HPCC Systems™

HPCC Data Handling

Boca Raton Documentation Team

LexisNexis® Risk Solutions

HPCC Data Handling

Boca Raton Documentation Team

We welcome your comments and feedback about this document via email to <docfeedback@hpccsystems.com> Please include **Documentation Feedback** in the subject line and reference the document name, page numbers, and current Version Number in the text of the message.

February 2012 Version 3.6.2.1

HPCC Data Handling

Introduction

There are a number of different ways in which data may be transferred to, from, or within an HPCC system. For each of these data transfers, there are a few key parameters that must be known.

Prerequisites for most file movements:

- Logical filename
- Physical filename
- Record size (fixed)
- Source directory
- Destination directory
- Dali IP address (source and/or destination)
- Landing Zone IP address

The above parameters are used for these major data handling methods:

- Import - Spraying Data from the Landing Zone to Thor
- Export - Despraying Data from Thor to Landing Zone
- Copy - Replicating Data from Thor to Thor (within same Dali File System)
- Copying Data from Thor to Thor (between different Dali File Systems)

Data Handling Terms

A *spray* or *import* is the relocation of a data file from one location (such as a Landing Zone) to a Data Refinery cluster. The term spray was adopted due to the nature of the file movement – the file is partitioned across all nodes within a cluster.

A *despray* or *export* is the relocation of a data file from a Data Refinery cluster to a single machine location (such as a Landing Zone). The term despray was adopted due to the nature of the file movement – the file is reassembled from its parts on all nodes in the cluster and placed in a single file on the destination.

A *copy* is the replication of a data file from one Data Refinery cluster to another cluster within the same environment.

A *Remote copy* is the replication of a data file from one Data Refinery cluster to another cluster in a different environment.

A *Landing Zone* (or Drop Zone) is a physical storage location defined in your system's environment. There can be one or more of these locations defined. A daemon (DaFileSrv) must be running on that server to enable file sprays and desprays.

Working with data files

Once you start working with your HPCC system, you will want to process some real data, this section shows you how to load data to your HPCC system.

Before you begin

First, you should consider the size of the data and the capacity of your system. A typical production HPCC system would have much more data capacity than a development system. The size of the files you wish to work with is limited by the size of your system.

Uploading a file

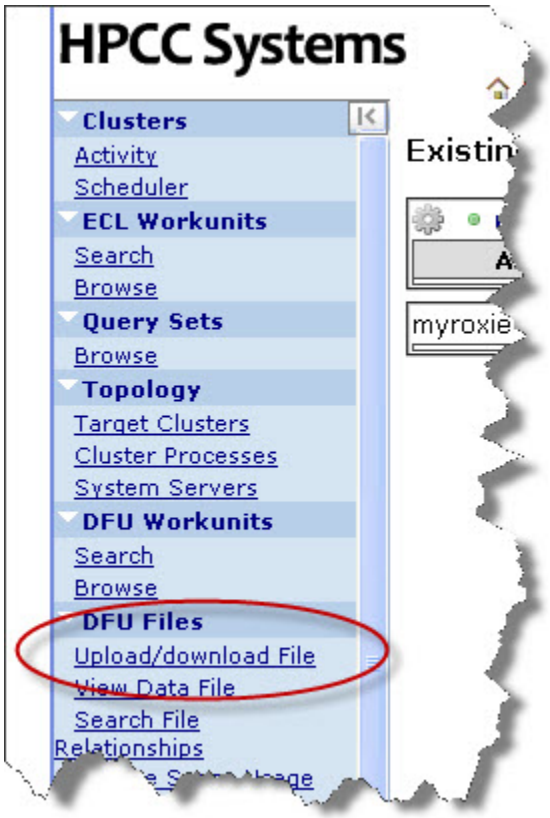
For smaller data files, maximum of 2GB, you can use the upload/download file utility in ECL Watch.

1. In your browser, go to the **ECL Watch** URL displayed example, http://nnn.nnn.nnn.nnn:8010, where nnn.nnn.nnn.nnn is your ESP Server's IP address.

	Your IP address could be different from the ones provided in the example images. Please use the IP address provided by **your** installation.

