

18 A 18

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

CI-BER tutorial

Richard Marciano Chien-Yi Hou

11/14/2013

constraints of

Our goal today...

• Learn how to access CI-BER data collection



Outline

- Part 1
 - Introduction to iRODS
 - What data do we have in CI-BER?
- Part 2
 - Hands on with icommands



What is iRODS?

- is software middleware that manages a highly controlled collection of distributed digital objects, while enforcing user-defined Management Policies across the multiple storage locations.
- open source, BSD license



How iRODS is used?

- A data grid for sharing data across collaborations
- a digital library for publishing data
- a preservation environment for long-term data retention
- a data workflow for data processing
- a system for federating real-time sensor data streams



Organize Distributed Data into a Sharable Collection

- Project repository
 - MotifNet manage collection of analysis products
- Institutional repository
 - Carolina Digital Repository for UNC collections
- Regional collaboration
 - RENCI Data Grid linking resources across North Carolina
- National collaboration
 - NSF Temporal Dynamics of Learning Center
 - Australian Research Collaboration Service
- National Library
 - French National Library
- National Archive
 - NARA Transcontinental Persistent Archive Prototype, Taiwan
- International collaboration
 - BaBar High Energy Physics (SLAC-IN2P3)



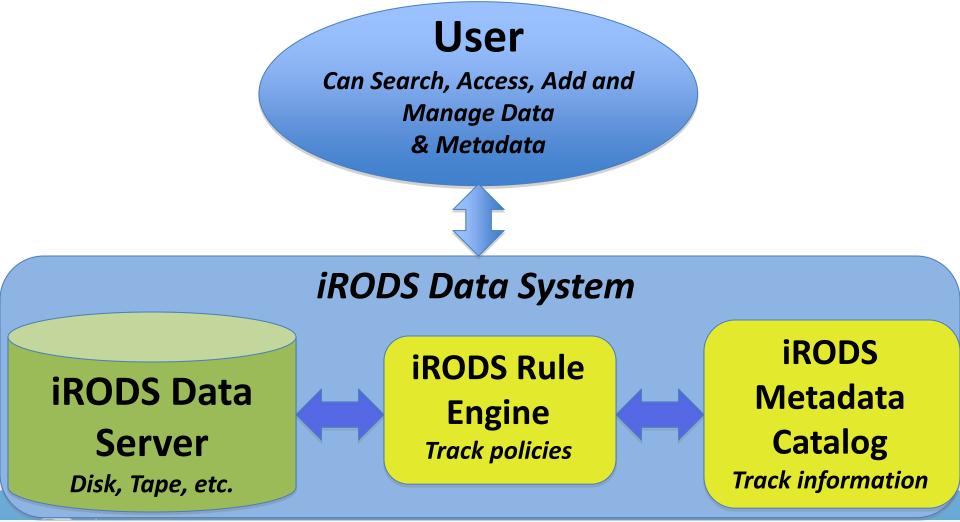


iRODS Glossaries

- Object
- Metadata
- Resource
- Micro-service
- Rule



Overview of iRODS Architecture



*Access data with Web-based Browser or iRODS GUI or Command Line clients.

