Analytics and Visualization of Big Data

Fadel M. Megahed

Lecture 04: Map-Reduce and an Introduction to AWS

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING

Department of Industrial and Systems Engineering

Spring 13

Chapter 02: Map-Reduce (Cont.)

What is a Distributed File System?

- Motivation for a Distributed File System
- Physical Organization of Compute Nodes
- Large-Scale File-System Organization

Map-Reduce

- The Map Tasks
- Grouping and Aggregation
- The Reduce Tasks
- Combiners
- Coordination
- Coping with Node Failures

Motivation Example: Reading 5M Books with a Program ☺

- Counting the number of times each distinct word or phrase appears in a large collection of documents.

Source: http://www.ted.com/playlists/56/making_sense_of_too_much_data.html

An Overview of Map-Reduce Programming

- It is a style of computing that has been implemented in several systems:
  - Google's own implementation → map-reduce
  - Several popular open-source implementations → Hadoop

- Map-reduce can manage large-scale computations in a way that is tolerant to hardware failures.

- All you need to write are two functions **Map** and **Reduce**, while the system:
  - Manages the parallel execution
  - Coordinates the tasks that execute Map and Reduce
  - Deals with the possibility that one of the tasks will fail

An Overview of Map-Reduce Programming

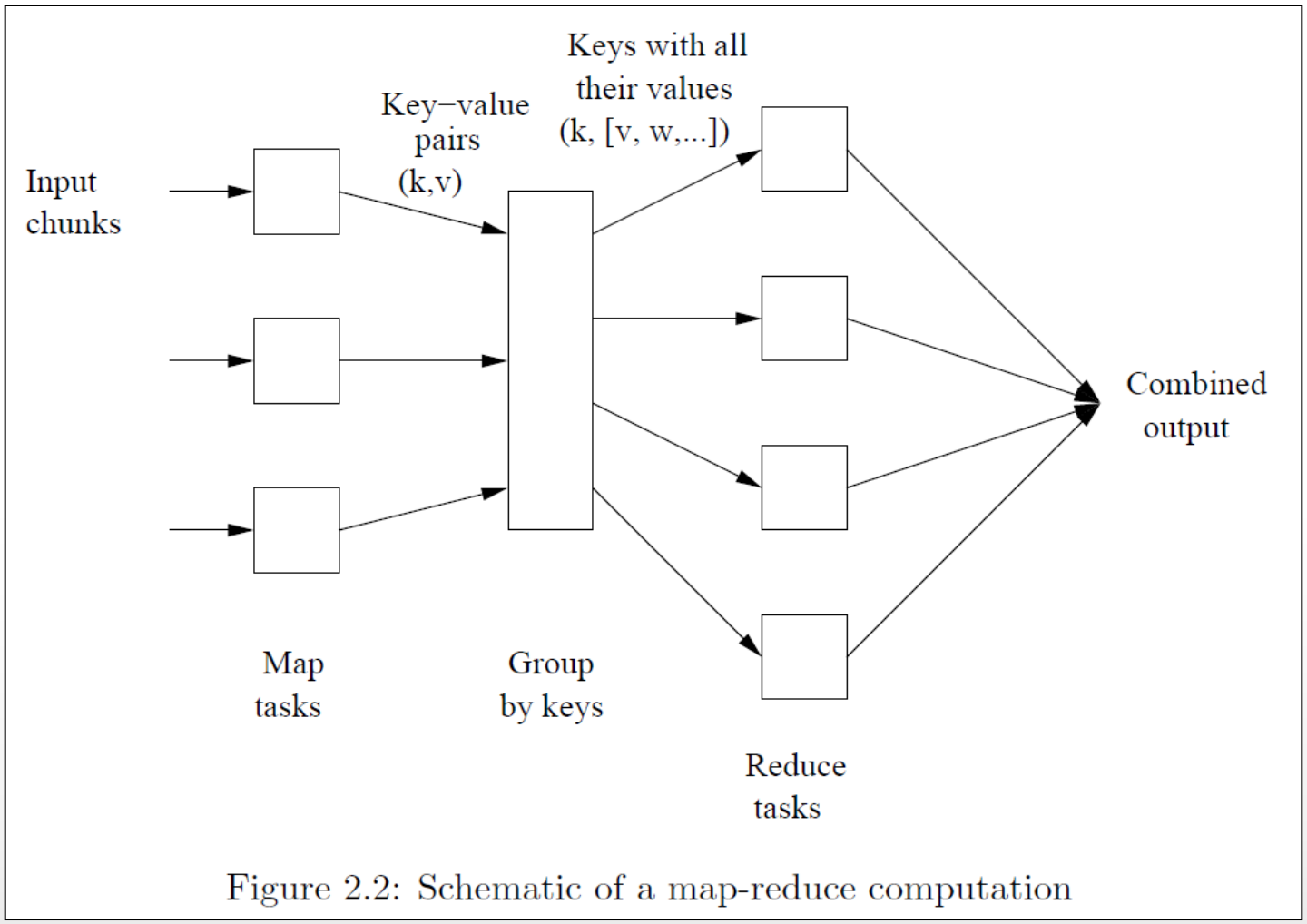
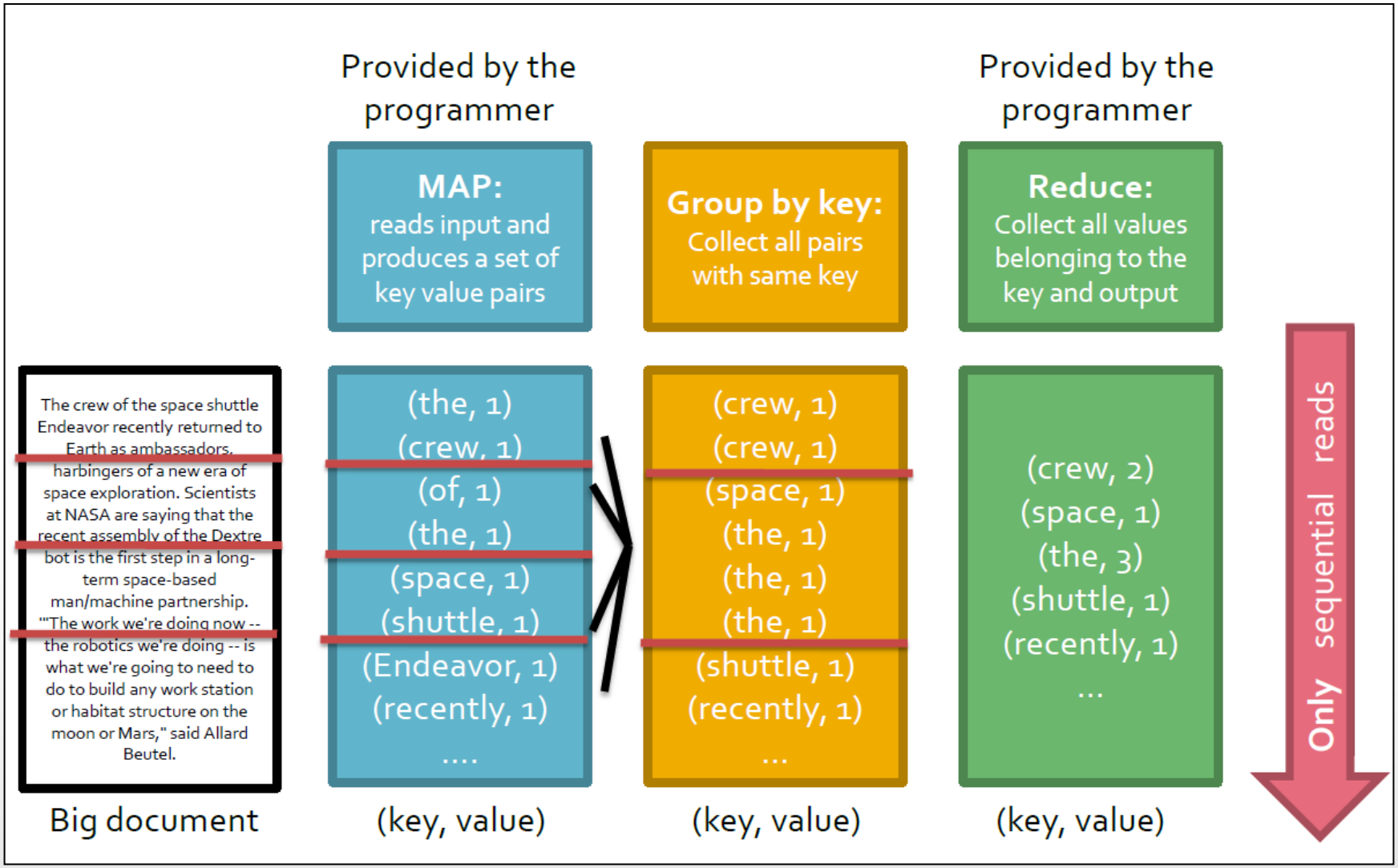


