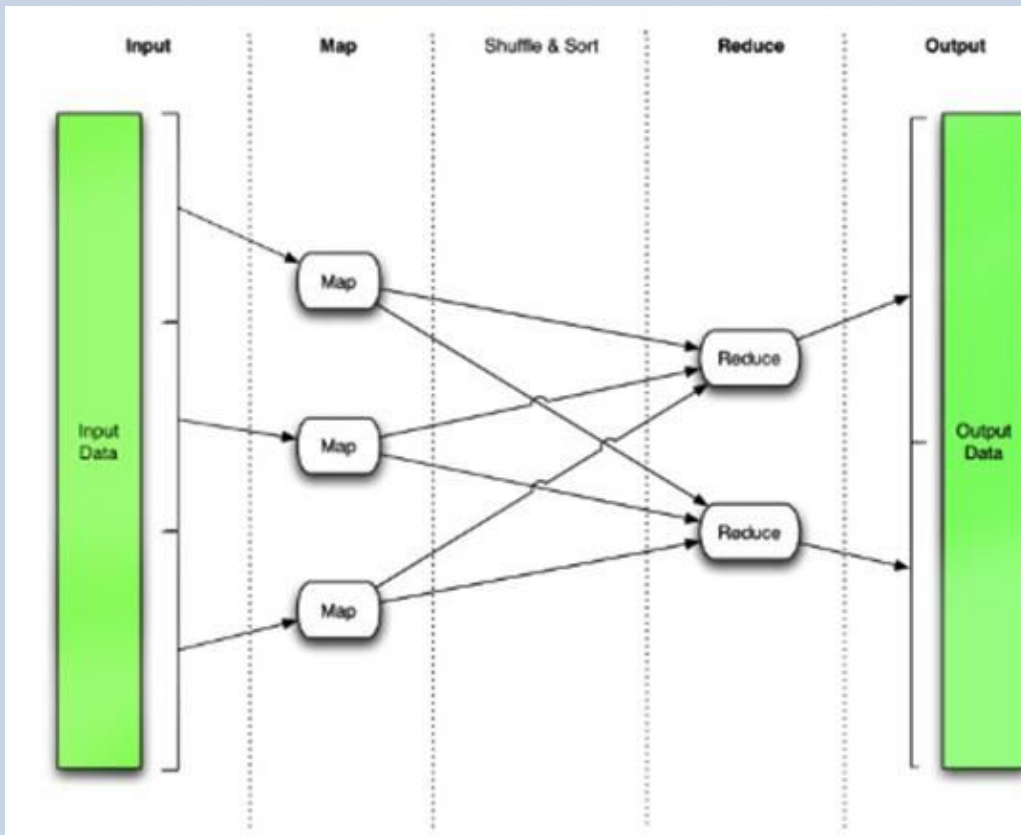


# MapReduce

## MapReduce là gì?

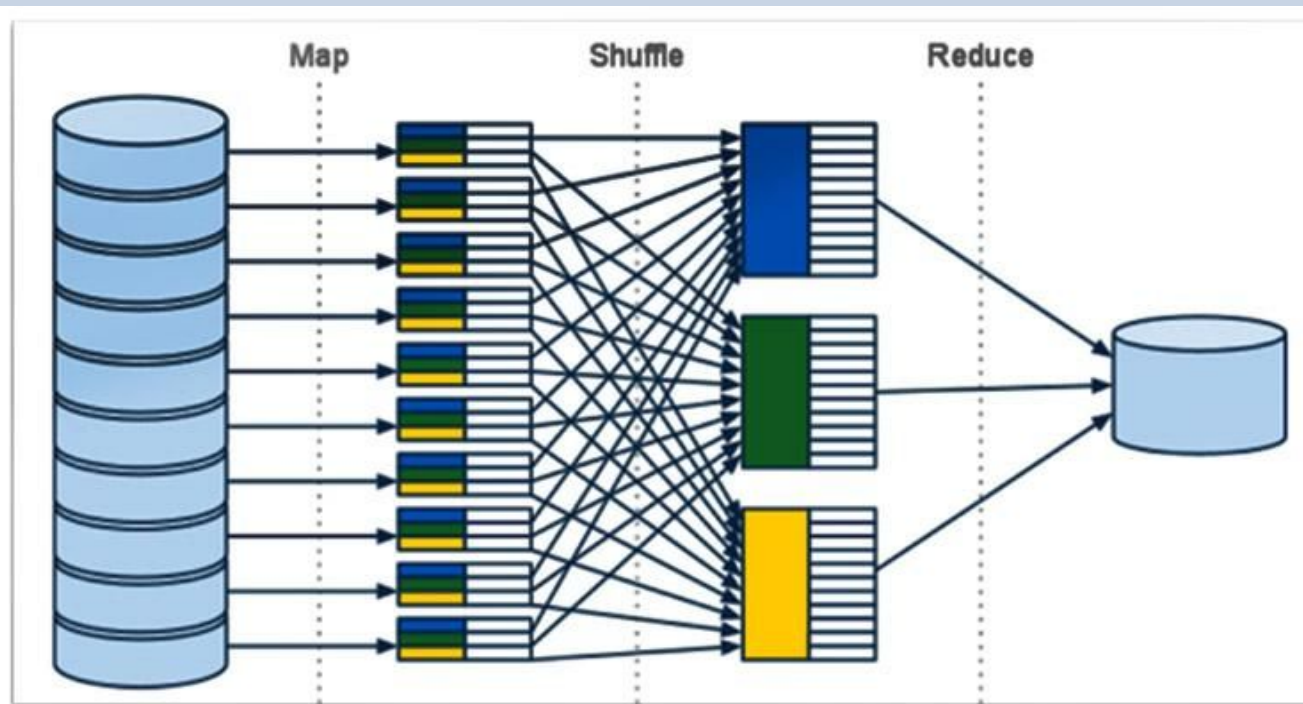
Là mô hình được thiết kế độc quyền bởi Google, nó có khả năng lập trình xử lý các tập dữ liệu lớn song song và phân tán thuật toán trên 1 cụm máy tính



