Simplified Columnar File Conversions with odapt

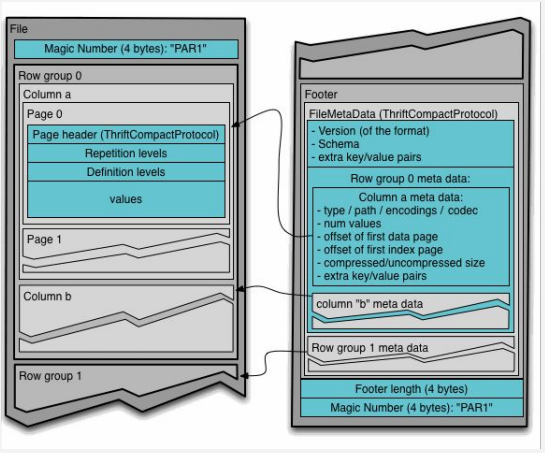
Zoë Bilodeau

Jim Pivarski

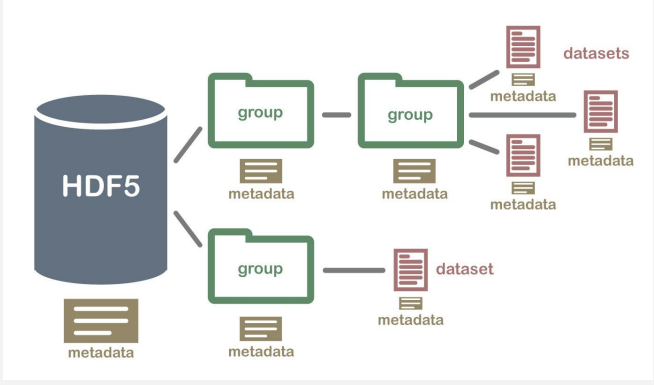
PRINCETON UNIVERSITY

Problems Addressed by odapt:

- Columnar file conversions require time and knowledge:
  - Multiple function calls, setup code
  - Multiple packages
  - File structures differ
  - Handling memory etc.



https://parquet.apache.org/docs/file-format/



https://www.neonscience.org/resources/learning-hub/tutorials/about-hdf5

= ?

Goals of odapt:

- Quick file conversions
  - Blocks of code -> single function call
  - Parameters for customization
  - Memory management
  - Address common problems

- User-oriented:
  - Choosing new functions and features
  - Proactive communication from developer side

Features and Design:

- Convert between ROOT, Parquet, Feather, and HDF5
  - ROOT to ROOT
  - Specific user requests; hadd-like capabilities

- Functions implemented using common tools
  - h5py
  - Uproot
  - Awkward Array
  - dask-awkward

- CLI

Examples:

To copy a ROOT file and remove branches:

```
>>> odapt.copy_root("destination.root", "file.root",
    drop_branches=["branch1", "branch2"], step_size=100)
```

To convert a ROOT file to a Parquet file:

```
>>> odapt.root_to_parquet("destination.root", "file.parquet")
```

Progress so Far:

- Histogram adding
- ROOT <-> ROOT
- Merge TTrees and add histograms
- Parquet -> ROOT (nearly done with ROOT -> Parquet)
- ~CLI

What's Next?

- Continue developing odapt
  - Finish core functionality
  - Continue implementing user requests
  - Get feedback and make changes!

- Broader scope of my work:
  - Work on Uproot
    - Dask-Awkward -> ROOT function
  - Act as go-between for users and Uproot (HSF trainings)

Thank you!

Questions or feedback for odapt? Find me here:

- Mattermost: CMS coffea-users channel
- Iris-HEP slack: dask-awkward and uproot-awkward channels
- https://github.com/zbilodea/odapt/issues