iRODS Rule Example

acPostProcForPut{

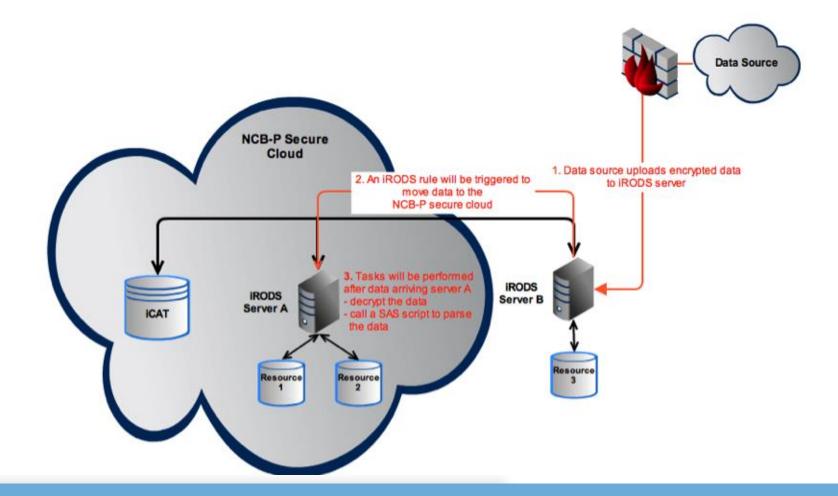
ON(\$rescName == "resource-uiuc"){

delay("<PLUSET>1m</PLUSET>"){

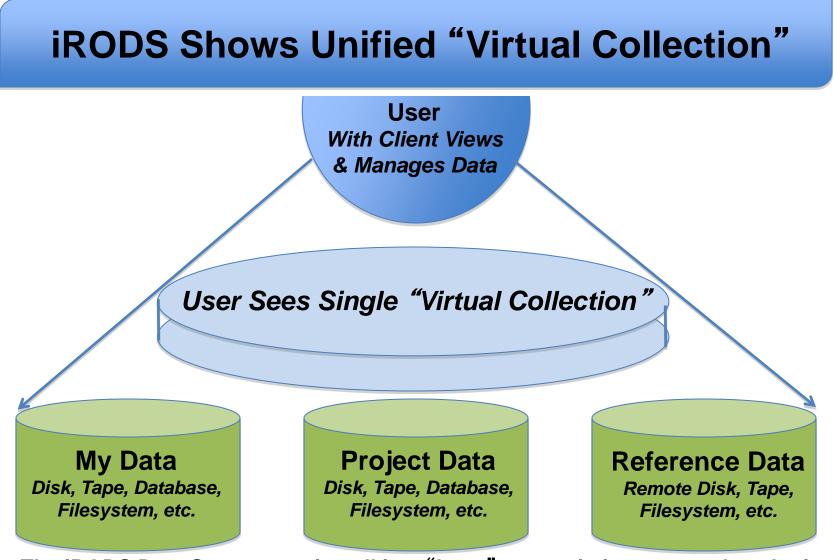
msiSysRepIDataObj("resource-unc","null");



iRODS Use Case Example – NCB-P



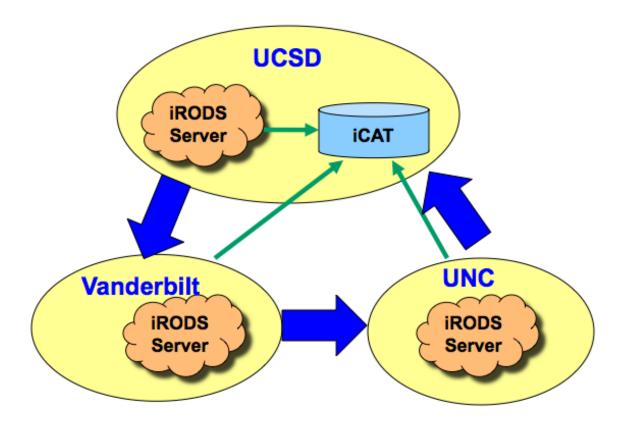




The iRODS Data System can install in a "layer" over existing or new data, letting you view, manage, and share part or all of diverse data in a unified Collection.



iRODS Use Case Example - TDLC





Data Grid Federation

- Motivation
 - Improve performance, scalability, and independence
- To initiate the federation, each Data Grid administrator establishes trust and creates a remote user
 - iadmin mkzone B remote Host:Port
 - iadmin mkuser rods#B rodsuser
- Use cases
 - Chained data grids
 NOAO
 - Master-slave data grids NIH BIRN
 - Central archive Ul
 - UK e-Science

– Deep archive

- NARA TPAP

- Replication

- NSF Teragrid



Accessing Data in Federated iRODS

Federated

irodsUser

(use iRODS clients)

"With access permissions"

"Finds the data"

"Gets data to user"

Two federated iRODS data grids

iRODS/ICAT system at University of North Carolina at Chapel Hill (renci zone) *iRODS/ICAT system at University of Texas at Austin (tacc zone)*

Federated irodsUsers can upload, download,

replicate, copy, share data, manage & track access, in either zone.



