Data Containers

pandas

Francesc Alted
Freelance Consultant
http://www.blosc.org/professional-services.html

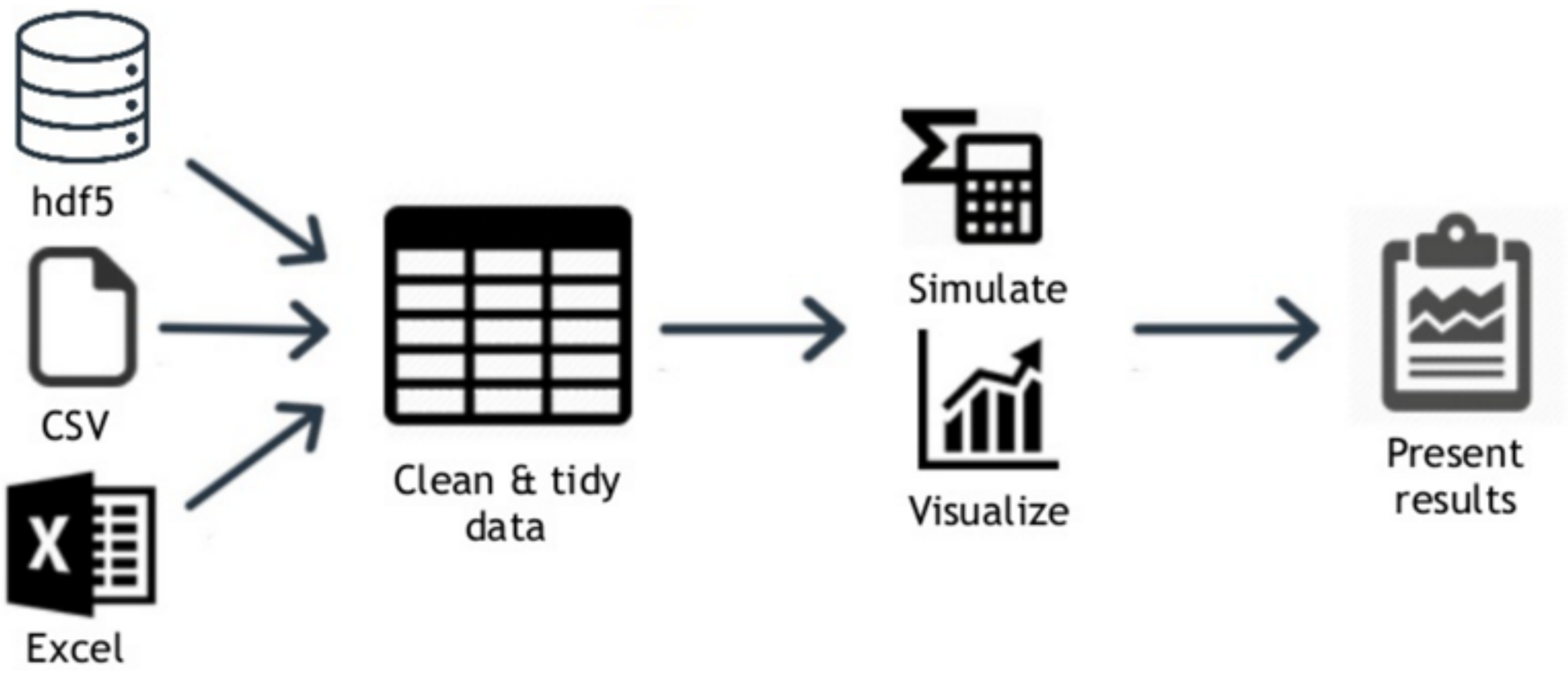
Advanced Scientific Programming in Python
Reading, UK
September, 2016

