DATA-DRIVEN COLLABORATION ENVIRONMENTS

INTEGRATING HUBZERO AND IRODS

Rajesh Kalyanam

Robert Campbell

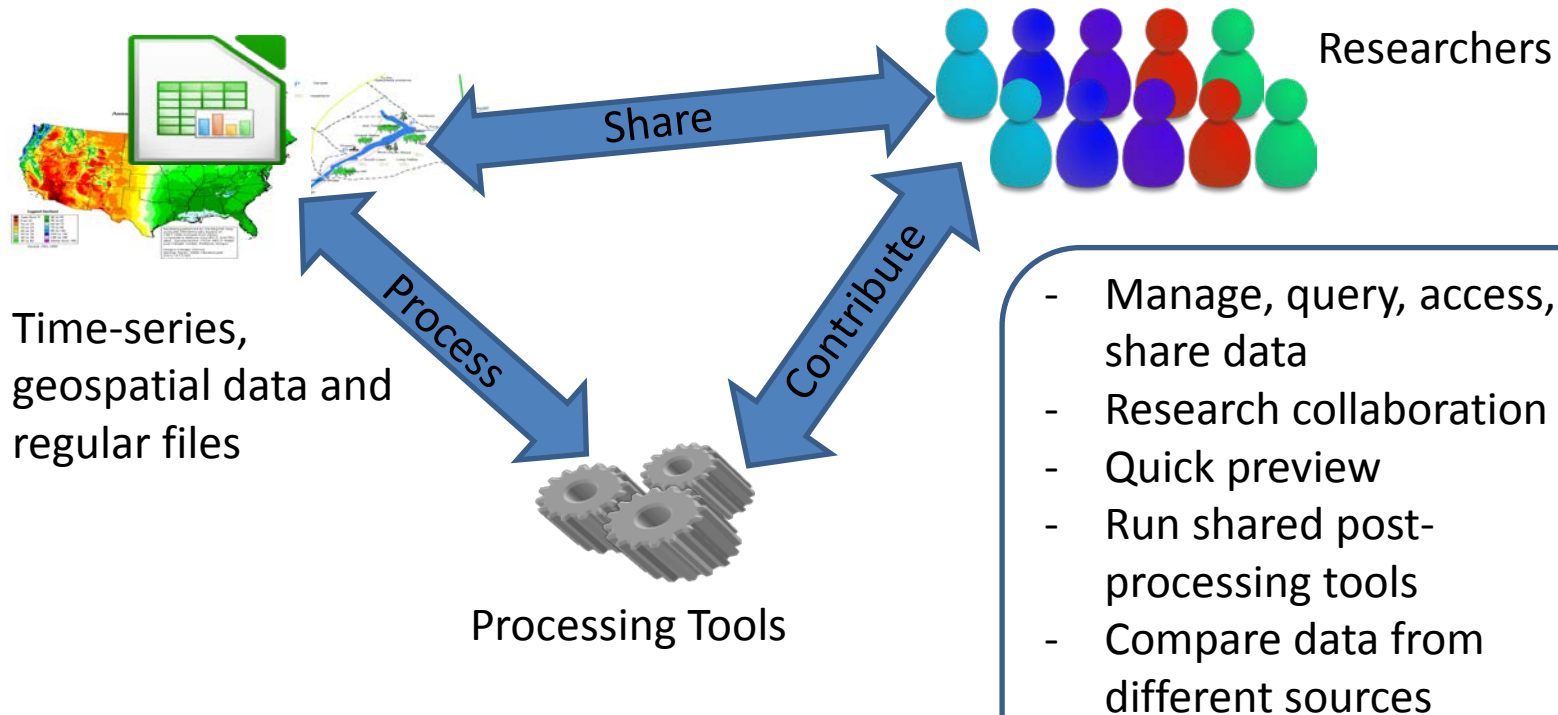**Lan Zhao**

Carol Song

RCAC, Purdue University

G·A·B·B·S
geospatial data analysis building blocks

PURDUE
UNIVERSITY

Researchers

Share

Process

Contribute

Time-series, geospatial data and regular files

Processing Tools

- Manage, query, access, share data
- Research collaboration
- Quick preview
- Run shared post-processing tools
- Compare data from different sources

PURDUE
UNIVERSITY

IDATA

- Hub component for scientific data management

- Supports familiar nested-folder file organization

- File sharing based on hub groups

- Support for metadata entry and metadata-based file query

- Preview, simple processing support for tabular, geospatial data

IDATA

IDATA FILE MANAGEMENT - IRODS

- Abstracts physical storage in iRODS
  - o Distributed data management system
  - o Developed at RENCI, DICE - UNC