2. From ECL Watch page, click on the **Upload/download File** link in the menu on the left side.

Figure 1. Upload/download



Once you click on the Upload/download file link, it will take you to the dropzones and files page, where you can choose to **Browse** your machine for a file to upload:

Figure 2. Dropzones

3. Press the **Browse** button to browse the files on your local machine, select the file to upload and then click **Open** button.

   The file you selected should appear in the **Select a file to upload** field.

4. Press on **Upload Now** to complete the file upload.

Uploading files with a Secure Copy Client

To upload a large file for processing to your virtual machine, you will need a tool that supports the secure copy protocol. In this section, we discuss using WinSCP. There are other tools available, but the steps are similar.

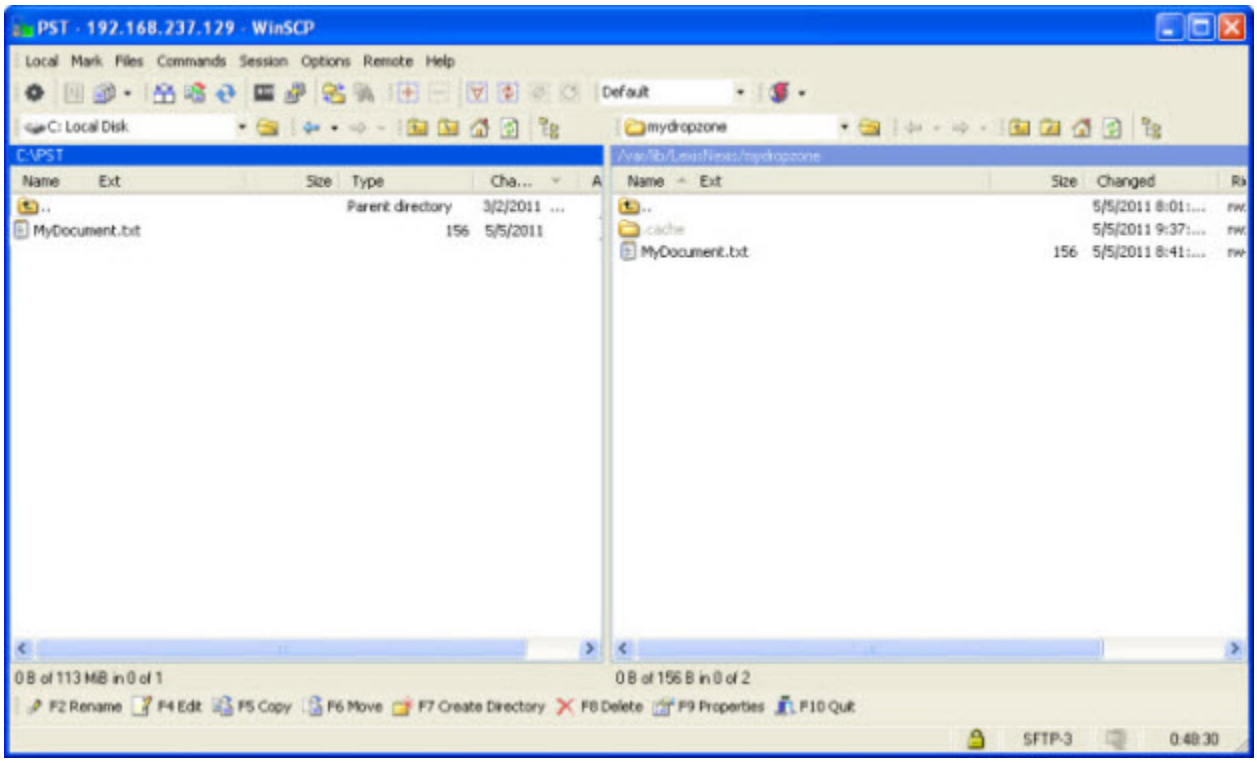
1. Open the WinSCP tool, and login to your Landing Zone node using the username and password given.

Login ID:	hpccdemo
Password:	hpccdemo

2. Once logged in, it should, navigate automatically to the landing zone folder. (/var/lib/LexisNexis/mydropzone)

3. Navigate to where your local file is in the left part of the window.

Figure 3. WinSCP



4. Select the data file to send and copy it to the landing zone, using drag-and-drop.

Data Handling Methods

There are several ways to spray, despray, or copy data files:

- The DFU interface in Ecl Watch
- The DFU Plus command line utility

  See the *Client Tools* manual for details

- Using ECL Code and FileServices library functions.

