Hadoop – Introduction

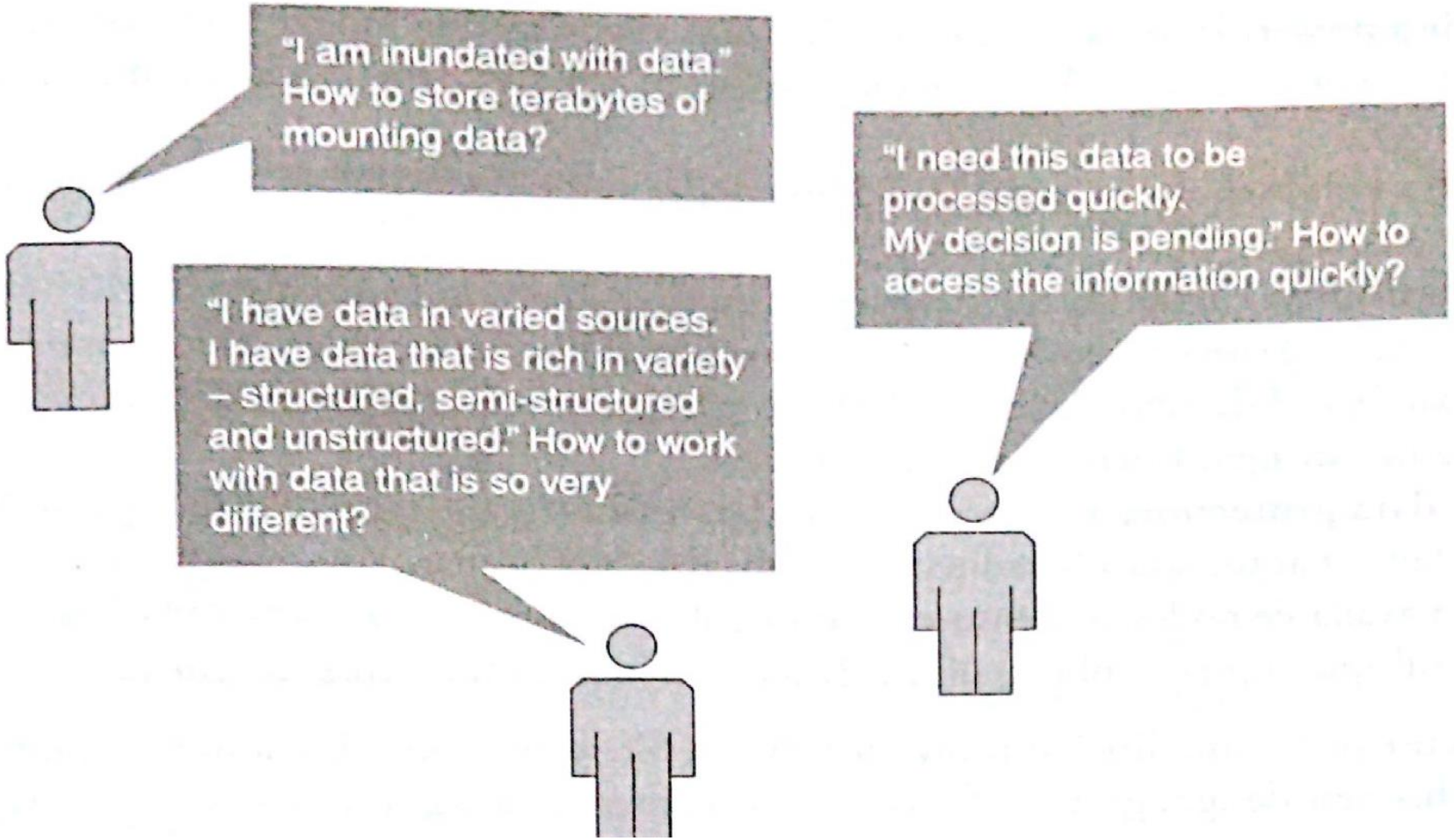
15CSE334- Bigdata Analytiics

A.Baskar

Introducing Hadoop

1. **Every day:**
   (a) NYSE (New York Stock Exchange) generates 1.5 billion shares and trade data.
   (b) Facebook stores 2.7 billion comments and Likes.
   (c) Google processes about 24 petabytes of data.
2. **Every minute:**
   (a) Facebook users share nearly 2.5 million pieces of content.
   (b) Twitter users tweet nearly 300,000 times.
   (c) Instagram users post nearly 220,000 new photos.
   (d) YouTube users upload 72 hours of new video content.
   (e) Apple users download nearly 50,000 apps.
   (f) Email users send over 200 million messages.
   (g) Amazon generates over $80,000 in online sales.
   (h) Google receives over 4 million search queries.
3. **Every second:**
   (a) Banking applications process more than 10,000 credit card transactions.

Challenges with Big volume, variety and velocity of data



What is Hadoop?

What exactly is Hadoop?

