Big Data Management & Analytics

EXERCISE 4 – MAPREDUCE, SPARK

23rd of November 2015

Sabrina Friedl
LMU Munich

1. Matrix Multiplication with MapReduce

REVISION AND EXAMPLE

MapReduce – Matrix Multiplication

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} \quad A \cdot B = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}$$

Can be written as $A = (I, J, V),\ B = (J, K, W)$ $\quad \text{where} \quad [0] = row, \quad [1] = column \quad \text{and} \quad [2] = values$

Steps

- 1. Map $(i, j, a_{ij}) \longrightarrow (j, (A, i, a_{ij})) \qquad (j, k, b_{jk}) \longrightarrow (j, (B, k, b_{jk}))$
- 2. Join $(j, (A, i, a_{ij})) \bowtie (j, (B, k, b_{jk})) \longrightarrow (j, [(A, i, a_{ij}), (B, k, b_{jk})])$
- 3. Map $(j, [(A, i, a_{ij}), (B, k, b_{jk})]) \longrightarrow ((i, k), (a_{ij}b_{jk}))$
- 4. ReduceByKey $((i, k), [(a_{ij}b_{jk})]) \longrightarrow ((i, k), \sum(a_{ij}b_{jk}))$

Matrix Multiplication - Example

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} = \begin{pmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix} \qquad A \cdot B = C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} 58 & 64 \\ 139 & 154 \end{pmatrix}$$

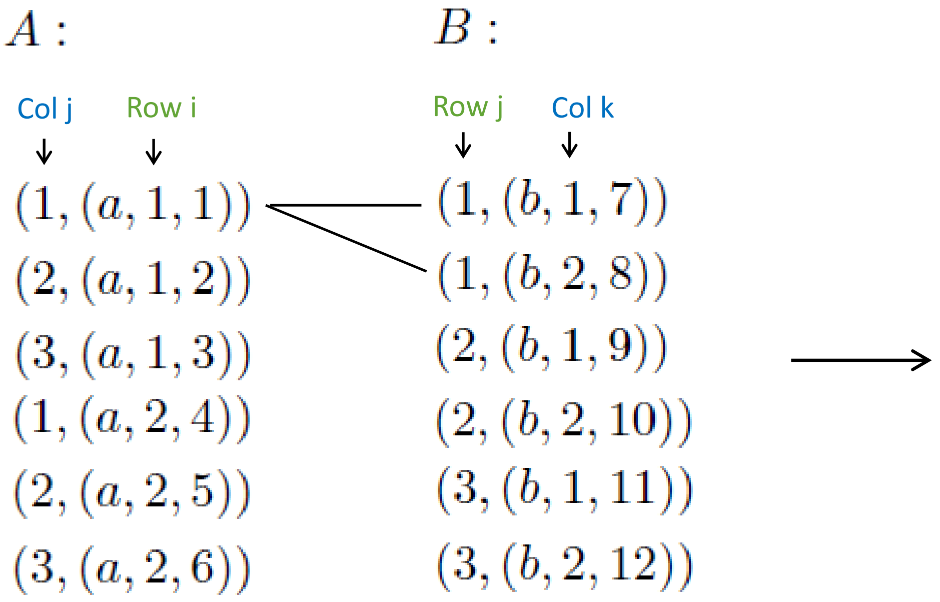
1. Map: $(i, j, a_{ij}) \longrightarrow (j, (A, i, a_{ij}))$, $\qquad (j, k, b_{jk}) \longrightarrow (j, (B, k, b_{jk}))$

$$\begin{aligned}
A: \quad & (\overset{\substack{\text{row}\\ \downarrow}}{1}, \overset{\substack{\text{col}\\ \downarrow}}{1}, 1) \longrightarrow (\overset{\substack{\text{col}\\ \downarrow}}{1}, (\overset{\substack{\text{ID}\\ \downarrow}}{a}, \overset{\substack{\text{row}\\ \downarrow}}{1}, 1)) \\
& (1, 2, 2) \longrightarrow (2, (a, 1, 2)) \\
& (1, 3, 3) \longrightarrow (3, (a, 1, 3)) \\[1ex]
& (2, 1, 4) \longrightarrow (1, (a, 2, 4)) \\
& (2, 2, 5) \longrightarrow (2, (a, 2, 5)) \\
& (2, 3, 6) \longrightarrow (3, (a, 2, 6))
\end{aligned}$$

