HPCC Systems®

HPCC Systems® Instant Cloud for AWS

Boca Raton Documentation Team

LexisNexis®
Risk Solutions

HPCC Systems® Instant Cloud for AWS

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.

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

2015 Version 5.4.2-1

Introduction

This guide provides details and guidance in running an HPCC Systems® Platform inside an Amazon Web Services (AWS) Elastic Cloud (EC2) using the Instant Cloud for AWS page.

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

This is useful for:

- Proof-of-concept
- Experimentation
- Learning
- Leveraging the HPCC Systems platform without incurring cost of hardware and administration
- Create and use an HPCC Systems 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 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

Using Instant Cloud Launch

Before you begin

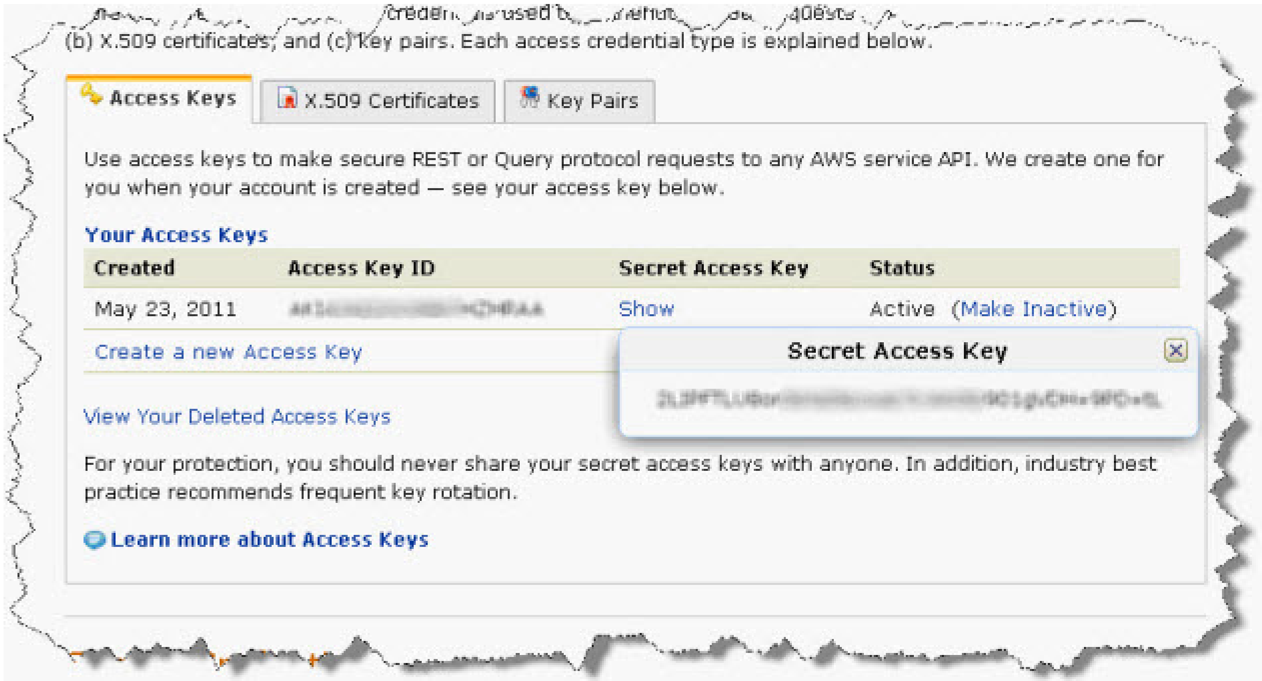
In this section, you will gather some information you need before beginning. This includes:

- 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



(b) X.509 certificates, and (c) Key pairs. Each access credential type is explained below.

Use access keys to make secure REST or Query protocol requests to any AWS service API. We create one for you when your account is created — see your access key below.

Your Access Keys

Created	Access Key ID	Secret Access Key	Status
May 23, 2011	[redacted]	Show	Active (Make Inactive)

Create a new Access Key

Secret Access Key

[redacted]

View Your Deleted Access Keys

For your protection, you should never share your secret access keys with anyone. In addition, industry best practice recommends frequent key rotation.

Learn more about Access Keys

Portions of this image are intentionally blurred

Access Key ID	
Secret Access Key	

Login

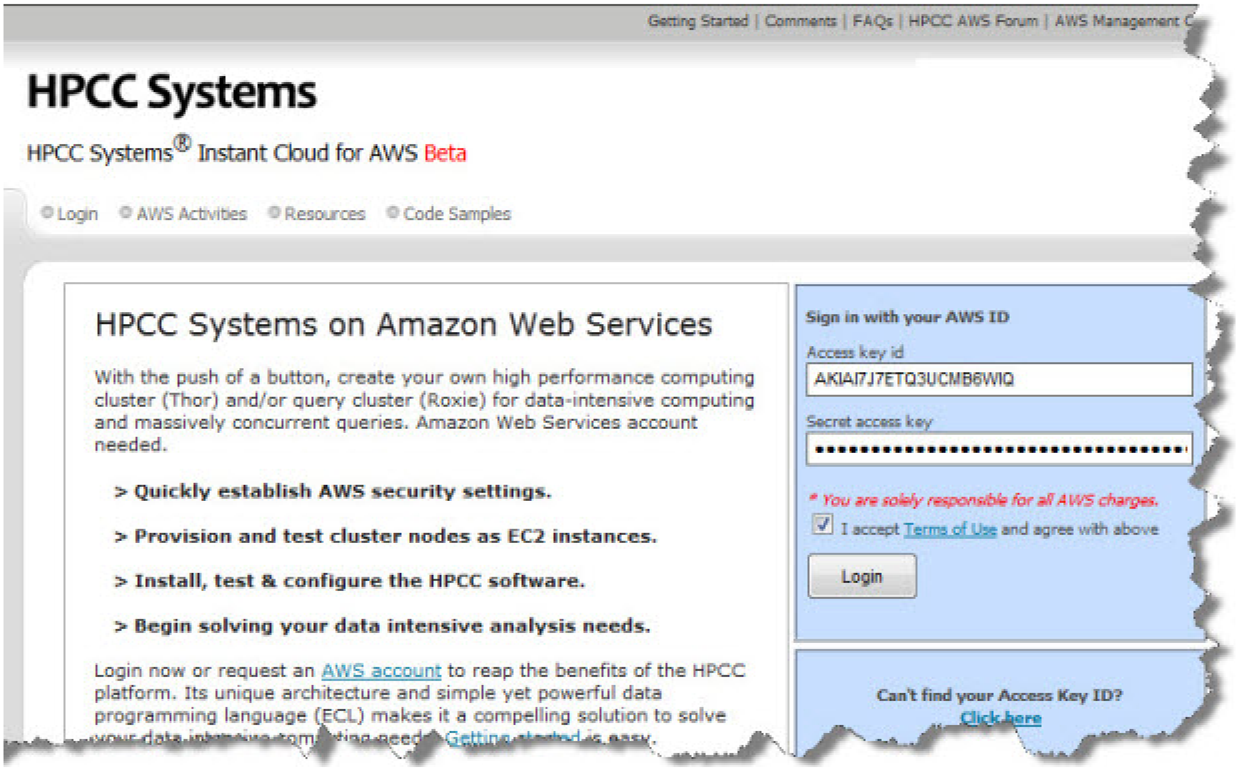
1. Open a browser and go to https://aws.hpccsystems.com/.