What It Is

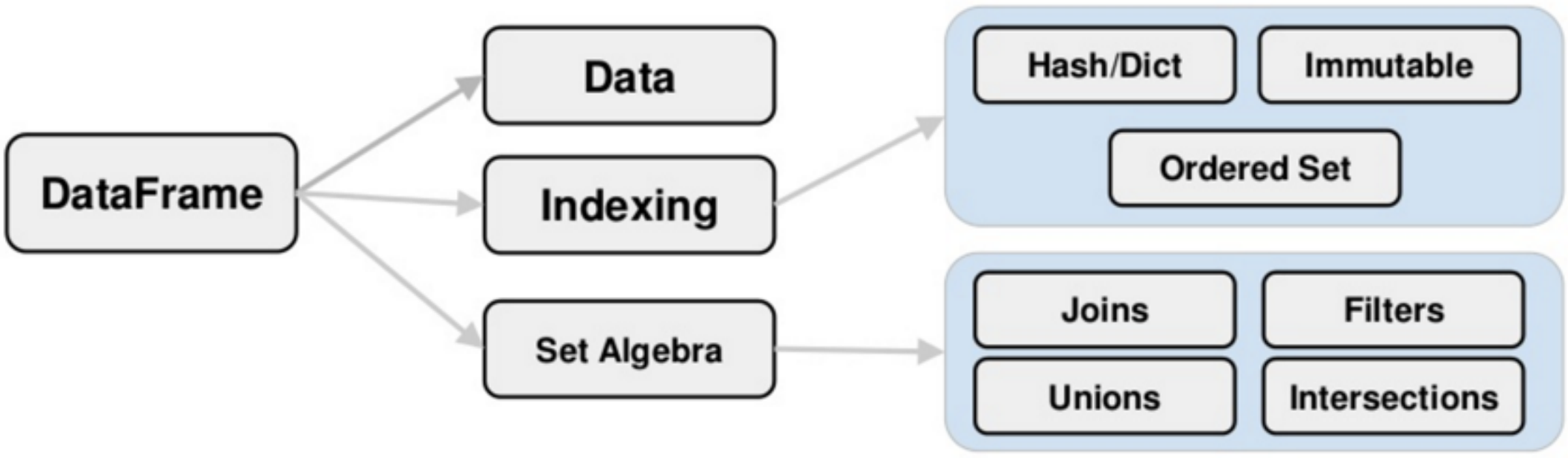
- pandas is a Python package providing fast, flexible and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive.

- It aims to be the fundamental high-level building block for doing practical, real world analysis in Python

Data. pandas. profit!

