is an **open-source software framework, used to manage BigData.**
- Its built and used by a global community of contributors and users.
- It’s not only a tool, it’s a **Framework of tools.**
- **Moving computation is cheaper than moving data.**
- Most important Hadoop sub-projects:
  - HDFS: Hadoop Distributed File System
  - MapReduce: A Programming Model
- **Hadoop“ is the name of a stuffed toy ELEPHANT that belonged to the son of its creator “DOUG CUTTING”.**

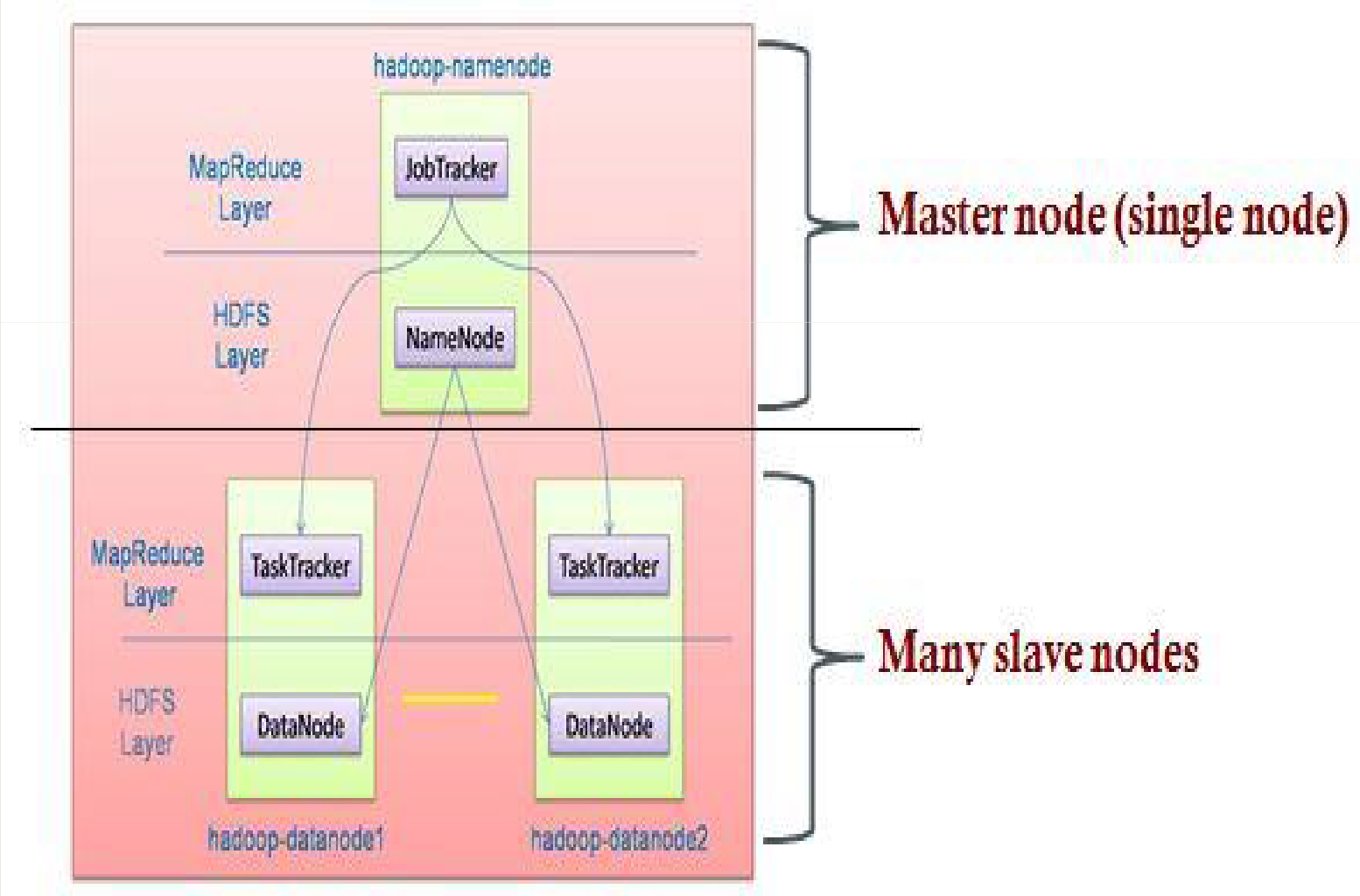
Hadoop Features

- **Scalable– New nodes can be added without changing data formats.**
- **Cost-effective– It parallelly processes huge datasets on large clusters of commodity computers.**
- **Efficient and Flexible- It is schema-less, and can absorb any type of data, from any number of sources.**