$$\begin{aligned}
B: \quad & (\overset{\substack{\text{row}\\ \downarrow}}{1}, \overset{\substack{\text{col}\\ \downarrow}}{1}, 7) \longrightarrow (\overset{\substack{\text{row}\\ \downarrow}}{1}, (\overset{\substack{\text{ID}\\ \downarrow}}{b}, \overset{\substack{\text{col}\\ \downarrow}}{1}, 7)) \\
& (1, 2, 8) \longrightarrow (1, (b, 2, 8)) \\[1ex]
& (2, 1, 9) \longrightarrow (2, (b, 1, 9)) \\
& (2, 2, 10) \longrightarrow (2, (b, 2, 10)) \\[1ex]
& (3, 1, 11) \longrightarrow (3, (b, 1, 11)) \\
& (3, 2, 12) \longrightarrow (3, (b, 2, 12))
\end{aligned}$$

2. Join: $(j,(A,i,a_{ij})) \bowtie (j,(B,k,b_{jk})) \longrightarrow (j,[(A,i,a_{ij}),(B,k,b_{jk})])$

$$\begin{pmatrix}\boxed{1} & 2 & 3\\ 4 & 5 & 6\end{pmatrix} \qquad \begin{pmatrix}\boxed{7} & \boxed{8}\\ 9 & 10\\ 11 & 12\end{pmatrix}$$