- Supports easy storage expandability

- File metadata stored as attribute-value pairs

- iCAT enables file query based on metadata

- PHP client APIs enable easy integration with Hub component code

- Geospatial Data Analysis Building Blocks
  - Reusable blocks for geo-augmented tools, geospatial file management, visualization
- Enable interoperability between hub data management, tools, publications
- Enable tool input, output, discovery, provenance
  - display list of applicable tools for a file
  - support tool result reproducibility
- Support richer keyword (include geospatial) search
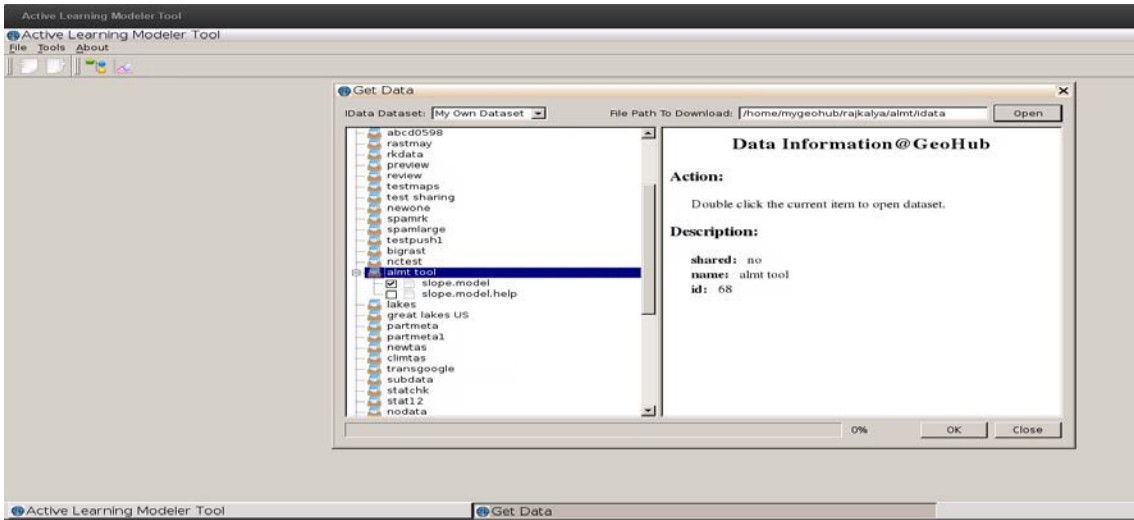
- Hub API controller exposes service interface to data management operations

- Tools invoke service methods for data listing, file upload, download and metadata

- Tool invocation URL arguments specify file input when launched from iData

METADATA CAPTURE – IRODS MICROSERVICES

```
#acPostProcForPut { ON($objPath like "\*txt") {writeLine("serverLog","File $objP
#acPostProcForPut { }
acPostProcForPut { on($objPath like "*.tif") { msiExtractGeoMetaTif; } }
acPostProcForPut { on($objPath like "*.nc") { msiExtractGeoMetaNc; } }
acPostProcForPut { on($objPath like "*.shx") { msiExtractGeoMetaShp; } }
acPostProcForPut { on($objPath like "*.shp") { msiExtractGeoMetaShp; } }
acPostProcForPut { on($objPath like "*.prj") { msiExtractGeoMetaShp; } }
acPostProcForPut { on($objPath like "*.dbf") { msiExtractGeoMetaShp; } }
acPostProcForCopy { }
acPostProcForFilePathReg { }
acPostProcForCreate { }
# acPostProcForOpen {writeLine("serverLog",$objPath); }
acPostProcForOpen { }
acPostProcForPhymv { }
acPostProcForRepl { }
# 9) acSetNumThreads - Rule to set the number of threads for a data transfer
# This rule supports condition based on $rescName so that different
```

Event-based file processing

```
/* Geo module microservices */
/* Service name                    # args             Service function */
{ "msiExtractGeoMetaTif",                0,                    (funcPtr) msiExtractGeoMetaTif }
{ "msiExtractGeoMetaNc",                 0,                    (funcPtr) msiExtractGeoMetaNc },
{ "msiExtractGeoMetaShp",                0,                    (funcPtr) msiExtractGeoMetaShp }
```

Microservice C-methods

```
        OGRSpatialReferenceH hSpatialRef, tSpatialRef;
        OGREnvelope hExtent;

        hDS = OGROpen( path, FALSE, NULL );

        if( hDS == NULL )
          {
            exit( 1 );
          }

        hLayer = OGR_DS_GetLayer( hDS, 0);

        OGR_L_ResetReading(hLayer);

        hSpatialRef = OGR_L_GetSpatialRef(hLayer);

        snprintf(metaname, sizeof metaname, "projection");
        snprintf(metavalue, sizeof metavalue, "%s", (char *)OSRGetAttrValue(hSpatialRef,"PROJECTION",0));
        addKeyVal (&kvdriver, metaname, metavalue);
```

Shapefile processing with GDAL C APIs

PURDUE
UNIVERSITY

INDEXED FILE SEARCH – APACHE SOLR

- DCMI metadata fields (+ geographic coverage)
- Captured by iRODS microservices
- Indexed into Solr

- All DCMI text fields added to keyword search index
- Geographic search uses Solr coverage field

PURDUE
UNIVERSITY

GOING FORWARD

TIGHTER IRODS – HUB INTEGRATION

- Storage option for hub Project Files
  - iDrop desktop client integration for large files
- Explore iRODS Fuse mounts
  - Can simplify file access, tool development
  - Will enable "local" file input/output in tool session (no service calls)
- Faceted search for hub files
  - iRODS microservice-based metadata capture
  - Import into Solr index and hooks for Solr faceted search
- Post-process files for compiled preview

GOING FORWARD

iRODS FEDERATION

- Federation enables file sharing across remote iRODS zones accessible as local collections

- Distinct hubs can support file sharing, collaboration across projects by federating iRODS storage

- Can enable remote file access in local hub tools
  - Potentially support tool workflows across hubs

GOING FORWARD

ENABLING RICHER TOOL WORKFLOWS

- Richer tool input filtering against iRODS-extracted metadata
  - For e.g. restricted to geographic region, require specific spreadsheet structure, images with 3 bands
- Tool output post-processing enabling subsequent tool input
  - For e.g. re-project geospatial data to standard projections
- Offline remote data download
- Implement data access protocols (OpenDAP) on top of iRODS for data subsetting

ACKNOWLEDGEMENTS

PURDUE
UNIVERSITY