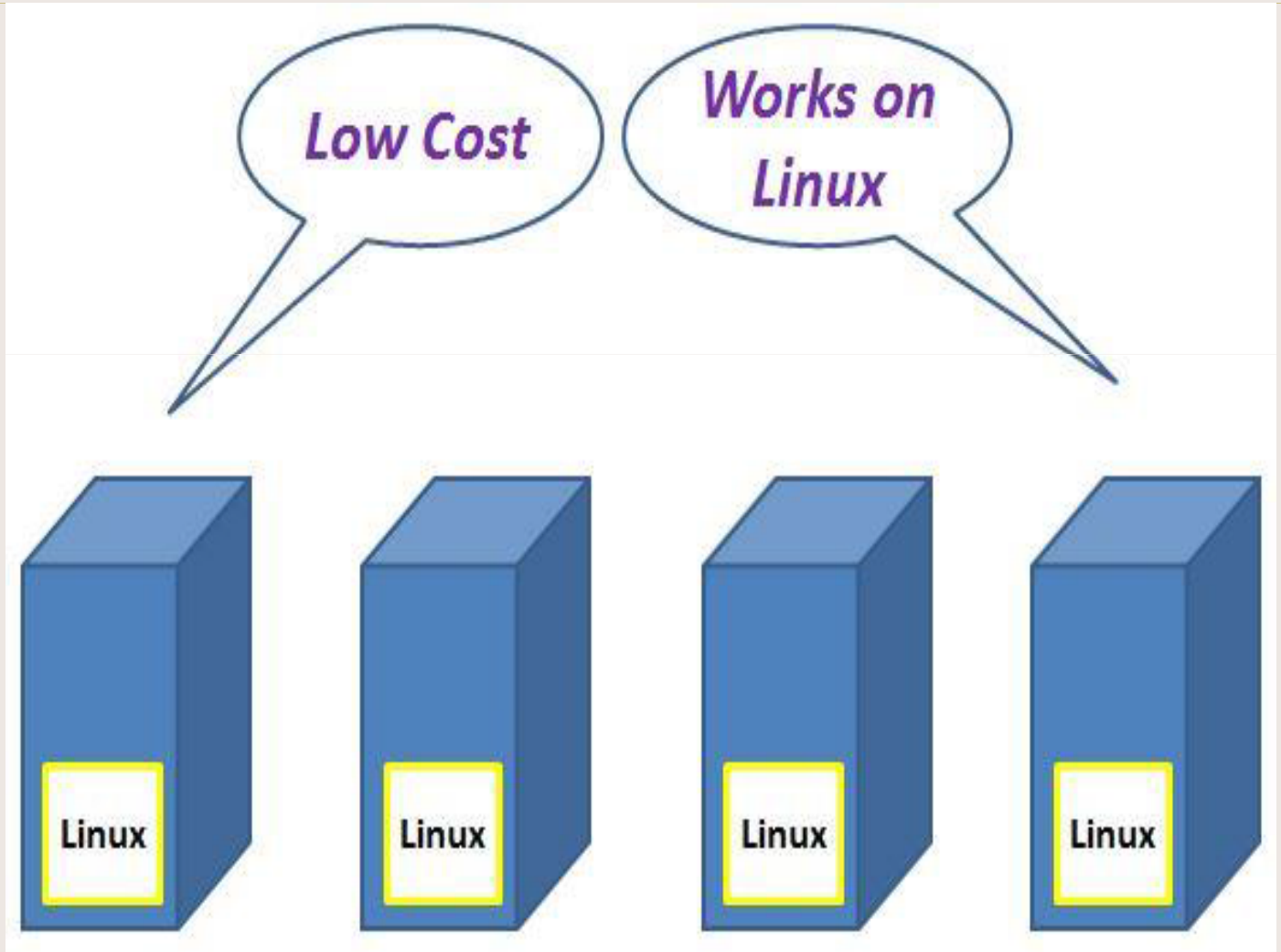
Hadoop Features

- F**ault-tolerant and Reliable- It handles failures of nodes easily because of Replication.**
- **Easy to use- It uses simple Map and Reduce functions to process the data.**
- It is developed in **Java but it can support Python & others too.**