$A:$

Col j ↓ Row i ↓

$(1,(a,1,1))$

$(2,(a,1,2))$

$(3,(a,1,3))$

$(1,(a,2,4))$

$(2,(a,2,5))$

$(3,(a,2,6))$

$B:$

Row j ↓ Col k ↓

$(1,(b,1,7))$

$(1,(b,2,8))$

$(2,(b,1,9))$

$(2,(b,2,10))$

$(3,(b,1,11))$

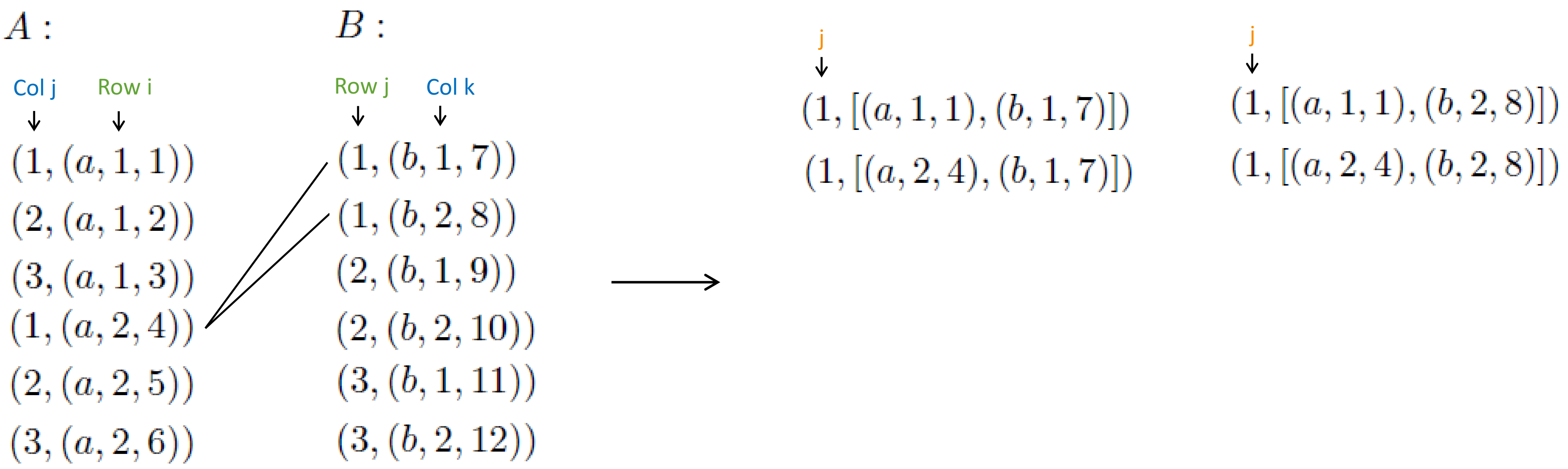
$(3,(b,2,12))$

$(\overset{\overset{j}{\downarrow}}{1},[(a,1,1),(b,1,7)])$ $\qquad$ $(\overset{\overset{j}{\downarrow}}{1},[(a,1,1),(b,2,8)])$

“Join over j“

2. Join: $(j, (A, i, a_{ij})) \bowtie (j, (B, k, b_{jk})) \longrightarrow (j, [(A, i, a_{ij}), (B, k, b_{jk})])$

$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ $\begin{pmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix}$



$A:$

Col j ↓ Row i ↓

$(1, (a, 1, 1))$

$(2, (a, 1, 2))$

$(3, (a, 1, 3))$

$(1, (a, 2, 4))$

$(2, (a, 2, 5))$

$(3, (a, 2, 6))$

$B:$

Row j ↓ Col k ↓

$(1, (b, 1, 7))$

$(1, (b, 2, 8))$

$(2, (b, 1, 9))$

$(2, (b, 2, 10))$

$(3, (b, 1, 11))$

$(3, (b, 2, 12))$

$\longrightarrow$

j ↓

$(1, [(a, 1, 1), (b, 1, 7)])$

$(1, [(a, 2, 4), (b, 1, 7)])$

j ↓

$(1, [(a, 1, 1), (b, 2, 8)])$

$(1, [(a, 2, 4), (b, 2, 8)])$

“Join over j“

2. Join: $(j,(A,i,a_{ij})) \bowtie (j,(B,k,b_{jk})) \longrightarrow (j,[(A,i,a_{ij}),(B,k,b_{jk})])$

$\begin{pmatrix}1 & \boxed{2} & 3\\ 4 & 5 & 6\end{pmatrix}$ $\begin{pmatrix}7 & 8\\ \boxed{9 \quad 10}\\ 11 & 12\end{pmatrix}$

