

# **Hadoop/MapReduce Computing Paradigm**

<http://developer.yahoo.com/hadoop/tutorial/module4.html>

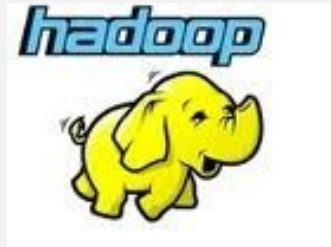
**Spring** 2015

Taken from:

WPI, Mohamed Eltabakh

# Large-Scale Data Analytics

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



vs.



- **Many enterprises are turning to Hadoop**
  - Especially applications generating **big data**
  - Web applications, social networks, scientific applications

# Why Hadoop is able to compete?



VS.



Scalability (petabytes of data, thousands of machines)



Flexibility in accepting all data formats (no schema)



Efficient and simple fault-tolerant mechanism



Commodity inexpensive hardware



Performance (tons of indexing, tuning, data organization tech.)



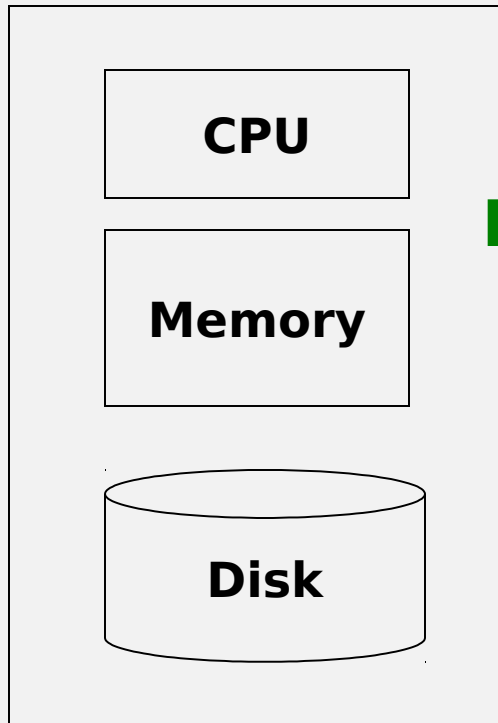
Features:

- Provenance tracking
- Annotation management
- ....

# 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*

# Single Node Architecture



**Machine Learning, Statistics**

**“Classical” Data Mining**

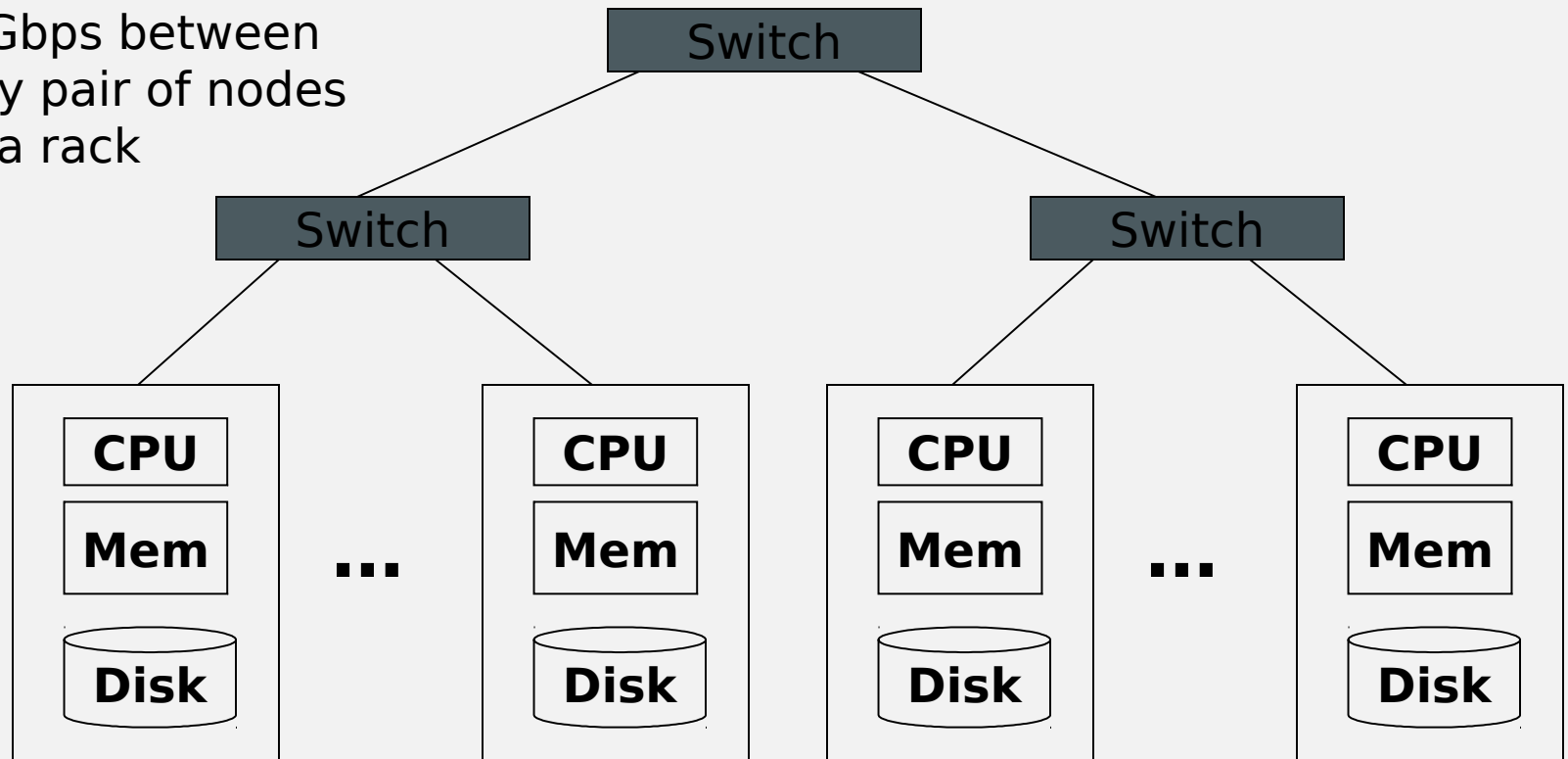
# Motivation: Google Example

- 20+ billion web pages x 20KB = 400+ TB
- 1 computer reads 30-35 MB/sec from disk
  - ~4 months to read the web
- ~1,000 hard drives to store the web
- Takes even more to **do** something useful with the data!
- **Today, a standard architecture for such problems is emerging:**
  - Cluster of commodity Linux nodes
  - Commodity network (ethernet) to connect them

# Cluster Architecture

2-10 Gbps backbone between racks

1 Gbps between  
any pair of nodes  
in a rack



Each rack contains 16-64 nodes

In 2011 it was guestimated that Google had 1M machines, <http://bit.ly/Shh0RO>  
J. Leskovec, A. Rajaraman, J. Ullman, Mining Massive Datasets, <http://www.mmms.org>







# Large-scale Computing

- **Large-scale computing** for **data mining** problems on **commodity hardware**
- **Challenges:**
  - **How do you distribute computation?**
  - **How can we make it easy to write distributed programs?**
  - **Machines fail:**
    - One server may stay up 3 years (1,000 days)
    - If you have 1,000 servers, expect to loose 1/day
    - People estimated Google had ~1M machines in 2011
      - 1,000 machines fail every day!

# Idea and Solution

- **Issue:** Copying data over a network takes time
- **Idea:**
  - Bring computation close to the data
  - Store files multiple times for reliability
- **Map-reduce** addresses these problems
  - Google's computational/data manipulation model
  - Elegant way to work with big data
  - **Storage Infrastructure - File system**
    - Google: GFS. Hadoop: HDFS
  - **Programming model**
    - Map-Reduce

# Storage Infrastructure

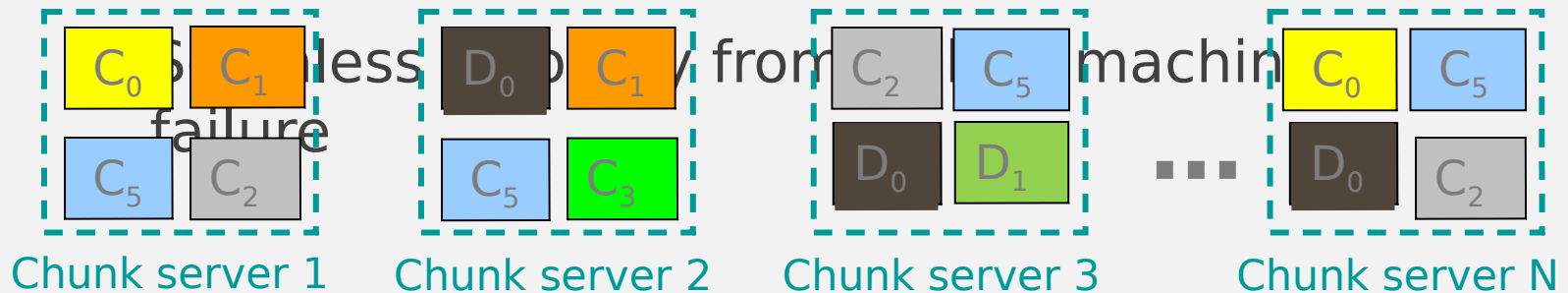
- **Problem:**
  - If nodes fail, how to store data persistently?
- **Answer:**
  - **Distributed File System:**
    - Provides global file namespace
    - Google GFS; Hadoop HDFS;
- **Typical usage pattern**
  - Huge files (100s of GB to TB)
  - Data is rarely updated in place
  - Reads and appends are common

# Distributed File System

- **Chunk servers**
  - File is split into contiguous chunks
  - Typically each chunk is 16-64MB
  - Each chunk replicated (usually 2x or 3x)
  - Try to keep replicas in different racks
- **Master node**
  - a.k.a. Name Node in Hadoop's HDFS
  - Stores metadata about where files are stored
  - Might be replicated
- **Client library for file access**
  - Talks to master to find chunk servers
  - Connects directly to chunk servers to access data

# Distributed File System

- **Reliable distributed file system**
- Data kept in “chunks” spread across machines
- Each chunk replicated on different machines

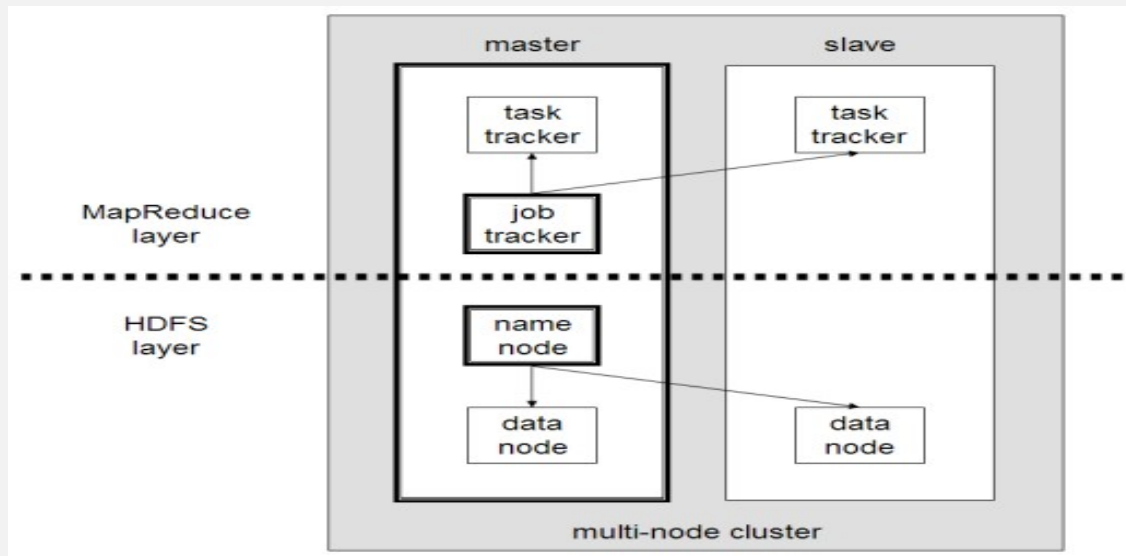


Bring computation directly to the data!