- **Hadoop** is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
- **Hadoop** is an open source, Java based framework used for storing and processing **big data.**
- A software ecosystem that allows for massively parallel computing

Is Big Data and Hadoop same?

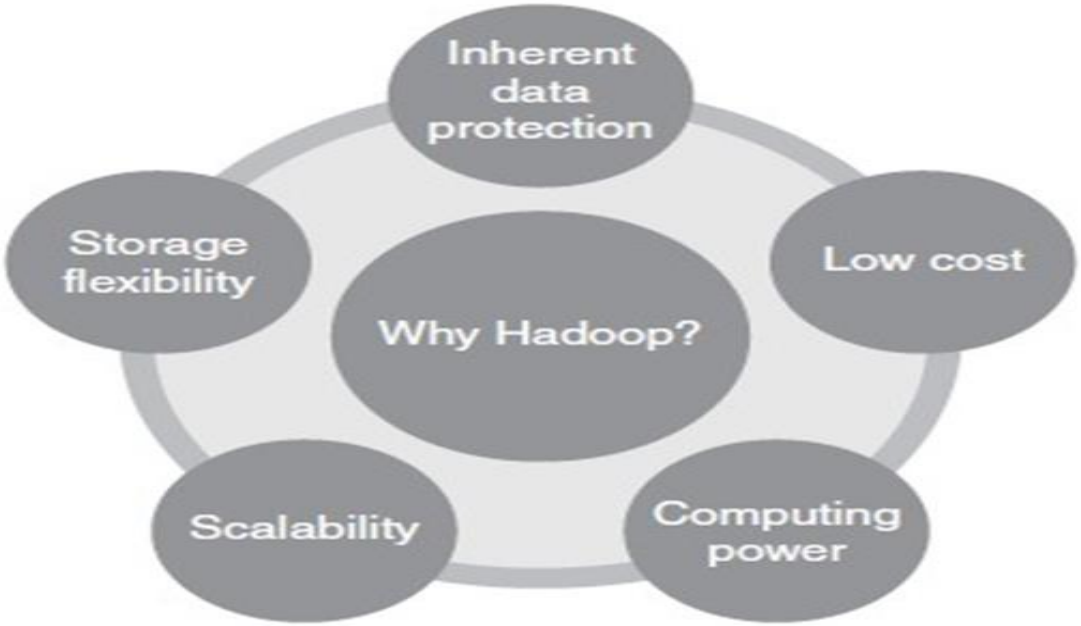
- **Hadoop** is a kind of framework that can handle the huge volume of **Big Data** and process it,
- whereas **Big Data** is just a large volume of the **Data** which can be in unstructured and structured **data**

Why Hadoop is called a big data technology?

- **Hadoop** is the **Big Data** operating system. Optimized for parallel processing using structured and unstructured **data**, using low hardware costs.
- It has the ability to process unstructured **data** that is more than 80% world's **data**.

Why Hadoop?

- The key consideration (the rationale behind its huge popularity) is:
- ***Its capability to handle massive amounts of data, different categories of data – fairly quickly.***
- The other considerations are :



Why Hadoop?

