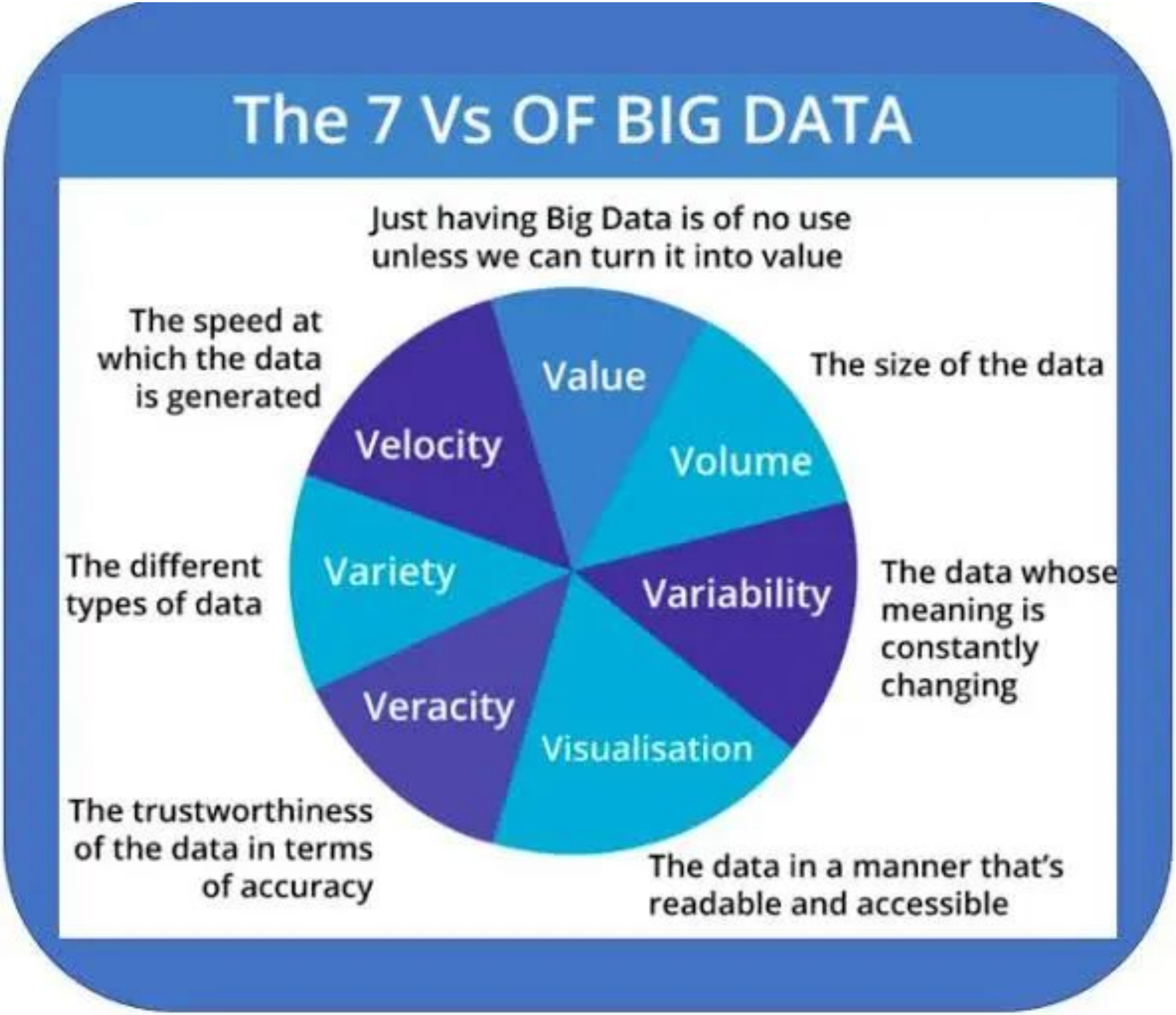
Big Data Analytics and Information Retrieval

Netflix Recommender System — A Big Data Case Study

What is Netflix and what do they do?

- Provides movie streaming through a subscription model
- Includes television shows and in-house produced content along with movies
- Company is heavily data-driven.
- Their data of tens of petabytes of data was moved to AWS in 2016
- Netflix stores approximately 105TB of data with respect to videos alone.
- By the end of 2019, Netflix has 1 million subscribers and 159 million viewers

Why was this a “big data” problem?



The 7 Vs OF BIG DATA

Just having Big Data is of no use unless we can turn it into value

The size of the data

The data whose meaning is constantly changing

The data in a manner that’s readable and accessible

The trustworthiness of the data in terms of accuracy

The different types of data

The speed at which the data is generated

1. Volume

This is the main characteristic of big data. The term volume here defines big data as "BIG".

With a massive amount of data generating daily, we know gigabytes is not enough to store such huge amount of data.

Because of this, now the data is stored in terms of Zettabytes, Exabytes, and Yottabytes. For instance, almost 50 hours of videos are uploaded on YouTube every single minute.

Now imagine how much data is being generated on YouTube itself.

4. Variability

Variability is different from the variety. Variability refers to the data which keeps on changing constantly.

Variability mainly focuses on understanding and interpreting the correct meanings of raw data.

For example – A soda shop may offer 6 different blends of soda, but if you get the same blend of soda every day and it tastes different every day, that is variability.

The same is in the case of data, and if it is continuously changing, then it can have an impact on the quality of your data.

Visualization

Visualization here refers to how you can present your data to the management for decision-making purposes.

We all know that data can be presented in many ways, such as excel files, word docs, graphical charts, etc.

Irrespective of the format, the data should be easily readable, understandable, and accessible, and that's why data visualization is important.

Veracity

If your data is not accurate, it is of no use, and here comes the concept of Veracity. It is all about making sure the data gathered by you is accurate and also keeping the bad data away from your systems.

It is also the trustworthiness or quality of data which a company received and processes to derive useful insights

Variety

Here variety means types of data sources. Big data can be of various types – structured, semi-structured, and unstructured.

In today’s world, the data which is generated in large quantities is unstructured data only like audio files, video, images, text files, etc.

Value

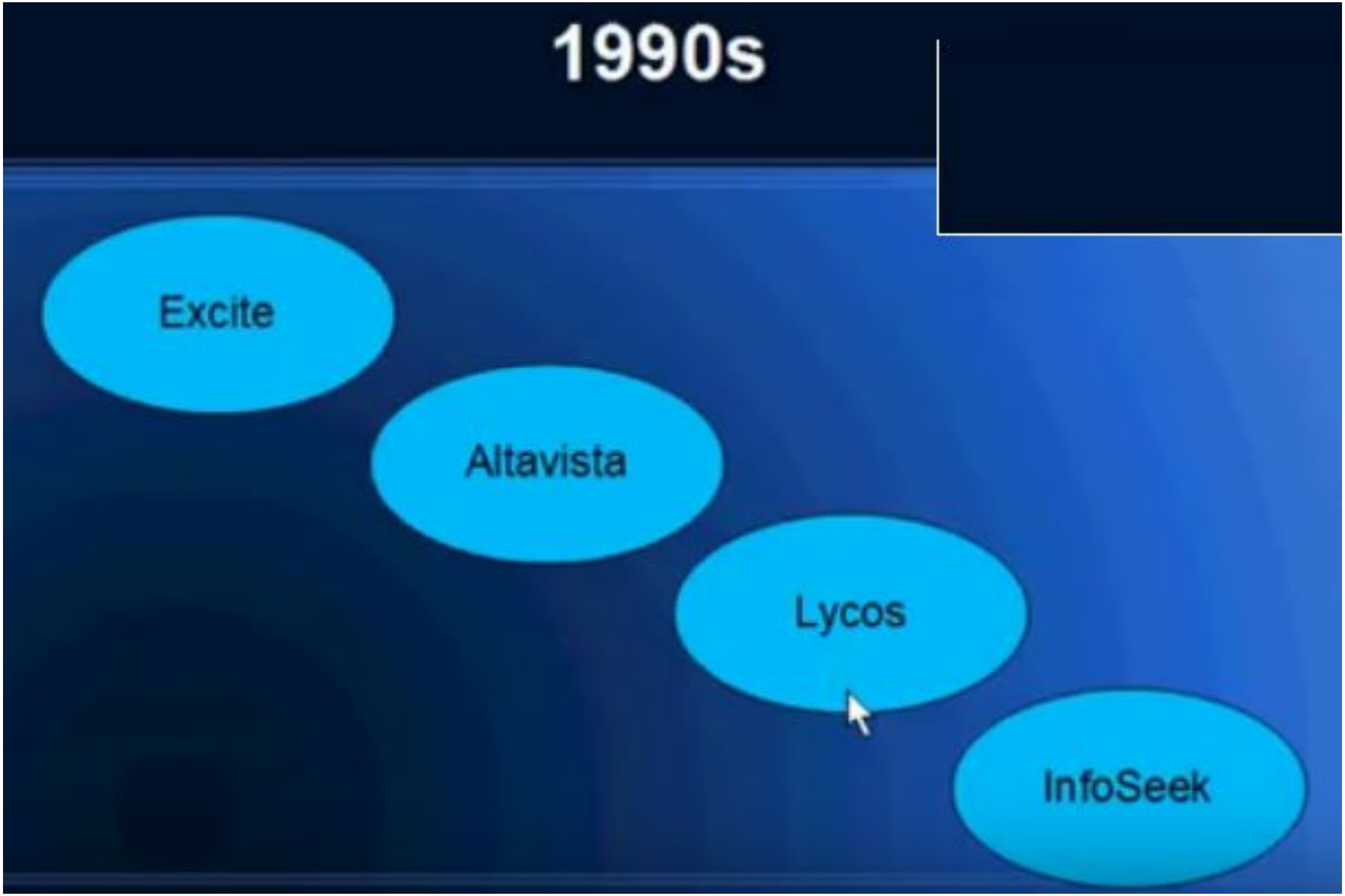
Value is known as the end game in big data. Every user needs to understand that the organization needs some value after efforts are made and resources are spent on the above mentioned V’s.

Big Data can help a user provide value if it is done and processed correctly

Velocity

Velocity here refers to how fast the data can be processed and accessed. For example, social media posts, YouTube videos, audio files, images that are uploaded in thousands every second should be accessible as early as possible.

