CC5212-1
Procesamiento Masivo de Datos
Otoño 2017

Lecture 4: MapReduce/Hadoop II

Aidan Hogan
aidhog@gmail.com

HADOOP: IN MORE DEPTH

Why “Hadoop” …

… and why an elephant and not, e.g., a rabbit?

Why “Hadoop” …

… and why an elephant and not, e.g., a rabbit?

… better luck next time, Tiddles.

HADOOP: UNDER THE HOOD

MapReduce Abstraction

Can we abstract any general framework?

Define input as a set of key value pairs $I \subseteq T_{IK} \times T_{IV}$

For example, $I = \{(1, \texttt{"soy una linea"}), (2, \texttt{"soy otra linea"})\}$
T_{IK} is the set of all `int`, T_{IV} is the set of all `string`

Define map as a function $I \to 2^M$ where $M \subseteq T_{MK} \times T_{MV}$

For example, map$(1, \texttt{"soy una linea"}) := \{(\texttt{"soy"}, 1), (\texttt{"una"}, 1), (\texttt{"linea"}, 1)\}$
T_{MK} is the set of all `string`, T_{MV} is the set of all `int`

Define reduce as a function $T_{MK} \times 2^{T_{MV}} \to 2^R$ where $R \subseteq T_{RK} \times T_{RV}$

For example, reduce$(\texttt{"soy"}, \{1, 1\}) := \{(\texttt{"soy"}, 2)\}$
T_{RK} is the set of all `string`, T_{RV} is the set of all `int`

In general, we must assume bags/multisets (sets with duplicates)

MapReduce: Main Idea

Given I

... compute map over all $i \in I$

... group resulting set by map key

... apply reduce over groups

← But how to implement this part in a distributed system

1. Partition by map key
2. Sort (in parallel) by map key
3. Apply reduce / Profit

... let's be more specific

MapReduce/Hadoop

1. Input

2. Map

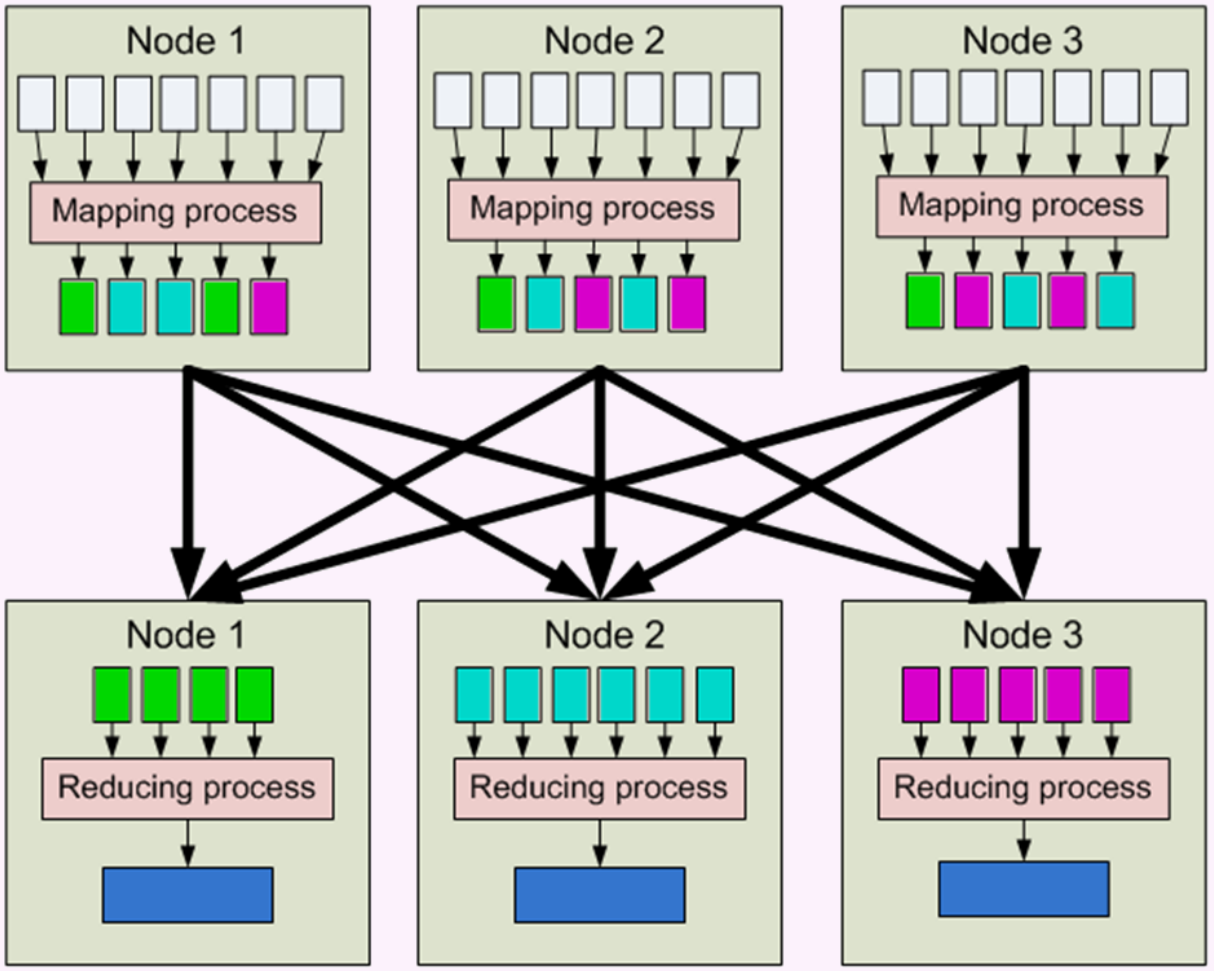
3. Partition [Sort]