Chunk servers also serve as compute servers

# What is Hadoop

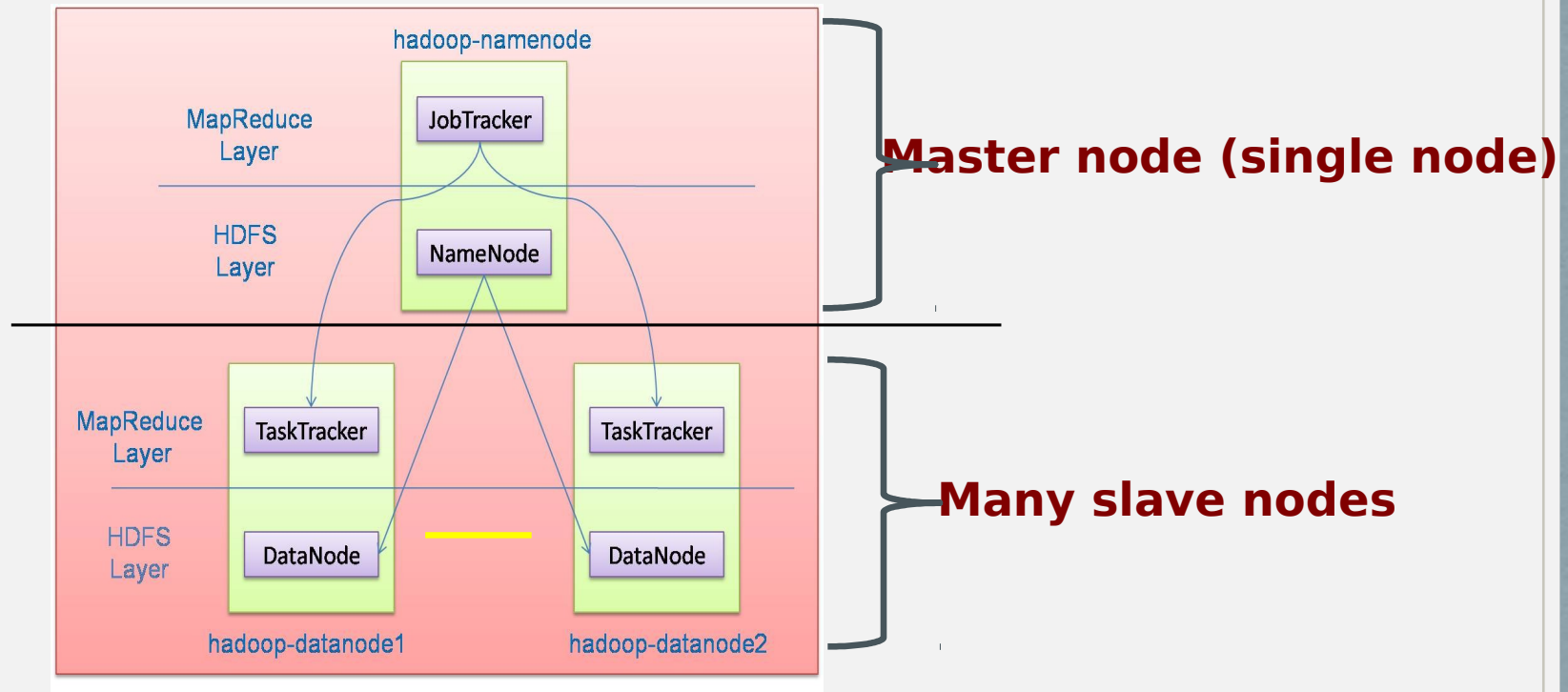
- **Hadoop framework consists on two main layers**
  - Distributed file system (HDFS)
  - Execution engine (MapReduce)





# Hadoop Master/Slave Architecture

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



# Design Principles of Hadoop

- Need to process big data
- Need to parallelize computation across thousands of nodes
- **Commodity hardware**
  - Large number of low-end cheap machines working in parallel to solve a computing problem
- This is in contrast to **Parallel DBs**
  - Small number of high-end expensive machines

# Commodity Clusters

- MapReduce is designed to efficiently process large volumes of data by connecting many commodity computers together to work in parallel
- A *theoretical* 1000-CPU machine would cost a very large amount of money, far more than 1000 single-CPU or 250 quad-core machines
- MapReduce ties smaller and more reasonably priced machines together into a single cost-effective *commodity cluster*

# Design Principles of Hadoop

- **Automatic parallelization & distribution**
  - Hidden from the end-user
- **Fault tolerance and automatic recovery**
  - Nodes/tasks will fail and will recover automatically
- **Clean and simple programming abstraction**
  - Users only provide two functions “map” and “reduce”

# How Uses MapReduce/Hadoop

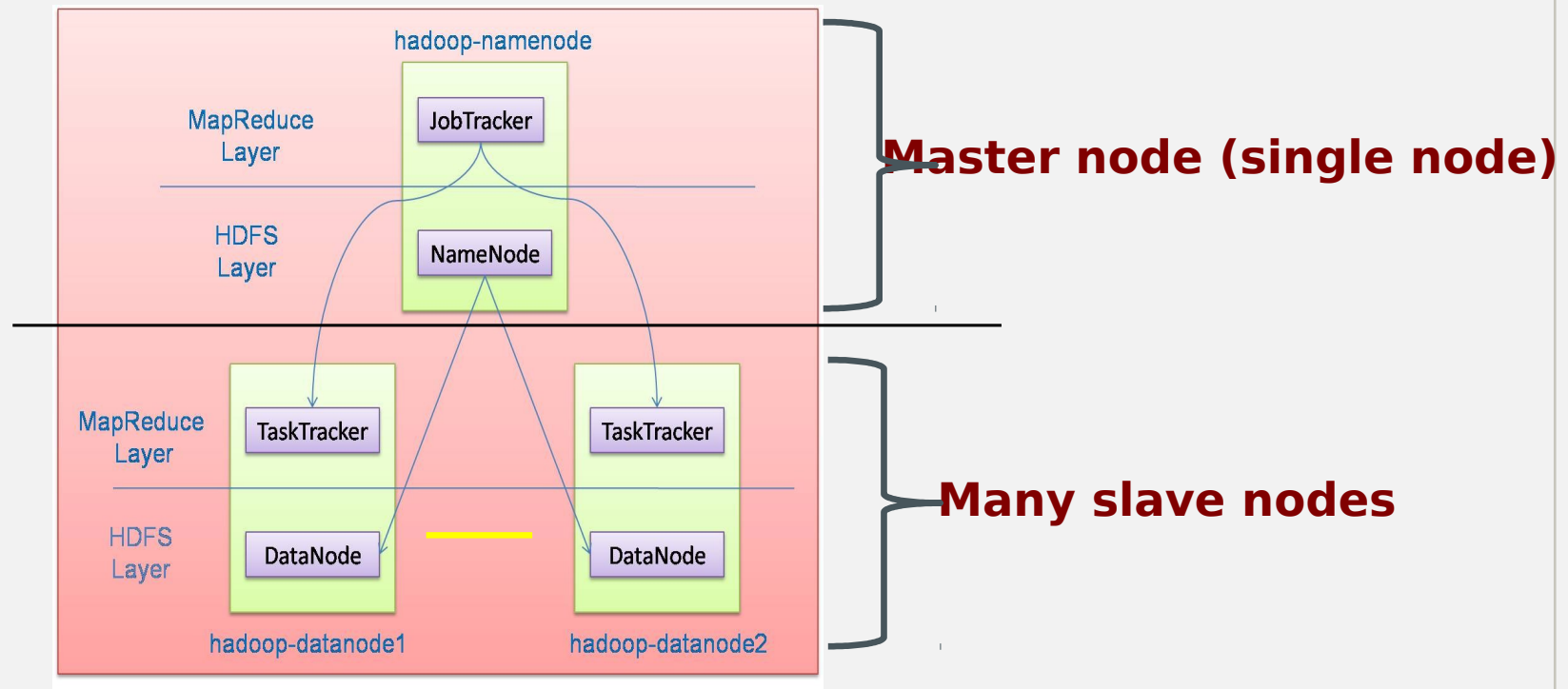
- Google: Inventors of MapReduce computing paradigm
- Yahoo: Developing Hadoop open-source of MapReduce
- IBM, Microsoft, Oracle
- Facebook, Amazon, AOL, NetFlex
- Many others + universities and research labs

# Hadoop: How it Works

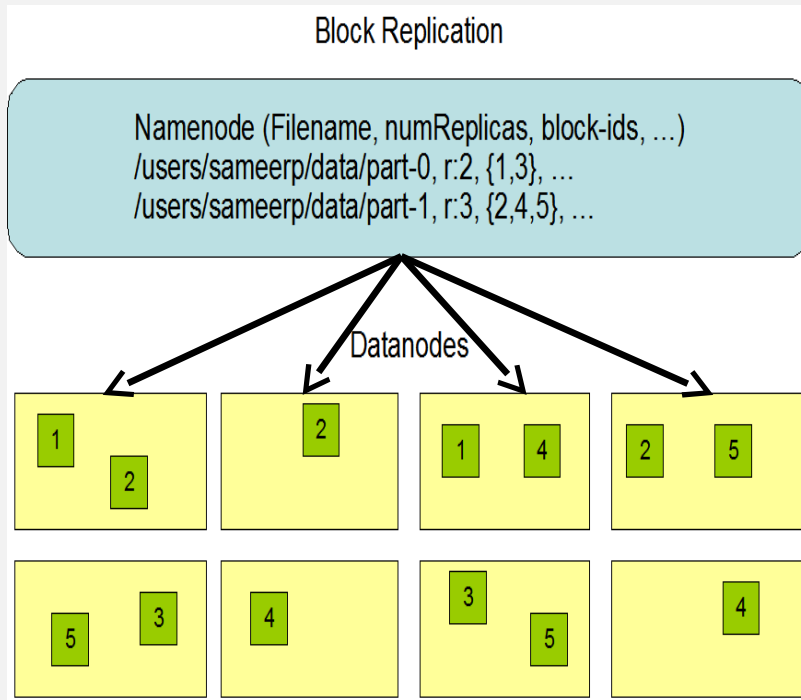


# Hadoop Architecture

- Distributed file system (HDFS)
- Execution engine (MapReduce)

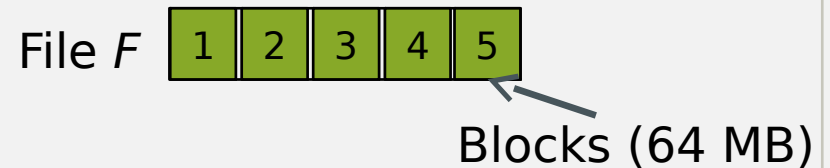


# Hadoop Distributed File System (HDFS)



## Centralized namenode

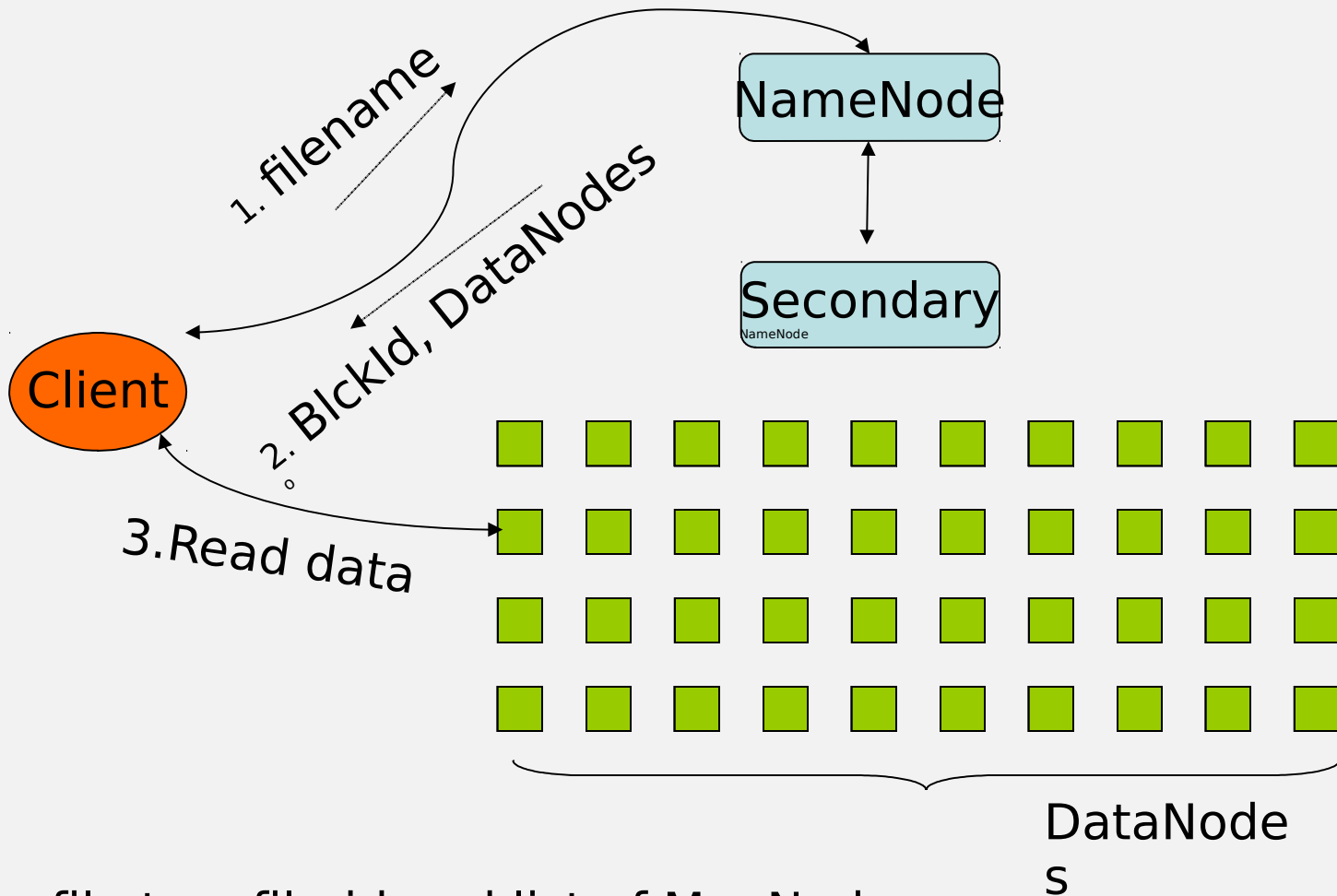
- Maintains metadata info about files



## Many datanode (1000s)

- Store the actual data
- Files are divided into blocks
- Each block is replicated *N* times  
(Default = 3)

