

HPCC Systems™

One-Click Thor™ within Amazon Web Services

Boca Raton Documentation Team



One-Click Thor™ within AWS

Boca Raton Documentation Team

Copyright © 2012 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.

Amazon Web Services, AWS, Amazon EC2, EC2, Amazon Elastic Compute Cloud, Amazon S3, Amazon Simple Storage Service, are trademarks, registered trademarks or trade dress of AWS in the U.S. and/or other countries.

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.

February 2012 Version 3.4.2 **BETA**

Introduction	4
Prerequisites and Assumptions	5
Using One-Click Launch	6
Before you begin	6
Login	7
Launch a Thor Cluster	9
Running ECL	18
Running ECL on your HPCC Systems Thor cluster	18
More ECL Examples	25
ECL Example: Anagram1	25
Anagram2	28
Data Handling	35
Using S3 buckets	35
Next Steps	37

Introduction

This guide provides details and guidance in running an HPCC Systems Thor cluster inside an Amazon Web Services (AWS) Elastic Cloud (EC2) using the One-Click Launch page.

This allows you to **instantiate** and run HPCC Systems Thor clusters of different sizes on the fly.

This is useful for:

- Proof-of-concept
- Experimentation
- Learning
- Leveraging the HPCC Systems Thor platform without incurring cost of hardware and administration
- Create and use an HPCC Systems Thor cluster immediately without purchasing and installing new hardware

You can create a small cluster for small tasks or larger clusters for larger jobs. This flexibility allows you to match cost and processing power to the job at hand.

Instantiating temporary EC2 nodes allows you to "rent" computing capacity without long term commitments. In this manner, you pay as you go instead of incurring large fixed costs at the start.

Keep in mind that you should terminate any unneeded instances to avoid paying for computing time you don't need. You are solely responsible for all AWS charges.

We are working with Amazon to improve the set up processes. **In the near future, we will have a guide to show how to run an HPCC Systems Thor cluster in an Amazon Elastic Map Reduce (EMR) type environment.** We expect that method to be more robust and easier to set up and operate. In addition, HPCC Systems is working on its own Enterprise Cloud offering, but it is not yet ready to be announced.



We suggest **reading** this document in its entirety before beginning.

Prerequisites and Assumptions

You will need:

- An Amazon Web Services account with EC2 enabled
- A workstation with Internet access to access the Amazon Web Services, This can be either a:
 - *Windows PC or*
 - *Linux workstation with the Eclipse IDE and the ECL plug-in for Eclipse (available soon)*
 - *Mac* workstation with the Eclipse IDE and the ECL plug-in for Eclipse (available soon)*
- A Web browser (Firefox, Internet Explorer, or Chrome)

Optionally, you could benefit from having:

- An SSH tool, such as PuTTY
- A key generation and conversion tool, like PuTTYGen
- A secure copy tool (such as, WinSCP)
- Familiarity in navigating Linux file systems



For detailed PuTTY/pcsp/PuTTYGen directions from Amazon, see:

<http://docs.amazonwebservices.com/AmazonEC2/gsg/2006-06-26/putty.html>

*Some required components have not yet been built for the Mac platform (for example, the eclcc compiler). For the time being you can use a Windows emulator or a Virtual Machine on a Mac.

Using One-Click Launch

Before you begin

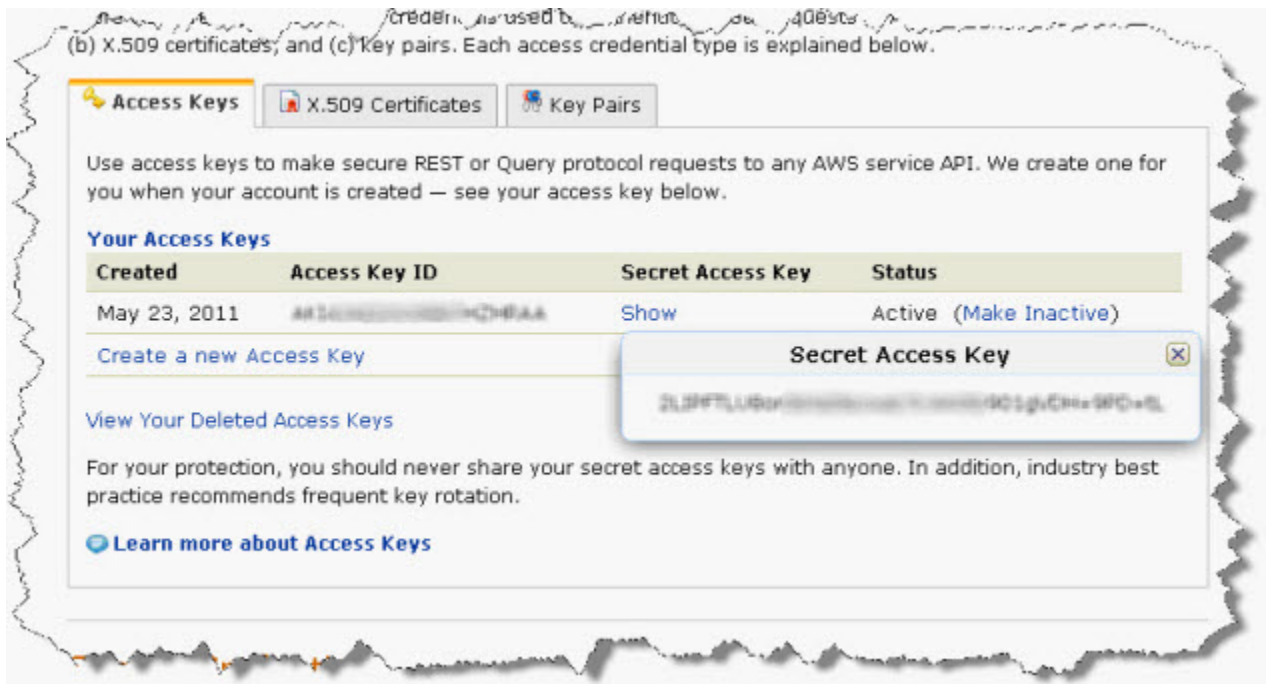
In this section, you will gather some information you need before beginning. This includes:

- Your AWS Access Key ID and Secret Access Key
- The size of the cluster you want.

Find your Amazon Access Key ID and Secret Access Key

1. Go to aws.amazon.com and login, if needed.
2. Select **Account**.
3. Select **Security Credentials**.
4. On the page, look for the section called **Access Credentials**.
5. Note your **Access Key ID** and your **Secret Access Key**.

Figure 1. Credentials



Portions of this image are intentionally blurred

Access Key ID	
Secret Access Key	

Login

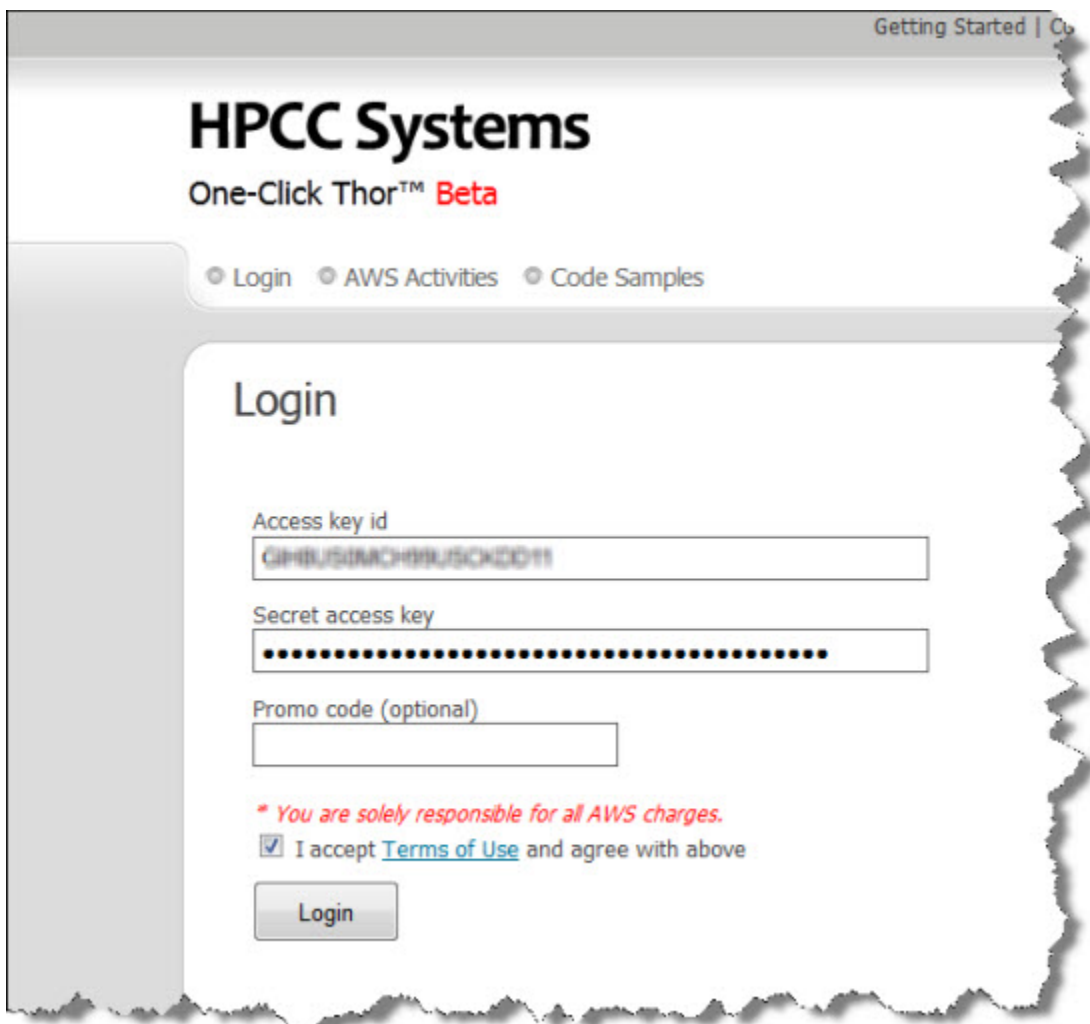
1. Open a browser and go to <https://aws.hpccsystems.com/>.