   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 link under **Can't find your Access Key ID?** to go to that section of the AWS Management Console..

Figure 2. Login


Getting Started | Comments | FAQs | HPCC AWS Forum | AWS Management C
HPCC Systems
HPCC Systems® Instant Cloud for AWS Beta
Login
AWS Activities
Resources
Code Samples
Login

HPCC Systems on Amazon Web Services

With the push of a button, create your own high performance computing cluster (Thor) and/or query cluster (Roxie) for data-intensive computing and massively concurrent queries. Amazon Web Services account needed.

> Quickly establish AWS security settings.

> Provision and test cluster nodes as EC2 instances.

> Install, test & configure the HPCC software.

> Begin solving your data intensive analysis needs.

Login now or request an AWS account to reap the benefits of the HPCC platform. Its unique architecture and simple yet powerful data programming language (ECL) makes it a compelling solution to solve your data intensive computing needs. Getting started is easy.

Sign in with your AWS ID

Access key id: AKIAI7J7ETQ3UCMB6WIQ

Secret access key: ••••••••••••••••••••••••••••••••••

* You are solely responsible for all AWS charges.

☑ I accept Terms of Use and agree with above

Can't find your Access Key ID?
Click here

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 HPCC Cluster** window displays.

Launch a New HPCC Cluster

In this section, you will launch a set of Ubuntu 12.04 machines to use for your HPCC Systems Thor platform. The Instant Cloud page uses the input you specify to create your cluster for you.

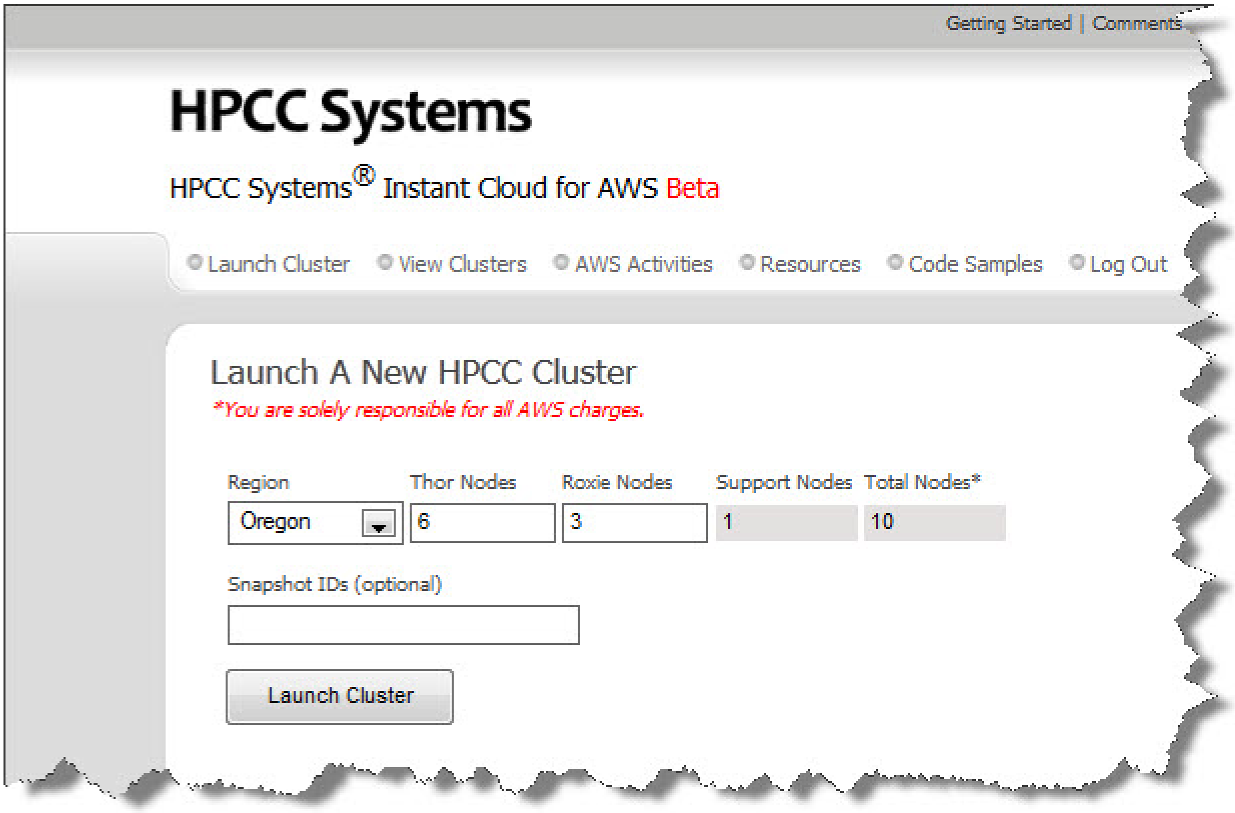
When you press the **Launch Cluster** button, it:

- 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 the provided AMI (ami-e01698d0).
- Gathers Private and Public IPs.
- Installs HPCC Systems platform packages.
- Configures the requested Thor Cluster, the requested Roxie Cluster, and the required 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 and the number of Roxie nodes to instantiate.

Figure 3. Launch a New Thor Cluster



Launch A New HPCC Cluster

*You are solely responsible for all AWS charges.

Region	Thor Nodes	Roxie Nodes	Support Nodes	Total Nodes*
Oregon	6	3	1	10

Snapshot IDs (optional)

