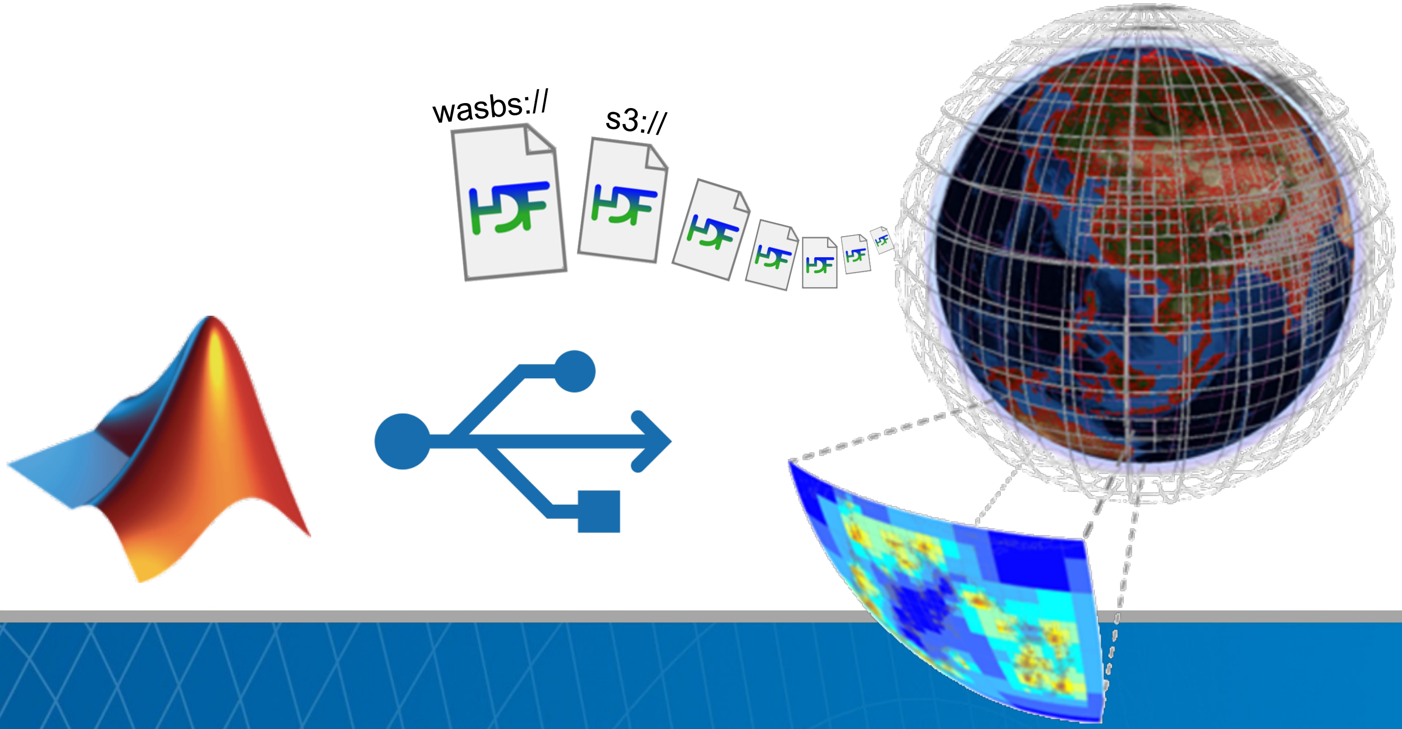
MATLAB Meets HDF5 in the Cloud



Ellen Johnson
Senior Software Engineer, MathWorks
HDF5 User Group 2021
October 14, 2021

Agenda

MathWorks®

Accelerating the pace of engineering and science

- MATLAB scientific data overview
- HDF5 in MATLAB
- What we’ve been up to
- Cloud workflows
- Demo
- Performance and compatibility
- Future work
- Wrap-up and Q&A

