Hierarchical data format for eddy-covariance data

David Durden[1], Cove Sturtevant[1], Natchaya Pingintha-Durden[1], Hongyan Luo[1], Andy Fox[2,3] and Stefan Metzger[1]

[1] National Ecological Observatory Network, Boulder, Colorado, USA
[2] Arizona State University, Tucson, Arizona, USA
[3] National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA

Background

Large data collecting networks have led to better understanding of environmental variation through an increase in available information. However, analyzing, curating, and archiving the observations with associated metadata for large datasets can be complicated. Tower networks, such as ICOS, Ameriflux, TERN, and NEON, illustrate the growing size of datasets from dispersed measurement sites. Eddy-covariance data from across the NEON network are expected to amount to 100 Gigabytes per day. The large throughputs of data between the database, the processing environment, and the data portal require an efficient file format.

The capability to process large data sets is reliant upon:

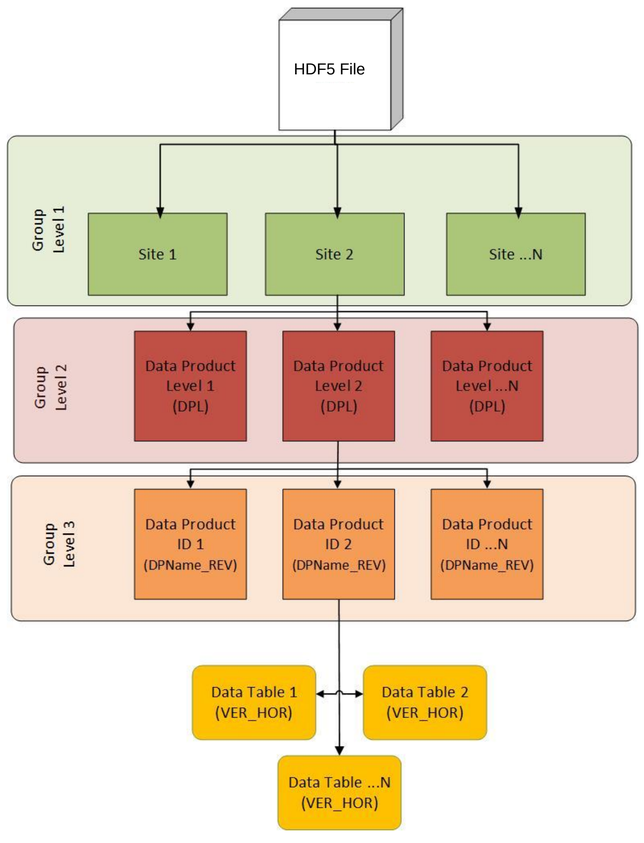
- efficient input and output of data
- data compressibility to reduce compute resource loads
- the ability to easily package and access metadata.

NEON HDF5 File Structure

The Hierarchical Data Format (HDF5) is a file format that can meet these needs. The "directory-like" structure of the HDF5 files provides intuitive navigation of the data based on the NEON data product naming convention.

NEON.DOM.SITE.DPL.PRNUM.REV.TERMS.HOR.VER.TMI

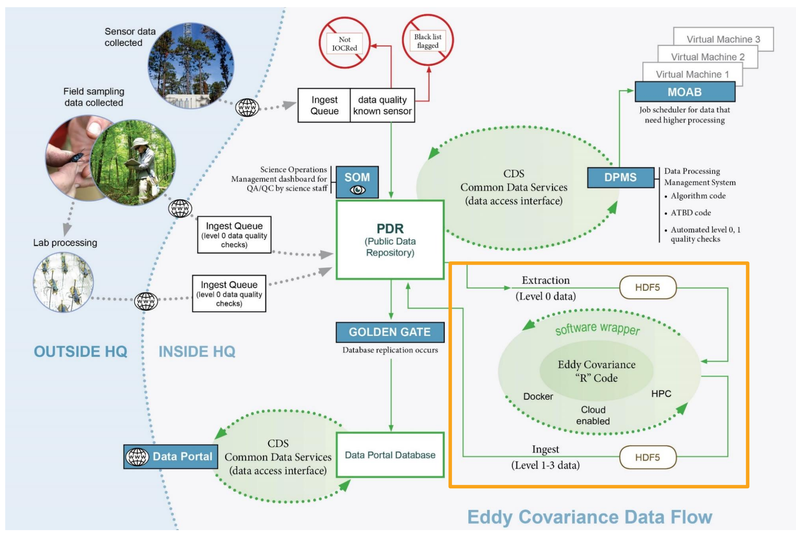
WHERE:
NEON=NEON
DOM=DOMAIN, e.g. D10
SITE=SITE, e.g. STER
DPL=DATA PRODUCT LEVEL, e.g. DP1
PRNUM = PRODUCT NUMBER =>5 digit number. Set in data products catalog.
TIS = 00000-09999
REV = REVISION, e.g 001.
TERMS=From NEON's controlled list of terms. Index is unique across products.
HOR = HORIZONTAL INDEX. Semi-controlled; AIS and TIS use different rules.
Examples: Tower=000, Hut = 700, DFIR=900.
VER = VERTICAL INDEX. Semi-controlled; AIS and TIS use different rules.
Examples: Ground level=000, second tower level=020.
TMI=TEMPORAL INDEX. Examples: 001=1 minute, 030=30 minute, 999=irregular intervals.