  See the *ECL Language Reference* for details.

Data Handling Using ECL Watch

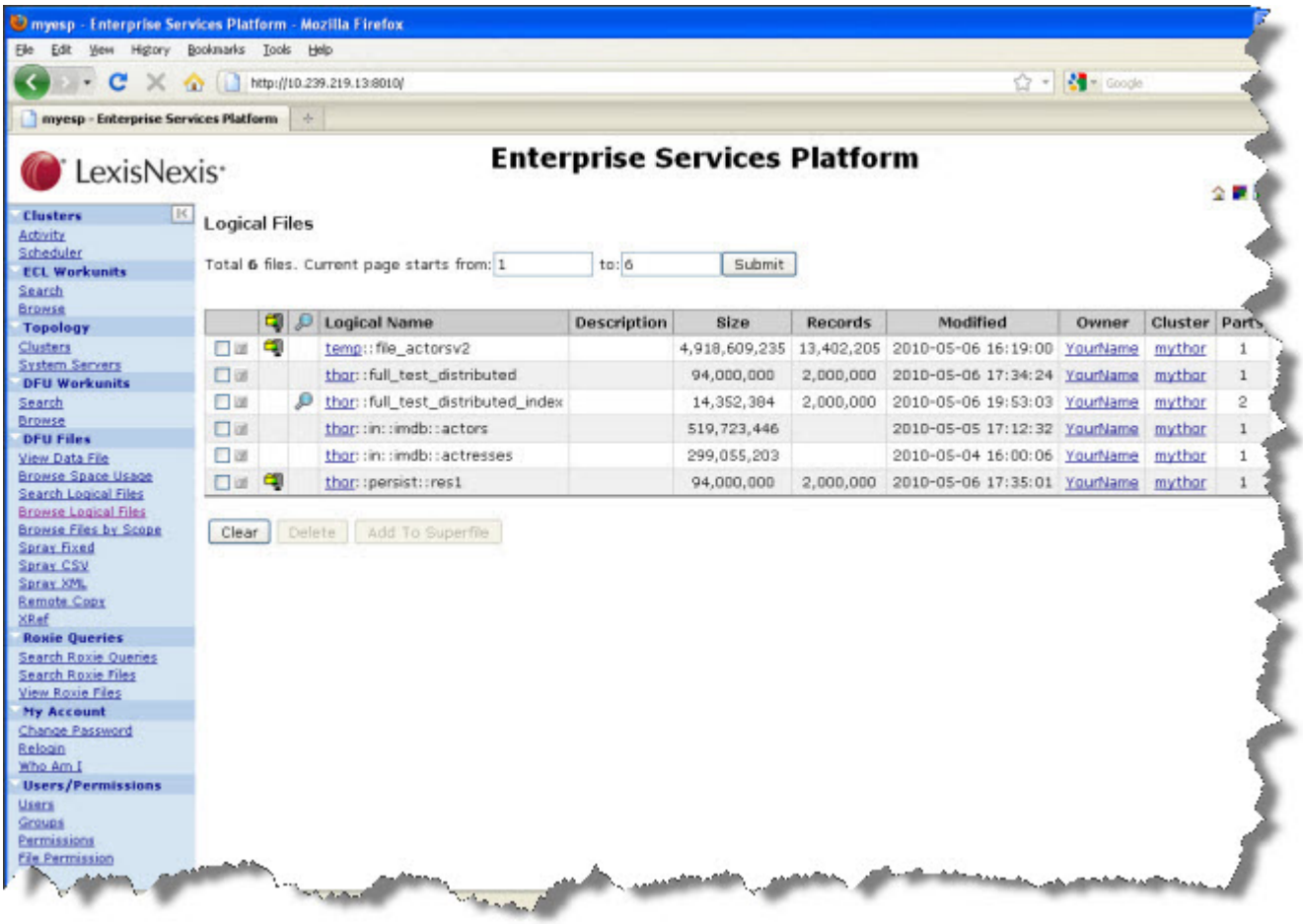
- Login to ECL Watch for the environment.