If not logged in, you will see the **Login** link at the top of the page. If you see a **Logout** link, you are already logged in.

2. Specify your **Access Key ID** and **Secret Access Key**. This information is never stored on our system.

If you don't have it handy, you can click on the **AWS Activities >> Access Keys** link from the top menu to go to that section of the AWS Management Console..

Figure 2. Login



Portions of this image are intentionally blurred

3. Check the box to accept the Terms of Use.
4. Press the **Login** button.

The **View Clusters** window displays. This shows any clusters you have started. From here, you can access the link to launch a new cluster.

5. Click on the **Launch Cluster** link at the top.

The **Launch a New Thor Cluster** window displays.

Launch a Thor Cluster

In this section, you will launch a set of Ubuntu 10.04 machines to use for your HPCC Systems Thor platform. The One-Click page, uses the input you specify to create your cluster for you.

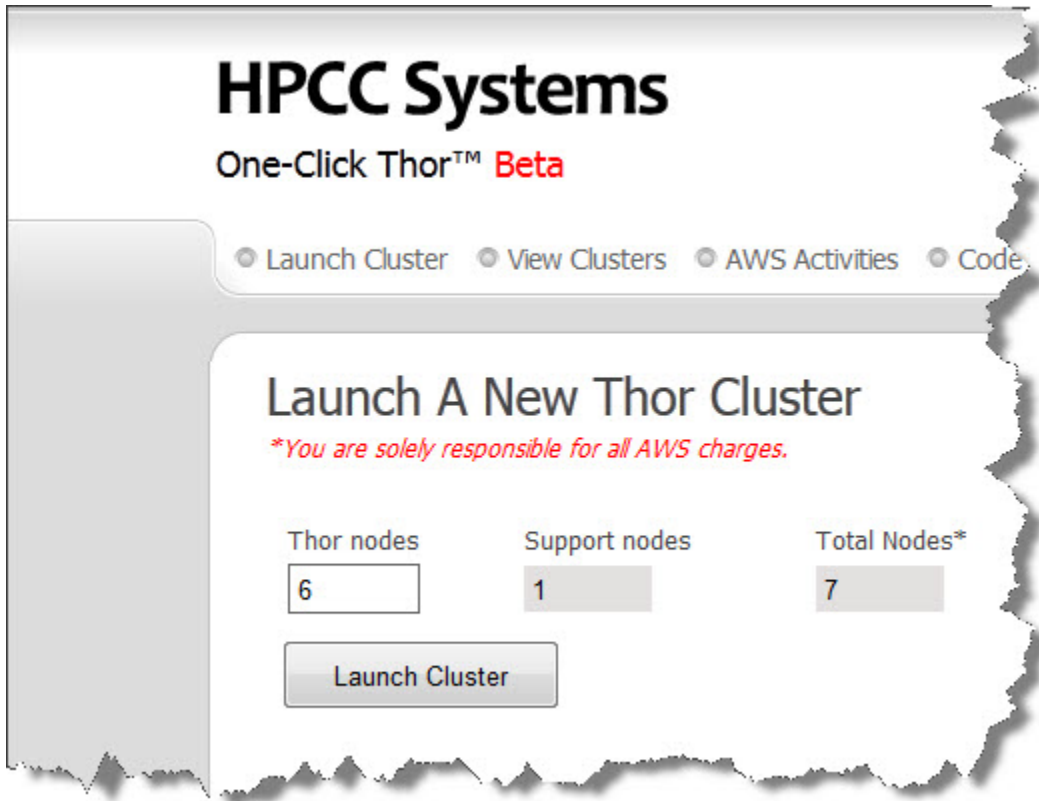
When you press the **Launch Cluster** button, it:

- Creates a unique Cluster Name.
- Creates a Security Group with the access to the TCP and UDP ports enabled.
- Creates a Key Pair.
- Launches the number of m1.large nodes requested using ami-fbbf7892.
- Gathers Private and Public IPs.
- Installs HPCC Systems platform packages (required and optional packages and their dependencies).
- Configures Cluster and Support nodes.
- Creates the internal user (HPCC).
- Propagates the environment.xml file to all nodes.
- Starts up all components.

Launch a New Thor Cluster

1. Specify the number of Thor nodes to instantiate.

Figure 3. Launch a New Thor Cluster



The screenshot shows the HPCC Systems One-Click Thor™ Beta interface. At the top, there are navigation links: Launch Cluster, View Clusters, AWS Activities, and Code. The main heading is "Launch A New Thor Cluster" with a red warning: "*You are solely responsible for all AWS charges.*" Below this, there are three input fields: "Thor nodes" with a value of 6, "Support nodes" with a value of 1, and "Total Nodes*" with a value of 7. A "Launch Cluster" button is positioned below these fields.

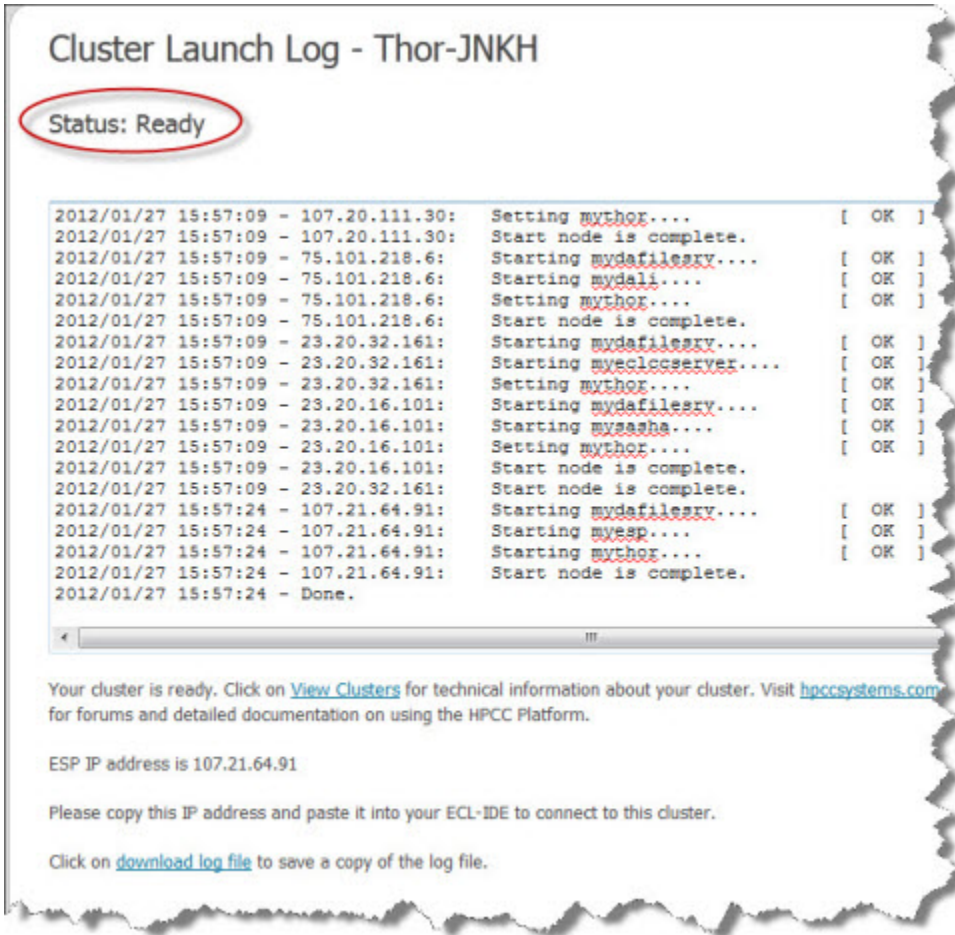
Thor nodes	Support nodes	Total Nodes*
6	1	7

2. Press the **Launch Cluster** button.

The **Cluster Launch Log** window displays. This shows details while it is launching (it auto-refreshes during launch or termination). It also shows your Cluster ID (a unique identifier) which can be useful to identify the cluster when you have more than one running.

3. Wait until the **Cluster Launch Log** says **Status: Ready** to indicate completion of the startup processes.

Figure 4. Launch Cluster



The screenshot displays the 'Cluster Launch Log - Thor-JNKH'. At the top, the status 'Status: Ready' is circled in red. Below this is a scrollable log window showing a series of system startup messages for various nodes, including 'Setting mythor...', 'Start node is complete.', and 'Starting mydfilesrv...'. Each message is followed by a timestamp and an IP address, and ends with '[OK]'. The log concludes with '2012/01/27 15:57:24 - Done.'. Below the log window, there is a message stating 'Your cluster is ready. Click on [View Clusters](#) for technical information about your cluster. Visit hpccsystems.com for forums and detailed documentation on using the HPCC Platform.' This is followed by the text 'ESP IP address is 107.21.64.91' and a request to 'Please copy this IP address and paste it into your ECL-IDE to connect to this cluster.' At the bottom, there is a link to 'Click on [download log file](#) to save a copy of the log file.'

4. Click on the [View Clusters](#) link to see running clusters.

This list has links for the ECL Watch page, the Launch Log page, the Configuration file for the cluster, a list of IPs, and the SSH Key.

It also has a link that allows you to **Terminate** the cluster instantiation.