2. Optionally, specify Snapshot ID(s) to attach data to your landing zone.

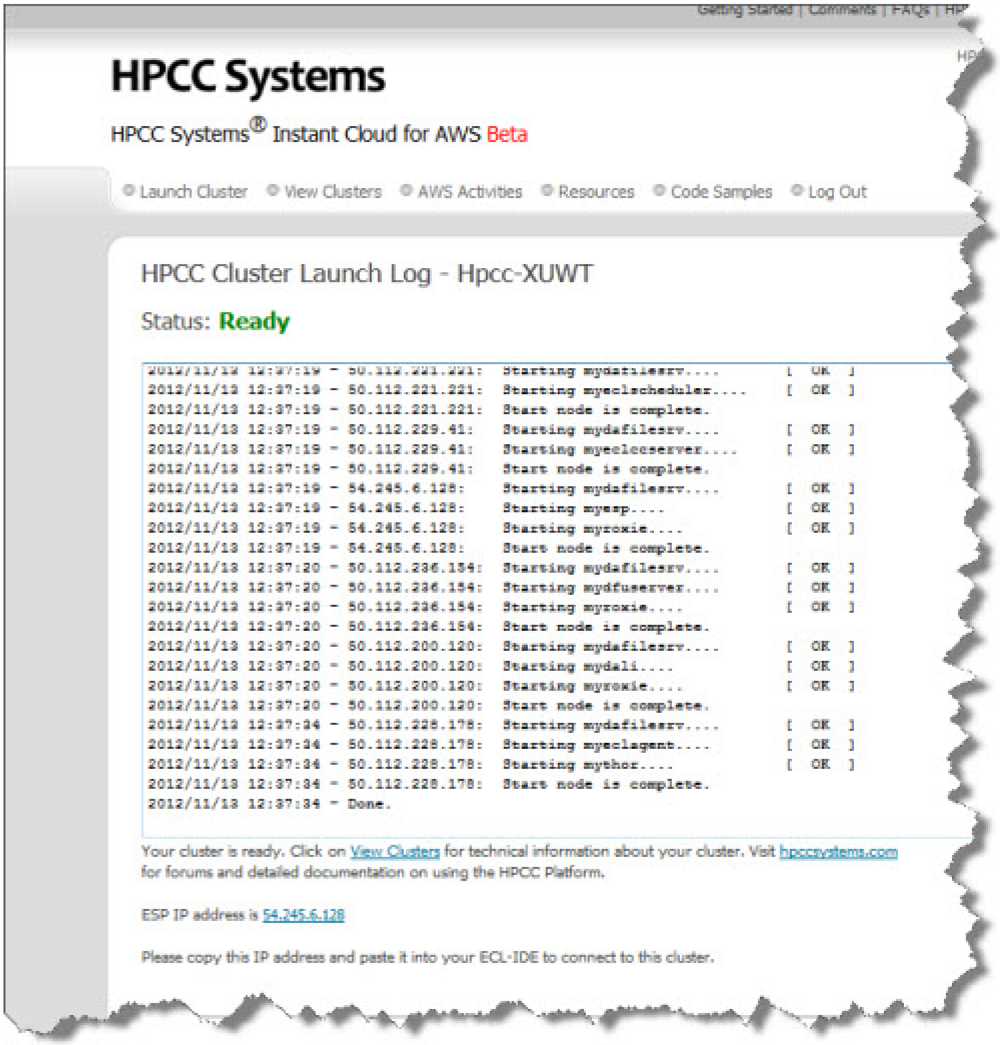
   This would be a previously saved "snapshot" of a landing zone data store.

3. 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.

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

Figure 4. Launch Cluster



HPCC Cluster Launch Log - Hpcc-XUWT

Status: Ready

```
2012/11/13 12:37:19 - 50.112.221.221:  Starting mydafilesrv....       [  OK  ]
2012/11/13 12:37:19 - 50.112.221.221:  Starting myeclscheduler....    [  OK  ]
2012/11/13 12:37:19 - 50.112.221.221:  Start node is complete.
2012/11/13 12:37:19 - 50.112.229.41:   Starting mydafilesrv....       [  OK  ]
2012/11/13 12:37:19 - 50.112.229.41:   Starting myeclccserver....     [  OK  ]
2012/11/13 12:37:19 - 50.112.229.41:   Start node is complete.
2012/11/13 12:37:19 - 54.245.6.128:    Starting mydafilesrv....       [  OK  ]
2012/11/13 12:37:19 - 54.245.6.128:    Starting myesp....             [  OK  ]
2012/11/13 12:37:19 - 54.245.6.128:    Starting myroxie....           [  OK  ]
2012/11/13 12:37:19 - 54.245.6.128:    Start node is complete.
2012/11/13 12:37:20 - 50.112.236.154:  Starting mydafilesrv....       [  OK  ]
2012/11/13 12:37:20 - 50.112.236.154:  Starting mydfuserver....       [  OK  ]
2012/11/13 12:37:20 - 50.112.236.154:  Starting myroxie....           [  OK  ]
2012/11/13 12:37:20 - 50.112.236.154:  Start node is complete.
2012/11/13 12:37:20 - 50.112.200.120:  Starting mydafilesrv....       [  OK  ]
2012/11/13 12:37:20 - 50.112.200.120:  Starting mydali....            [  OK  ]
2012/11/13 12:37:20 - 50.112.200.120:  Starting myroxie....           [  OK  ]
2012/11/13 12:37:20 - 50.112.200.120:  Start node is complete.
2012/11/13 12:37:34 - 50.112.228.178:  Starting mydafilesrv....       [  OK  ]
2012/11/13 12:37:34 - 50.112.228.178:  Starting myeclagent....        [  OK  ]
2012/11/13 12:37:34 - 50.112.228.178:  Starting mythor....            [  OK  ]
2012/11/13 12:37:34 - 50.112.228.178:  Start node is complete.
2012/11/13 12:37:34 - Done.
```

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.

ESP IP address is 54.245.6.128

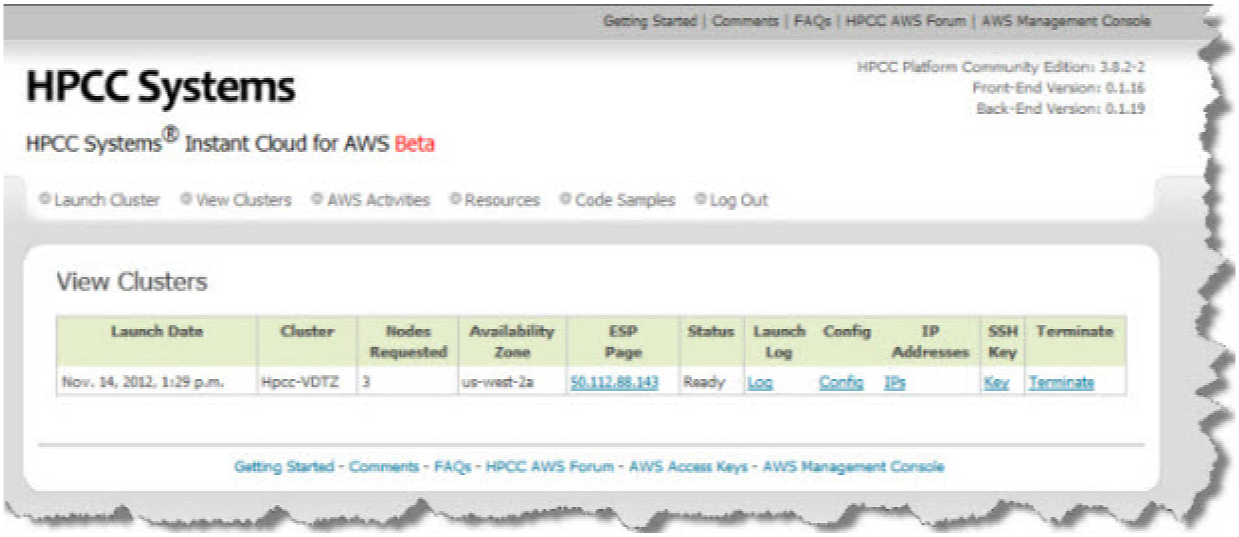
Please copy this IP address and paste it into your ECL-IDE to connect to this cluster.

