NYU | Center for Data Science

DS-GA 1007
Programming for Data Science

Lecture 12
pandas III + SQL II – Operations on Tables

Dask package in Python splits large tables into small tables and large tasks into dependent small tasks

DS-GA 1007
Programming for Data Science

Lecture 12
pandas III + SQL II – Operations on Tables

Dask package in Python allows us to scale pandas operations to large datasets

DS-GA 1007
Programming for Data Science

Lecture 12
pandas III + SQL II – Operations on Tables

Announcements

- Homework 9 due **Monday November 25** at 11:59pm
- Project
  - Milestone due **Thursday November 28** at 11:59pm
  - Background and Plans
- Labs
  - Submit on Jupyter Hub under Assignments tab
  - Access scores from Submitted Assignments under Assignments tab

Agenda

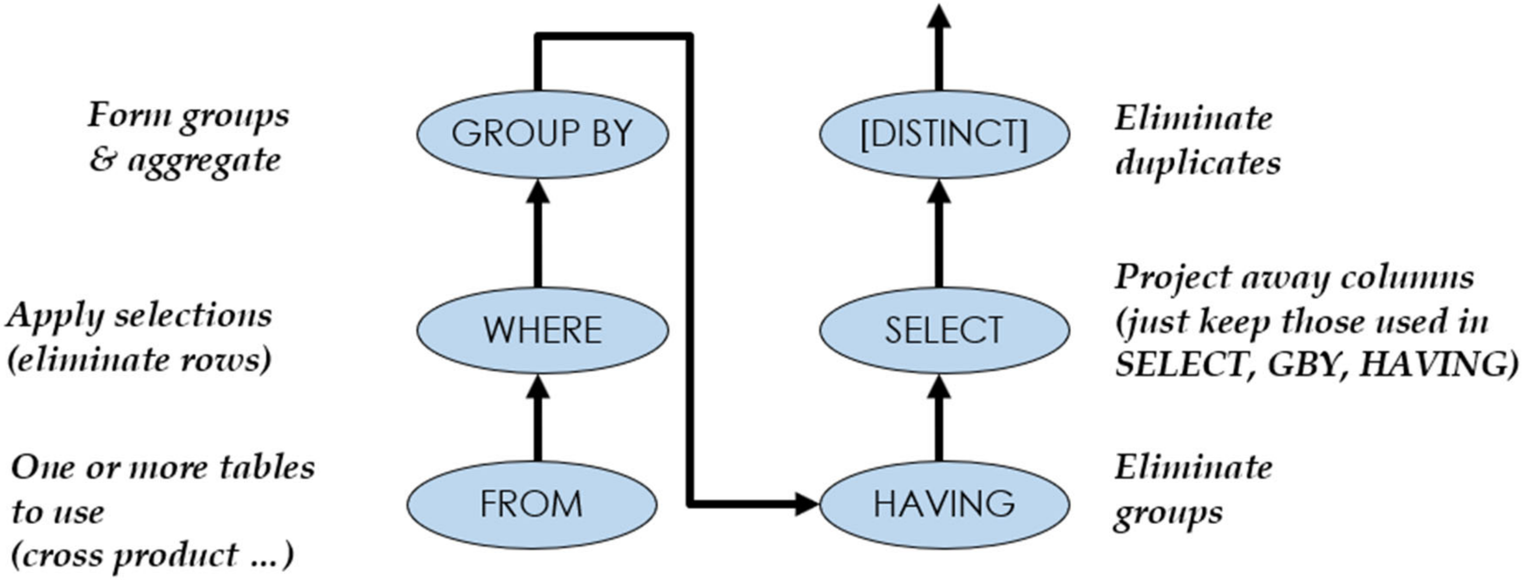
- Lesson
  - pandas
  - SQL
  - Dask
- Demos
  - Joins
  - Grouping
  - Pivot / Unpivot
  - Comparing running time and memory

Objectives

- How can we combine multiple tables through joins?
- Group together rows by values in columns
- Pivot between rows and columns in a table
- How can we access large datasets?