- With this new paradigm the data can be manged by Hadoop systems as follows:
  - Distributes data and duplicates chunk of data across several nodes
  - Locally available computing resource is used to process each chunk of data parallelly.
  - Hadoop handles failures smartly and automatically

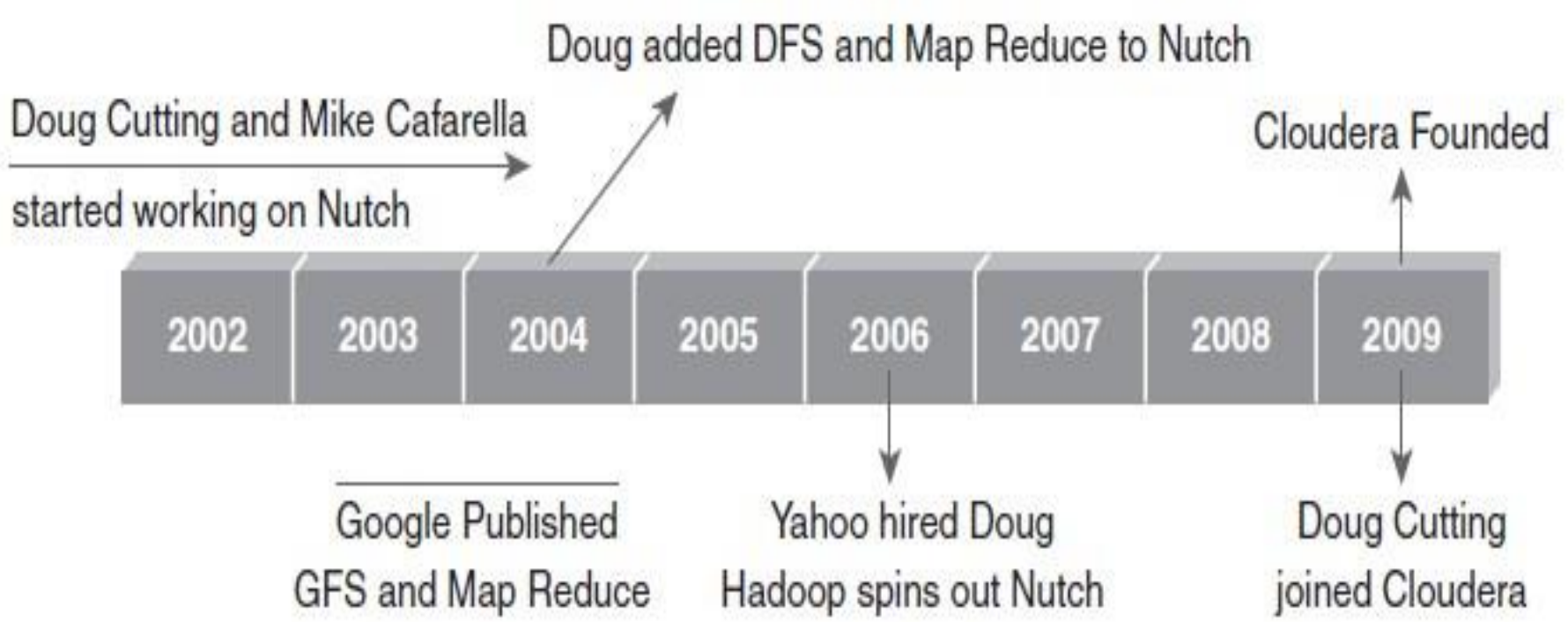
RDBMS Vs Hadoop

PARAMETERS	RDBMS	HADOOP
System	Relational Database Management System.	Node Based Flat Structure.
Data	Suitable for structured data.	Suitable for structured, unstructured data. Supports variety of data formats in real time such as XML, JSON, text based flat file formats, etc.
Processing	OLTP	Analytical, Big Data Processing
Choice	When the data needs consistent relationship.	Big Data processing, which does not require any consistent relationships between data.
Processor	Needs expensive hardware or high-end processors to store huge volumes of data.	In a Hadoop Cluster, a node requires only a processor, a network card, and few hard drives.
Cost	Cost around $10,000 to $14,000 per terabytes of storage.	Cost around $4,000 per terabytes of storage.

