Big Data Management and Analytics
Assignment 4

Parts of the slides are based on work by Sabrina Friedl

Assignment 4-1

- Given two matrices A,B:

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

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

- A,B can be rewritten as:

$$A = (I, J, V), B = (J, K, W) \ \ where \ \ [0] \coloneqq row, [1] \coloneqq column \ and \ [2] = values$$

Assignment 4-1

(a) Describe the steps which are required to perform a matrix multiplication using MapReduce.

Steps:

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

Matrix Multiplication - Example

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

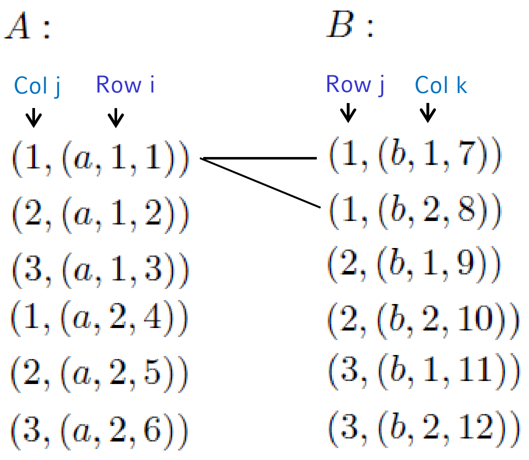
1. Map: $(i, j, a_{ij}) \longrightarrow (j, (A, i, a_{ij}))$, $\qquad (j, k, b_{jk}) \longrightarrow (j, (B, k, b_{jk}))$

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

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

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

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



$A:$

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 ↓

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

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

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

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

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

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

$\longrightarrow$

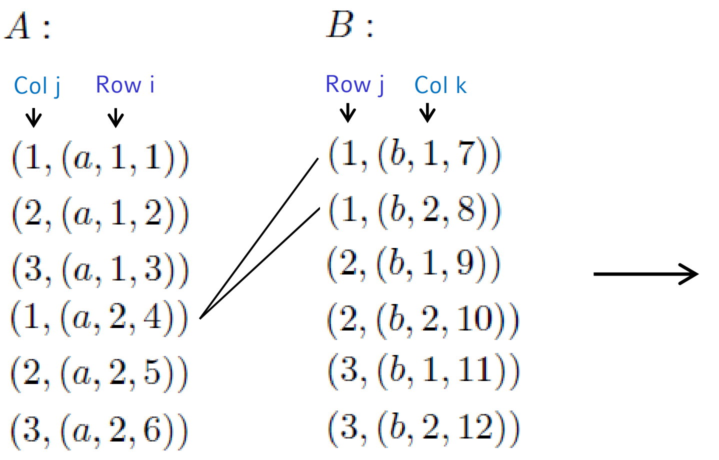
j ↓ $(1,[(a,1,1),(b,1,7)])$

j ↓ $(1,[(a,1,1),(b,2,8)])$

“Join over j“

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

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



$A:$

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

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

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

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

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

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

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

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 ↓

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

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

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

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

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

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

"Join over j"