# MapReduce

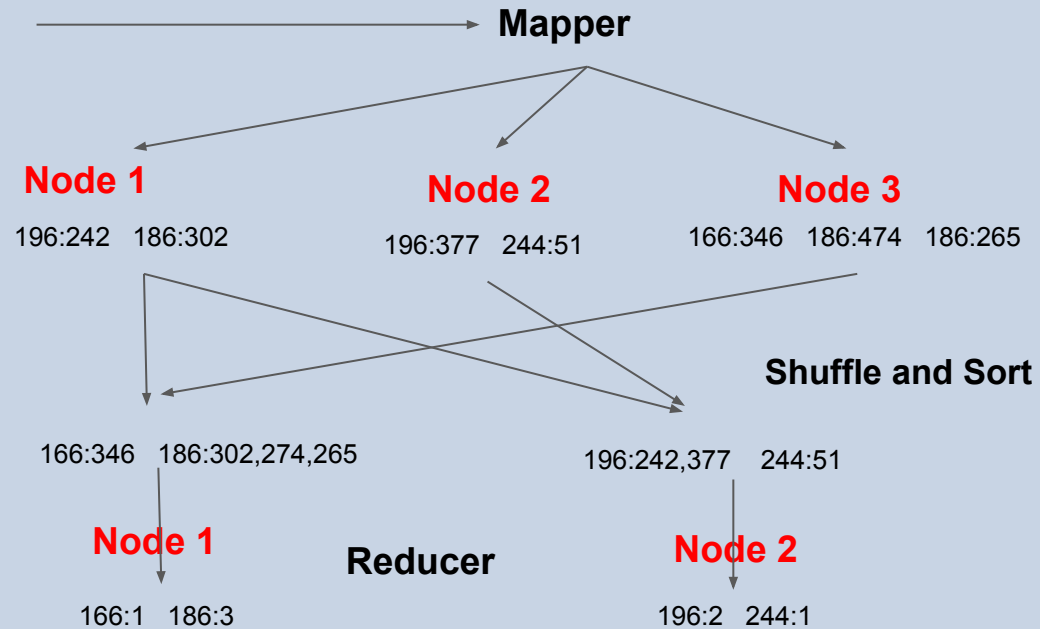
## Các thủ tục của MapReduce:

- Map
- Reduce
- Bước trung gian Shuffle



# MapReduce

User ID	Movie ID	Rating	Timestamp
196	242	3	24676
186	302	3	27349
196	377	1	38678
244	51	2	92758
166	346	1	73846
186	474	4	23756
186	265	2	37563



# MapReduce

## Matrix-Vector multiplication MapReduce

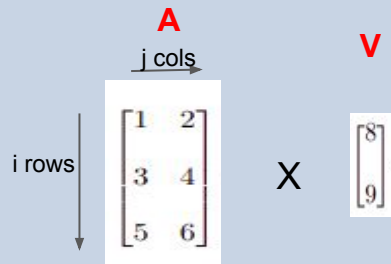
$$\begin{array}{c} \text{i rows} \downarrow \end{array}
 \begin{array}{c} \text{A} \\ \xrightarrow{\text{j Cols}} \end{array}
 \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}
 \times
 \begin{array}{c} \text{V} \\ \begin{bmatrix} 8 \\ 9 \end{bmatrix} \end{array}$$

Cách nhân ma trận thông thường

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 1.8 + 2.9 \\ 3.8 + 4.9 \\ 5.8 + 6.9 \end{bmatrix}$$

# MapReduce

## Nhân ma trận đối với MapReduce



Giả sử ta có ma trận **A** có kích thước (i, j)

Với mỗi điểm tương ứng trên ma trận có tọa độ **aij**

Và vector **V** với độ dài n.