Search Engines in 1990s



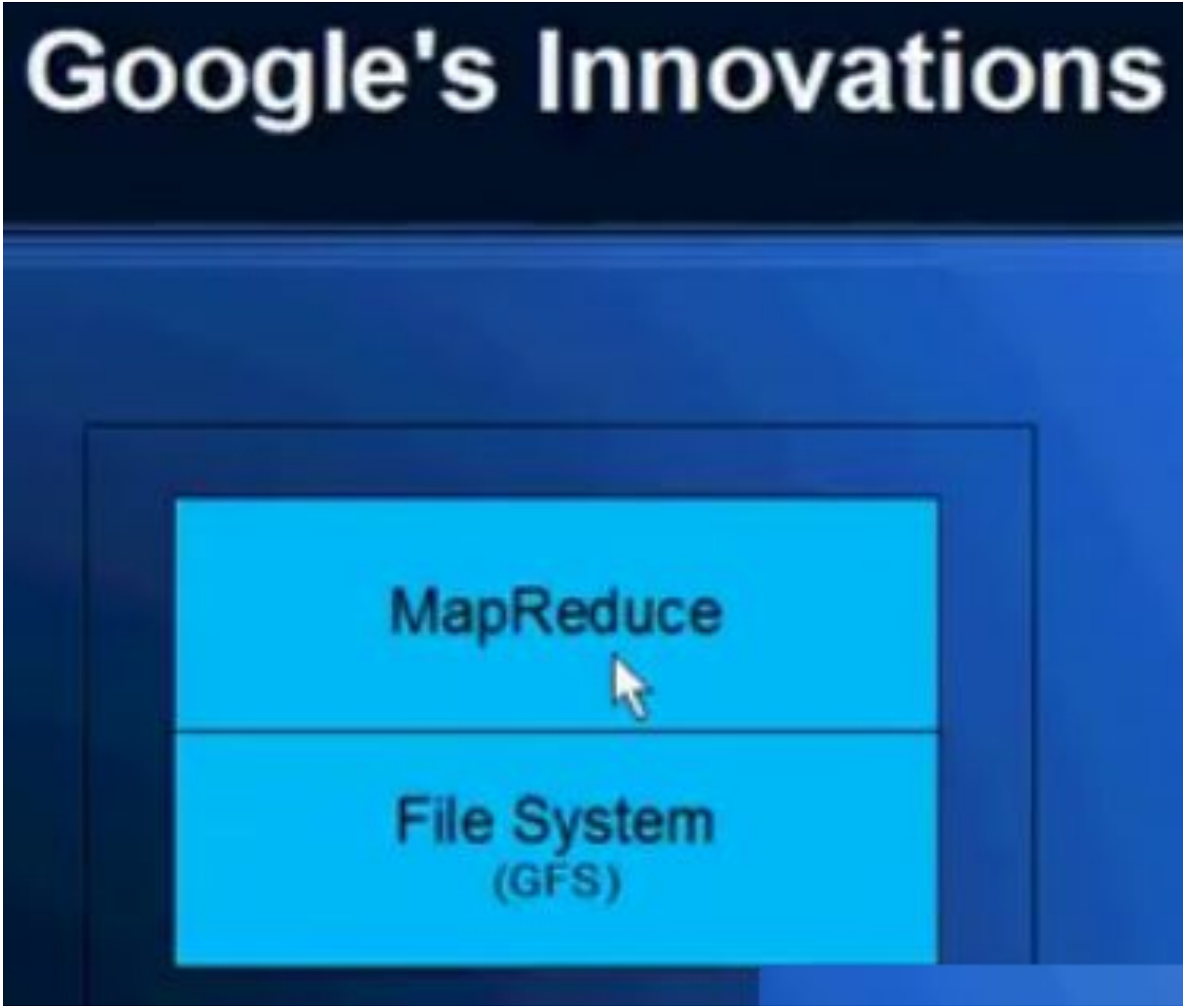
1990s

Excite

Altavista

Lycos

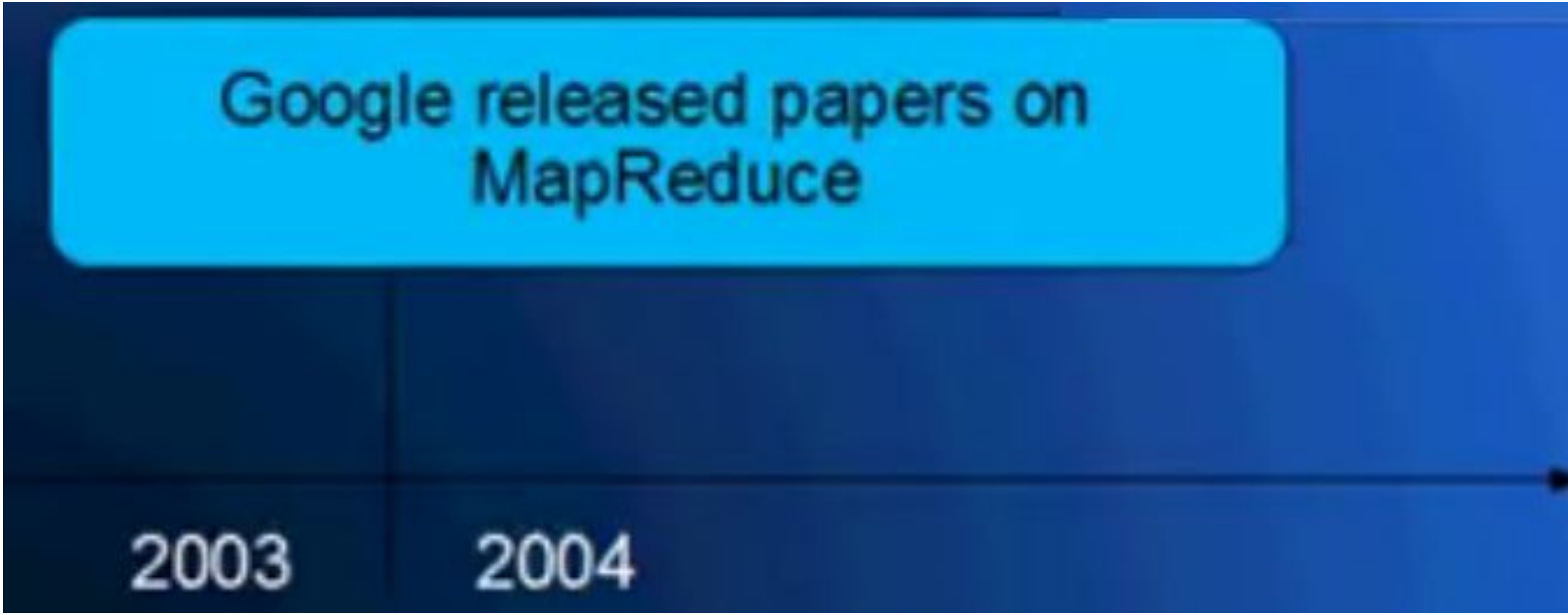
InfoSeek



Google's Innovations

MapReduce

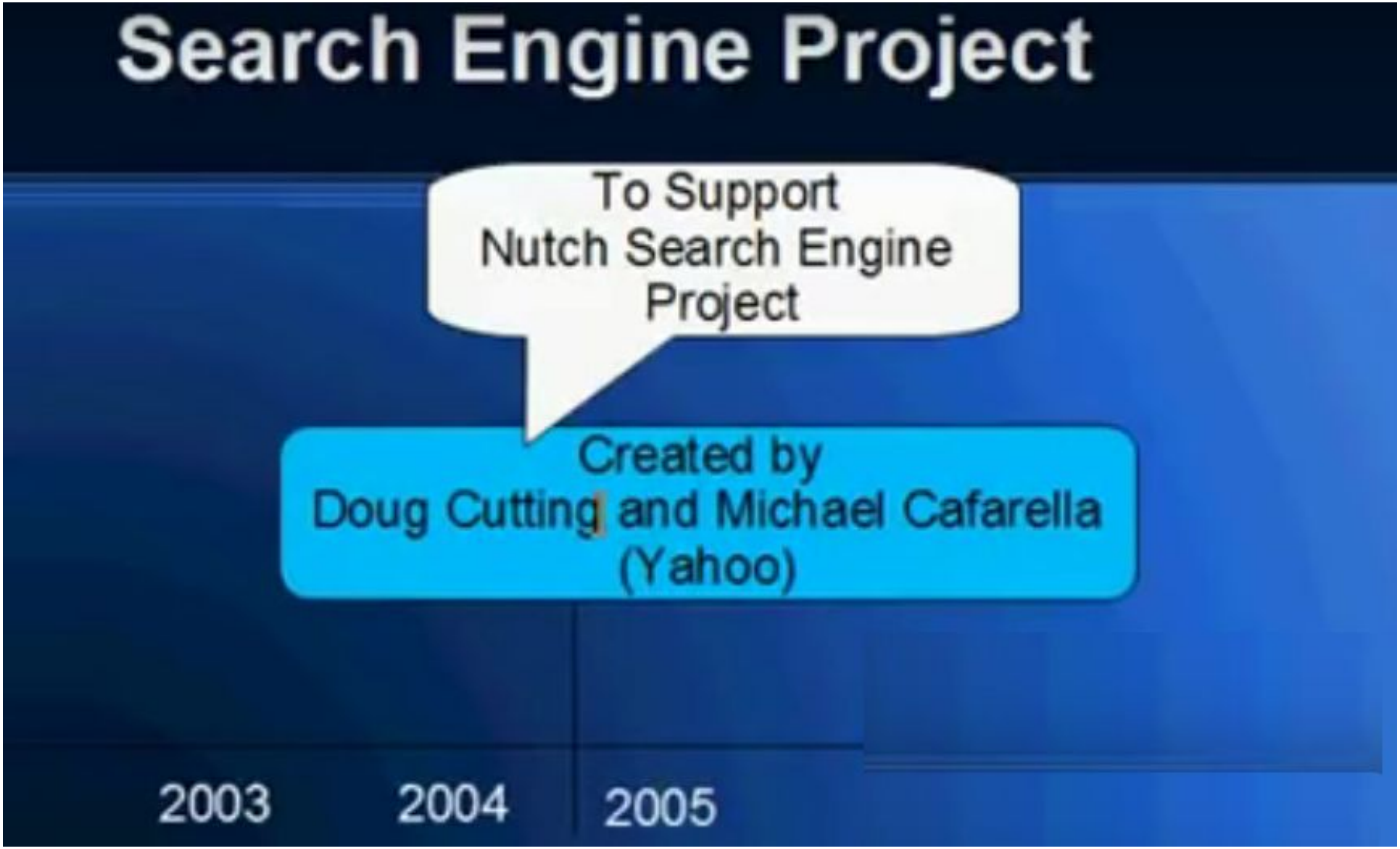
File System
(GFS)



Google released papers on MapReduce

2003

2004

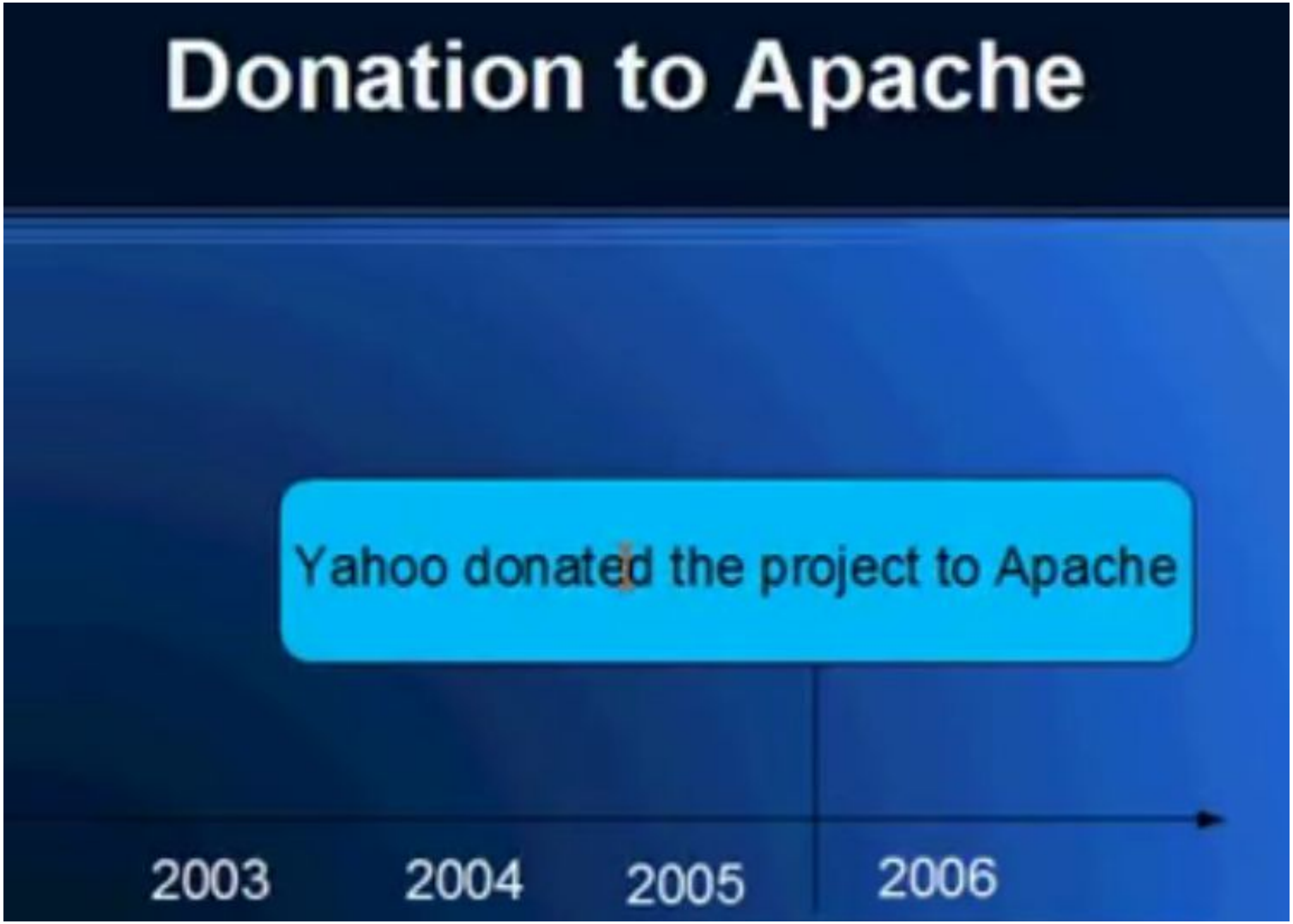


Search Engine Project

To Support
Nutch Search Engine
Project

Created by
Doug Cutting and Michael Cafarella
(Yahoo)

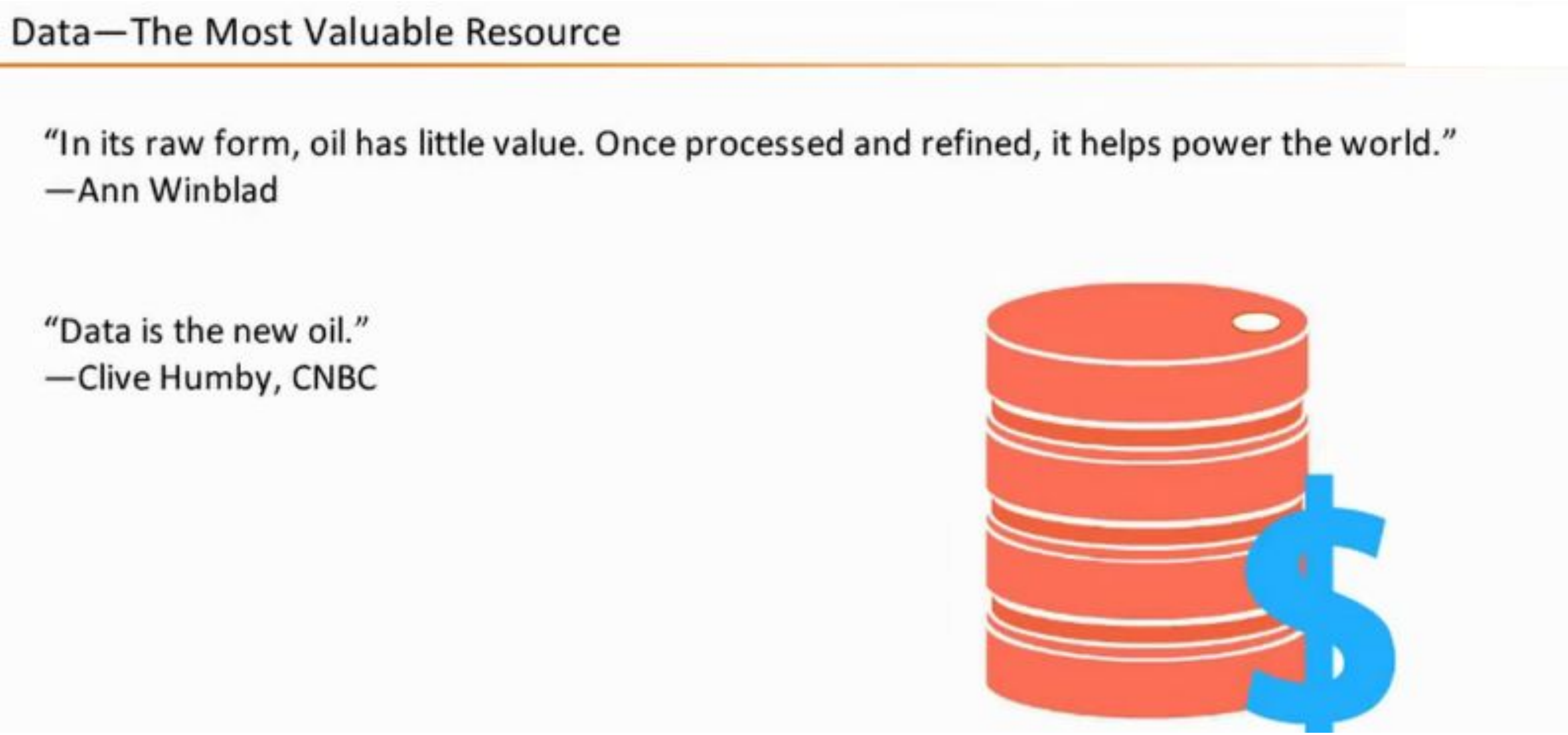
2003 2004 2005



Donation to Apache

Yahoo donated the project to Apache

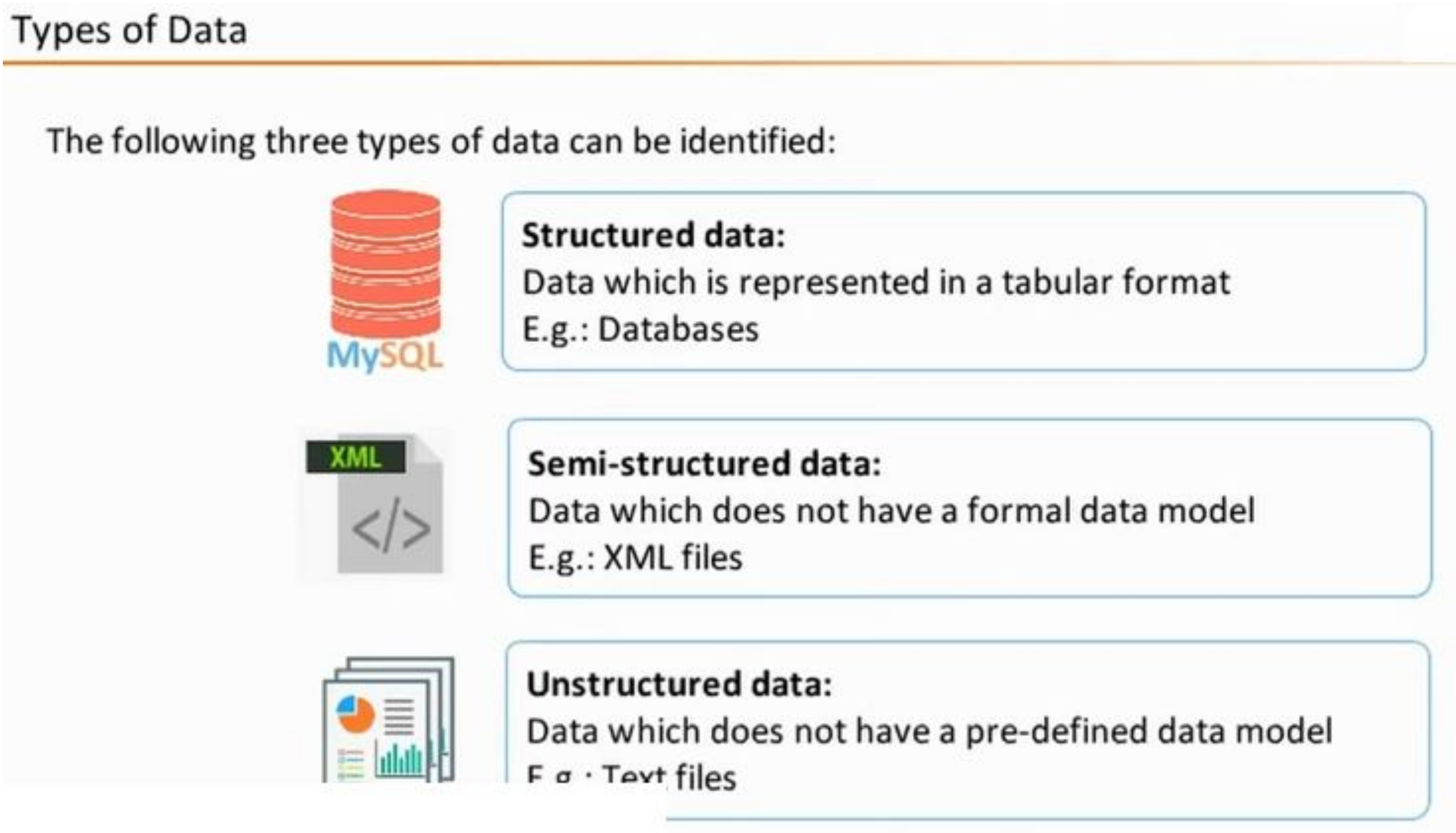
2003 2004 2005 2006



Data—The Most Valuable Resource

"In its raw form, oil has little value. Once processed and refined, it helps power the world."
—Ann Winblad

