Map Reduce

Map Reduce

- Algoritmo eficiente para computación paralela/distribuida
- Paradigma de programación - mezcla de datos con controlador
- Sacrificios para acelerar computación

Computación Distribuida

- Datos tan grandes que no caben en un computador
- Repartidos en muchos servidores
- Cada servidor no conoce lo que tiene el resto
- No podemos dar el lujo de comunicar todos los datos

Map Reduce

Ejemplo: ver cuantas veces ocurre cada palabra en un texto T

¿Cómo hacer esto en un archivo de 100 petabytes?

Map Reduce

- **Map**: recibe datos y genera pares key – values
- **Reduce**: recibe pares con el mismo key y los agrega
- **Shuffle**: transfiere los datos desde mappers a reducers

Map Reduce

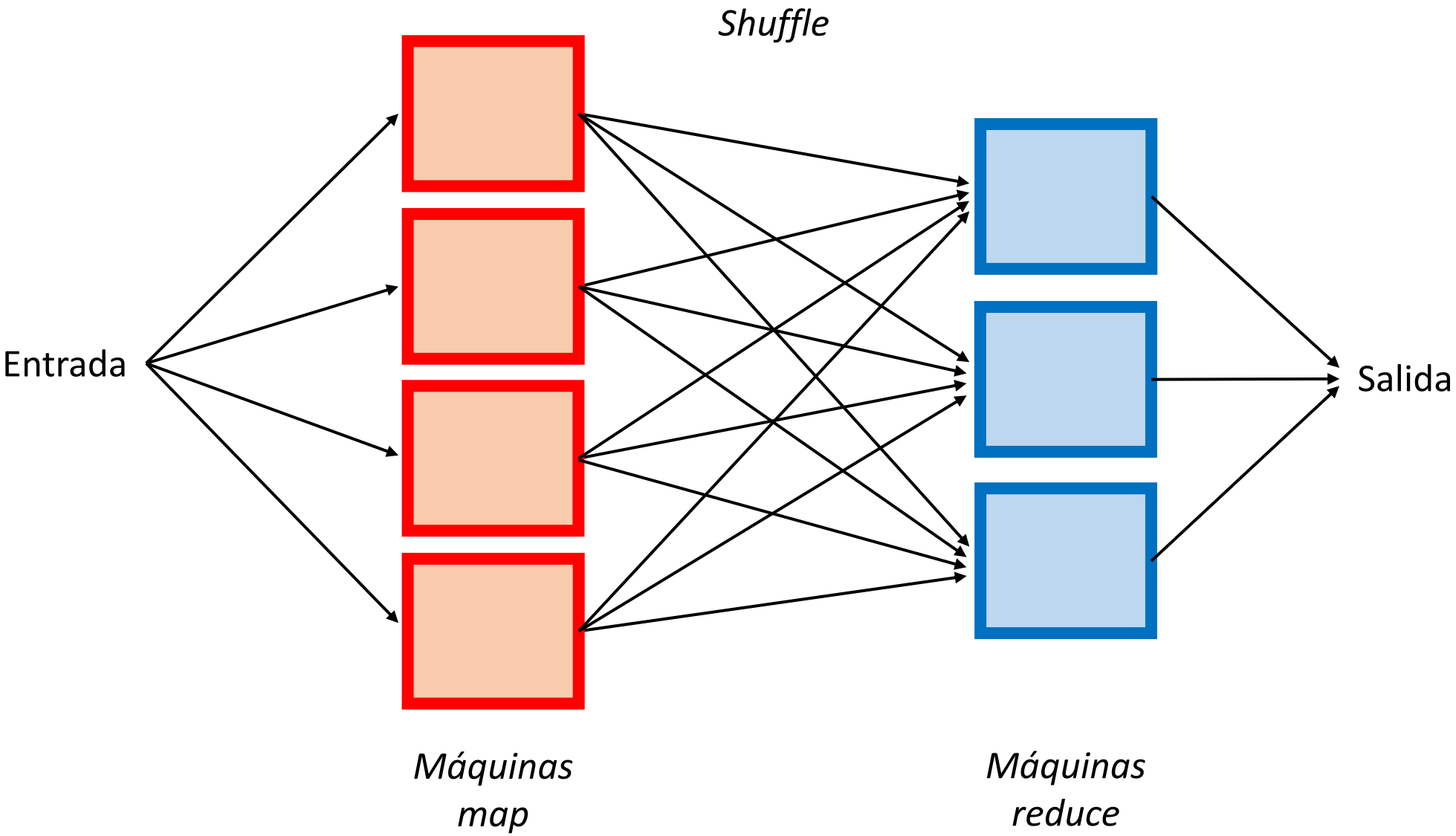
Arquitectura

Mappers:

- Nodos encargados de hacer Map
- Reciben parte del documento y lo envían a los reducers (a través de **shuffle**)

Reducers:

- Nodos encargados de hacer Reduce
- Reciben los Map y los agregan
- El output es la unión de cada Reducer


Shuffle
Entrada
Salida
Máquinas map
Máquinas reduce

Map Reduce

Ejemplo

¿Cuántas veces cada palabra ocurre en un archivo de texto grande?

Map Reduce

Ejemplo

Map:

- Recibe un pedazo de texto
- Por cada palabra, emite el par (palabra, número de ocurrencias)

Reduce:

- Cada reduce recibe todos los pares asociados a la misma palabra
- Junta todos estos pares y suma las ocurrencias