# HDFS Architecture

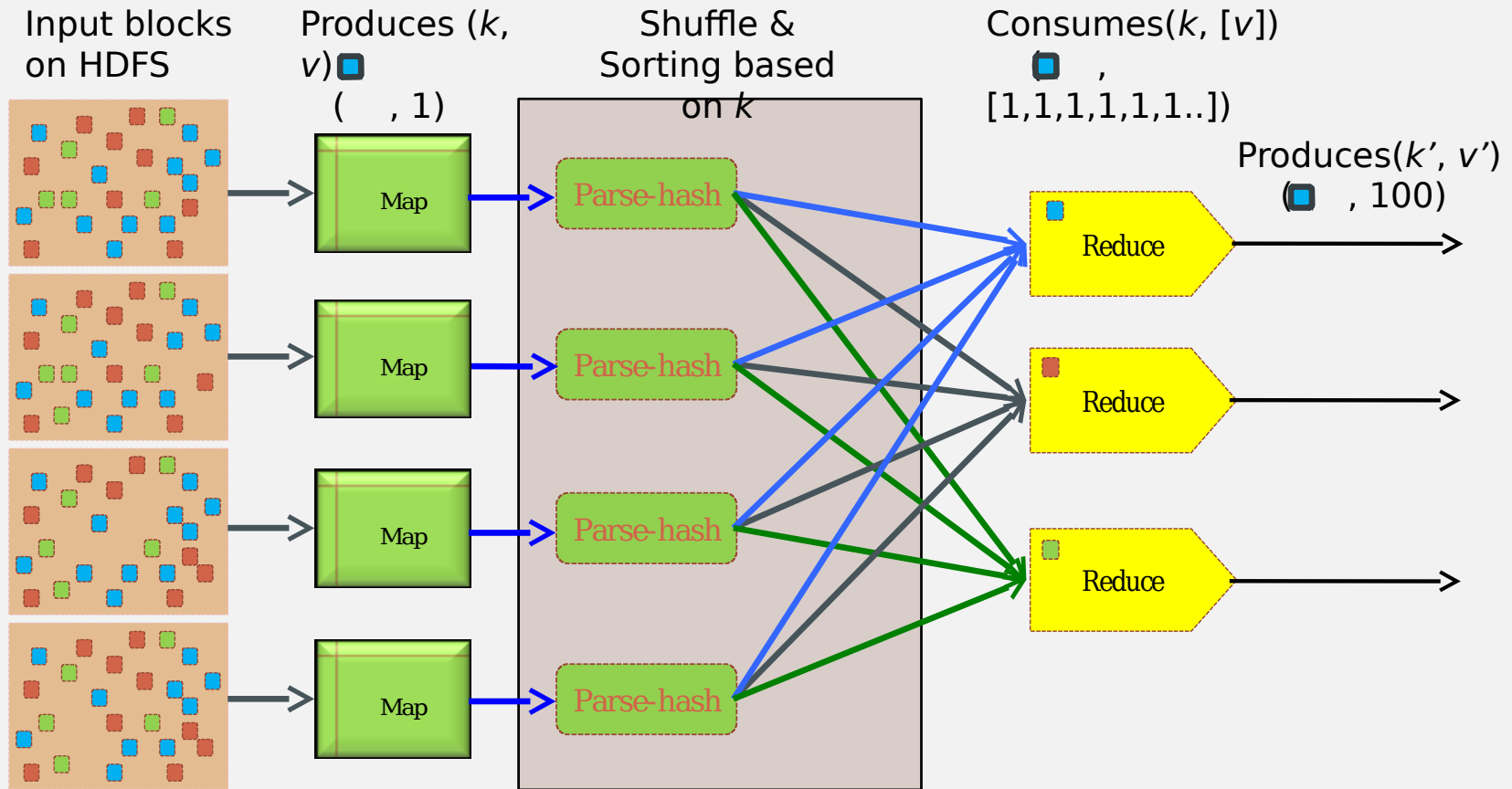


Maps a file to a file-id and list of MapNodes  
Maps a block-id to a physical location on disk

# Main Properties of HDFS

- ***Large:*** A HDFS instance may consist of thousands of server machines, each storing part of the file system's data
- ***Replication:*** Each data block is replicated many times (default is 3)
- ***Failure:*** Failure is the norm rather than exception
- ***Fault Tolerance:*** Detection of faults is quick, and automatic recovery from them is a core architectural goal of HDFS
  - Namenode is consistently checking Datanodes

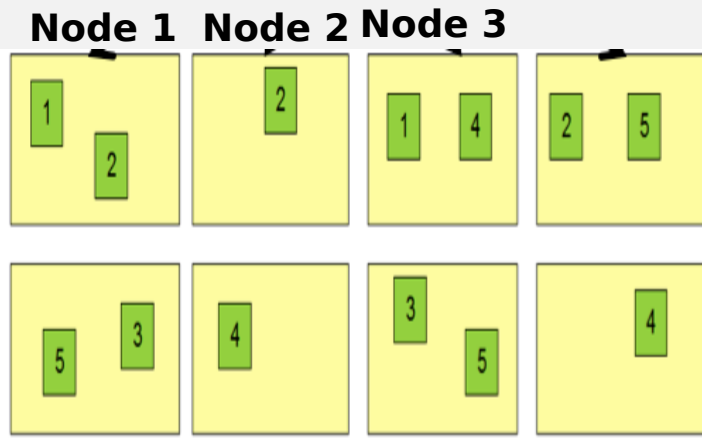
# Map-Reduce Execution Engine (Example: Color Count)



***Users only provide the “Map” and “Reduce” functions***

# 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)

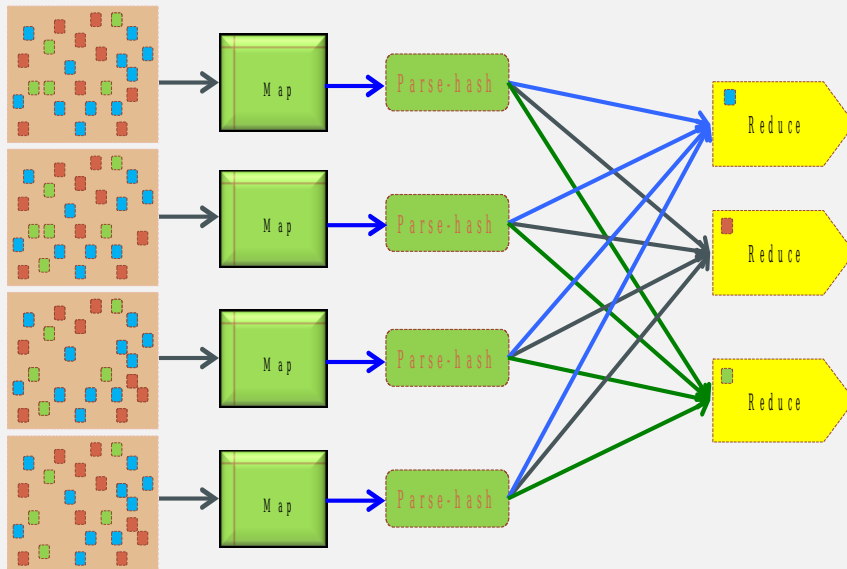


- This file has 5 Blocks ✉ run 5 map tasks
- Where to run the task reading block “1”
  - *Try to run it on Node 1 or Node 3*



# 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*

# Isolated Tasks

- MapReduce divides the workload into multiple *independent tasks* and schedule them across cluster nodes
- A work performed by each task is done *in isolation* from one another
- The amount of communication which can be performed by tasks is mainly limited for scalability reasons

# Key-Value Pairs

- Mappers and Reducers are users' code (provided functions)
- Just need to obey the Key-Value pairs interface
- **Mappers:**
  - Consume <key, value> pairs
  - Produce <key, value> pairs
- **Reducers:**
  - Consume <key, <list of values>>
  - Produce <key, value>
- **Shuffling and Sorting:**
  - Hidden phase between mappers and reducers
  - Groups all similar keys from all mappers, sorts and passes them to a certain reducer in the form of <key, <list of values>>

# Programming Model: MapReduce

## Warm-up task:

- We have a huge text document
- Count the number of times each distinct word appears in the file
- **Sample application:**
  - Analyze web server logs to find popular URLs

# Task: Word Count

## Case 1:

- File too large for memory, but all `<word, count>` pairs fit in memory

## Case 2:

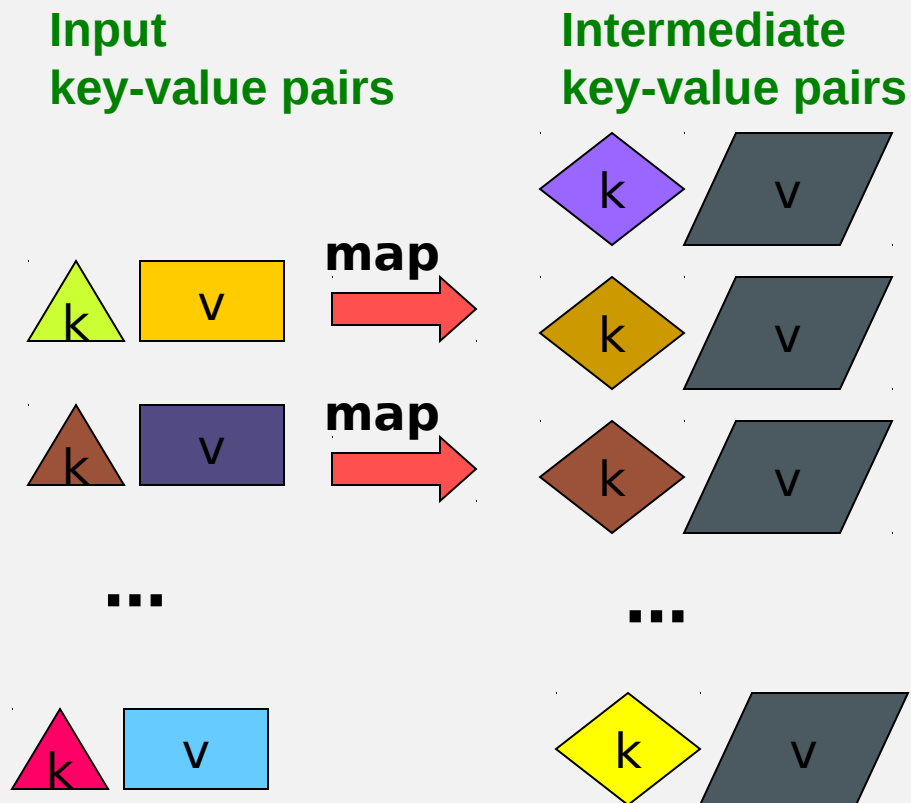
- Count occurrences of words:
  - `words(doc.txt) | sort | uniq -c`
    - where `words` takes a file and outputs the words in it, one per a line
- Case 2 captures the essence of **MapReduce**
  - Great thing is that it is naturally parallelizable