- The CI-BER project
 CI-BER (CyberInfrastructure for Billions of **Electronic Records):**
 - Funded by NARA / NSF (2010-2013)
 - See: <u>http://ci-ber.blogspot.com/</u>
 - **Big data management** project based on the integration of heterogeneous datasets:
 - 1. Testbed collection of 100M files and 50TB of data with content from over 100 federal agencies.
 - 2. Comprises nearly 6,000 file types ranging from a few files to tens of millions of files each, and including diverse file types (text, desktop publishing, databases, audio, video, GIS, XML, etc.), and historical, cultural, social science, and scientific content.
 - 3. The CI-BER testbed has primarily been used to support the development of scalable record visualization of e-records (geo-analytics) and test the development of national

federated infrastructure.



1. Testbed collection

marciano@ciber.renci.org:1249 | Sign Out

Collections «	Select All 🏠 Browse Up 💿 New 🗸 🤤 Delete 🔻 🔕 Upload 🗸 🛛 More 🔻 Search By Name						
ia, <mark>} cib</mark> ia, } ciberZone	Name 🔻	Resource	Size	Date Modified	1		
in the home	RG 013 - Records of the National Mediation Board			December 10, 2010, 3:42			
🕀 🦳 chien-yi	RG 015 - Records of the Department of Veterans Affairs			December 10, 2010, 3:46			
🖃 😋 ciberAdmin 👜 🦳 maconrad	RG 016 - Records of the Office of the Secretary of Agriculture			December 10, 2010, 3:48			
⊕ Communication myTest	RG 021 - Records of District Courts of the United States			December 10, 2010, 3:51			
🕀 🧰 escott	RG 022 - Records of the U.S. Fish and Wildlife Service			December 10, 2010, 4:17			
⊕ 🚞 jefferson ⊕ 🚞 maconrad	RG 024 - Records of the Bureau of Naval Personnel			December 10, 2010, 4:23			
marciano	RG 025 - Records of the National Labor Relations Board			December 10, 2010, 4:32			
🕀 🧰 public	RG 026 - Records of the U.S. Coast Guard			December 10, 2010, 4:39			
⊕ 🛅 rlopez ⊕ 🛅 trash	RG 029 - Records of the Bureau of Census			December 13, 2010, 7:11			
	RG 034 - Records of the Federal Deposit Insurance Corporation			December 15, 2010, 3:59			
	RG 038 - Records of the Office of the Chief of Naval Operations			December 14, 2010, 4:56			
	RG 046 - Records of the U.S. Senate			December 14, 2010, 6:08			
	-			December 14, 2010, 6:11			
	RG 047 - Records of the Social Security Administration			December 14, 2010, 6:11			
	RG 049 - Records of the Bureau of Land Management						
	RG 051 - Records of the Office of Management and Budget			December 14, 2010, 6:12			
	RG 053 - Records of the Bureau of the Public Debt			December 14, 2010, 6:13			
	RG 056 - General Records of the Department of the Treasury			December 14, 2010, 6:16			
	RG 057 - Records of the U.S. Geological Survey			November 29, 2010, 5:05			
	RG 058 - Records of the Internal Revenue Service			December 14, 2010, 6:18			
	RG 059 - General Records of the Department of State			December 15, 2010, 8:17			
	RG 060 - General Records of the Department of Justice			December 14, 2010, 6:26			
	RG 064 - Records of the National Archives and Records Administration			December 15, 2010, 2:48			