INPUT

hola que hola año zzz hola que zzz que

Máquina map1

Máquina map2

Máquina map3

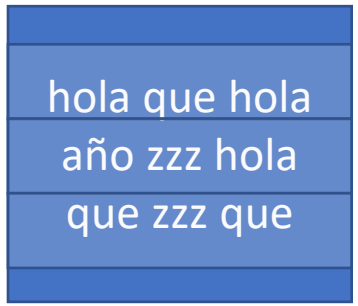
Máquina reduce1

Palabras A-M

Máquina reduce2

Palabras N-Z

SEPARAR INPUT



Máquina map1

Máquina map2

Máquina map3

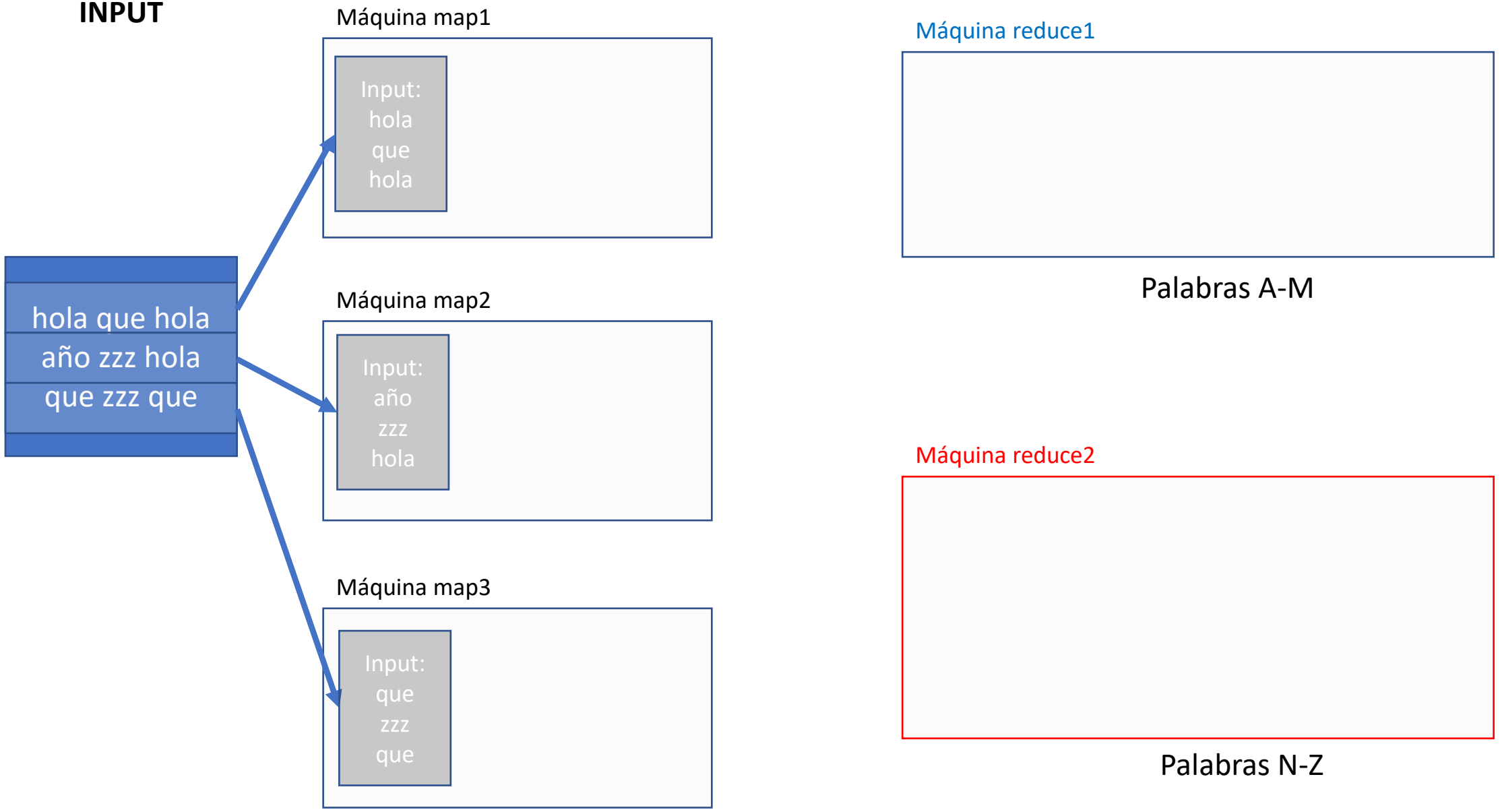
Máquina reduce1

