

# **h5**

## An Object Oriented Interface to HDF5

---

Mario Annau

21 May 2016

# Data Challenges

- Store large amounts of data, e.g. tick data
- Retrieve subsets of data into memory
- Programming language independent
- Minimal setup requirements, single client
- High performance

# HDF5 Hierarchical Data Format

- Binary, language independent file format
- Developed by NCSA<sup>1</sup> and the tri-labs, supported by NASA<sup>2</sup>
- First released in 1998
- HDF5 building blocks are *Groups*, *Datasets* and *Attributes*
- *Datasets* are stored in *Groups*, similar to folders in file system
- Addition of metadata using *Attributes*

---

<sup>1</sup>National Center for Supercomputing Applications

<sup>2</sup>National Aeronautics and Space Administration

## Existing HDF5 Packages

- History of packages: **rhdf5** (**h5r**, **hdf5**, **activeH5**)
- Packages suffer in terms of usability and speed
- No active packages on CRAN

## h5: An Object Oriented Interface to HDF5

- Intuitive to use through (subset) operators
- **Rcpp** to interface library
- 200+ test cases with a coverage of more than 80%
- Available on CRAN and Github for all major platforms

## First Steps

```
f <- h5file('test.h5')
g1 <- f['group1']
g1['mat'] <- matrix(1:9, nrow = 3)
g1['mat2'] <- matrix(11:19, nrow = 3)
h5attr(g1, 'attr1') <- 'This is Group 1'
f['group2/mat3'] <- matrix(21:29, nrow = 3)
sapply(list.datasets(f), function(x) f[x][, 1])
```

```
##           /group1/mat /group1/mat2 /group2/mat3
## [1,]                1             11           21
## [2,]                2             12           22
## [3,]                3             13           23
```

```
h5close(f)
```

## Extract Time Series Created from Pandas

### Python:

```
from pandas import date_range, DataFrame
from numpy import random
t = date_range('2010-01-01', '2016-01-01', freq='D').date
randmat = random.standard_normal((len(t), 3))
df = DataFrame(randmat, index=t)
df.to_hdf('ex-pandas.h5', 'testset')
```

### R:

```
f <- h5file('ex-pandas.h5', 'r')
dates <- as.Date(f['testset/axis1'][1:3],
  origin = '0001-01-01') - 1
zoo(f['testset/block0_values'][1:3, ], order.by=dates)
```

## Conclusion and Outlook

- **h5** directly maps R data types and objects to HDF5
- Facilitates data exchange between languages like R, Python and Matlab.
- Available on CRAN for all major platforms
- Support for *data.frames* planned
- Project is open for contributors, pull requests!

<https://github.com/mannau/h5>