SQL Commands

```
SELECT      [DISTINCT]  target-list
FROM        relation-list
WHERE       qualification
GROUP BY    grouping-list
HAVING      group-qualification
```



Form groups & aggregate

Apply selections (eliminate rows)

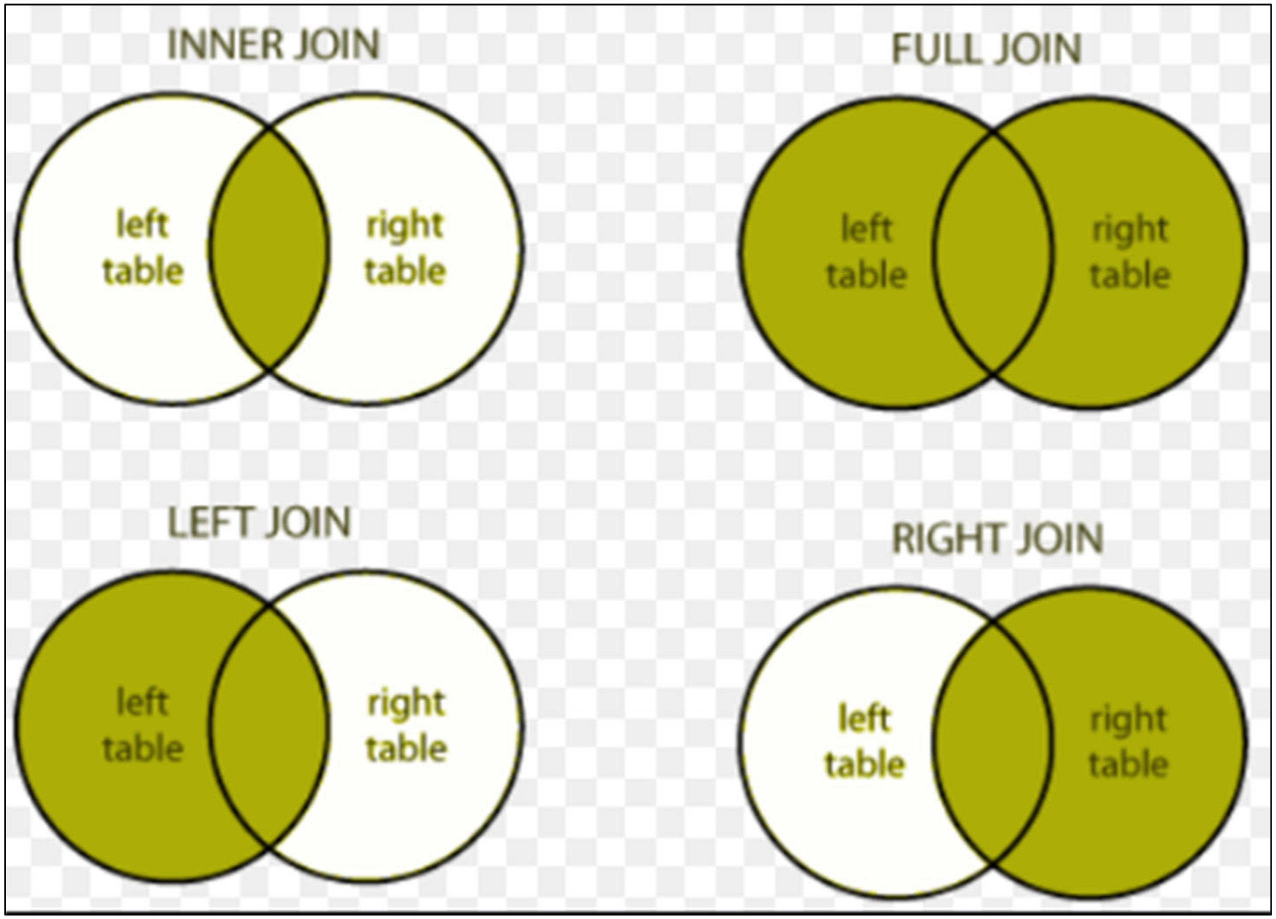
One or more tables to use (cross product ...)