4. Shuffle

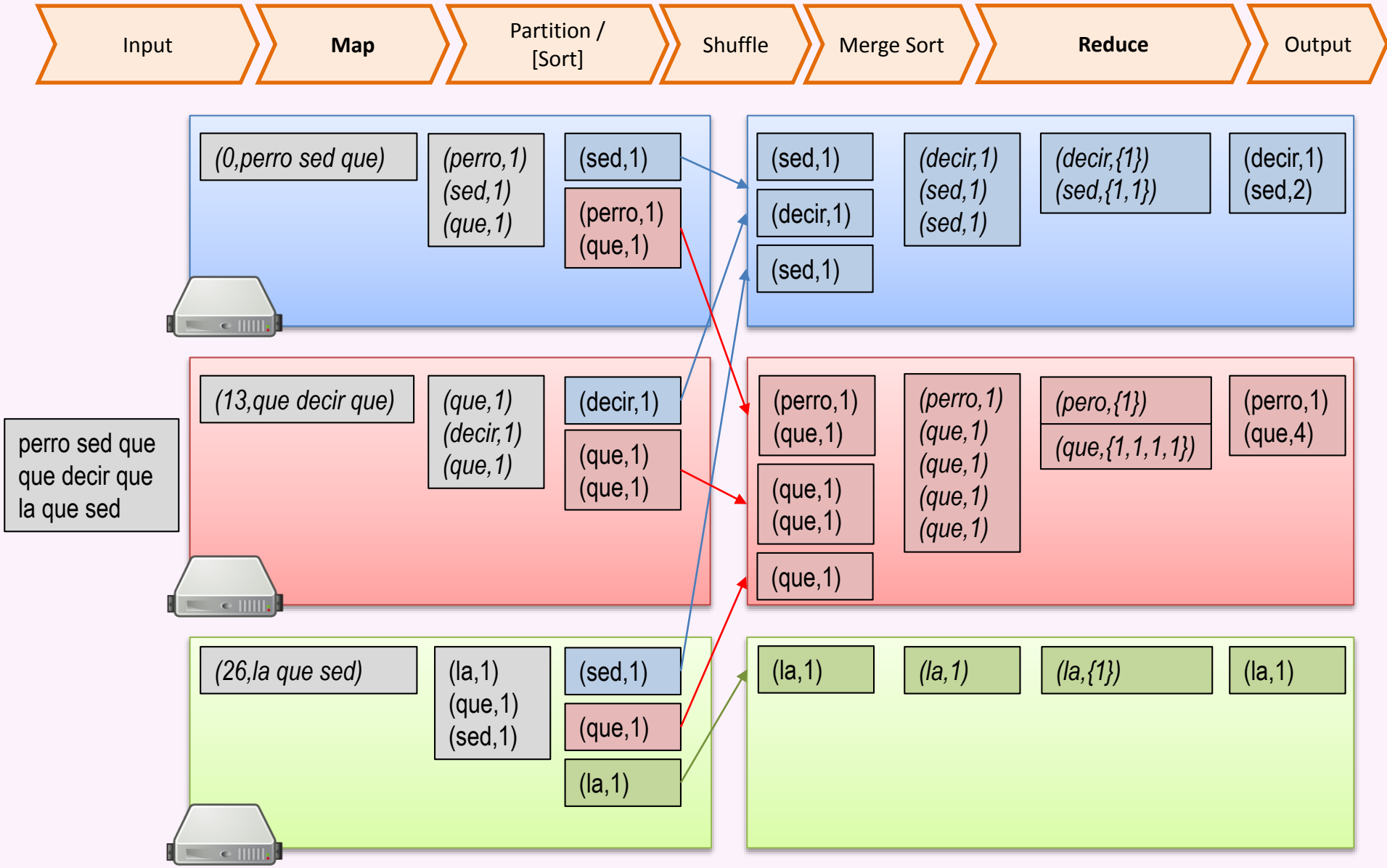
5. Merge Sort

6. Reduce

7. Output

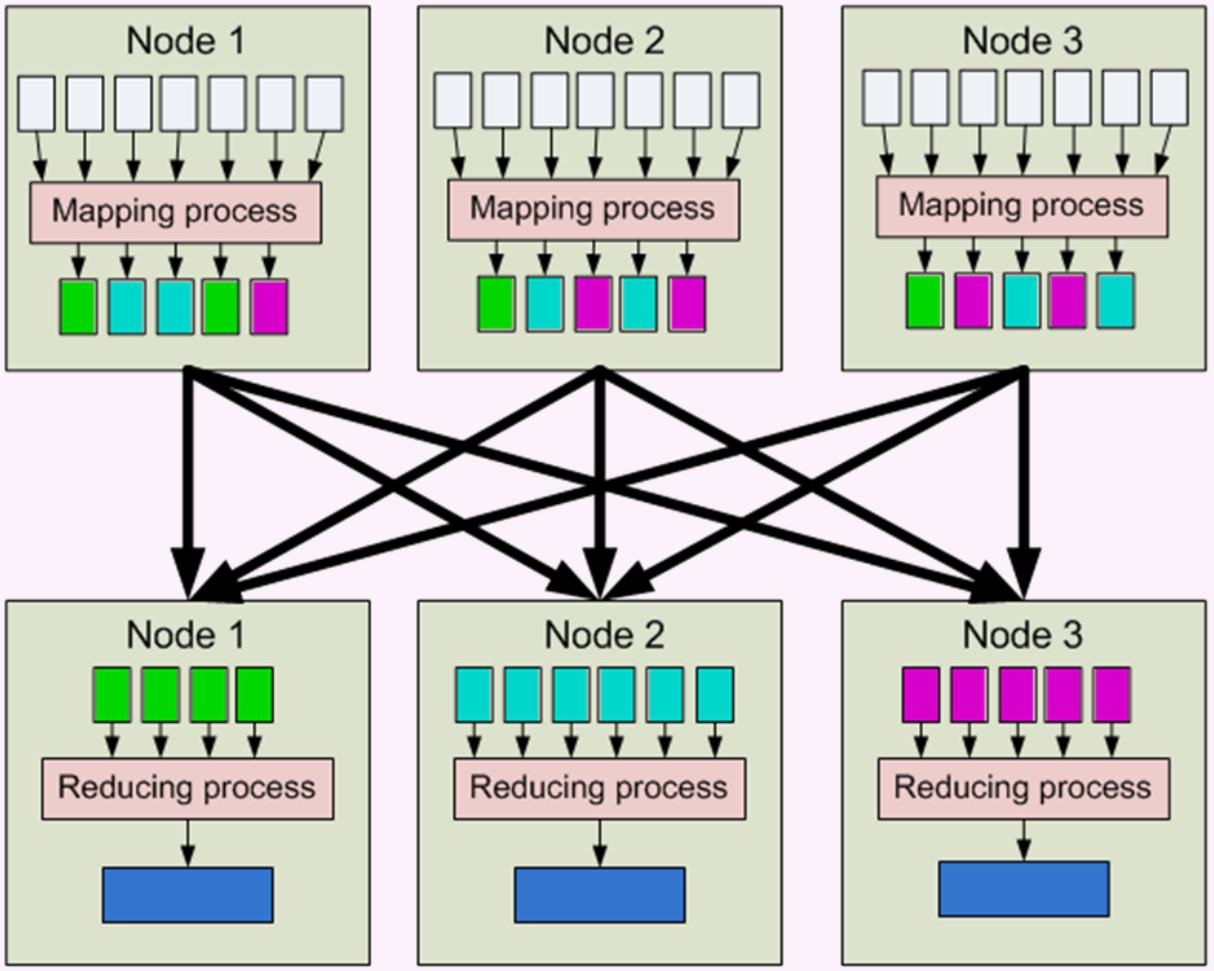


MapReduce/Hadoop: Counting Words

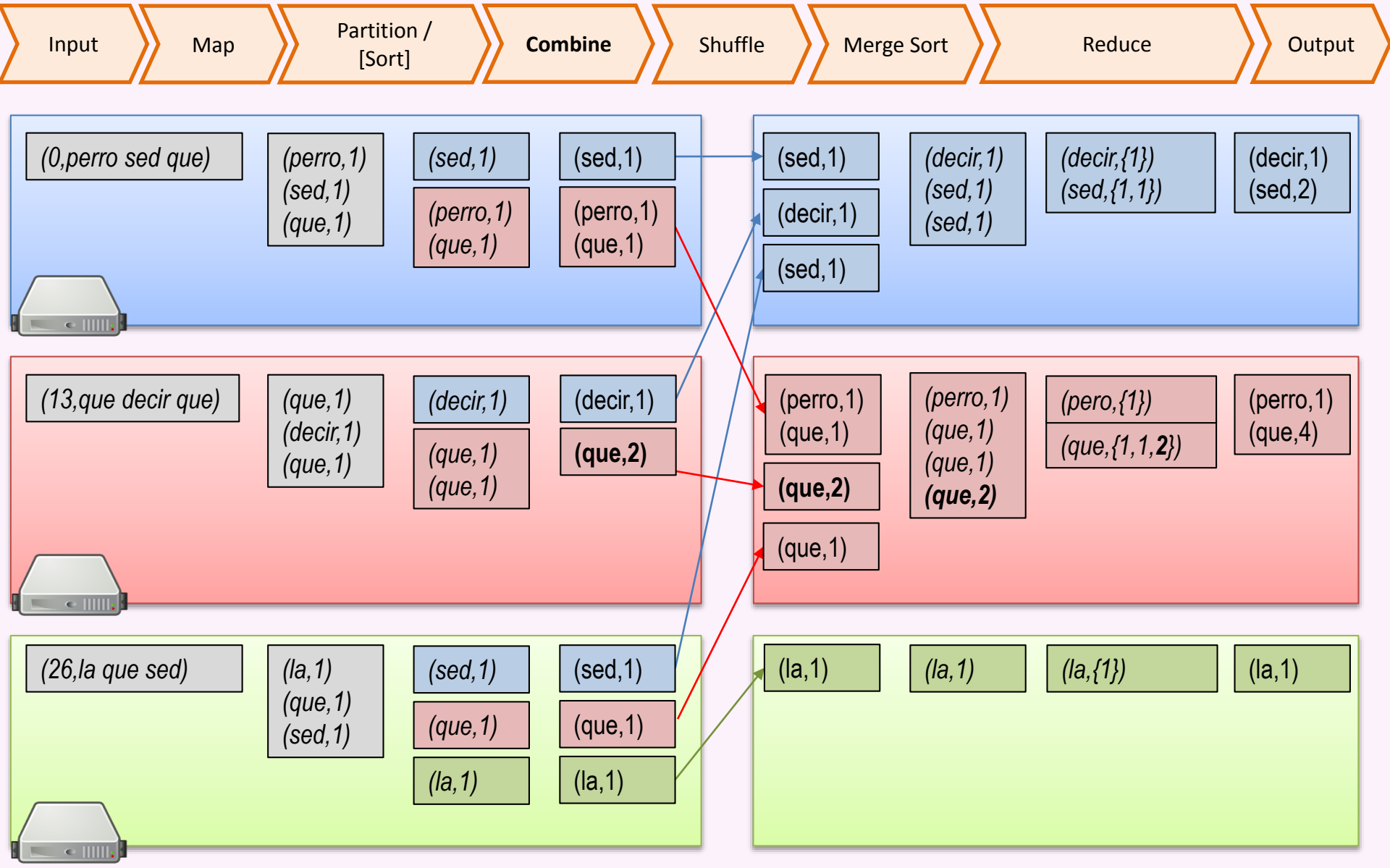