$A:$

Col j ↓ Row i ↓

$(1,(a,1,1))$

$(2,(a,1,2))$

$(3,(a,1,3))$

$(1,(a,2,4))$

$(2,(a,2,5))$

$(3,(a,2,6))$

$B:$

Row j ↓ Col k ↓

$(1,(b,1,7))$

$(1,(b,2,8))$

$(2,(b,1,9))$

$(2,(b,2,10))$

$(3,(b,1,11))$

$(3,(b,2,12))$

⟶

j ↓

$(1,[(a,1,1),(b,1,7)])$

$(1,[(a,2,4),(b,1,7)])$

$(2,[(a,1,2),(b,1,9)])$

j ↓

$(1,[(a,1,1),(b,2,8)])$

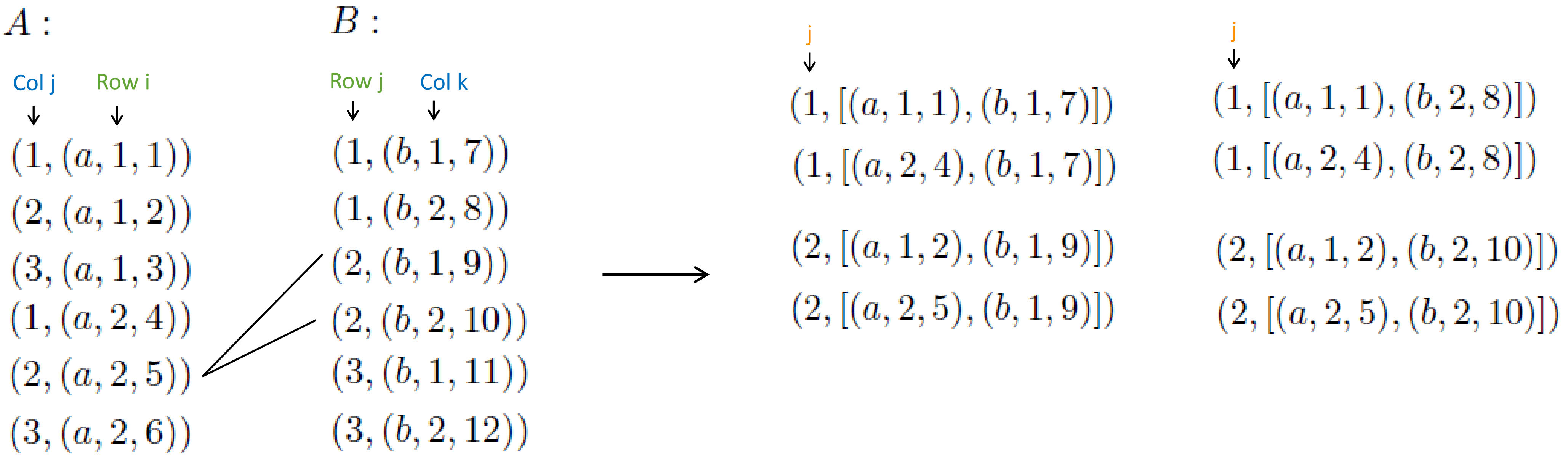
$(1,[(a,2,4),(b,2,8)])$

$(2,[(a,1,2),(b,2,10)])$

“Join over j“

2. Join: $(j,(A,i,a_{ij})) \bowtie (j,(B,k,b_{jk})) \longrightarrow (j,[(A,i,a_{ij}),(B,k,b_{jk})])$

$$\begin{pmatrix}1 & \boxed{\begin{matrix}2\\5\end{matrix}} & 3\\ 4 & & 6\end{pmatrix} \qquad \begin{pmatrix}7 & 8\\ \boxed{9 \quad 10}\\ 11 & 12\end{pmatrix}$$



$A:$

Col j ↓ Row i ↓

$(1,(a,1,1))$

$(2,(a,1,2))$

$(3,(a,1,3))$

$(1,(a,2,4))$

$(2,(a,2,5))$

$(3,(a,2,6))$

$B:$

Row j ↓ Col k ↓

$(1,(b,1,7))$

$(1,(b,2,8))$

$(2,(b,1,9))$

$(2,(b,2,10))$

$(3,(b,1,11))$

$(3,(b,2,12))$

$\longrightarrow$

j ↓

$(1,[(a,1,1),(b,1,7)])$

$(1,[(a,2,4),(b,1,7)])$

$(2,[(a,1,2),(b,1,9)])$

$(2,[(a,2,5),(b,1,9)])$

j ↓

$(1,[(a,1,1),(b,2,8)])$

$(1,[(a,2,4),(b,2,8)])$

$(2,[(a,1,2),(b,2,10)])$

$(2,[(a,2,5),(b,2,10)])$

“Join over j“

2. Join: $(j,(A,i,a_{ij})) \bowtie (j,(B,k,b_{jk})) \longrightarrow (j,[(A,i,a_{ij}),(B,k,b_{jk})])$

$$\begin{pmatrix} 1 & 2 & \boxed{3} \\ 4 & 5 & 6 \end{pmatrix} \qquad \begin{pmatrix} 7 & 8 \\ 9 & 10 \\ \boxed{11 \quad 12} \end{pmatrix}$$