Eliminate duplicates

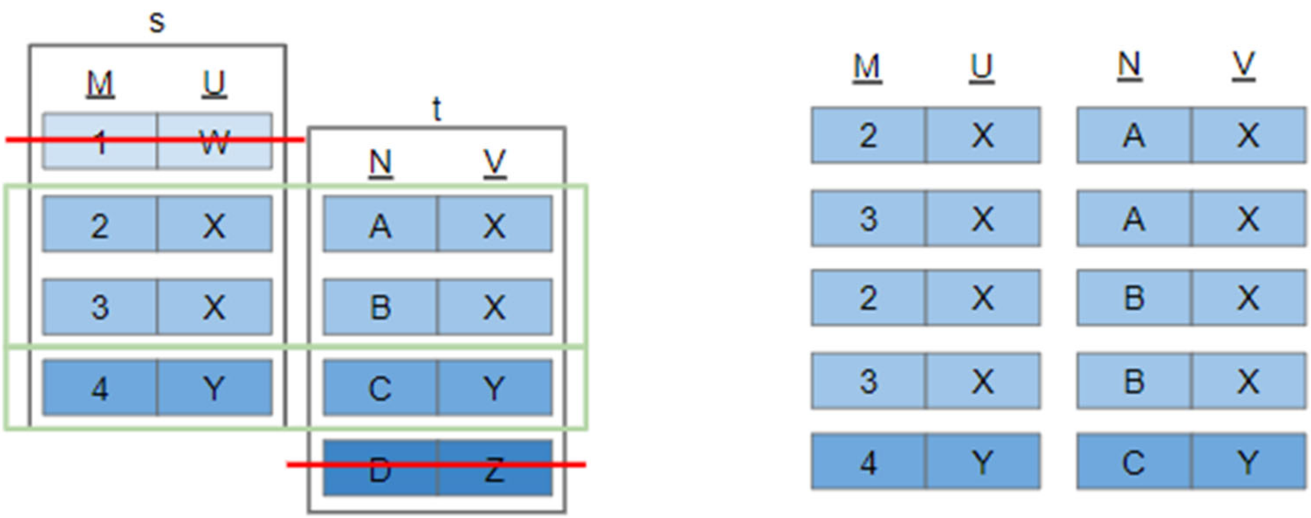
Project away columns (just keep those used in SELECT, GBY, HAVING)

Eliminate groups

SQL Joins


INNER JOIN
left table
right table
FULL JOIN
left table
right table
LEFT JOIN
left table
right table
RIGHT JOIN
left table
right table