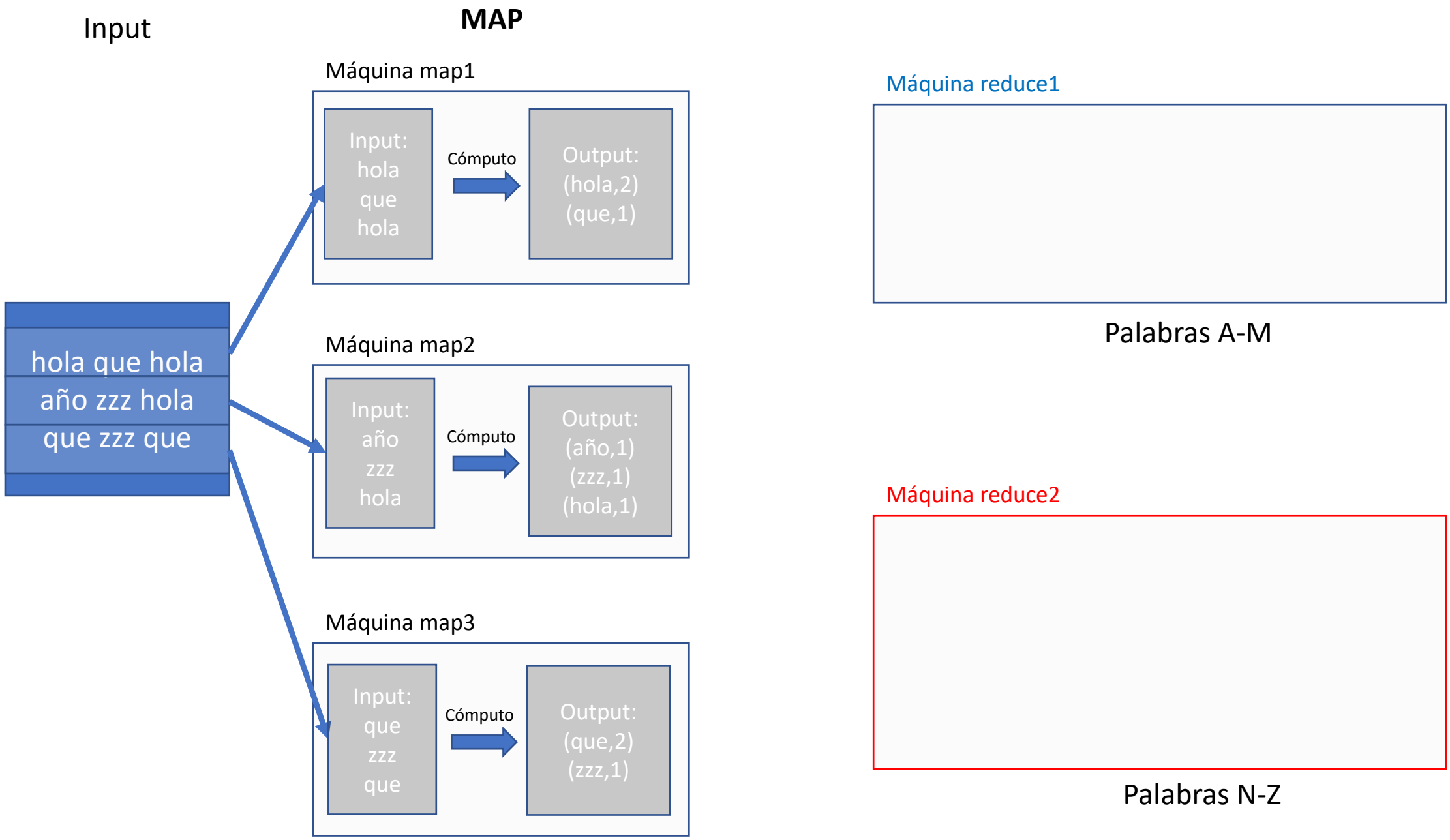
Palabras A-M

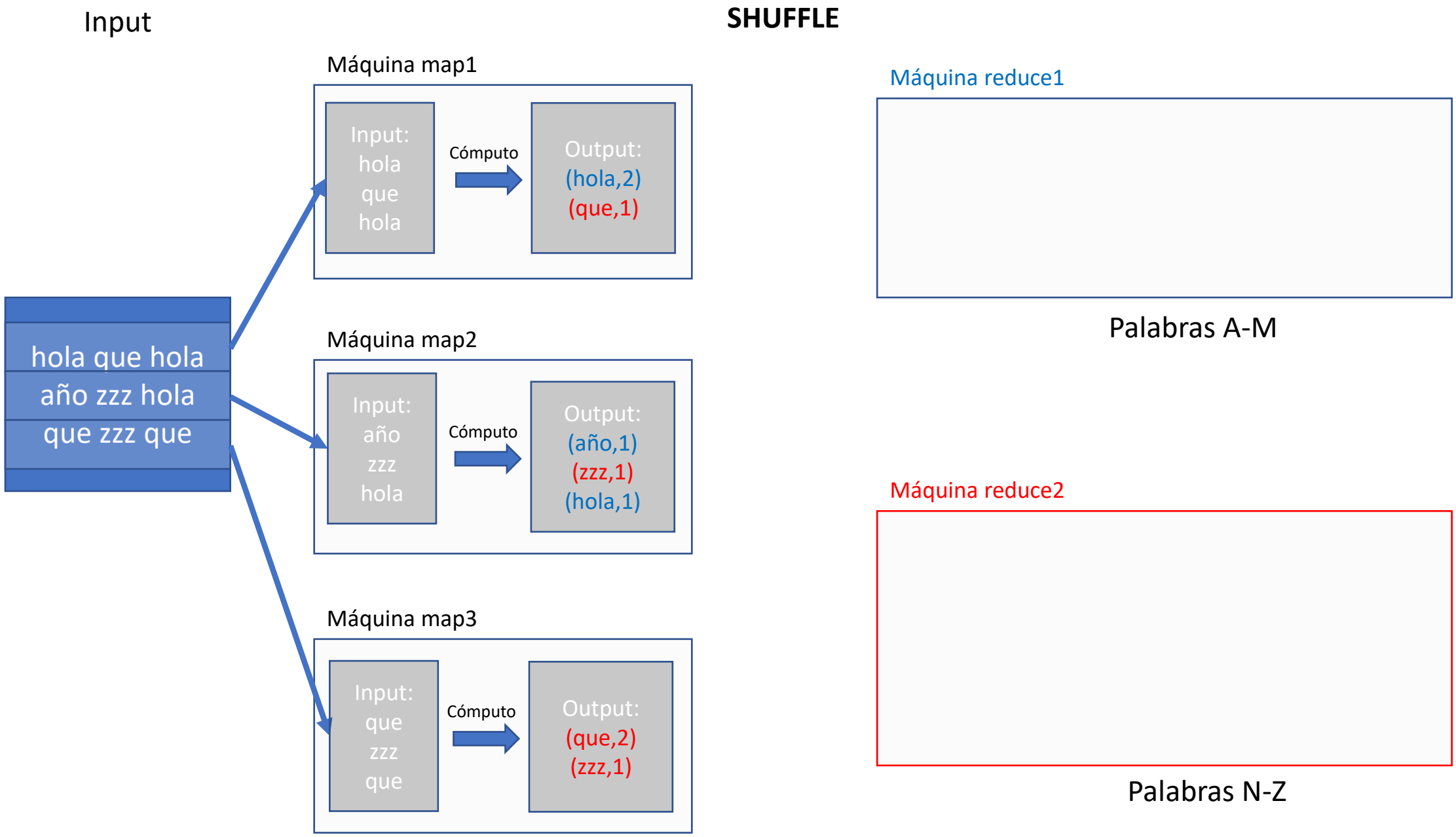
Máquina reduce2

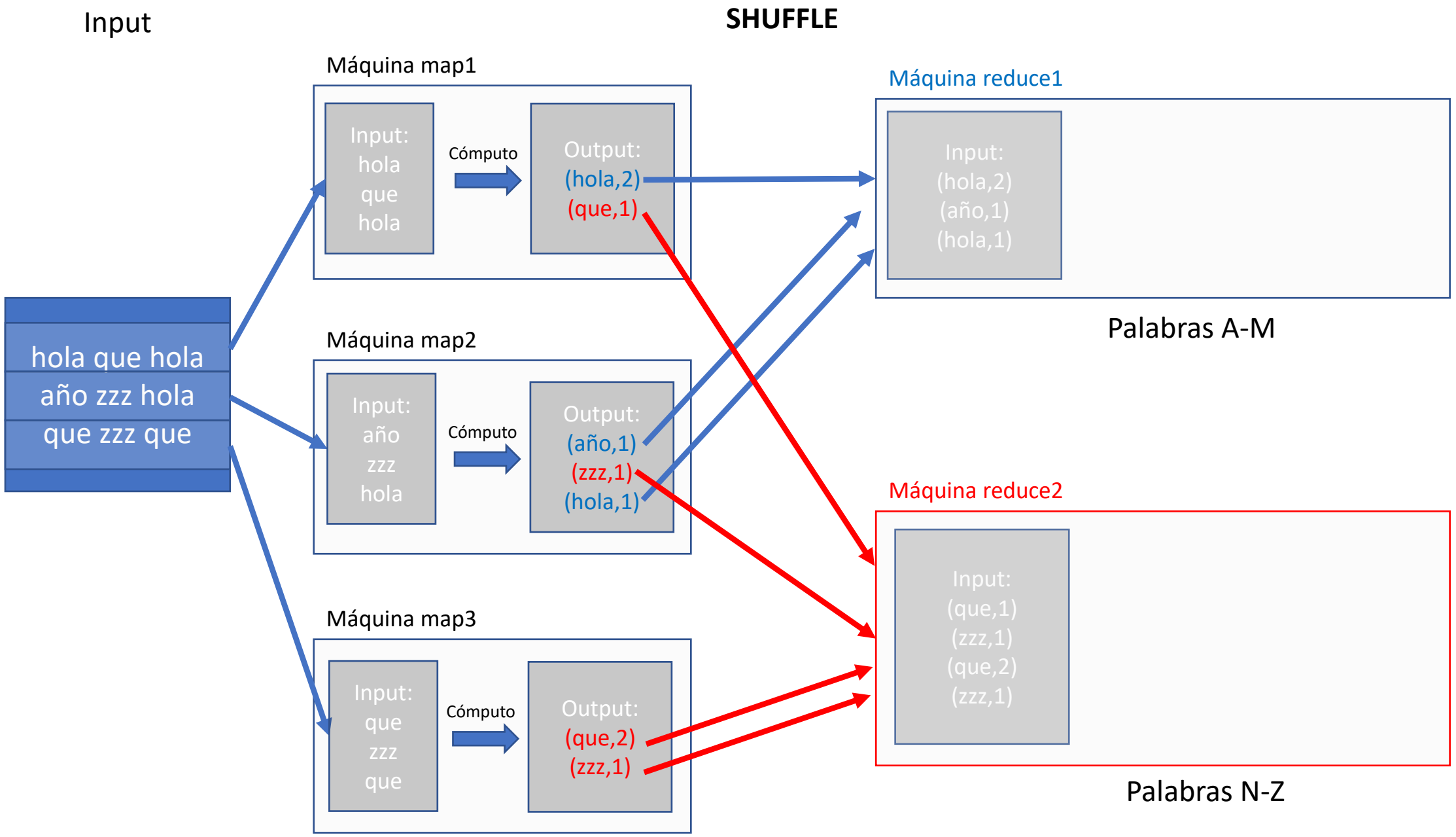
Palabras N-Z

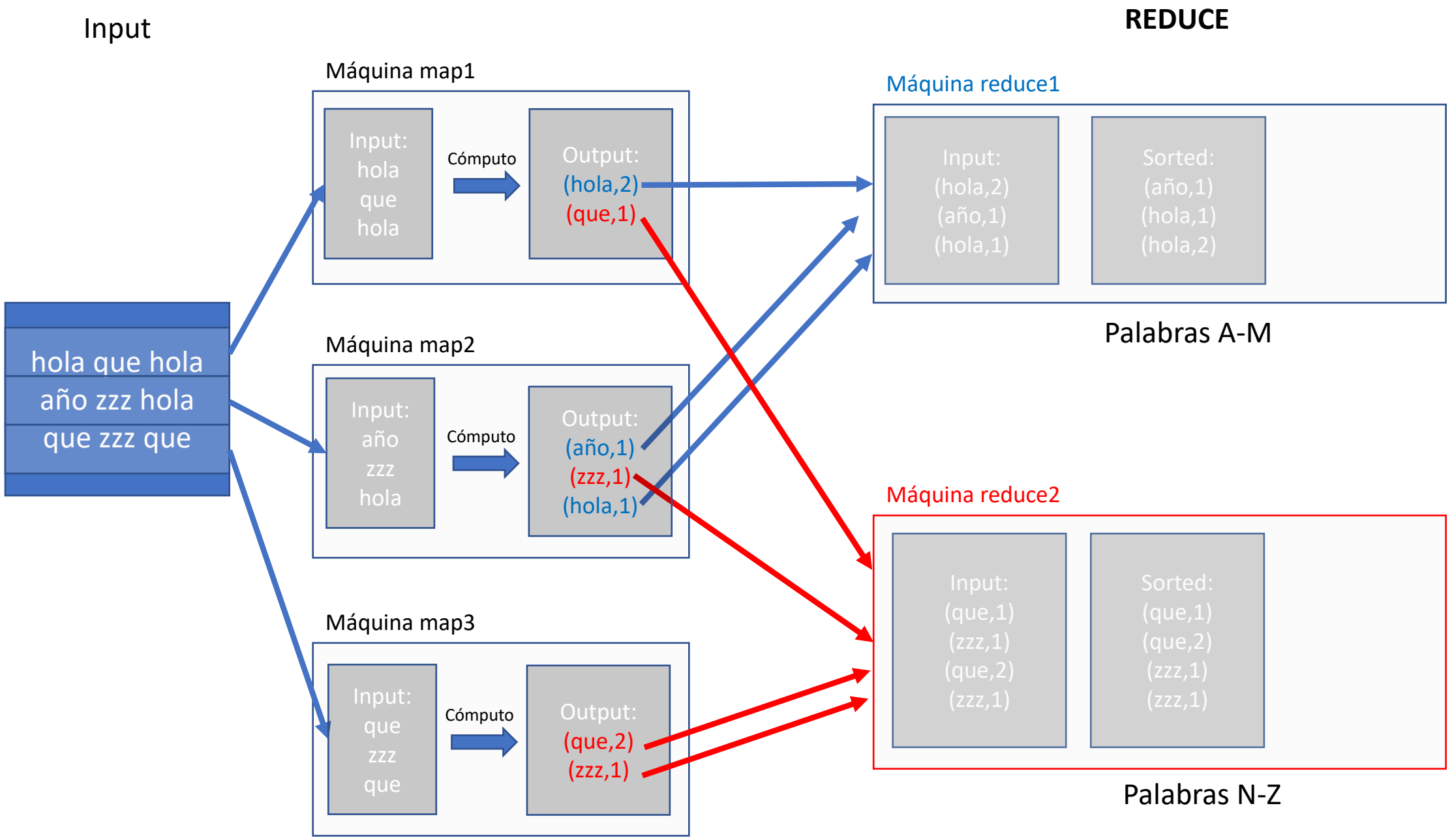
SEPARAR INPUT

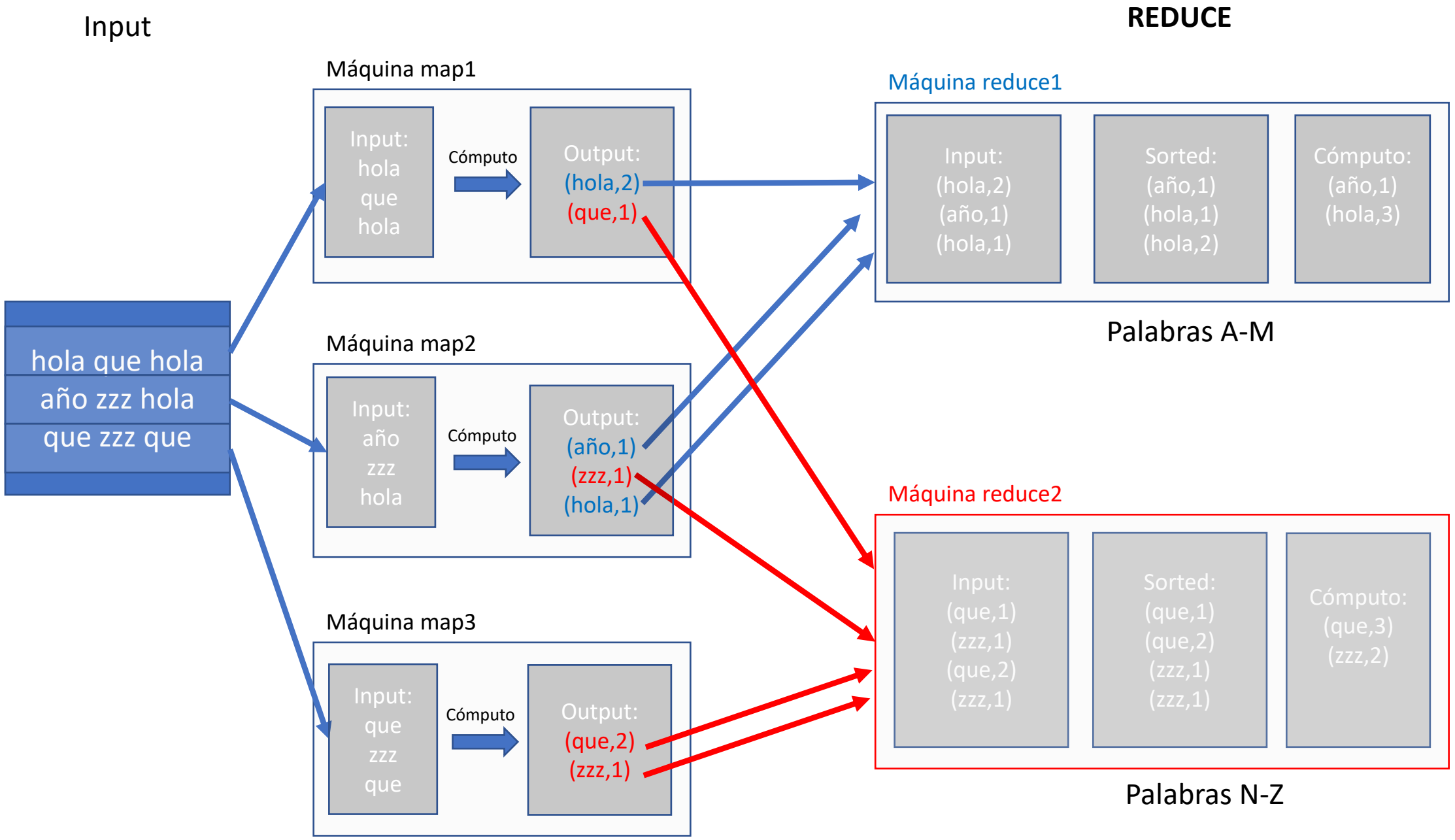


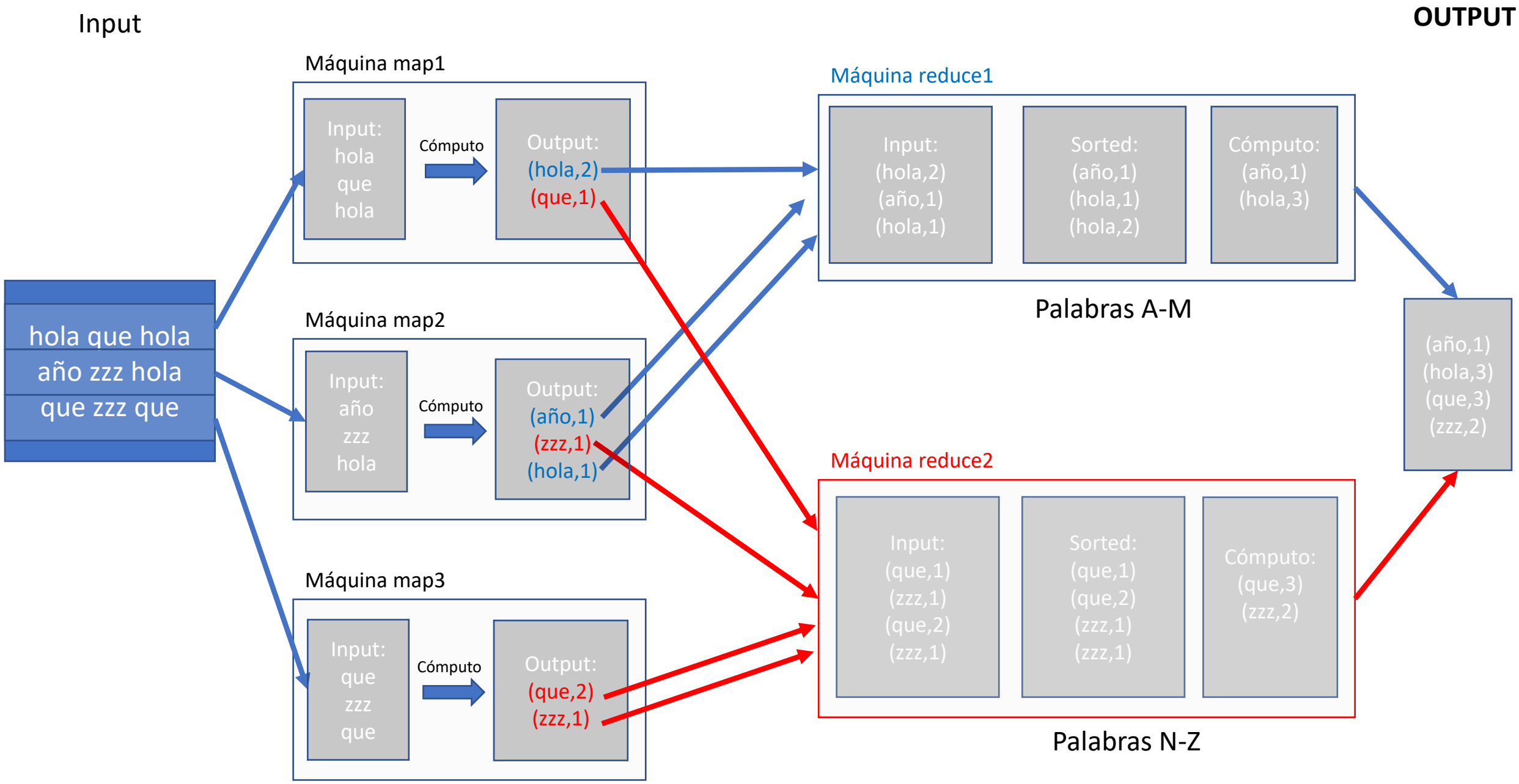

Input
MAP
Máquina map1
Input:
hola
que
hola
Cómputo
Output:
(hola,2)
(que,1)
hola que hola
año zzz hola
que zzz que
Máquina map2
Input:
año
zzz
hola
Cómputo
Output:
(año,1)
(zzz,1)
(hola,1)
Máquina map3
Input:
que
zzz
que
Cómputo
Output:
(que,2)
(zzz,1)
Máquina reduce1
Palabras A-M
Máquina reduce2
Palabras N-Z

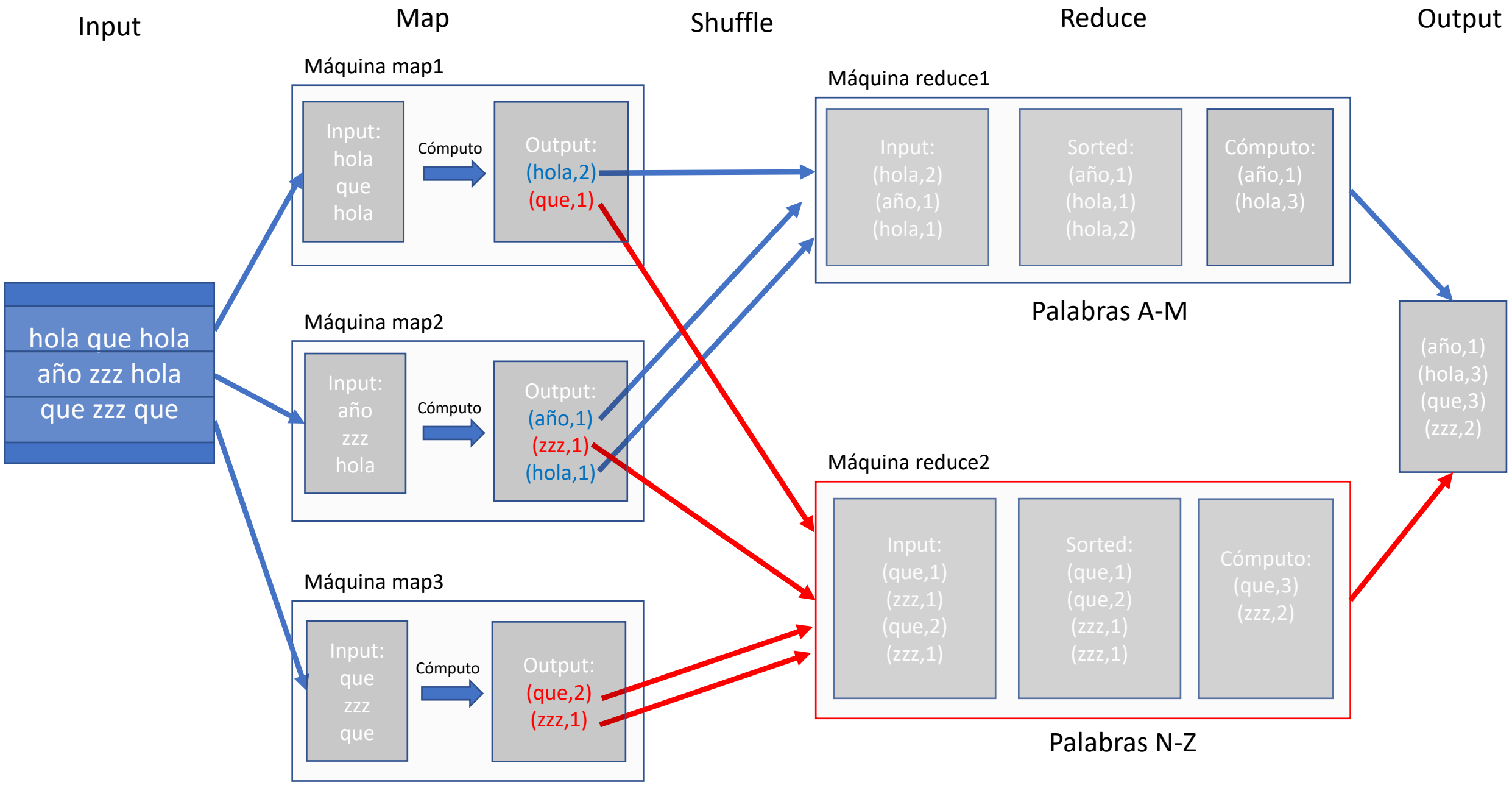

Input
SHUFFLE
Máquina map1
Input:
hola
que
hola
Cómputo
Output:
(hola,2)
(que,1)
Máquina map2
Input:
año
zzz
hola
Cómputo
Output:
(año,1)
(zzz,1)
(hola,1)
Máquina map3
Input:
que
zzz
que
Cómputo
Output:
(que,2)
(zzz,1)
hola que hola
año zzz hola
que zzz que
Máquina reduce1
Palabras A-M
Máquina reduce2
Palabras N-Z


