HPCC Systems® Data Handling

Boca Raton Documentation Team

HPCC Systems® Data Handling

Boca Raton Documentation Team
Copyright © 2026 HPCC Systems®. All rights reserved

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.

LexisNexis and the Knowledge Burst logo are registered trademarks of Reed Elsevier Properties Inc., used under license.

HPCC Systems® is a registered trademark of LexisNexis Risk Data Management Inc.

Other products, logos, and services may be trademarks or registered trademarks of their respective companies.

All names and example data used in this manual are fictitious. Any similarity to actual persons, living or dead, is purely coincidental.

2026 Version 10.2.16-1

HPCC Systems® Data Handling 4
Introduction 4
Data Handling Terms 5
File Movement 6
Working with data files 7
Data Handling Methods 9
Data Handling Using ECL Watch 12
HPCC Systems® Data Backups 24
Introduction 24
Dali data 25
Environment Configuration files 26
Thor data files 27
Roxie data files 29
Attribute Repositories 30
Landing Zone files 31

© 2026 HPCC Systems®. All rights reserved

HPCC Systems® Data Handling

Introduction

There are a number of different ways in which data may be transferred to, from, or within an HPCC Systems cluster. 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)

© 2026 HPCC Systems®. All rights reserved

Data Handling Terms

A *spray* or *import* is the relocation of a data file from one location (such as a Landing Zone) to a 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 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 cluster to another cluster within the same environment.

A *Remote copy* is the replication of a data file from one 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.

© 2026 HPCC Systems®. All rights reserved

File Movement

The ESP file service can request whether an import (spray), despray, or copy command is a **push** or a **pull** command.

These processes run as DFU workunits and communicate with service components, either running as dynamically launched ftslave processes (the default behavior on Bare-Metal deployments), or as tasks running on a dafilesrv service (the only option on a Containerized deployment).

- **pull** means that the commands will notionally run per target part and pull from the source.
- **push** means the commands will notionally run per source part and push to the target.

In Bare-Metal deployments, where the compute is coupled to the data, ftslave processes are launched on either the source or the target nodes depending on whether pushing or pulling.

In a Containerized system, the commands always run on a dedicated dafilesrv spray-service where neither the source nor the target are local.

The default setting in a bare-metal system is useFtSlave=true, an option of dfuserver, but can be configured off so Bare-Metal will use dafilesrv as a service, too.

It is generally recommended to leave the decision to whether to pull or push up to the platform. In Bare-Metal, it may decide that it has to push because running ftslave on the target side is not supported such as when the the Landing Zone is a Windows server. It may also choose to pull when it concludes it is more efficient to do so. In a Containerized system, some target planes may not support multiple processes writing concurrently, and therefore pull is forced. If you implicitly request pull or push and the platform cannot satisfy the request, it will warn and continue with the method that is supported.

In addition, the **noCommon** option is available. The noCommon option prevents the job/dfuserver from "commoning up" the individual partition commands into a single request. Commoning up means a single ftslave or service command will process multiple partitions sequentially. Not commoning up, means multiple concurrent partitions, that would otherwise be grouped, will operate in parallel as individual ftslaves or individual spray service commands.

In Bare-Metal, noCommon is on by default, and will mean that any partition command that is destined for the same host, that would otherwise be grouped into a single command, will be left as multiple concurrent commands. For example, if spraying a file to a Thor cluster in Bare-Metal, and the Thor group has several partitions per node (because slavesPerNode>1), then with the default noCommon=true and pull mode, each of the multiple partition pull commands for each node would run in parallel.

In Containerized deployments, noCommon is currently not supported and will be ignored.

