Blosc and PyTables: What's New

Francesc Alted

PyData NYC 2019, New York, November 2019

What is Blosc?

- High performance compressor for binary data
- Can potentially go faster than memory copies
- Used in different projects (PyTables, Zarr, bcolz, the Julia language and probably many more). HDF5 is planning in adopting it as a core compressor.

https://blosc.org

What’s New in Blosc2

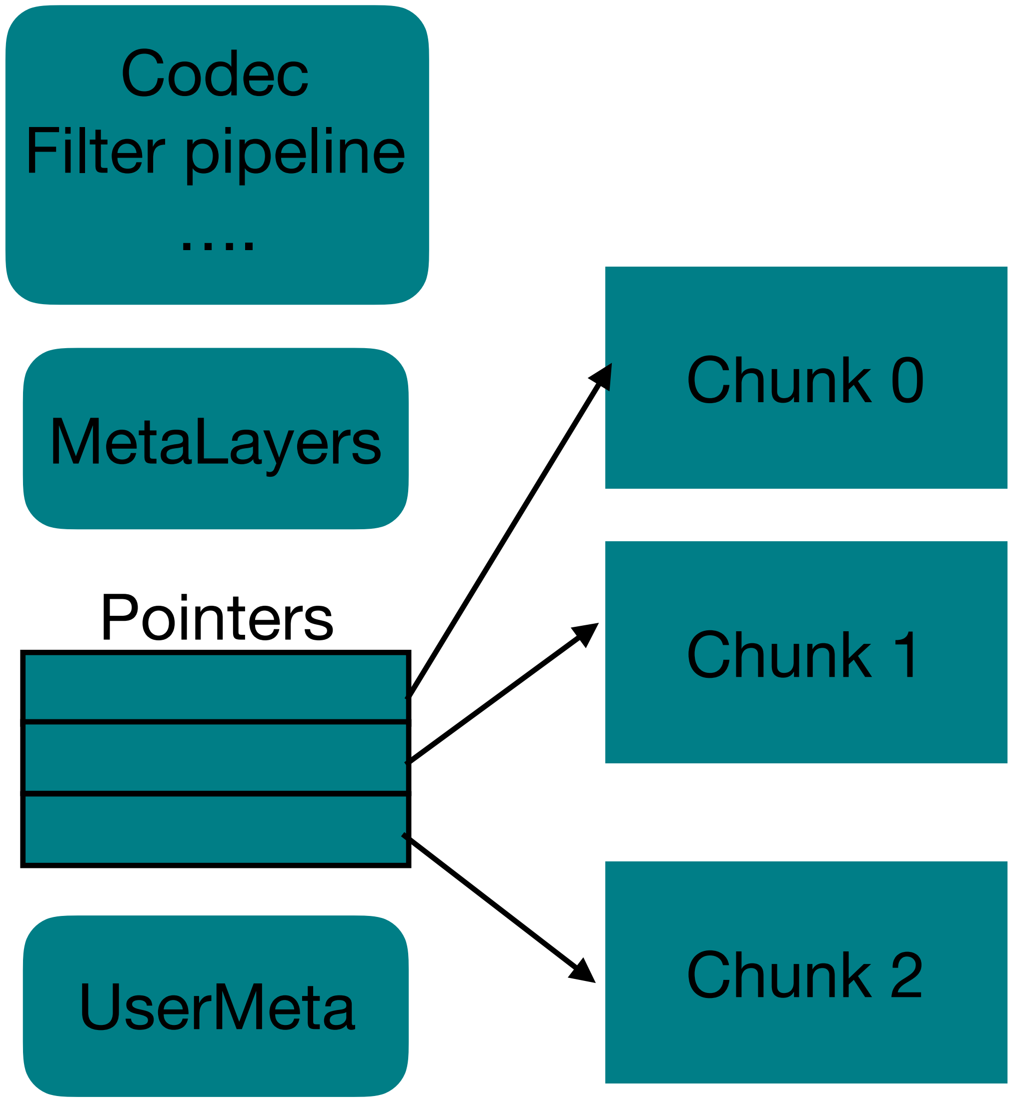
- Blosc2 entered beta stage: 4 beta releases are out already. Please help us testing the package!
- NumFOCUS provided a small development grant ($5000 USD) that we used for putting Blosc2 in beta
- Persistent format is in beta and well defined, but still missing some parts of the implementation (fingerprint)

New Features in Blosc2

- Enlargeable 64-bit containers: in-memory or on-disk
- New compression codecs
- New filters
- Metalayers
- User metadata

Containers in Blosc2

Super-chunk



Codec
Filter pipeline
....