Figure 2.2: Schematic of a map-reduce computation

Source: A. Rajaraman, J. Leskovec, J.D. Ullman. (2012). "Mining of Massive Datasets". http://i.stanford.edu/~ullman/mmds.html

The Use of Map-Reduce for Word Counting



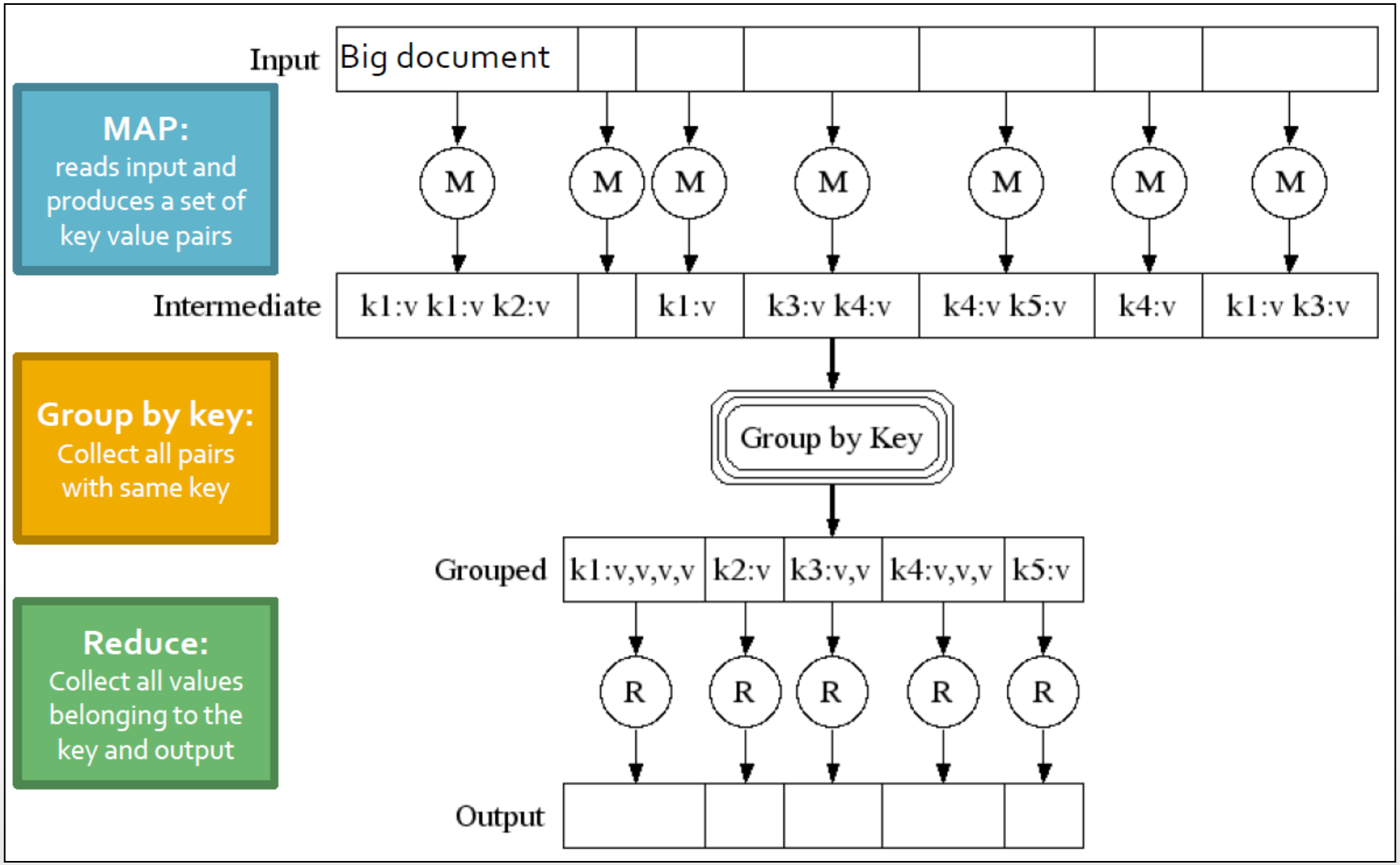
Source: Figure from Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

Map-Reduce: Environment

Map-Reduce environment takes care of:

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

Map-Reduce: Another Representation



Source: Figure from Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

Data Flow in a Map-Reduce Environment

- **Input and final output** are stored on **a distributed file system (DFS)**:
  - 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 map reduce task

Source: Slide is adapted from Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

Coordination – How compute nodes, tasks and files interact?