© 2026 HPCC Systems®. All rights reserved

Working with data files

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

Before you begin

First, you should consider the size of the data and the capacity of your system. A typical production HPCC Systems platform 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

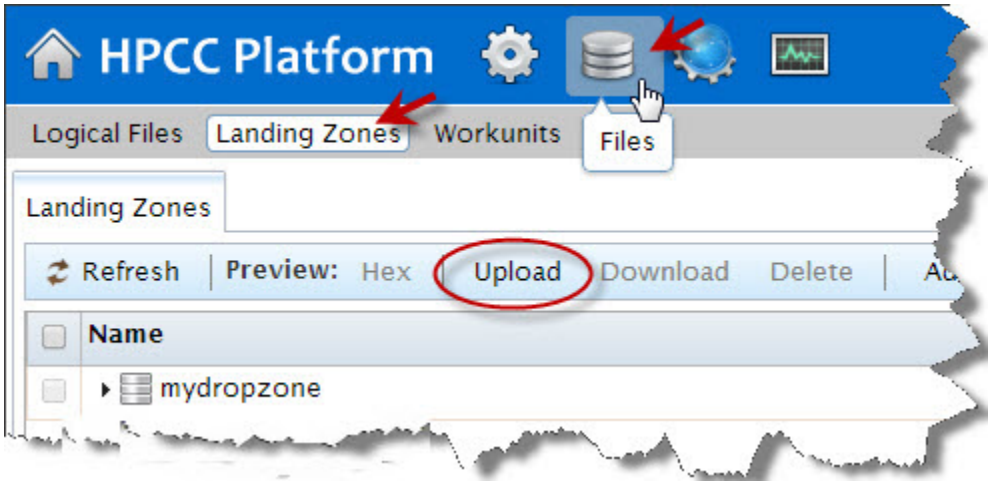
For smaller data files, maximum of 2GB, you can use the upload/download file utility in ECL Watch.

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 the ECL Watch home page, press the **Files** icon link at the top, then press the **Landing Zones** button from the sub-menu, and then click on the **Upload** action button as illustrated in the following image.

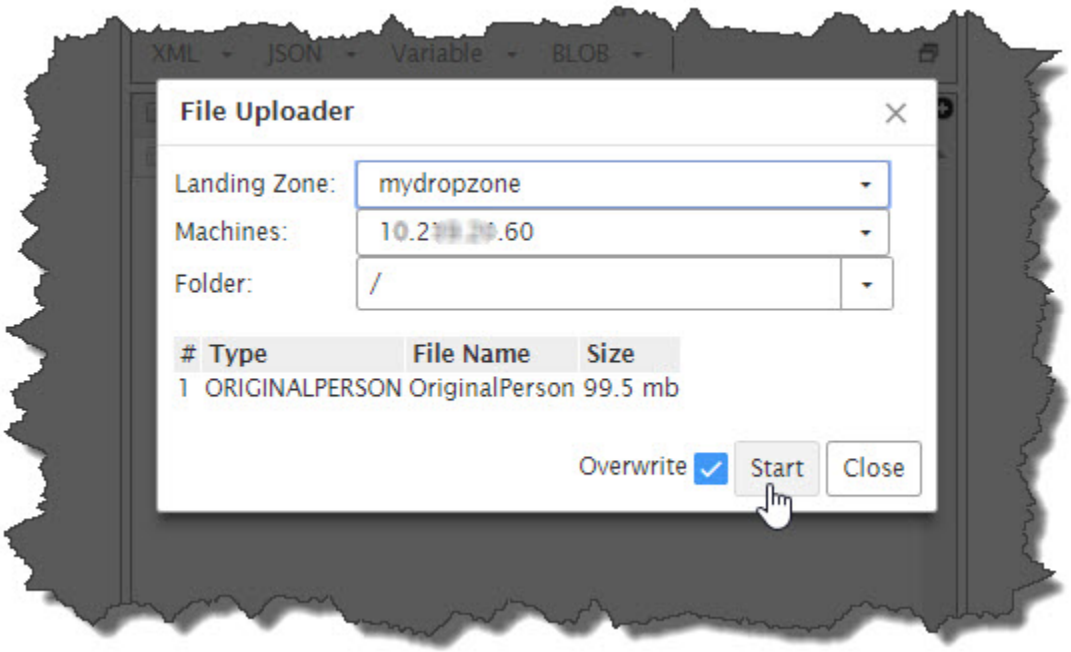
Figure 1. Upload/download



A system file dialog opens. Browse and select the file to upload and then press the **Open** button.

© 2026 HPCC Systems®. All rights reserved

Figure 2. Dropzones



3. Once you choose the file to upload, it will take you to the **File Uploader** dialog, where you can choose the landing zone, machine, and/or folder for the file to upload.