Figure 5. View Clusters

HPCC Systems
One-Click Thor™ Beta

HPCC Systems Thor Community Edition: 3.4
Front-End Version: 0.1.9
Back-End Version: 0.1.8

Launch Cluster View Clusters AWS Activities Code Samples Log Out

View Clusters

Launch Date	Cluster ID	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Jan. 27, 2012, 11:08 a.m.	Thor-MSRT	4	us-east-1a	ESP	Ready	Log	Config	IPs	Key	Terminate
Jan. 27, 2012, 3:48 p.m.	Thor-JNKH	7	us-east-1d	ESP	Ready	Log	Config	IPs	Key	Terminate

[Getting Started](#) - [Comments](#) - [FAQs](#) - [HPCC AWS Forum](#) - [AWS Access Keys](#) - [AWS Management Console](#)

LexisNexis® Risk Solutions
HPCC Systems
1000 Alderman Dr.
Atlanta, GA 30309
info@hpccsystems.com
US: 1-877-316-9669
Fax: 1-678-415-7200
© 2012 HPCC Systems. All rights reserved.
*By using this service, you agree to the Terms of Service.

Terminating your instances

If you need to save your data, you must to despray it first and save it off of your cluster before shutting down. More information about Data Handling in an HPCC Systems platform are available in the *Data Handling* manual. See the Next Steps section for details on downloading other manuals.

To terminate your cluster:

1. Open the **View Clusters** page using the link at the top of a page.

Figure 6. Running Clusters

HPCC Systems
One-Click Thor™ Beta

HPCC Systems Thor Community Edition: 3.4.1
Front-End Version: 0.1.9
Back-End Version: 0.1.8

Launch Cluster View Clusters AWS Activities Code Samples Log Out

View Clusters

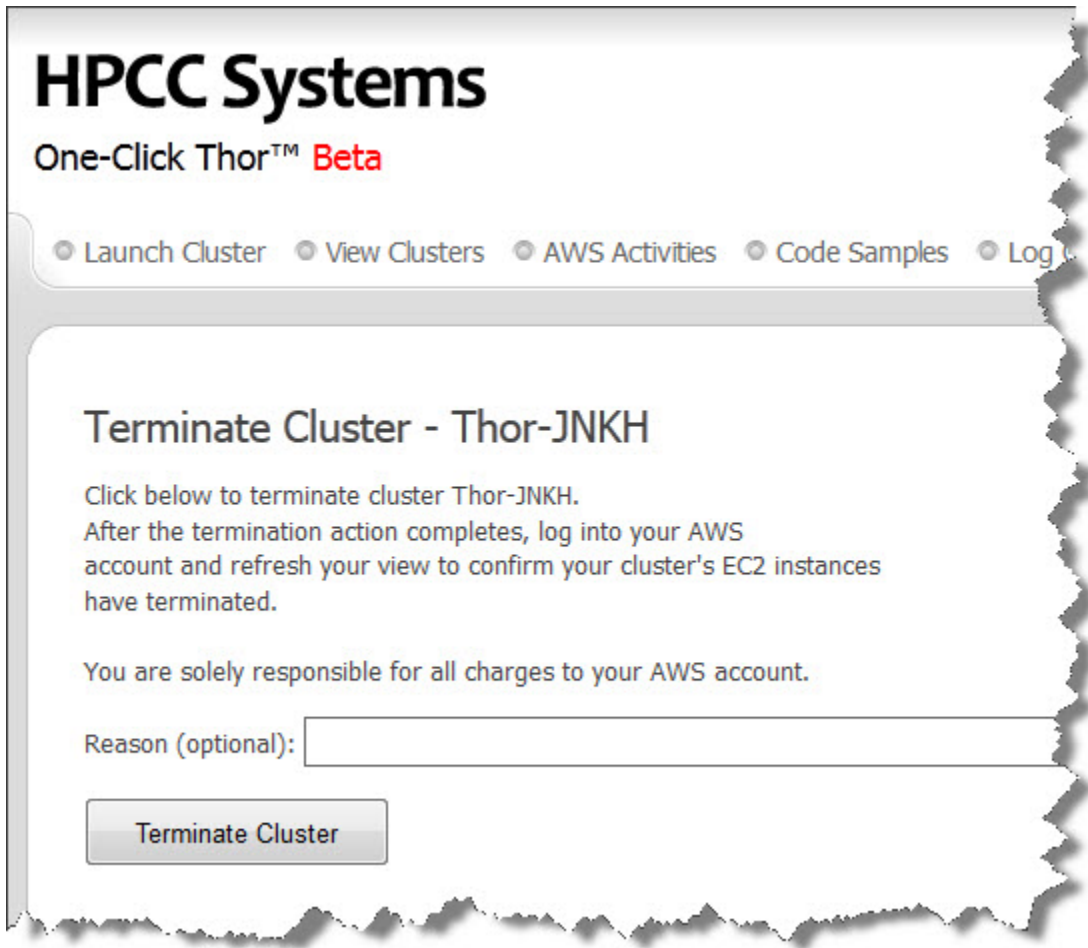
Launch Date	Cluster ID	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Jan. 27, 2012, 11:08 a.m.	Thor-MSRT	4	us-east-1a	ESP	Ready	Log	Config	IPs	Key	Terminate
Jan. 27, 2012, 3:48 p.m.	Thor-JNKH	7	us-east-1d	ESP	Ready	Log	Config	IPs	Key	Terminate

[Getting Started](#) - [Comments](#) - [FAQs](#) - [HPCC AWS Forum](#) - [AWS Access Keys](#) - [AWS Management Console](#)

LexisNexis® Risk Solutions
HPCC Systems
1000 Alderman Dr.
Marietta, GA 30066
info@hpccsystems.com
US: 1.877.316.9669
Tel: 1.678.498.2200
© 2012 HPCC Systems. All rights reserved.
* By using this service, you agree to the terms of service.

2. Click on the **Terminate** link next to the cluster you wish to close.

Figure 7. Terminate Cluster

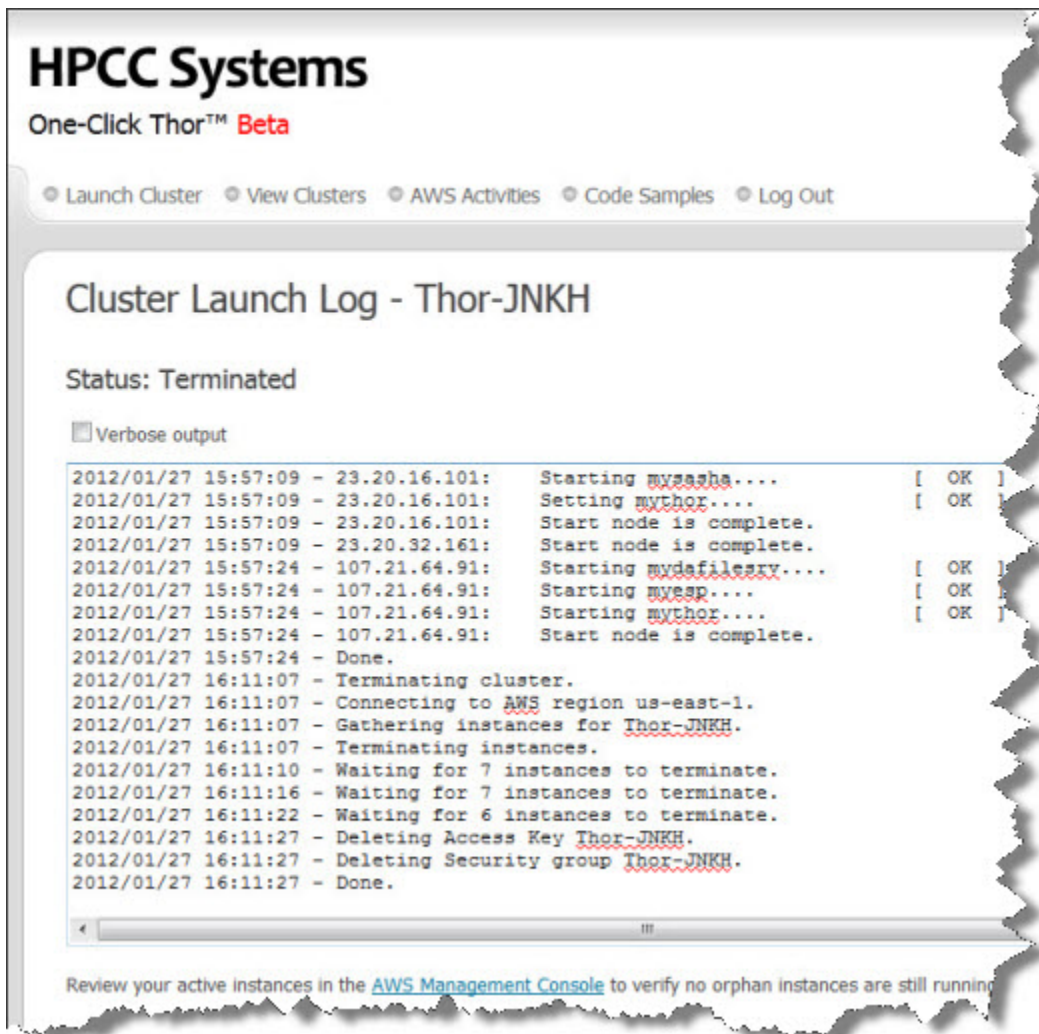


3. Press the **Terminate Cluster** button and confirm when prompted.

The **Launch Log** page displays and shows activity while terminating.

4. Wait until the Cluster Launch Log says **Status: Terminated**.

Figure 8. Terminated Cluster



5. Optionally, go to the AWS management console to confirm your instances have properly terminated.

You are solely responsible for all charges to your AWS account.

