Data Management in Release 2

The User's and Administrator's Points of View

Peter.Kunszt@cern.ch

EDG Conference Barcelona, 12 -15 May 2003

Outline

➢ Data Management Tools

➢ API's } For users

➢ Services

➢ Admin tools } For admins

Overview: Data Management for the User

➢ Users want to

- ➢ access their files transparently
- ➢ find the locations of their files
- ➢ query and browse the catalogs
- ➢ group files into collections
- ➢ make data available at many sites
- ➢ delete files and data

Overview: Data Management for the Administrator

- Administrators want to

    - installation, configuration

    - validation and functionality testing

    - detect inconsistencies, reproduce problems

    - clean up catalogs from erroneous and data

    - clean up data stores

    - subscribe to data to be replicated automatically (R2.1)

Naming : Everything is a URI

> All file names are valid URIs, i.e. they have a scheme and a scheme-specific part, separated by a colon. URLs are also valid URIs. URI jargon:

```
[scheme:]scheme-specific-part[#fragment]
[scheme:][//authority][path][?query][#fragment]
```

> User's view : Logical File Name. Scheme = lfn

```
lfn:some_string
```

restrictions: being a valid URI excludes certain characters like space, $, :, %, etc. These may be escaped using % (%20 is the space character).

> Catalog view : GUID. Scheme = guid

```
guid:73e16e74-26b0-11d7-b1e0-c5c68d88236a
```

Naming (Cont.)

> Storage Resource Manager view: Site File Names and Site URL. Scheme : srm

```
srm://srmhost.domain/path/to/data/file.dat
```

This is called a SURL. The SFN is just the *path* part of this URI, i.e. `path/to/data/file.dat` in this case.

RLS stores SURL – GUID mappings

RMC stores GUID – LFN mappings

User's Command Line Interface

> The EDG Replica Manager is the user's interface to all Data Management functions. It is installed on UI and WN in

$EDG_LOCATION/bin/edg-replica-manager

edg-rm as a short-hand.

> Command similar to cvs:

edg-rm [options] command [command-options]

Options for all commands

- General options for all commands are:

  - -h, --help   print a help screen on all commands or help on commands if command is given

  - -v, --verbose   print informative messages while executing

  - -i, --insecure   connect insecurely to the services

  - --vo=VO   set the vo to 'VO'. Necessary for insecure operations and until we integrate with VOMS **currently this option is mandatory**.

  - --config=file   load configuration from 'file' instead of the default config file at $EDG_LOCATION/etc/edg-replica-manager/edg-replica-manager.conf

Bringing a file in the Grid

➢ Make a file available in the grid:

```
copyAndRegisterFile  sourceURL [command-options]
```

Copy the file given by source and register it in the RLS. The command will return the GUID of the registered file. Valid source URL schemes are file, gsiftp, http, ftp. For a local file you have to specify the full file path using the file scheme. Short form: cr, creg

➢ Options

  ➢ -d destination   specify the host of the destination SE or even the full URL of the destination SURL.

  ➢ -l logical-filename   specify a logical file name

  ➢ -p protocol specify the protocol to be used (currently only gsiftp)

  ➢ -n nstreams specify the number of streams to be used (def: 8)

➢ Example

edg-rm --vo=wpsix cr file:/home/pkunszt/test.dat –l lfn:pkunszt-test1

Bringing a file in the Grid 2.

> If a file is already available through an SRM but is not registered yet, it can be registered without the need to copy it using

```
registerFile surl [command-options]
```

Register the file given by the SURL it in the RLS. The command will return the GUID of the registered file. Short form: rf

> Options

  > -l logical-filename   specify a logical file name

> Example

edg-rm --vo=wpsix rf
srm:/lxshare0408.cern.ch/flatfiles/wpsix/test1.dat –l lfn:pkunszt-test1

Listing files available on the Grid

- Listing files:

  ```
  listReplicas lfn-or-guid-or-surl
  ```

  list all SURLs that correspond to the given lfn, guid or surl

- List closest file:

  ```
  listBestFile lfn-or-guid
  ```

  list the SURL that is 'closest' in terms of network cost.

- Getting the GUID

  ```
  listGUID lfn-or-surl
  ```

  retrieve the GUID corresponding to the given lfn or surl.

Getting a file from the Grid

- To get a local copy of a file that is 'on the Grid' use the command

  ```
  copyFile sourceURL destURL [command-options]
  ```

  Copy a file. The source may be any kind of file, an LFN, GUID, SURL or other (http, ftp, gsiftp, file), but the destination can only be a non-grid file, i.e. of the scheme file, gsiftp. The command is to a large extent identical to globus-url-copy but it can also take LFNs and GUIDs as source files. Short form: cp

- Options

  - -f, --force   overwrite destination if it exists

  - -p, --protocol prot   use the given protocol. Currently only gsiftp is supported.

  - -n, --nstreams n   set the number of parallel streams. Defaults to 8.

- Example

  edg-rm --vo=wpsix cp lfn:peter-test1 file:/home/pkunszt/test.dat

---

Copying files in the Grid

- To create a new replica at a given site use

  ```
  replicateFile sourceURL [command-options]
  ```

  Copy a file to a grid site. Unlike copyFile, the destination may **only** be a SURL or SE host. The source has to be an LFN, GUID or SURL. Short form: rep

- Options

  - -d, --destination dest   destination SE host or SURL. If it is omitted, the default SE is used as destination

  - -p, --protocol prot   use the given protocol. Currently only gsiftp is supported.

  - -n, --nstreams n   set the number of parallel streams. Defaults to 8.

- Example

  edg-rm --vo=wpsix rep lfn:peter-test1 –d lxshare0408.cern.ch

  The command `getBestFile` is identical, the source has to be GUID or LFN. The best source to copy from is determined through the optimizer.

Removing and unregistering

➢ To completely remove a file from the Grid, use

```
deleteFile sourceURL [command-options]
```

Delete a file. If the source is an LFN or GUID, the destination SE host (from where to delete the file) has to be given. Or the SURL can be specified directly. For GUIDs the `--all` flag may be set, in which case all replicas of the given GUID are removed. Short form: del

➢ Options

  ➢ -d, --destination dest   destination SE host.

  ➢ --all   remove all replicas (only valid with GUIDs)

➢ Example

edg-rm --vo=wpsix del lfn:peter-test1 –d lxshare0408.cern.ch

Similarly, files may be unregistered (the copy will not be physically removed) using `unregisterFile`.

Miscellaneous Commands

- ➢ Getting the Server and Client version

    getVersion

- ➢ Printing the information that is used from the info providers

    printInfo

- ➢ Listing the contents of a directory on an SRM

    listDirectory

Environment Variables: Logging behavior

➢ EDG_RM_LOGPROPERTIES   by default logging is **disabled**. the default logging properties are read from

`$EDG_LOCATION/etc/edg-replica-manager/log4j.properties`

to change it as a user: copy this file and edit. Line of interest:

`log4j.logger.org.edg.data=OFF`

Set this to one of FATAL INFO WARN DEBUG which are the available logging level. DEBUG gives the most verbose information. Set EDG_RM_LOGPROPERTIES to your local copy.

➢ EDG_RM_LOGFILE   by default, the log file goes to

`/tmp/edg-replica-manager-username.log`

where username is the unix user that the command is run under. Set EDG_RM_LOGFILE to change this.