"Data is the new oil."
—Clive Humby, CNBC



Types of Data

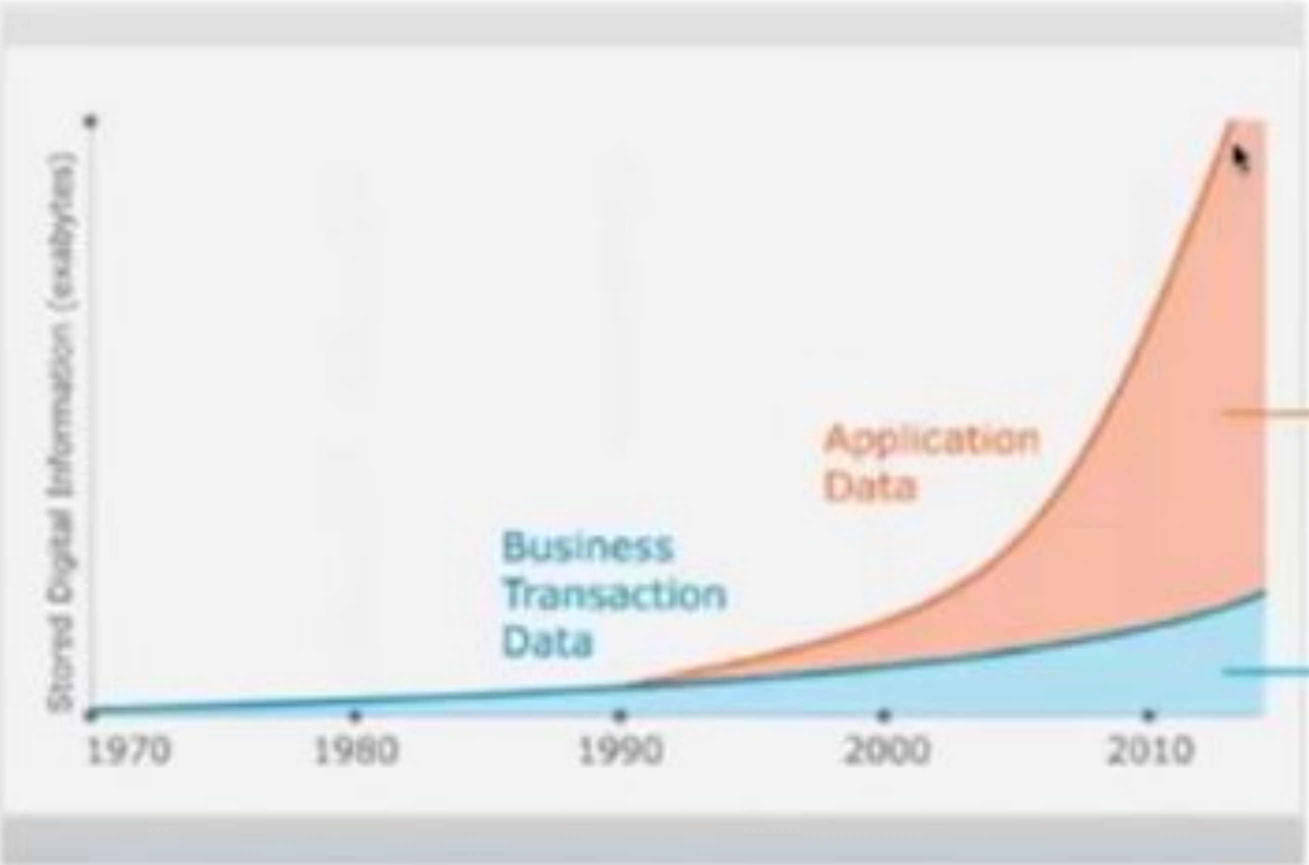
The following three types of data can be identified:

Structured data:
Data which is represented in a tabular format
E.g.: Databases

Semi-structured data:
Data which does not have a formal data model
E.g.: XML files

Unstructured data:
Data which does not have a pre-defined data model
E.g.: Text files

Un-Structured Data is Exploding



Complex, Unstructured

Relational

- 2,500 exabytes of new information in 2012 with Internet as primary driver
- Digital universe grew by 62% last year to 800K petabytes and will grow to 1.2 "zettabytes" this year

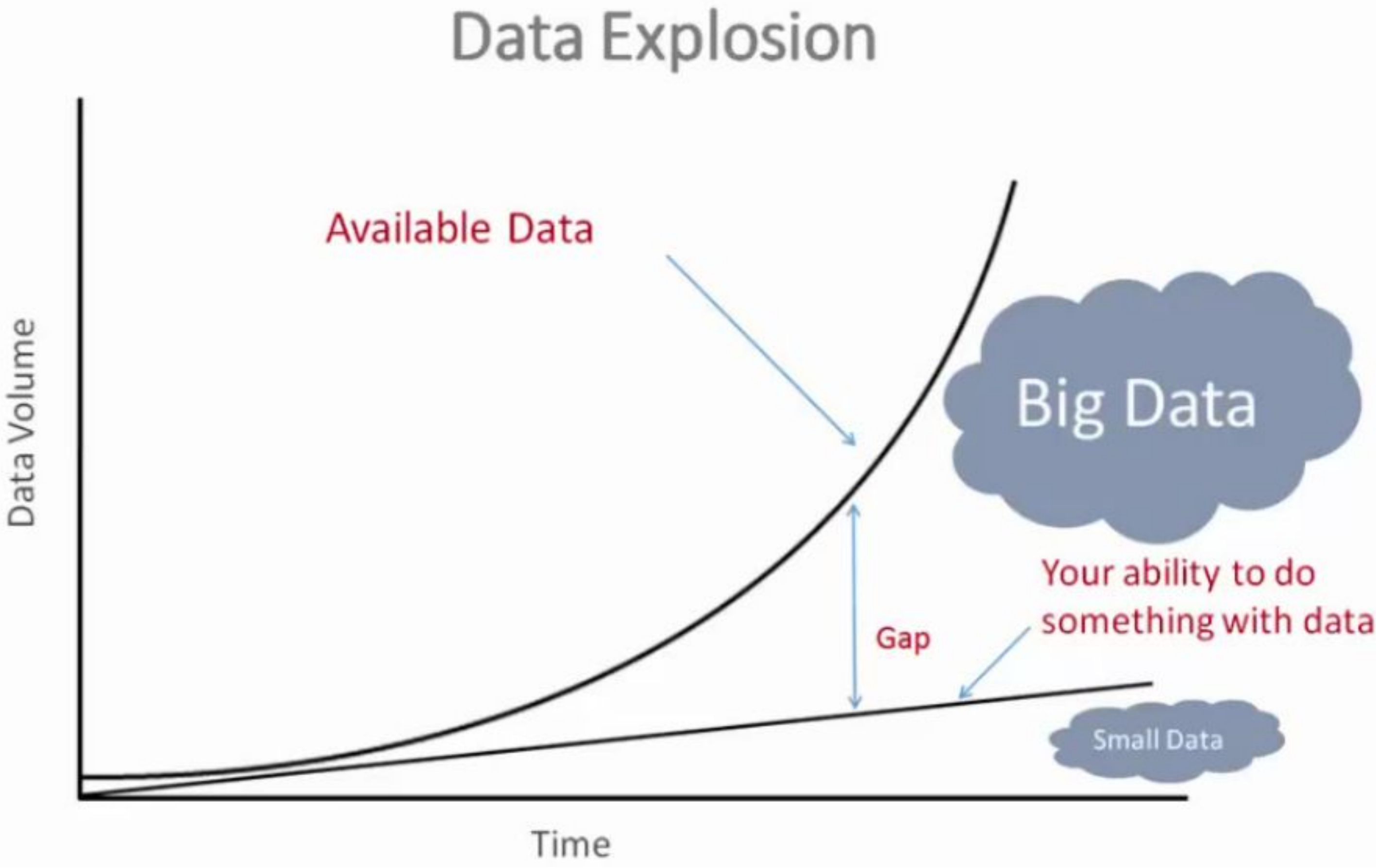
The Data Explosion

2.5 quintillion bytes (2.3 Trillion GB) of data is created every day

----

90% of data in the world was created in the last two years

Source: IBM



How Big Data is Different ?

- Automatically generated by Sensors
- Huge volumes generated by Social Media
- Generated by IOT
- Continuous stream of data

NYSE generates about one terabyte of new trade data per day to Perform stock trading analytics to determine trends for optimal trades.

Problems with Conventional Approaches

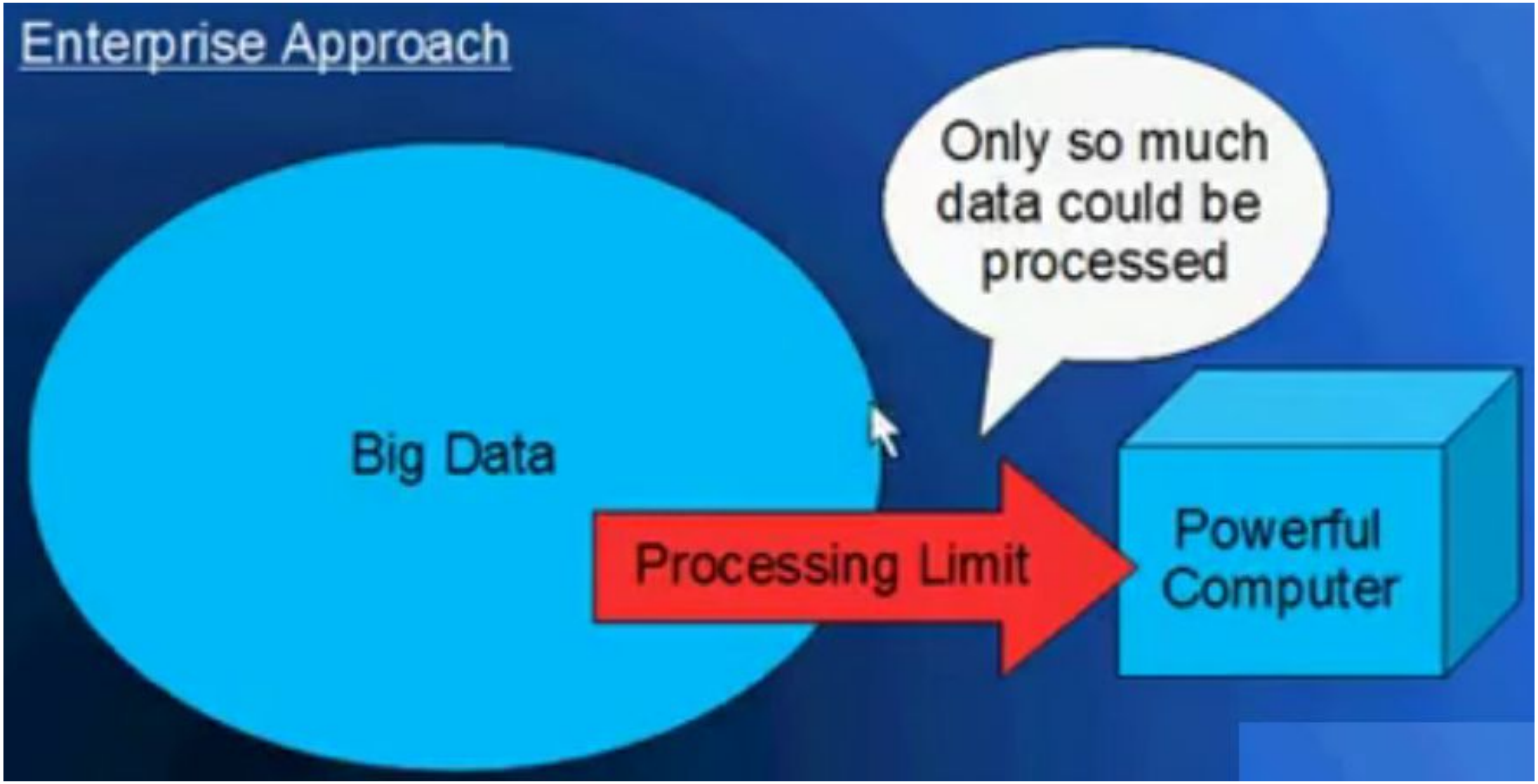
- Limited Storage Capacity
- Limited Processing Capacity
- No Scalability
- Single Point of Failure
- Sequential Processing
- RDBMS can Handle Structured Data
- Requires Pre-processing of Data

Companies Using Hadoop

- Facebook
- Yahoo
- Amazon
- eBay
- American Airlines
- The New York Times
- Federal Reserve Board
- IBM
- Orbitz