5. 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


Getting Started | Comments | FAQs | HPCC AWS Forum | AWS Management Console
HPCC Systems
HPCC Systems® Instant Cloud for AWS Beta
HPCC Platform Community Edition: 3.8.2-2
Front-End Version: 0.1.16
Back-End Version: 0.1.19
Launch Cluster
View Clusters
AWS Activities
Resources
Code Samples
Log Out

View Clusters

Launch Date	Cluster	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Nov. 14, 2012, 1:29 p.m.	Hpcc-VDTZ	3	us-west-2a	50.112.88.143	Ready	Log	Config	IPs	Key	Terminate

Getting Started - Comments - FAQs - HPCC AWS Forum - AWS Access Keys - AWS Management Console

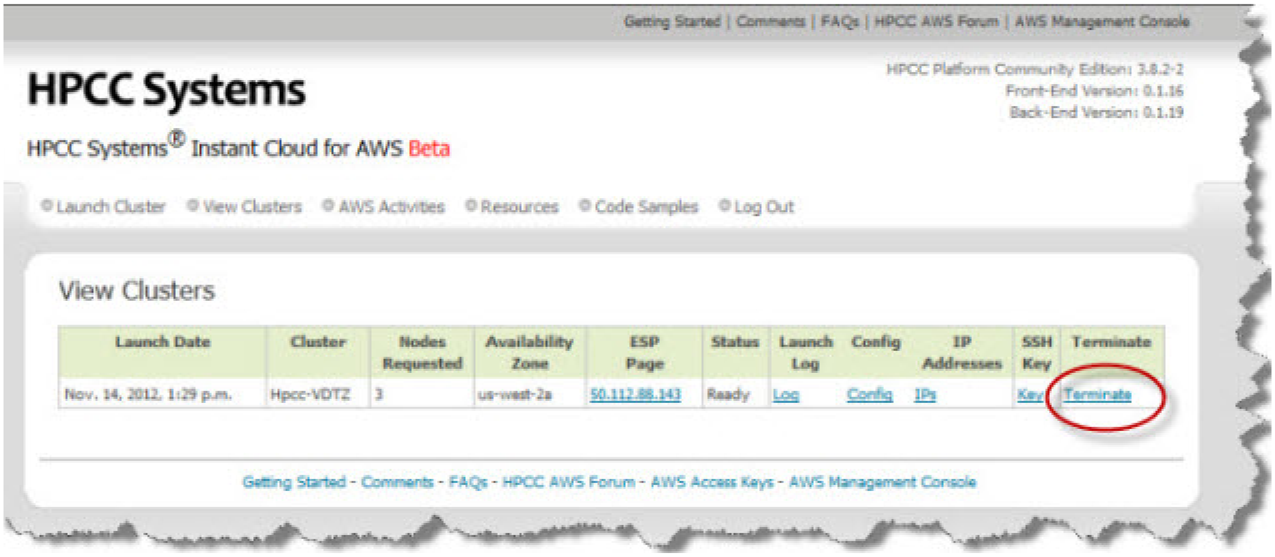
Terminating your instances

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.

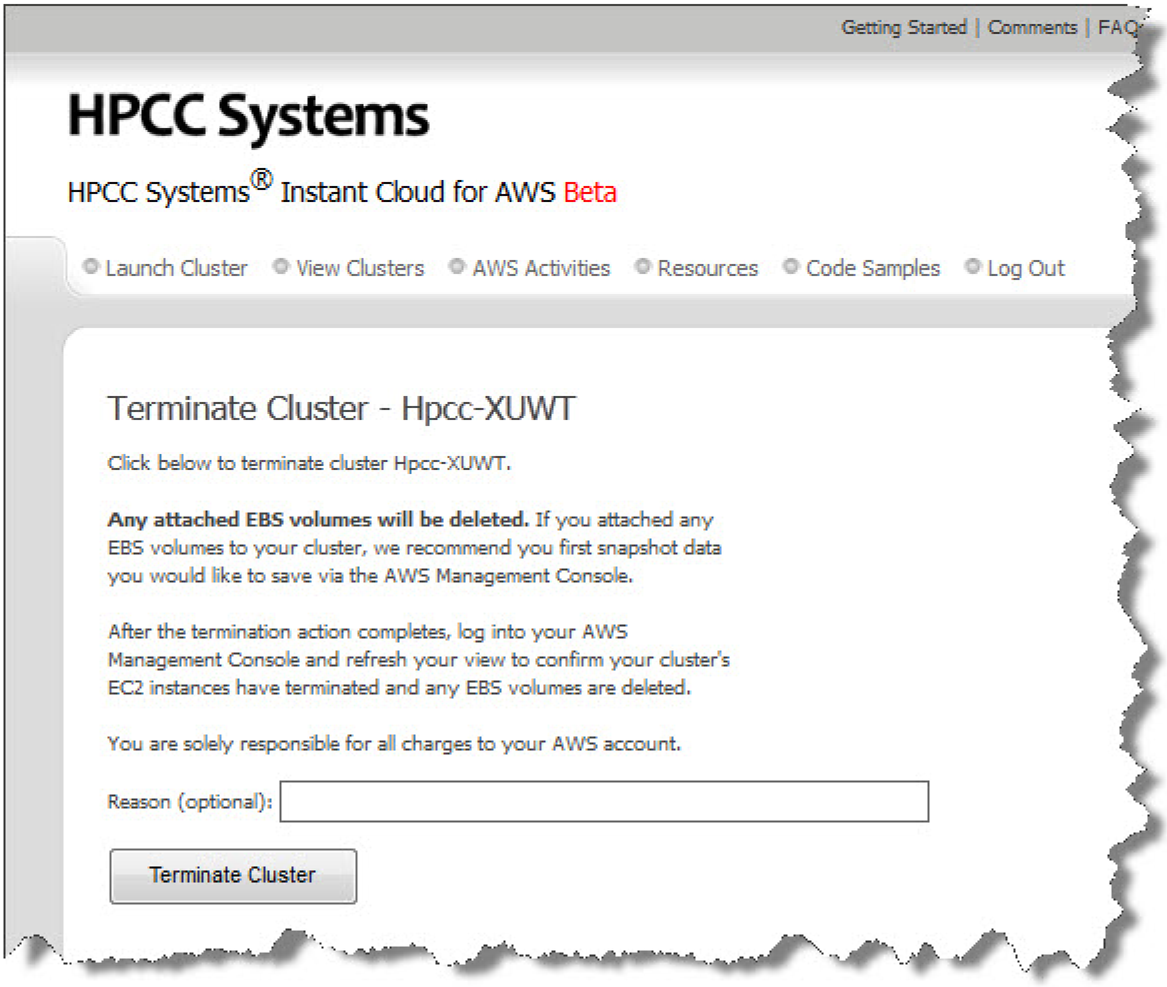
Figure 6. Running Clusters


Getting Started | Comments | FAQs | HPCC AWS Forum | AWS Management Console
HPCC Systems
HPCC Platform Community Edition: 3.8.2-2
Front-End Version: 0.1.16
Back-End Version: 0.1.19
HPCC Systems® Instant Cloud for AWS Beta
Launch Cluster
View Clusters
AWS Activities
Resources
Code Samples
Log Out
View Clusters
Getting Started - Comments - FAQs - HPCC AWS Forum - AWS Access Keys - AWS Management Console

Launch Date	Cluster	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Nov. 14, 2012, 1:29 p.m.	Hpcc-VDTZ	3	us-west-2a	50.112.88.143	Ready	Log	Config	IPs	Key	Terminate

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

Figure 7. Terminate Cluster



Terminate Cluster - Hpcc-XUWT

Click below to terminate cluster Hpcc-XUWT.