  The URL is the IP address where the ESP Server is installed plus the port to which the wssmc service is bound. The default port is 8010. For example:

```
http://<ESPserverIP>:8010/
```

- Click on the **Browse Logical Files** hyperlink below **DFU Files** in the menu on the left.

  The Logical Files page displays showing all files with logical entries in the Dali Server's Distributed File System.



  From this page, you can despray or copy any file.

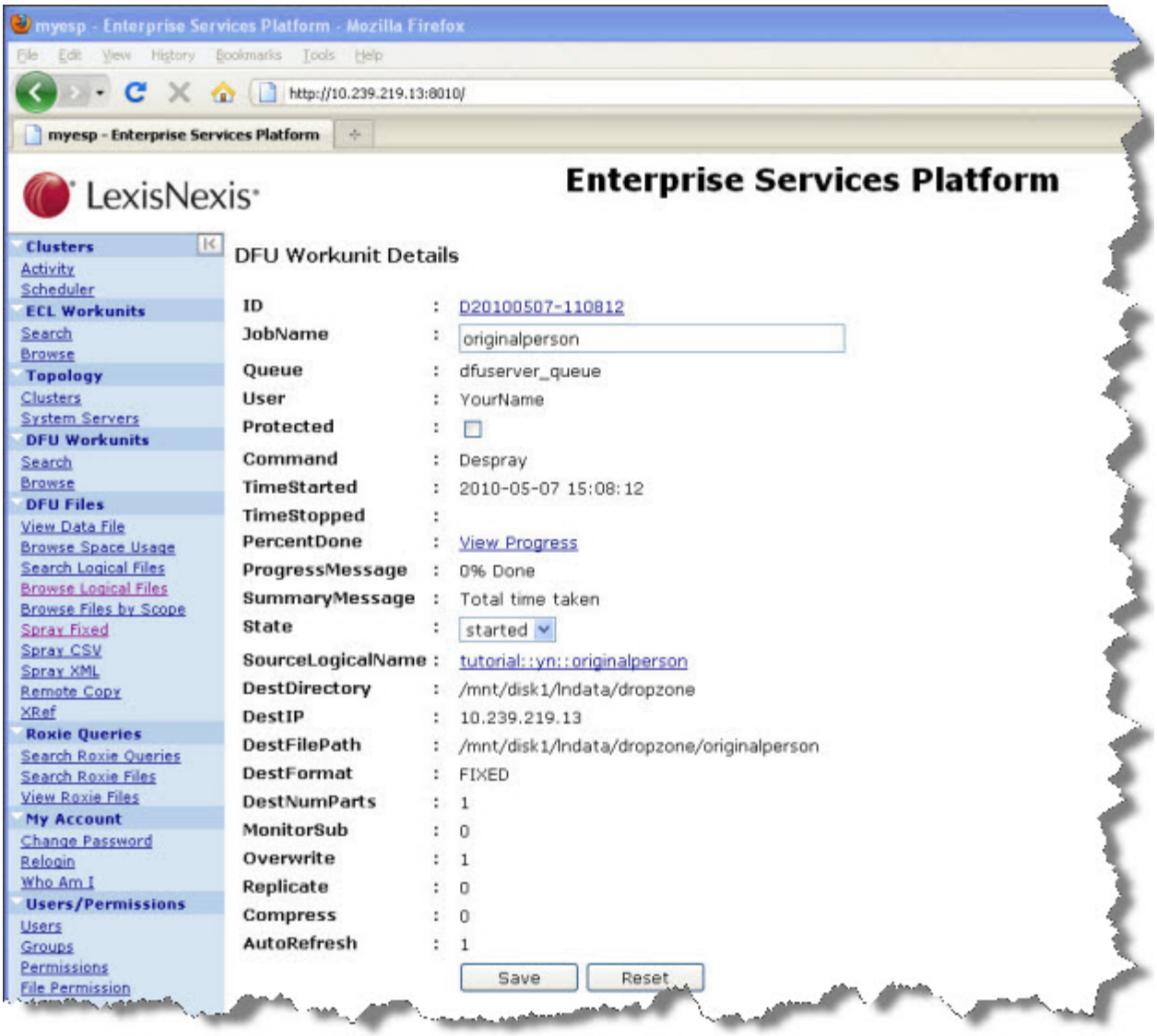
Desprays

- Locate the file to despray in the list of files, then select the arrow graphic on the left hand side, then select **Despray** from the pop-up menu.
- Check the **Source** information that is already filled in.
- Provide **Destination** information.