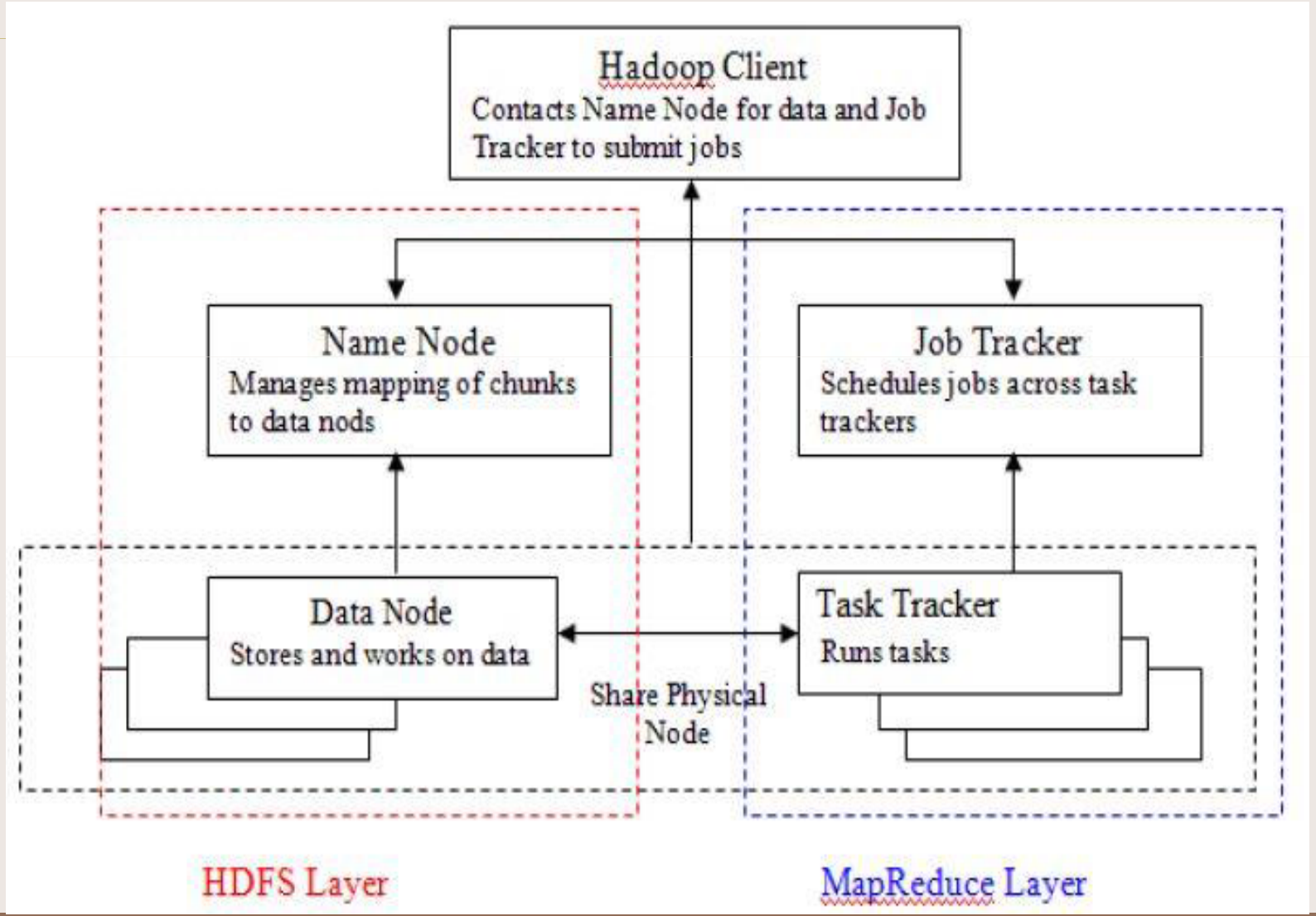
Hadoop Architcture



Hadoop Distributed



Hadoop Data Flow



Hadoop Core

Hadoop core has two major components:

1.HDFS

- a.Name Node
- b.Secondary Name Node
- c.Data Node

2.MapReduce Engine

- a.Job Tracker
- b.Task Tracker

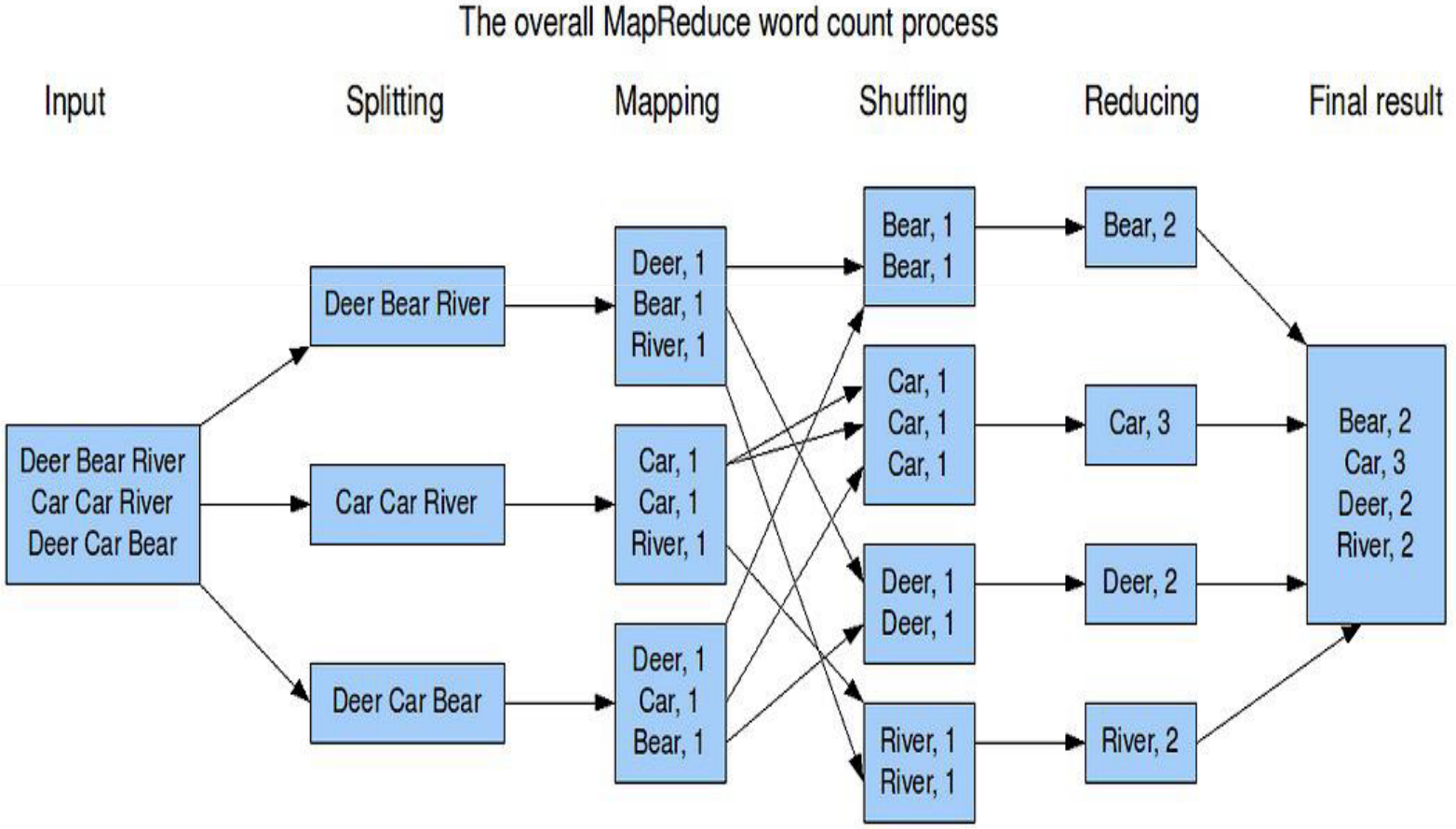
HDFS

- Pioneered by Google File System (GFS)
- It consists of three major components -
- **Name Node**
- It is responsible for the distribution of the data throughout the Hadoop cluster.
- **Secondary Name Node (Backup Node)**
- It regularly contacts Name Node and maintains an up to date snapshots of Name Node's directory information.
- **Data Node**
- It is responsible to store the chunk of data that is assigned to it by the Name Node.

Map Reduce

- Pioneered by Google, Popularized by Yahoo (Apache).
- It consists of two major components –
- **Job Tracker**
- It is responsible for scheduling the task to slave nodes.
- So it consults the Name Node and assigns the task to the nodes which has the data on which task would be performed.
- **Task Tracker**
- It has the actual logic to perform the task, so it performs Map and Reduce functions on the data assigned to it by Master Node.

Map Reduce Example



Hadoop Advantage

- **Moving Computation is far better than Moving Data**
- Runs on commodity hardware, No special HW needed.
- It's a Master/Slave architecture
- It handles all types of node failures by live Heartbeats
- It handles assigning tasks to nodes
- It has Rack awareness between nodes
- The Programmers only need to concentrate on getting business values from BigData

Hadoop Limitations

- Not suitable, if data is too small.
- Not suitable, if there is a **dependency between the data.**
- Not suitable, if Job cannot be divided into small chunks.
- Not suitable, to process real-time and stream-based processing.

Cloud Computing

- Cloud computing is basically the provision of on-demand computing services.
