Policy-Based Data Management

NSF Software Development for Cyberinfrastructure ~ Data Intensive Cyber Environments Center, UNC at Chapel Hill

Enable Data-Driven Science

- Manage time dependent data
  - Compare current observation with prior observation
  - Compare current observation with simulation of the phenomena
- Federate repositories from multiple independent projects
  - Base research on integration of data from different communities
  - Federate collections across multiple federal agencies
- Develop reference collections
  - Promote use within education
  - Promote repurposing of collections for interdisciplinary research
- Virtualize the data life cycle
  - Enable federation-based sustainability
- Build national scale data cyber infrastructure
  - Enable interoperation across heterogeneous data management systems

Data Life Cycle

Each data life cycle stage re-purposes the original collection

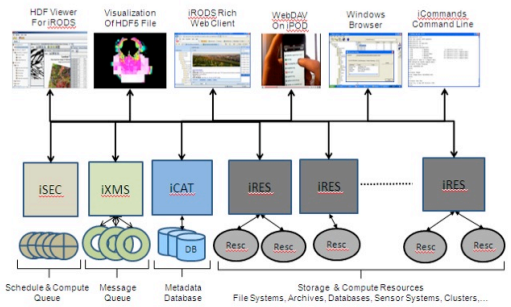
Project Collection	Data Grid	Data Processing Pipeline	Digital Library	Reference Collection	Federation
Private	*Shared*	*Analyzed*	*Published*	*Preserved*	*Sustained*
Local Policy	Distribution Policy	Service Policy	Description Policy	Representation Policy	Re-purposing Policy

Stages correspond to *addition of new policies* for a *broader community,* virtualizing data life cycle through policy evolution

Policy-based Data Management

- Turn Management Policies into computer actionable Rules
- Turn management processes into remotely executable computer procedures
  - Apply a workflow at the storage system to filter, subset, manipulate data
  - Minimize size of data moved over the network
  - Automate administrative tasks
  - Compose procedural workflows from operations encapsulated in micro-services
- Organize state information as metadata on logical name spaces for files, users, storage, rules, and policies

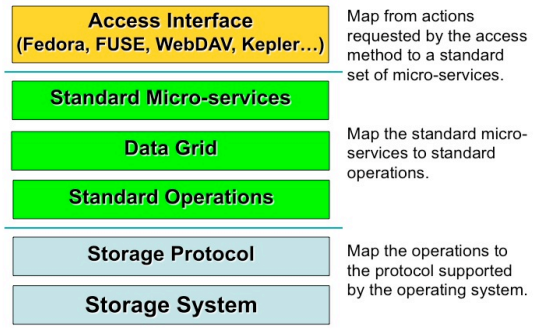
iRODS Distributed Data Management



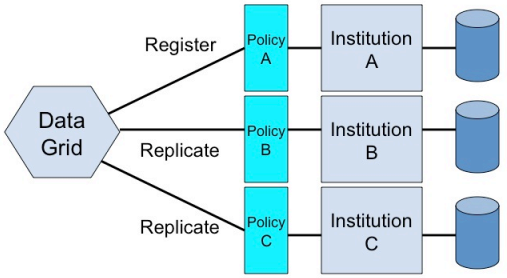
Policies

- Administrative
  - Retention, disposition, distribution, replication, deletion, registration, synchronization, integrity checks, IRB approval flags, addition of users, addition of resources
- Ingestion and Access
  - Metadata extraction, logical organization, derived data product generation, redaction, time-dependent access controls
- Validation
  - Authenticity checks, chain of custody, repository trustworthiness, audit trails

Data Virtualization

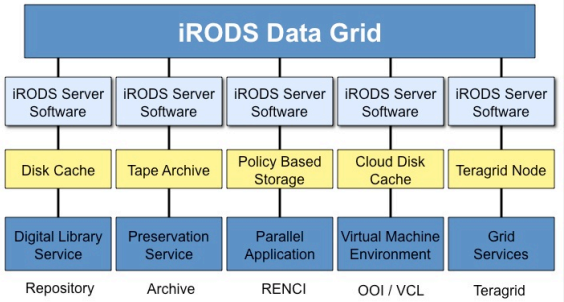


Federation Based Sustainability



Virtualizing sustainability through re-purposing of collections

Integrating Across Repository - Archive Supercomputer - Cloud - Grid



iRODS software is available under a BSD open source license at the iRODS wiki https://www.irods.org.