Scientific Data in MATLAB

Scientific data formats

- HDF5, HDF4, HDF-EOS2
- NetCDF (with OPeNDAP)
- FITS, CDF, BIL, BIP, BSQ

Image file formats

- TIFF, JPEG, PNG, JPEG2000, HDR, *and more*

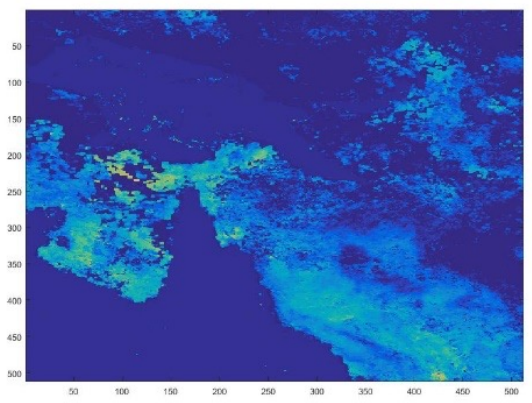
Vector data file formats

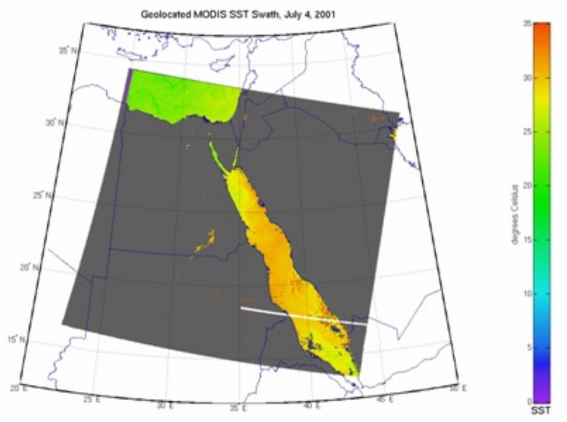
- ESRI Shapefiles, KML, GPS *and more*

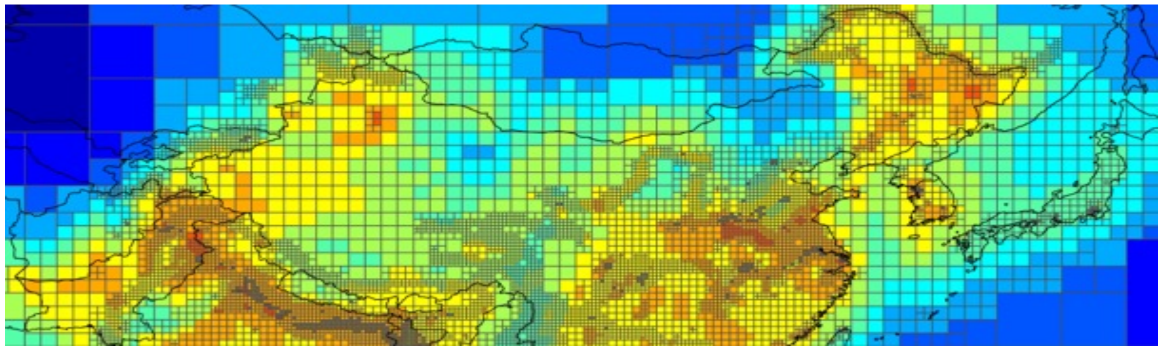
Raster data file formats

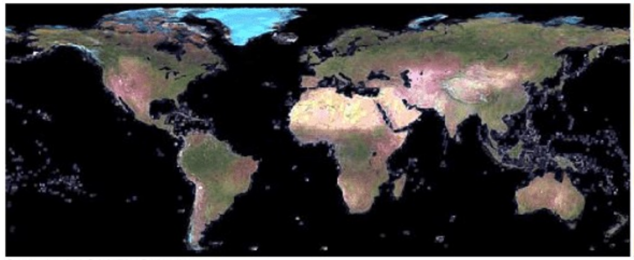
- GeoTIFF, NITF, USGS and SDTS DEM, NIMA DTED, *and more*

Web Map Service (WMS)