Distributed computing challenges

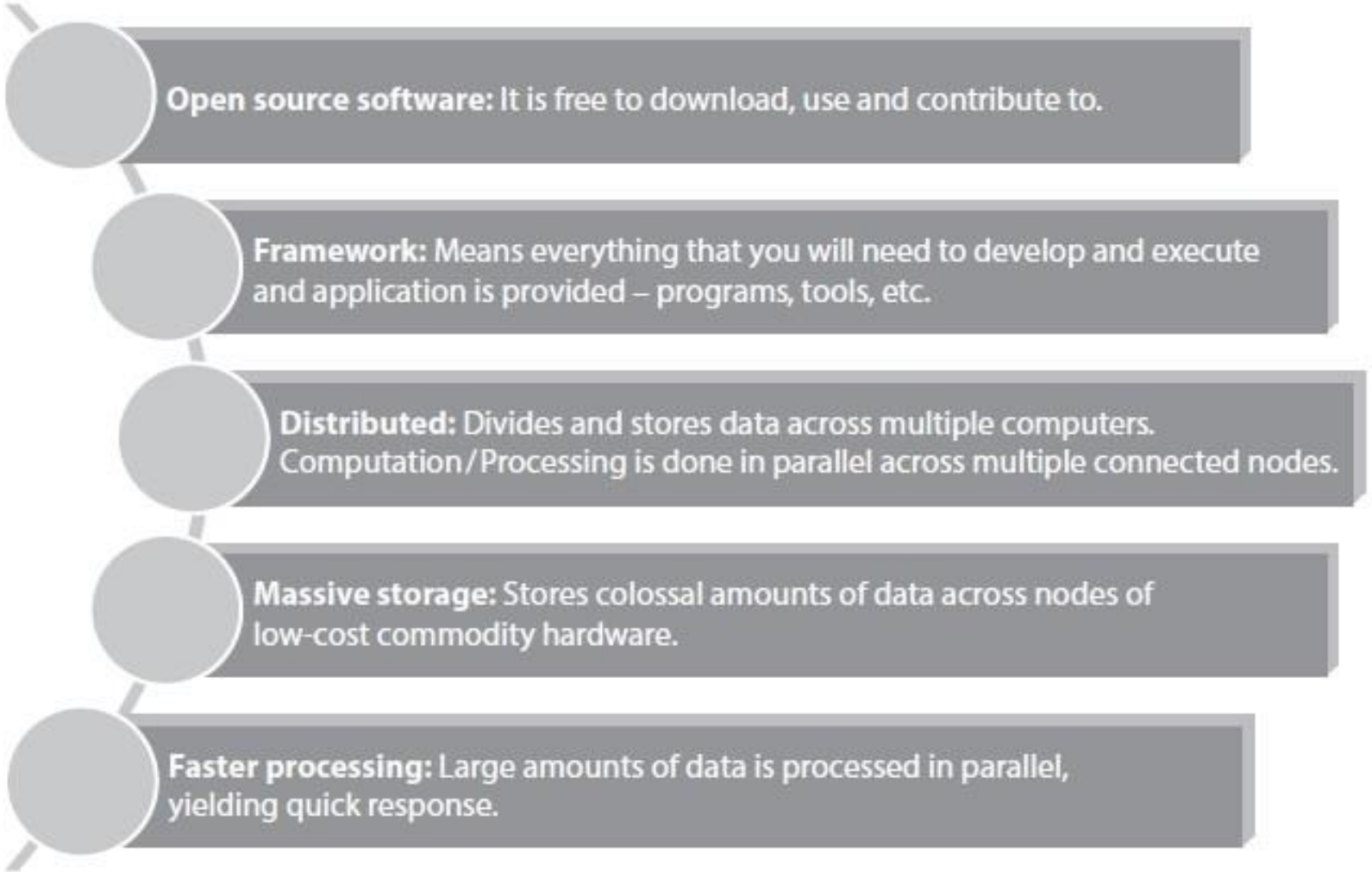
- Hardware Failure
  - Replication Factor
- How to Process This Gigantic Store of Data?
  - MapReduce