4. Press the **Start** button to complete the file upload.

Uploading files with a Secure Copy Client

To upload a large file for processing to your system, you will need a tool that supports the secure copy protocol (SCP). There are several free tools available, and the steps are quite similar.

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

Login ID:	hpccdemo
Password:	hpccdemo

   Ensure you are using the secure protocol, SFTP, connected to your Landing Zone's IP or hostname through port 22 in your SCP tool.

2. Once logged in, set the destination to the landing zone folder. (/var/lib/HPCCSystems/mydropzone)

3. Navigate to the location where your local file is.

4. Select the data file to send and copy it to your landing zone.

© 2026 HPCC Systems®. All rights reserved

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.

© 2026 HPCC Systems®. All rights reserved

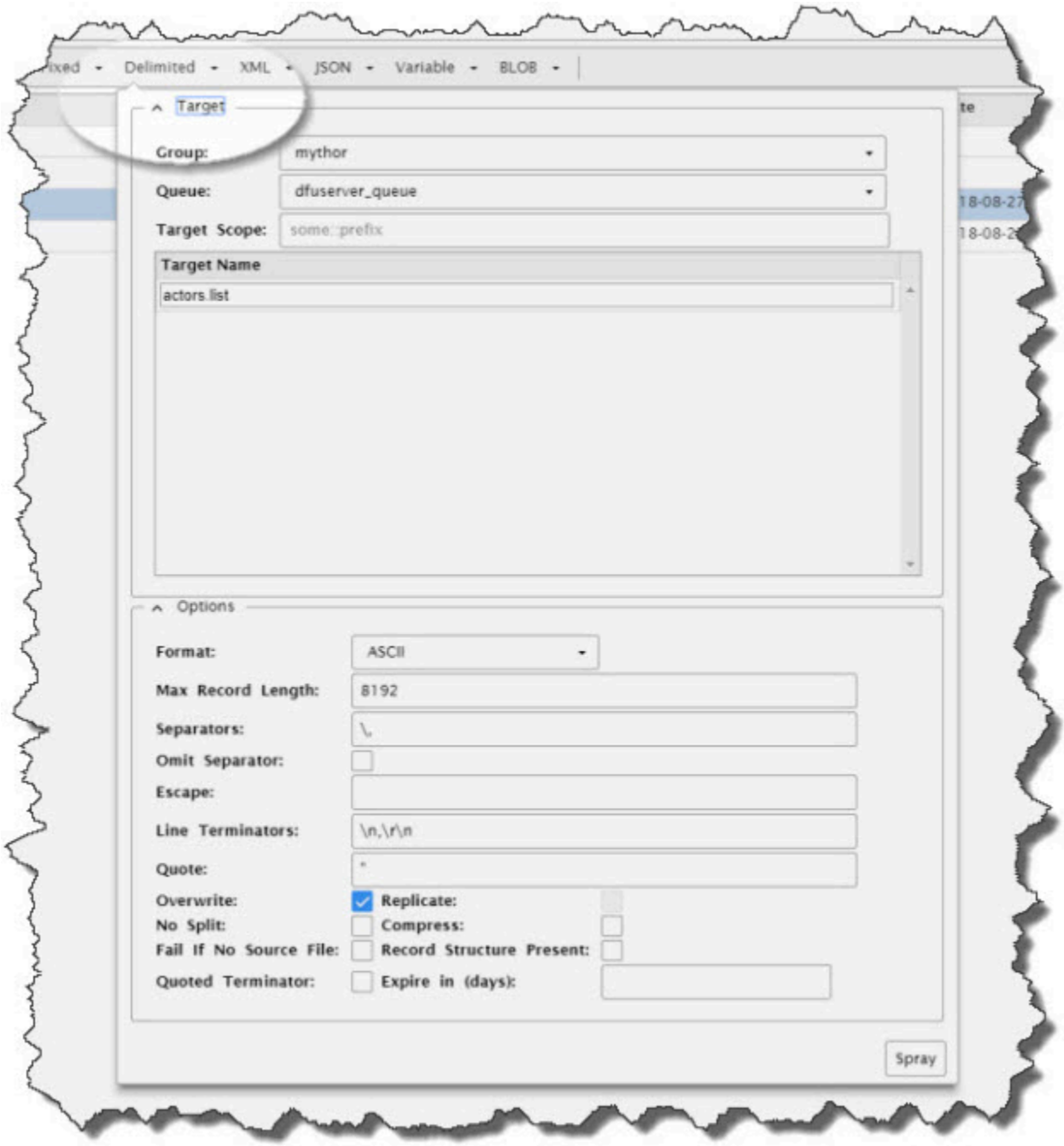
Spray Files

One way you can spray files to your clusters is from the **Landing Zone** page in ECL Watch.

