Python Workshop Series Session 7:
I/O with HDF5

Nick Featherstone

Research Computing

Slides: https://github.com/ResearchComputing/Python_Spring_2018

Outline

- Overview
- HDF5 File Creation/Reading
- Attributes
- Groups
- Subgroups

Useful HDF5 References

- General HDF5 (C++/Fortran):
  https://support.hdfgroup.org/HDF5/doc/index.html
- h5py (Python):
  http://www.h5py.org/
- h5 (R):
  https://cran.r-project.org/web/packages/hdf5r/index.html

Overview: Why HDF5

- Hierarchical Data Format (File structured as mini file system)
- Standardized data structure
- Data + Metadata
- Portable (your HDF5 files can be read in Fortran,C, & R too)
- Parallel (underpins parallel I/O layer in many HPC applications)
- *Relatively* easy to use (esp. in Python and R)

Important Note

- For today, it is best NOT to work in Jupyter.
- Code spanning multiple cells can cause file-related issues.
- Instead, run sample programs directly from the command prompt.
- e.g., python create_hdf5.py

Getting Started

- First, make sure you can import the h5py and numpy modules

```
import h5py
import numpy as np
```

- If this fails:
  - source activate idp
  - conda install h5py

h5py: File Creation

- First, create a file object:

```
filename = 'test.hdf5'
f = h5py.File(filename, "w")     w = 'write'
```

- create_hdf5.py
- File objects and their methods provide a high-level interface for interacting with a file (open, close, flush, etc.)
- Data can be added to the file by creating HDF5 datasets associated with the file.

h5py: Datasets

- Next, add a dataset to the file:

```
dname1="Integers"
ndata1= (100,)  note the comma!
dset1 = f.create_dataset(dname1,  ndata1, dtype='int32')
```

- Specify a name
- Specify dimensions (ndata1)
- ndata1 must be a tuple (you need the comma; otherwise it's cast as an 'int' and isn't iterable).
- Use numpy datatypes

h5py: Datasets

- Populate the dataset (works like a NumPy array):

```
dset1[:] = np.arange(1,101, dtype='int32')
```

- Datasets can be multidimensional

```
dname2='Reals'
ndata2=(2, 2)
dset2 = f.create_dataset(dname2, ndata2, dtype='float64')
dset2[0,:] = np.array( [2.1, 3.0 ], dtype='float64')
dset2[1,:] = np.array( [55.0, -73.01 ], dtype='float64')
```

Attributes

- Small, named pieces of data directly attached to group and dataset objects
- Basically a small dictionary with scalar or NumPy array values

```
dset1.attrs[ 'month' ]=7
dset1.attrs[ 'year' ]=2017
```

- Once we're finished describing our data, close the file

```
f.close()          close the file
```

h5py: File Inquiry

- We can examine a file's contents at the command line using h5dump (a tool that is part of HDF5)

h5dump –n test.hdf5	display table of contents
h5dump –B test.hdf5	display values
h5dump -h	display additional options

H5py: Reading Data

- When reading a file, we can access our datasets by treating the file object as a dictionary (read_hdf5.py):

```
import h5py
import numpy as np
f = h5py.File( 'test.hdf5' , 'r')    r = "read"
integers = f[ 'Integers' ]
reals = f[ 'Reals' ]
print(  integers[:]  )
f.close()
```

HDF5: File Structure

- HDF5 files organized around:
  - Groups:
    - folder-like containers that hold datasets and other groups
    - Think “directories”
    - Subgroups → “subdirectories”
  - Datasets:
    - array-like collections of data
- In Python:
  - Datasets behave like NumPy arrays
  - Groups behave like dictionaries

HDF5: File Structure

- The File object:
  - Represents the “root” group of the file:
  - Analogous to “/” on a POSIX file system
  - Each group has a name attribute

```
import h5py
f = h5py.File( ‘test.hdf5’, ‘r’)
print( f.name )   displays ‘/’
f.close()
```

H5Py: Groups

- Creating groups (create_groups.py)
  - Groups can be created within the “root” group of the file:
  - Use create_group method

```
import h5py
f = h5py.File( ‘new.hdf5’ , ‘w’ )
ff1 = f.create_group( “folder1” )
ff2 = f.create_group( “folder2” )
f.close()
```

```
h5dump –n new.hdf5        display table of contents
```

H5Py: Groups

- We can open a file and add data to existing groups (modify_groups.py):

```
import h5py
f = h5py.File( 'new.hdf5', 'r+')        r+ = "read/write"
ff1 = f[ "folder1" ]
npts=10
ndata=( npts, )
dname='data_range_1'
dset1=ff1.create_dataset(dname,ndata,dtype='int32')
dset1[:]=np.arange(1, npts+1, dtype='int32')
f.close()
```

```
h5dump –n new.hdf5        display table of contents
```

H5Py: Subgroups

- We can create groups within groups (and add data; subgroups.py):

```
import h5py
f = h5py.File( 'new.hdf5', 'r+')        r+ = "read/write"
ff1 = f[ "folder1" ]
fsub = ff1.create_group( "subfolder1" )
ndata=(10, )
dname2='data_set_2'
dset2=fsub.create_dataset(dname, ndata, dtype='int32' )
dset2[:]=np.arange(21, 31, dtype='int32' )
f.close()
```

```
h5dump –n new.hdf5        display table of contents
```

H5Py: Subgroups

- We can create a full path of subgroups at once:

```
import h5py
f = h5py.File( 'new.hdf5', 'r+')        r+ = "read/write"
ff2 = f[ "folder2" ]
sfstring='subfolder2/subfolder3/subfolder4'
fsub2 = ff2.create_group( sfstring )
f.close()
```

```
h5dump –n new.hdf5        display table of contents
```