2. Geo-analytics prototype

Python Libs	iPhone 🚺 Netflix	SWeather - Durham	WunderMap Interact	XSo2i.jpg (897×120	Work	Read Later	Subscribe to this fee	🚯 Triangle Zope and P	🎸 Durham, NC Events	imgur: the simple in	» 🚞 Other Bookmarks
herbic ofr02	26 ofr026 ut_uta cd10)4 sd_mr aquipł sd	Lesofr02- ofr96- land 338_s447_w			Z	A.		\$		Sh-)
ofr99- ofr96 263_w450_	6- co_stl <mark>ofr97-</mark> ha24 c 574_o	l_sd_mcca_preut	_we ofr99- co_lfh con 77_gw	n 🍛			вс	AB	SK TA	7	J oc frand
ofr99- ut_w 78_po	a ds102 ofr027 ofr98 414	3- ut_wa ut_da ofi 44	r96- countjut_saruss 4_rŧ	bil			Vange		T ND	ON	1 And
co_ge geo2	5 landfil ofr96- ofr02 448_c	26 ut_car ofr02- ofr 338_o26	r99- aquife places ofr0 5	28				OR ID	SD WY	MN Minisobelis W MU MU Datest	Toronto NY MA
ag_chofr02 338_	e- ut_irorerf1 ofr02 s 338_	2- frame <mark>cir92</mark> da s	area co_weofr96- co_ 448_w	ge			San IT	W UT	Denver of America CO KS	MO St. Louis	PA Without States WV Menindus States VA
ds102_glac ofr0281_oc	c_ofr99- ut_cache 198_wlcł ti	ofr0288_etunit	nasqan2k					AZ Phoenix San Drago	NM TX Dollar	AR TN MS AL Atlanta LA S	SC Bermuda
ofr96- 453_aqbour	n ofr99-463_contou	Irs	huc250k					24		interest of the second second	A
cont1996				~ a					Mexico- ouatoria?- MCouatoria?		Bahamas Luba Sand Gilcos Asman Sands Sand
cont1994	ofr99-463_sites									Beize Guatemala Elisalvador Nicaragua	jameire Stylenening S. A.
cont1992	ofr99-463_surface	es								Par	name Arna ke kengan Venezuela Newsco
cont2000				0							Data CC-By-SA by OpenSt



- CI-BER directly contributes to one of the objectives of the White House Big Data Initiative:
 - Advanced technologies that support Big Data and data analytics:
 - Setting up of a big data test-bed for managing billion-record scale distributed collections.
 - Exploring computational infrastructure (Hadoop, noSQL)
 - Conducting scalable data analysis. The CI-BER team have built open-source interfaces to big data test-beds based on
 interactive map interfaces, where users can interact with records at scale from an iPad, cell phone, or web browser.
 Software at:
 - » github: <u>https://github.com/jeffheard/geoanalytics</u>
 - Relevant papers include:
 - » Cyberinfrastructure tools for Big Geographic Data", OGRS (Open Source Geospatial Research & Education Symposium). Oct. 24-26, 2012 Yverdon-les-Bains, Switzerland. (<u>http://youtu.be/5XCpED0qrJ4</u>)
 - » IEEE LDAV 2011 Symposium on Large-scale Data Analysis and Visualization: "A System for Scalable Visualization of Geographic Archival Records", Oct. 2011, Providence, RI. (<u>http://www.slideshare.net/richardjmarciano/a-system-for-scalable-visualization-of-geographic-archival-records</u>)
- CI-BER directly contributes to two of the objectives of the White House Big Data Initiative:
 - Educate and expand the Big Data workforce:
 - <u>Scholars:</u>
 - IEEE BigData 2013 workshop on Big Humanities Data, workshop on Oct. 8, 2013 in Santa Clara, <u>http://www.ischool.drexel.edu/bigdata/bigdata2013/topics.htm</u>
 - Professionals:
 - Workshops on CI-BER conducted at national professional meetings attended by archivists, records managers, and IT staff: NAGARA 2013 and SAA 2013.
 - Graduate Students and post-docs:
 - NSF Pan-American Advanced Studies Institutes Program (PASI), workshop taught on Jul. 24, 2013, <u>http://artcaonline.org/news/new-collaborations-at-the-center-of-pasi-2013/</u>
 - Virtual School on Computational Science and Engineering (VSCSE), Data Intensive Summer School, workshop taught on Jul. 8, 2013, <u>http://www.vscse.org/summerschool/2013/bigdata.html</u>
 - <u>Undergraduate Students:</u>
 - Duke University Bass Connections initiative for undergraduate scholars. Workshops in the fall of 2013 and spring of 2014, <u>http://today.duke.edu/2013/05/bassconnx</u>
 - Improve key outcomes in public awareness and education through demonstrations of Big Data applications:
 - <u>Citizens:</u>
 - CI-BER has launched a citizen-led initiative with the city of Asheville, N.C., including citizen groups, non-profits, city
 organizations, and universities, exploring the potential of crowdsourcing around the integration of urban renewal data, and
 census, economic, and planning content. The partnership is with the Southside Community Advisory Board and seeks to
 remap a historically African-American neighborhood in Asheville, adversely impacted in the 70s and 80s during urban
 renewal.