MapReduce/Hadoop: Combiner

1. Input
2. **Map**
3. Partition [Sort]
   (“*Combine*”)
4. Shuffle
5. Merge Sort
6. **Reduce**
7. Output

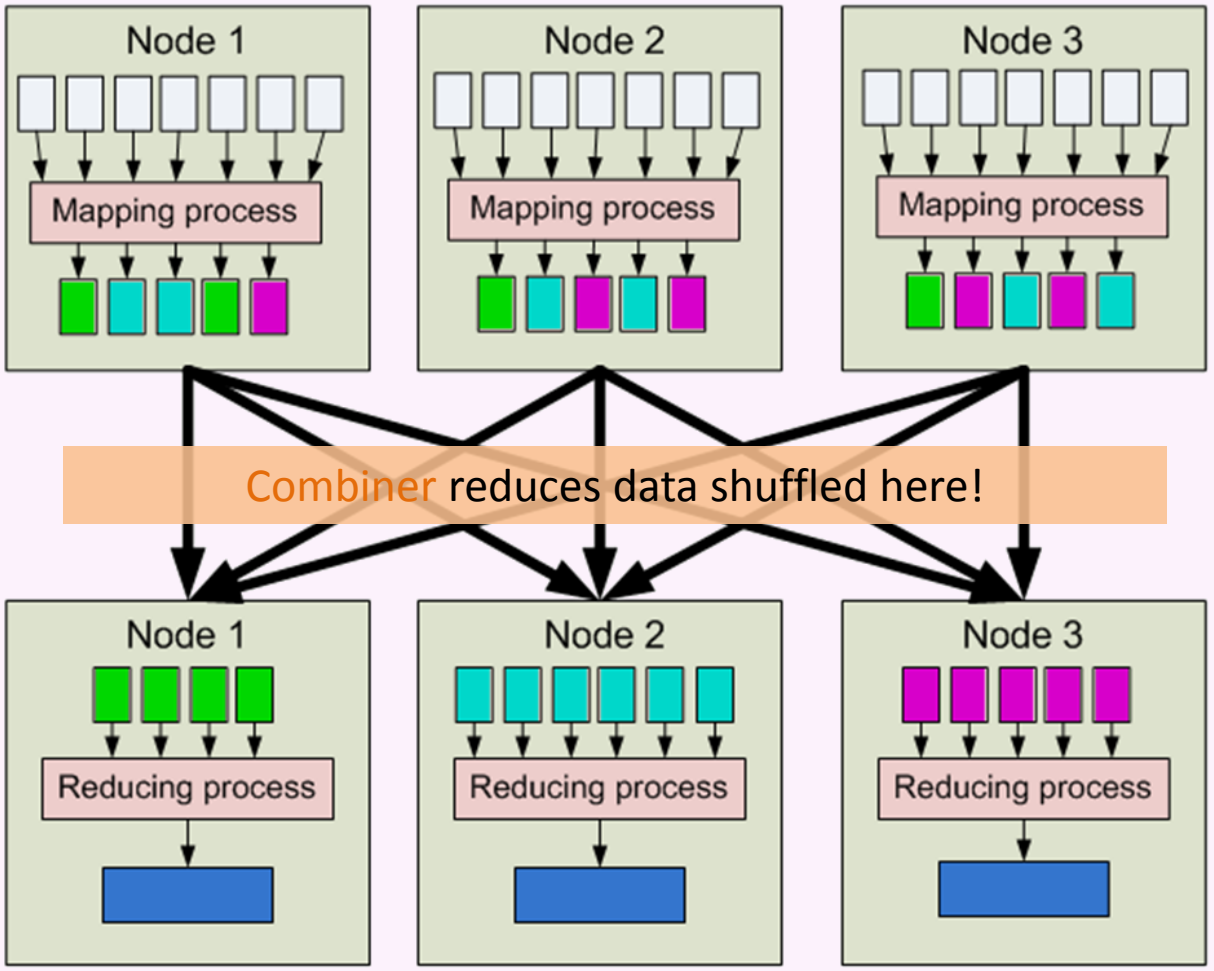


MapReduce/Hadoop: Combiner



MapReduce/Hadoop: Combiner

1. Input
2. **Map**
3. Partition [Sort]
   (“*Combine*”)
4. Shuffle
5. Merge Sort
6. **Reduce**
7. Output



IN-CLASS EXERCISES

Supermarket Example

MapReduce: Supermarket Example

ReceiptItems

Receipt ID	Item ID
R1401	I306
R1401	I306
R1401	I504
R1402	I007
R1402	I306
R1403	I306
R1403	I504
...	...

ReceiptTimes

Receipt ID	Time
R1403	19:00
R1401	18:59
R1402	19:01
...	...