⟶

j ↓

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

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

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

j ↓

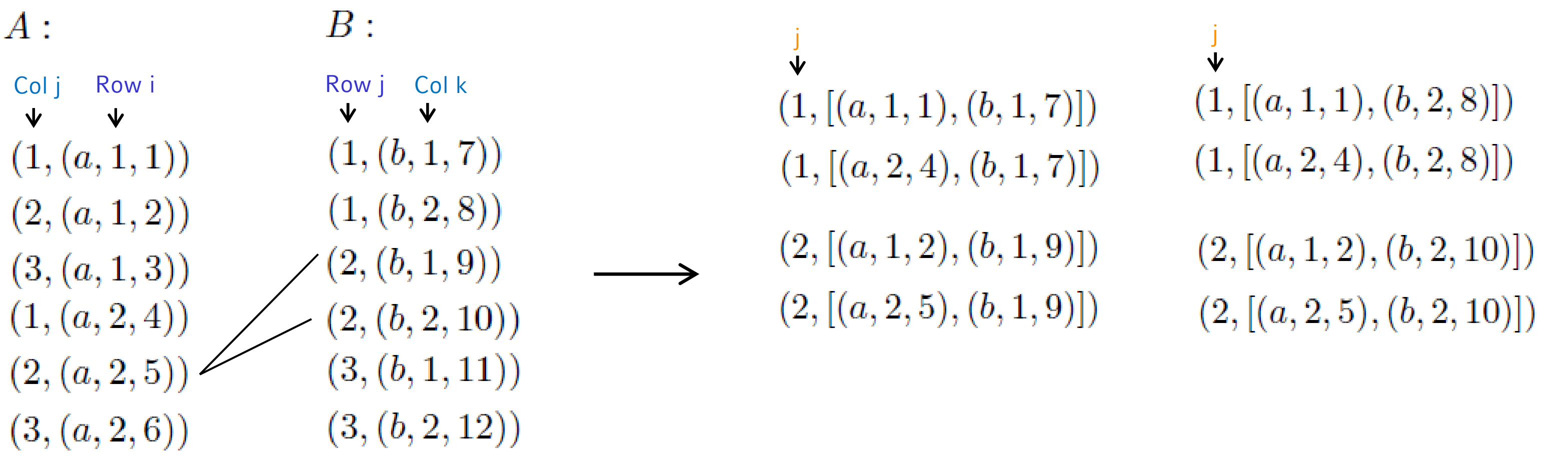
$(1, [(a, 1, 1), (b, 2, 8)])$

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

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

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

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



$A:$

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

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

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

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

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

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

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

$A:$

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 ↓

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

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

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

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

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

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

$\longrightarrow$

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

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

$A:$

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 ↓

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

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

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

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

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

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

“Join over j“

⟶

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 ↓

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

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

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

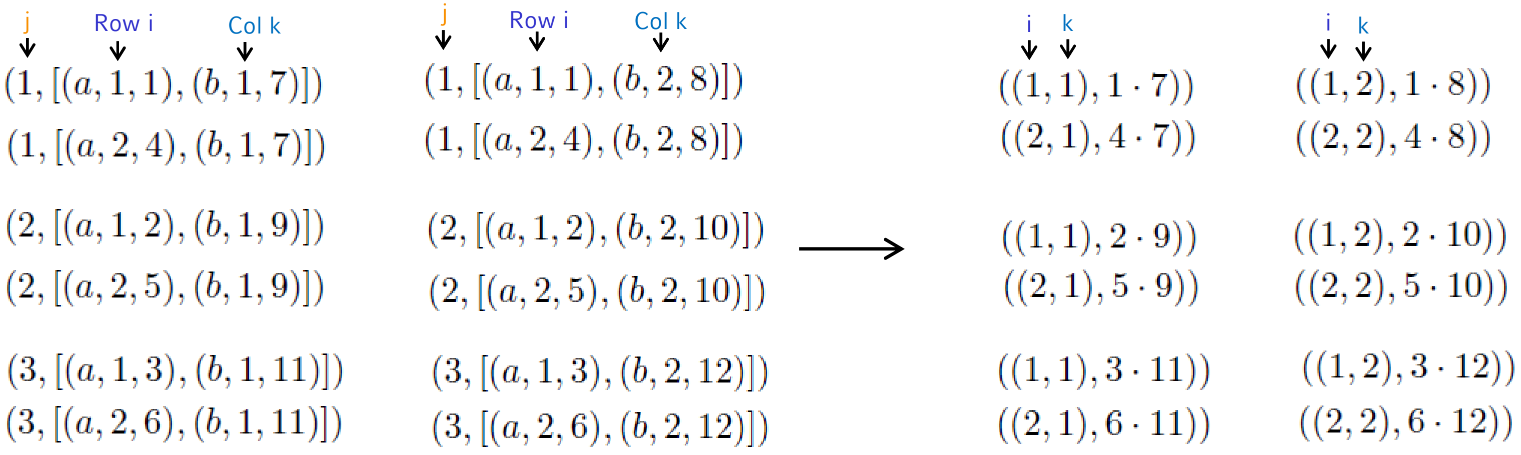
$(2,[(a,2,5),(b,2,10)])$

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

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

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

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



$(1, [(a, 1, 1), (b, 1, 7)])$ $(1, [(a, 1, 1), (b, 2, 8)])$ $((1, 1), 1 \cdot 7))$ $((1, 2), 1 \cdot 8))$

