The Journal of Open Source Software

h5preserve: Thin wrapper around h5py, inspired by camel

James Tocknell[1]

1 Macquarie University, Sydney, Australia

DOI: 10.21105/joss.00581

Software

- Review
- Repository
- Archive

Submitted: 07 February 2018
Published: 20 February 2018

Introduction

h5preserve is a wrapper around h5py (Collette 2013) and hdf5 (The HDF Group 1997), providing easier reading and writing of native and user-created python types to hdf5 files, transparently dealing with multiple versions of the data. The two python libraries most similar to h5preserve are pickle and Camel (Eevee (Lexy Munroe) 2017).

Origin of h5preserve

h5preserve was created by the author after two previous attempts to write libraries which would allow easy addition of new and changed data structures produced by some modelling and visualisation code he wrote for his thesis (both of which failed due to evolving complexity). Inspired by Camel (Eevee (Lexy Munroe) 2017), h5preserve focuses on providing an simple interface to read and write different and evolving versions of data structures produced by the modelling and analysis of in-memory datasets.

Why use h5preserve

The security flaws in pickle are well known (these are explicitly called out at the very top of the pickle documentation in a big red warning) and design flaws in pickle have been brought up in both PyCon talks (Alex Gaynor 2014) and blog posts (Lexy Munroe 2015) by well known members of the Python community, which inspired the creation of Camel (Eevee (Lexy Munroe) 2017). camel uses YAML as the output format, which is ideal for textual data, but not for numerical data or multidimensional arrays. h5preserve takes the design elements of camel, and ports them to hdf5, making it easy to use the design philosophy of camel, with the multidimensional array support of hdf5.

Being built on hdf5, and with scientific and numerical use in mind, h5preserve has support for complex numerical data and multidimensional arrays via numpy (Van Der Walt, Colbert, and Varoquaux 2011), which other serialisation formats (e.g. CSV, JSON or YAML) may not represent as effectively. h5preserve makes it easy to split out the interaction with hdf5 files from the main logic of your code, allowing for the creation of data classes via libraries such as namedtuples without the need to reimplement or modify existing libraries.

Why not to use h5preserve

As h5preserve is designed to hide the underlying hdf5 file, large files where memory usage is a concern do not work well with h5preserve. In this case, h5preserve provides easy

access to the underlying h5py objects, or you may want to look at using pytables (Alted, Vilata, and others 2002), which provides a more database-like interface to hdf5 files.

References

Alex Gaynor. 2014. "Pickles Are for Delis, Not Software." https://www.youtube.com/watch?v=7KnfGDajDQw.

Alted, Francesc, Ivan Vilata, and others. 2002. "PyTables: Hierarchical Datasets in Python." http://www.pytables.org/.

Collette, Andrew. 2013. *Python and HDF5: Unlocking Scientific Data*. O'Reilly Media, Inc.

Eevee (Lexy Munroe). 2017. "Camel: Python Serialization for Adults." https://github.com/eevee/camel.

Lexy Munroe. 2015. "Don't Use Pickle — Use Camel / Fuzzy Notepad." Blog. *Fuzzy Notebook*. https://eev.ee/release/2015/10/15/dont-use-pickle-use-camel/.

The HDF Group. 1997. "Hierarchical Data Format, Version 5." http://www.hdfgroup.org/HDF5.

Van Der Walt, Stefan, S Chris Colbert, and Gael Varoquaux. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science & Engineering* 13 (2):22–30. https://doi.org/10.1109/MCSE.2011.37.