Other Tasks

View Clusters

The **View Clusters** page provides access to each cluster's Launch Date/Time, Cluster ID, Number of Nodes, Zone, ECL Watch Page, Status, Launch Log, Config File, IP Addresses, and SSH Key.

It also provides a link to terminate a cluster with a single click.

Figure 9. Running Clusters

Launch Date	Cluster ID	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Jan. 4, 2012, 2:31 p.m.	Thor-GY6D	6	None		Failed	Log			Key	
Jan. 5, 2012, 9:14 a.m.	Thor-SZ5Y	1	us-east-1a	ESP	Ready	Log	Config	IPs	Key	Terminate
Jan. 5, 2012, 9:16 a.m.	Thor-MANC	6	us-east-1a	ESP	Ready	Log	Config	IPs	Key	Terminate

[Comments](#) - [FAQs](#) - [Forum](#) - [AWS Access Keys](#) - [AWS Management Console](#) - [AWS Health Dashboard](#) - [AWS Limit Increase Request](#)

Manage your SSH keys

The SSH Key management page allows you to download your cluster's SSH key (.PEM file) to use to authenticate an SSH session, such as a console session using PuTTY. It also provides a means to delete it from the One-Click system.

1. Open the **Running Clusters** page using the [View Clusters](#) Link at the top of a page.

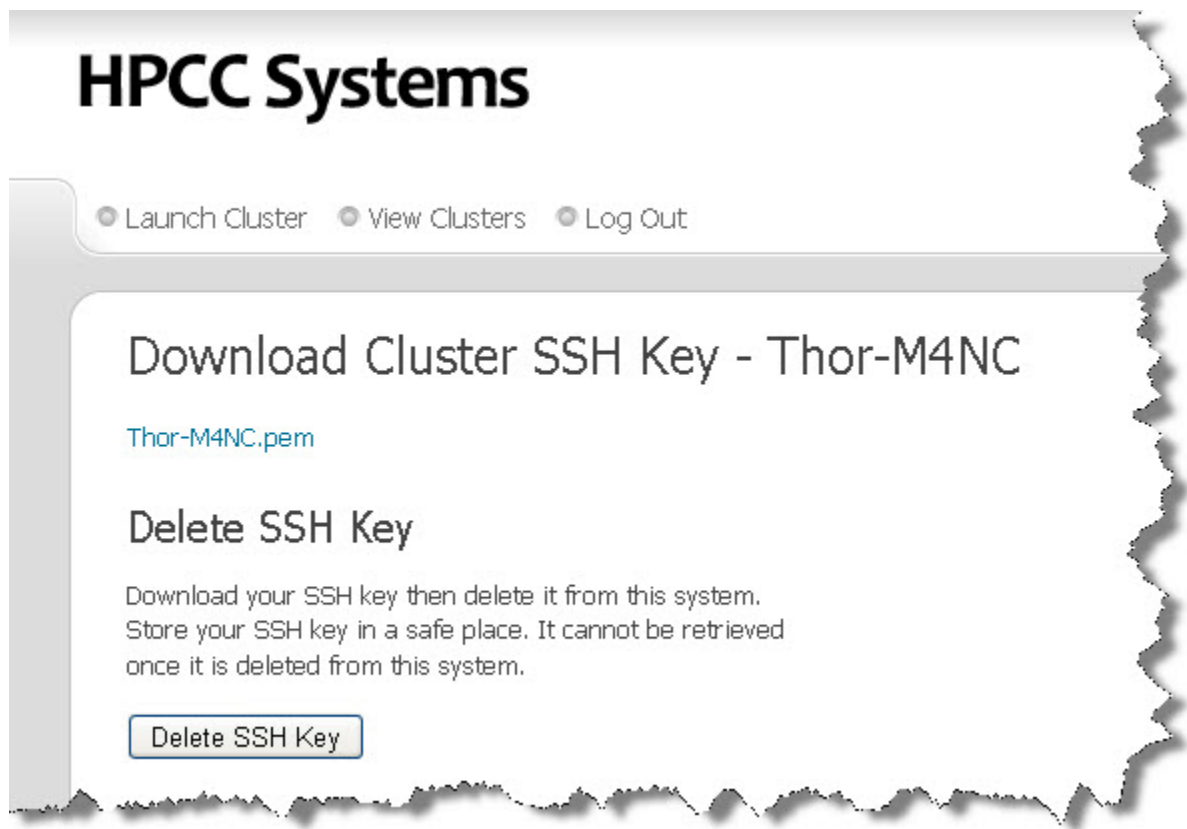
Figure 10. Running Clusters

Launch Date	Cluster ID	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Jan. 4, 2012, 2:31 p.m.	Thor-GY6D	6	None		Failed	Log			Key	
Jan. 5, 2012, 9:14 a.m.	Thor-SZ5Y	1	us-east-1a	ESP	Ready	Log	Config	IPs	Key	Terminate
Jan. 5, 2012, 9:16 a.m.	Thor-MANC	6	us-east-1a	ESP	Ready	Log	Config	IPs	Key	Terminate

[Comments](#) - [FAQs](#) - [Forum](#) - [AWS Access Keys](#) - [AWS Management Console](#) - [AWS Health Dashboard](#) - [AWS Limit Increase Request](#)

2. Click on the [Key](#) link next to the cluster.

Figure 11. Key Management



3. Click on the [pem file](#) link to download the key.

You should store this file in a safe place.

4. Press the **Delete SSH Key** button to delete the SSH key from the One-Click system.

Note: This does not remove the keys from your running cluster. It only removes it from the One-Click system and prevents further downloads of the key. Once deleted, there is no way to retrieve the key.

Running ECL

Running ECL on your HPCC Systems Thor cluster

After your platform is running, and you can now create and run some ECL¹ code using either ECL IDE, the command line ECL compiler, or the ECLPlus tool.

Install the ECL IDE and HPCC Client Tools

You only need to install the ECL IDE once. If you have already installed it, you can skip this section. .

1. In a Web browser, connect to ECL Watch using http://<PUBLIC_DNS>:8010 (where PUBLIC_DNS is the public DNS name of your ESP server).



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

¹Enterprise Control Language (ECL) is a declarative, data centric programming language used to manage all aspects of the massive data joins, sorts, and builds that truly differentiate HPCC (High Performance Computing Cluster) from other technologies in its ability to provide flexible data analysis on a massive scale.

- From ECL Watch page, click on the **Browse** link under **Resources** in the menu on the left side.

Figure 12. ECL Watch Resource Page

The screenshot shows the ECL Watch interface. On the left is a navigation menu with categories like Clusters, ECL Workunits, Query Sets, Topology, DFU Workunits, DFU Files, Roxie Queries, and Resources. The 'Resources' menu item is circled in red. The main content area has a header 'EclWatch' and a message: 'Click a link below to download a version from your installation. You may visit <http://hpccsystems.com/download> for other versions.'

graphcontrol:

Name	Version	Description
Graph Control Installer	20110523	Windows installer for Graph Control

documentation:

Name	Version	Description
HPCC Data Tutorial	3.0.4	HPCC Data Tutorial
ECL Language Reference	3.0.4	ECL Language Reference
ECL Programmers Guide	3.0.4	ECL Programmers Guide
Six Degrees of Kevin Bacon Example	3.0.4	Six Degrees of Kevin Bacon Example
HPCC Client Tools	3.0.4	HPCC Client Tools
HPCC Data Handling	3.0.4	HPCC Data Handling
Original Person File Zip	2.0	Original Person File Zip
Original Person Compressed Tar file	2.0	Original Person Compressed Tar File
HPCC Configuration Manager	3.0.4a	HPCC Configuration Manager
Rapid Data Delivery Engine Reference	3.0.4	Rapid Data Delivery Engine Reference

clienttools:


Name	Version	Description
ECL IDE Installer	ver-6_0_1_5	Windows installer for ECL IDE

Note: The order of items on the page may vary depending on the order in which optional packages are installed.

- Click on the **ECL IDE Installer** link. When prompted, save this file to your PC and then run it—do not run directly from your browser. This is a Windows installer for a Windows application.
- You can close your browser, if desired.
- Install the ECL IDE, following the prompts in the installer. Once the ECL IDE is installed successfully, you can proceed.

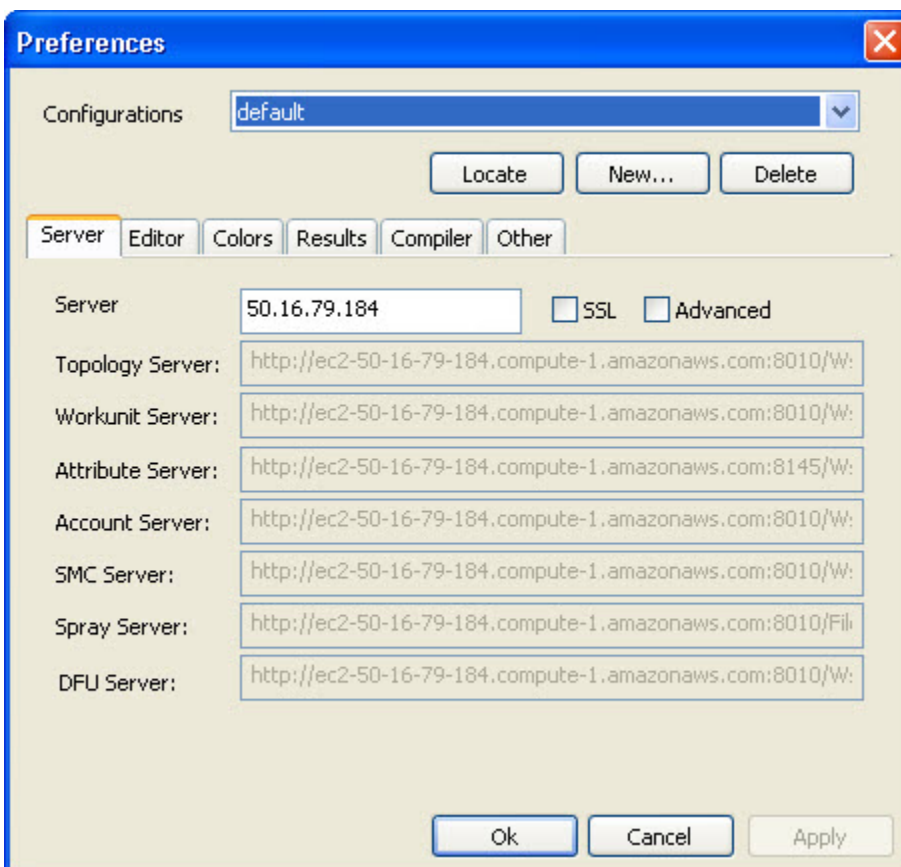
Running a basic ECL program from the ECL IDE

1. Open the ECL IDE on your Windows workstation, from your start menu. (**Start >> All Programs >> HPCC-Systems >> ECL IDE**).