# MapReduce: Overview

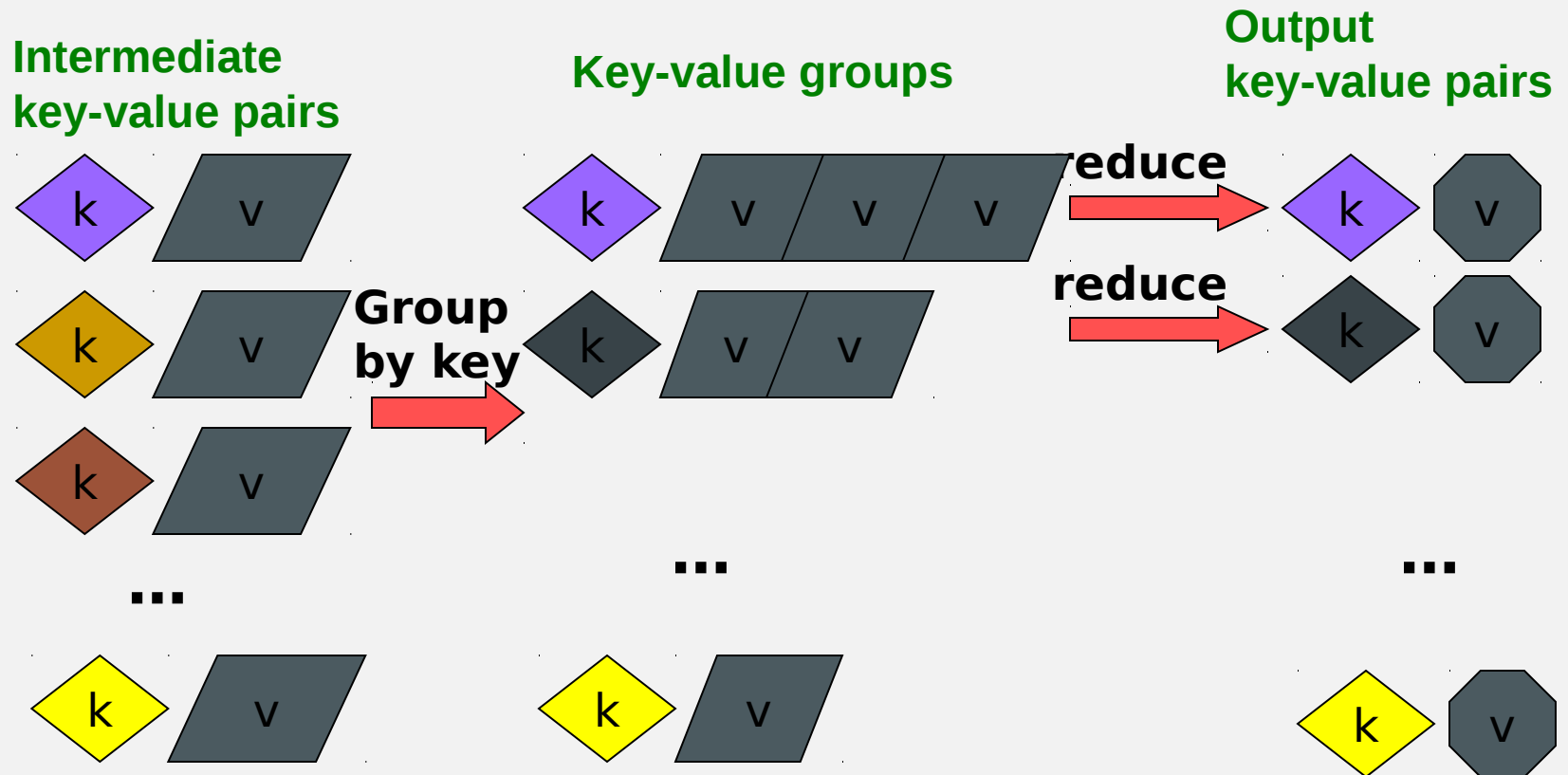
- Sequentially read a lot of data
- **Map:**
  - Extract something you care about
- **Group by key:** Sort and Shuffle
- **Reduce:**
  - Aggregate, summarize, filter or transform
- Write  
**Outline stays the same, Map and Reduce change to fit the problem**



# MapReduce: The Map Step



# MapReduce: The Reduce Step



# More Specifically

- **Input:** a set of key-value pairs
- Programmer specifies two methods:
  - **Map( $k, v$ )**  $\rightarrow \langle k', v' \rangle^*$ 
    - Takes a key-value pair and outputs a set of key-value pairs
      - E.g., key is the filename, value is a single line in the file
    - There is one Map call for every  $(k, v)$  pair
  - **Reduce( $k', \langle v' \rangle^*$ )**  $\rightarrow \langle k', v'' \rangle^*$ 
    - **All values  $v'$  with same key  $k'$  are reduced together and processed in  $v'$  order**
    - There is one Reduce function call per unique key  $k'$

# MapReduce: Word Counting

Provided by the  
programmer

## MAP:

Read input and  
produces a set  
of key-value  
pairs

(The, 1)  
(crew, 1)  
(of, 1)  
(the, 1)  
(space, 1)  
(shuttle, 1)  
(Endeavor, 1)  
(recently, 1)  
...

(key, value)

## Group by key:

Collect all pairs  
with same key

(crew, 1)  
(crew, 1)  
(space, 1)  
(the, 1)  
(the, 1)  
(the, 1)  
(shuttle, 1)  
(recently, 1)  
...

(key, value)

Provided by the  
programmer

## Reduce:

Collect all  
values  
belonging to  
the key and  
output

(crew, 2)  
(space, 1)  
(the, 3)  
(shuttle, 1)  
(recently, 1)  
...

(key, value)

Only sequential reads

The crew of the space shuttle Endeavor recently returned to Earth as ambassadors, harbingers of a new era of space exploration. Scientists at NASA are saying that the recent assembly of the Dextre bot is the first step in a long-term space-based man/machine partnership. "The work we're doing now -- the robotics we're doing -- is what we're going to need .....

Big document

# Word Count Using MapReduce

```
map(key, value):
```

```
// key: document name; value: text of the document
```

```
  for each word w in value:
```

```
    emit(w, 1)
```

```
reduce(key, values):
```

```
// key: a word; value: an iterator over counts
```

```
  result = 0
```

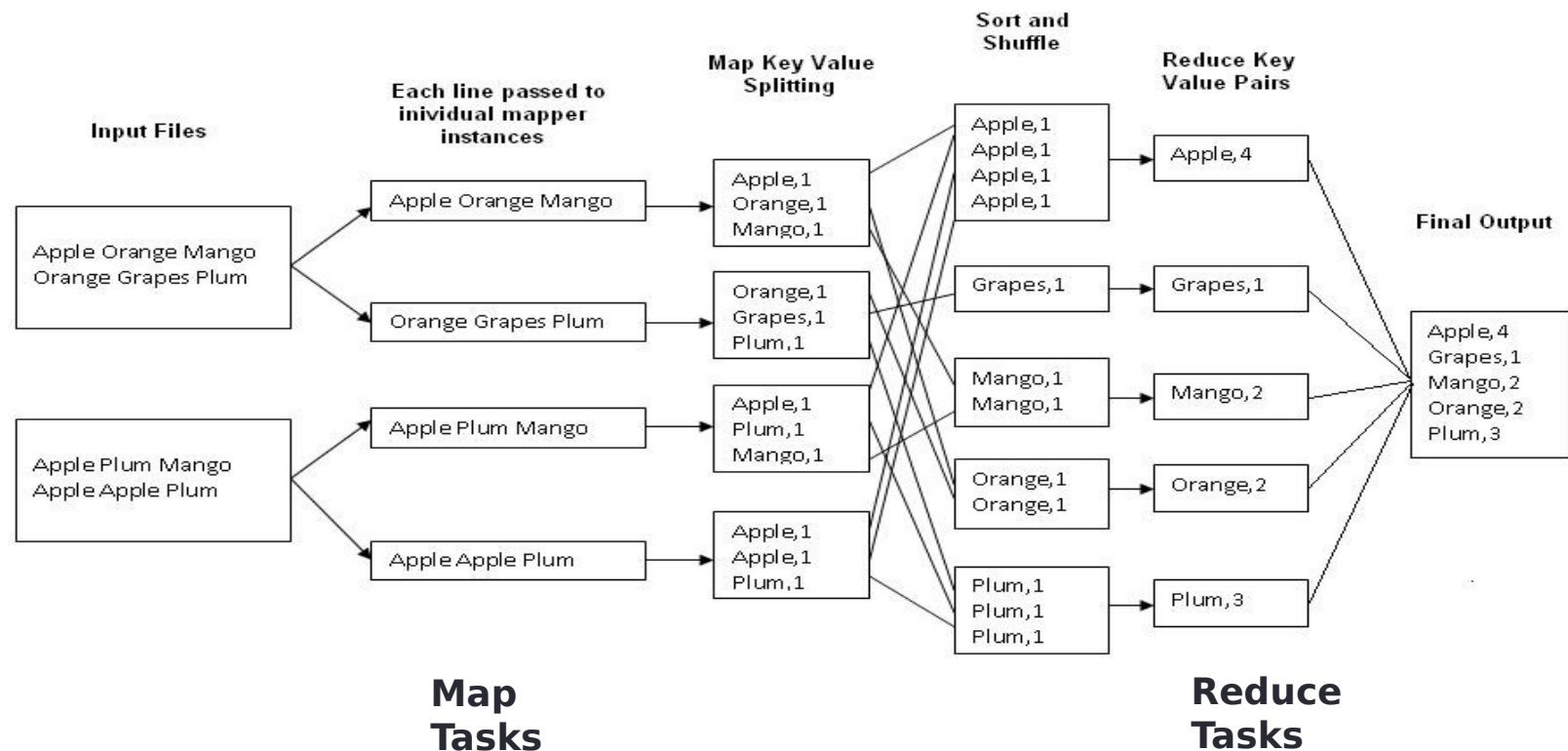
```
  for each count v in values:
```

```
    result += v
```

```
  emit(key, result)
```

# Example: Word Count

- Job: Count the occurrences of each word in a data set**



# Map-Reduce: Environment

## Map-Reduce environment takes care of:

- Partitioning the input data
- Scheduling the program's execution across a set of machines
- Performing the **group by key** step
- Handling machine failures
- Managing required inter-machine communication

# Map-Reduce: A

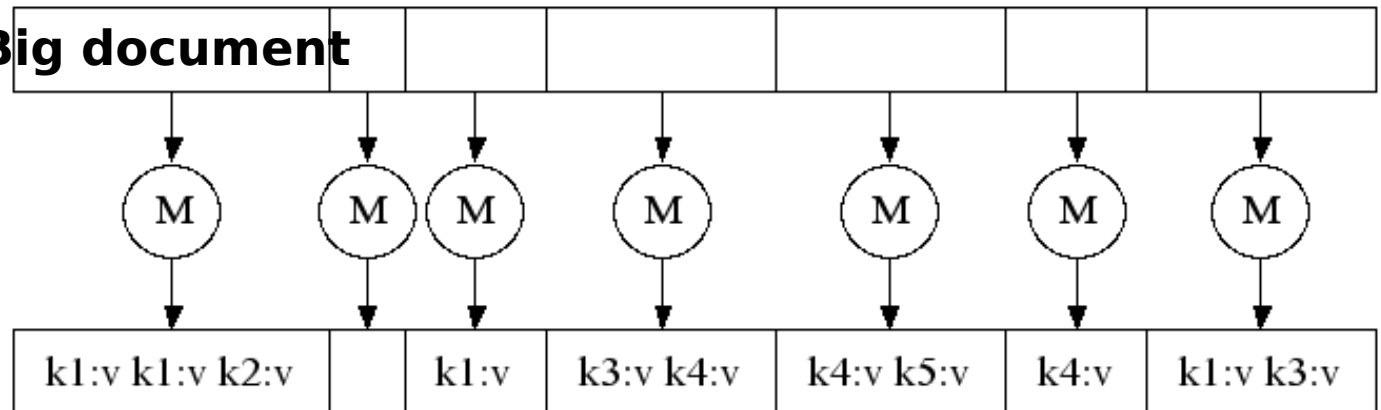
distributed system

Input **Big document**

## MAP:

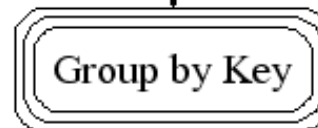
Read input and produces a set of key-value pairs

Intermediate

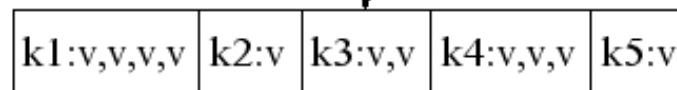


## Group by key:

Collect all pairs with same key (Hash merge, Shuffle, Sort, Partition)