$(1, [(a, 2, 4), (b, 1, 7)])$ $(1, [(a, 2, 4), (b, 2, 8)])$ $((2, 1), 4 \cdot 7))$ $((2, 2), 4 \cdot 8))$

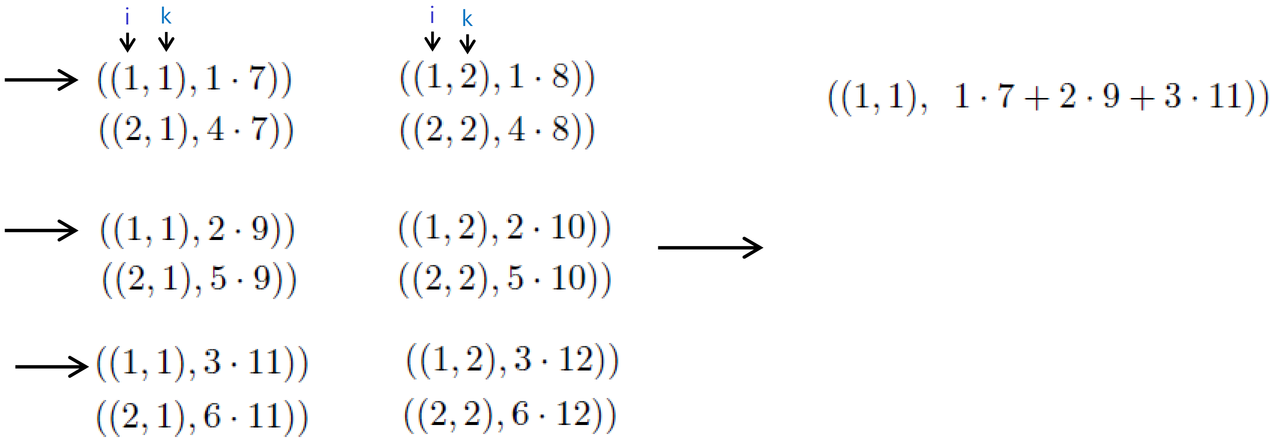
$(2, [(a, 1, 2), (b, 1, 9)])$ $(2, [(a, 1, 2), (b, 2, 10)])$ $\longrightarrow$ $((1, 1), 2 \cdot 9))$ $((1, 2), 2 \cdot 10))$

$(2, [(a, 2, 5), (b, 1, 9)])$ $(2, [(a, 2, 5), (b, 2, 10)])$ $((2, 1), 5 \cdot 9))$ $((2, 2), 5 \cdot 10))$

$(3, [(a, 1, 3), (b, 1, 11)])$ $(3, [(a, 1, 3), (b, 2, 12)])$ $((1, 1), 3 \cdot 11))$ $((1, 2), 3 \cdot 12))$

$(3, [(a, 2, 6), (b, 1, 11)])$ $(3, [(a, 2, 6), (b, 2, 12)])$ $((2, 1), 6 \cdot 11))$ $((2, 2), 6 \cdot 12))$