SQL Inner Join



s

M	U
~~1~~	~~W~~
2	X
3	X
4	Y

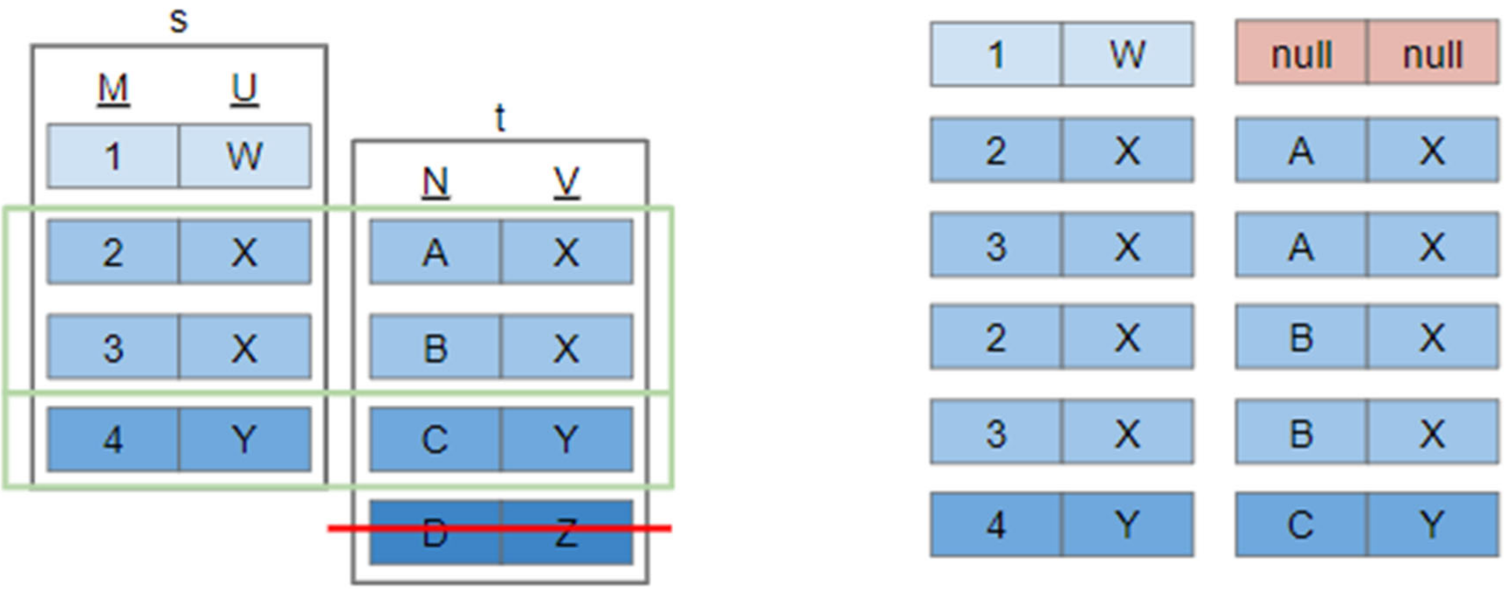
t

N	V
A	X
B	X
C	Y
~~D~~	~~Z~~

M	U	N	V
2	X	A	X
3	X	A	X
2	X	B	X
3	X	B	X
4	Y	C	Y

```
SELECT * FROM s JOIN t ON s.u = t.v;
SELECT * FROM s INNER JOIN t ON s.u = t.v;
SELECT * FROM s, t WHERE s.u = t.v;
```

SQL Left Outer Join



s

M	U
1	W
2	X
3	X
4	Y

t

N	V
A	X
B	X
C	Y
~~D~~	~~Z~~

1	W	null	null
2	X	A	X
3	X	A	X
2	X	B	X
3	X	B	X
4	Y	C	Y

```
SELECT * FROM s LEFT JOIN t ON s.u = t.v;
```

SQL Cross Join

s

M	U
1	W
2	X
3	X
4	Y

t

N	V
A	X
B	X
C	Y
D	Z

```
SELECT * FROM s, t;
```

M	U	N	V
1	W	A	X
2	X	A	X
3	X	A	X
4	Y	A	X
1	W	B	X
2	X	B	X
3	X	B	X
4	Y	B	X

(to be continued ...)

(... continued)

