Scaling NumPy and Pandas

Paweł Kordek
SOFTWARE ENGINEER

@pawel_kordek https://kordek.github.io

Arrays and DataFrames

Objectives

Dataset

Demos

Dask's Array and DataFrame Objective

Provide drop-in, scalable replacements for NumPy and Pandas.

Dask APIs

Bags

Arrays and DataFrames

Dask APIs

Bags	Arrays and DataFrames
High-level	Much more specialized

Dask APIs

Bags	Arrays and DataFrames
High-level	Much more specialized
Does not 'cover' existing library	Expose functionalities of existing libs

Dask APIs

Bags	Arrays and DataFrames
High-level	Much more specialized
Does not 'cover' existing library	Expose functionalities of existing libs
Requires code changes	Minimize code changes

Arrays and DataFrames

Use original libraries underneath

Code changes only in few places

BUT they lack some parts of the original APIs

Dataset

Cellular Network Usage

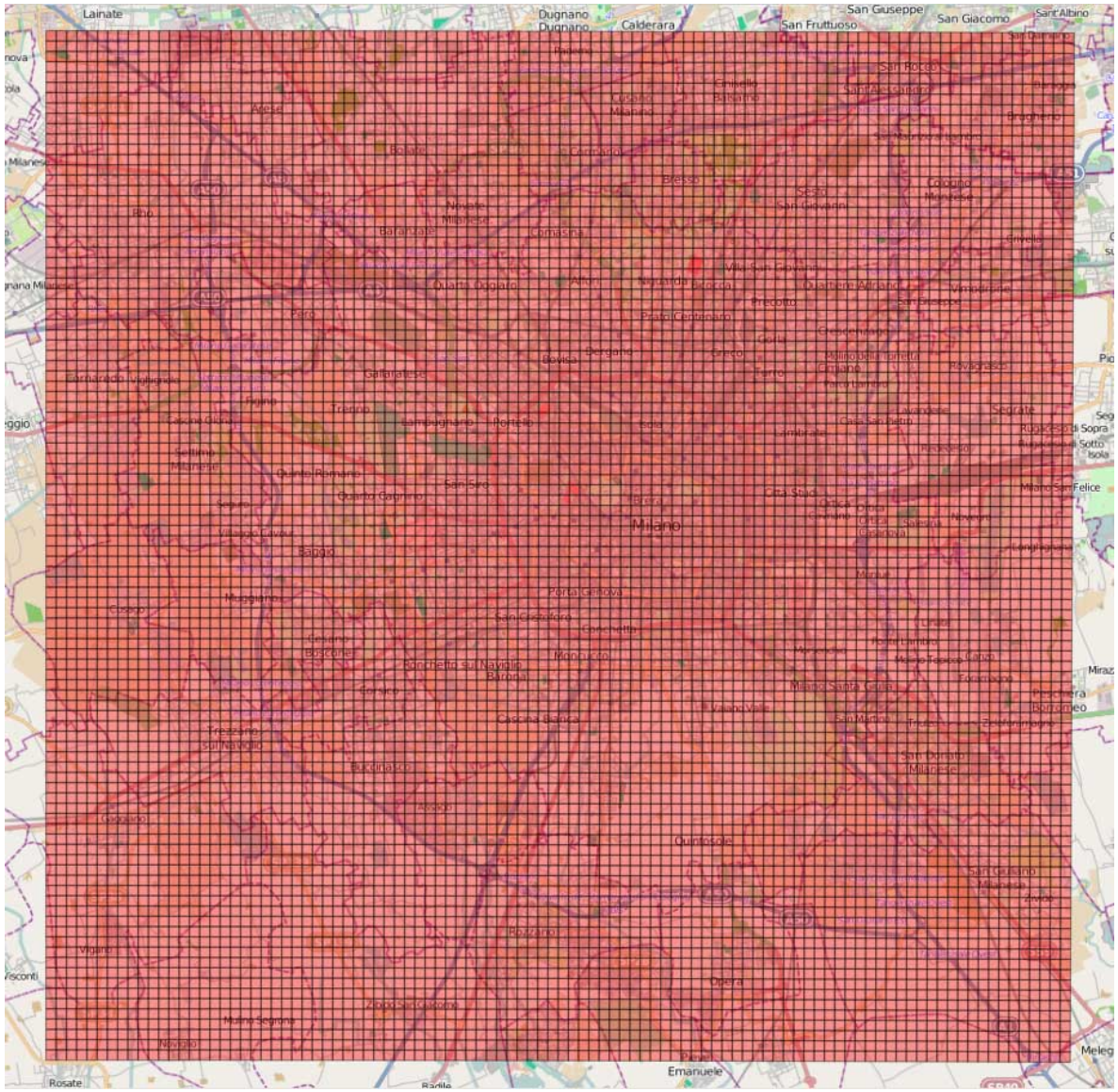
One month

One day

ID of the square in the Milano Grid

square

10							


Lainate
Dugnano
Calderara
San Giuseppe
San Fruttuoso
San Giacomo
Sant'Albino