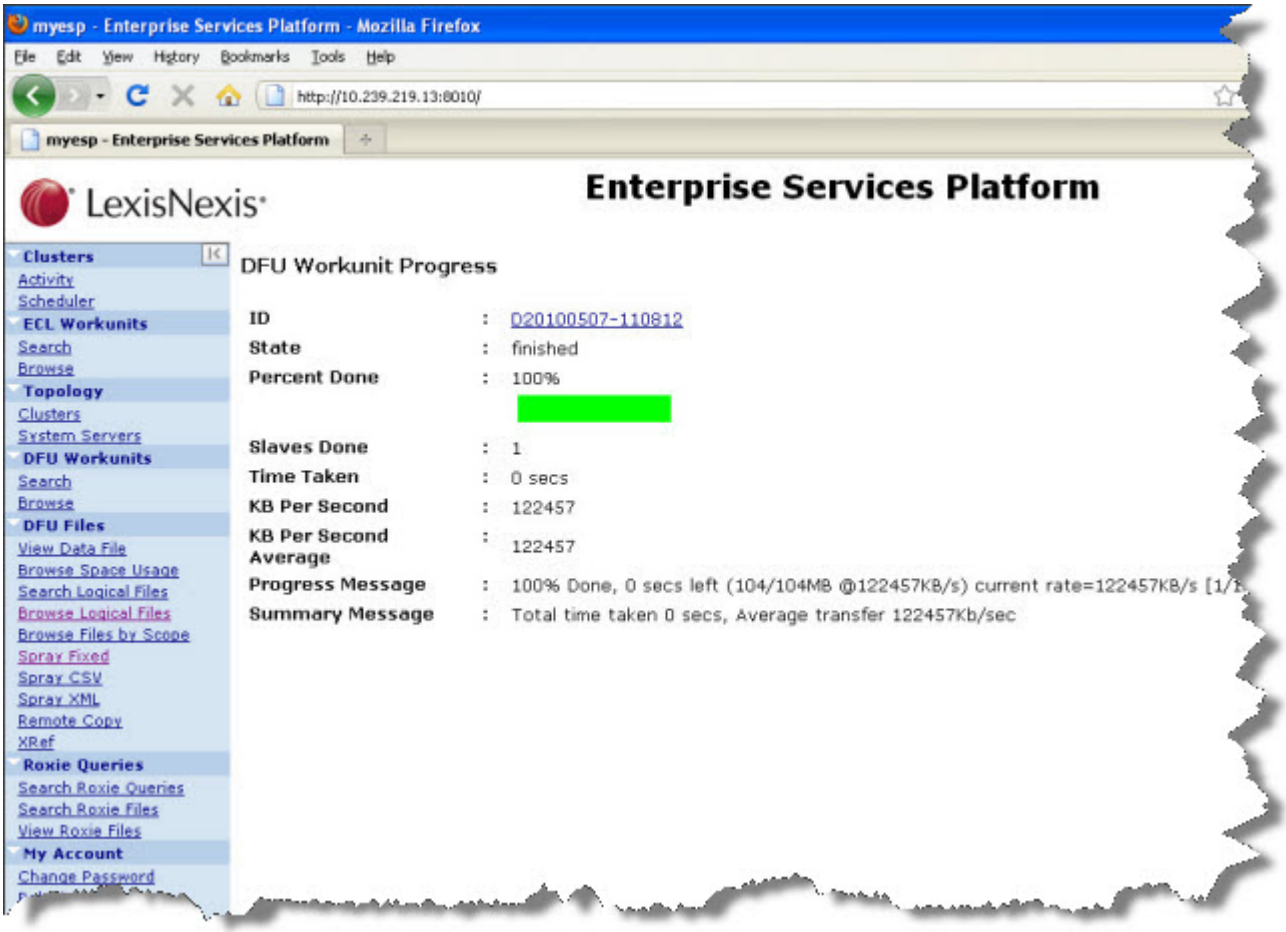
Destination-Machine	Use the drop-list to select the machine to despray to. The items in the list are landing zones defined in the system's confguration.
Destination-IP Address	This is prefilled based upon the selected machine.
Destination-Local Path	Provide the complete file path of the destination including file name and extention.
Destination-Network Path	The complete network path of the destination including file name and extension. (read only)
Overwrite	Check this box to overwrite a file with the same name if it exists.