M	U	N	V
1	W	C	Y
2	X	C	Y
3	X	C	Y
4	Y	C	Y
1	W	D	Z
2	X	D	Z
3	X	D	Z
4	Y	D	Z

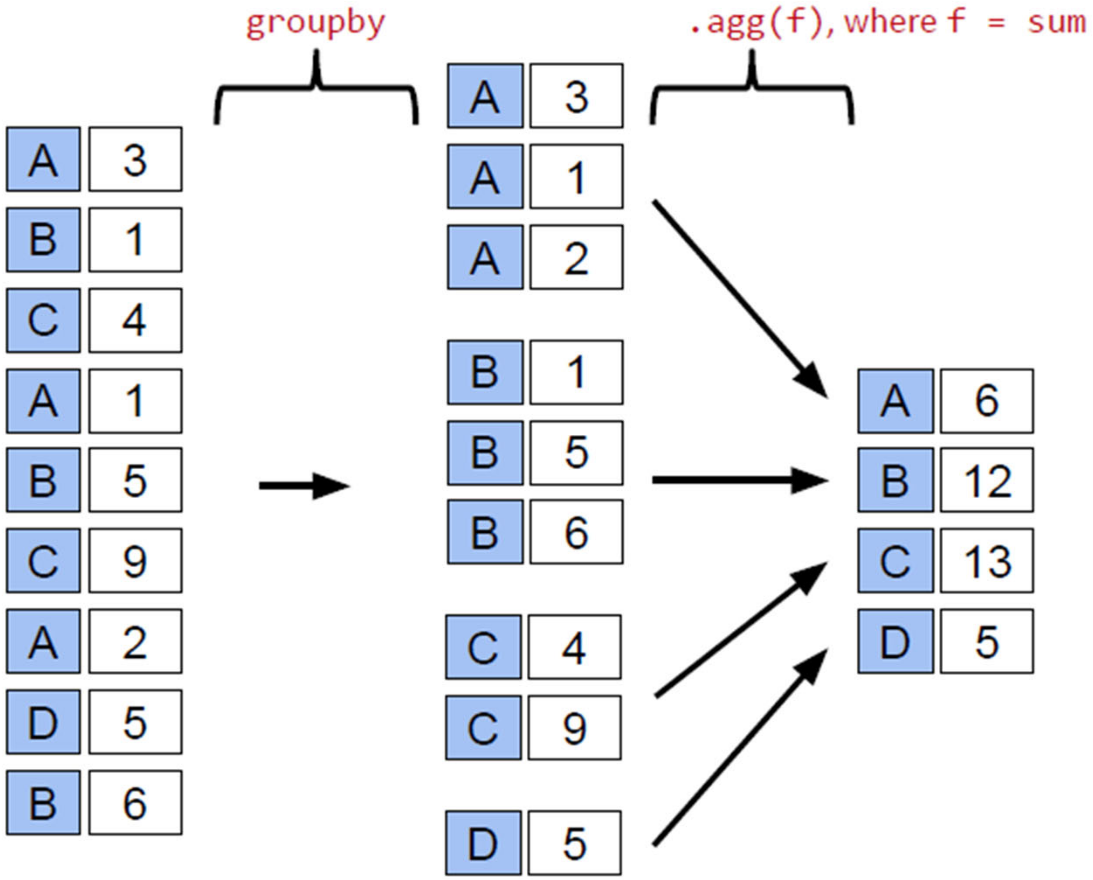
Agenda

- Lesson
  - pandas
  - SQL
  - Dask
- Demos
  - Joins
  - Grouping
  - Pivot / Unpivot
  - Comparing running time and memory

Objectives

- How can we combine multiple tables through joins?
- Group together rows by values in columns
- Pivot between rows and columns in a table
- How can we access large datasets?