THE UNIVERS Apublic workshop was held on Aug. 3, 2013, http://www.citizen-times.com/article/20130804/LIVING/308040026/

of NORTH CAROLINA at CHAPEL HILL

Break for 5 minutes



iRODS Clients



Clients

- icommands unix shell commands
- C I/O library
- Jargon Java I/O library
- PRODS PHP iRODS client API
- iDrop desktop and web GUI
- Python load library
- Fedora / DSpace digital library
- Taverna / Kepler workflows
- Web / Windows browser
- WebDAV Davis / FUSE
- Parrot / SAGA
- GridFTP



EST-ful API – based on Jargon Core (planning...) of North Carolina at Chapel Hill 21

iRODS Web Browser

https://irods.org/web/

-Account Info	ormation
Host/IP :	iren2.renci.org
Port :	1249
Username:	chienyi
Password:	•••••



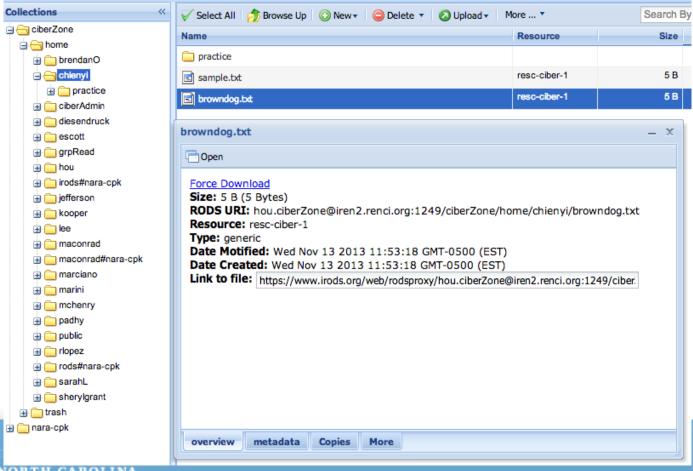
Web browser

- Directory style listing of collections
- List metadata
- List resources
- Load and retrieve files
- Replicate files
- Invoke interactive rule execution



iRODS Web Browser Practice

• Get browndog.txt from 'chienyi'



TH of NORTH CAROLINA at CHAPEL HILL

Installation of icommands



iRODS installation

- Download appropriate installation manual from iRODS Wiki <u>http://irods.org</u>
- Installation procedure will take

 Up to 30 minutes for
 server/catalog/clients
 - Up to 10 minutes for server/clients
 - About 3 minutes for clients
- We will do a client install



iRODS - Unix/Linux/Mac Installation

- <u>https://www.irods.org/download3.2.html</u>
 Tar file
 - Installation script (Linux, Solaris, Mac OS X)
 - Automated download of PostgreSQL, ODBC
 - Installation of PostgreSQL, ODBC, iRODS
 - Initiation of iRODS collection



iRODS Installation- Unix

- Unpack the release tar file
 - gzip -d irods.tgz
 - tar xf irods.tar
- It will prompt for a few parameters



iRODS Source Distribution

- INSTALL.txt
- LICENSE.txt
- Makefile
- README.txt
- Configure
- Vault
- irodsctl
- irodssetup

- COPYRIGHT
- CVS
- bin
- clients
- config
- doc
- install
- installLogs
- lib
- modules
- nt
- scripts
- server



irodssetup

- Set up iRODS
- -----
- iRODS is a flexible data archive management system that supports many different site configurations. This script will ask you a few questions, then automatically build and configure iRODS.
- There are four main components to iRODS:
- 1. An iRODS server that manages stored data.
- 2. An iCAT catalog that manages metadata about the data.
- 3. A database used by the catalog.
- 4. A set of 'i-commands' for command-line access to your data.
- You can build some, or all of these, in a few standard configurations. For new users, we recommend that you build everything.



- iRODS configuration setup
- This script prompts you for key iRODS configuration options.
- Default values (if any) are shown in square brackets [] at each
- prompt. Press return to use the default, or enter a new value.
- For flexibility, iRODS has a lot of configuration options. Often
- the standard settings are sufficient, but if you need more control
- enter yes and additional questions will be asked.
- Include additional prompts for advanced settings [no]?