Input
SHUFFLE
hola que hola
año zzz hola
que zzz que
Máquina map1
Input:
hola
que
hola
Cómputo
Output:
(hola,2)
(que,1)
Máquina map2
Input:
año
zzz
hola
Cómputo
Output:
(año,1)
(zzz,1)
(hola,1)
Máquina map3
Input:
que
zzz
que
Cómputo
Output:
(que,2)
(zzz,1)
Máquina reduce1
Input:
(hola,2)
(año,1)
(hola,1)
Palabras A-M
Máquina reduce2
Input:
(que,1)
(zzz,1)
(que,2)
(zzz,1)
Palabras N-Z


Input
REDUCE
Máquina map1
Input:
hola
que
hola
Cómputo
Output:
(hola,2)
(que,1)
Máquina map2
Input:
año
zzz
hola
Cómputo
Output:
(año,1)
(zzz,1)
(hola,1)
Máquina map3
Input:
que
zzz
que
Cómputo
Output:
(que,2)
(zzz,1)
hola que hola
año zzz hola
que zzz que
Máquina reduce1
Input:
(hola,2)
(año,1)
(hola,1)
Sorted:
(año,1)
(hola,1)
(hola,2)
Palabras A-M
Máquina reduce2
Input:
(que,1)
(zzz,1)
(que,2)
(zzz,1)
Sorted:
(que,1)
(que,2)
(zzz,1)
(zzz,1)
Palabras N-Z


Input
REDUCE
Máquina map1
Input:
hola
que
hola
Cómputo
Output:
(hola,2)
(que,1)
Máquina map2
Input:
año
zzz
hola
Cómputo
Output:
(año,1)
(zzz,1)
(hola,1)
Máquina map3
Input:
que
zzz
que
Cómputo
Output:
(que,2)
(zzz,1)
hola que hola
año zzz hola
que zzz que
Máquina reduce1
Input:
(hola,2)
(año,1)
(hola,1)
Sorted:
(año,1)
(hola,1)
(hola,2)
Cómputo:
(año,1)
(hola,3)
Palabras A-M
Máquina reduce2
Input:
(que,1)
(zzz,1)
(que,2)
(zzz,1)
Sorted:
(que,1)
(que,2)
(zzz,1)
(zzz,1)
Cómputo:
(que,3)
(zzz,2)
Palabras N-Z


Input
OUTPUT
Máquina map1
Input:
hola
que
hola
Cómputo
Output:
(hola,2)
(que,1)
Máquina map2
Input:
año
zzz
hola
Cómputo
Output:
(año,1)
(zzz,1)
(hola,1)
Máquina map3
Input:
que
zzz
que
Cómputo
Output:
(que,2)
(zzz,1)
hola que hola
año zzz hola
que zzz que
Máquina reduce1
Input:
(hola,2)
(año,1)
(hola,1)
Sorted:
(año,1)
(hola,1)
(hola,2)
Cómputo:
(año,1)
(hola,3)
Palabras A-M
Máquina reduce2
Input:
(que,1)
(zzz,1)
(que,2)
(zzz,1)
Sorted:
(que,1)
(que,2)
(zzz,1)
(zzz,1)
Cómputo:
(que,3)
(zzz,2)
Palabras N-Z
(año,1)
(hola,3)
(que,3)
(zzz,2)


Input
Map
Shuffle
Reduce
Output
hola que hola
año zzz hola
que zzz que
Máquina map1
Input:
hola
que
hola
Cómputo
Output:
(hola,2)
(que,1)
Máquina map2
Input:
año
zzz
hola
Cómputo
Output:
(año,1)
(zzz,1)
(hola,1)
Máquina map3
Input:
que
zzz
que
Cómputo
Output:
(que,2)
(zzz,1)
Máquina reduce1
Input:
(hola,2)
(año,1)
(hola,1)
Sorted:
(año,1)
(hola,1)
(hola,2)
Cómputo:
(año,1)
(hola,3)
Palabras A-M
Máquina reduce2
Input:
(que,1)
(zzz,1)
(que,2)
(zzz,1)
Sorted:
(que,1)
(que,2)
(zzz,1)
(zzz,1)
Cómputo:
(que,3)
(zzz,2)
Palabras N-Z
(año,1)
(hola,3)
(que,3)
(zzz,2)

Map Reduce

Ejemplo: Join

¿Cómo hago un join con Map Reduce?

- Modelo: un archivo con el nombre de la tabla y sus tuplas
- Map: Agrupo por el atributo que hace el join
- Reduce: Hago el producto cruz para las tablas distintas

INPUT

R

A	B
1	1
1	3
3	2
3	3

S

B	C
2	4
2	7
3	8
3	9

Máquina map1

Máquina map2

Máquina reduce llave 1

Máquina reduce llave 2

Máquina reduce llave 3

INPUT

MAP

R

A	B
1	1
1	3
3	2
3	3

S

B	C
2	4
2	7
3	8
3	9

Máquina map1

Input:
(R,1,1)
(R,1,3)
(R,3,2)
(S,2,4)

Máquina map2

Máquina reduce llave 1

Máquina reduce llave 2

Máquina reduce llave 3

INPUT

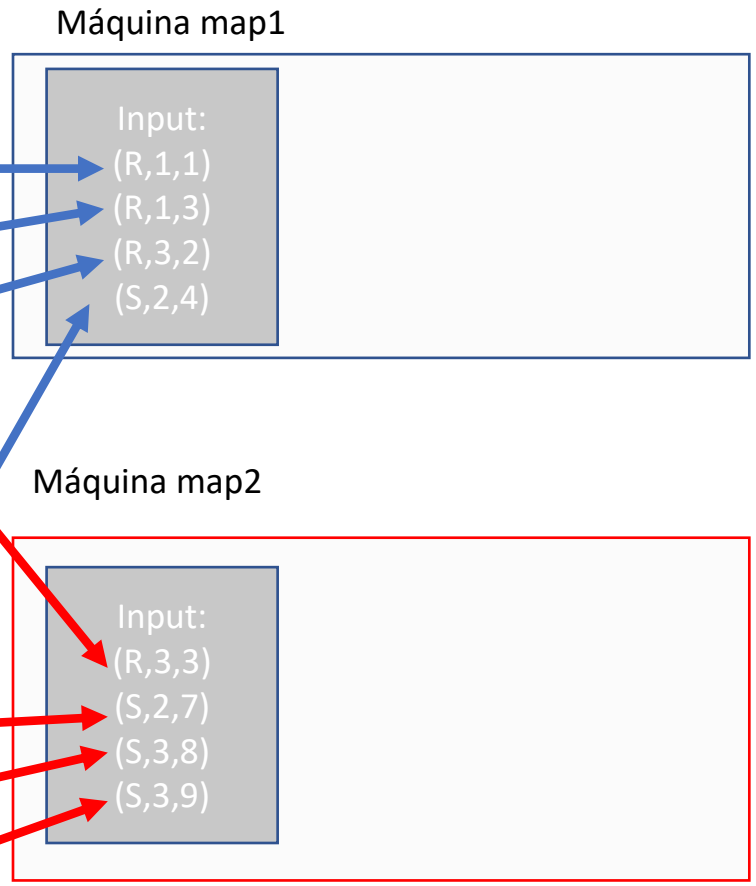
MAP

R

A	B
1	1
1	3
3	2
3	3

S

B	C
2	4
2	7
3	8
3	9



Máquina reduce llave 1

Máquina reduce llave 2

Máquina reduce llave 3

INPUT

MAP

R

A	B
1	1
1	3
3	2
3	3

S

B	C
2	4
2	7
3	8
3	9

Máquina map1

Input:
(R,1,1)
(R,1,3)
(R,3,2)
(S,2,4)