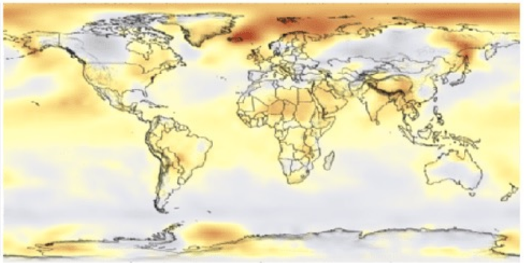








Courtesy NASA/JPL-Caltech

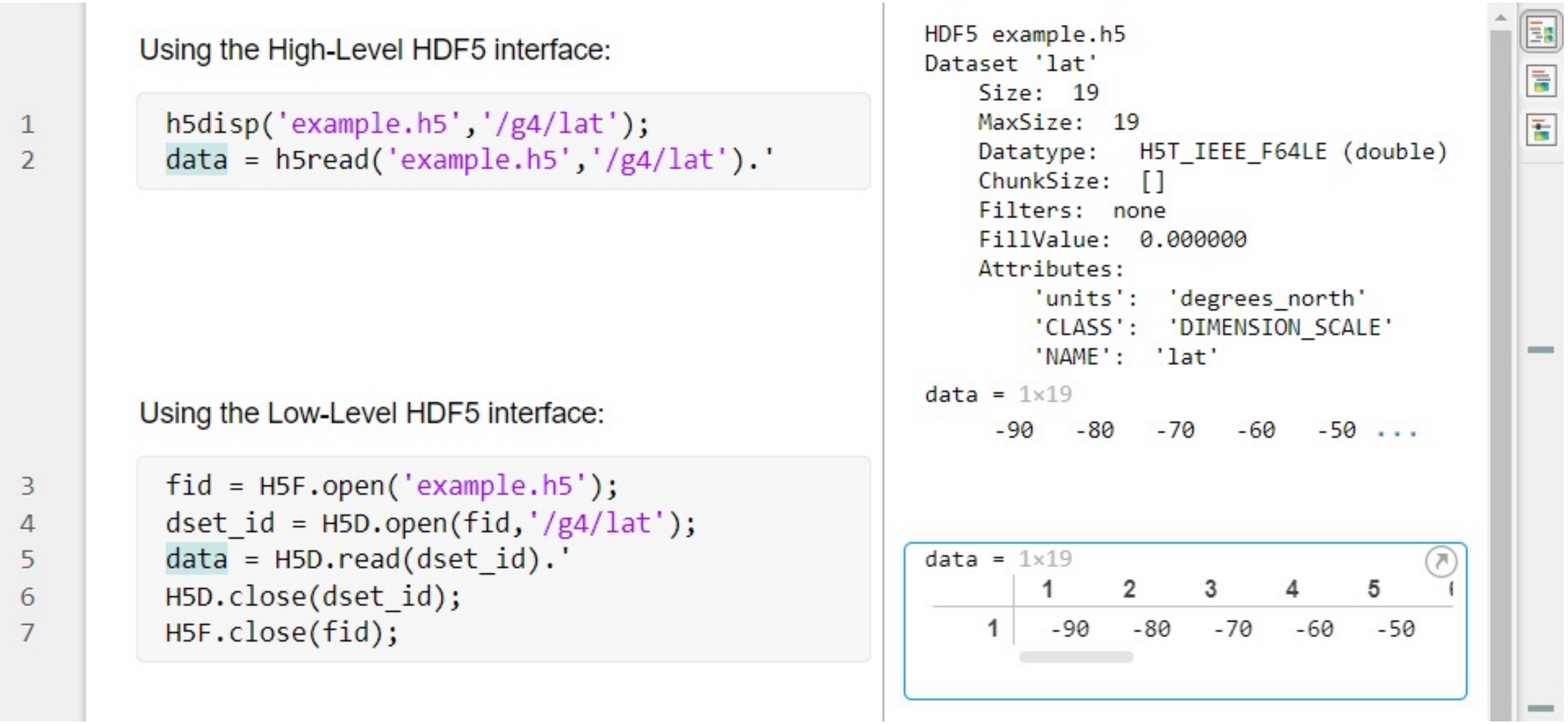


Courtesy NASA/Goddard Space Flight Center Scientific Visualization Studio

HDF5 in MATLAB

Two HDF5 interfaces

- High-level (HL) : Ease-of-use, less control
- Low-level (LL) : Wraps HDF5 C library, more control



Using the High-Level HDF5 interface:

```
h5disp('example.h5','/g4/lat');
data = h5read('example.h5','/g4/lat').'
```

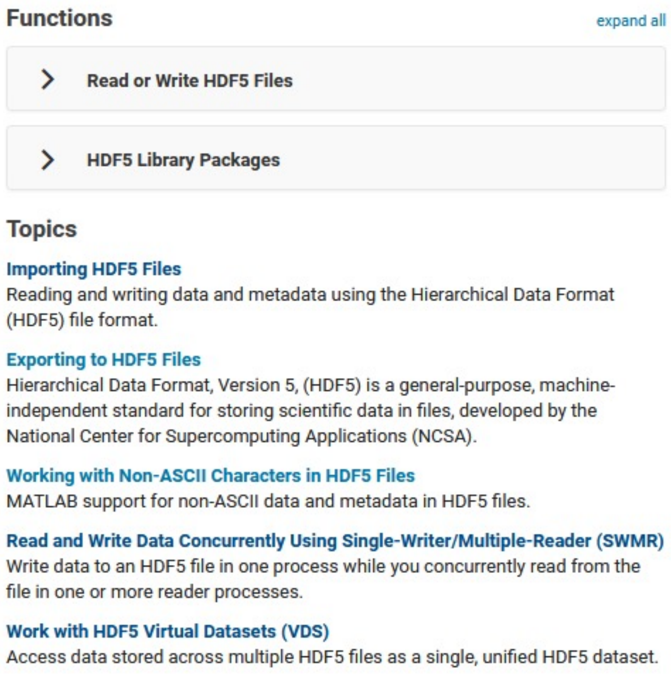
Using the Low-Level HDF5 interface:

```
fid = H5F.open('example.h5');
dset_id = H5D.open(fid,'/g4/lat');
data = H5D.read(dset_id).'
H5D.close(dset_id);
H5F.close(fid);
```

```
HDF5 example.h5
Dataset 'lat'
    Size:  19
    MaxSize:  19
    Datatype:   H5T_IEEE_F64LE (double)
    ChunkSize:  []
    Filters:  none
    FillValue:  0.000000
    Attributes:
        'units':  'degrees_north'
        'CLASS':  'DIMENSION_SCALE'
        'NAME':  'lat'
data = 1x19
     -90    -80    -70    -60    -50 ...
```

```
data = 1x19
           1      2      3      4      5
     1   -90    -80    -70    -60    -50
```

HDF5 in MATLAB



Functions

expand all

> Read or Write HDF5 Files

> HDF5 Library Packages

Topics

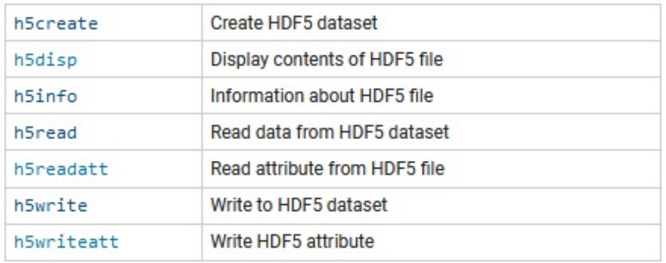
Importing HDF5 Files
Reading and writing data and metadata using the Hierarchical Data Format (HDF5) file format.

Exporting to HDF5 Files
Hierarchical Data Format, Version 5, (HDF5) is a general-purpose, machine-independent standard for storing scientific data in files, developed by the National Center for Supercomputing Applications (NCSA).

Working with Non-ASCII Characters in HDF5 Files
MATLAB support for non-ASCII data and metadata in HDF5 files.