BDA Experiment 1

Non Hadoop Approach



Enterprise Approach

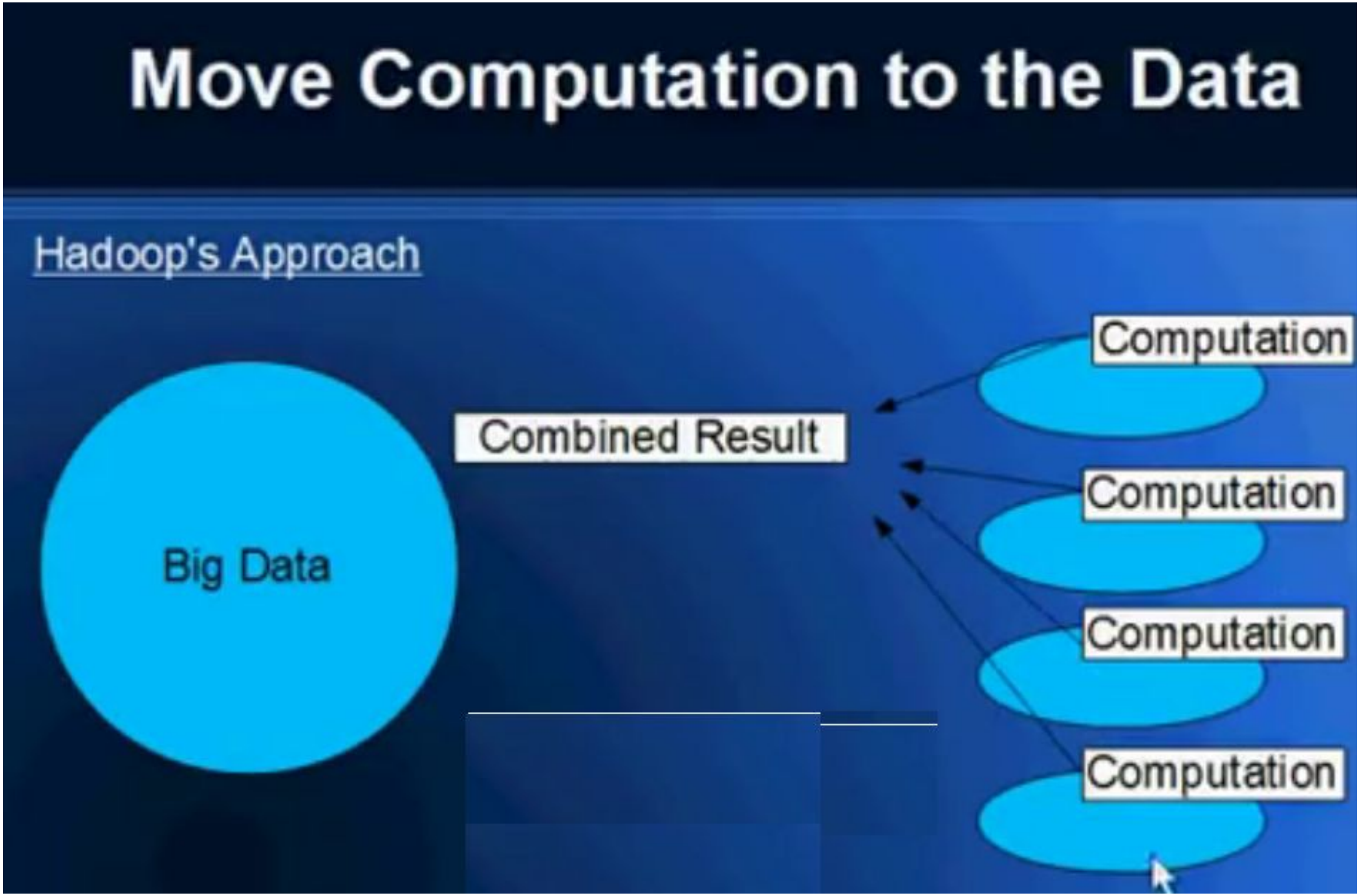
Big Data

Processing Limit

Powerful Computer

Only so much data could be processed

Hadoop Approach



Move Computation to the Data

Hadoop's Approach

Big Data

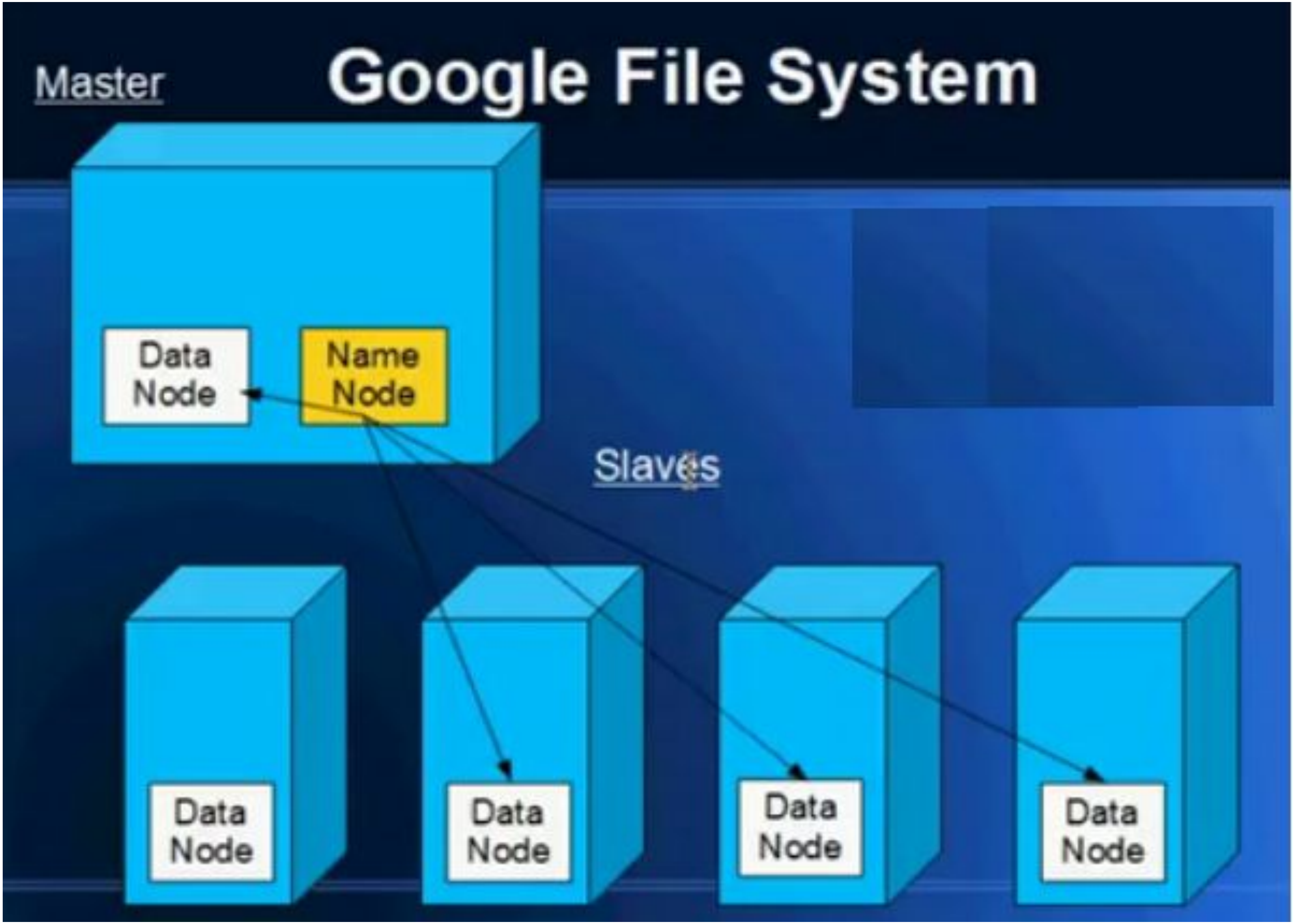
Combined Result

Computation

Computation

Computation

Computation



Google File System

Handling Limitations of Big Data

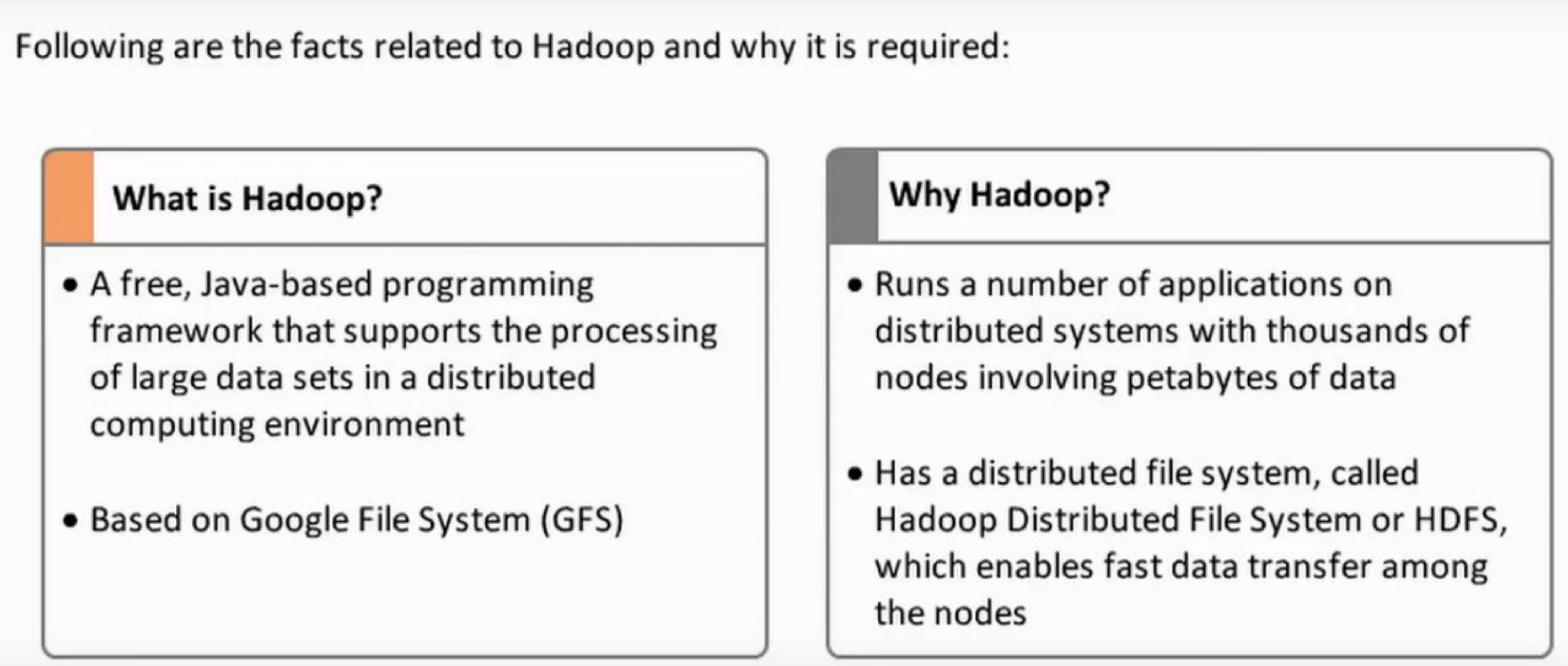
Following are the challenges that need to be addressed by Big Data technology:

How to handle the system uptime and downtime

- Using commodity hardware for data storage and analysis
- Maintaining a copy of the same data across clusters

How to combine data accumulated from all systems

- Analyzing data across different machines
- Merging of data



Following are the facts related to Hadoop and why it is required:

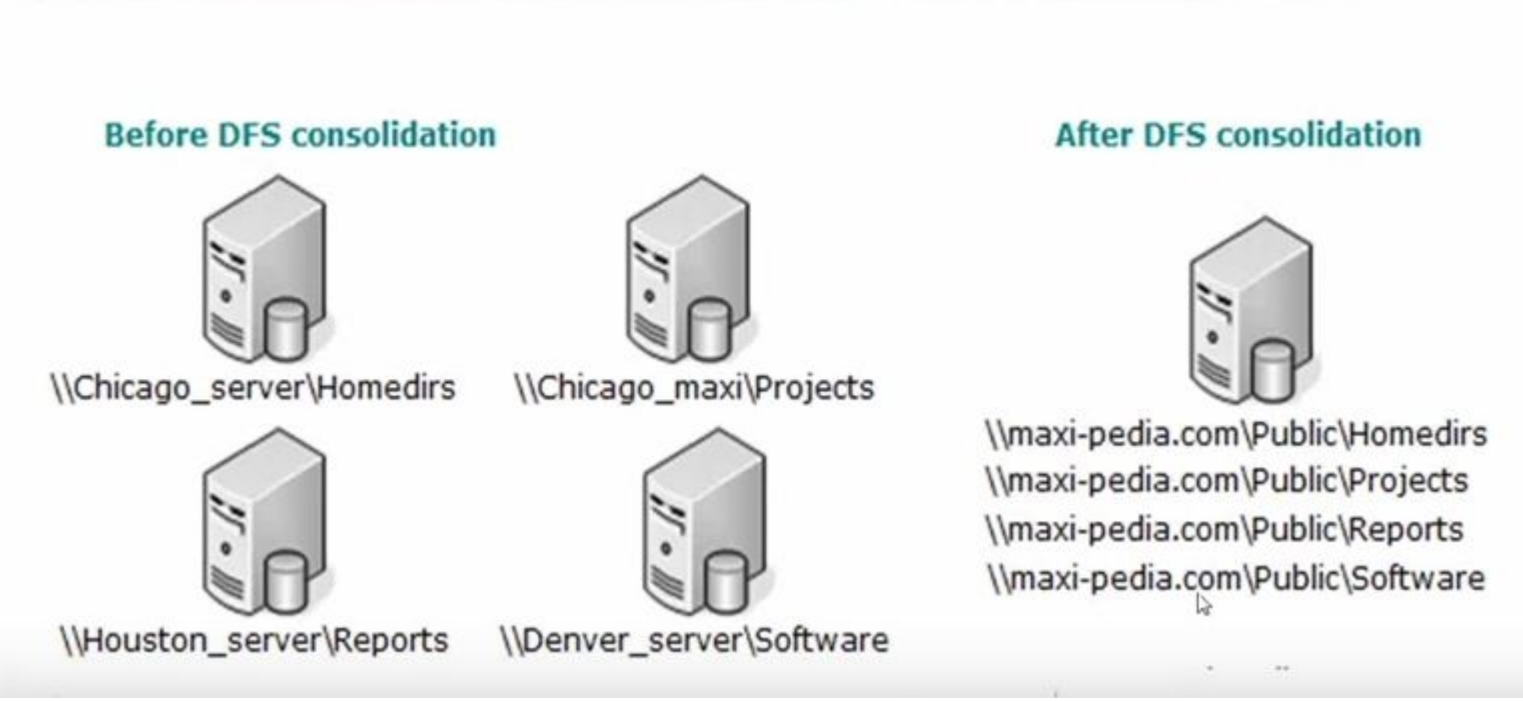
What is Hadoop?

- A free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment
- Based on Google File System (GFS)

Why Hadoop?

- Runs a number of applications on distributed systems with thousands of nodes involving petabytes of data
- Has a distributed file system, called Hadoop Distributed File System or HDFS, which enables fast data transfer among the nodes

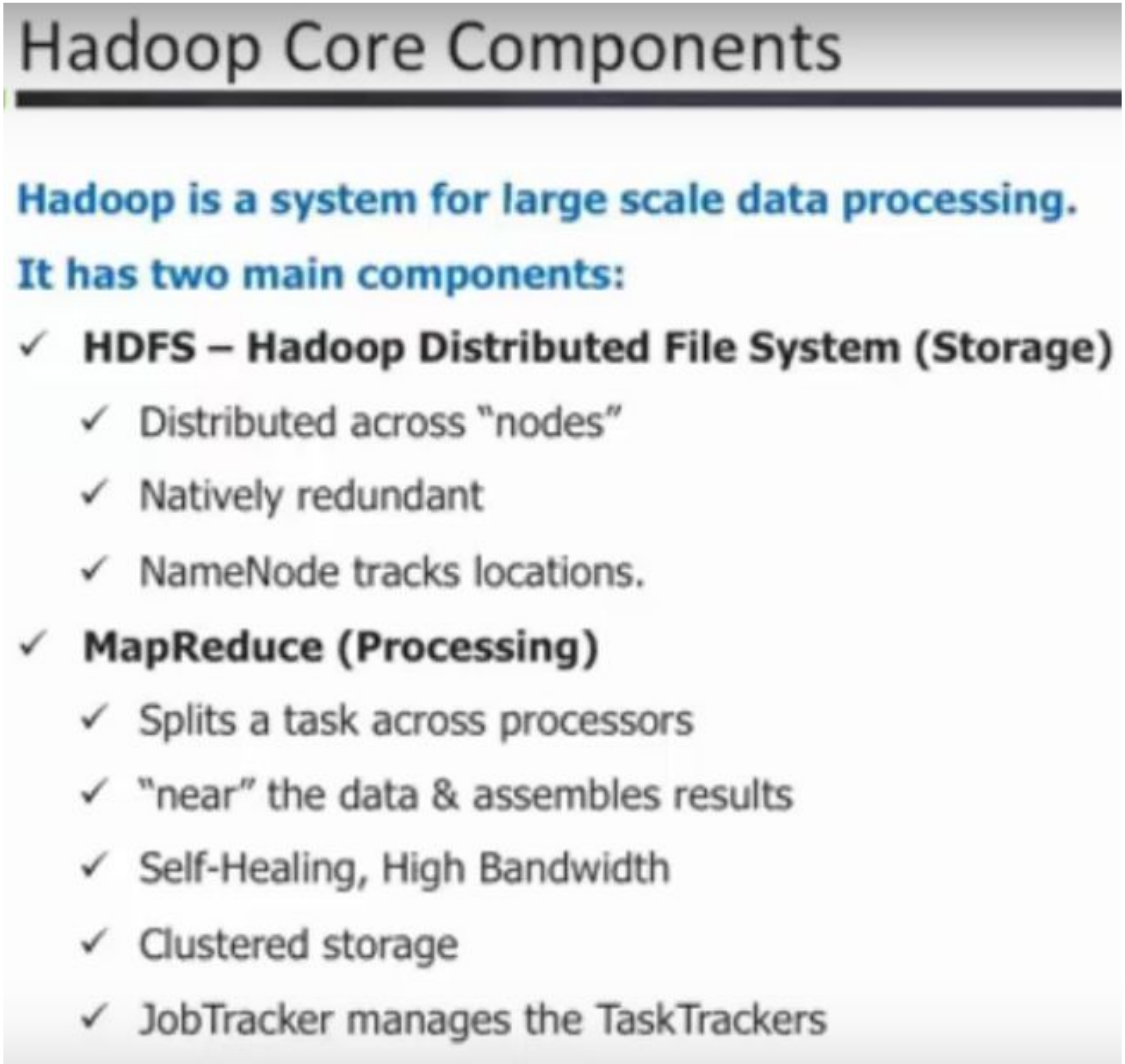
What Is Distributed File System? (DFS)



Before DFS consolidation

After DFS consolidation

\\Chicago_server\Homedirs
\\Chicago_maxi\Projects
\\Houston_server\Reports
\\Denver_server\Software
\\maxi-pedia.com\Public\Homedirs
\\maxi-pedia.com\Public\Projects
\\maxi-pedia.com\Public\Reports
\\maxi-pedia.com\Public\Software



Hadoop Core Components

Hadoop is a system for large scale data processing.