Cómputo

Output:
1: (R,1)
3: (R,1)
2: (R,3)
2: (S,4)

Máquina map2

Input:
(R,3,3)
(S,2,7)
(S,3,8)
(S,3,9)

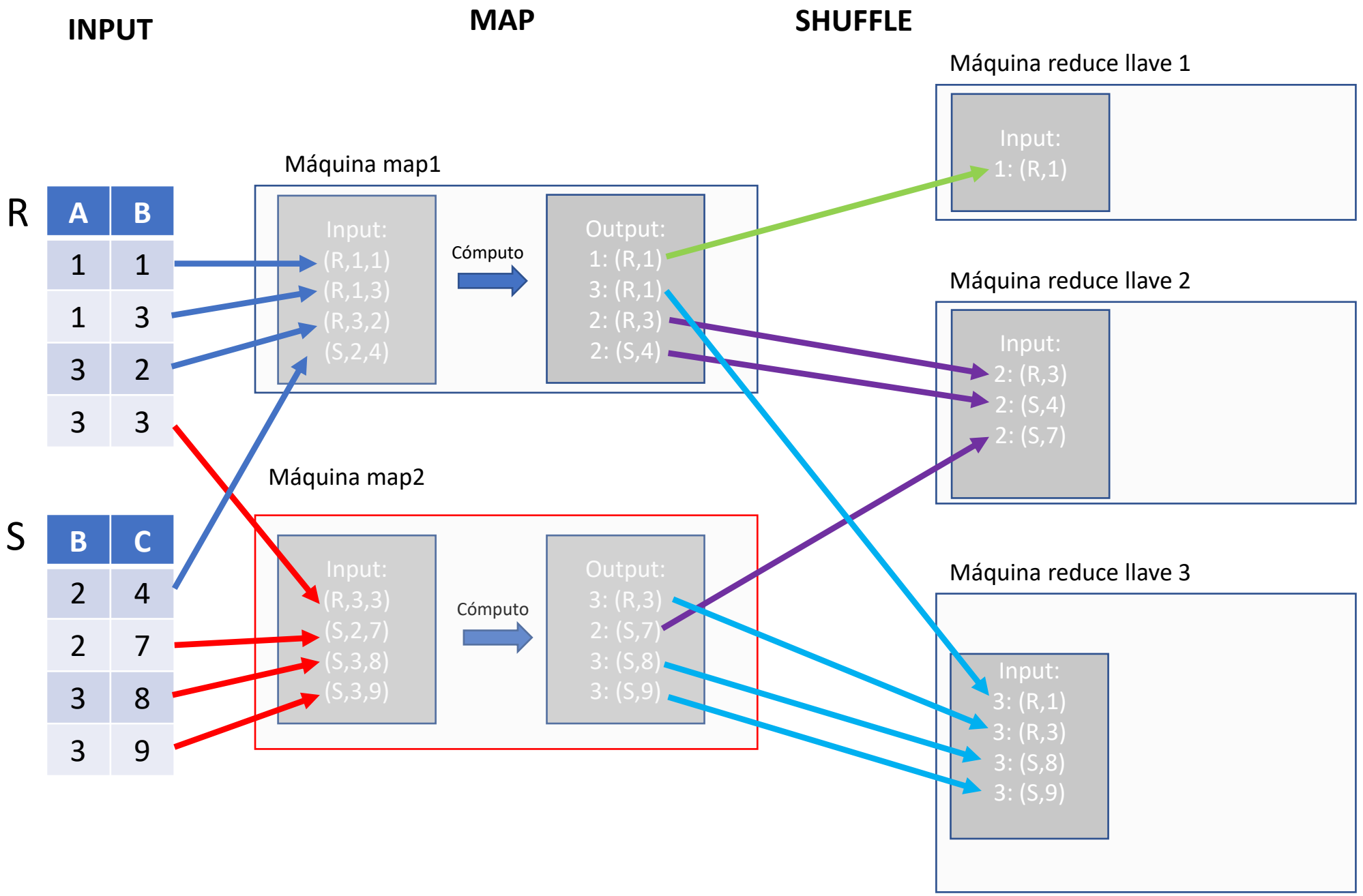
Cómputo

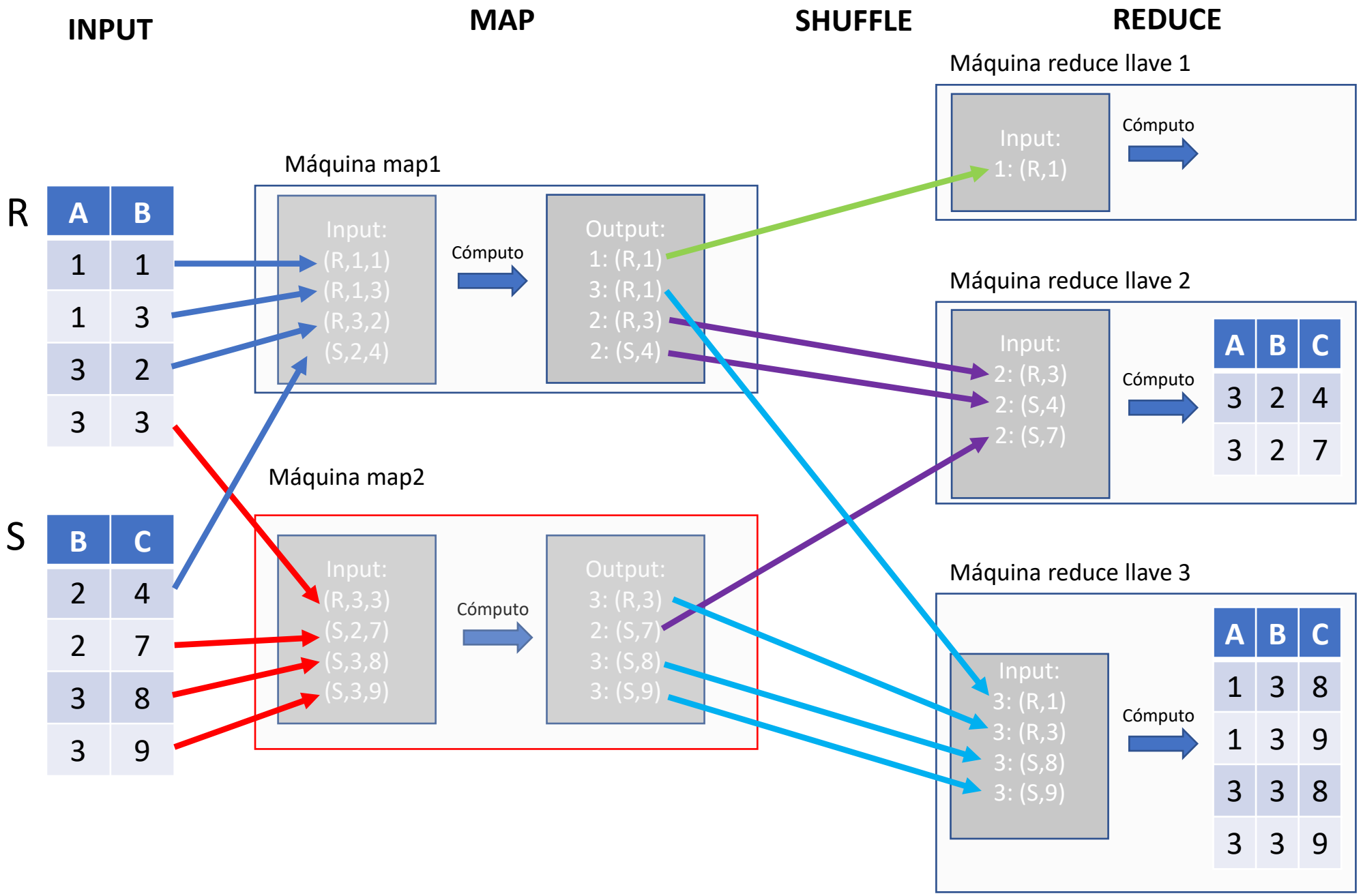
Output:
3: (R,3)
2: (S,7)
3: (S,8)
3: (S,9)