Any attached EBS volumes will be deleted. If you attached any EBS volumes to your cluster, we recommend you first snapshot data you would like to save via the AWS Management Console.

After the termination action completes, log into your AWS Management Console and refresh your view to confirm your cluster's EC2 instances have terminated and any EBS volumes are deleted.

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

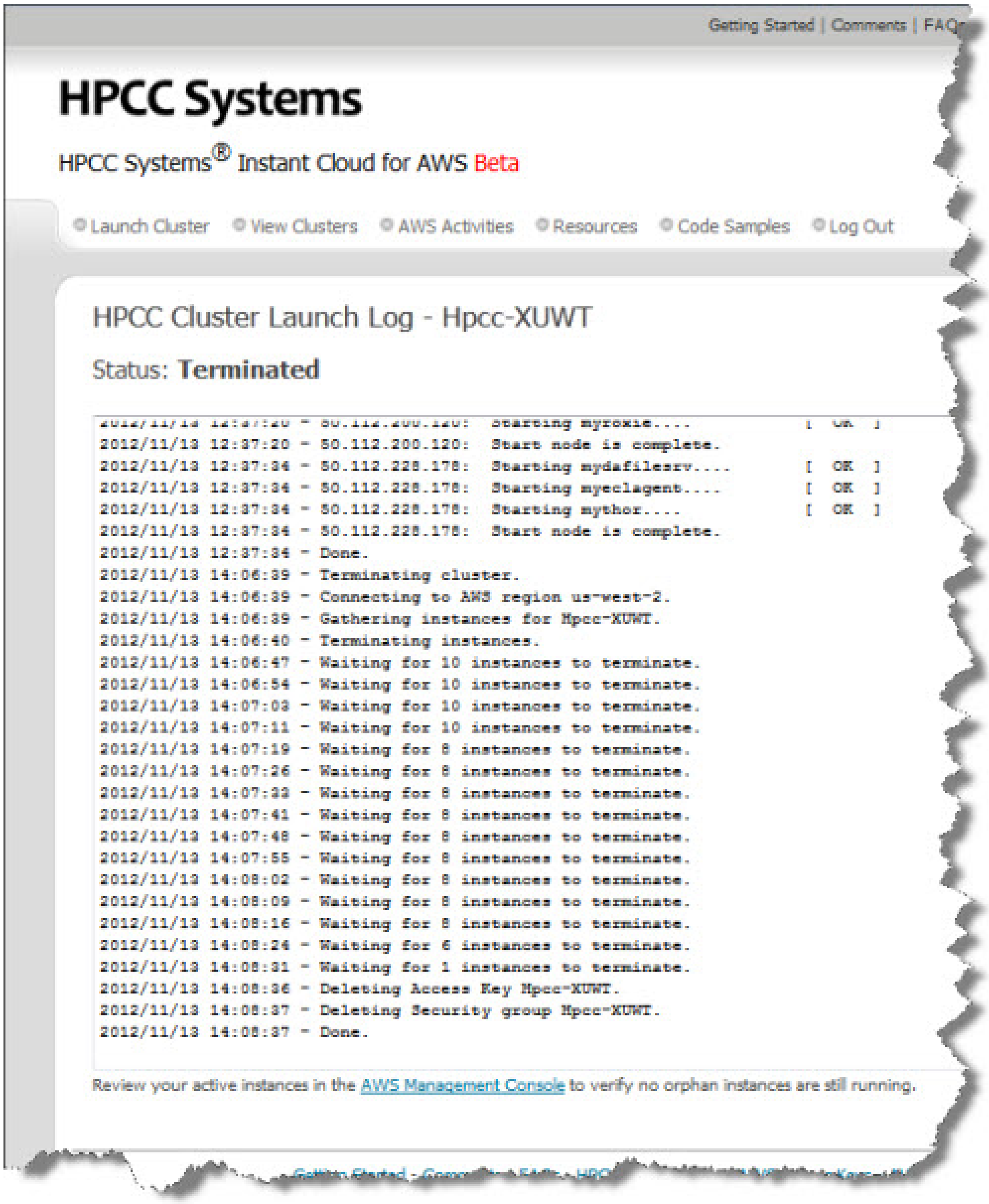
Reason (optional):

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



HPCC Cluster Launch Log - Hpcc-XUWT

Status: **Terminated**

```
2012/11/13 12:37:20 - 50.112.200.120:  Starting myroxie....              [  OK  ]
2012/11/13 12:37:20 - 50.112.200.120:  Start node is complete.
2012/11/13 12:37:34 - 50.112.228.178:  Starting mydafilesrv....          [  OK  ]
2012/11/13 12:37:34 - 50.112.228.178:  Starting myeclagent....           [  OK  ]
2012/11/13 12:37:34 - 50.112.228.178:  Starting mythor....               [  OK  ]
2012/11/13 12:37:34 - 50.112.228.178:  Start node is complete.
2012/11/13 12:37:34 - Done.
2012/11/13 14:06:39 - Terminating cluster.
2012/11/13 14:06:39 - Connecting to AWS region us-west-2.
2012/11/13 14:06:39 - Gathering instances for Hpcc-XUWT.
2012/11/13 14:06:40 - Terminating instances.
2012/11/13 14:06:47 - Waiting for 10 instances to terminate.
2012/11/13 14:06:54 - Waiting for 10 instances to terminate.
2012/11/13 14:07:03 - Waiting for 10 instances to terminate.
2012/11/13 14:07:11 - Waiting for 10 instances to terminate.
2012/11/13 14:07:19 - Waiting for 8 instances to terminate.
2012/11/13 14:07:26 - Waiting for 8 instances to terminate.
2012/11/13 14:07:33 - Waiting for 8 instances to terminate.
2012/11/13 14:07:41 - Waiting for 8 instances to terminate.
2012/11/13 14:07:48 - Waiting for 8 instances to terminate.
2012/11/13 14:07:55 - Waiting for 8 instances to terminate.
2012/11/13 14:08:02 - Waiting for 8 instances to terminate.
2012/11/13 14:08:09 - Waiting for 8 instances to terminate.
2012/11/13 14:08:16 - Waiting for 8 instances to terminate.
2012/11/13 14:08:24 - Waiting for 6 instances to terminate.
2012/11/13 14:08:31 - Waiting for 1 instances to terminate.
2012/11/13 14:08:36 - Deleting Access Key Hpcc-XUWT.
2012/11/13 14:08:37 - Deleting Security group Hpcc-XUWT.
2012/11/13 14:08:37 - Done.
```

Review your active instances in the AWS Management Console to verify no orphan instances are still running.

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.