- Press the **Submit** button.

The **DFU Workunit** displays.



- Press the **Refresh** button periodically until the status of your request indicates it is **Finished** or click on the **View Progress** hyperlink to see a progress indicator.



The Progress window shows a green progress bar indicating the percentage of completion, as well as other information related to the operation.

If a job fails, information related to the cause of the failure also displays.

Spray Fixed

- Click on the **Spray Fixed** hyperlink below **DFU** in the menu on the left.

  The **Spray Fixed** page displays.

- Fill in **Source** information (**Machine, IP, File Path, and record length)** and the **Destination** information (**Group**and **Label).**

	Source:
Machine	Use the drop-list to select the machine where the source file is located.
IP	IP address of machine from which to spray. This is automatically completed when you select the **Source Machine.**
Local Path	The file path of source file to spray.
Record Length	The size of each record.

Destination:	
Group	Select the name of THOR cluster to spray to.
Label	The logical name that you choose for the file.
Options:	
Overwrite	Check this box to overwrite files of the same name.
Replicate	Check this box to create backup copies of all file parts exist in the backup directory (by convention on the secondary drive of the node following in the cluster).
Compress	Check this box to compress the files.

- Press the **Submit** button.

  The **DFU Workunit** displays.

- Press the **Refresh**button periodically until the status of your request indicates it is **Finished** or click on the **View Progress** hyperlink to see a progress indicator.

You can use the Choose File button to look up files on the selected source machine.

Spray CSV

- Click on the **Spray CSV** hyperlink below **DFU** in the menu on the left.

  The **Spray CSV** page displays.

- Fill in **Source** information (**Machine, IP, File Path, and record information)** and the **Destination** information (**Group** and **Label**).

Source:	
Machine	Use the drop-list to select the machine where the source file is located.
IP	IP address of machine from which to spray. This is automatically completed when you select the **Source Machine.**
Local Path	The file path of source file to spray.
Max Record Length	The length of longest record in the file.
Separator	The character used as a separator in the source file.
Line Terminator	The character used as a line terminator in the source file.
Quote	The character used as a quote in the source file.
Destination:	
Group	Select the name of THOR cluster to spray to.
Label	The logical name that you choose for the file.
Options:	
Overwrite	Check this box to overwrite files of the same name.
Replicate	Check this box to create backup copies of all file parts exist in the backup directory (by convention on the secondary drive of the node following in the cluster).
Compress	Check this box to compress the files.

- Press the **Submit** button.

The **DFU Workunit** displays.

- Press the **Refresh** button periodically until the status of your request indicates it is **Finished**or click on the **View Progress** hyperlink to see a progress indicator.

You can use the Choose File button to look up files on the selected source machine.

Spray XML

- Click on the **Spray XML**hyperlink below **DFU**in the menu on the left.

  The **Spray XML** page displays.

- Fill in **Source** information (**Machine, IP, File Path, and record information)**and the **Destination** information (**Group** and **Label).**

Source:	
Machine	Use the drop-list to select the machine where the source file is located.
IP	IP address of machine from which to spray. This is automatically completed when you select the **Source Machine.**
Local Path	The file path of source file to spray.
Format	Select the file format from the drop-list.
Max Record Length	The length of longest record in the file.
Row Tag	The record separator tag in the XML file
Destination:	
Group	Select the name of THOR cluster to spray to.
Label	The logical name that you choose for the file.
Options:	
Overwrite	Check this box to overwrite files of the same name.
Replicate	Check this box to create backup copies of all file parts exist in the backup directory (by convention on the d: drive of the node following in the cluster).
Compress	Check this box to compress the files.

- Press the **Submit** button.

  The **DFU Workunit** displays.

- Press the **Refresh** button periodically until the status of your request indicates it is **Finished** or click on the **View Progress** hyperlink to see a progress indicator.