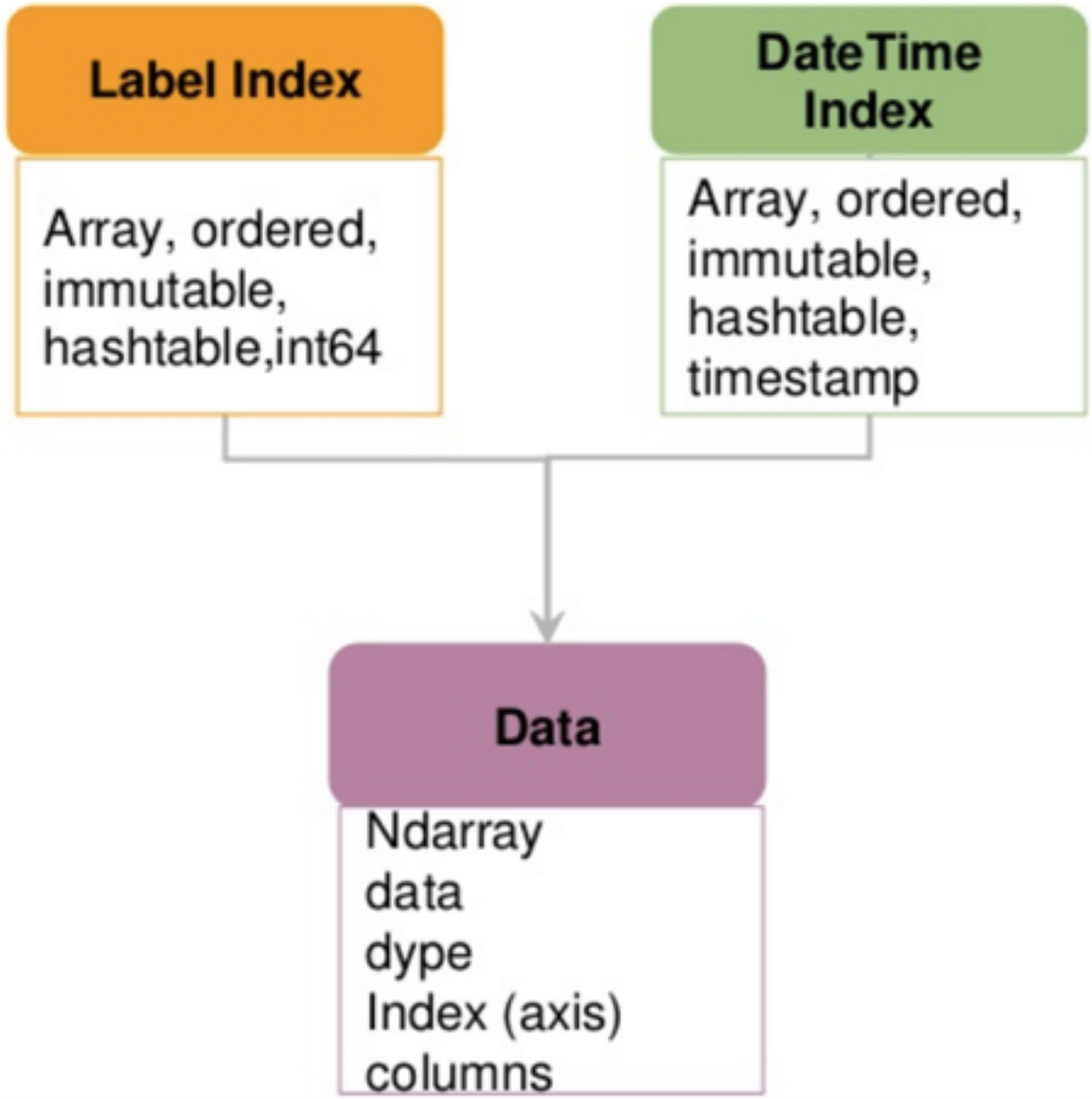


pandas DataFrames



Indexing a DataFrame

Year	Total	Gas	Liquid	Solid
1997	250255	12561	66649	159191
1998	255310	12990	71750	158106
1999	271548	11549	77852	169087
2000	281389	11974	82834	172812
...				