- iRODS configuration (advanced)
- -----
- iRODS consists of clients (e.g. i-commands) with at least one iRODS
- server. One server must include the iRODS metadata catalog (iCAT).
- For the initial installation, you would normally build the server with
- the iCAT (an iCAT-Enabled Server, IES), along with the i-commands.
- After that, you might want to build another Server to support another
- storage resource on another computer (where you are running this now).
- You would then build the iRODS server non-ICAT, and configure it with
- the IES host name (the servers connect to the IES for ICAT operations).
- If you already have iRODS installed (an IES), you may skip building
- the iRODS server and iCAT, and just build the command-line tools.





- iRODS can make use of the Grid Security Infrastructure (GSI)
- authentication system in addition to the iRODS secure
- password system (challenge/response, no plain-text).
- In most cases, the iRODS password system is sufficient but
- if you are using GSI for other applications, you might want
- to include GSI in iRODS. Both the clients and servers need
- to be built with GSI and then users can select it by setting
- irodsAuthScheme=GSI in their .irodsEnv files (or still use
- the iRODS password system if they want).
- Include GSI [no]? no



- Confirmation
- -----
- Please confirm your choices.
-) ------
- GSI not selected
- Build iRODS command-line tools
- Save configuration (irods.config) [yes]?
- Saved.
- Start iRODS build [yes]?



- Build and configure
- ------
- Preparing...
- Configuring iRODS...
- Step 1 of 4: Enabling modules...
- properties
- Step 2 of 4: Verifying configuration...
 - No database configured.
- Step 3 of 4: Checking host system...
- Host OS is Mac OS X.
- Perl: /usr/bin/perl
- C compiler: /usr/bin/gcc (gcc)
- Flags: none
- Loader: /usr/bin/gcc
 - Flags: none
- Archiver: /usr/bin/ar
- Ranlib: /usr/bin/ranlib
- 64-bit addressing not supported and automatically disabled.



- Step 4 of 4: Updating configuration files...
- Updating config.mk...
- Created /iRODS/config/config.mk
- Updating platform.mk...
- Created /iRODS/config/platform.mk
- Updating irods.config...
- Updating irodsctl...
- Compiling iRODS...
- Step 1 of 2: Compiling library and i-commands...
- Step 2 of 2: Compiling tests...
- Done!



- To use the iRODS command-line tools, update your PATH:
- For csh users:
- set path=(/iRODS/clients/icommands/bin \$path)
- For sh or bash users:
- PATH=/iRODS/clients/icommands/bin:\$PATH
- Please see the iRODS documentation for additional notes on how
- to manage the servers and adjust the configuration.
- Change the path name to your installation path



Environment Variables

- In home directory
 - $-cd \sim /.irods$
 - -vi.irodsEnv
- Default values to describe settings for interacting with your data grid



Environment File

iRODS personal configuration file.
#

iRODS server host name:irodsHost 'iren2.renci.org'# iRODS server port number:irodsPort 1249

Default storage resource name: irodsDefResource 'resc-ciber-1' # Home directory in iRODS: irodsHome '/ciberZone/home/chienyi' # Current directory in iRODS: irodsCwd '/ciberZone/home/chienyi' # Account name: irodsUserName 'chienyi' # Zone:

irodsZone 'ciberZone'



User Configuration

- To use the iRODS 'i-commands', update your PATH:
- For csh users:
 - set path=(/storage-site/iRODS/clients/icommands/bin \$path)
- For sh or bash users:
 - PATH=/storage-site/iRODS/clients/icommands/bin:\$PATH



icommands I

- iinit: login
- UNIX like commands:
 - ipasswd: change password
 - ils: list files/directories
 - icd: change directory
 - imkdir: create directory
 - icp: copy files/direcotries
 - imv: change name of file/directory
 - ichmod: modify permission of file/directory



icommands II

- iget: download files/directories
- iput: upload files/directories
- imeta: show/edit metadata
- ireg: register files/directories to iCAT
- irepl: replicate files to other resources
- irule: execute a rule file