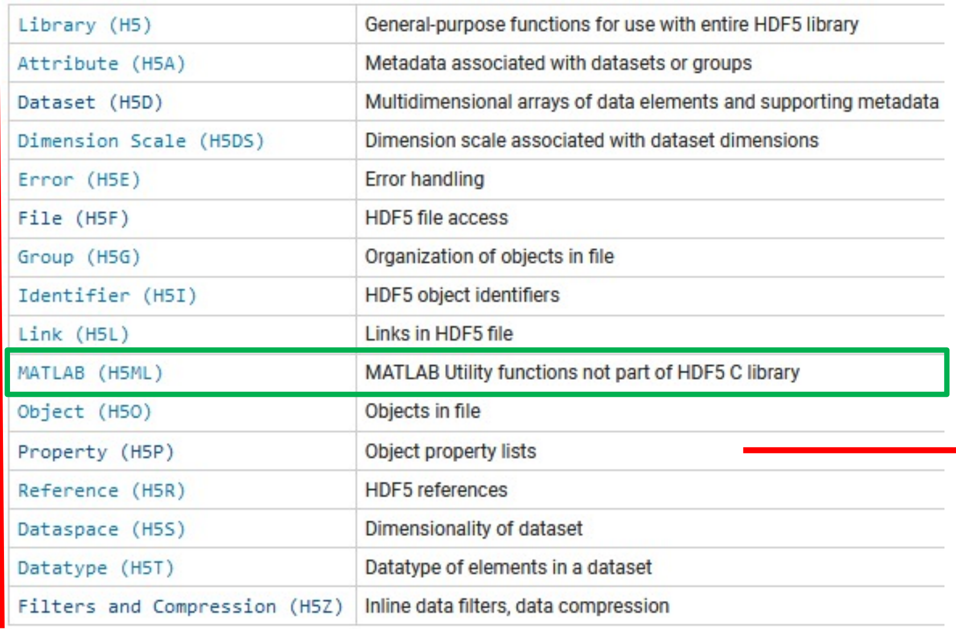
Read and Write Data Concurrently Using Single-Writer/Multiple-Reader (SWMR)
Write data to an HDF5 file in one process while you concurrently read from the file in one or more reader processes.

Work with HDF5 Virtual Datasets (VDS)
Access data stored across multiple HDF5 files as a single, unified HDF5 dataset.



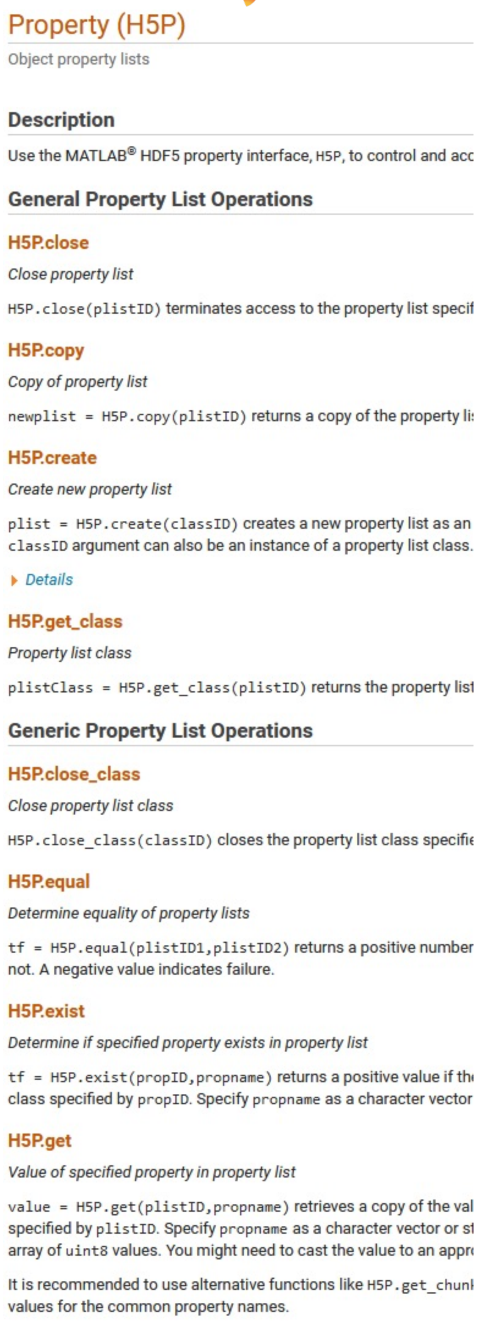
`h5create`	Create HDF5 dataset
`h5disp`	Display contents of HDF5 file
`h5info`	Information about HDF5 file
`h5read`	Read data from HDF5 dataset
`h5readatt`	Read attribute from HDF5 file
`h5write`	Write to HDF5 dataset
`h5writeatt`	Write HDF5 attribute

7 high-level functions



`Library (H5)`	General-purpose functions for use with entire HDF5 library
`Attribute (H5A)`	Metadata associated with datasets or groups
`Dataset (H5D)`	Multidimensional arrays of data elements and supporting metadata
`Dimension Scale (H5DS)`	Dimension scale associated with dataset dimensions
`Error (H5E)`	Error handling
`File (H5F)`	HDF5 file access
`Group (H5G)`	Organization of objects in file
`Identifier (H5I)`	HDF5 object identifiers
`Link (H5L)`	Links in HDF5 file
`MATLAB (H5ML)`	MATLAB Utility functions not part of HDF5 C library
`Object (H5O)`	Objects in file
`Property (H5P)`	Object property lists
`Reference (H5R)`	HDF5 references
`Dataspace (H5S)`	Dimensionality of dataset
`Datatype (H5T)`	Datatype of elements in a dataset
`Filters and Compression (H5Z)`	Inline data filters, data compression

~330 low-level functions



Property (H5P)

Object property lists

Description

Use the MATLAB® HDF5 property interface, `H5P`, to control and acc

General Property List Operations

H5P.close

Close property list

`H5P.close(plistID)` terminates access to the property list specif

H5P.copy

Copy of property list

`newplist = H5P.copy(plistID)` returns a copy of the property li

H5P.create

Create new property list

`plist = H5P.create(classID)` creates a new property list as an `classID` argument can also be an instance of a property list class.