It has two main components:

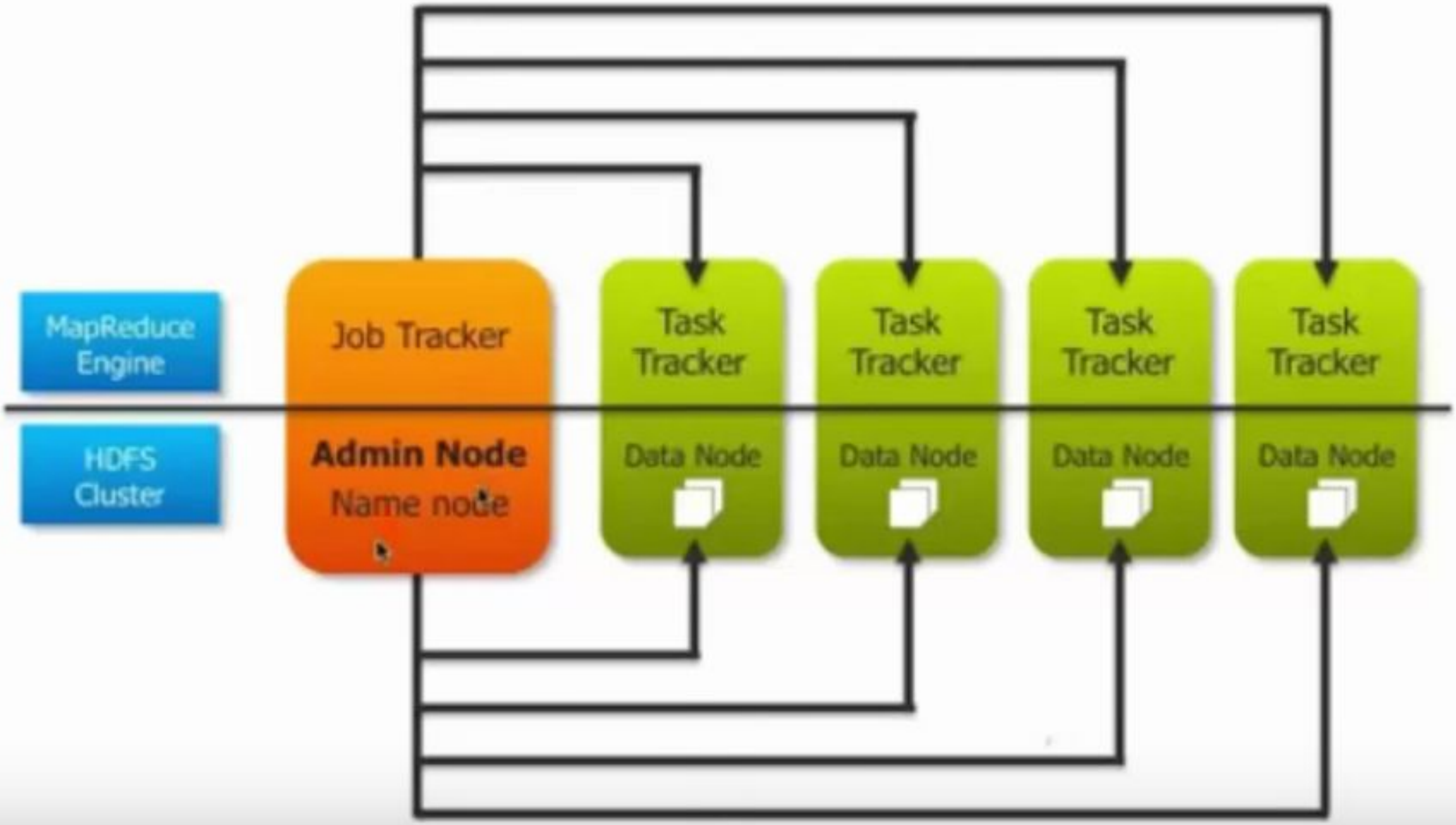
- ✓ **HDFS – Hadoop Distributed File System (Storage)**
  - ✓ Distributed across "nodes"
  - ✓ Natively redundant
  - ✓ NameNode tracks locations.
- ✓ **MapReduce (Processing)**
  - ✓ Splits a task across processors
  - ✓ "near" the data & assembles results
  - ✓ Self-Healing, High Bandwidth
  - ✓ Clustered storage
  - ✓ JobTracker manages the TaskTrackers

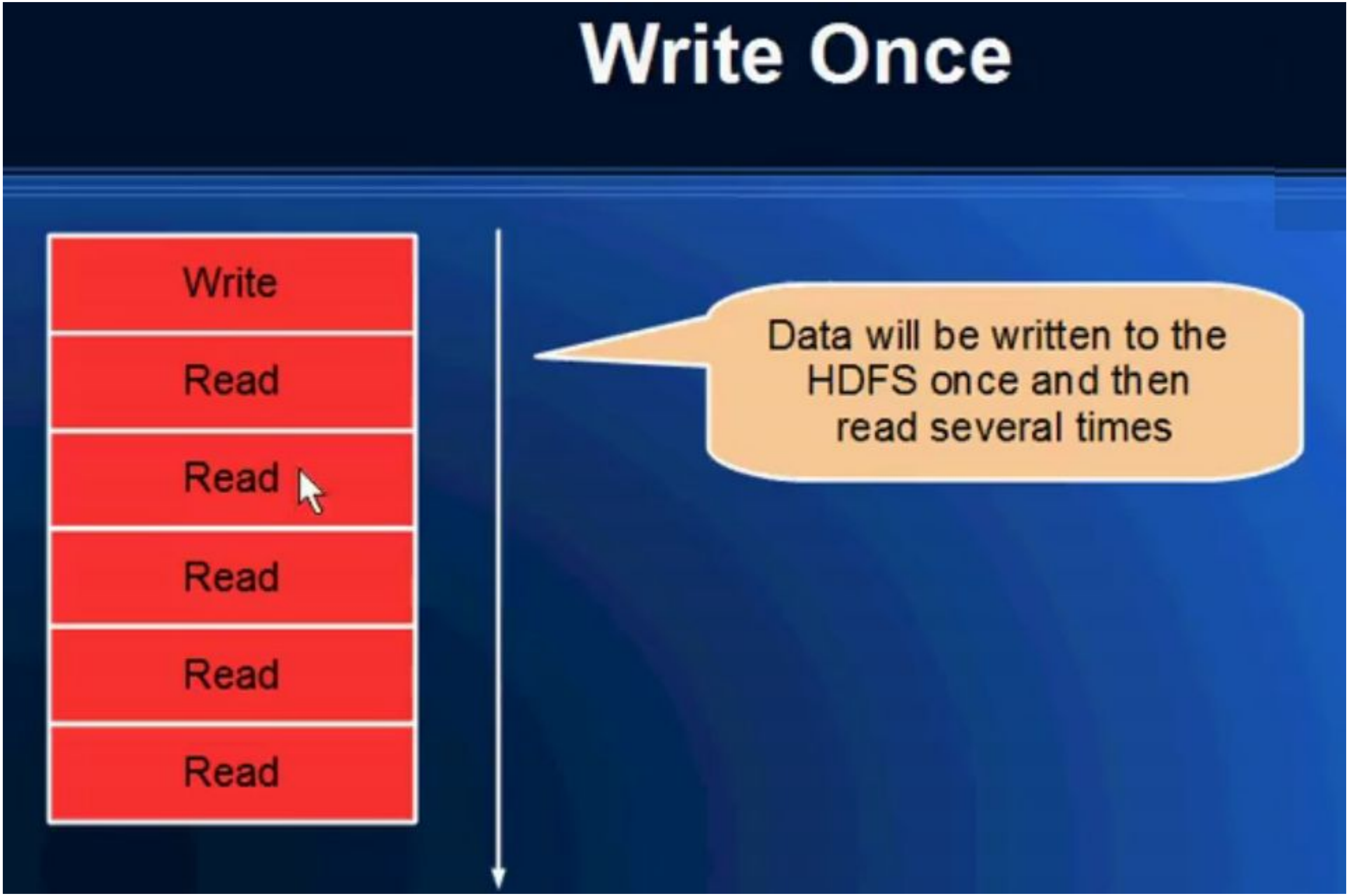
/

NameNode DataNode

JobTracker TaskTracker

Hadoop Core Components (Contd.)





Write Once

Write

Read

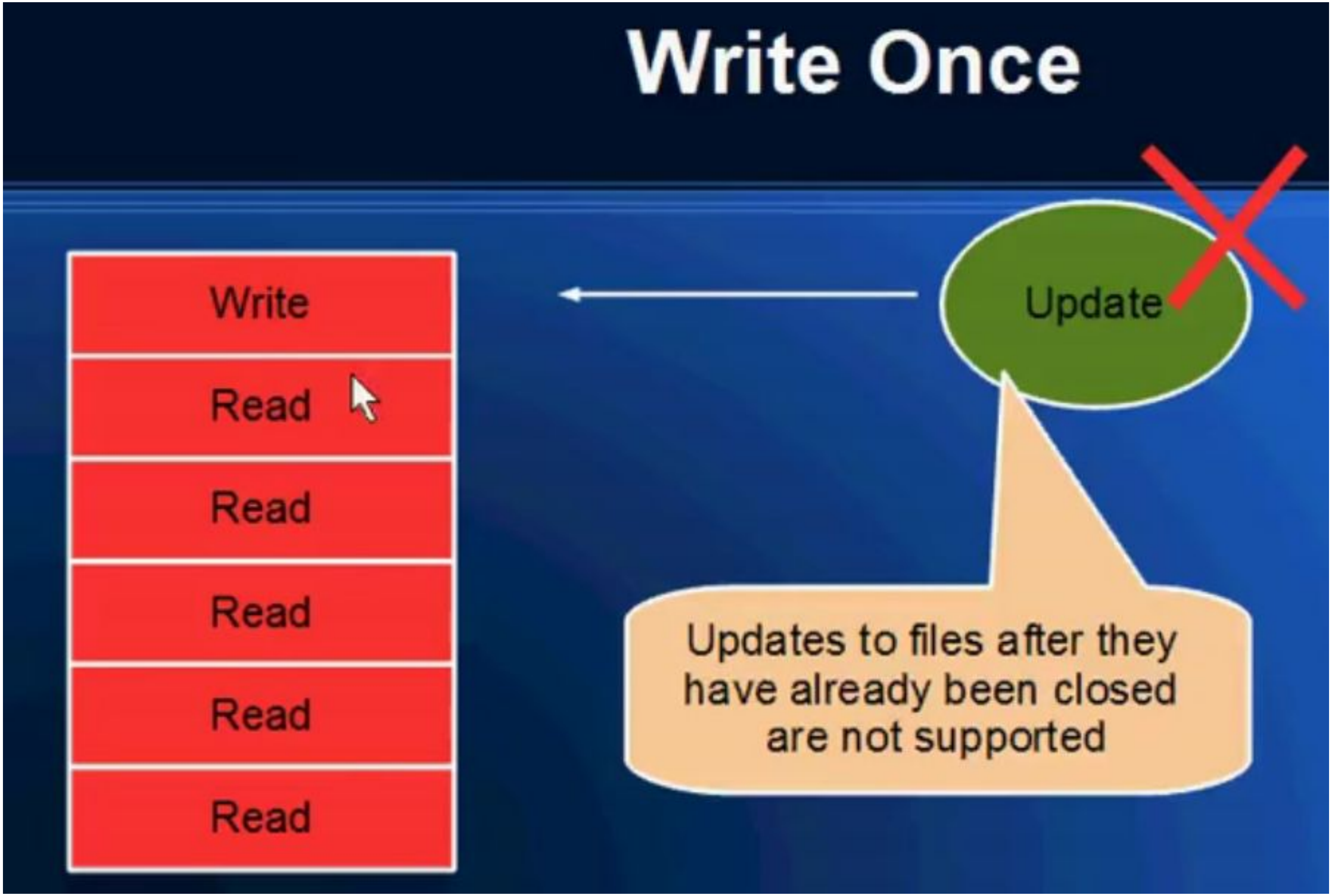
Read

Read

Read

Read

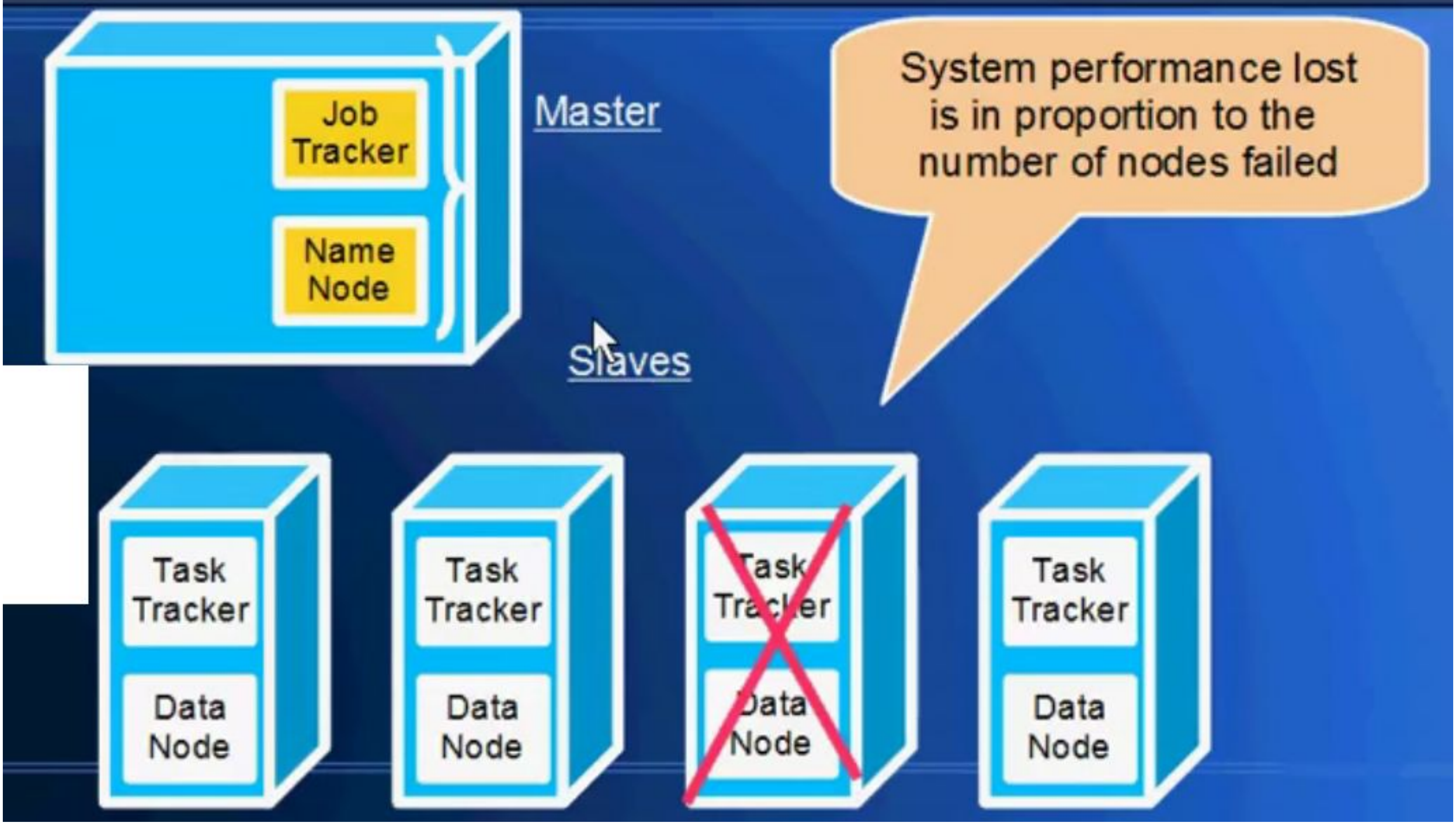
Data will be written to the HDFS once and then read several times