Check out ihelp or for more icommands



icommands practice

- 1. Change your password
- Download sample1.txt and sample2.txt from /ciberZone/home/chienyi
- 3. Upload a file to your home directory
- 4. Add "author" metadata to the file
- 5. Change access permission to allow others to get your file



icommand Practice - Answers

- ipasswd
- iget /ciberZone/home/chienyi/sample1.txt
- iput sample3.txt .
- imeta add -d sample2.txt author hou
- ichmod read grpRead sample2.txt



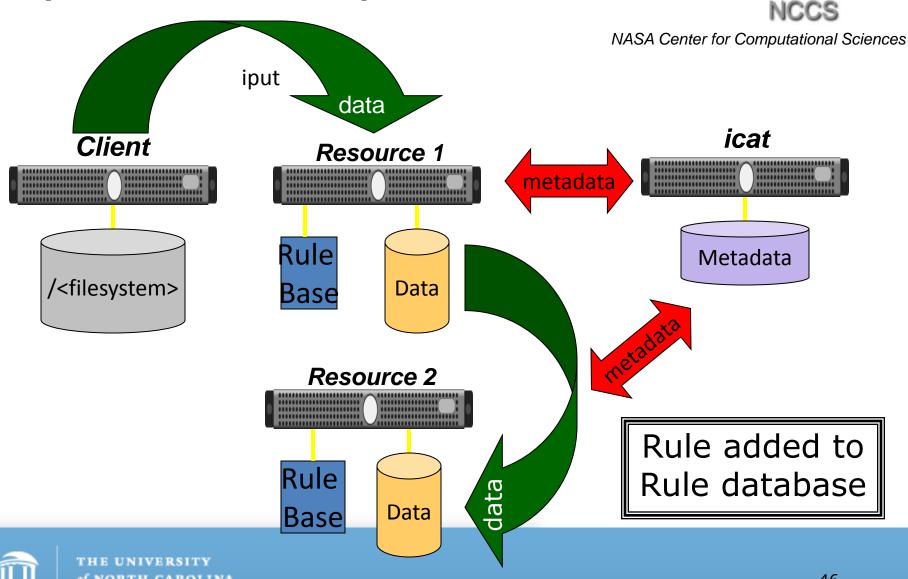
Cyberintegrator

	Hist	ory
	🖌 🗸 Tools	
	Search	Q
	iget file netstat ils	
	ils test 2 iget 1 iget	
	Ð	≌ +
	See Workflows	
	Search	Q
т	O-O untitled O-O ils test2 NORTH'CAROLINA	
oj at		



iput With Replication

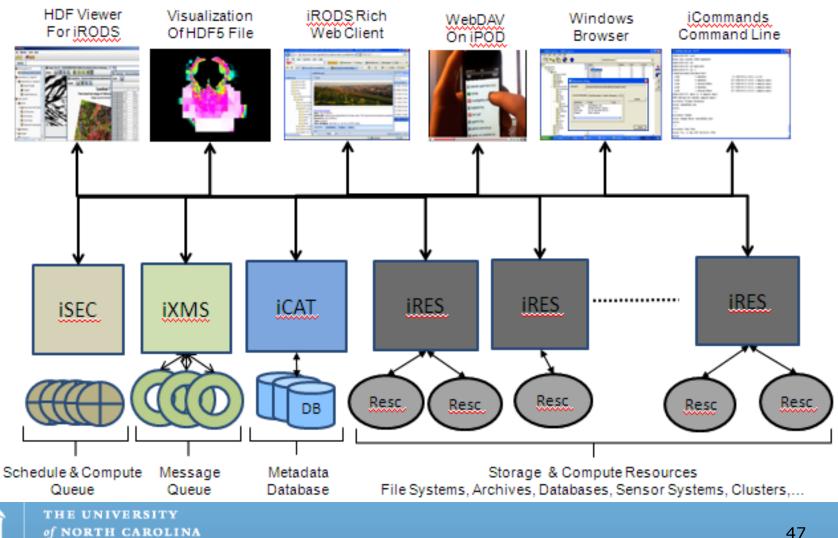




of NORTH CAROLINA at CHAPEL HILL

46

iRODS Distributed Data Management



at CHAPEL HILL