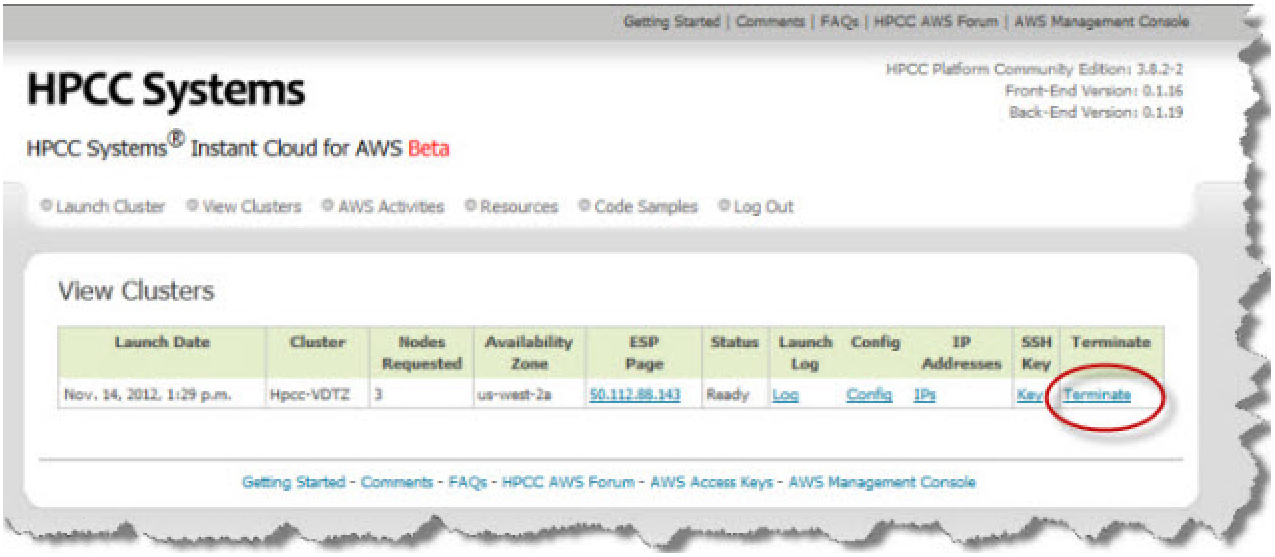
Other Tasks

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


Getting Started | Comments | FAQs | HPCC AWS Forum | AWS Management Console
HPCC Systems
HPCC Platform Community Edition: 3.8.2-2
Front-End Version: 0.1.16
Back-End Version: 0.1.19
HPCC Systems® Instant Cloud for AWS Beta
Launch Cluster
View Clusters
AWS Activities
Resources
Code Samples
Log Out

View Clusters

Launch Date	Cluster	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Nov. 14, 2012, 1:29 p.m.	Hpcc-VDTZ	3	us-west-2a	50.112.88.143	Ready	Log	Config	IPs	Key	Terminate

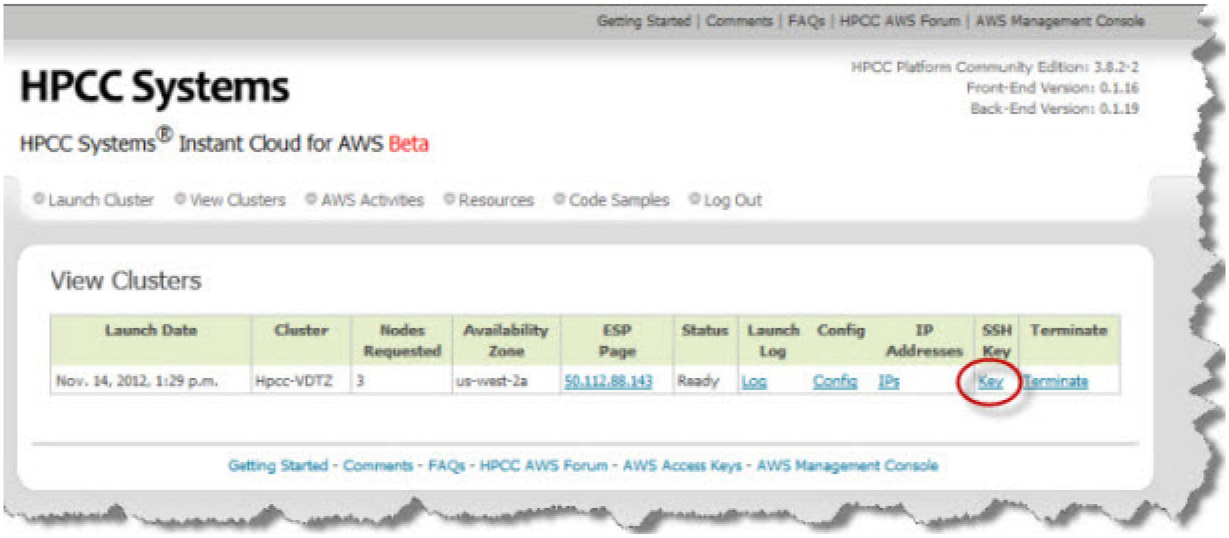
Getting Started - Comments - FAQs - HPCC AWS Forum - AWS Access Keys - AWS Management Console

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 **View Clusters** page using the **View Clusters** Link at the top of a page.