Milano
Quinto Stampi
Opera
Rosate
Emanuele

Milano Grid

Start of the 10 minute interval

square	interval_start						
10	1386670200						

Incoming SMS activity

square	interval_start		sms_in				
10	1386670200		0.19600				

Outgoing SMS activity

square	interval_start		sms_in	sms_out			
10	1386670200		0.19600	0.73170			

Incoming call activity

square	interval_start		sms_in	sms_out	call_in		
10	1386670200		0.19600	0.73170	0.68417		

Outgoing call activity

square	interval_start		sms_in	sms_out	call_in	call_out	
10	1386670200		0.19600	0.73170	0.68417	0.59840	

Mobile data activity

square	interval_start		sms_in	sms_out	call_in	call_out	data
10	1386670200		0.19600	0.73170	0.68417	0.59840	5.80175

Phone's country prefix

square	interval_start	country	sms_in	sms_out	call_in	call_out	data
10	1386670200	39	0.19600	0.73170	0.68417	0.59840	5.80175

Phone's country prefix

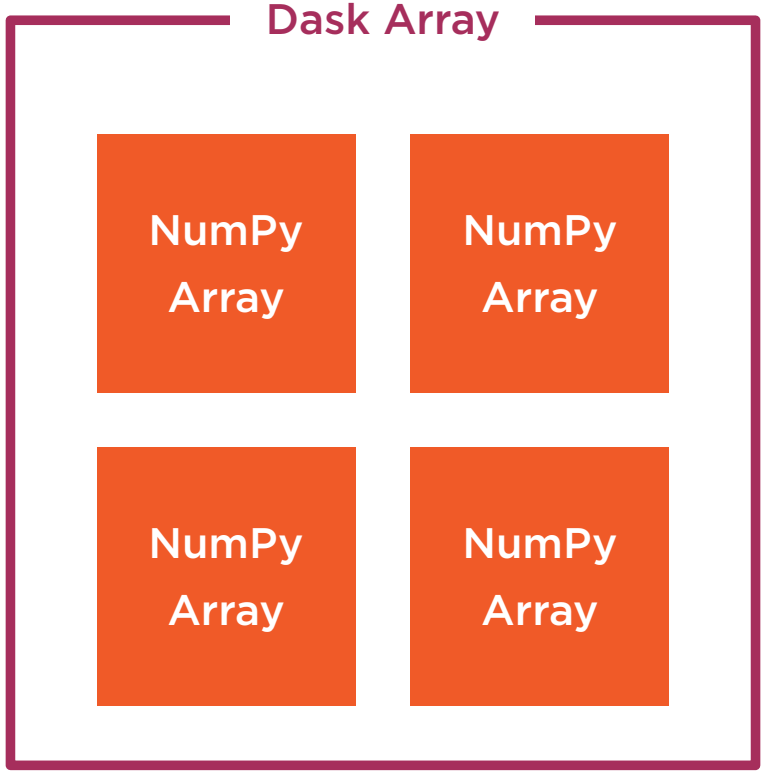
Identifier variables

square	interval_start	country	sms_in	sms_out	call_in	call_out	data
10	1386670200	39	0.19600	0.73170	0.68417	0.59840	5.80175