Group



groupby

.agg(f), where f = sum

A	3
B	1
C	4
A	1
B	5
C	9
A	2
D	5
B	6

A	3
A	1
A	2
B	1
B	5
B	6
C	4
C	9
D	5

A	6
B	12
C	13
D	5

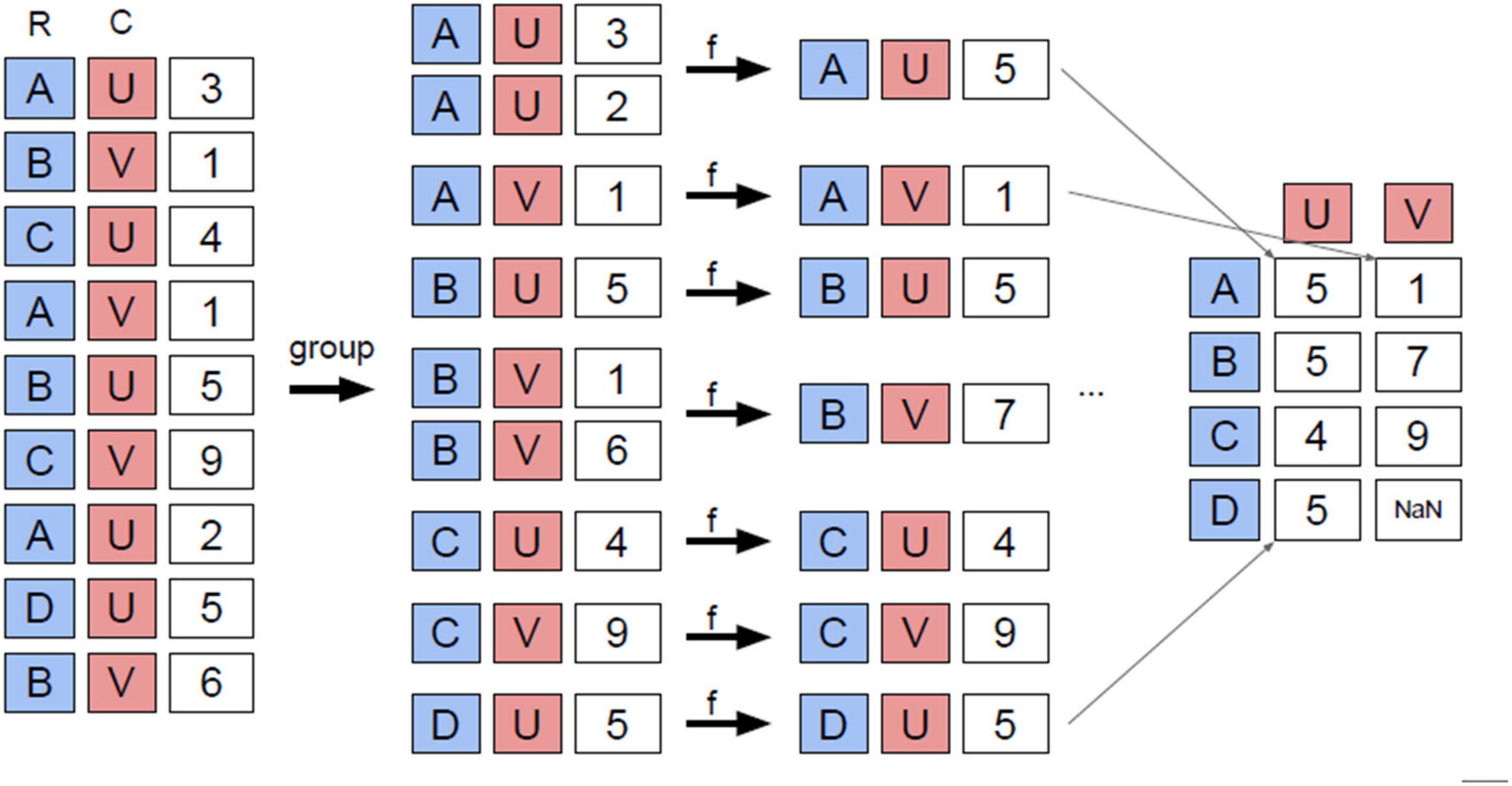
Agenda

- Lesson
  - pandas
  - SQL
  - Dask
- Demos
  - Joins
  - Grouping
  - Pivot / Unpivot
  - Comparing running time and memory

Objectives

- How can we combine multiple tables through joins?
- Group together rows by values in columns
- Pivot between rows and columns in a table
- How can we access large datasets?

Pivot



R	C	
A	U	3
B	V	1
C	U	4
A	V	1
B	U	5
C	V	9
A	U	2
D	U	5
B	V	6

group

A	U	3
A	U	2
A	V	1
B	U	5
B	V	1
B	V	6
C	U	4
C	V	9
D	U	5

f

A	U	5
A	V	1
B	U	5
B	V	7
C	U	4
C	V	9
D	U	5

...

