Data Containers

Introduction

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

Advanced Scientific Programming in Python
Reading, UK
September, 2016

Goals

- Get in contact with well known (and less well known) data containers, showing advantages and disadvantages

- Profile memory consumption of data containers

- Use high-performance compression as a way to alleviate the I/O bandwidth bottleneck

Main Data Containers that We Are Going to Visit

- Basic standard Python containers: tuples, lists, dicts
- NumPy (old friend)
- pandas (a de-facto standard from data analysis)
- bcolz (an in-memory, on-disk, compressed container)
- PyTables (an HDF5-based container)
- SQLite (as an example of relational databases)

Let’s Start With Some Exercises

Go to:
https://python.g-node.org/wiki/data_containers

and follow the instructions there.

Take Home Messages

- Fortunately, there is a huge range of data containers out there for your convenience

- **Experiment** with them and use the ones that are more convenient for your own use.

- Due to computer evolution, compression is becoming more and more important for data containers meant for big datasets.