Figure 10. Running Clusters

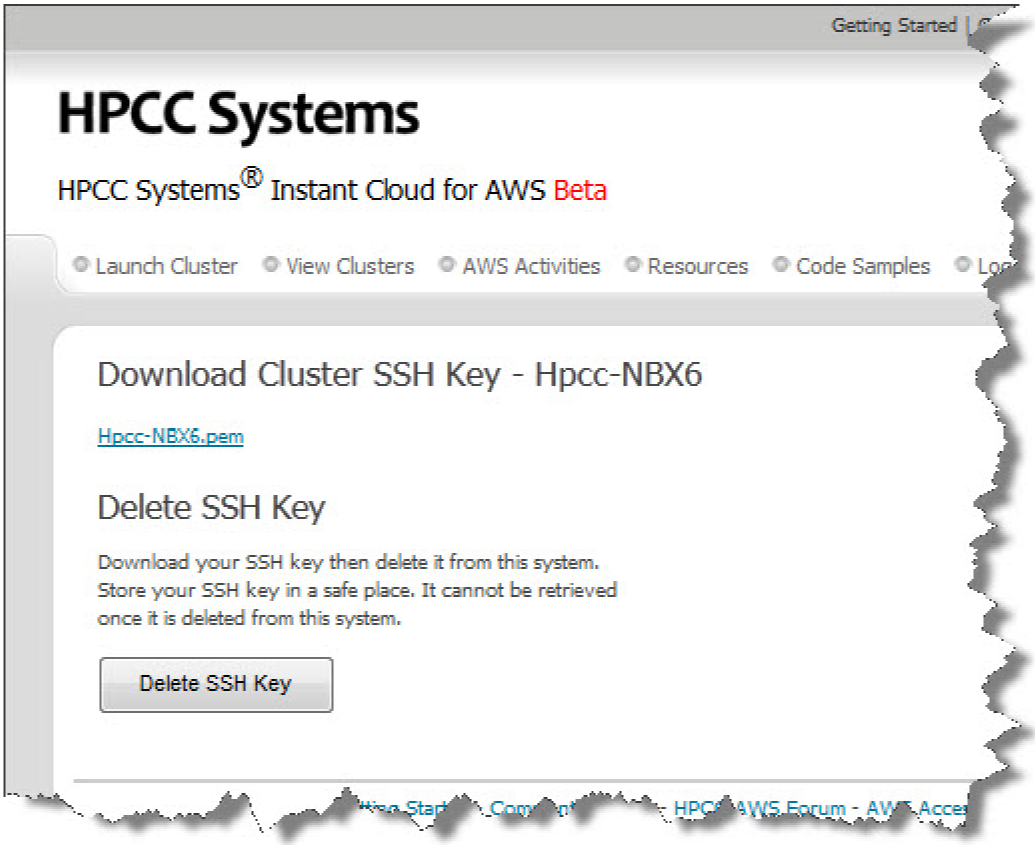

Getting Started | Comments | FAQs | HPCC AWS Forum | AWS Management Console
HPCC Systems
HPCC Systems® Instant Cloud for AWS Beta
HPCC Platform Community Edition: 3.8.2-2
Front-End Version: 0.1.16
Back-End Version: 0.1.19
Launch Cluster
View Clusters
AWS Activities
Resources
Code Samples
Log Out
Getting Started - Comments - FAQs - HPCC AWS Forum - AWS Access Keys - AWS Management Console

View Clusters

Launch Date	Cluster	Nodes Requested	Availability Zone	ESP Page	Status	Launch Log	Config	IP Addresses	SSH Key	Terminate
Nov. 14, 2012, 1:29 p.m.	Hpcc-VDTZ	3	us-west-2a	50.112.88.143	Ready	Log	Config	IPs	Key	Terminate

2. Click on the **Key** link next to the cluster.

Figure 11. Key Management



Download Cluster SSH Key - Hpcc-NBX6

Hpcc-NBX6.pem

Delete SSH Key

Download your SSH key then delete it from this system. Store your SSH key in a safe place. It cannot be retrieved once it is deleted from this system.

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 Instant Cloud 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 cluster

After your platform is running, and you can now create and run some ECL[1] code using either ECL IDE, the command line ECL compiler, or the ECLPlus tool.

Install the ECL IDE and HPCC Client Tools

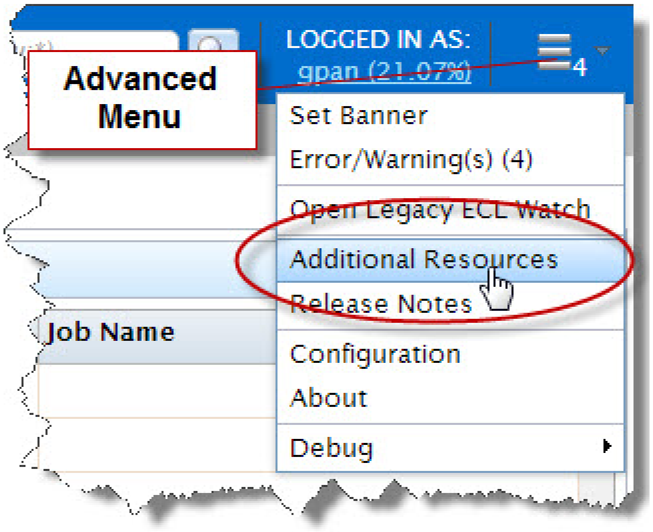
You only need to install the ECL IDE once. If you have already installed it, you can skip this section. .

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.

2. From the ECL Watch Advanced menu, select on the **Additional Resources** link.

Figure 12. ECL Watch Resource Page



Follow the link to the HPCC System's portal download page.

3. Click on the **ECL IDE** link. (on the right hand side in the Download column, under the Free Community Edition heading)

4. Follow the instructions on the web page to install the ECL IDE.

[1] **E**nterprise **C**ontrol **L**anguage (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.

5. Install the ECL IDE, following the prompts in the installation program. Once the ECL IDE is installed successfully, you can proceed.

Running a basic ECL program from the ECL IDE

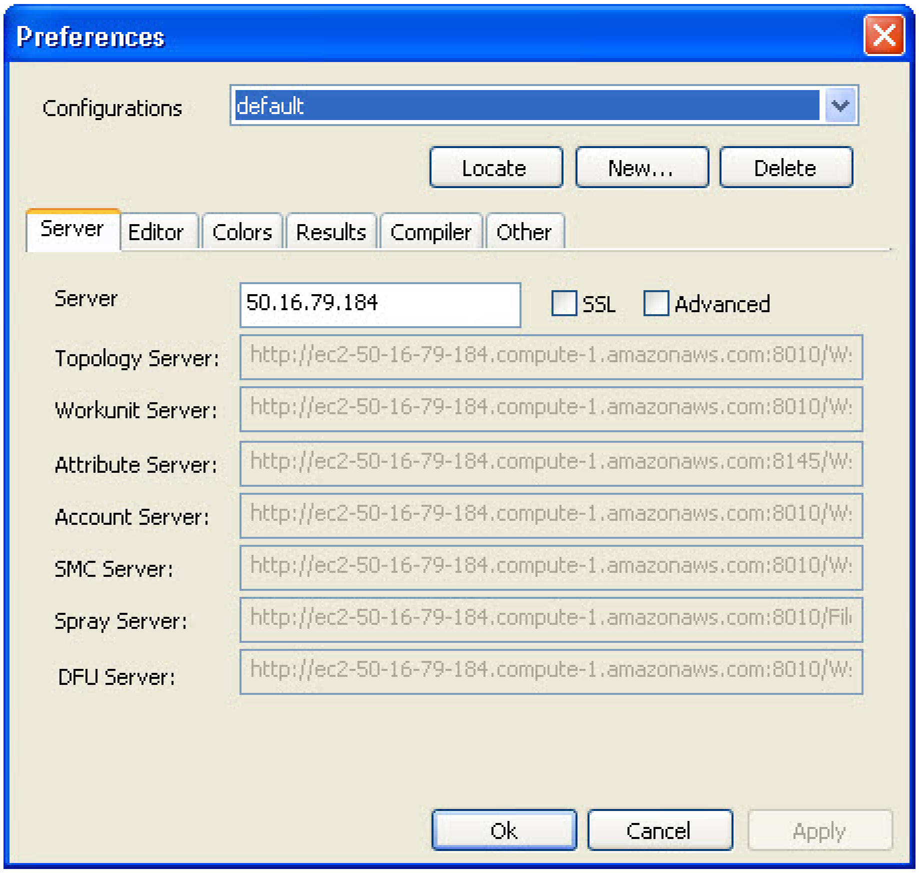
1. Open the ECL IDE on your Windows workstation, from your start menu. (**Start** >> **All Programs** >> **HPCCSystems** >> **ECL IDE**).