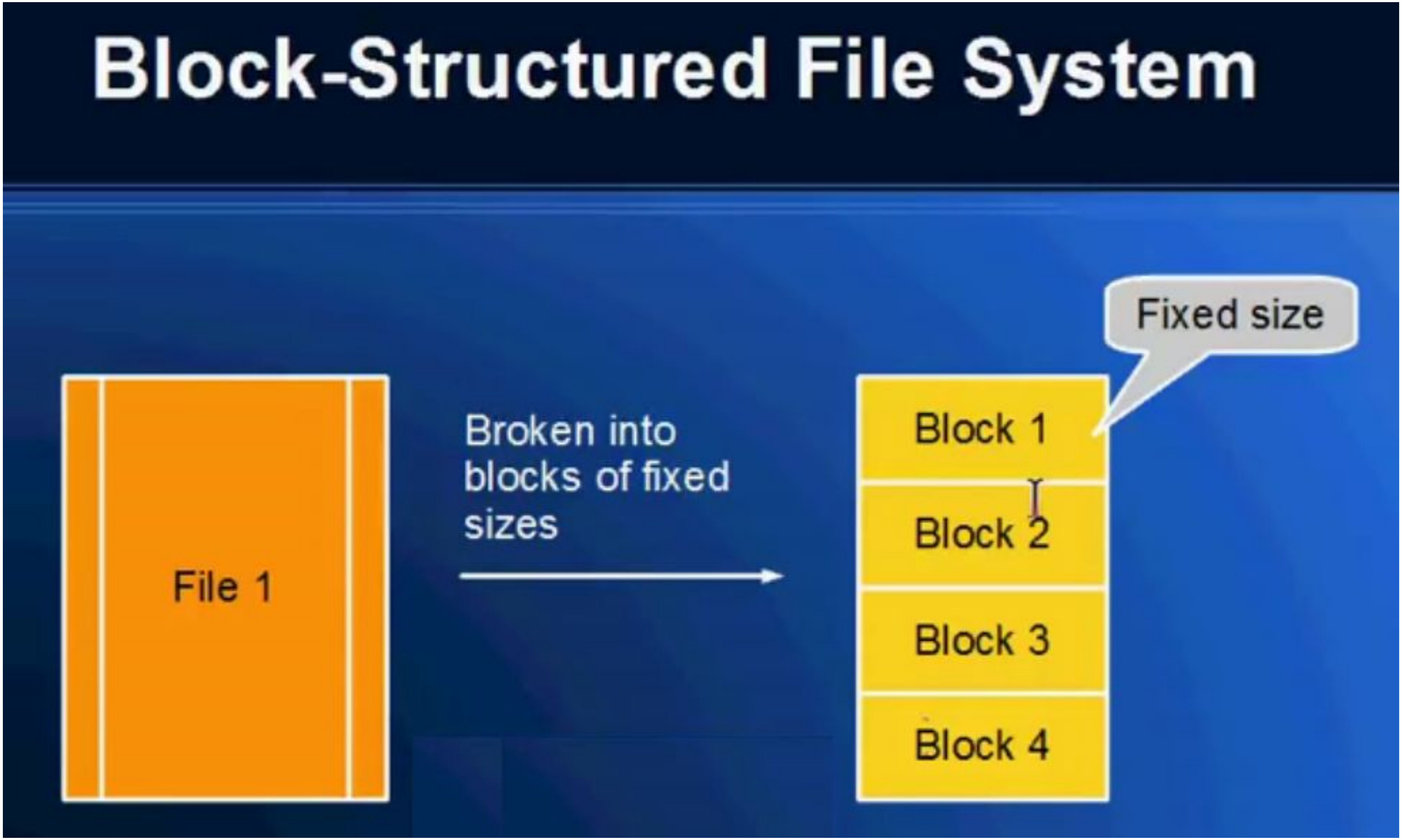
Write Once

Updates to files after they have already been closed are not supported

Performance Lost in Failures



System performance lost is in proportion to the number of nodes failed



Block-Structured File System

File 1

Broken into blocks of fixed sizes

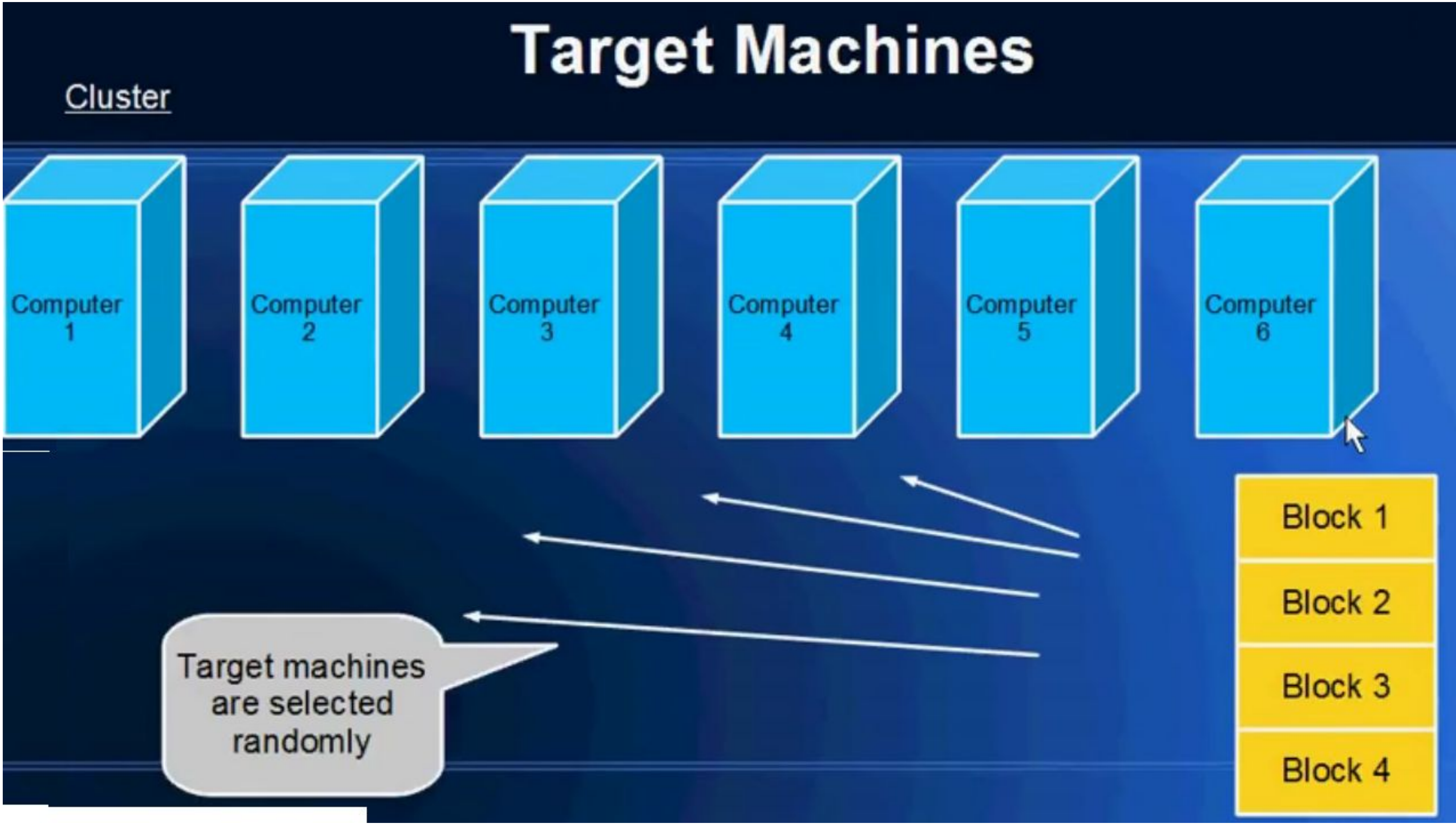
Block 1

Block 2

Block 3

Block 4

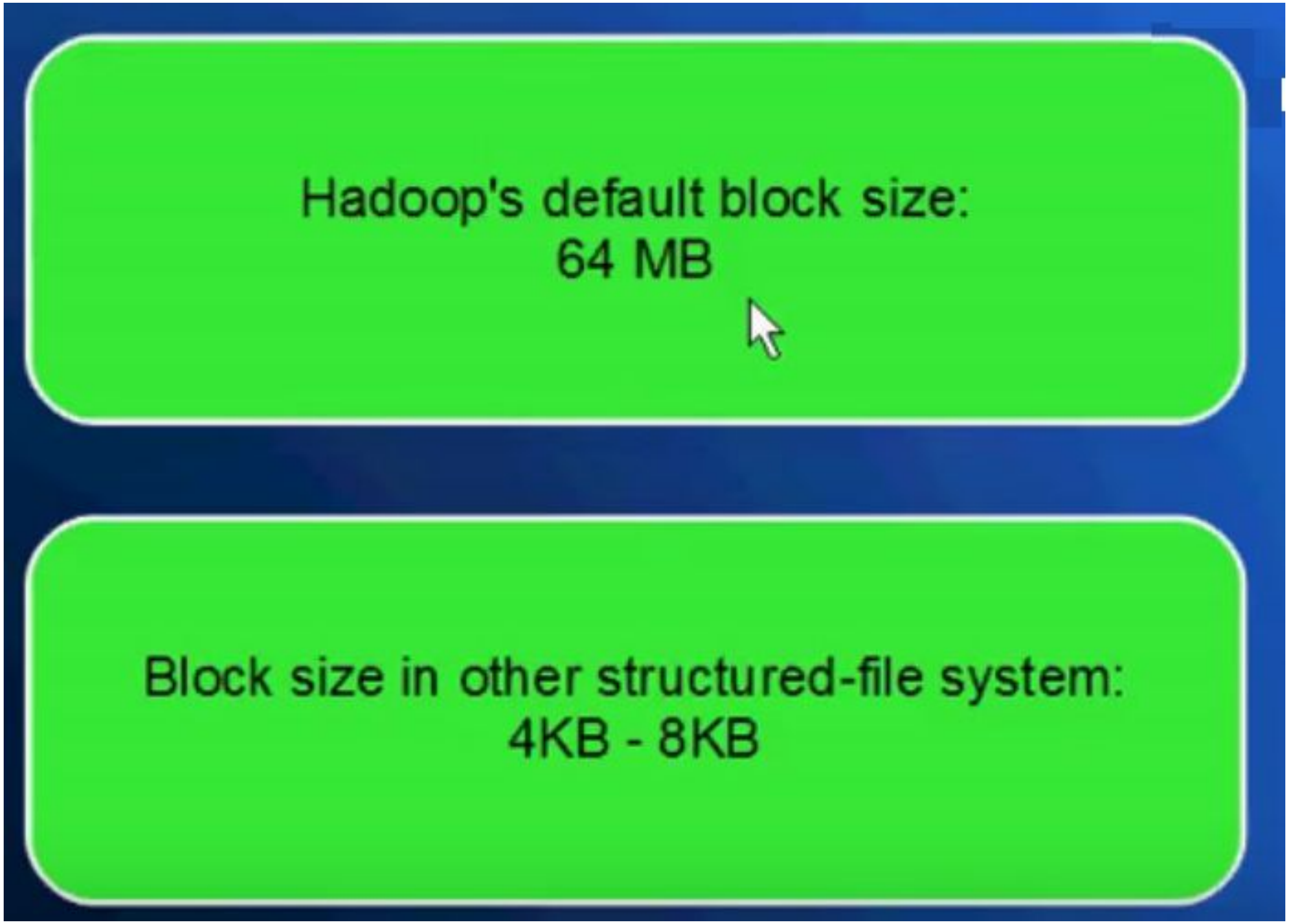
Fixed size



Target Machines

Cluster

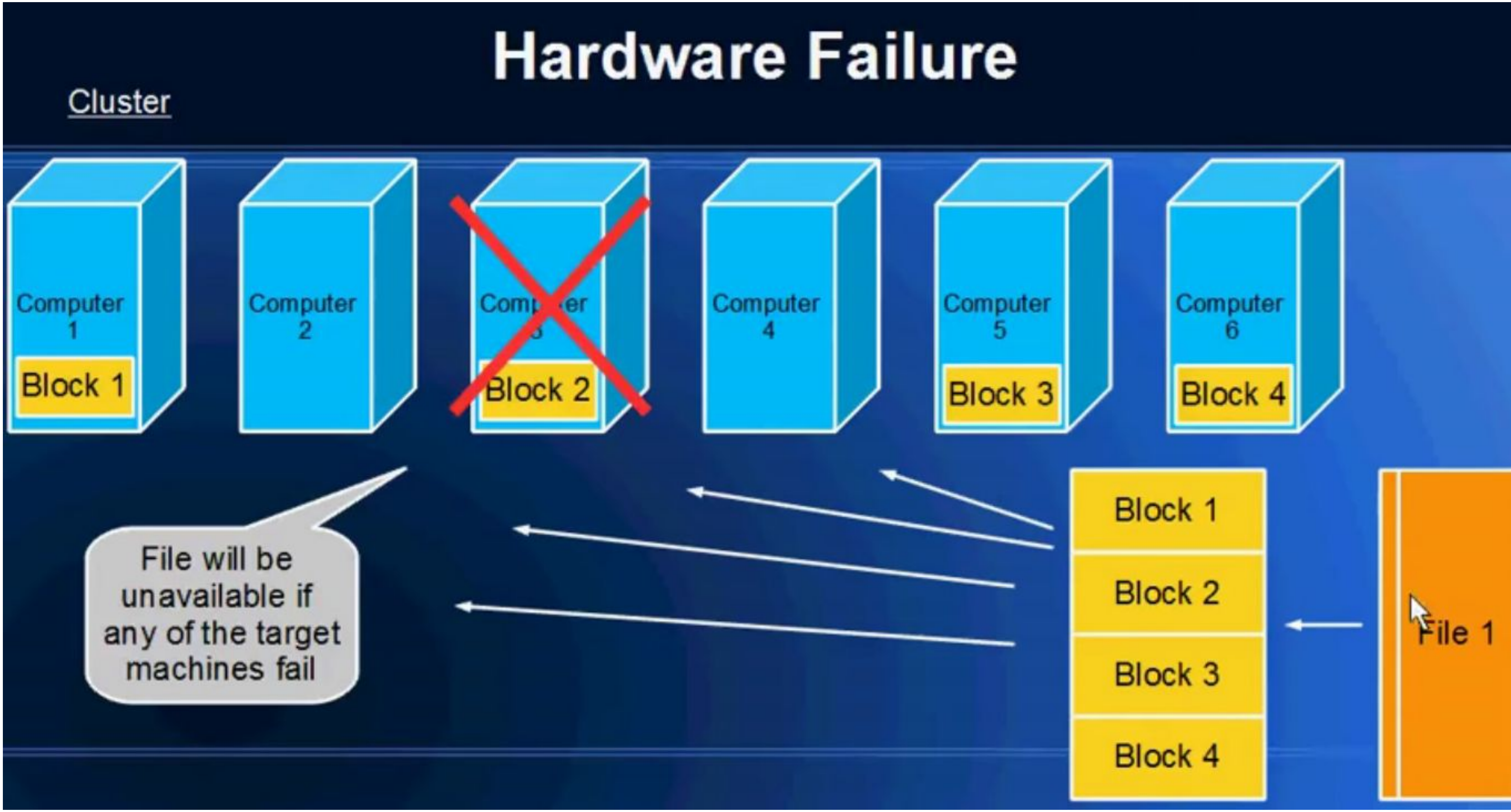
Target machines are selected randomly



Hadoop's default block size:
64 MB

Block size in other structured-file system:
4KB - 8KB

Hardware Failure



Cluster

Computer 1

Block 1

Computer 2

Computer 3

Block 2

Computer 4

Computer 5

Block 3

Computer 6

Block 4

File will be unavailable if any of the target machines fail

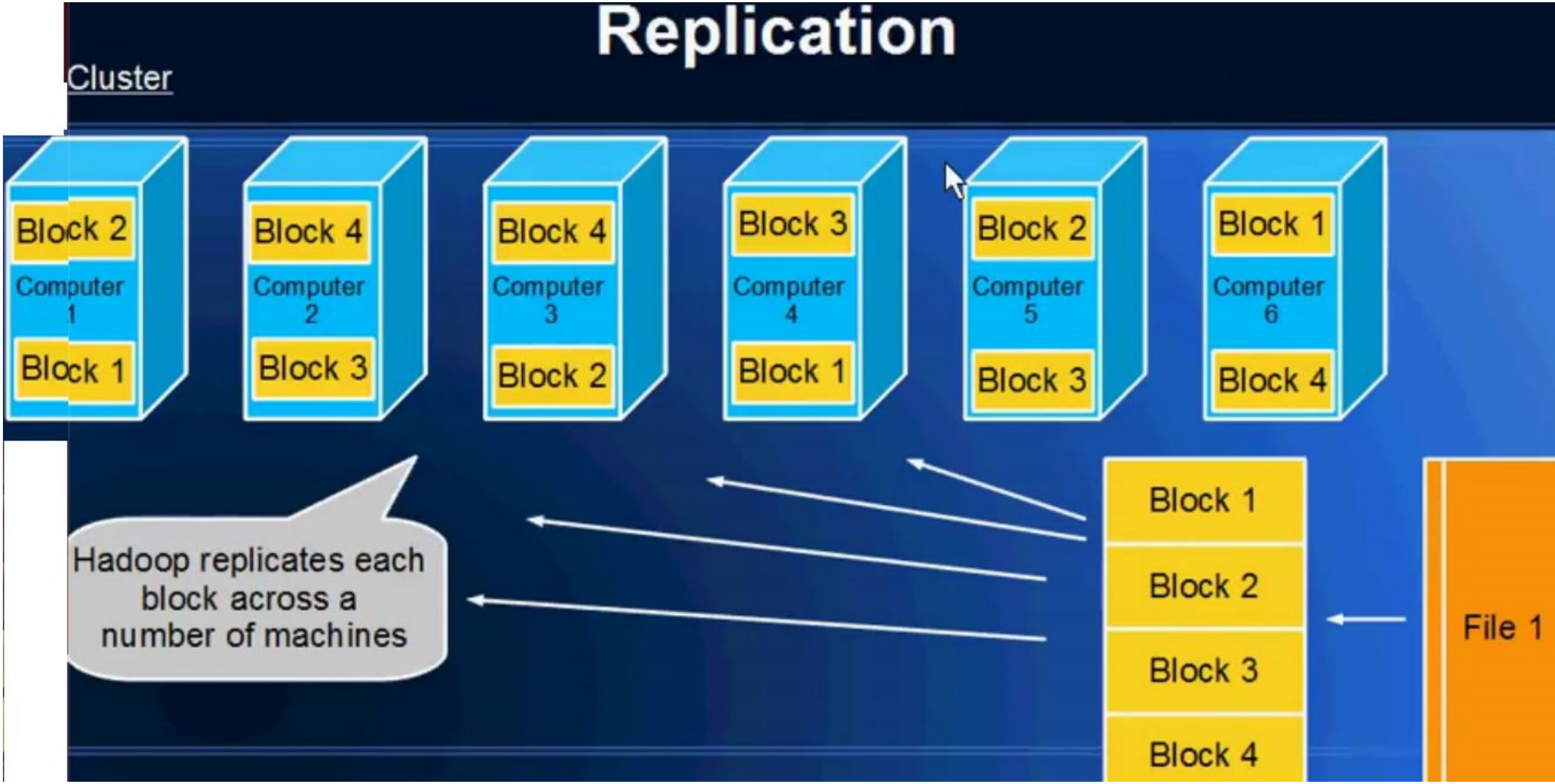
Block 1

Block 2

Block 3

Block 4

File 1



Replication

Cluster

Block 2
Computer 1
Block 1

Block 4
Computer 2
Block 3

Block 4
Computer 3
Block 2

Block 3
Computer 4
Block 1

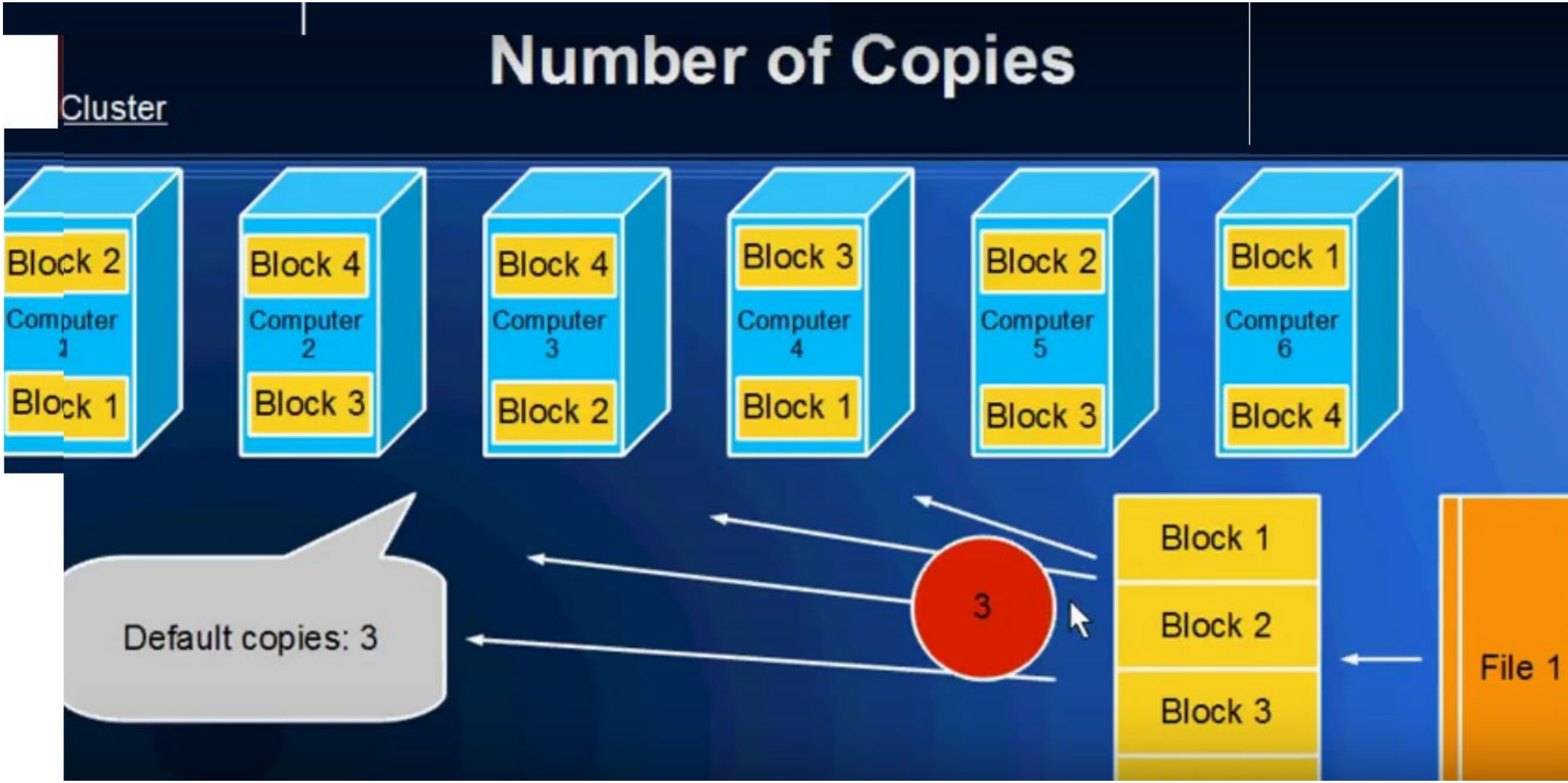
Block 2
Computer 5
Block 3

Block 1
Computer 6
Block 4

Hadoop replicates each block across a number of machines

Block 1
Block 2
Block 3
Block 4

File 1



Cluster

Number of Copies

Block 2
Computer 1
Block 1

Block 4
Computer 2
Block 3

Block 4
Computer 3
Block 2

Block 3
Computer 4
Block 1