- There are quite a lot of things that come under it, ranging from applications to storage and processing capabilities, usually via the internet and pay on a basis as well.
- **Components involved**
  - Operating system
  - Virtualization
  - Networking

Cloud Players

- Amazon
- Google
- Azure (Microsoft)
- Private cloud offerings

Cloud applications

- Specialized applications VM
- Custom Cluster of Machines
- Cloud based testing tools
- Data Centers
- Data processing
- Analyzers (DNA Processors, Space Image processing)

AI

- AI is when Machines
- – Exhibit intelligence
- – Perceive their environment
- – Take actions/make decision to maximize chance of success at a goal

Computers

- Made with algorithms
- Knowledgeable ONLY about what taught
- Control ONLY what we give them control of
- Aware of nuances and can continue to learn more
- Do very boring work for humans repeatedly
- Often make better, more consistent decisions than humans
- Be efficient, won’t get tired

How AI can be applied

- Subject Matter Experts (SME's) Availability
- – Lawyers
- – Machinists
- – Insurance adjusters
- – Physicians
- • Usually not experienced in machine learning
- – Need close collaboration with those making algorithms

AI ?

- Creating an AI requires
- Algorithms
- Documents
- Ground truth (annotation)
- Teaching
- Iteration
- Repeat

Machine Learning

- Machine learning creates more highly trained specialists
- Not an “all **knowing**” being

Inputs..

- What intelligence does the system need?
- What is the AI perceiving in their environment?
- What actions are taken to maximize chance of success at goal?
  - Intelligence?
  - Perception?
  - Action/Decision?

Examples

- Understanding Human Speech
- Speech Generation ?
- Decision making : Self Diving cars
- Image recognition and Processing
- Sound recognition
- Text analysis
- Automation for Repetitive work
- Process Retina images
- Deep Learning and Deep Fakes
- Optical Character Recognition