ItemDetails

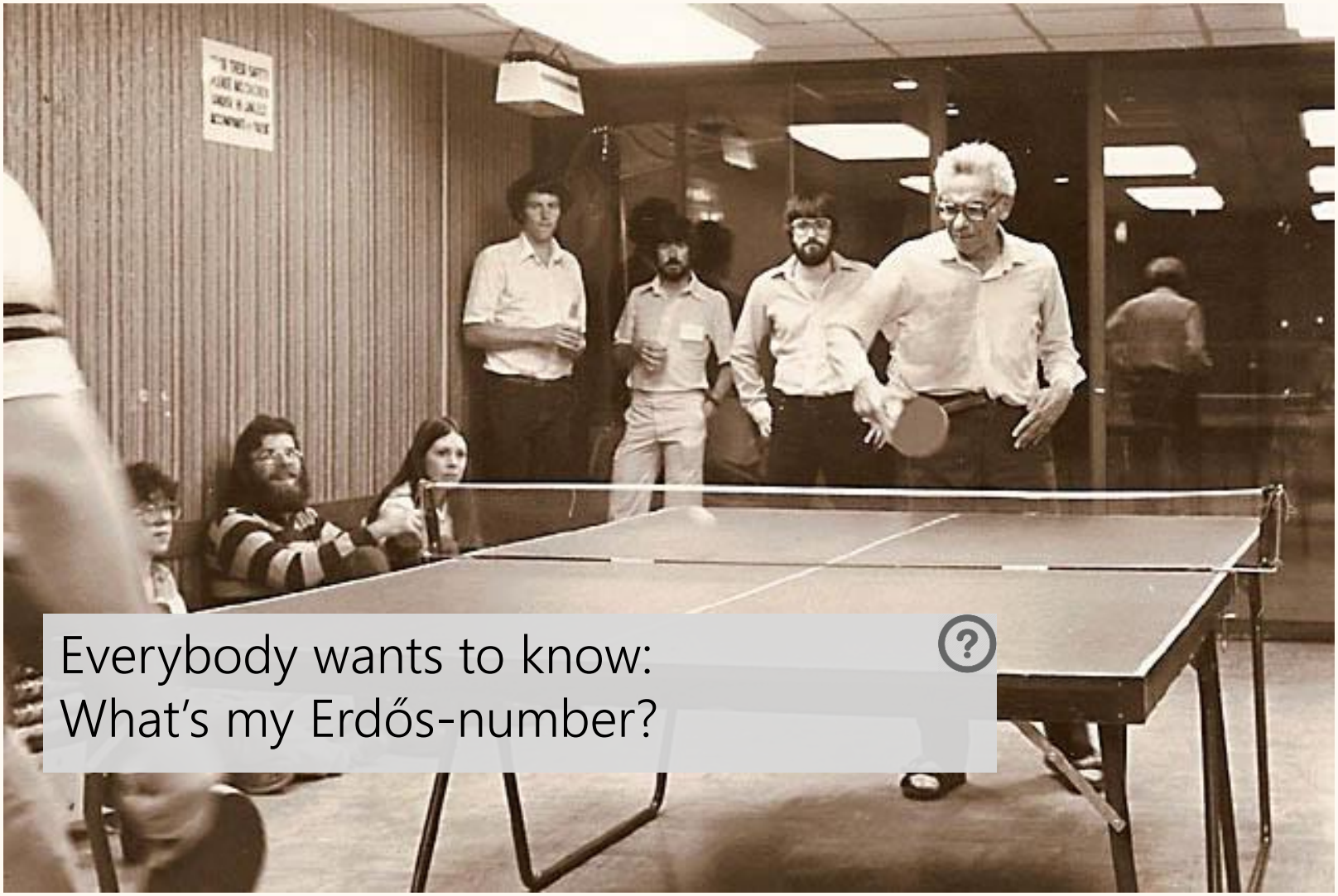
Item ID	Name	Price ($)
I306	Zanahoria 500g	500
I504	CocaCola 3L	1400
I007	Comfort	1200
...	...	

Compute total sales per hour of the day?

Output

Hour	Total
...	...
18:00–18:59	$2400
19:00–19:59	$3600
...	...