Block 2
Computer 5
Block 3

Block 1
Computer 6
Block 4

Default copies: 3

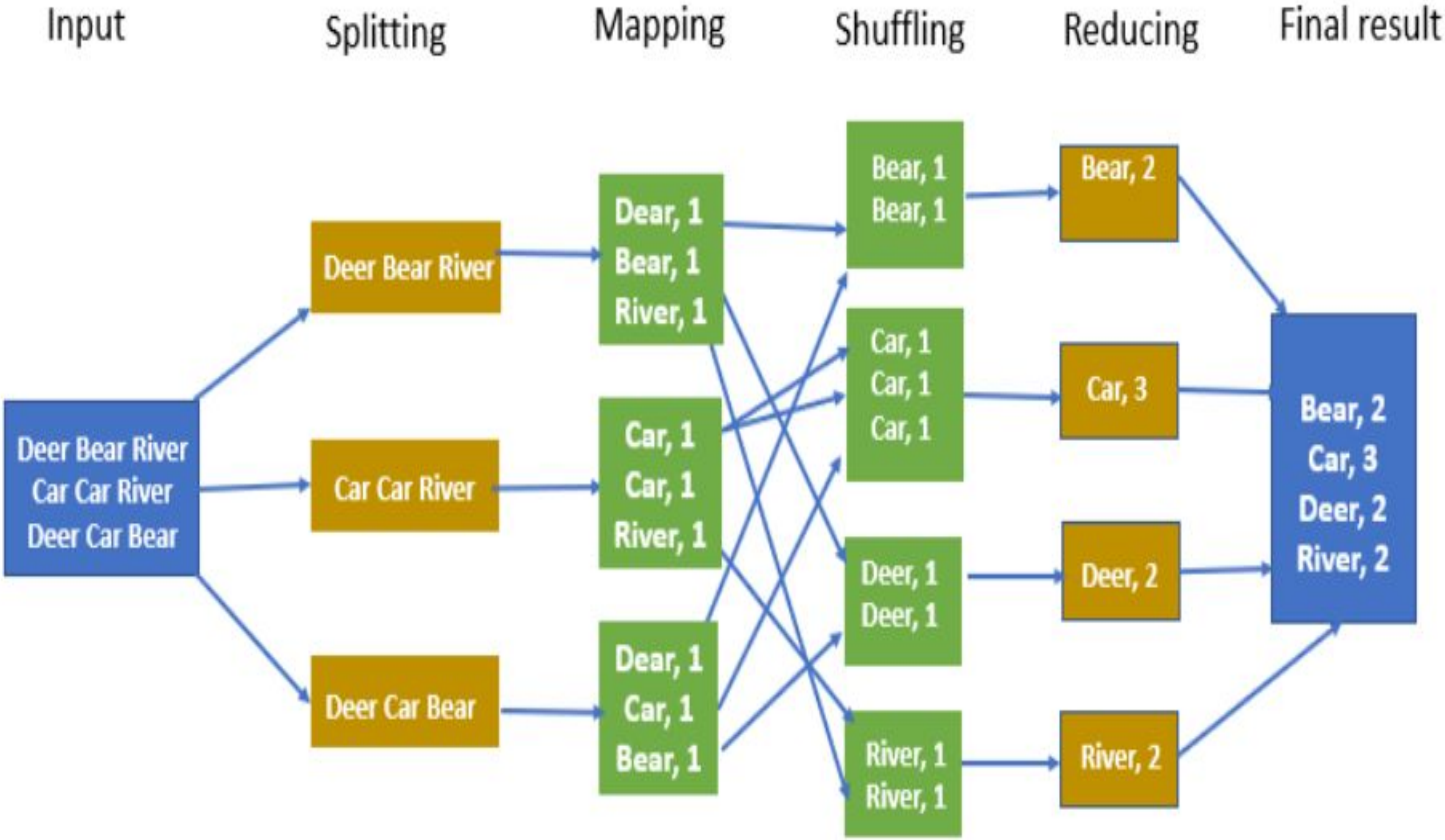
3

Block 1
Block 2
Block 3

File 1

BDA Expt No 4

The overall MapReduce word count process



Input

Splitting

Mapping

Shuffling

Reducing

Final result

Deer Bear River
Car Car River
Deer Car Bear

Deer Bear River

Car Car River

Deer Car Bear

Dear, 1
Bear, 1
River, 1

Car, 1
Car, 1
River, 1

Dear, 1
Car, 1
Bear, 1

Bear, 1
Bear, 1

Car, 1
Car, 1
Car, 1

Deer, 1
Deer, 1

River, 1
River, 1

Bear, 2

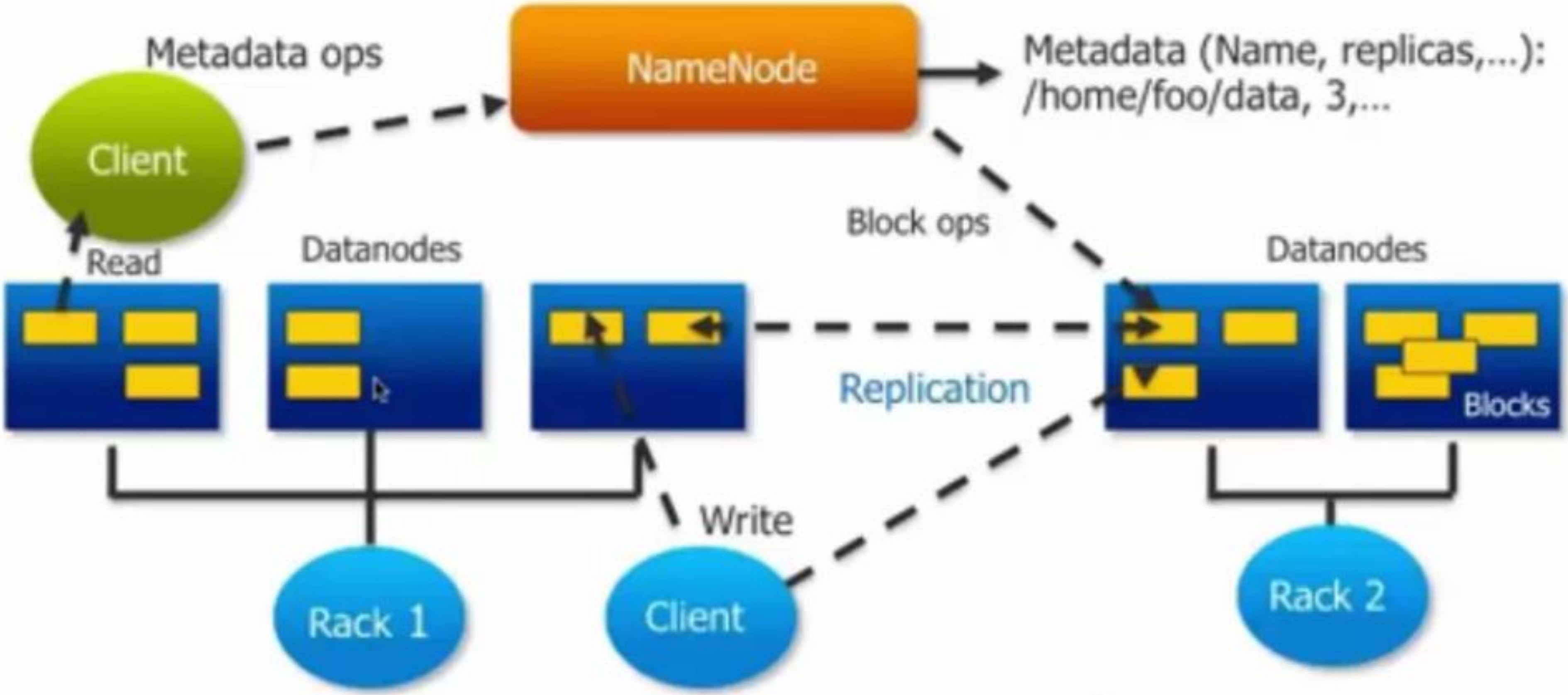
Car, 3

Deer, 2

River, 2

Bear, 2
Car, 3
Deer, 2
River, 2

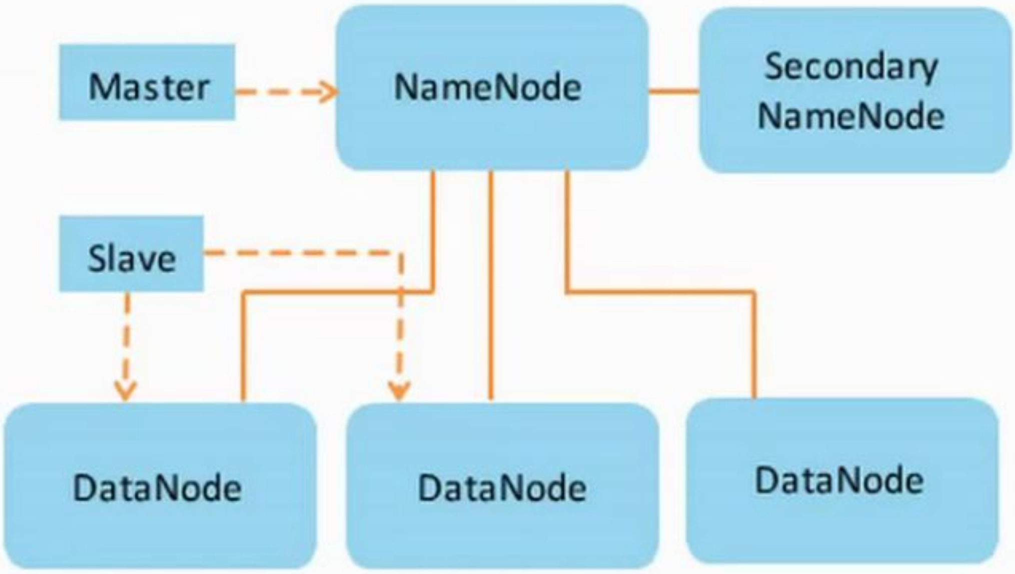
HDFS Architecture



HDFS Architecture

HDFS architecture can be summarized as follows:

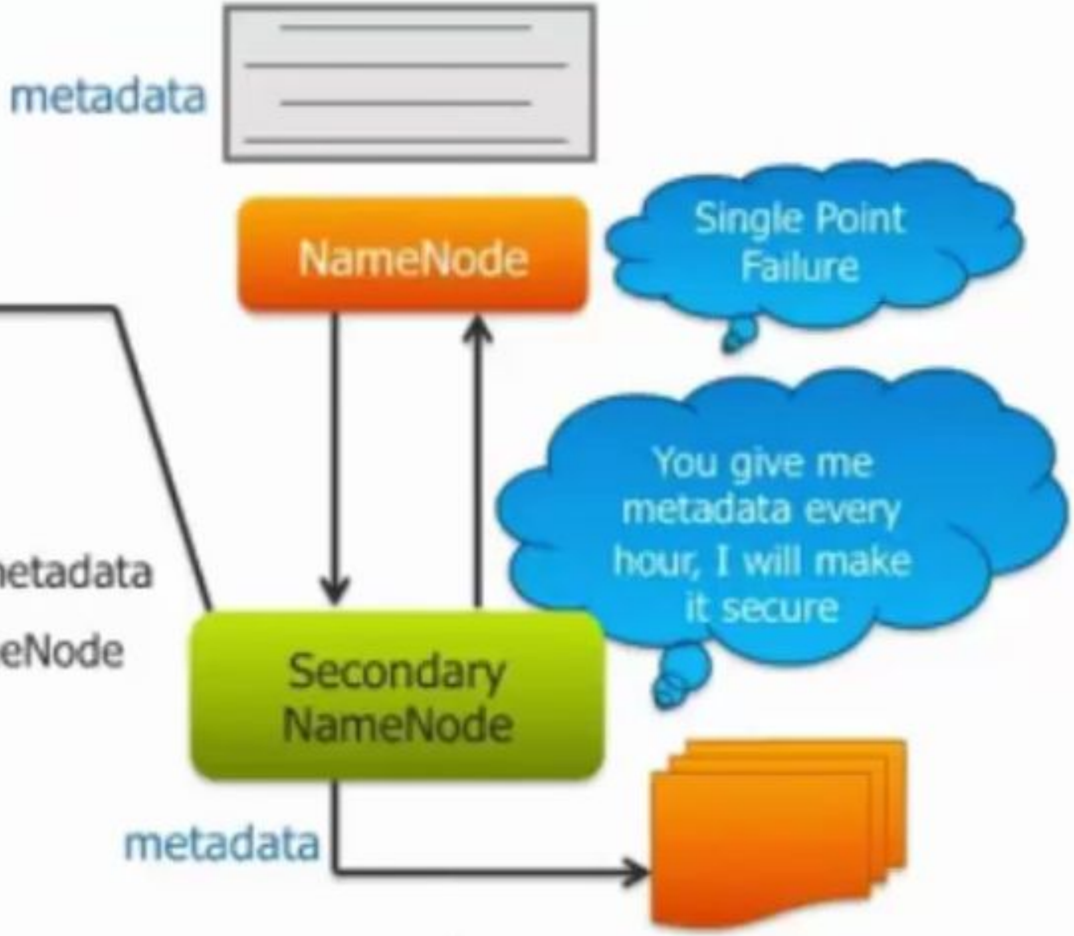
- NameNode and the Secondary NameNode services constitute the master service. DataNode service is the slave service.
- The master service is responsible for accepting a job from clients and ensures that the data required for the operation will be loaded and segregated into chunks of data blocks.
- HDFS exposes a file system namespace and allows user data to be stored in files. A file is split into one or more blocks stored and replicated in DataNodes. The data blocks are then distributed to the DataNode systems within the cluster. This ensures that replicas of the data are maintained.