You can use the Choose File button to look up files on the selected source machine.

Copy

- Locate the file to copy in the list of files, then click on the arrow icon, then select **Copy** from the pop-up menu..

- Fill in **Destination** and **Options** information.

Destination:	
Group	Select the name of THOR cluster to copy to.
Note	You can only choose from THOR clusters within the current environment.
Logical File	The logical name for the copied file.
File Mask	Automatically updated based on logical file name entered.
Options:	
Replicate	Check this box to create backup copies of all file parts exist in the backup directory (by convention on the second drive of the node following in the cluster).
Wrap	Check this box to keep the number of parts the same and wrap if the target cluster is smaller that the original.
Overwrite	Check this box to overwrite files of the same name.
Compress	Check this box to compress the files.
Retain Superfile Structure	Check this box to retain the superfile structure.

- Press the **Submit** button.

  The **DFU Workunit** displays.

- Press the **Refresh** button periodically until the status of your request indicates it is **Finished** or click on the **View Progress** hyperlink to see a progress indicator.

Remote Copy

Remote Copy allows you to copy data from a Thor cluster outside your environment to the one in your environment.

- Click on the **Remote Copy** hyperlink below **DFU** in the menu on the left.

  The **Copy File** page displays.

- Fill in **Source, Destination,** and **Options** information.

Source:	
Logical File	The logical file name in the remote environment.
Source Dali	The Dali Server in the remote environment
Source Username	A valid user in the remote environment
Source Password	The password for the user in the remote environment
Destination:	
Group	Select the name of THOR cluster to copy to.
Note	You can only choose from THOR clusters within the current environment.
Logical Name	The logical name for the copied file.
Options:	
Replicate	Check this box to create backup copies of all file parts exist in the backup directory (by convention on the d: drive of the node following in the cluster).
Wrap	Check this box to keep the number of parts the same and wrap if the target cluster is smaller that the original.

Overwrite	Check this box to overwrite files of the same name.
Compress	Check this box to compress the files.
Retain Superfile Structure	Check this box to retain the superfile structure.

- Press the **Submit** button.

  The **DFU Workunit** displays.

- Press the **Refresh** button periodically until the status of your request indicates it is **Finished** or click on the **View Progress** hyperlink to see a progress indicator.

Modifying a DFU Workunit

From the DFU Workunit page, you can modify and run any DFU Spray, Despray, Copy, or Remote Copy action using the modified settings. This allows you to run similar jobs without filling in similar details again. It also allows you to correct any errors that might have caused a DFU workunit to fail.

From any DFU Workunit page:

- Press the Modify button.

  The page for the original action displays.

- Modify the details, as needed.

- Press the **Submit** button.

  The **DFU Workunit** displays.

- Press the **Refresh** button periodically until the status of your request indicates it is **Finished** or click on the **View Progress** hyperlink to see a progress indicator.

HPCC Data Backups

Introduction

This section covers critical system data that requires regular backup procedures to prevent data loss.

There are

- The System Data Store (Dali data)
- Environment Configuration files
- Data Refinery (Thor) data files
- Rapid Data Delivery Engine (Roxie) data files
- Attribute Repositories
- Landing Zone files

Dali data

The Dali Server data is typically mirrored to its backup node. This location is specified in the environment configuration file using the Configuration Manager.

Since the data is written simultaneously to both nodes, there is no need for a manual backup procedure.

Environment Configuration files

There is only one active environment file, but you may have many alternative configurations.

Configuration manager only works on files in the /etc/HPCCSystems/source/ folder. To make a configuration active, it is copied to /etc/HPCCSystems/environment.xml on all nodes.

Configuration Manager automatically creates backup copies in the /etc/HPCCSystems/source/backup/ folder.

Thor data files

Thor clusters are normally configured to automatically replicate data to a secondary location known as the mirror location. Usually, this is on the second drive of the subsequent node.

If the data is not found at the primary location (for example, due to drive failure or because a node has been swapped out), it looks in the mirror directory to read the data. Any writes go to the primary and then to the mirror. This provides continual redundancy and a quick means to restore a system after a node swap.

A Thor data backup should be performed on a regularly scheduled basis and on-demand after a node swap.