1. Select the file from your drop zone by checking the box next to it.
2. Select the appropriate drop menu option for the type of spray you want.

   For example, to spray a delimited file, select the **Delimited** action button.

Figure 3. Landing Zone Spray



© 2026 HPCC Systems®. All rights reserved

3. Fill in the values as appropriate for the spray.

4. Press the **Spray** button to spray the file(s).

Spray multiple files

You can choose to spray multiple files with the multi-file spray feature. This is useful for spraying a number of files of the same type using the same spray options.

Fixed (length) files can have different record lengths and XML files can have different row tags which must be specified individually for each file. To specify these differences select the files you want to spray and the spray type. You will then see the files listed. Enter the record length or row tag information for each file if using the Fixed or XML spray types, then check all other applicable options and Spray.

© 2026 HPCC Systems®. All rights reserved

Data Handling Using ECL Watch

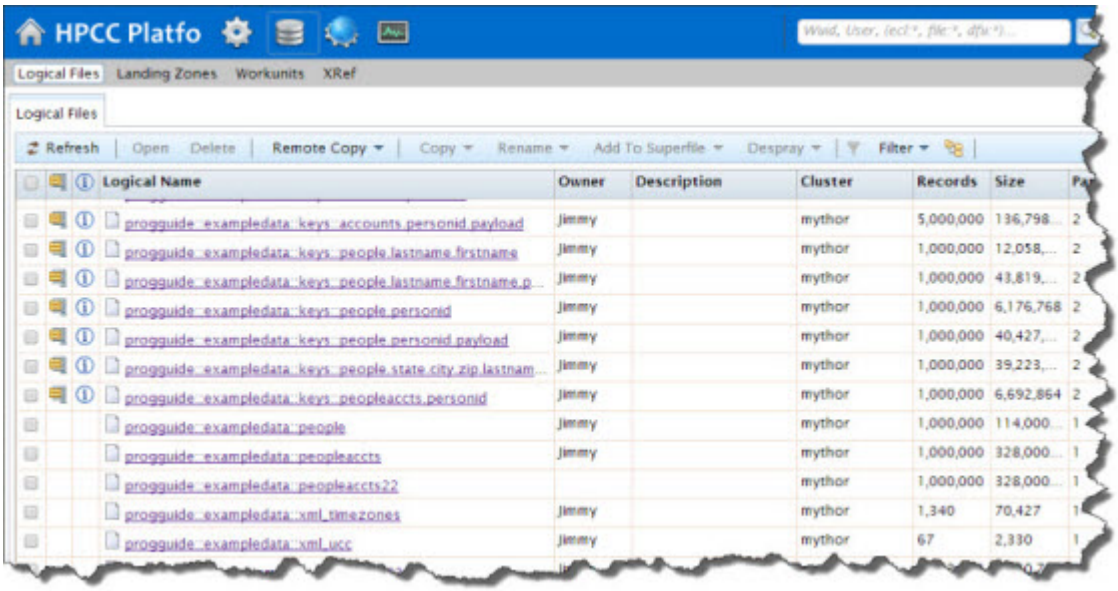
- Login to ECL Watch for the environment.

  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 **Files** icon, then click on **Logical Files**.

  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.

© 2026 HPCC Systems®. All rights reserved

Spray Fixed

- Click on the **Files** icon, then click the **Landing Zones** link from the navigation sub-menu.
- Click on the arrow next to your dropzone to expand the list.

  The files on your drop zone display.
- Check the checkboxes for the file(s) you want to spray, then press the Spray: **Fixed** action button.