▸ Details

H5P.get_class

Property list class

`plistClass = H5P.get_class(plistID)` returns the property list

Generic Property List Operations

H5P.close_class

Close property list class

`H5P.close_class(classID)` closes the property list class specifie

H5P.equal

Determine equality of property lists

`tf = H5P.equal(plistID1,plistID2)` returns a positive number not. A negative value indicates failure.

H5P.exist

Determine if specified property exists in property list

`tf = H5P.exist(propID,propname)` returns a positive value if th class specified by `propID`. Specify `propname` as a character vector

H5P.get

Value of specified property in property list

`value = H5P.get(plistID,propname)` retrieves a copy of the val specified by `plistID`. Specify `propname` as a character vector or st array of `uint8` values. You might need to cast the value to an appr

It is recommended to use alternative functions like `H5P.get_chunl` values for the common property names.

What We’ve Been Up To

R2015a

1.8.12 upgrade

Reading datasets with **D**ynamically **L**oaded **F**ilters

1.10.2 attempted upgrade…*oops, performance regressions* ☹

Meanwhile…while sorting that out…

R2020b

HDF5 Interface: Cloud-enabled

- S3 and Azure: Read/Write
- Hadoop: Read-only
- Enabled for all HL and LL functions

Amazon S3

Microsoft Azure Blob Storage

HDFS

R2021a

MAT-file v7.3 save/load: Cloud-enabled

1.10.6 attempted upgrade…*still regressions, but devised a solution* ☺

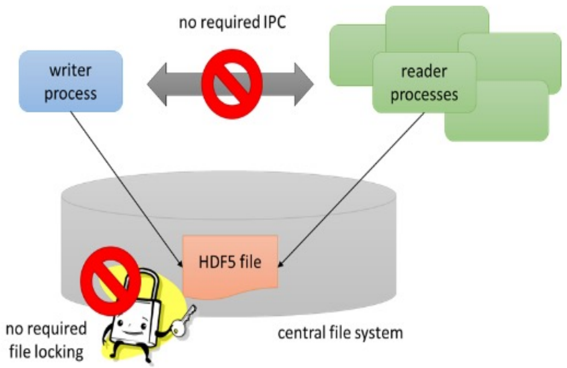
What’s New in R2021b

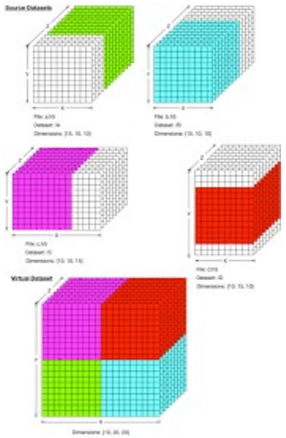
R2021b

MATLAB now on HDF5 1.10.7

New low-level functions:

```
SWMR       Fine Tuning MDC
VDS        Partial Edge Chunk
```





Shipping binaries for both 1.10.7 and 1.8.12 (Interim solution)

Or: How I Learned to Stop Worrying and Love GNU Export Maps

- 1.10.7 for MATLAB HDF5 interface
- 1.8.12 for MAT-file v.7.3 to avoid 1.10 regressions
- Consulting with THG and MathWorks teams on solution

Goal: Ship one version and stay current with HDF5 releases

Functional Details

New functions added to LL interface

- Added ~30 new functions across the 16 APIs
- Provides fine-grained control of **SWMR, VDS, Partial Edge Chunk, Metadata Cache**

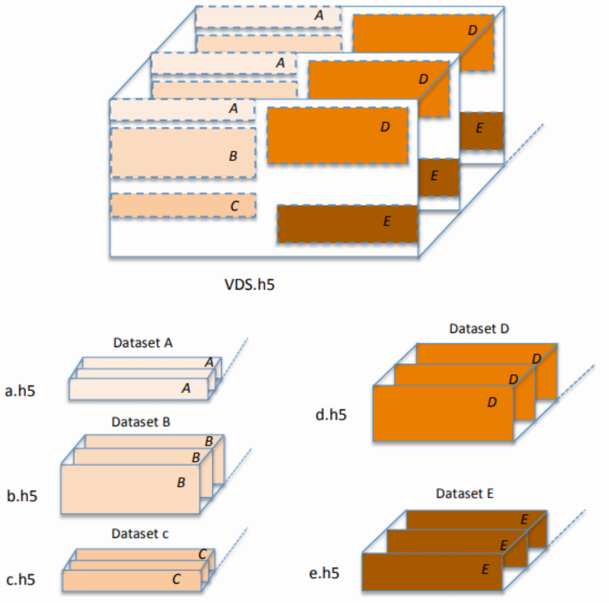
Modified existing functions to work with 1.10.7

- Including H5P.set_libver_bounds (for new high/low values)