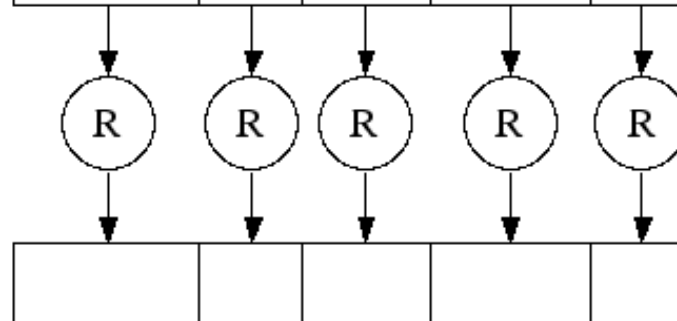
Grouped



## Reduce:

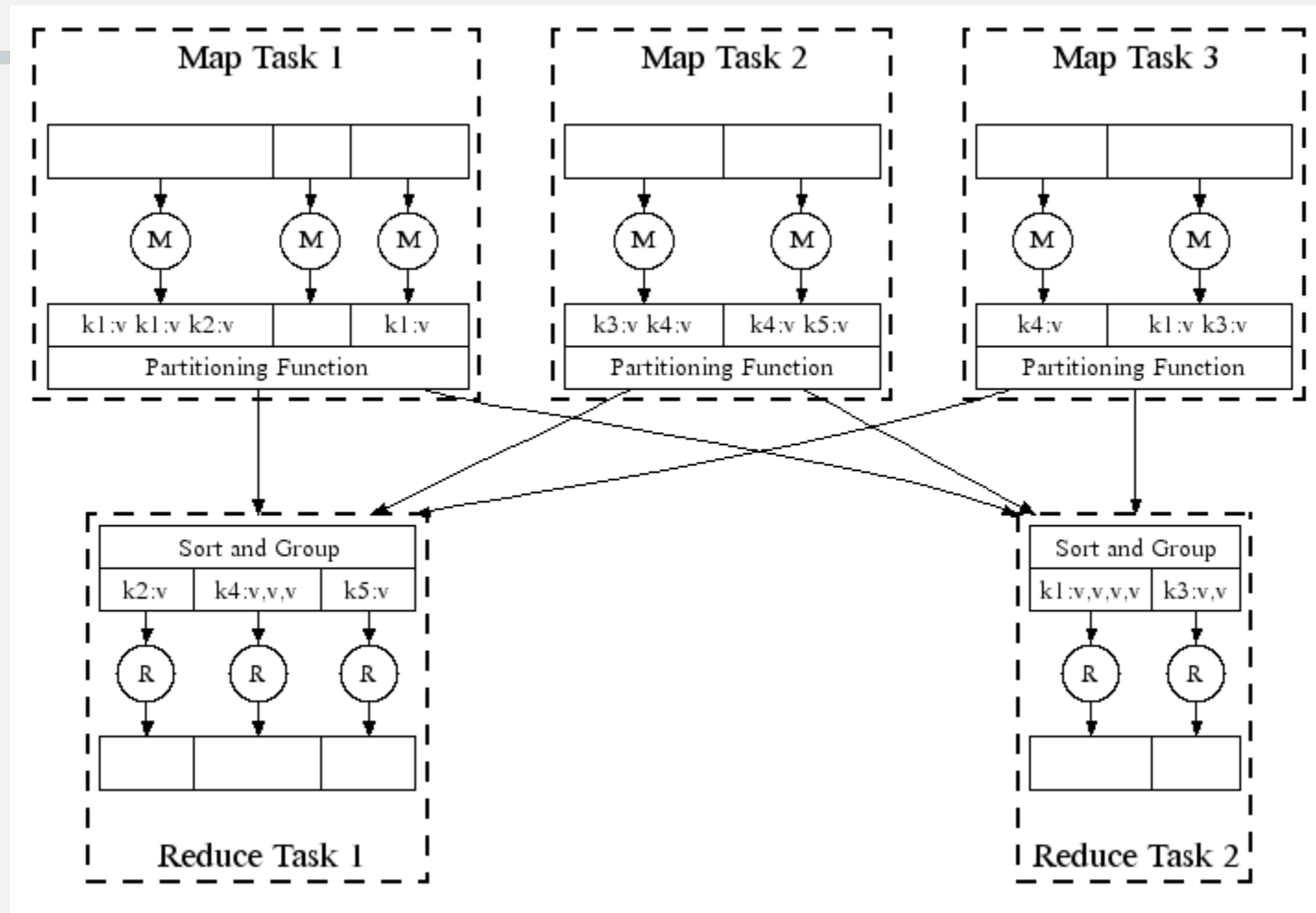
Collect all values belonging to the key and output

Output





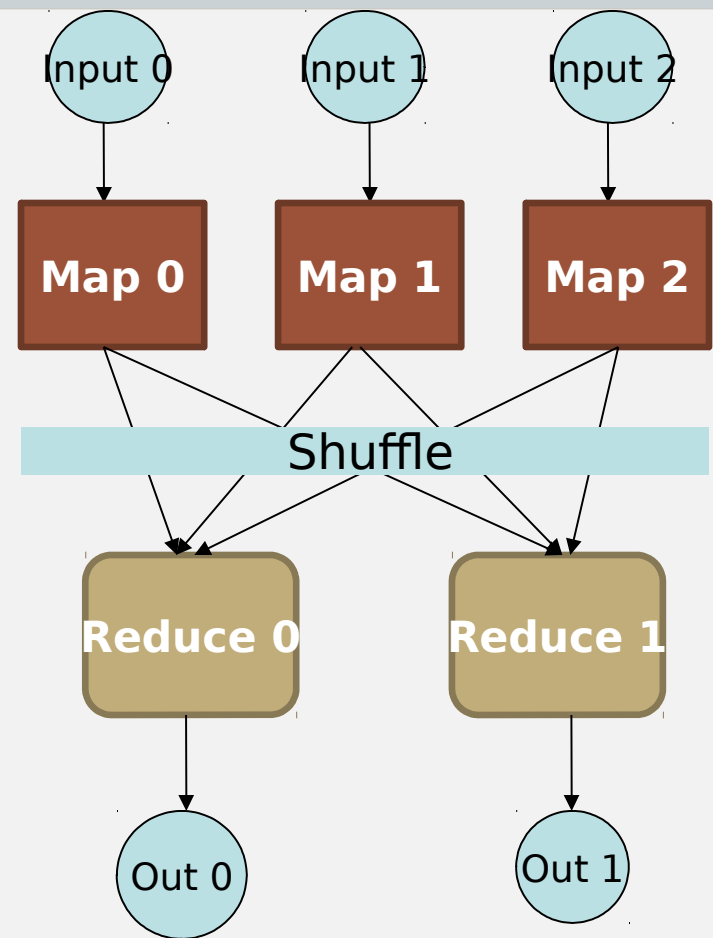
# Map-Reduce: In Parallel



**All phases are distributed with many tasks doing the work**

# Map-Reduce

- Programmer specifies:
  - Map and Reduce and input files
- **Workflow:**
  - Read inputs as a set of key-value-pairs
  - **Map** transforms input kv-pairs into a new set of k'v'-pairs
  - Sorts & Shuffles the k'v'-pairs to output nodes
  - All k'v'-pairs with a given k' are sent to the same **reduce**
  - **Reduce** processes all k'v'-pairs grouped by key into new k''v''-pairs
  - Write the resulting pairs to files
- All phases are distributed with many tasks doing the work



# Data Flow

- **Input and final output are stored on a distributed file system (FS):**
  - Scheduler tries to schedule map tasks “close” to physical storage location of input data
- **Intermediate results are stored on local FS of Map and Reduce workers**
- **Output is often input to another MapReduce task**

# Coordination: Master

- **Master node takes care of coordination:**
  - **Task status:** (idle, in-progress, completed)
  - **Idle tasks** get scheduled as workers become available
  - When a map task completes, it sends the master the location and sizes of its  $R$  intermediate files, one for each reducer
  - Master pushes this info to reducers
- Master pings workers periodically to detect failures

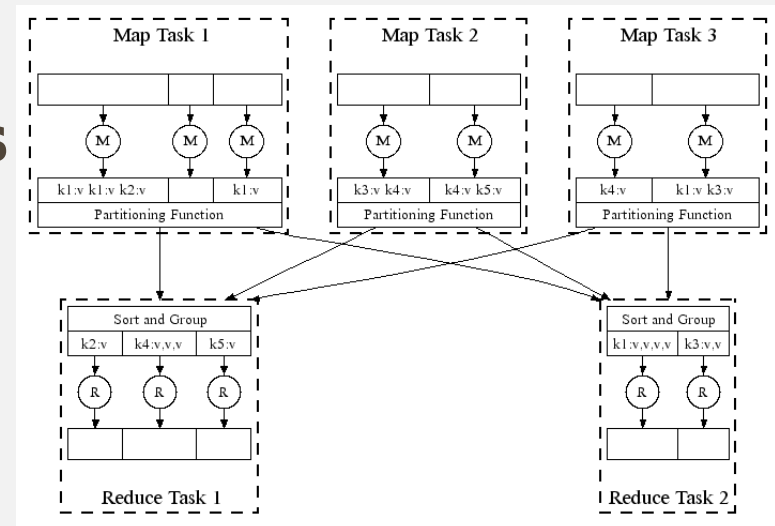
# Refinement: Combiners

- Often a Map task will produce many pairs of the form  $(k, v_1), (k, v_2), \dots$  for the same key  $k$ 
  - E.g., popular words in the word count example

- **Can save network time pre-aggregating values the mapper:**

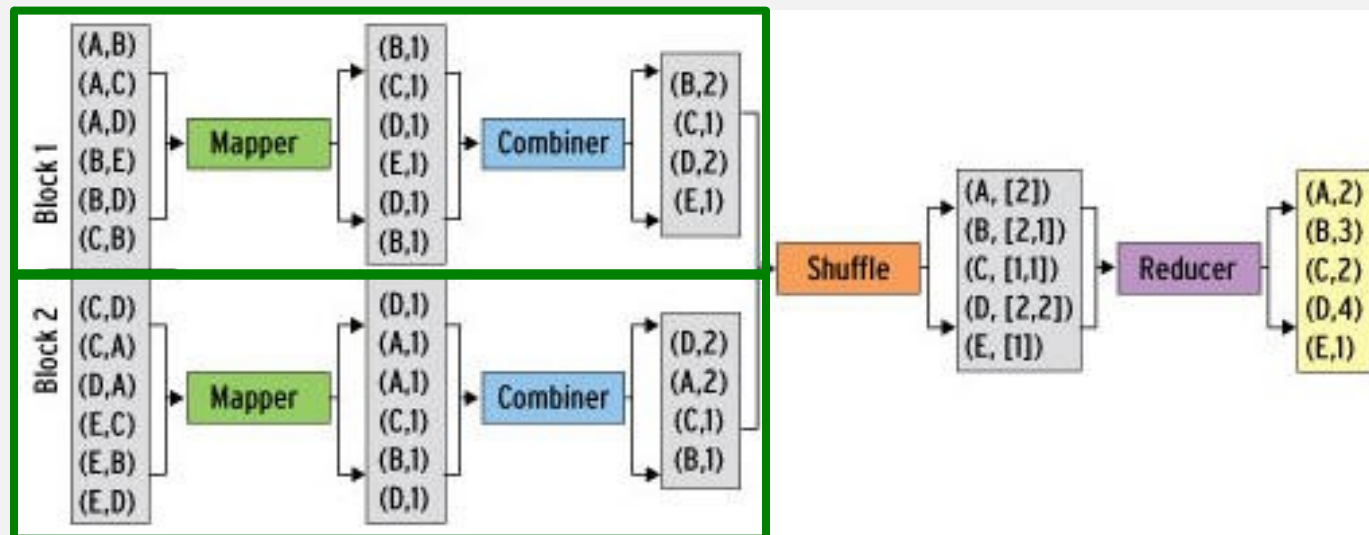
- $\text{combine}(k, \text{list}(v_1)) \rightarrow v_2$
- Combiner is usually same as the reduce function

- Works only if reduce function is commutative and associative



# Refinement: Combiners

- **Back to our word counting example:**
  - Combiner combines the values of all keys of a single mapper (single machine):