  The **Spray Fixed** dialog displays.
- Fill in relevant details:

Target	
Group	Select the name of cluster to spray to. You can only select a cluster in your environment.
Queue	Select the queue for the spray.
Target Scope	The prefix for the logical file
Target Name	The logical filename to create. This is pre-filled with the name of the source file on the landing zone, but can be changed.
Record Length	The size of each record.
Options:	
Overwrite	Check this box to overwrite files of the same name.
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
Compress	Check this box to compress the files.
Expire in (days)	An integer value indicating the number of days before automatically removing the file. If omitted, the default is -1 (never expires).
No Split	Check this box to prevent splitting file parts to multiple target parts.
Fail if no source file	Check this box to allow the spray to fail if no source file is found.

- Press the **Spray** button.

  A **DFU Workunit** tab displays for each job. You can see the progress of each spray operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Spray Delimited

- Click on the **Files** icon, then click the **Landing Zones** link from the navigation sub-menu.
- Click on the arrow next to your dropzone to expand the list.

  The files on your drop zone display.
- Check the checkboxes for the file(s) you want to spray, then press the Spray: **Delimited** action button.

  The **Spray Delimited** page displays.
- Fill in relevant details:

Target	
Group	Select the name of cluster to spray to. You can only select a cluster in your environment.
Queue	Select the queue for the spray.
Target Scope	The prefix for the logical file
Target Name	The logical filename to create. This is pre-filled with the name of the source file on the landing zone, but can be changed.
Options:	
Format	Select the format from the droplist
Max Record Length	The length of longest record in the file.
Separators	The character(s) used as a separator in the source file.
Omit Separator	Check this box to omit the separator.
Escape	A null-terminated string containing the CSV escape characters.
Line Terminators	The character(s) used as a line terminators in the source file.
Quote	The character used as a quote in the source file.
Overwrite	Check this box to overwrite files of the same name.
No Split	Check this box to prevent splitting file parts to multiple target parts.
Fail if no source file	Check this box to allow the spray to fail if no source file is found.
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
Compress	Check this box to compress the files.
Expire in (days)	An integer value indicating the number of days before automatically removing the file. If omitted, the default is -1 (never expires).
Quoted Terminator	Check this box to indicate that the terminator character can be included in a quoted field. If unchecked, it allows quicker partitioning of the file (avoiding a complete file scan).
Record Structure Present	Flag indicating whether to derive the record structure from the header of the file.

- Press the **Spray** button.

© 2026 HPCC Systems®. All rights reserved

A **DFU Workunit** tab displays for each job. You can see the progress of each spray operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Spray XML

- Click on the **Files** icon, then click the **Landing Zones** link from the navigation sub-menu.
- Click on the arrow next to your dropzone to expand the list.

  The files on your drop zone display.
- Check the checkboxes for the file(s) you want to spray, then press the Spray: **XML**action button.

  The **Spray XML** dialog displays.
- Fill in relevant details:

Target	
Group	Select the name of cluster to spray to. You can only select a cluster in your environment.
Queue	Select the queue for the spray.
Target Scope	The prefix for the logical file
Target Name	The logical filename to create. This is pre-filled with the name of the source file on the landing zone, but can be changed.
Row Tag	The tag name of the row delimiter. Required.
Options:	
Format	Select the format from the droplist
Max Record Length	The length of longest record in the file.
Overwrite	Check this box to overwrite files of the same name.
No Split	Check this box to prevent splitting file parts to multiple target parts.
Expire in (days)	An integer value indicating the number of days before automatically removing the file. If omitted, the default is -1 (never expires).
Fail if no source file	Check this box to allow the spray to fail if no source file is found.
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
Compress	Check this box to compress the files.

- Press the **Spray** button.

  A **DFU Workunit** tab displays for each job. You can see the progress of each spray operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Spray JSON

- Click on the **Files** icon, then click the **Landing Zones** link from the navigation sub-menu.
- Click on the arrow next to your dropzone to expand the list.

  The files on your drop zone display.
- Check the checkboxes for the file(s) you want to spray, then press the Spray: **JSON** action button.