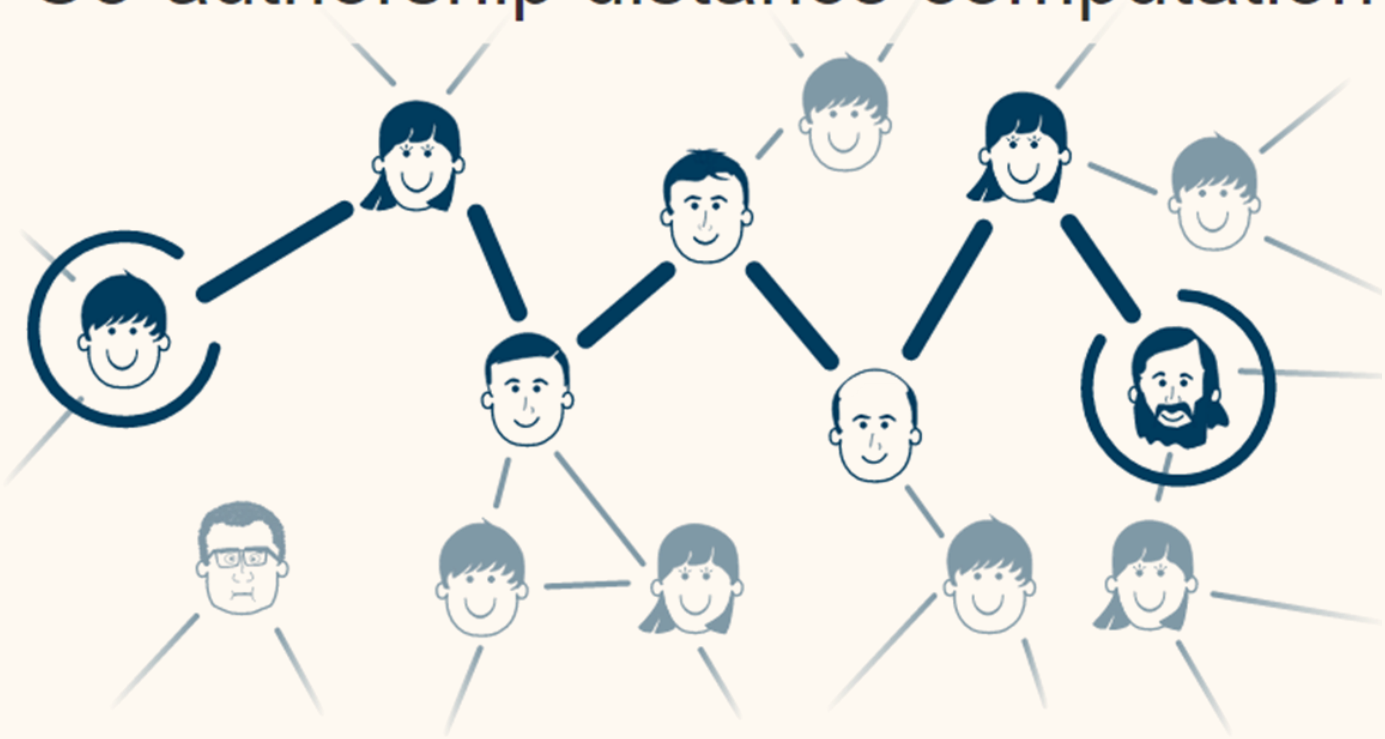
Erdős-number Example



Everybody wants to know:
What's my Erdős-number?

csauthors.net

Co-authorship distance computation



Find the path between two authors:

Paul Erdős

Erdős-number Example

Input: Authors.tsv

Author	Paper
Aidan Hogan	IMGpedia: Enriching the Web of Data with Image Content Analysis
Benjamin Bustos	IMGpedia: Enriching the Web of Data with Image Content Analysis
Benjamin Bustos	Scalability of Non-Rigid 3D Shape Retrieval
H. Li	A new class of Ramsey-Turán problems
H. Li	Scalability of Non-Rigid 3D Shape Retrieval
Paul Erdős	Random induced graphs
Richard H. Schelp	A new class of Ramsey-Turán problems
Richard H. Schelp	Random induced graphs
Sebastian Ferrada	IMGpedia: Enriching the Web of Data with Image Content Analysis
...	...

Who has an Erdős-number of 3?

Output

Author
Sebastián Ferrada
Aidan Hogan
...

Questions?