History of Hadoop

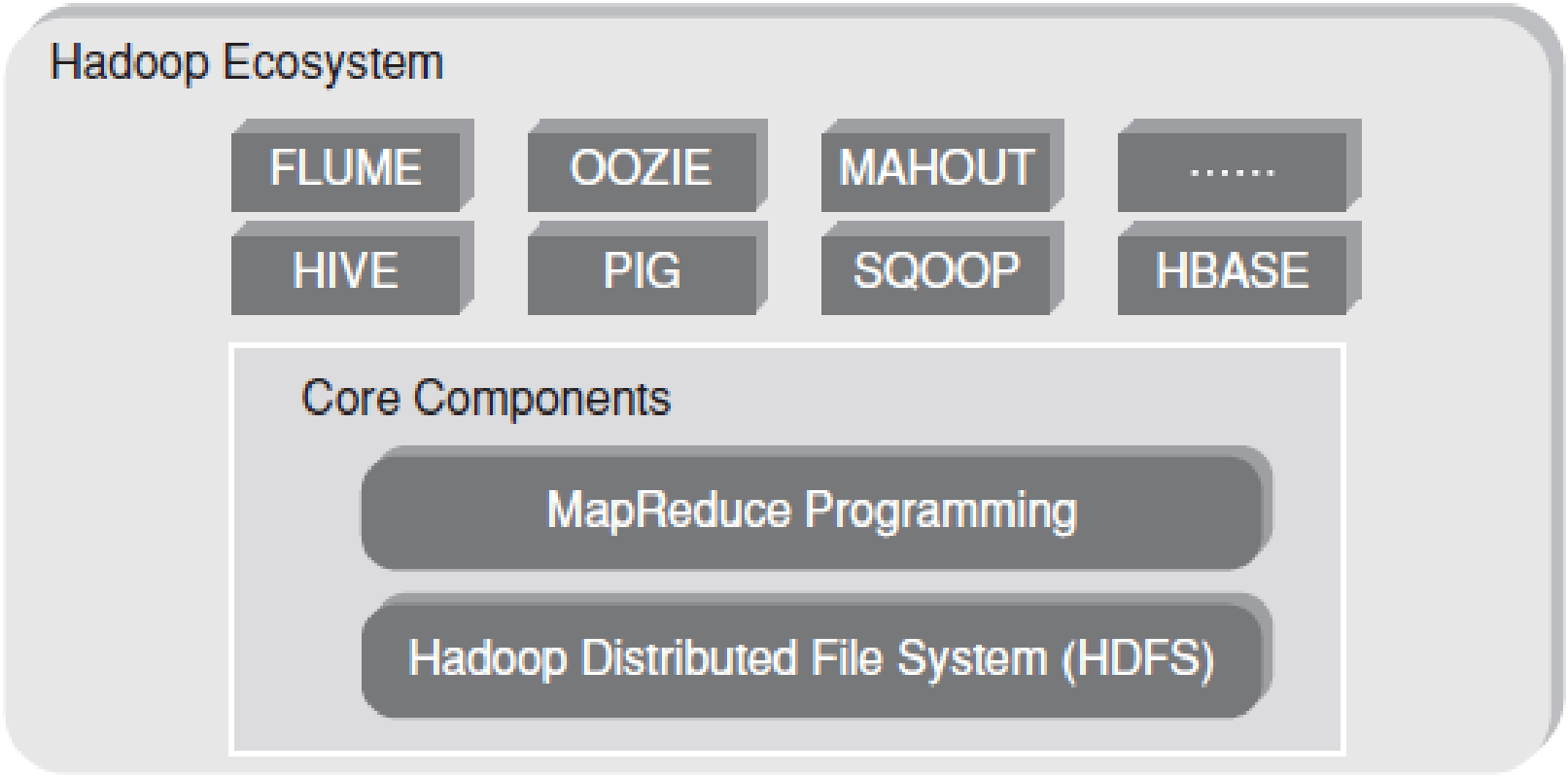


Hadoop Overview

Key Aspects of Hadoop



Hadoop Components



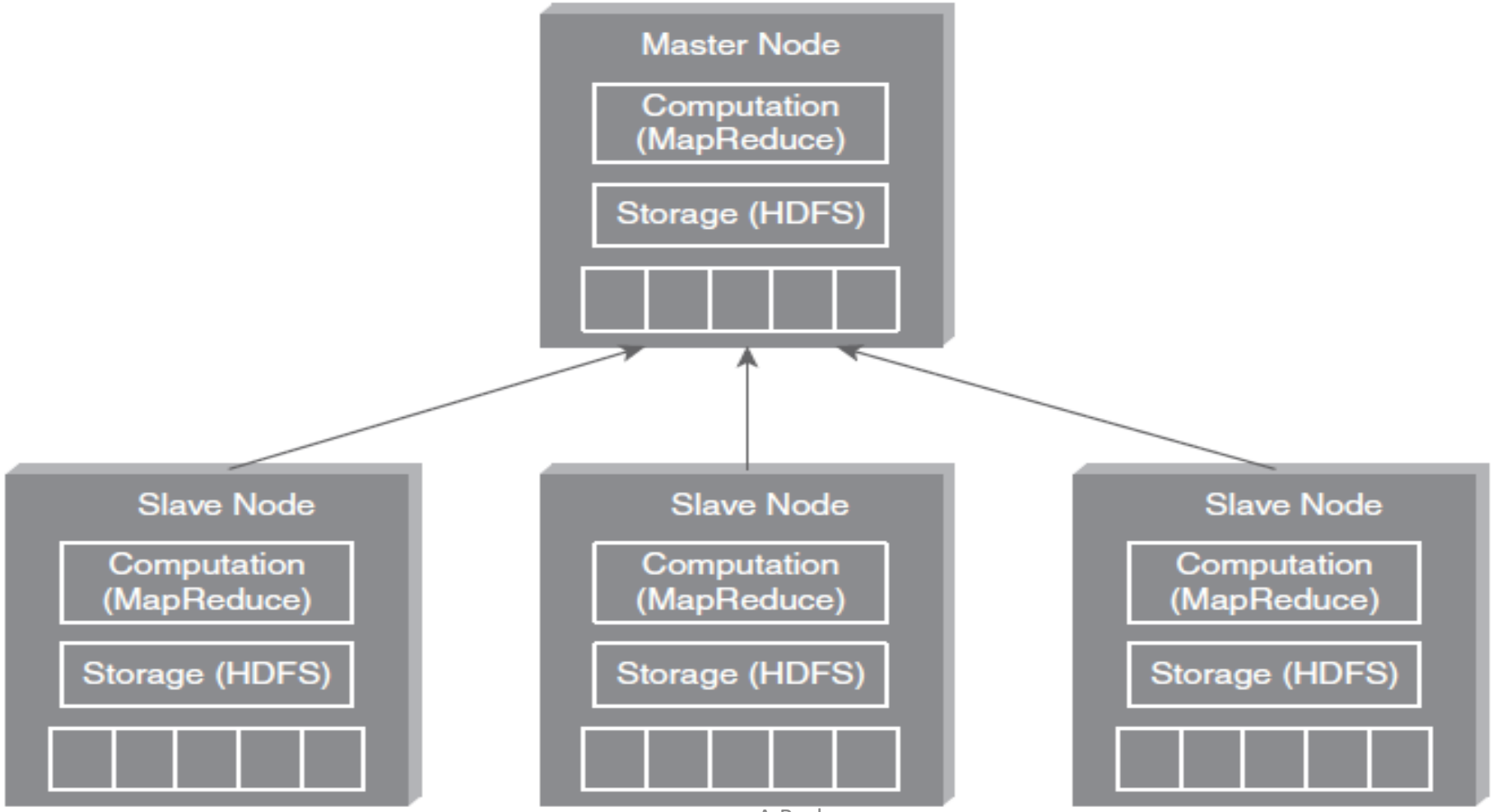
Hadoop Components

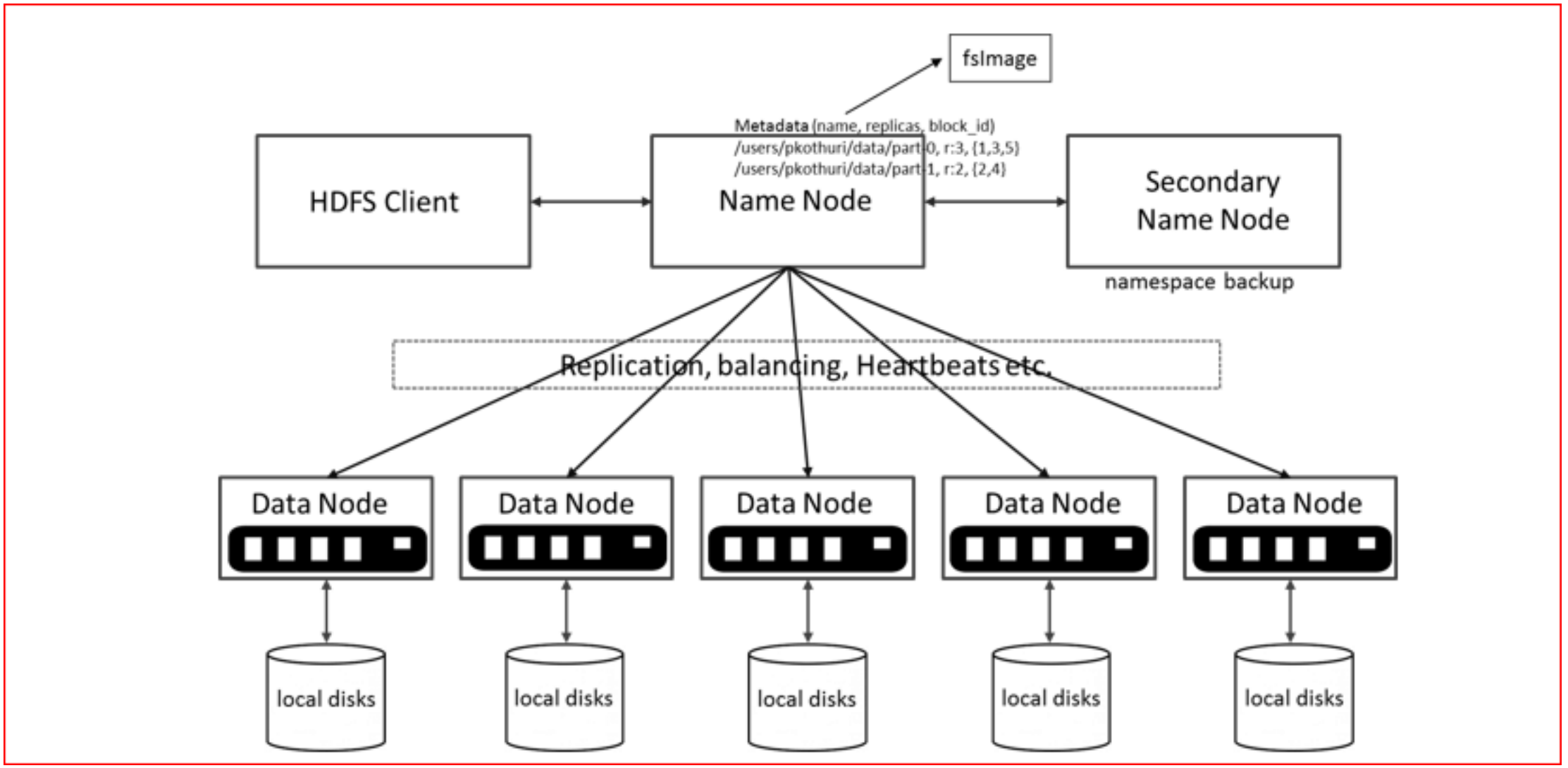
Hadoop Core Components:

HDFS:
(a)Storage component.
(b)Distributes data across several nodes.
(c)Natively redundant.

MapReduce:
(a)Computational framework.
(b)Splits a task across multiple nodes.
(c)Processes data in parallel.

Hadoop High Level Architecture




fsImage
Metadata (name, replicas, block_id)
/users/pkothuri/data/part-0, r:3, {1,3,5}
/users/pkothuri/data/part-1, r:2, {2,4}
HDFS Client
Name Node
Secondary Name Node
namespace backup
Replication, balancing, Heartbeats etc.
Data Node
Data Node
Data Node
Data Node
Data Node
local disks
local disks
local disks
local disks
local disks

Hadoop Distributors