- Normally a compute node (worker) handles either map tasks or reduce tasks but not both.

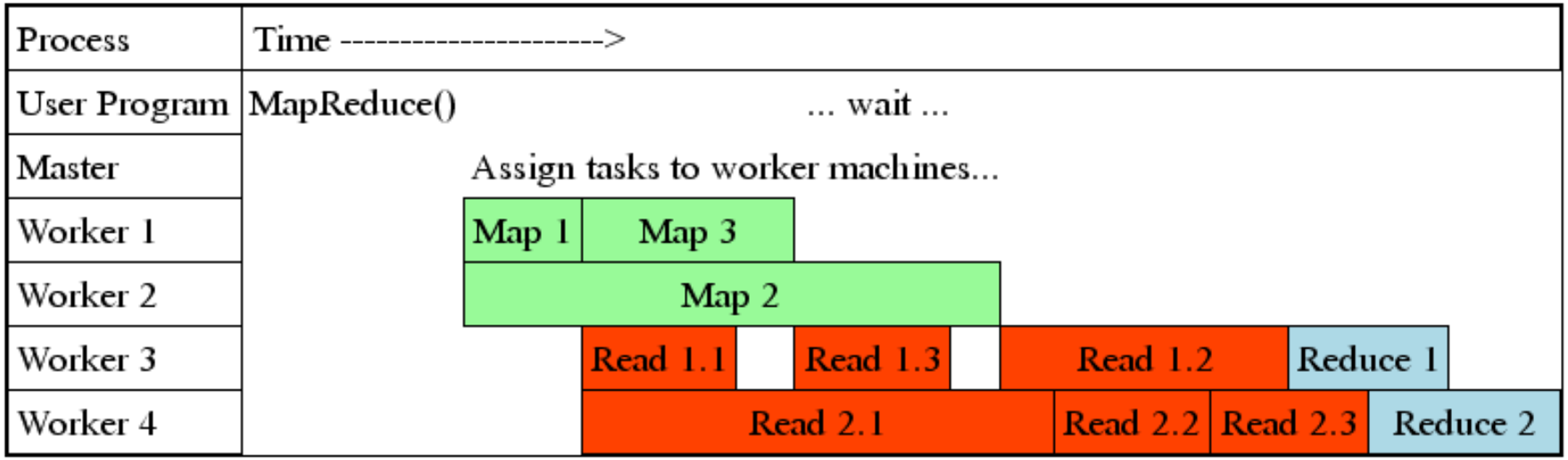
- The master node has many responsibilities:
  - Task status: (idle, in-progress, completed)
  - Idle tasks get scheduled as workers become available
  - When a map task completes, it sends the master the location and sizes of its R intermediate files, one for each reducer
  - Master pushes this info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

Coping with Node 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
- **Master failure**
  - Map-Reduce task is aborted and client is notified (In case of one master node)

Similarities of Map-Reduce to MRP???

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



Source: Slide is adapted from Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

An Introduction to Amazon Web Services

Amazon Web Services

- AWS) is an '*Infrastructure as a Service*' (IaaS) provider.
- AWS' cloud based computing offers dynamically scalable computing, storage, and data access "on demand" over the Internet, with "pay as you go" pricing for the hardware and software that is delivered.
- AWS includes over two dozen cloud-related services, including their popular elastic computing (EC2) and storage (S3) capabilities. Broadly speaking, their cloud services can be grouped into 3 categories, infrastructure services, platform services and supporting services.

Source: http://aws.amazon.com/

Amazon Web Services – Getting Started

Source: http://aws.amazon.com/

Analytics and Visualization of Big Data

Fadel M. Megahed

Lecture 04: Map-Reduce and an Introduction to AWS

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING

Department of Industrial and Systems Engineering

Spring 13