Manual backup

To run a backup manually, follow these steps:

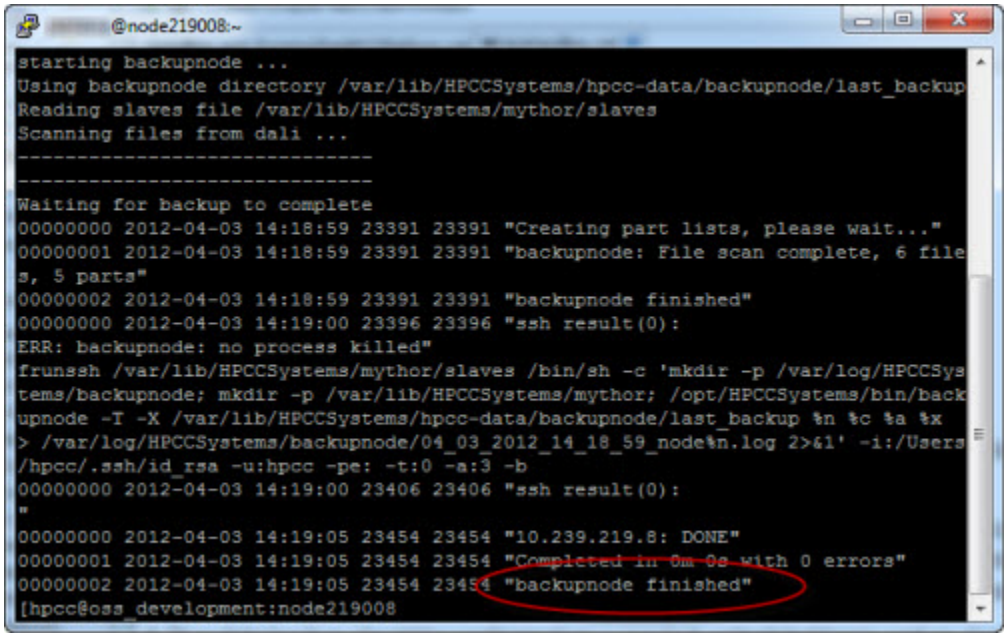
1. Login to the Thor Master node.

   If you don't know which node is your Thor Master node, you can look it up using ECL Watch.

2. Run this command:

   ```
   sudo su hpcc
   /opt/HPCCSystems/bin/start_backupnode <thor_cluster_name>
   ```

   This starts the backup process.



   Wait until completion. It will say "backupnode finished" as shown above.

3. Run the XREF utility in ECL Watch to verify that there are no orphan files or lost files.

Scheduled backup

The easiest way to schedule the backup process is to create a cron job. Cron is a daemon that serves as a task scheduler.

Cron tab (short for CRON TABle) is a text file that contains a the task list. To edit with the default editor, use the command:

```
sudo crontab -e
```

Here is a sample cron tab entry:

```
30 23 * * * /opt/HPCCSystems/bin/start_backupnode mythor
```

30 represents the minute of the hour.

23 represents the hour of the day

The asterisks (*) represent every day, month, and weekday.

mythor is the clustername

To list the tasks scheduled, use the command:

```
sudo crontab -l
```

Roxie data files

Roxie data is protected by three forms of redundancy:

- Original Source Data File Retention: When a query is deployed, the data is typically copied from a Thor cluster's hard drives. Therefore, the Thor data can serve as backup, provided it is not removed or altered on Thor. Thor data is typically retained for a period of time sufficient to serve as a backup copy.
- Peer-Node Redundancy: Each Slave node typically has one or more peer nodes within its cluster. Each peer stores a copy of data files it will read.
- Sibling Cluster Redundancy: Although not required, Roxie deployments may run multiple identically-configured Roxie clusters. When two clusters are deployed for Production each node has an identical twin in terms of data and queries stored on the node in the other cluster.

This provides multiple redundant copies of data files.

Attribute Repositories

Attribute repositories are stored on ECL developer's local hard drives. They can contain a significant number of hours of work and therefore should be regularly backed up. In addition, we suggest using some form of source version control, too.

Landing Zone files

Landing Zones contain raw data for input. They can also contain output files. Depending on the size or complexity of these files, you may want to retain copies for redundancy.