  The dialog displays.
- Fill in relevant details:

Target	
Group	Select the name of cluster to spray to. You can only select a cluster in your environment.
Queue	Select the queue for the spray.
Target Scope	The prefix for the logical file
Target Name	The logical filename to create. This is pre-filled with the name of the source file on the landing zone, but can be changed.
Row Path	The path specifier to the JSON content. The default takes the root level content as an array of objects to be treated as rows.
Options:	
Format	Select the format from the droplist
Max Record Length	The length of longest record in the file.
Overwrite	Check this box to overwrite files of the same name.
No Split	Check this box to prevent splitting file parts to multiple target parts.
Fail if no source file	Check this box to allow the spray to fail if no source file is found.
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
Expire in (days)	An integer value indicating the number of days before automatically removing the file. If omitted, the default is -1 (never expires).
Compress	Check this box to compress the files.

- Press the **Spray** button.

  A **DFU Workunit** tab displays for each job. You can see the progress of each spray operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Spray Variable

- Click on the **Files** icon, then click the **Landing Zones** link on the navigation sub-menu.
- Click on the arrow next to your dropzone to expand the list.

  The files on your drop zone display.
- Check the checkboxes for the file(s) you want to spray, then press the Spray: **Variable** action button.

  The Spray **Variable** dialog displays.
- Fill in relevant details:

Target	
Group	Select the name of cluster to spray to. You can only select a cluster in your environment.
Queue	Select the queue for the spray.
Target Scope	The prefix for the logical file
Target Name	The logical filename to create. This is pre-filled with the name of the source file on the landing zone, but can be changed.
Options:	
Source Type	Select the source type from the drop list. Values: recfmv, recfmvb, Variable, or Variable Big-endian.
Overwrite	Check this box to overwrite files of the same name.
No Split	Check this box to prevent splitting file parts to multiple target parts.
Fail if no source file	Check this box to allow the spray to fail if no source file is found.
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
Expire in (days)	An integer value indicating the number of days before automatically removing the file. If omitted, the default is -1 (never expires).
Compress	Check this box to compress the files.

- Press the **Spray** button.

  A **DFU Workunit** tab displays for each job. You can see the progress of each spray operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Spray Blob

- Click on the **Files** icon, then click the **Landing Zones** link on the navigation sub-menu.
- Click on the arrow next to your dropzone to expand the list.

  The files on your drop zone display.
- Check the checkboxes for the file(s) you want to spray, then press the Spray: **BLOB** action button.

  The Spray **BLOB** dialog displays.
- Fill in relevant details:

Target	
Group	Select the name of cluster to spray to. You can only select a cluster in your environment.
Queue	Select the queue for the spray.
Target Name	The logical target name to create. Required. You must provide a target name.
Source Path	The path to the file. This is pre-filled with the name of the selected source file(s) on the landing zone, but can be changed. Supports wildcards.
Options:	
Blob Prefix	The prefix for the file.
Overwrite	Check this box to overwrite files of the same name.
No Split	Check this box to prevent splitting file parts to multiple target parts.
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
Compress	Check this box to compress the files.
Expire in (days)	An integer value indicating the number of days before automatically removing the file. If omitted, the default is -1 (never expires).
Fail if no source file	Check this box to allow the spray to fail if no source file is found.