Khi đó tích của ma trận **A** với vector **V** được xác định bằng công thức :

$$x_i = \sum_{j=1}^n m_{ij} v_j$$

MAP Function:

**map(key, value)**

for every (i,j, **aij**) in value

return ( i, **aij \* vj** )

SHUFFLE

REDUCE Function:

**reduce(key, value)**

result = 0

for values in values

result += value

return (key, value)

i	j	aij	i aij * Vj
1	1	aij = 1	→ (1, 1 x 8)
1	2	aij = 2	→ (1, 2 x 9)
2	1	aij = 3	→ (2, 3 x 8)
2	2	aij = 4	→ (2, 4 x 9)
3	1	aij = 5	→ (3, 5 x 8)
3	2	aij = 6	→ (3, 6 x 9)

1, ((1 x 8), (2 x 9))

2, ((3 x 8), (4 x 9))

3, ((5 x 8), (6 x 9))

1, ((1 x 8) + (2 x 9))

2, ((3 x 8) + (4 x 9))

3, ((5 x 8) + (6 x 9))

# MapReduce

## Matrix multiplication

Ta có:

M là 1 ma trận

$m_{ij}$  :  $i$  = rows,  $j$  = cols

N là 1 ma trận

$n_{jk}$  :  $j$  = rows,  $k$  = cols

Đặt :

$P = M * N$

$p_{ik}$  :  $i$  = rows,  $k$  = cols

**M**

$$\begin{bmatrix} 2 & 3 \\ 4 & 9 \end{bmatrix}$$

**N**

$$\begin{bmatrix} 5 & 6 & 8 \\ 2 & 4 & 7 \end{bmatrix}$$

**x**

**= P**

$$\begin{bmatrix} 2.5 + 3.2 & 2.6 + 3.4 & 2.8 + 3.7 \\ 4.5 + 9.2 & 4.6 + 9.4 & 4.8 + 9.7 \end{bmatrix}$$

# MapReduce

## Matrix multiplication using MapReduce

$$\begin{bmatrix} 2 & 3 \\ 4 & 9 \end{bmatrix} \cdot \begin{bmatrix} 5 & 6 & 8 \\ 2 & 4 & 7 \end{bmatrix} = \begin{bmatrix} 2.5 + 3.2 & 2.6 + 3.4 & 2.8 + 3.7 \\ 4.5 + 9.2 & 4.6 + 9.4 & 4.8 + 9.7 \end{bmatrix}$$

Bước 1:

MAP Function:

Đưa từng phần tử của 2 ma trận vào theo cặp key-value

REDUCE Function:

$(j, (m, i, m_{ij}))$

1, (m, 1,  $m_{11}$ ) = 2

1, (m, 2,  $m_{21}$ ) = 4

2, (m, 1,  $m_{12}$ ) = 3

2, (m, 2,  $m_{22}$ ) = 9

$(j, (n, k, n_{jk}))$

1, (n, 1,  $n_{11}$ ) = 5

1, (n, 2,  $n_{12}$ ) = 6

1, (n, 3,  $n_{13}$ ) = 8

2, (n, 1,  $n_{21}$ ) = 2

2, (n, 2,  $n_{22}$ ) = 4

2, (n, 3,  $n_{23}$ ) = 7

$(j, (i, k, m_{ij} * n_{jk}))$

1, (1, 1,  $2*5$ )

1, (1, 2,  $2*6$ )

1, (1, 3,  $2*8$ )

1, (2, 1,  $4*5$ )

1, (2, 2,  $4*6$ )

1, (2, 3,  $4*8$ )

$(j, (i, k, m_{ij} * n_{jk}))$

2, (1, 1,  $3*2$ )

2, (1, 2,  $3*4$ )

2, (1, 3,  $3*7$ )

2, (2, 1,  $9*2$ )

2, (2, 2,  $9*4$ )

2, (2, 3,  $9*7$ )

Bước 2:

MAP Function: là output của reduce ở bước 1.

Mapper2

(1, [(1, 1,  $2*5$ ), (1, 2,  $2*6$ ), (1, 3,  $2*8$ ), (2, 1,  $4*5$ ), (2, 2,  $4*6$ ), (2, 3,  $4*8$ ) ] )

(2, [(1, 1,  $3*2$ ), (1, 2,  $3*4$ ), (1, 3,  $3*7$ ), (2, 1,  $9*2$ ), (2, 2,  $9*4$ ), (2, 3,  $9*7$ ) ] )

SHUFFLE2

[((1,1),  $2*5$ ), ((1,2),  $2*6$ ), ((1,3),  $2*8$ ), ((2,1),  $4*5$ ), ((2,2),  $4*6$ ), ((2,3),  $4*8$ )]

[((1,1),  $3*2$ ), ((1,2),  $3*4$ ), ((1,3),  $3*7$ ), ((2,1),  $9*2$ ), ((2,2),  $9*4$ ), ((2,3),  $9*7$ )]

REDUCE2 : Sẽ nhóm theo key-value ( i,k), V  
trong đó V là giá trị hàng i cột k của P

((1,1),  $2*5 + 3*2$ )

((1,2),  $2*6 + 3*4$ )

((1,3),  $2*8 + 3*7$ )

((2,1),  $4*5 + 9*2$ )

((2,2),  $4*6 + 9*4$ )

((2,3),  $4*8 + 9*7$ )