Create and access Virtual Datasets

whether stored locally or cloud

- S3, Azure, Hadoop



New Functions Mapped to HDF5 Features

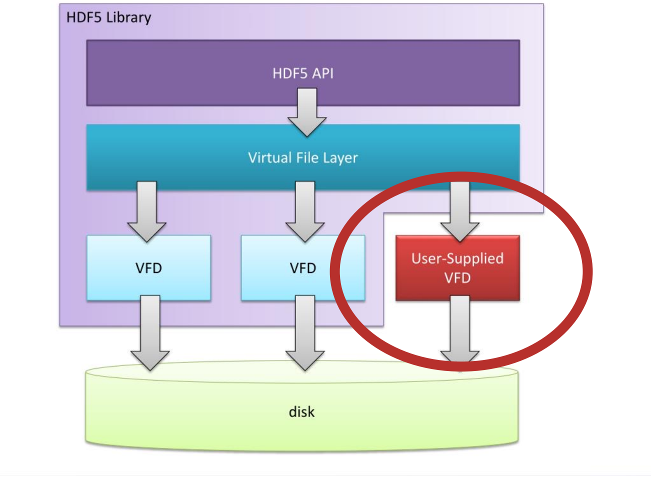
HDF5 Feature	MATLAB Function
SWMR	H5F.start_swmr_write H5O.disable_mdc_flushes H5O.enable_mdc_flushes H5O.are_mdc_flushes_disabled
VDS	H5P.set_virtual H5P.get_virtual_count H5P.get_virtual_vspace H5P.get_virtual_srcspace H5P.get_virtual_dsetname H5P.get_virtual_filename H5P.set_virtual_printf_gap H5P.gset_virtual_printf_gap H5P.set_virtual_view H5P.get_virtual_view H5S.is_regular_hyperslab H5S.get_regular_hyperslab
Fine Tuning the MDC	H5F.get_metadata_read_retry_info H5P.get_metadata_read_attempts H5P.set_metadata_read_attempts H5F.get_intent H5D.flush H5D.refresh H5G.flush H5G.refresh H5O.flush H5O.refresh H5T.flush H5T.refresh
Partial Edge Chunk	H5P.get_chunk_opts H5P.set_chunk_opts