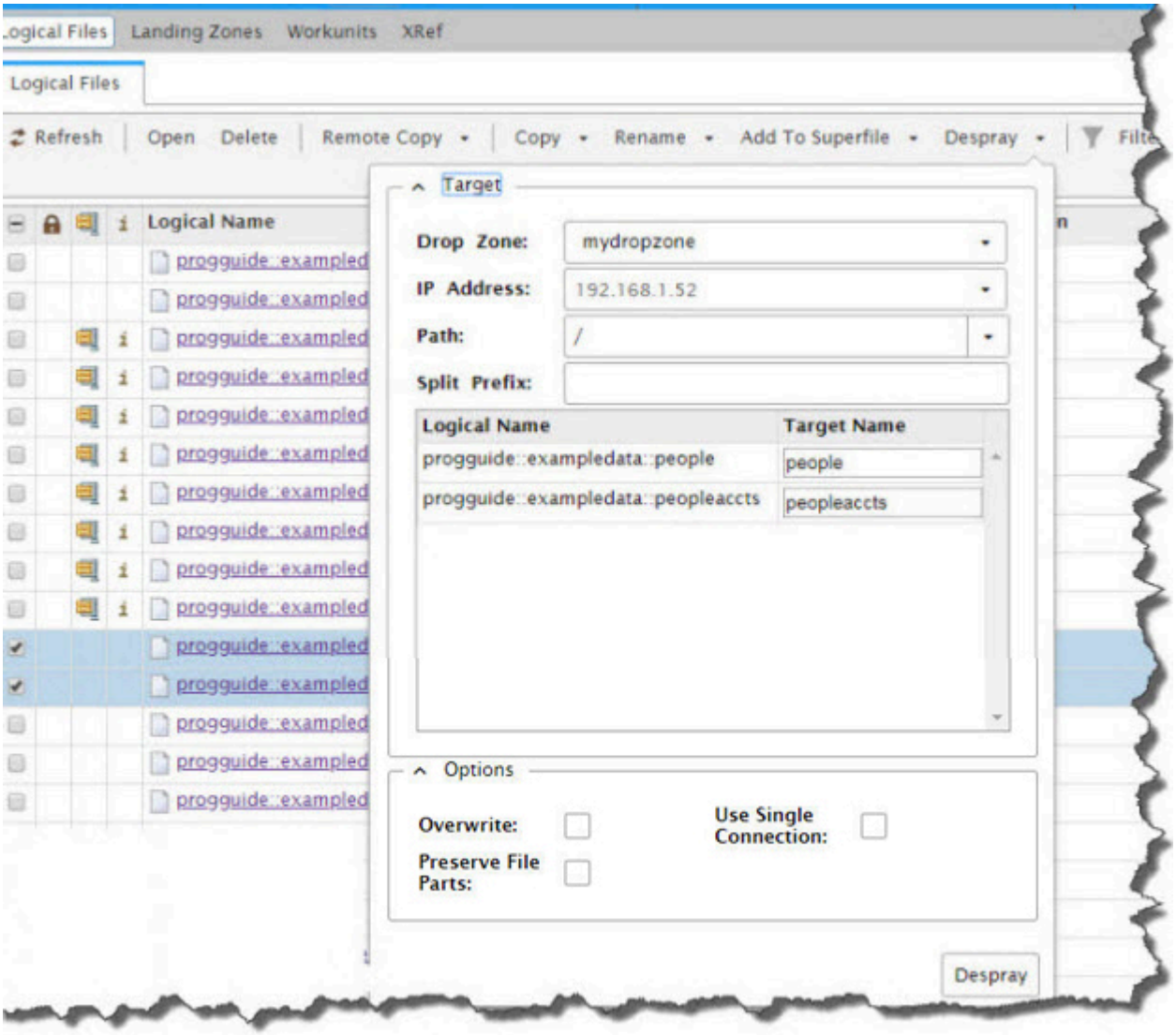
- Press the **Spray** button.

  A **DFU Workunit** tab displays for each job. You can see the progress of each spray operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Desprays

- Locate the file(s) to despray in the list of files, then Press the the **Despray** action button.



- Provide **Destination** information.

Drop Zone	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.
IP Address	This is prefilled based upon the selected machine.
Path	Provide the complete file path of the destination including file name and extention.
Split Prefix	Prefix
Overwrite	Check this box to overwrite a file with the same name if it exists.
Use Single Connection	Check this box to use a single network connection to despray.

© 2026 HPCC Systems®. All rights reserved

Preserve File Parts	Check this box to preserve the original file parts and write multiple files to the landing zone.

- Press the **Despray** button.

  A DFU Workunit tab for each job opens. You can see the progress of each despray operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Copy

- Click on the **Files** icon, then click the Logical Files button on the navigation bar.
- Select the file(s) to copy in the list of files, then click on the **Copy** action button.
- Fill in **Destination** and **Options** information.

Target:	
Group	Select the name of cluster to copy to. You can only select a cluster in your environment.
Target Name	The logical filename to create. This is pre-filled with the name of the source file on the landing zone, but can be changed.
Options:	
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
Wrap	Check this box to keep the number of parts the same and wrap if the target cluster is smaller that the original.
No Split	Check this box to prevent splitting file parts to multiple target parts.
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.
Preserve Compression	Check this box to preserve the compression of the original file when copying

- Press the **Copy** button.