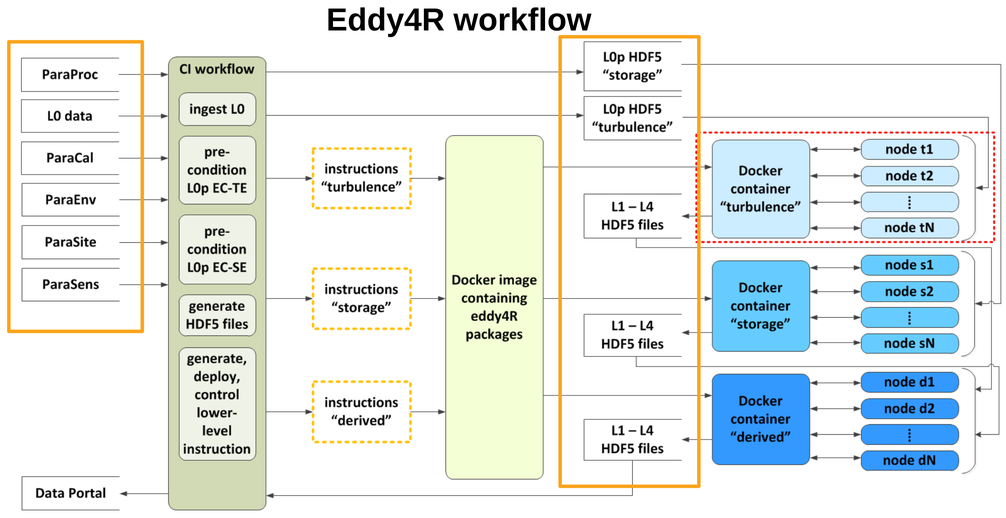
Contact Information: ddurden@BattelleEcology.org

Eddy-covariance flux data flow

NEON data processing schematic



Eddy4R workflow

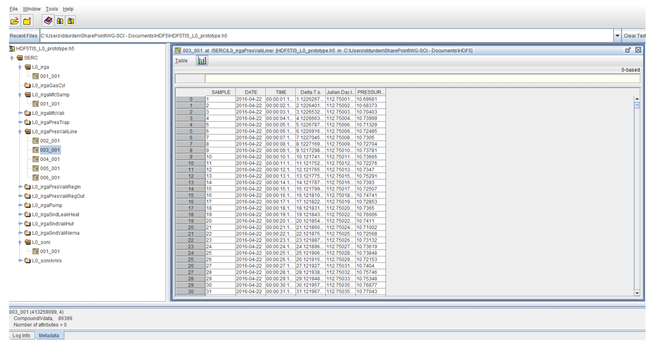


HDF5 files will be used for input/output to the eddy-covariance processing scheme. Metadata and data are packaged together with data in data tables and metadata as attributes.

NEON HDF5 Performance

The NEON standard HDF5 file structure and metadata attributes allow users to explore larger data sets in an intuitive "directory-like" structure. Additionally, HDF5 allows multiple NEON data products to be packaged into a single file and expands possibilities for data provenance where various levels of data products can be packaged together.

- Timeframe:
  - 4/22/2016 -5/03/2016
- File size for 1 day (4/22/2016):
  - Compressed = 398 MB
  - Uncompressed = 1.84 GB
  - Data Compression Ratio ~ 4.5:1
- Metadata: Units and variable names



Test datasets approximated 1 day of calibrated raw (L0p) IRGA data

- "**compound**": single dataset with each row having many numeric float values and a single string value
- "**simple**": one dataset with each row having many numeric float values, second dataset with each row having a single string value

Results for **COMPOUND** dataset:

	Compressed	Non-compressed
Read	45 secs	4.25 secs
Write	621 secs	11.25 secs
Size	78 MB	266 MB

Results for **SIMPLE** dataset

	Compressed	Non-compressed
Read	1.45 secs	0.75 secs
Write	21.45 secs	4 secs
Size	21 MB	266 MB

NEON is a project sponsored by the National Science Foundation and operated under cooperative agreement by Battelle.

Battelle
The Business of Innovation