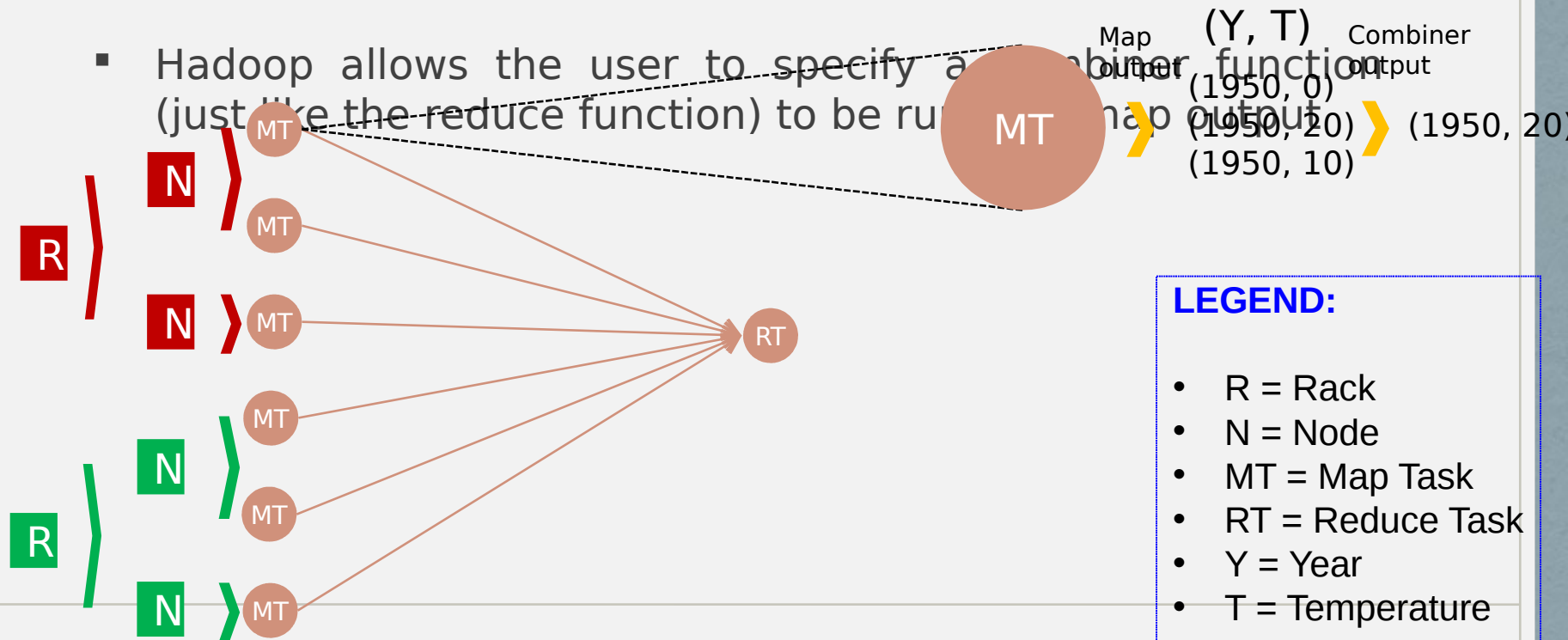
- Much less data needs to be copied and shuffled!

# Combiner Functions

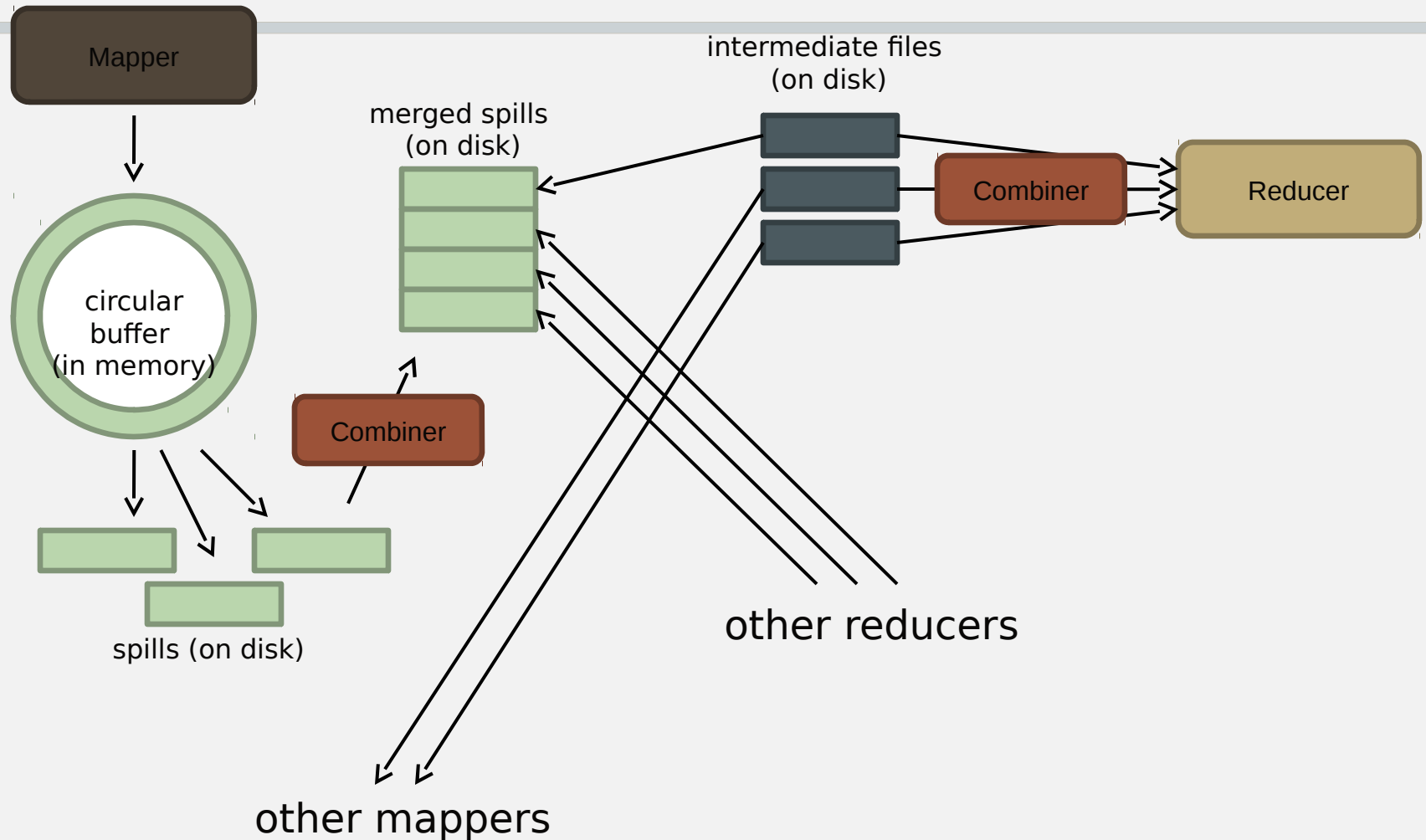
- MapReduce applications are limited by the bandwidth available on the cluster

- It pays to minimize the data shuffled between map and reduce tasks

- Hadoop allows the user to specify a combiner function (just like the reduce function) to be run



# Shuffle and Sort





# Shuffle and Sort in Hadoop

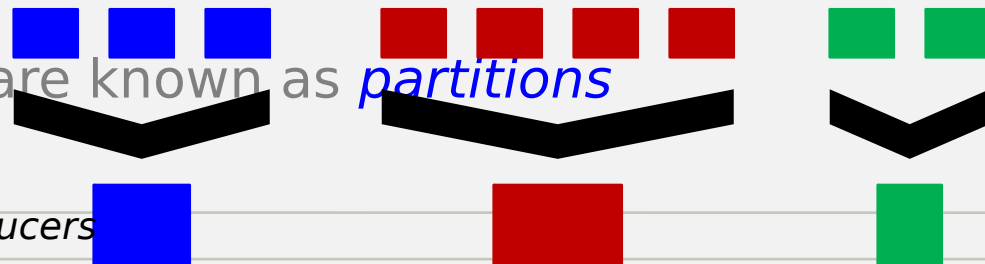
- Probably the most complex aspect of MapReduce!
- Map side
  - Map outputs are buffered in memory in a circular buffer
  - When buffer reaches threshold, contents are “spilled” to disk
  - Spills merged in a single, partitioned file (sorted within each partition): combiner runs here
- Reduce side
  - First, map outputs are copied over to reducer machine
  - “Sort” is a multi-pass merge of map outputs (happens in memory and on disk): combiner runs here
  - Final merge pass goes directly into reducer

# Partitions

- In MapReduce, intermediate output values are not usually reduced together
- *All values with the same key are presented to a single Reducer together*
- More specifically, a different subset of intermediate key space is assigned to each Reducer

*Different colors represent different keys (potentially) from different Mappers*

These subsets are known as *partitions*



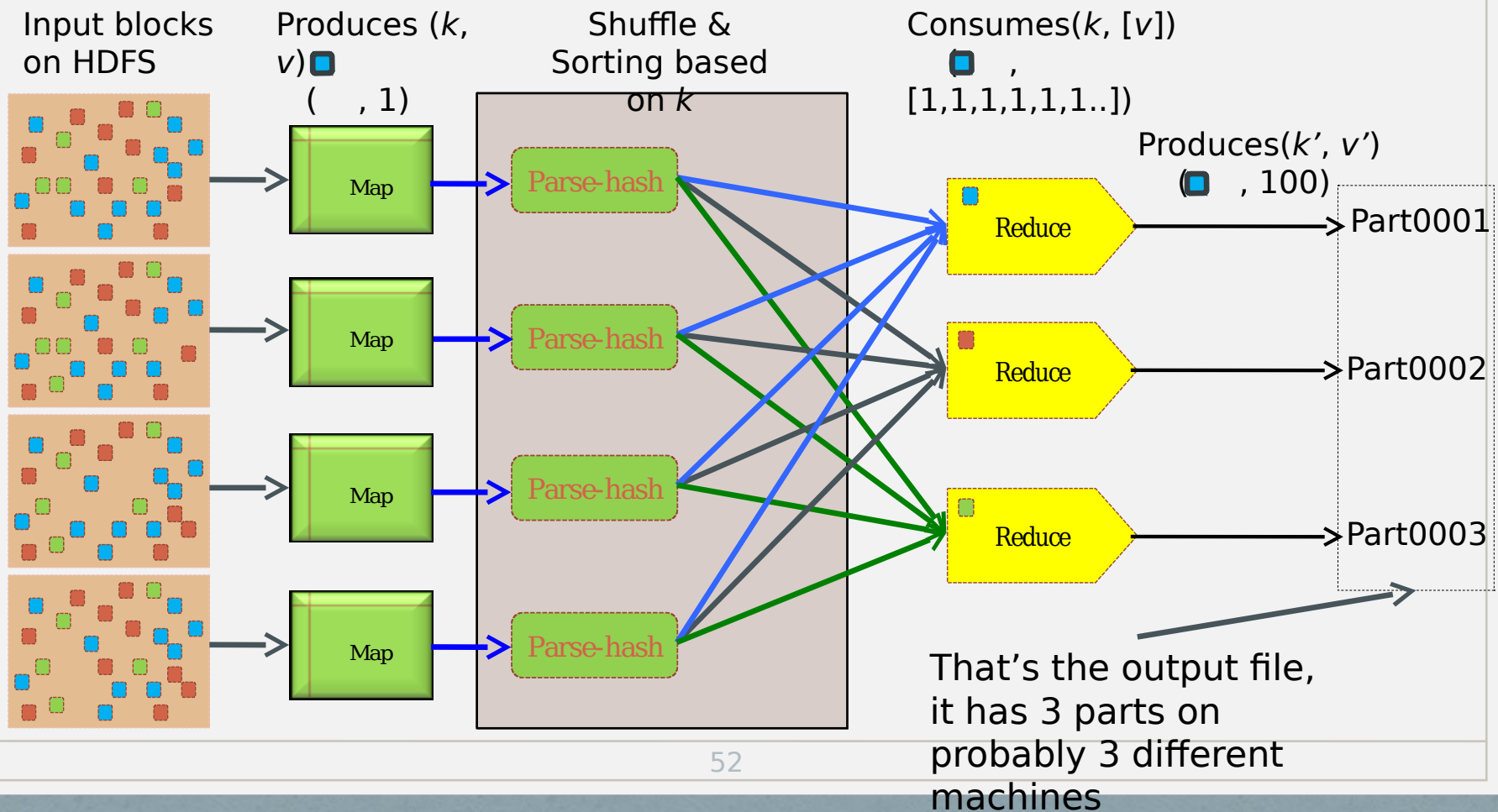
*Partitions are the input to Reducers*

# Refinement: Partition Function

- **Want to control how keys get partitioned**
  - Inputs to map tasks are created by contiguous splits of input file
  - Reduce needs to ensure that records with the same intermediate key end up at the same worker
- **System uses a default partition function:**
  - $\text{hash}(\text{key}) \bmod R$
- **Sometimes useful to override the hash function:**
  - E.g.,  $\text{hash}(\text{hostname}(\text{URL})) \bmod R$  ensures URLs from a host end up in the same output file