Phone's country prefix

square	interval_start	country	Measured variables				
			sms_in	sms_out	call_in	call_out	data
10	1386670200	39	0.19600	0.73170	0.68417	0.59840	5.80175

Arrays


Dask Array
NumPy Array
NumPy Array
NumPy Array
NumPy Array

```python
import dask.array as da
import numpy as np

a = da.from_array(np.genfromtxt("file.txt"))
```

Creating Array

Directly from NumPy array.

```
import dask.array as da
import numpy as np

a = da.from_array(np.genfromtxt("file.txt"))

a = da.from_array(np.load("file.npy"))
```

Creating Array

Directly from NumPy array.

```
import dask.array as da

a = da.from_npy_stack("src_dir")

da.to_npy_stack("dst_dir", a)
```

Storing and Loading Array

Dask + NumPy binary format.

```
import dask.array as da

a = da.from_npy_stack("src_dir")
```

Keeping Data in Native Formats

Shortest ramp-up times.

Why No Direct Support for Plaintext?

Such data is usually stored in binary formats.

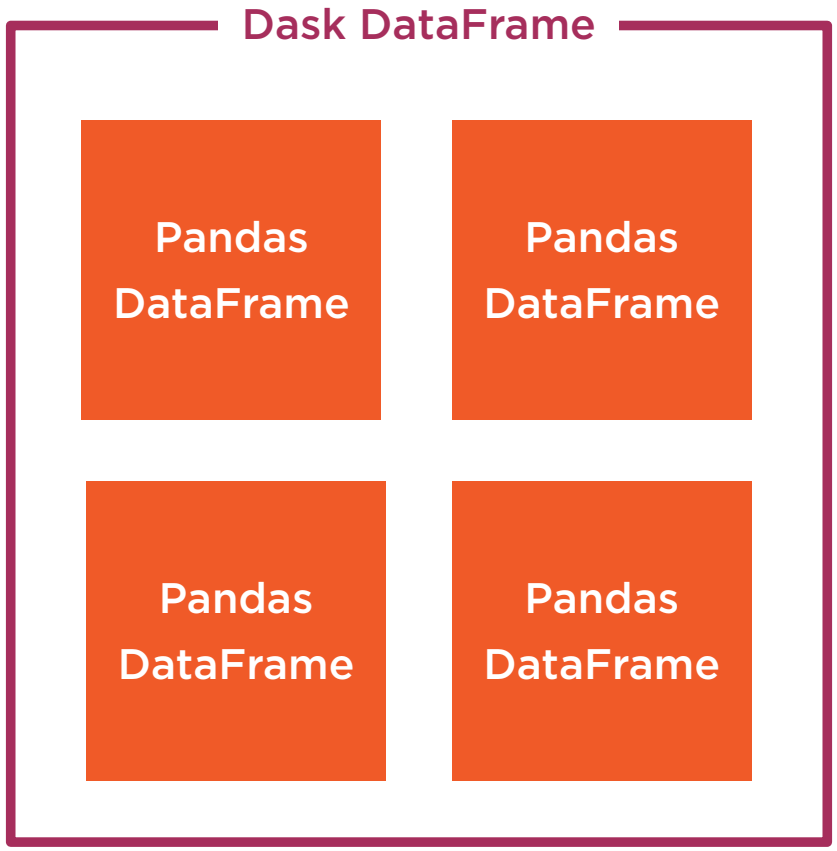
Performance difference is huge.

```
import dask.array as da

a = da.from_npy_stack("src_dir")

a + 4  # NumPy-style calculations
```

Keeping Data in Native Formats

Shortest ramp-up times.

DataFrames


Dask DataFrame
Pandas DataFrame
Pandas DataFrame
Pandas DataFrame
Pandas DataFrame

Task

Find hourly distribution of the network usage.

Summary

Arrays and DataFrames larger than memory

Easy transition

Coverage is extensive