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

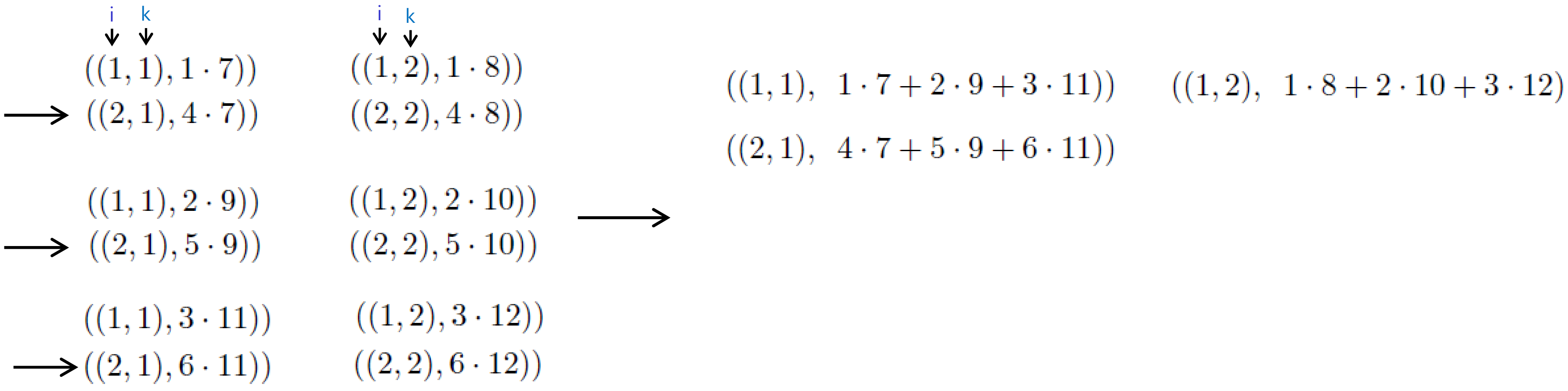


$$\begin{array}{llll}
\longrightarrow ((1,1), 1\cdot 7)) & ((1,2), 1\cdot 8)) & & ((1,1),\ \ 1\cdot 7 + 2\cdot 9 + 3\cdot 11)) \\
((2,1), 4\cdot 7)) & ((2,2), 4\cdot 8)) & & \\
\longrightarrow ((1,1), 2\cdot 9)) & ((1,2), 2\cdot 10)) & \longrightarrow & \\
((2,1), 5\cdot 9)) & ((2,2), 5\cdot 10)) & &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

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: $(lambda\ \ x, y : x + y)$



i k

$((1,1), 1 \cdot 7))$ $((1,2), 1 \cdot 8))$

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

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

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

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

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

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

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

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

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

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: $i \cdot k$

Assignment 4-2

(a) Extend the word count task by computing the average occurrences of each word in a set of documents.

Steps:

- 1. Partition : Split text into block of words
- 2. Map: apply a counter on each word
- 3. Shuffle & Sort: put words which are the same into their own block
- 4. Reduce: sum up the # of occurrences
- 5. Map: divide every number of occurrences by the total # of words

Assignment 4-2

(b) Now compute the standard deviation given the number of occurences of every word. Describe the steps which are necessary for the task using MapReduce

Steps:

…

- 5. Map: divide every number of occurrences by the total # of words
- 6. Reduce: sum all relative occurrences and divide them by the total # of distinct words
- 7. Map: substract from all the values in 5. the computed average (calculate deviations)
- 8. Reduce: sum up all the calculated deviations and divide them by the number of distinct words (calculate variance) and take the square root

Assignment 4-2 (c)

- Partition:

How
much
ground
would
a
groundhog
hog
if
a
groundhog

- Map:

(How,1)
(much,1)
(ground,1)
(would,1)
(a,1)
(groundhog,1)
(hog,1)
(if,1)
(a,1)
(groundhog,1)

Assignment 4-2

- Partition:

could
hog
ground
a
groundhog
would
hog
all
the
ground

- Map:

(could)
(hog,1)
(ground,1)
(a,1)
(groundhog,1)
(would,1)
(hog,1)
(all,1)
(the,1)
(ground,1)

Assignment 4-2

- Partition:

he
could
hog
if
a
groundhog
could
hog
ground

- Map:

(he)
(could,1)
(hog,1)
(if,1)
(a,1)
(groundhog,1)
(could,1)
(hog,1)
(ground,1)

Assignment 4-2

- Shuffle & Sort:

(how, 1)

(much, 1)

(ground, 1)
(ground, 1)

(would, 1)
(would, 1)

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

(groundhog, 1)
(groundhog, 1)
(groundhog, 1)
(groundhog, 1)

(hog, 1)
(hog, 1)
(hog, 1)
(hog, 1)
(hog, 1)

(if, 1)
(if, 1)

(could, 1)
(could, 1)
(could, 1)

(all, 1)

(the, 1)

(he, 1)

Assignment 4-2

- Reduce:

(how, 1)

(much, 1)

(ground, 2)

(would, 2)

(a, 4)

(groundhog, 4)

(hog,5)

(if, 2)

(could, 3)

(all, 1)

(the, 1)

(he, 1)

Assignment 4-2

- Map: (total # of words: 12)

(how, 0.083)

(much, 0.083)

(ground, 0.166)

(would, 0.166)

(a, 0.333)

(groundhog, 0.333)

(hog, 0.416)

(if, 0.166)

(could, 0.25)