# Example 2: Color Count

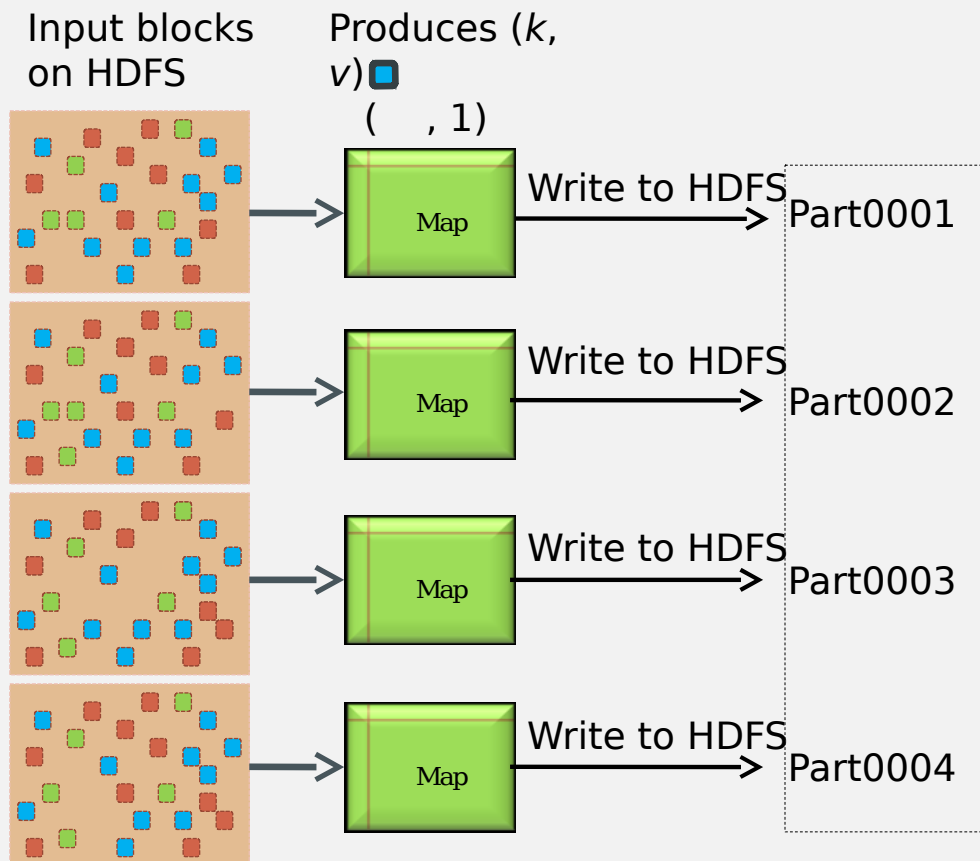
**Job: Count the number of each color in a data set**



# Example 3: Color Filter

**Job: Select only the blue and the green**

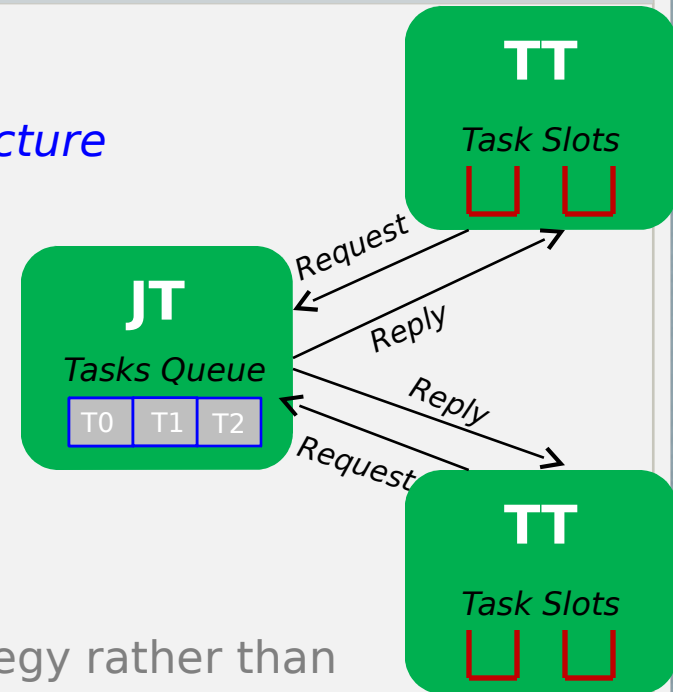
- Each map task will select only the blue or green colors
- No need for reduce phase



That's the output file, it has 4 parts on probably 4 different machines

# Task Scheduling in MapReduce

- MapReduce adopts a *master-slave architecture*
- The master node in MapReduce is referred to as *Job Tracker* (JT)
- Each slave node in MapReduce is referred to as *Task Tracker* (TT)
- MapReduce adopts a *pull scheduling* strategy rather than a *push one*
  - I.e., JT does not push map and reduce tasks to TTs but rather TTs pull them by making pertaining requests



# Map and Reduce Task Scheduling

- Every TT sends a *heartbeat message* periodically to JT encompassing a request for a map or a reduce task to run

## I. Map Task Scheduling:

- JT satisfies requests for map tasks via attempting to schedule mappers in the *vicinity* of their input splits (i.e., it considers locality)

## II. Reduce Task Scheduling:

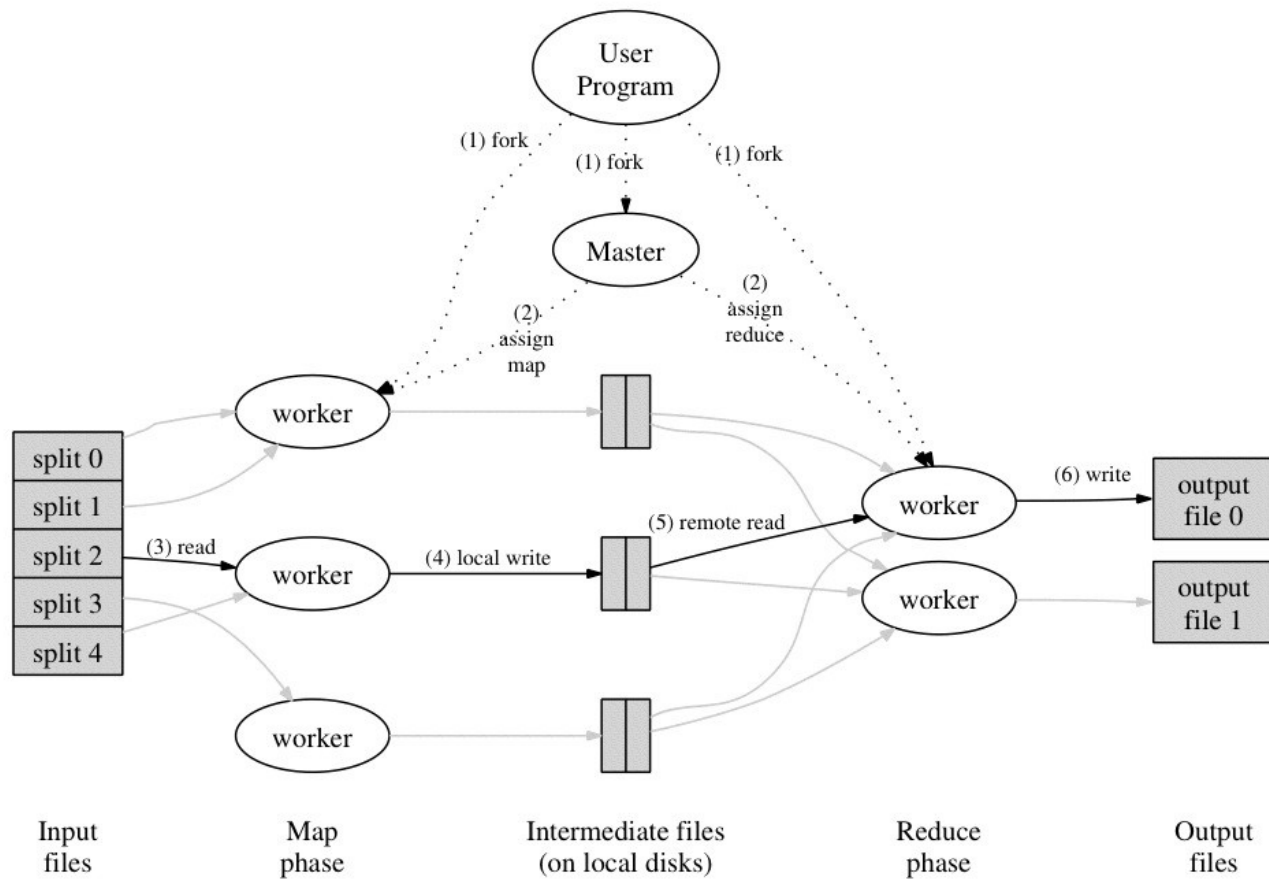
- However, JT simply assigns the next yet-to-run reduce task to a requesting TT regardless of TT's network location and its implied effect on the reducer's shuffle time (i.e., it does not consider locality)

# Job Scheduling in MapReduce

- In MapReduce, an application is represented as a *job*
- A job encompasses multiple map and reduce tasks
- MapReduce in Hadoop comes with a choice of schedulers:
  - The default is the *FIFO scheduler* which schedules jobs in order of submission
  - There is also a multi-user scheduler called the *Fair scheduler* which aims to give every user a fair share of the cluster capacity over time



# Execution Overview



# How many Map and Reduce jobs?

- $M$  map tasks,  $R$  reduce tasks
- **Rule of a thumb:**
  - Make  $M$  much larger than the number of nodes in the cluster
  - One DFS chunk per map is common
  - Improves dynamic load balancing and speeds up recovery from worker failures
- **Usually  $R$  is smaller than  $M$** 
  - Because output is spread across  $R$  files

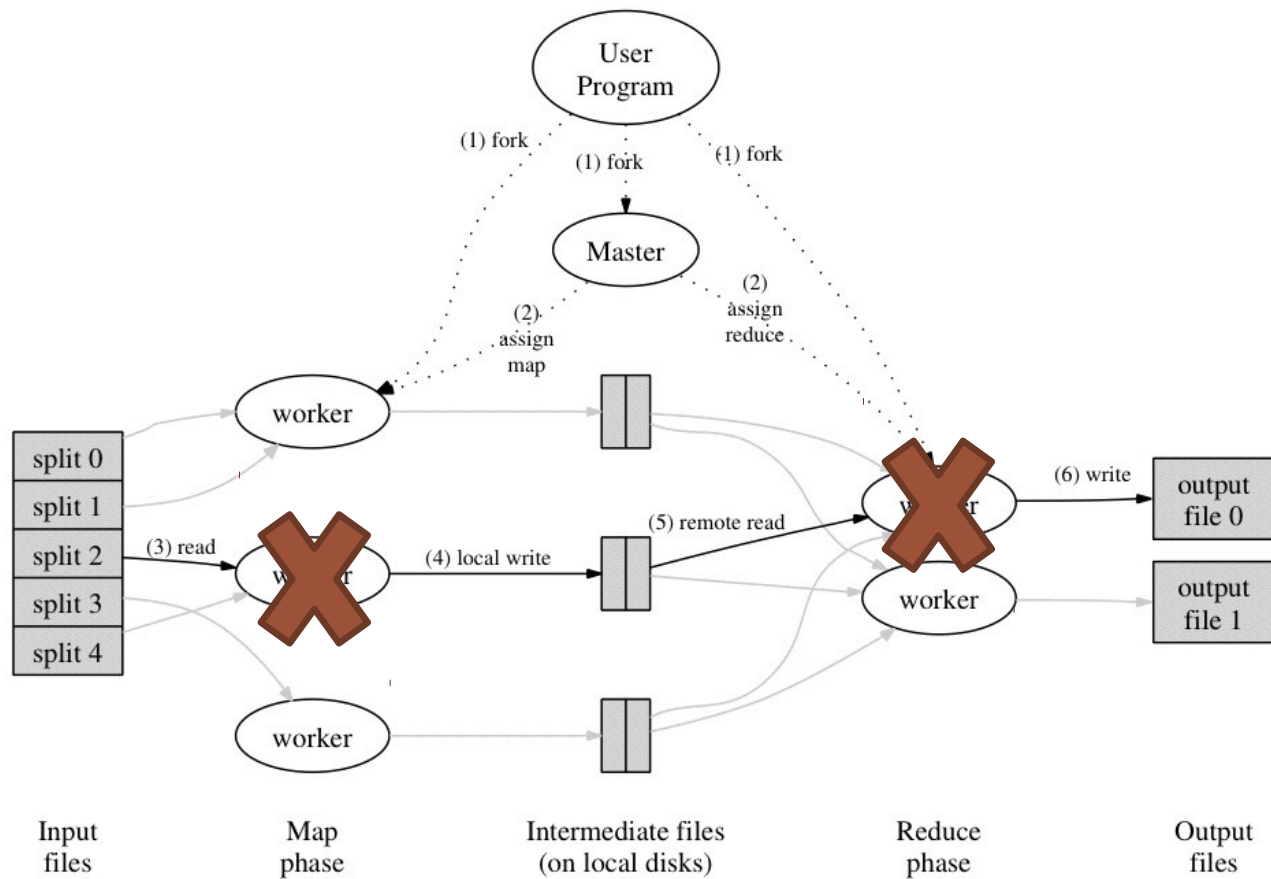
# Dealing with Failures

- **Map worker failure**
  - Map tasks completed or in-progress at worker are reset to idle
  - Reduce workers are notified when task is rescheduled on another worker
- **Reduce worker failure**
  - Only in-progress tasks are reset to idle
  - Reduce task is restarted
- **Master failure**
  - MapReduce task is aborted and client is notified