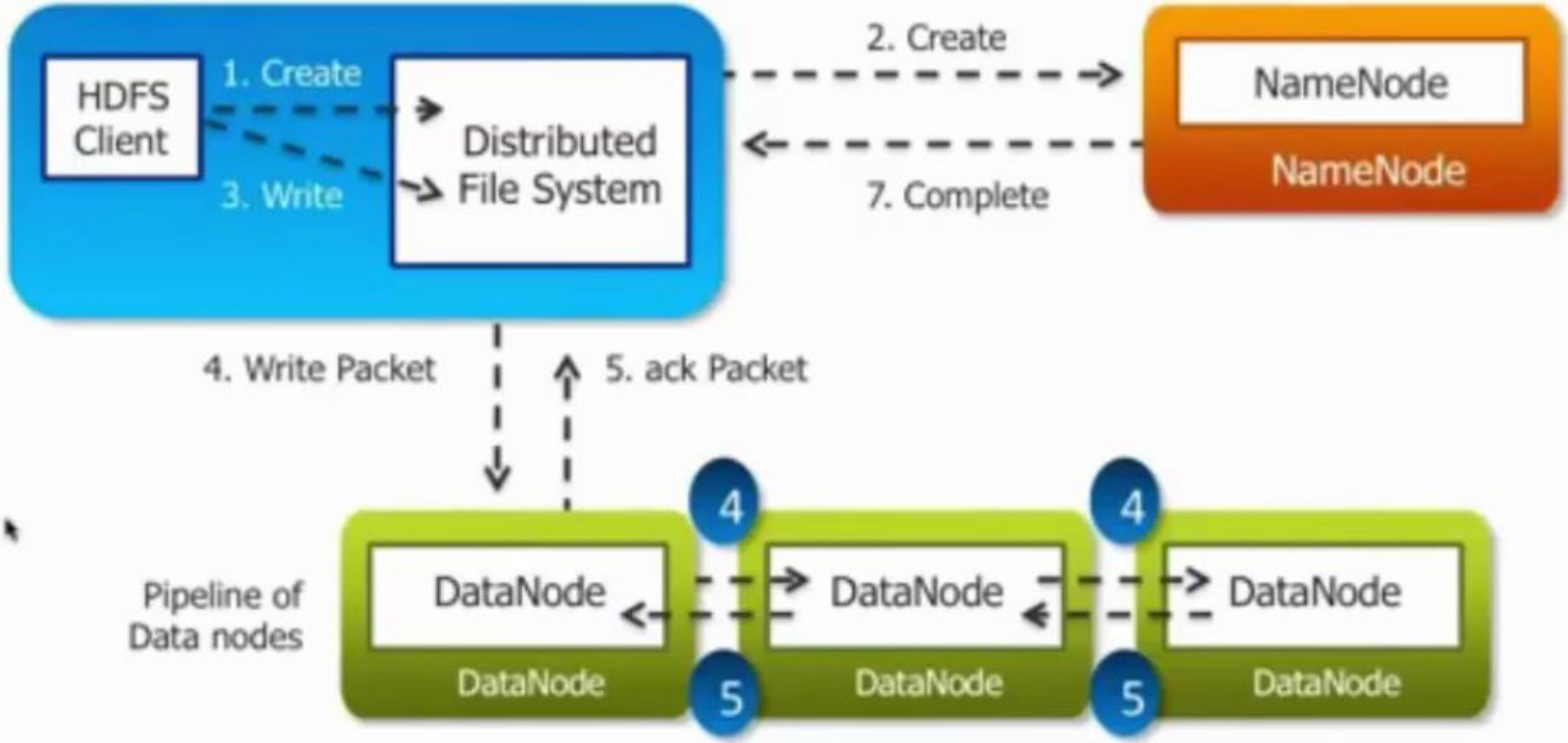
HDFS Cluster

Secondary Name Node

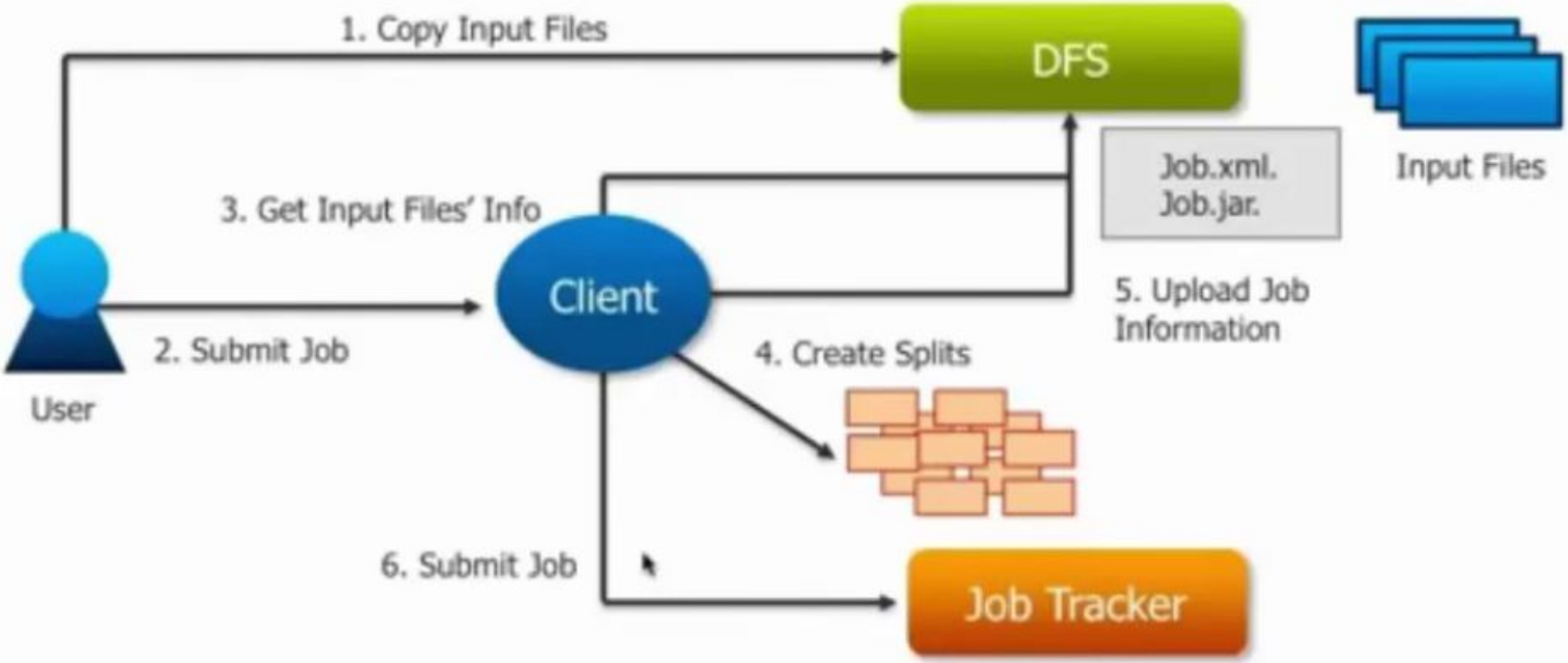
- ✓ Secondary NameNode:
  - ✓ Not a hot standby for the NameNode
  - ✓ Connects to NameNode every hour*
  - ✓ Housekeeping, backup of NemeNode metadata
  - ✓ Saved metadata can build a failed NameNode



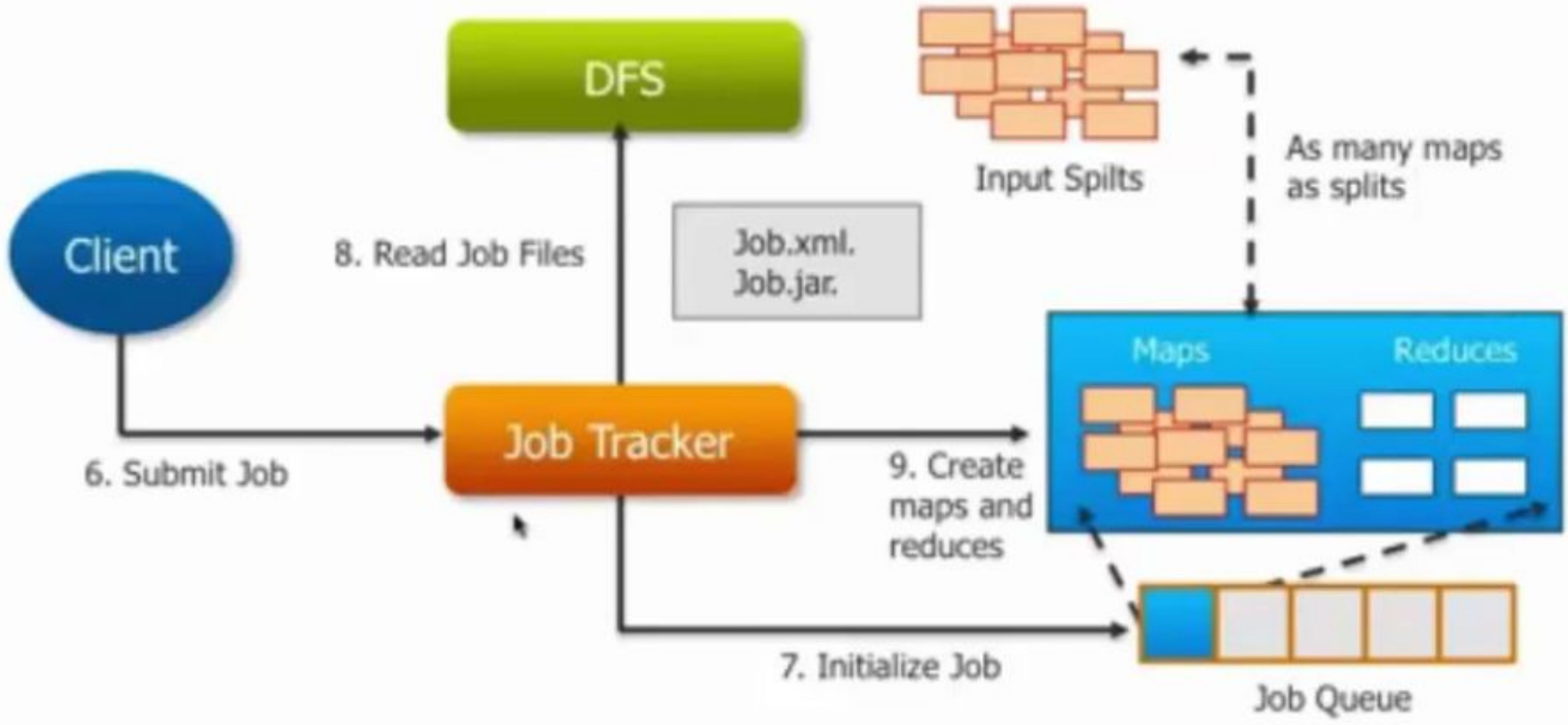
Anatomy of A File Write



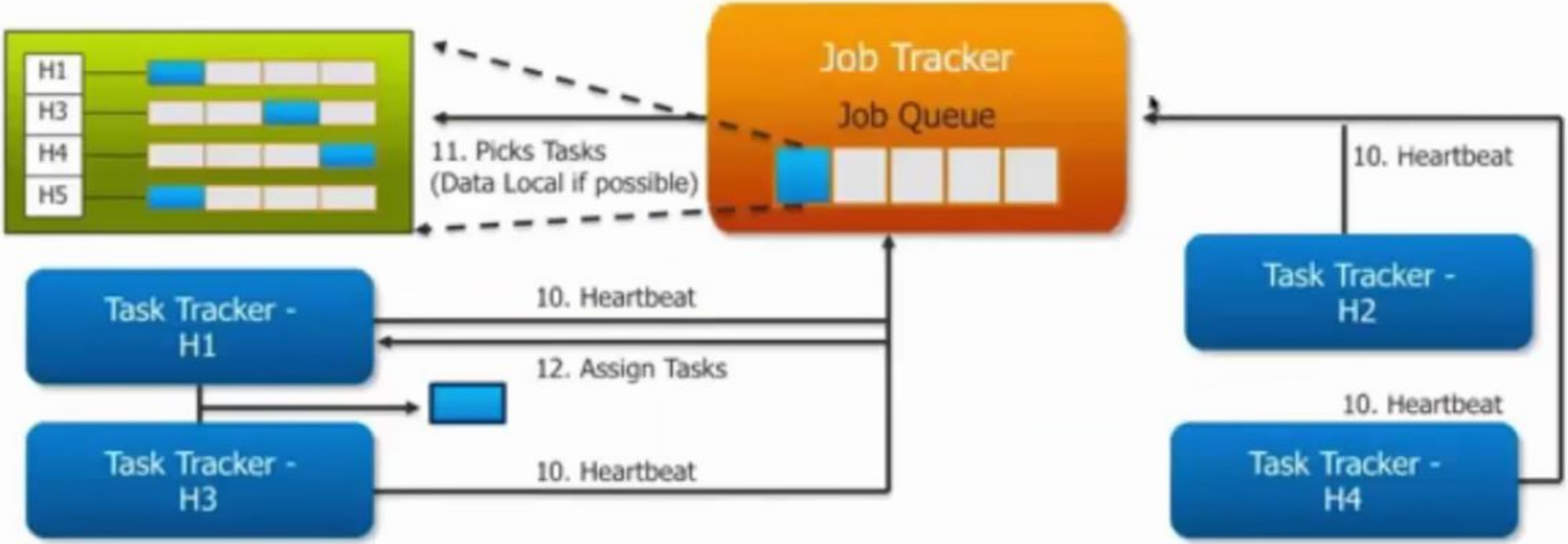
JobTracker


1. Copy Input Files
DFS
Input Files
Job.xml.
Job.jar.
3. Get Input Files' Info
Client
User
2. Submit Job
5. Upload Job Information
4. Create Splits
6. Submit Job
Job Tracker

JobTracker (Contd.)



JobTracker (Contd.)



BDA Experiment 3

Hadoop Installation

Installing Hadoop

Hadoop runs on Linux/Unix based Systems

Hadoop Install Modes

- Standalone
- Pseudo-Distributed
- Fully Distributed

Standalone

Runs on a single node

A single JVM process

Local File System for Storage

HDFS and YARN do not run

Standalone

Used to test MapReduce programs before running them on a cluster

Pseudo-Distributed

Runs on a single node

2 JVM processes to simulate 2 nodes

HDFS for storage

YARN for managing tasks

Pseudo-Distributed

Used as a fully-fledged test environment

Fully Distributed

Manual configuration of a cluster is complicated

Usually use enterprise editions

- Cloudera, MapR, Hortonworks