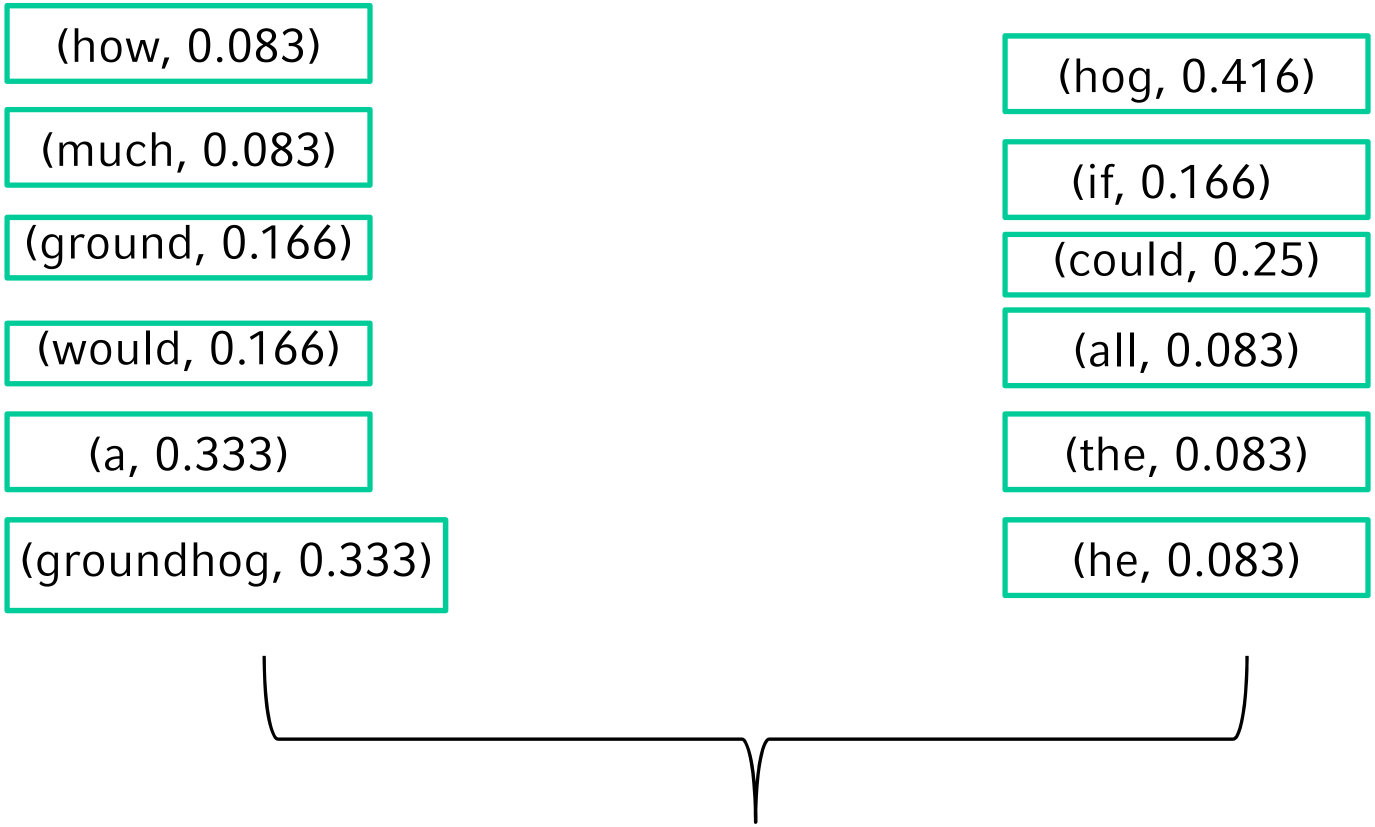
(all, 0.083)

(the, 0.083)

(he, 0.083)

Assignment 4-2

- Reduce: (total # of words: 12)



(how, 0.083)

(much, 0.083)

(ground, 0.166)

(would, 0.166)

(a, 0.333)

(groundhog, 0.333)

(hog, 0.416)

(if, 0.166)

(could, 0.25)

(all, 0.083)

(the, 0.083)

(he, 0.083)

$$\frac{\Sigma}{12} = 2.245/12 = 0.187$$