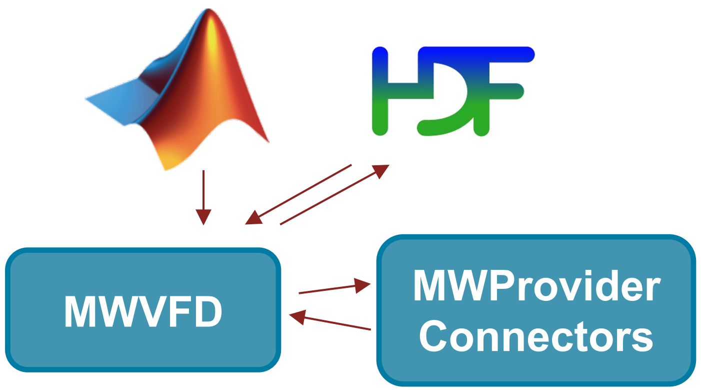
Cloud Data Access

Wrote **in-house VFD**

Use in-house provider architecture

Callbacks to HDF5 library





S3 and Azure: Read/Write

Hadoop: Read only

Amazon S3

HDFS

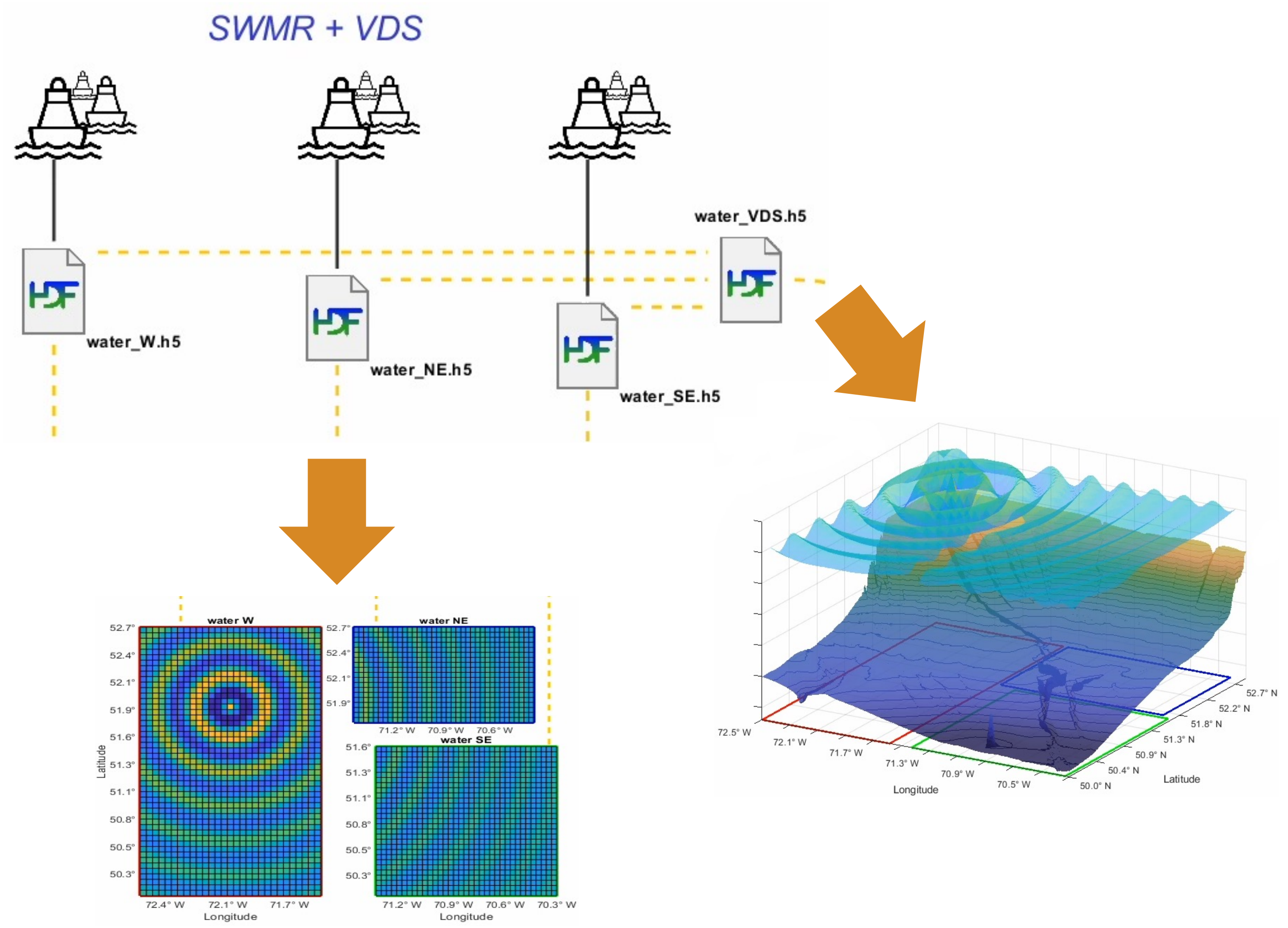
Microsoft Azure Blob Storage

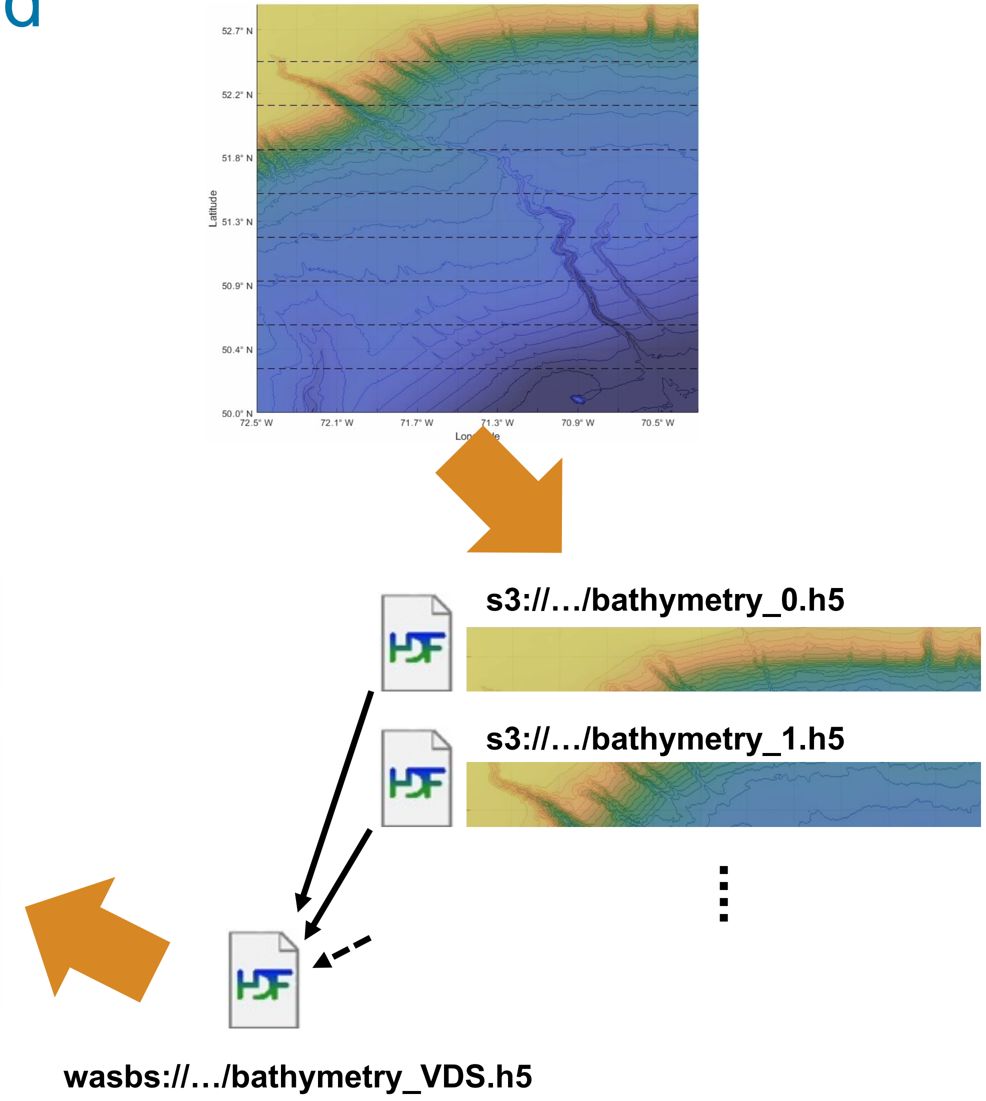
Support in **High and Low-level interfaces**