	You can create a shortcut on your desktop to provide quick access to the ECL IDE.
---	---

2. On the Login Window, press the **Preferences** button.
3. In the **Server** entry control, type the Public IP of your ESP Server of your ESP server) then press the **Ok** button.

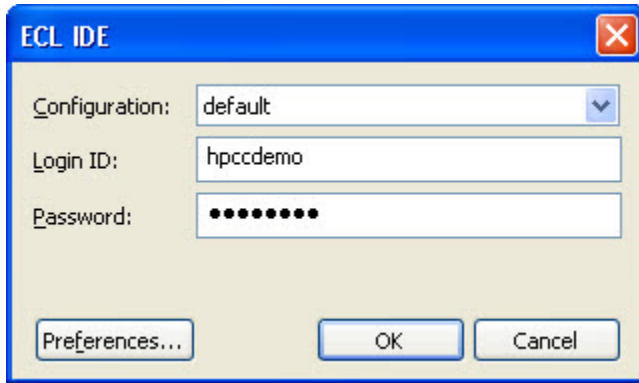
Figure 13. Login Window



4. Enter the **Login ID** and **Password** provided in the Login dialog.

Login ID	hpccdemo
Password	hpccdemo

Figure 14. Login Window



5. Open a new **Builder Window** (CTRL+N) and write the following code:

```
OUTPUT('Hello World');
```

This could also be written as:

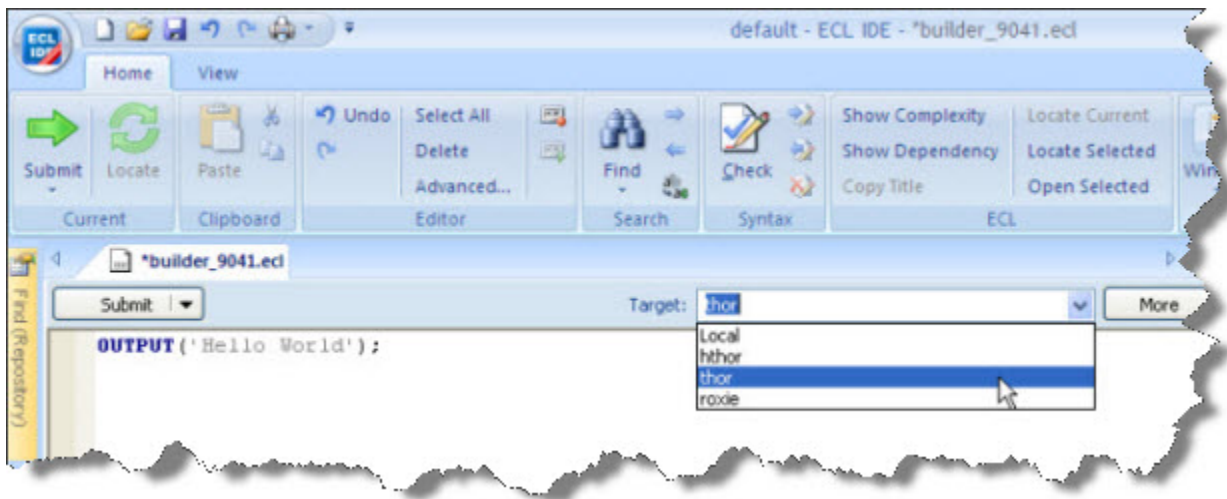
```
'Hello World';
```

In the second program listing, the OUTPUT keyword is omitted. This is possible because the language is declarative and the OUTPUT action is implicit.

6. Select **thor** as your target cluster.

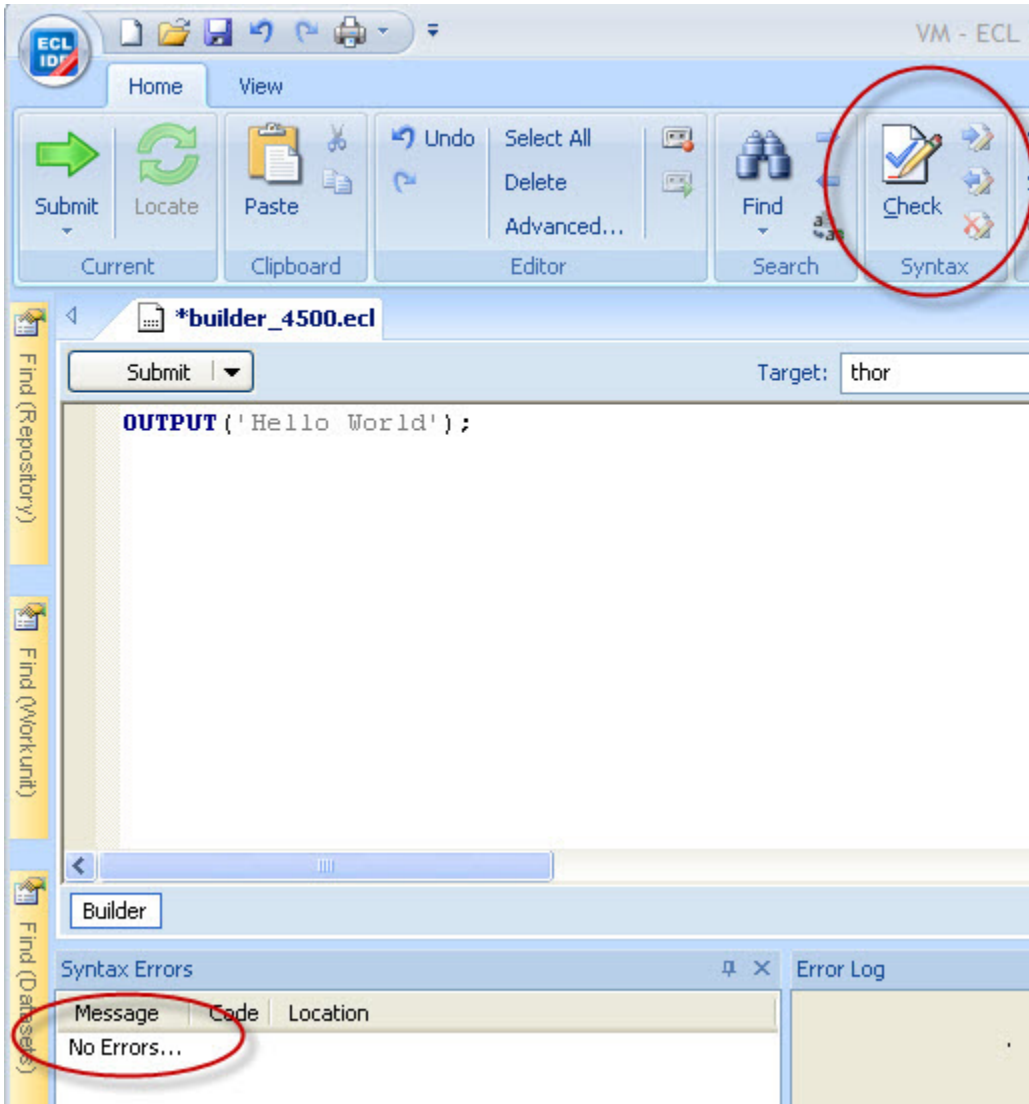
Thor is the Data Refinery component of your HPCC. It is a disk based massively parallel computer cluster, optimized for sorting, manipulating, and transforming massive data.

Figure 15. Select target



7. Press the syntax check button on the main toolbar (or press F7).

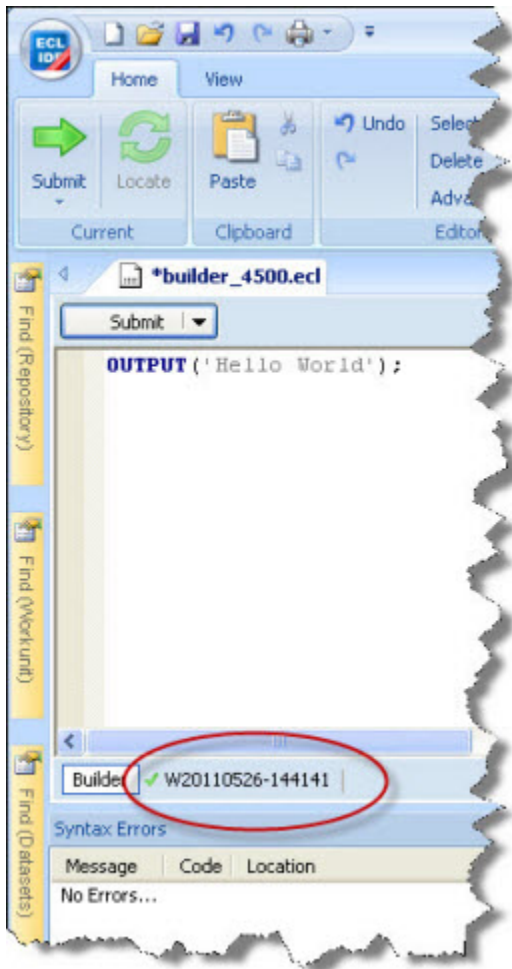
Figure 16. Syntax Check



A successful syntax check displays the "No Errors" message.