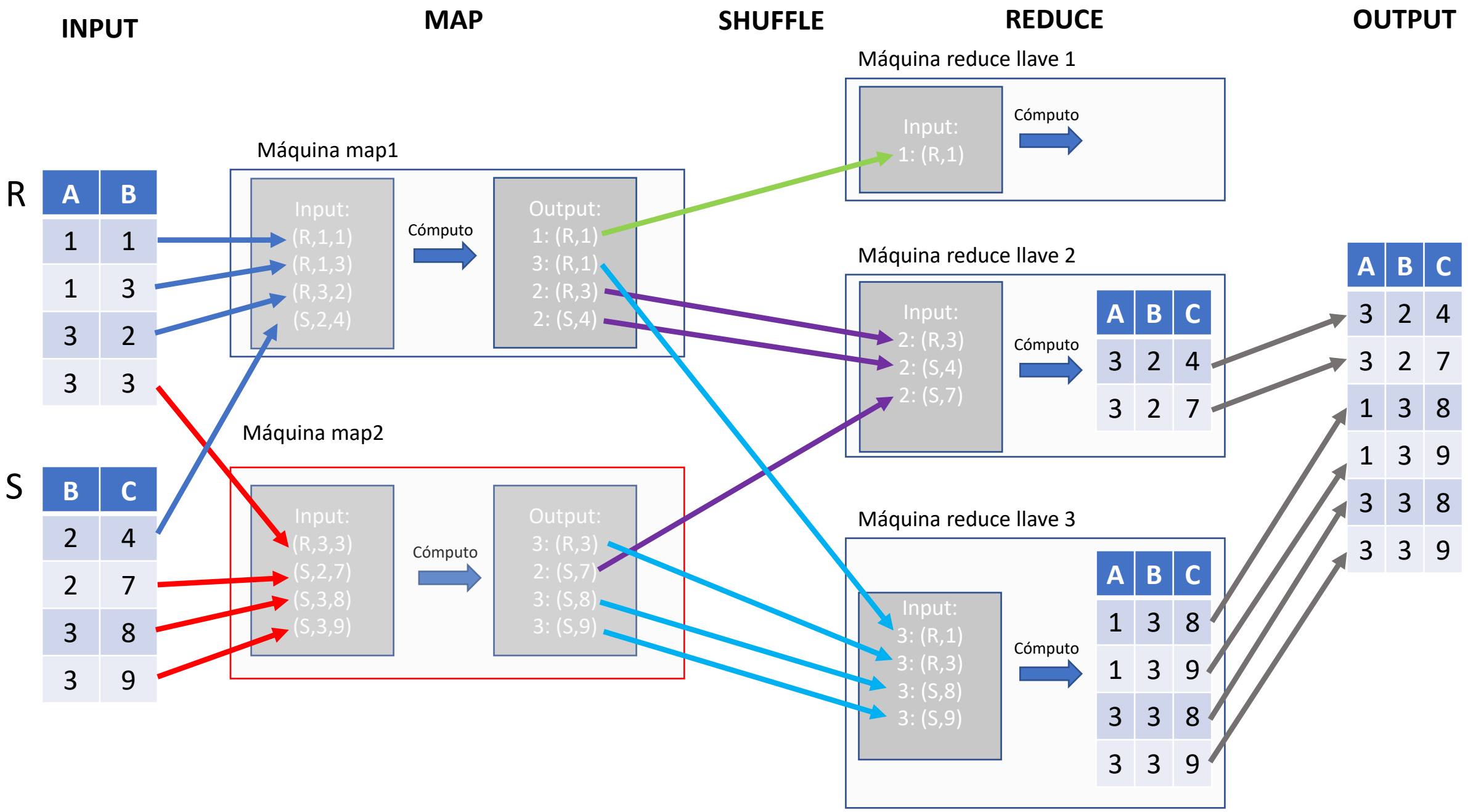
Máquina reduce llave 1

Máquina reduce llave 2

Máquina reduce llave 3


INPUT
MAP
SHUFFLE
R
A
B
1
1
1
3
3
2
3
3
S
B
C
2
4
2
7
3
8
3
9
Máquina map1
Input:
(R,1,1)
(R,1,3)
(R,3,2)
(S,2,4)
Cómputo
Output:
1: (R,1)
3: (R,1)
2: (R,3)
2: (S,4)
Máquina map2
Input:
(R,3,3)
(S,2,7)
(S,3,8)
(S,3,9)
Cómputo
Output:
3: (R,3)
2: (S,7)
3: (S,8)
3: (S,9)
Máquina reduce llave 1
Input:
1: (R,1)
Máquina reduce llave 2
Input:
2: (R,3)
2: (S,4)
2: (S,7)
Máquina reduce llave 3
Input:
3: (R,1)
3: (R,3)
3: (S,8)
3: (S,9)


INPUT
MAP
SHUFFLE
REDUCE
R
A B
1 1
1 3
3 2
3 3
S
B C
2 4
2 7
3 8
3 9
Máquina map1
Input:
(R,1,1)
(R,1,3)
(R,3,2)
(S,2,4)
Cómputo
Output:
1: (R,1)
3: (R,1)
2: (R,3)
2: (S,4)
Máquina map2
Input:
(R,3,3)
(S,2,7)
(S,3,8)
(S,3,9)
Cómputo
Output:
3: (R,3)
2: (S,7)
3: (S,8)
3: (S,9)
Máquina reduce llave 1
Input:
1: (R,1)
Cómputo
Máquina reduce llave 2
Input:
2: (R,3)
2: (S,4)
2: (S,7)
Cómputo
A B C
3 2 4
3 2 7
Máquina reduce llave 3
Input:
3: (R,1)
3: (R,3)
3: (S,8)
3: (S,9)
Cómputo
A B C
1 3 8
1 3 9
3 3 8
3 3 9


INPUT
MAP
SHUFFLE
REDUCE
OUTPUT
R
A B
1 1
1 3
3 2
3 3
S
B C
2 4
2 7
3 8
3 9
Máquina map1
Input:
(R,1,1)
(R,1,3)
(R,3,2)
(S,2,4)
Cómputo
Output:
1: (R,1)
3: (R,1)
2: (R,3)
2: (S,4)
Máquina map2
Input:
(R,3,3)
(S,2,7)
(S,3,8)
(S,3,9)
Cómputo
Output:
3: (R,3)
2: (S,7)
3: (S,8)
3: (S,9)
Máquina reduce llave 1
Input:
1: (R,1)
Cómputo
Máquina reduce llave 2
Input:
2: (R,3)
2: (S,4)
2: (S,7)
Cómputo
A B C
3 2 4
3 2 7
Máquina reduce llave 3
Input:
3: (R,1)
3: (R,3)
3: (S,8)
3: (S,9)
Cómputo
A B C
1 3 8
1 3 9
3 3 8
3 3 9
A B C
3 2 4
3 2 7
1 3 8
1 3 9
3 3 8
3 3 9

Map Reduce

No es un descubrimiento nuevo, pero recientemente se ha visto calzar perfectamente con las necesidades de las grandes BD

Es la arquitectura más importante en sistemas que reciben grandes bases de datos

- Apache Hadoop: La implementación open source de Map -Reduce, presente en muchos sistemas con computación distribuida