	U	V
A	5	1
B	5	7
C	4	9
D	5	NaN

Agenda

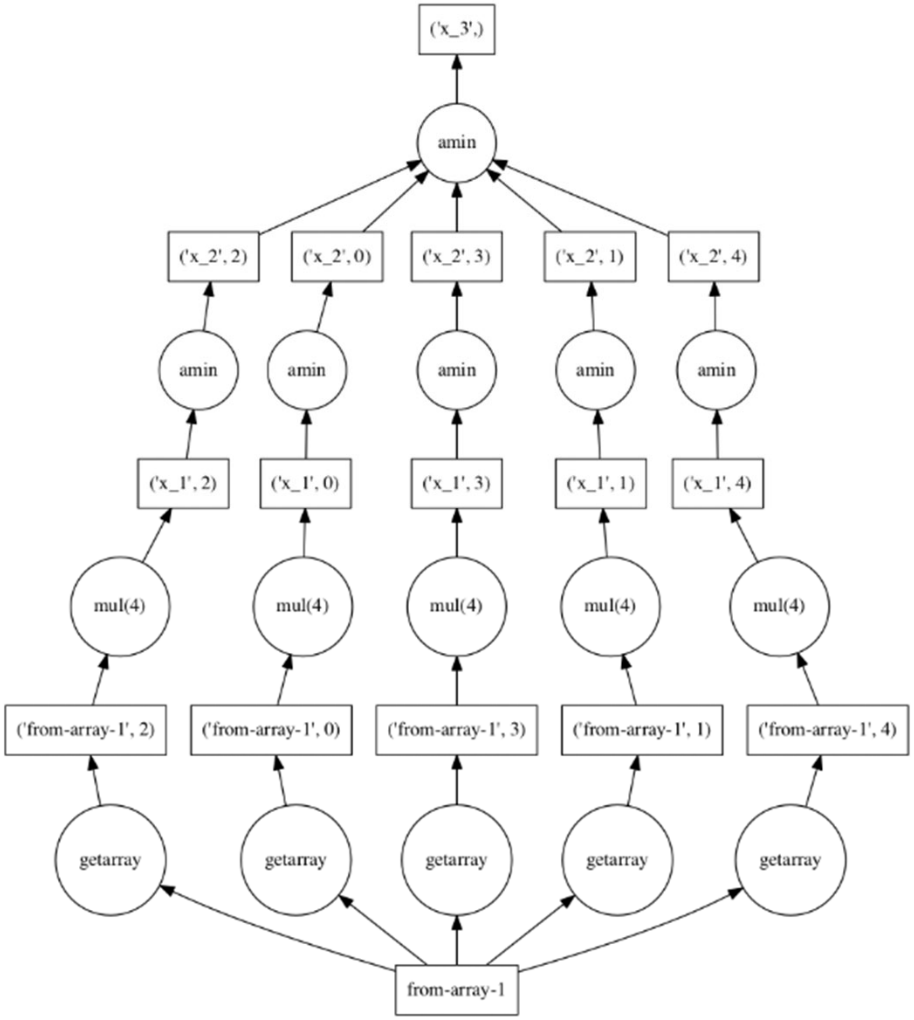
- Lesson
  - pandas
  - SQL
  - Dask
- Demos
  - Joins
  - Grouping
  - Pivot / Unpivot
  - ➡ Comparing running time and memory

Objectives

- How can we combine multiple tables through joins?
- Group together rows by values in columns
- Pivot between rows and columns in a table
- How can we access large datasets?

Dask

:[5]:


('x_3',)
amin
('x_2', 2)
('x_2', 0)
('x_2', 3)
('x_2', 1)
('x_2', 4)
amin
amin
amin
amin
amin
('x_1', 2)
('x_1', 0)
('x_1', 3)
('x_1', 1)
('x_1', 4)
mul(4)
mul(4)
mul(4)
mul(4)
mul(4)
('from-array-1', 2)
('from-array-1', 0)
('from-array-1', 3)
('from-array-1', 1)
('from-array-1', 4)
getarray
getarray
getarray
getarray
getarray
from-array-1

Summary

- pandas and SQL
  - Joining
  - Grouping
  - Pivot
- Dask
  - Running Time
  - Memory