8. Press the **Submit** button (or press ctrl+enter).

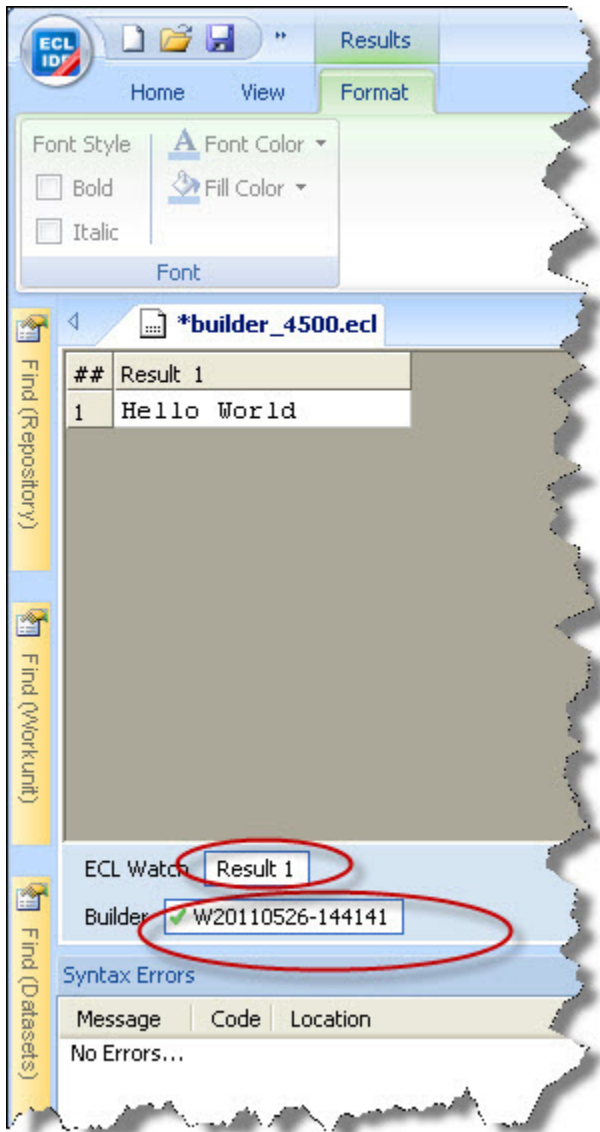
Figure 17. Completed job



The green check mark indicates successful completion.

9. Click on the workunit number tab and then on the Result 1 tab to see the output.

Figure 18. Completed job output



More ECL Examples

This section contains additional ECL examples you can use on your HPCC Systems Thor platform. You can run these on a single-node system or a larger multi-node cluster.

ECL Example: Anagram1

This example takes a STRING and produces every possible anagram from it. This code is the basis for a second example which evaluates which of these are actual words using a word list data file.

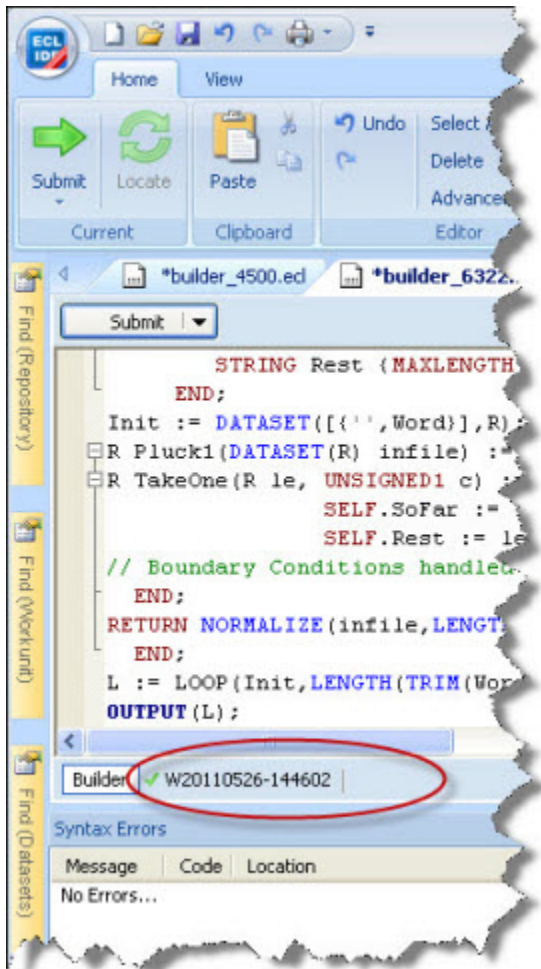
1. Open the ECL IDE (**Start >> All Programs >> HPCC Systems >> ECL IDE**) and login to your HPCC.
2. Open a new **Builder Window** (CTRL+N) and write the following code:

```
STRING Word := 'FRED' :STORED('Word');
R := RECORD
    STRING SoFar {MAXLENGTH(200)};
    STRING Rest {MAXLENGTH(200)};
END;
Init := DATASET(['',Word],R);
R Pluck1(DATASET(R) infile) := FUNCTION
R TakeOne(R le, UNSIGNED1 c) := TRANSFORM
    SELF.SoFar := le.SoFar + le.Rest[c];
    SELF.Rest := le.Rest[..c-1]+le.Rest[c+1..];
// Boundary Conditions handled automatically
END;
RETURN NORMALIZE(infile,LENGTH(LEFT.Rest),TakeOne(LEFT,COUNTER));
END;
L := LOOP(Init,LENGTH(TRIM(Word)),Pluck1(ROWS(LEFT)));
OUTPUT(L);
```

3. Select **thor** as your target cluster.
4. Press the syntax check button on the main toolbar (or press F7)

5. Press the **Submit** button (or press ctrl+enter).

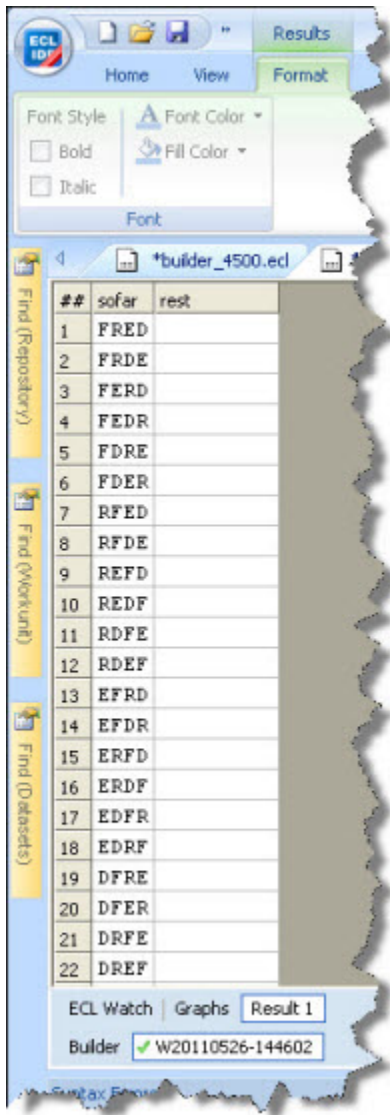
Figure 19. Completed job



The green check mark indicates successful completion.

- Click on the workunit number tab and then on the Result 1 tab to see the output.

Figure 20. Completed job output



Anagram2

In this example, we will download an open source data file of dictionary words, spray¹ that file to our Thor cluster, then validate our anagrams against that file so that we determine which are valid words. The validation step uses a JOIN of the anagram list to the dictionary file. Using an index and a keyed join would be more efficient, but this serves as a simple example.

Download the word list

We will download the word list from <http://wordlist.sourceforge.net/>

1. Download the *Official 12 Dicts* Package. The files are available in tar.gz or ZIP format.
2. Extract the **2of12.txt** file to a folder on your local machine.

Load the Dictionary File to your Landing Zone

In this step, you will copy the data files to a location from which it can be sprayed to your HPCC Thor cluster. A Landing Zone is a storage location attached to your HPCC. It has a utility running to facilitate file spraying to a cluster.

For smaller data files, maximum of 2GB, you can use the upload/download file utility in ECL Watch. This data file is only ~400 kb.

Next you will distribute (or Spray) the dataset to all the nodes in the HPCC Thor cluster. The power of the HPCC comes from its ability to assign multiple processors to work on different portions of the data file in parallel. Even though the VM Edition only has a single node, the data must be sprayed to the cluster.

1. In a Web browser, connect to ECL Watch using http://<PUBLIC_DNS>:8010 (where PUBLIC_DNS is the public DNS name of your ESP server).

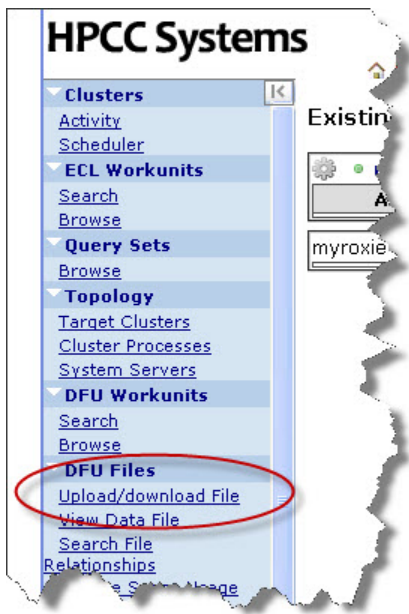


Your IP address could be different from the ones provided in the example images. Please use the IP address provided by **your** installation.