$A:$

Col j ↓ Row i ↓

$(1,(a,1,1))$

$(2,(a,1,2))$

$(3,(a,1,3))$

$(1,(a,2,4))$

$(2,(a,2,5))$

$(3,(a,2,6))$

$B:$

Row j ↓ Col k ↓

$(1,(b,1,7))$

$(1,(b,2,8))$

$(2,(b,1,9))$

$(2,(b,2,10))$

$(3,(b,1,11))$

$(3,(b,2,12))$

$\longrightarrow$

j ↓

$(1,[(a,1,1),(b,1,7)])$

$(1,[(a,2,4),(b,1,7)])$

$(2,[(a,1,2),(b,1,9)])$

$(2,[(a,2,5),(b,1,9)])$

$(3,[(a,1,3),(b,1,11)])$

j ↓

$(1,[(a,1,1),(b,2,8)])$

$(1,[(a,2,4),(b,2,8)])$

$(2,[(a,1,2),(b,2,10)])$

$(2,[(a,2,5),(b,2,10)])$

$(3,[(a,1,3),(b,2,12)])$

“Join over j“

2. Join: $(j,(A,i,a_{ij})) \bowtie (j,(B,k,b_{jk})) \longrightarrow (j,[(A,i,a_{ij}),(B,k,b_{jk})])$

$$\begin{pmatrix}1 & 2 & \boxed{3}\\ 4 & 5 & \boxed{6}\end{pmatrix} \qquad \begin{pmatrix}7 & 8\\ 9 & 10\\ \boxed{11} & \boxed{12}\end{pmatrix}$$

$A:$

Col j ↓ Row i ↓

$(1,(a,1,1))$

$(2,(a,1,2))$

$(3,(a,1,3))$

$(1,(a,2,4))$

$(2,(a,2,5))$

$(3,(a,2,6))$

$B:$

Row j ↓ Col k ↓

$(1,(b,1,7))$

$(1,(b,2,8))$

$(2,(b,1,9))$

$(2,(b,2,10))$

$(3,(b,1,11))$

$(3,(b,2,12))$

"Join over j"

$\longrightarrow$

j ↓

$(1,[(a,1,1),(b,1,7)])$

$(1,[(a,2,4),(b,1,7)])$

$(2,[(a,1,2),(b,1,9)])$

$(2,[(a,2,5),(b,1,9)])$

$(3,[(a,1,3),(b,1,11)])$

$(3,[(a,2,6),(b,1,11)])$

j ↓

$(1,[(a,1,1),(b,2,8)])$

$(1,[(a,2,4),(b,2,8)])$

$(2,[(a,1,2),(b,2,10)])$

$(2,[(a,2,5),(b,2,10)])$

$(3,[(a,1,3),(b,2,12)])$

$(3,[(a,2,6),(b,2,12)])$

Number of key-value pairs: $i \cdot j \cdot k$

3. Map: $(j, [(A, i, a_{ij}), (B, k, b_{jk})]) \longrightarrow ((i, k), (a_{ij} b_{jk}))$

$$\begin{array}{ll}
(\overset{\substack{\text{j}\\ \downarrow}}{1}, [(a, \overset{\substack{\text{Row i}\\ \downarrow}}{1}, 1), (b, \overset{\substack{\text{Col k}\\ \downarrow}}{1}, 7)]) & (\overset{\substack{\text{j}\\ \downarrow}}{1}, [(a, \overset{\substack{\text{Row i}\\ \downarrow}}{1}, 1), (b, \overset{\substack{\text{Col k}\\ \downarrow}}{2}, 8)]) \\
(1, [(a, 2, 4), (b, 1, 7)]) & (1, [(a, 2, 4), (b, 2, 8)]) \\
\\
(2, [(a, 1, 2), (b, 1, 9)]) & (2, [(a, 1, 2), (b, 2, 10)]) \\
(2, [(a, 2, 5), (b, 1, 9)]) & (2, [(a, 2, 5), (b, 2, 10)]) \\
\\
(3, [(a, 1, 3), (b, 1, 11)]) & (3, [(a, 1, 3), (b, 2, 12)]) \\
(3, [(a, 2, 6), (b, 1, 11)]) & (3, [(a, 2, 6), (b, 2, 12)])
\end{array}
\longrightarrow
\begin{array}{ll}
((\overset{\substack{\text{i}\\ \downarrow}}{1}, \overset{\substack{\text{k}\\ \downarrow}}{1}), 1 \cdot 7)) & ((\overset{\substack{\text{i}\\ \downarrow}}{1}, \overset{\substack{\text{k}\\ \downarrow}}{2}), 1 \cdot 8)) \\
((2, 1), 4 \cdot 7)) & ((2, 2), 4 \cdot 8)) \\
\\
((1, 1), 2 \cdot 9)) & ((1, 2), 2 \cdot 10)) \\
((2, 1), 5 \cdot 9)) & ((2, 2), 5 \cdot 10)) \\
\\
((1, 1), 3 \cdot 11)) & ((1, 2), 3 \cdot 12)) \\
((2, 1), 6 \cdot 11)) & ((2, 2), 6 \cdot 12))
\end{array}$$