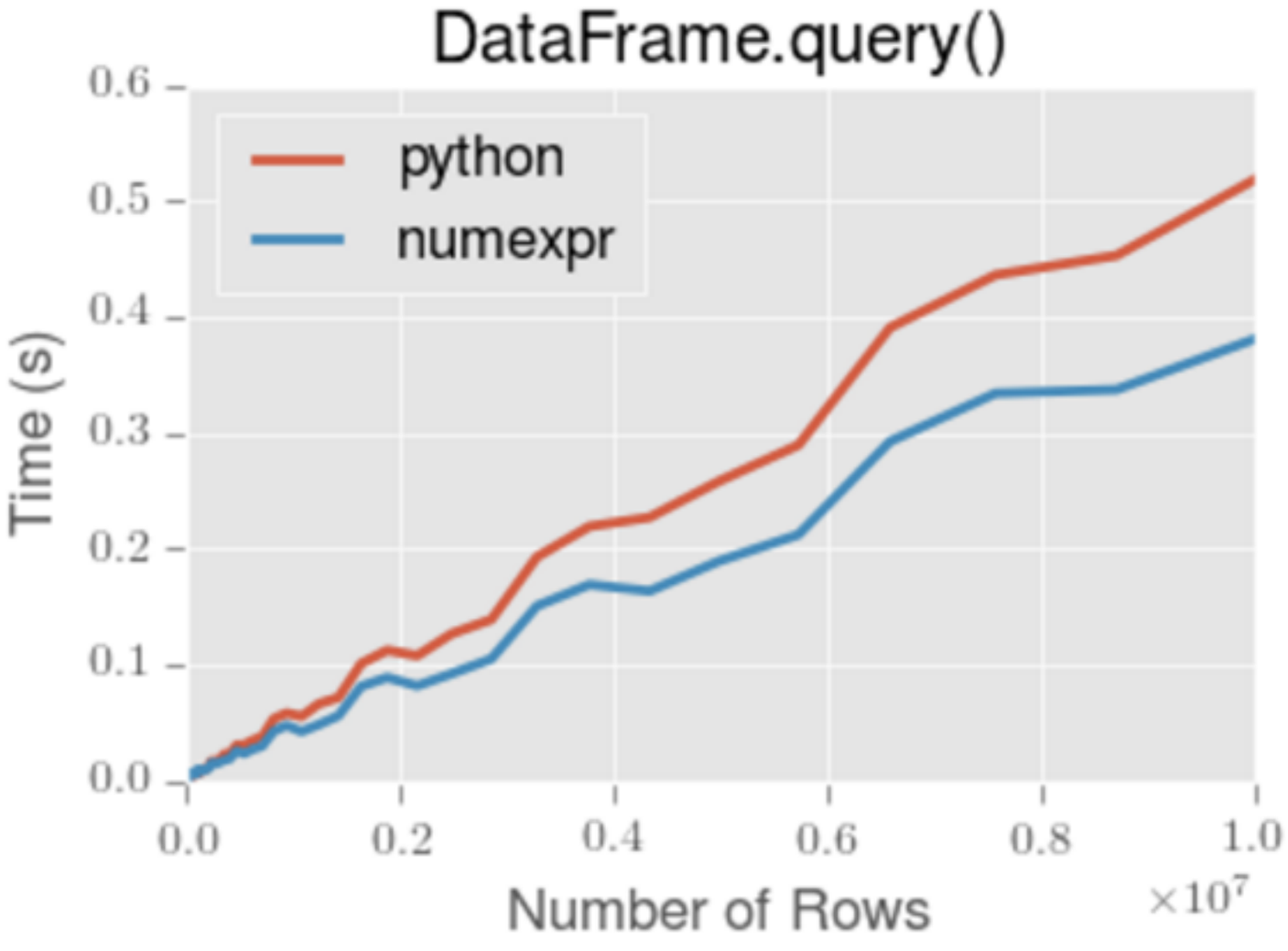


Querying a DataFrame

Year	Total	Gas	Liquid	Solid
1997	250255	12561	66649	159191
1998	255310	12990	71750	158106
1999	271548	11549	77852	169087
2000	281389	11974	82834	172812
...				

```
df.query('(Year == 1997) & (Gas < Solid)')
```

DataFrame.query() can use numexpr for performance



Joining 2 DataFrames

left

	A	B	key
0	A0	B0	K0
1	A1	B1	K1
2	A2	B2	K2
3	A3	B3	K3

right

	C	D	key
0	C0	D0	K0
1	C1	D1	K1
2	C2	D2	K2
3	C3	D3	K3

Result

	A	B	key	C	D
0	A0	B0	K0	C0	D0
1	A1	B1	K1	C1	D1
2	A2	B2	K2	C2	D2
3	A3	B3	K3	C3	D3

- Many-to-Many join case (Cartesian Product)
- This case is simple (one unique key combination)

I/O Connectors

- SQL
- Excel
- CSV
- HDF5 (PyTables)
- BigQuery
- Relational Databases
- JSON