¹A *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.

- From ECL Watch page, click on the [Upload/download File](#) link in the menu on the left side.

Figure 21. 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 22. Dropzones and Files



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

The file you selected should appear in the **Select a file to upload:** field. The data file is named: **2of12.txt**.

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

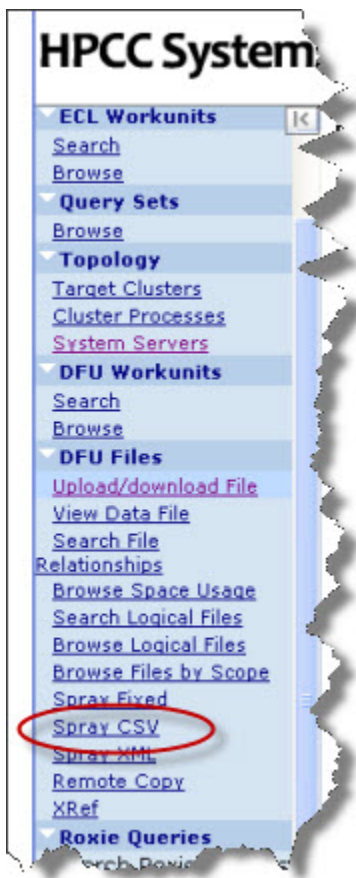
Spray the Data File to your *Thor Cluster*

To use the data file in our HPCC Thor system, we must “spray” it to all the nodes. A *spray* or *import* is the relocation of a data file from one location (such as a Landing Zone) to multiple file parts on nodes in a cluster.

The distributed or sprayed file is given a *logical-file-name* as follows: `~thor::word_list_csv` The system maintains a list of logical files and the corresponding physical file locations of the file parts.

1. In a Web browser, connect to ECL Watch using `http://<PUBLIC_DNS>:8010` (where PUBLIC_DNS is the public DNS name of your ESP server).
2. Click on the Spray CSV hyperlink under the DFU Files menu on the left.

Figure 23. Spray CSV



The **DFU Spray CSV** page displays.

3. Select mydropzone in the Source **Machine/dropzone** drop-list.

The IP Address is automatically filled and the Local Path is partially filled with the default folder on your landing zone. Note: The Community Edition typically only has one landing zone defined.

4. Complete the Local Path to include the complete file name or use the **Choose File** button to select the file from a list of files in the folder. The file is **2of12.txt**.

5. Fill in the rest of the parameters (if they are not filled in already).
 - Max Record Length 8192
 - Separator \,
 - Line Terminator \n,\r\n
 - Quote: ''
6. Fill in the Label using the Logical File name desired: thor::word_list_csv
7. Make sure the **Overwrite** and **Replicate** boxes are checked.

Figure 24. Spray the File

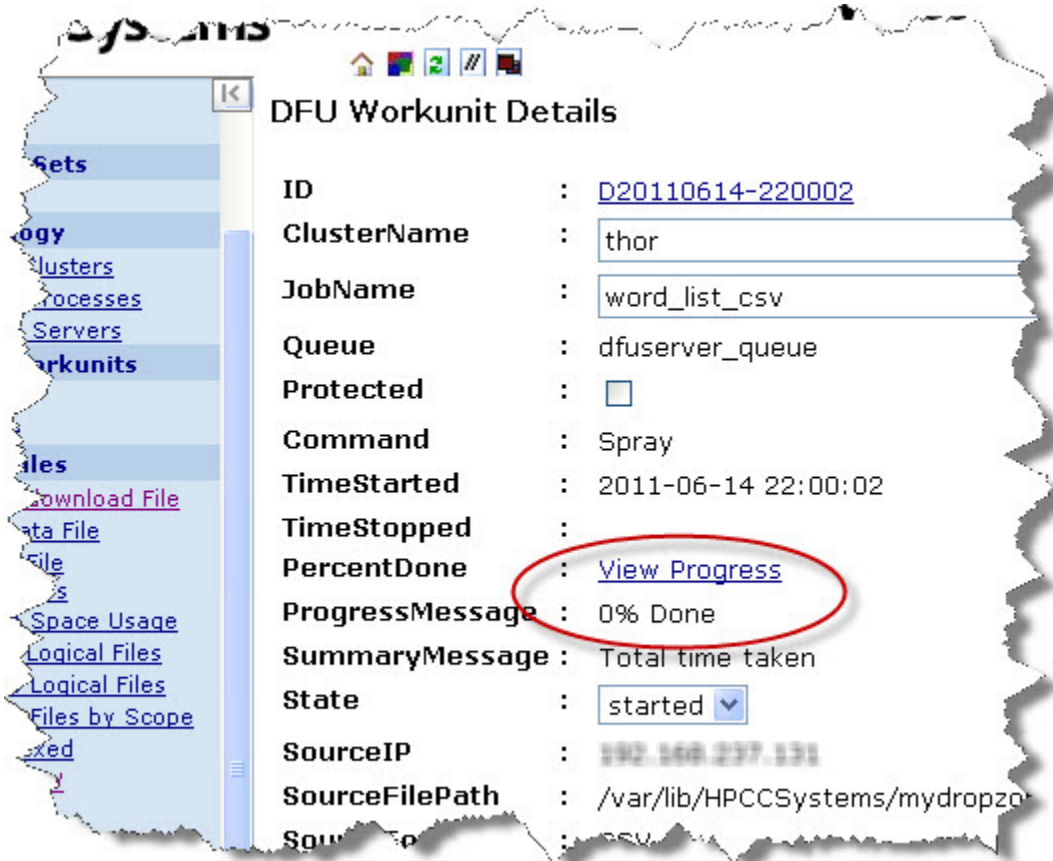
The screenshot shows the EclWatch web interface for configuring a 'Spray CSV' operation. The left sidebar contains a navigation menu with categories like Search, Query Sets, Topology, Target Clusters, Cluster Processes, System Servers, DFU Workunits, and DFU Files. The main content area is titled 'Spray CSV' and contains the following fields and options:

- Source:**
 - Machine/dropzone: localhost/mydropzone
 - IP Address: 192.168.237.131
 - Local Path: /var/lib/HPCCSystems/mydropzone/2of12.txt (with a 'Choose File' button)
 - Network Path: //192.168.237.131/var/lib/HPCCSystems/mydropzone/2of12.txt
 - Format: ASCII
 - Max Record Length: 8192
 - Separator: \,
 - Line Terminator: \n,\r\n
 - Quote: ''
- Destination:**
 - Group: mythor
 - Label: thor::word_list_csv
 - Mask: thor::word_list_csv_\$\$\$_of_\$\$\$_
 - Prefix: (empty)
- Options:**
 - Overwrite:
 - Replicate:
 - No Split:
 - Compress:

A 'Submit' button is located at the bottom of the configuration area.

8. Press the **Submit** button

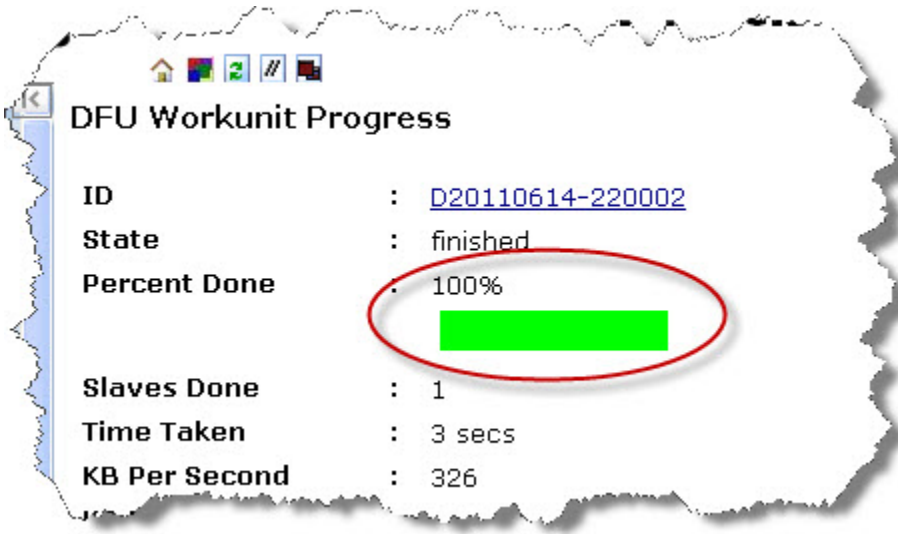
Figure 25. View Progress



9. Click the **View Progress** link

10.The Workunit progress page displays.