4. ReduceByKey: $(lambda\ \ x, y : x + y)$

$$\begin{array}{lllll}
\longrightarrow & ((\overset{\overset{i}{\downarrow}}{1},\overset{\overset{k}{\downarrow}}{1}),1\cdot 7)) & ((\overset{\overset{i}{\downarrow}}{1},\overset{\overset{k}{\downarrow}}{2}),1\cdot 8)) & & ((1,1),\ \ 1\cdot 7+2\cdot 9+3\cdot 11)) \\
 & ((2,1),4\cdot 7)) & ((2,2),4\cdot 8)) & & \\
\longrightarrow & ((1,1),2\cdot 9)) & ((1,2),2\cdot 10)) & \longrightarrow & \\
 & ((2,1),5\cdot 9)) & ((2,2),5\cdot 10)) & & \\
\longrightarrow & ((1,1),3\cdot 11)) & ((1,2),3\cdot 12)) & & \\
 & ((2,1),6\cdot 11)) & ((2,2),6\cdot 12)) & &
\end{array}$$

4. ReduceByKey: $(lambda\ \ x, y : x + y)$

$$\begin{tikzcd}
((\overset{\overset{i}{\downarrow}}{1},\overset{\overset{k}{\downarrow}}{1}),1\cdot 7)) \arrow[r] & ((\overset{\overset{i}{\downarrow}}{1},\overset{\overset{k}{\downarrow}}{2}),1\cdot 8)) & & ((1,1),\ \ 1\cdot 7+2\cdot 9+3\cdot 11)) & ((1,2),\ \ 1\cdot 8+2\cdot 10+3\cdot 12) \\
((2,1),4\cdot 7)) & ((2,2),4\cdot 8)) & & & \\
((1,1),2\cdot 9)) \arrow[r] & ((1,2),2\cdot 10)) & {} \arrow[r] & {} & \\
((2,1),5\cdot 9)) & ((2,2),5\cdot 10)) & & & \\
((1,1),3\cdot 11)) \arrow[r] & ((1,2),3\cdot 12)) & & & \\
((2,1),6\cdot 11)) & ((2,2),6\cdot 12)) & & &
\end{tikzcd}$$

4. ReduceByKey: $(lambda\ \ x, y : x + y)$

$$((\overset{i}{1},\overset{k}{1}),1\cdot 7)) \qquad ((\overset{i}{1},\overset{k}{2}),1\cdot 8))$$

$$\longrightarrow ((2,1),4\cdot 7)) \qquad ((2,2),4\cdot 8))$$

$$((1,1),2\cdot 9)) \qquad ((1,2),2\cdot 10))$$

$$\longrightarrow ((2,1),5\cdot 9)) \qquad ((2,2),5\cdot 10))$$

$$((1,1),3\cdot 11)) \qquad ((1,2),3\cdot 12))$$

$$\longrightarrow ((2,1),6\cdot 11)) \qquad ((2,2),6\cdot 12))$$

$$\longrightarrow$$

$$((1,1),\ \ 1\cdot 7+2\cdot 9+3\cdot 11)) \qquad ((1,2),\ \ 1\cdot 8+2\cdot 10+3\cdot 12)$$