MetaLayers

Pointers

Chunk 0

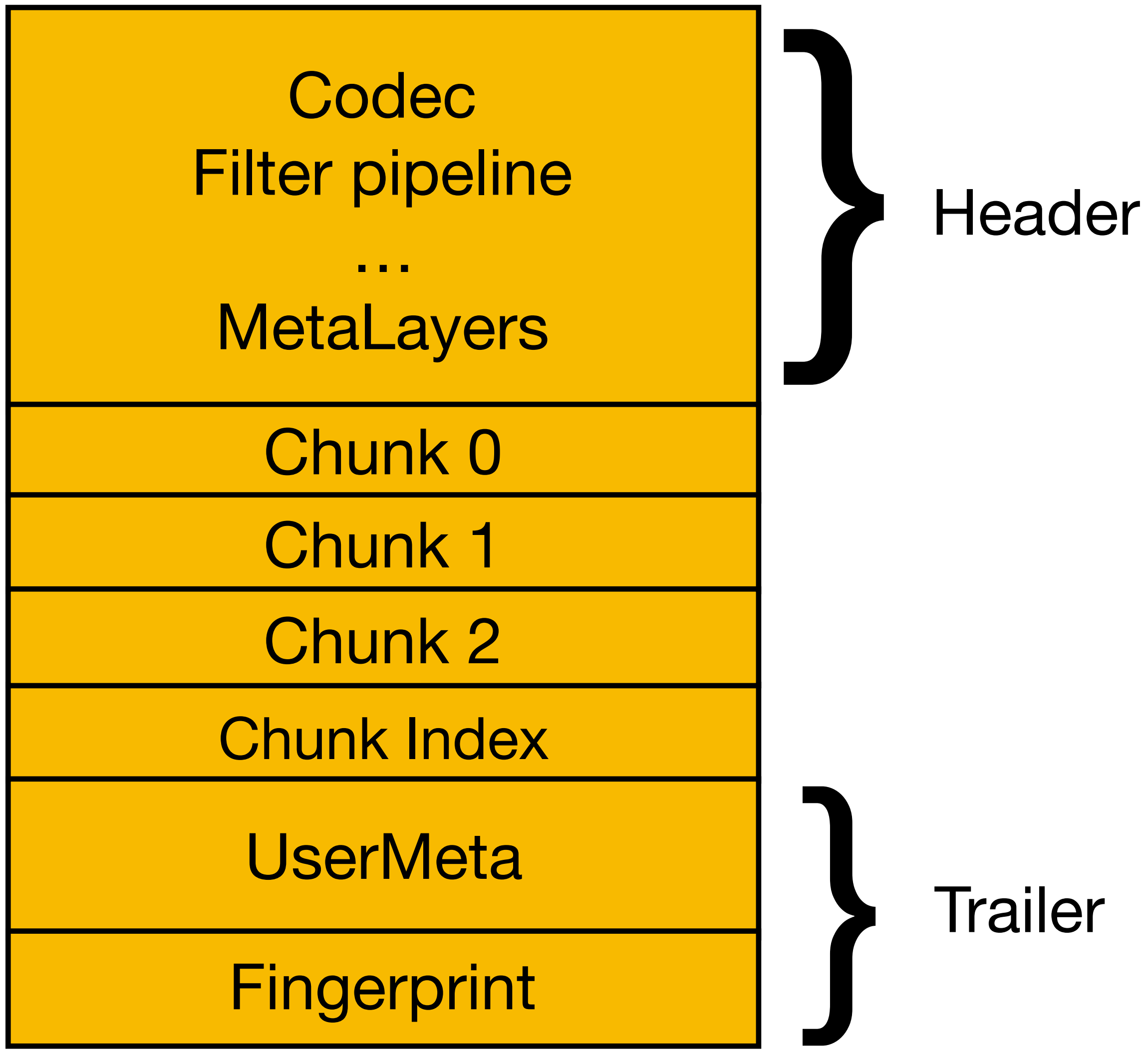
Chunk 1

Chunk 2

UserMeta

- Sparse
- In-memory

Frame



Codec
Filter pipeline
...
MetaLayers

Header

Chunk 0

Chunk 1

Chunk 2

Chunk Index

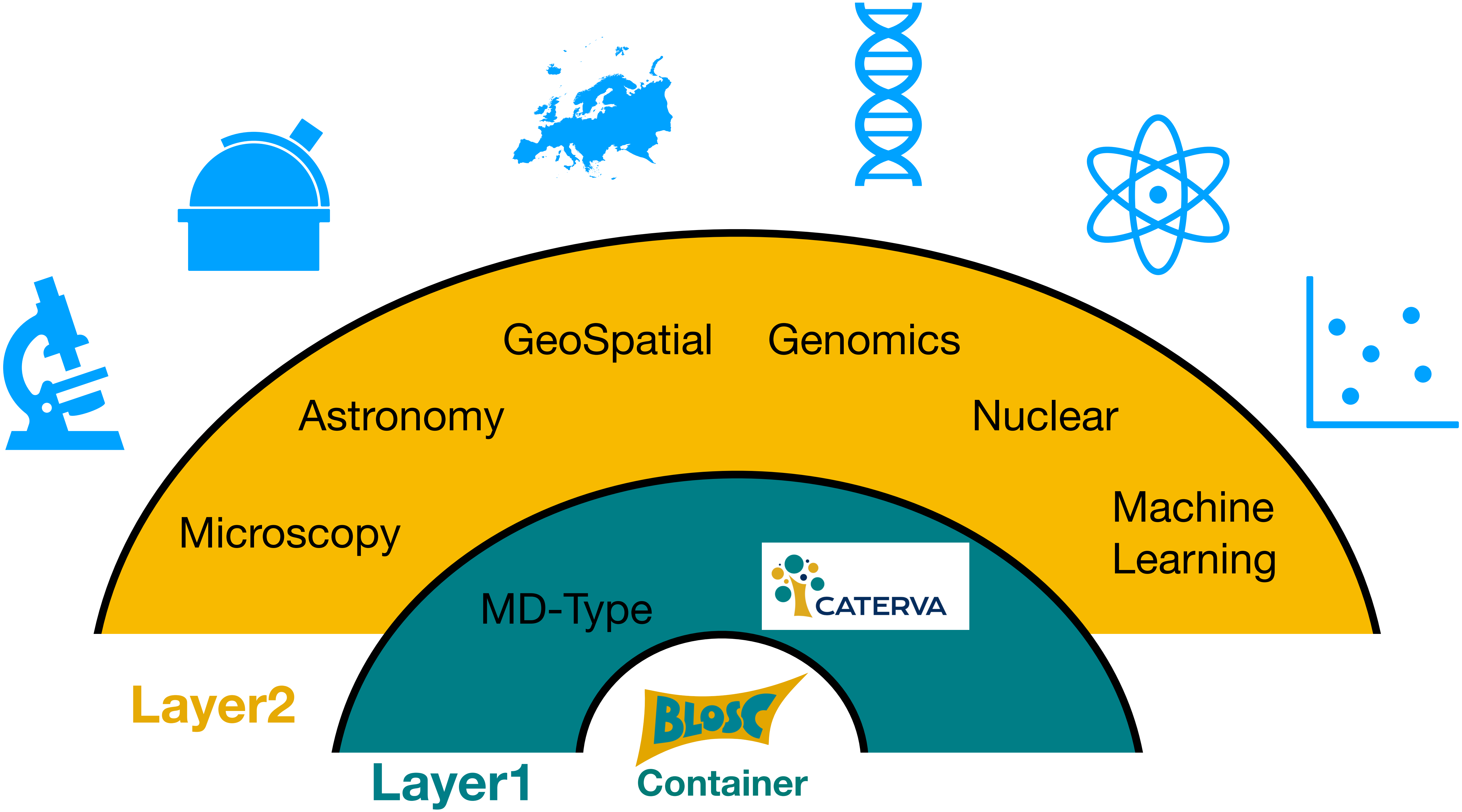
UserMeta

Fingerprint

Trailer

- Sequential
- In-memory / On-disk

MetaLayers in Blosc2



GeoSpatial

Genomics

Astronomy

Nuclear

Microscopy

Machine Learning

MD-Type

CATERVA

Layer2

Layer1

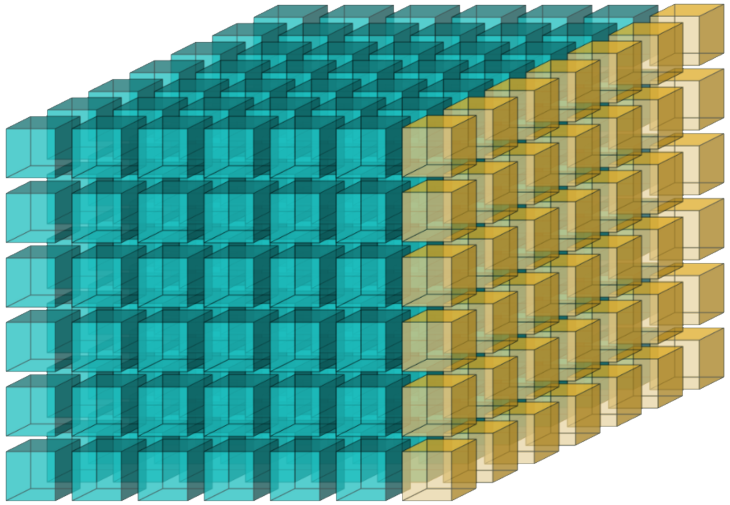
BLOSC

Container

Multiple layers to target different data aspects

Caterva: A New Multidimensional Container Based on Blosc2

- Thin layer on top of Blosc2 containers
- Open source C library
- cat4py: Python wrapper for Caterva



CATERVA

Main Contributors to Blosc2 / Caterva

- Alberto Sabater (documentation on Blosc2)
- Aleix Alcacer (Caterva)