Figure 26. Workunit Progress



Run the ECL program on Thor

1. Open a new **Builder Window** (CTRL+N) and write the following code:

```
IMPORT Std;
layout_word_list := record
  string word;
end;
File_Word_List := dataset('~thor::word_list_csv', layout_word_list,
                        CSV(heading(1), separator(',', quote(''))));
STRING Word := 'teacher' :STORED('Word');
STRING SortString(STRING input) := FUNCTION
  OneChar := RECORD
    STRING c;
  END;
  OneChar MakeSingle(OneChar L, unsigned pos) := TRANSFORM
    SELF.c := L.c[pos];
  END;
  Split := NORMALIZE(DATASET([input], OneChar), LENGTH(input),
    MakeSingle(LEFT, COUNTER));
  SortedSplit := SORT(Split, c);
  OneChar Recombine(OneChar L, OneChar R) := TRANSFORM
    SELF.c := L.c+R.c;
  END;
  Recombined := ROLLUP(SortedSplit, Recombine(LEFT, RIGHT), ALL);
  RETURN Recombined[1].c;
END;

STRING CleanedWord := SortString(TRIM(Std.Str.ToUpperCase(Word)));

R := RECORD
  STRING SoFar {MAXLENGTH(200)};
  STRING Rest {MAXLENGTH(200)};
END;
Init := DATASET(['', CleanedWord], R);
R Pluck1(DATASET(R) infile) := FUNCTION
  R TakeOne(R le, UNSIGNED l c) := TRANSFORM
    SELF.SoFar := le.SoFar + le.Rest[c];
    SELF.Rest := le.Rest[..c-1]+le.Rest[c+1..];
    // Boundary Conditions
    // handled automatically
  END;
  RETURN DEDUP(NORMALIZE(infile, LENGTH(LEFT.Rest), TakeOne(LEFT, COUNTER)));
END;
L := LOOP(Init, LENGTH(CleanedWord), Pluck1(ROWS(LEFT)));
ValidWords := JOIN(L, File_Word_List,
LEFT.SoFar=Std.Str.ToUpperCase(RIGHT.Word), TRANSFORM(LEFT));
OUTPUT(CleanedWord);
COUNT(ValidWords);
OUTPUT(ValidWords)
```

2. Select **thor** as your target cluster.
3. Press the syntax check button on the main toolbar (or press F7)
4. Press the **Submit** button.
5. When it completes, select the Workunit tab, then select the Result tab.
6. Examine the result.

Data Handling

This section explains data handling in an AWS configuration. More information about Data Handling in an HPCC Systems platform are available in the *Data Handling* manual.

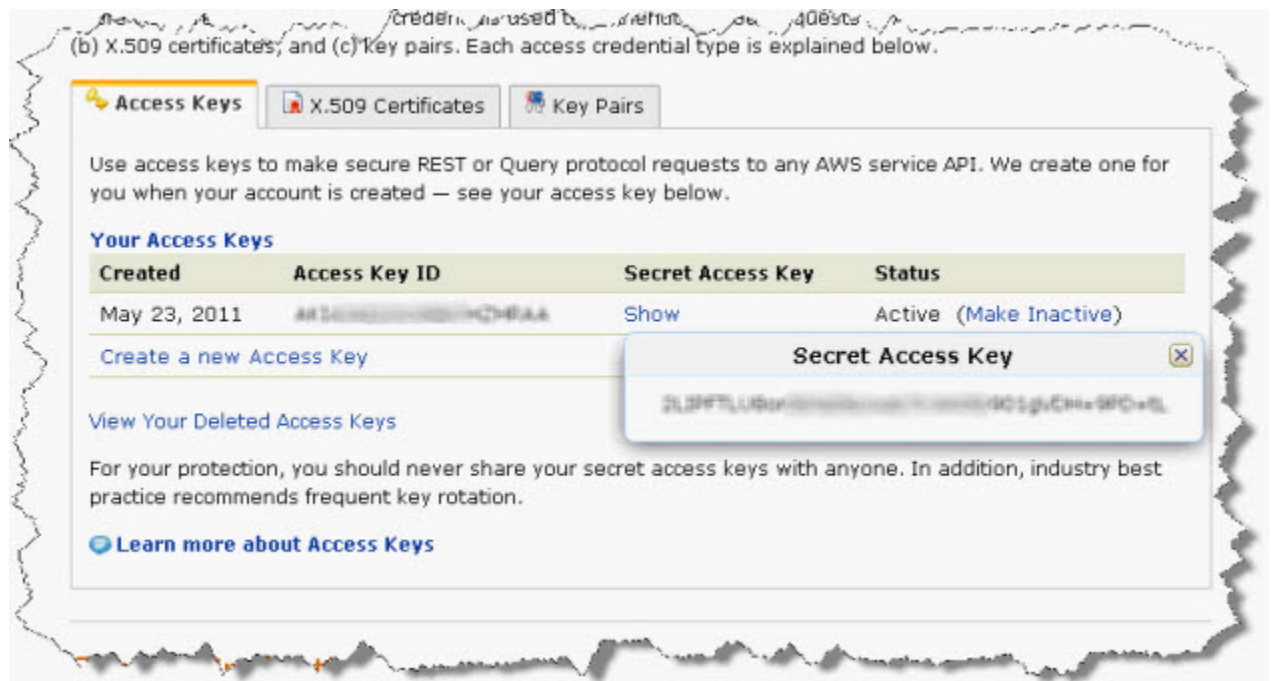
Using S3 buckets

S3 buckets provide a means of persistent storage inside Amazon Web Services. You must configure your AWS account to have an active Access Key pair enable and create S3 buckets. Once you have created and activated your access key pair and and created a unique S3 bucket, you will use these for all future instantiations.

Find your Amazon Access Key ID and Secret Access Key

1. Go to aws.amazon.com and login, if needed.
2. Select **Account**.
3. Select **Security Credentials**.
4. On the page, look for the section called **Access Credentials**.
5. Note your **Access Key ID** and your **Secret Access Key**.

Figure 27. Credentials



Portions of this image are intentionally blurred

Access Key ID	
Secret Access Key	

Install and Configure S3 packages on your Landing Zone node

To move files to or from S3 storage, the S3 packages must be installed and configured on your Landing Zone node.

1. Open a console window and connect to the Landing Zone (LZ) node
2. Run these commands:

```
sudo apt-get install s3cmd  
s3cmd --configure
```

3. Enter your **Access Key**
4. Enter your **Secret Access Key**
5. Leave encrypt password blank
6. Leave path to GPG program blank
7. Answer the question Use HTTPS?
 - Enter no to improve performance
 - Enter yes if you are concerned about data privacy
8. Leave proxy server blank
9. Enter **Yes** to Test Access
10. Enter **Yes** to Save Settings

Creating and Using S3 Buckets

To store data on S3, you must create a bucket that is unique to the whole s3 system. Once created, this bucket persists even when you close a instances of servers.

You can despray a file from Thor to your landing zone, then copy to an S3 bucket to for persistent storage. Later, you can copy files from the S3 bucket to a landing zone and spray the file to a Thor cluster. More information about Data Handling in an HPCC Systems platform are available in the *Data Handling* manual.

Create a bucket

```
s3cmd mb s3://your-unique-bucket-name
```

List Buckets

```
s3cmd ls
```

Upload a file to a bucket

```
s3cmd put myfile.csv s3://your-unique-bucket-name
```

Retrieve a file from a bucket

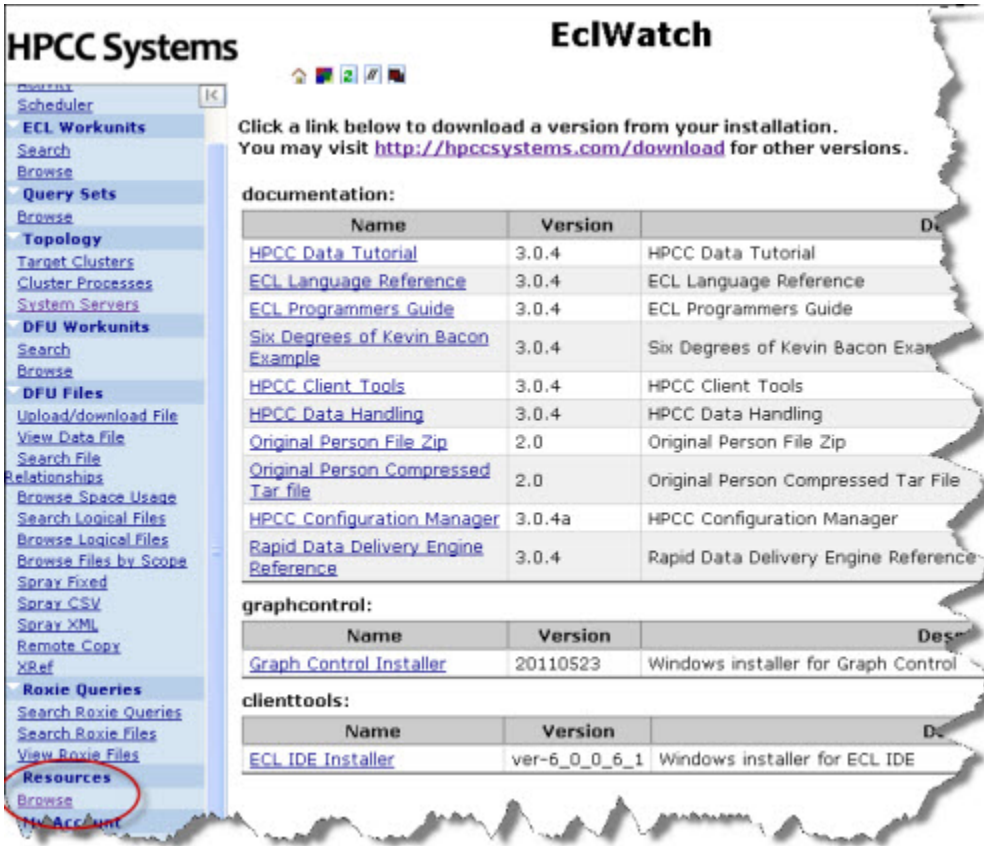
```
s3cmd get s3://your-unique-bucket-name/myfile.csv myfile.csv
```

See <http://s3tools.org/s3cmd> for more information on how to use s3cmd

Next Steps

From ECL Watch page, click on the **Resources** link in the menu on the left side.

Figure 28. ECL Watch Resource Page



To familiarize yourself with what your system can do, we recommend following the steps in:

- The **HPCC Data Tutorial**
- **The Six Degrees of Kevin Bacon** example
- Read **Using Config Manager** to learn how to configure an HPCC platform using Advanced View.
- Use your new skills to process your own massive dataset!

The HPCC Systems Portal (HPCCSystems.com) is also a valuable resource for more information including:

- Video Tutorials
- Additional examples
- White Papers
- Documentation
- User Forums