  A **DFU Workunit** tab displays for each job. You can see the progress of each copy operation on the tab. If a job fails, information related to the cause of the failure also displays.

© 2026 HPCC Systems®. All rights reserved

Remote Copy

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

- Click on the **Files** icon, then click the Logical Files button on the navigation bar.
- Click on the **Remote Copy** link

  The **Copy File** page displays.
- Fill in **Source, Destination,** and **Options** information.

Source:	
Dali	The Dali Server in the remote environment
User ID	The Username to use to authenticate on the Remote environment (if needed)
Password	The password to use to authenticate on the Remote environment (if needed)
Logical File	The logical filename in the remote environment.
Destination:	
Group	Select the name of cluster to copy to. You can only select a cluster in your environment.
Logical Name	The logical name for the copied file.
Options:	
Replicate	Check this box to create backup copies of all file parts in the backup directory (by convention on the secondary drive of the node following in the cluster). **This option is only available on systems where replication has been enabled.**
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.
No Split	Check this box to prevent splitting file parts to multiple target parts.
Retain Superfile Structure	Check this box to retain the superfile structure.

- Press the **Submit** button.

  A **DFU Workunit** tab displays. You can see the progress of the copy operation on the tab. If a job fails, information related to the cause of the failure also 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.

© 2026 HPCC Systems®. All rights reserved

HPCC Systems® 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

© 2026 HPCC Systems®. All rights reserved

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.

© 2026 HPCC Systems®. All rights reserved

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.

© 2026 HPCC Systems®. All rights reserved

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:

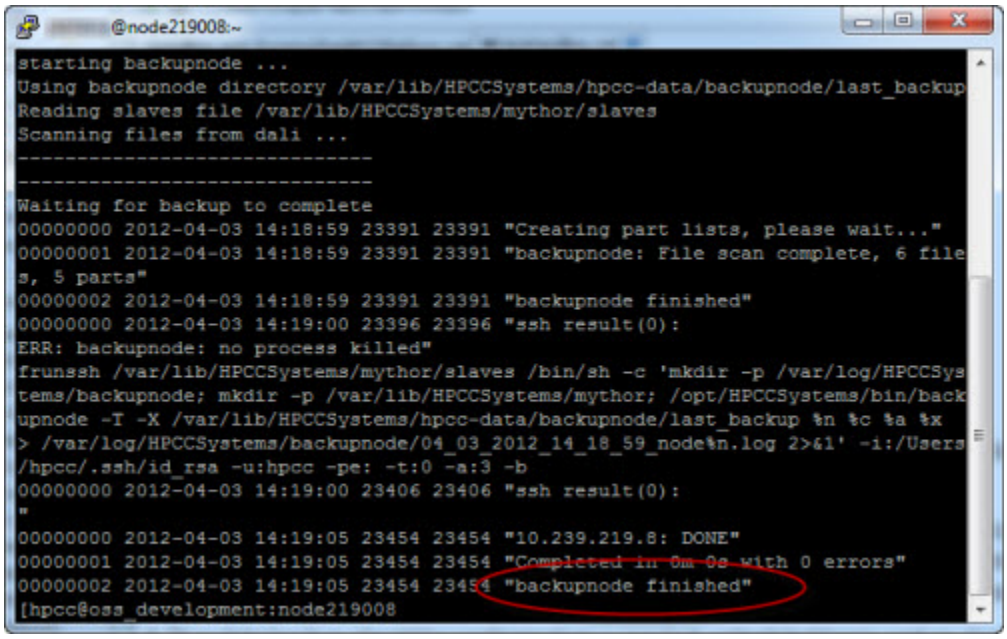
1. Login to the Thor Master node.

   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.

© 2026 HPCC Systems®. All rights reserved

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 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
```

© 2026 HPCC Systems®. All rights reserved

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.

© 2026 HPCC Systems®. All rights reserved

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.

© 2026 HPCC Systems®. All rights reserved

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.

© 2026 HPCC Systems®. All rights reserved