- Jerome Kieffer (PowerPC/Altivec support in Blosc2)
- Francesc Alted (architecture and implementation)

What’s New in PyTables

- Release 3.5.x (March 2019)
  - Better support for native HDF5 files with padding in compound types
- Release 3.6.x (October 2019)
  - Full Python 3.8 support. Dropped 2.7 support.
  - Bugfix for HDF5 files/types with padding

Producing Wheels

- We are currently providing wheels because it is convenient for the users
- However, creating wheels has been driving us nuts in the last releases
  - The wheelbuilder (macpython/pytables-wheels) works fine for both Linux and MacOSX (thanks to Matthew Brett et al).
  - Windows.... sigh....

What’s Up With Windows Wheels?

- 3.6.1 failed twice: Windows/Conda broken and Python 3.8 broken due to building sdist with outdated Cython
- We just cannot currently rely on wheels that are automatically generated during a new build to actually work for Windows
- We have now created a CI testrepo (tomkooij/pytables-testpypi) to automatically test new wheels.

Do Library Maintainers Should Take Care of the Wheels?

- Creating binaries with complex dependencies as HDF5 is a challenging task (compiling the HDF5 library itself exceeds the < 1 hour limit in appveyor).
- I am not convinced that PyTables developers or maintainers should be responsible of doing that.

Release Managers During 2019

- Francesc Alted released PyTables 3.5.0 and 3.5.1.
- Tom Kooij released 3.5.2, 3.6.0 and 3.6.1 (he needed to create the wheels for Windows on his own laptop).

Support Us!

- If you use Blosc or PyTables, please support us by donating to any of the projects via NumFOCUS:

NUMFOCUS
OPEN CODE = BETTER SCIENCE

Thanks and Enjoy Data!