$$((2,1),\ \ 4\cdot 7+5\cdot 9+6\cdot 11))$$

4. ReduceByKey: $(lambda\ \ x, y : x + y)$

$$\begin{array}{lcl}
((\overset{\overset{i}{\downarrow}}{1}, \overset{\overset{k}{\downarrow}}{1}), 1 \cdot 7)) & & ((\overset{\overset{i}{\downarrow}}{1}, \overset{\overset{k}{\downarrow}}{2}), 1 \cdot 8)) \\
((2, 1), 4 \cdot 7)) & \longrightarrow & ((2, 2), 4 \cdot 8)) \\
 & & \\
((1, 1), 2 \cdot 9)) & & ((1, 2), 2 \cdot 10)) \\
((2, 1), 5 \cdot 9)) & \longrightarrow & ((2, 2), 5 \cdot 10)) \\
 & & \\
((1, 1), 3 \cdot 11)) & & ((1, 2), 3 \cdot 12)) \\
((2, 1), 6 \cdot 11)) & \longrightarrow & ((2, 2), 6 \cdot 12))
\end{array}
\quad \longrightarrow \quad
\begin{array}{ll}
((1, 1),\ \ 1 \cdot 7 + 2 \cdot 9 + 3 \cdot 11)) & ((1, 2),\ \ 1 \cdot 8 + 2 \cdot 10 + 3 \cdot 12) \\
((2, 1),\ \ 4 \cdot 7 + 5 \cdot 9 + 6 \cdot 11)) & ((2, 2),\ \ 4 \cdot 8 + 5 \cdot 10 + 6 \cdot 12)
\end{array}$$

4. ReduceByKey: $(lambda\ \ x, y : x + y)$

$$
\begin{array}{ll}
((\overset{\overset{i}{\downarrow}}{1}, \overset{\overset{k}{\downarrow}}{1}), 1 \cdot 7)) & ((\overset{\overset{i}{\downarrow}}{1}, \overset{\overset{k}{\downarrow}}{2}), 1 \cdot 8)) \\
((2, 1), 4 \cdot 7)) & ((2, 2), 4 \cdot 8)) \\
 & \\
((1, 1), 2 \cdot 9)) & ((1, 2), 2 \cdot 10)) \\
((2, 1), 5 \cdot 9)) & ((2, 2), 5 \cdot 10)) \\
 & \\
((1, 1), 3 \cdot 11)) & ((1, 2), 3 \cdot 12)) \\
((2, 1), 6 \cdot 11)) & ((2, 2), 6 \cdot 12))
\end{array}
\longrightarrow
\begin{array}{ll}
((1, 1),\ \ 1 \cdot 7 + 2 \cdot 9 + 3 \cdot 11)) & ((1, 2),\ \ 1 \cdot 8 + 2 \cdot 10 + 3 \cdot 12) \\
((2, 1),\ \ 4 \cdot 7 + 5 \cdot 9 + 6 \cdot 11)) & ((2, 2),\ \ 4 \cdot 8 + 5 \cdot 10 + 6 \cdot 12)
\end{array}
$$

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} 58 & 64 \\ 139 & 154 \end{pmatrix}$$

Number of elements: $i \cdot k$

2. KMeans with MapReduce

Revision

MapReduce - KMeans

Randomly initialize k centers:

$$\mu^{(0)} = \mu_1^{(0)}, \ldots, \mu_k^{(0)}$$

Classify: Assign each point $j \in \{1, \ldots, m\}$ to nearest centre:

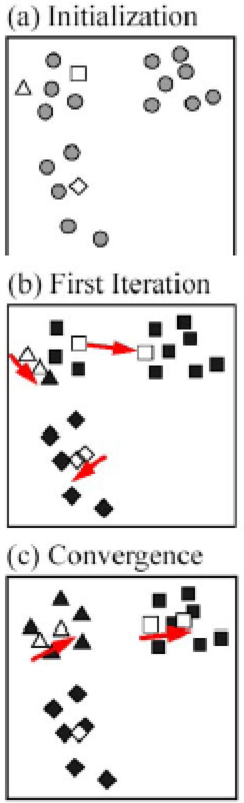
$$z^j \leftarrow \arg\min_i ||\mu_i - x^j||_2^2$$

Map

Recenter: μ_i becomes centroid of its points:

$$\mu_i^{(t+1)} \leftarrow \arg\min_\mu \sum_{j:z^j=i} ||\mu - x^j||_2^2$$

Reduce


(a) Initialization
(b) First Iteration
(c) Convergence