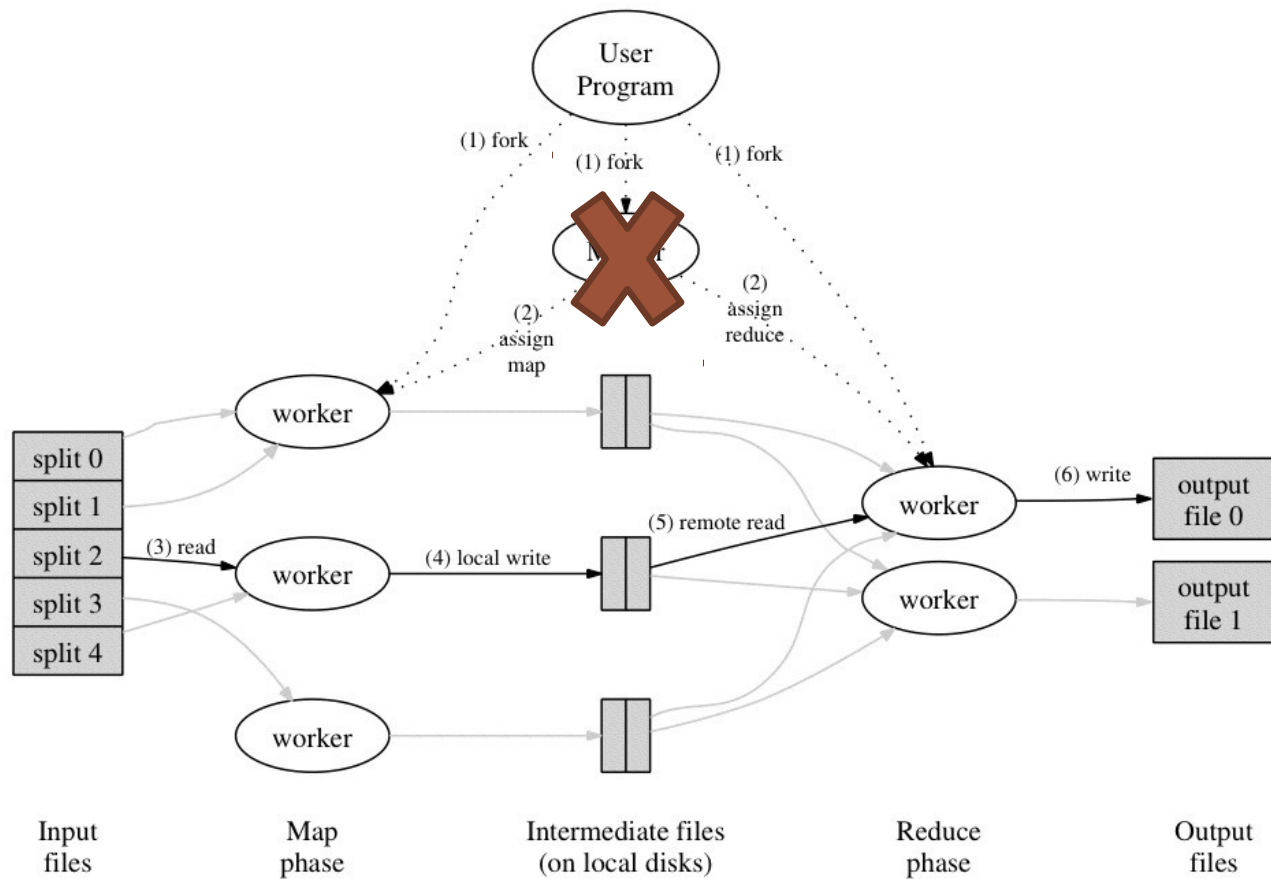
# Fault Tolerance in Hadoop

- MapReduce can guide jobs toward a successful completion even when jobs are run on a large cluster where probability of failures increases
- The primary way that MapReduce achieves fault tolerance is through *restarting tasks*
- If a TT fails to communicate with JT for a period of time (by default, 1 minute in Hadoop), JT will assume that TT in question has crashed
  - If the job is still in the map phase, JT asks another TT to re-execute all Mappers that previously ran at the failed TT
  - If the job is in the reduce phase, JT asks another TT to re-execute all Reducers that were in progress on the failed TT

# Worker Failure



# Master Failure



# Fault Tolerance / Workers

Handled via re-execution

- Detect failure via periodic heartbeats
- Re-execute completed + in-progress *map* tasks
  - *Why????*
- Re-execute in progress *reduce* tasks
- Task completion committed through master

Robust:

lost 1600/1800 machines once  finished o

Semantics in presence of failures: see paper

# Refinements: Backup Tasks

- **Problem**

- Slow workers significantly lengthen the job completion time:
  - Other jobs on the machine
  - Bad disks
  - Weird things

- **Solution**

- Near end of phase, spawn backup copies of tasks
  - Whichever one finishes first “wins”

- **Effect**

- Dramatically shortens job completion time

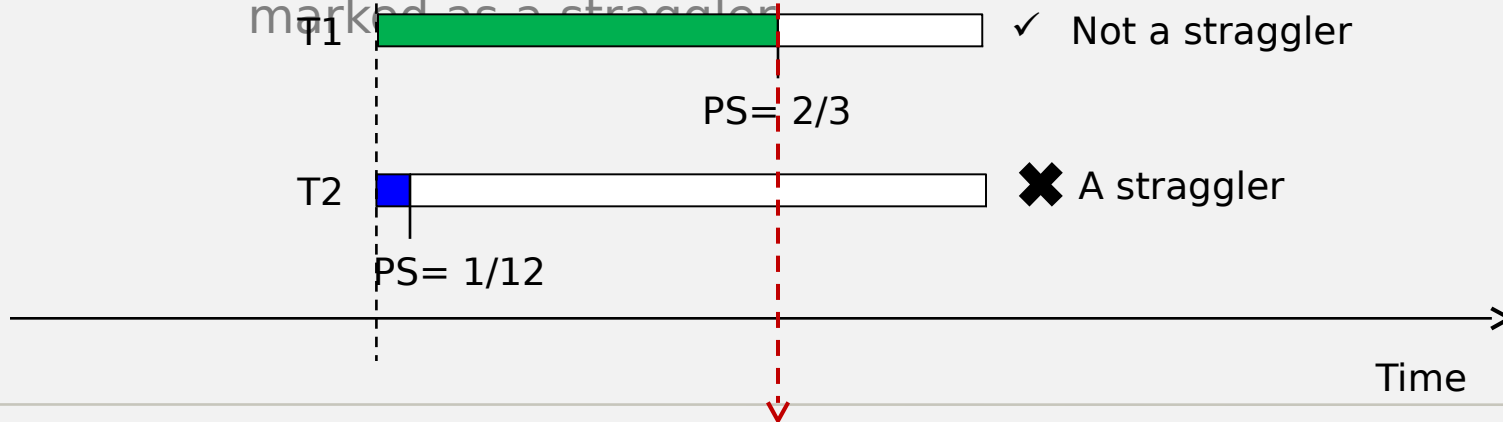


# Speculative Execution

- A MapReduce job is dominated by the slowest task
- MapReduce attempts to locate slow tasks (*stragglers*) and run redundant (*speculative*) tasks that will optimistically commit before the corresponding stragglers
- This process is known as *speculative execution*
- Only one copy of a straggler is allowed to be speculated
- Whichever copy (among the two copies) of a task commits first, it becomes the definitive copy, and the other copy is killed by JT

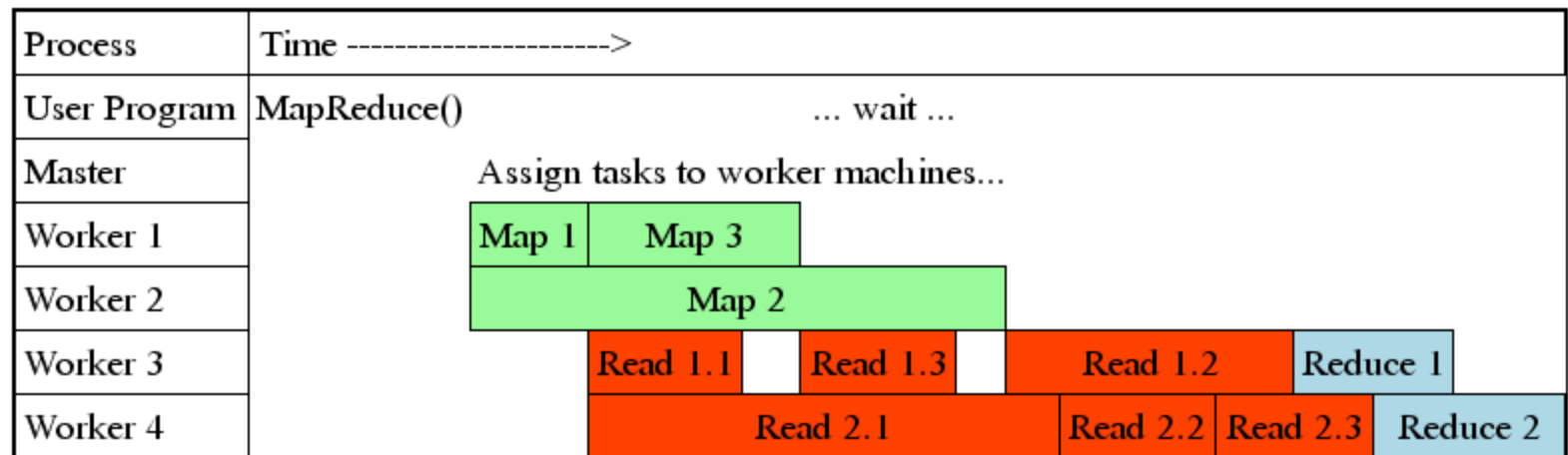
# Locating Stragglers

- How does Hadoop locate stragglers?
  - Hadoop monitors each task progress using a *progress score* between 0 and 1
  - If a task's progress score **is less than** (average - 0.2), and the task has run for at least 1 minute, it is marked as a straggler



# Task Granularity & Pipelining

- **Fine granularity tasks:** map tasks >> machines
  - Minimizes time for fault recovery
  - Can do pipeline shuffling with map execution
  - Better dynamic load balancing



# Bigger Picture: Hadoop vs. Other Systems

## Distributed Databases

## Hadoop

### Computing Model

- Notion of transactions
- Transaction is the unit of work
- ACID properties, Concurrency control

- Notion of jobs
- Job is the unit of work
- No concurrency control

### Data Model

- Structured data with known schema
- Read/Write mode

- Any data will fit in any format
- (un)(semi)structured
- ReadOnly mode

### Cost Model

- Expensive servers

- Cheap commodity machines

### Fault Tolerance

- Failures are rare
- Recovery mechanisms

- Failures are common over thousands of machines
- Simple yet efficient fault tolerance

## Cloud Computing

### Key Characteristics

Efficiency, optimizations, fine-tuning

A computing model where any computing infrastructure can run on the cloud

- Hardware & Software are provided as remote services
- Elastic: grows and shrinks based on the user's demand
- Example: Amazon EC2