*including new **SWMR** and **VDS** functions!*

```
>> h5create('s3://h5test/myfile.h5','/ds1',[200 Inf],'ChunkSize',[20 20],'Deflate',9)
>> h5write('wasbs://h5test/myfile.h5','/ds1',rand(200,500),[1 1],[200 500])
>> h5read('hdfs://h5test/myfile.h5','/ds1')
```

Demo – MATLAB Meets HDF5 in the Cloud





GEBCO Gridded Bathymetry Data: https://www.gebco.net/data_and_products/gridded_bathymetry_data/
GEBCO Compilation Group (2020) GEBCO 2020 Grid (doi:10.5285/a29c5465-b138-234d-e053-6c86abc040b9)

Demo

Performance

Performance benchmarks with 1.10.7 vs 1.8.12

Improvements

- h5write, h5create, many low-level functions: minimal/moderate improvements

Regressions

- h5info: Substantial regressions with highly-nested groups with small datasets
- Working with THG to determine if same issue as MAT-file v7.3

Future work

- Optimize h5read, h5info
- More workflow-based performance tests

Compatibility in R2021b

Linux-only: Filter plugins with calls to core HDF5 library must be rebuilt with our shipping HDF5 1.10.7 shared library to avoid symbol collisions

- Option 1: Rebuild plugin with /matlab/bin/glnxa64/libhdf5.so.103.3.0
- Option 2: Build 1.10.7 using our GNU export map, then rebuild plugin with this binary.
- Documented on MATLAB Answers

Interim solution until we ship one version again

H5P.set_libver_bounds

- low/high = latest/latest will create incompatible files with earlier MATLAB versions

Future Work and Community Engagement

Highest priority

- Ship one HDF5 version
- Writing datasets using Dynamically Loaded Filters – *coming soon!*
- VDS and SWMR support in high-level interface
- Improved experience with filter plugins
- Performance

Community Engagement

- Continue working with THG (long-standing collaboration)
- Earth/Climate Data Providers – please host more HDF5 data on cloud!
- High-energy physics community – provide SWMR and VDS feedback, wish-lists

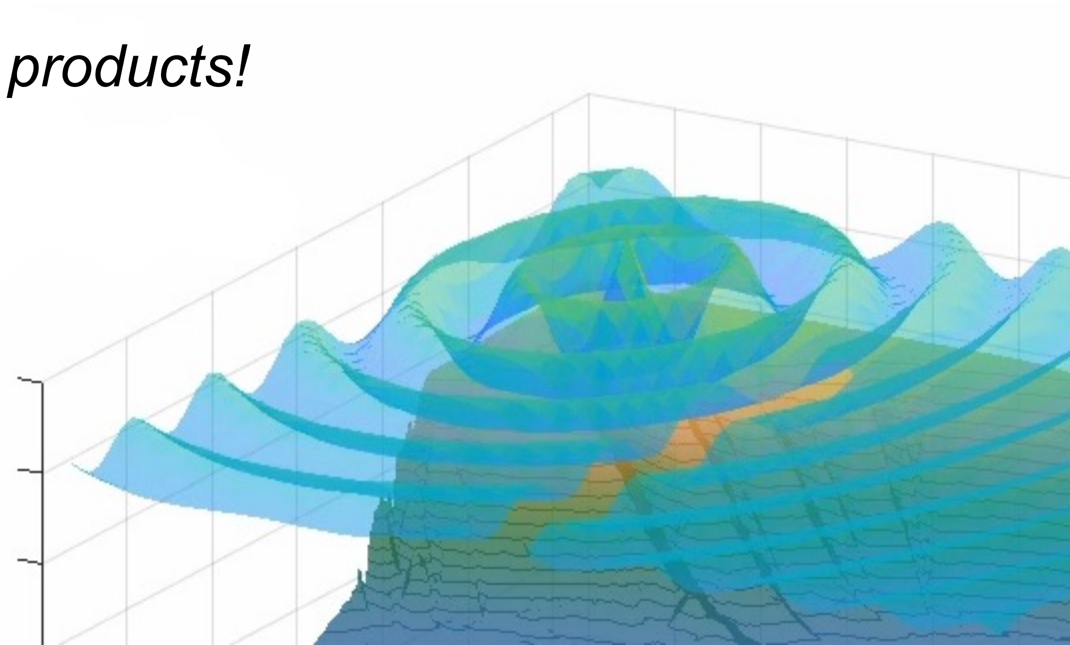
Wrap-up and Q&A

- MATLAB now current with latest HDF5 version on 1.10 branch
- New SWMR and VDS capabilities
- Linux Filter Plugin compatibility

We love hearing feedback – it helps us improve our products!

Reach out with any questions or wish-lists!

- ellenj@mathworks.com



Acknowledgements

- GEBCO Gridded Bathymetry Data: https://www.gebco.net/data_and_products/gridded_bathymetry_data/
  GEBCO Compilation Group (2020) GEBCO 2020 Grid (doi:10.5285/a29c5465-b138-234d-e053-6c86abc040b9)
- The HDF Group: www.hdfgroup.com
- HDF5 VDS RFC: https://portal.hdfgroup.org/display/HDF5/RFC+HDF5+Virtual+Dataset

Thank You!