> 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



Configurations: default

Server: 50.16.79.184 □ SSL □ Advanced

Topology Server: http://ec2-50-16-79-184.compute-1.amazonaws.com:8010/W:

Workunit Server: http://ec2-50-16-79-184.compute-1.amazonaws.com:8010/W:

Attribute Server: http://ec2-50-16-79-184.compute-1.amazonaws.com:8145/W:

Account Server: http://ec2-50-16-79-184.compute-1.amazonaws.com:8010/W:

SMC Server: http://ec2-50-16-79-184.compute-1.amazonaws.com:8010/W:

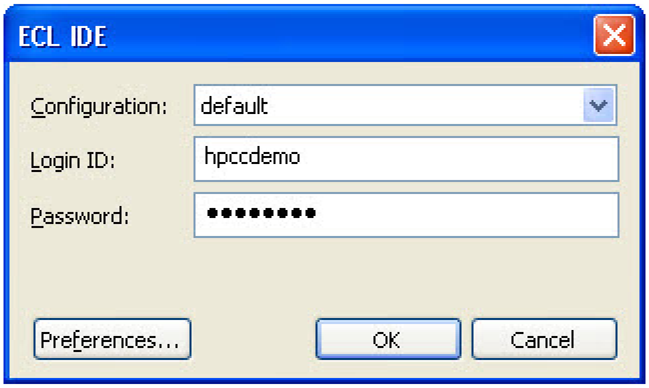
Spray Server: http://ec2-50-16-79-184.compute-1.amazonaws.com:8010/Fil

DFU Server: http://ec2-50-16-79-184.compute-1.amazonaws.com:8010/W:

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

Login ID	**hpccdemo**
Password	**hpccdemo**

Figure 14. Login Window



Configuration: default

Login ID: hpccdemo

Password: ●●●●●●●●●

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

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

This could also be written as:

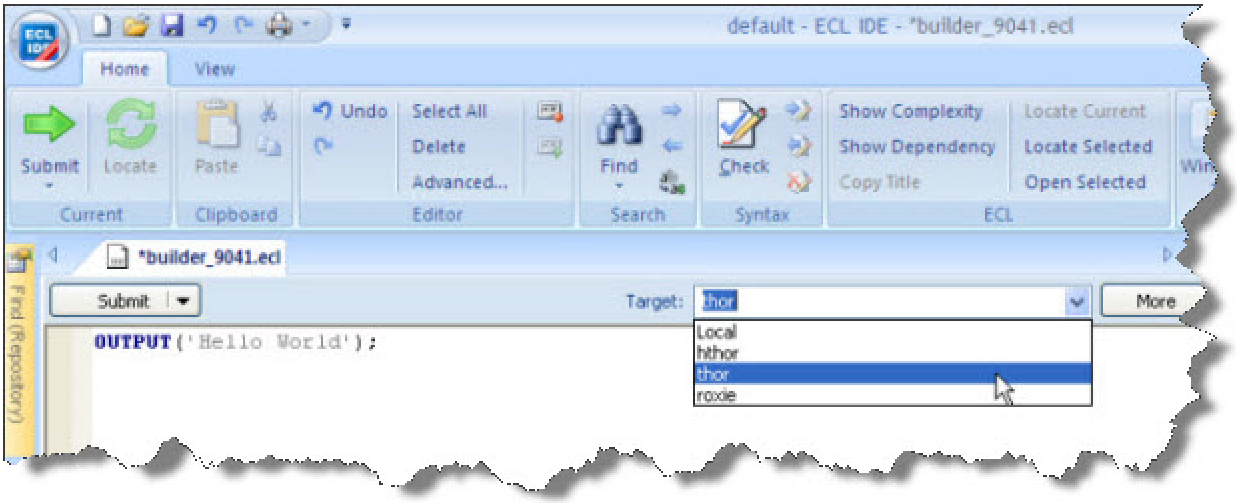
```
'Hello World';
```

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

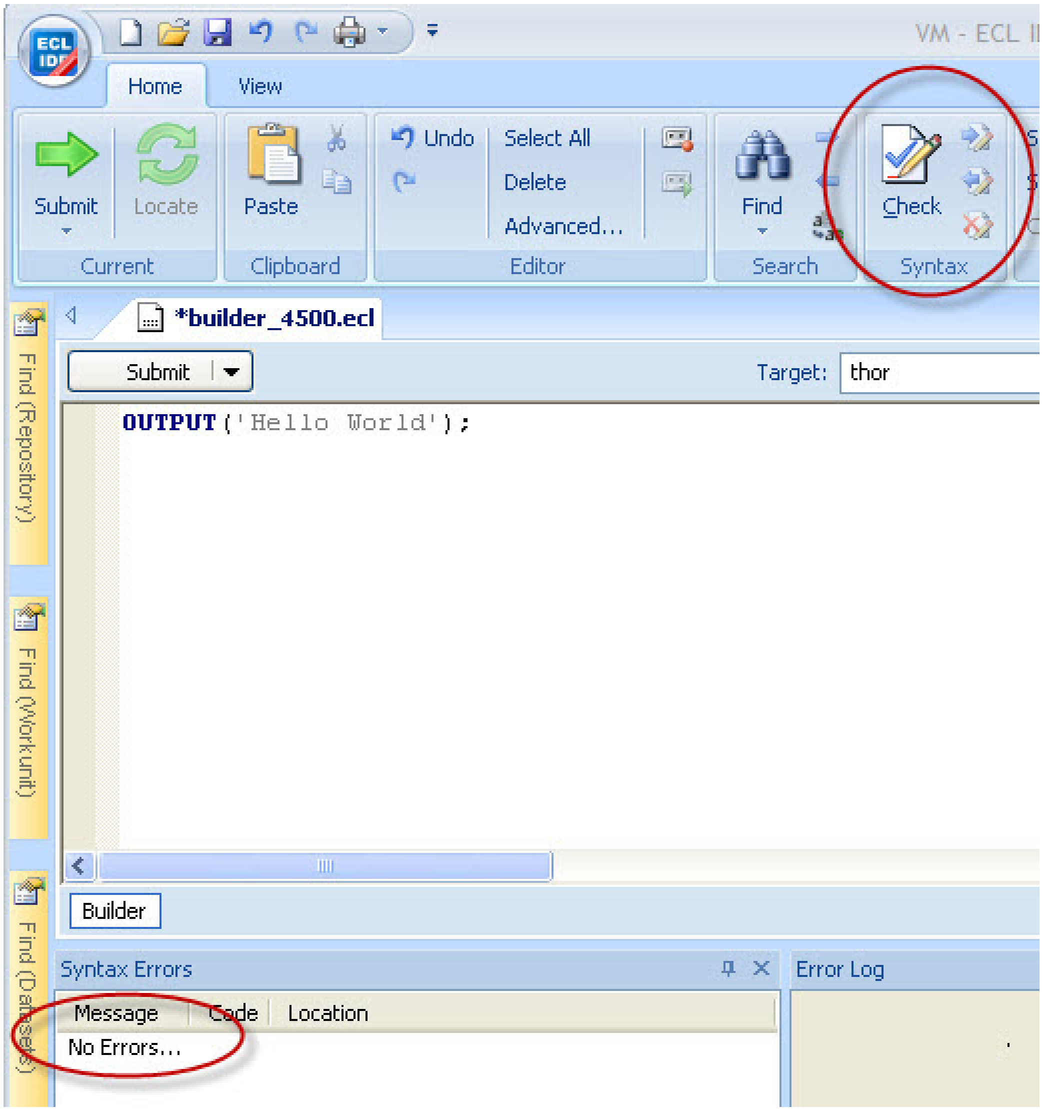


Target: thor

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

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

Figure 16. Syntax Check



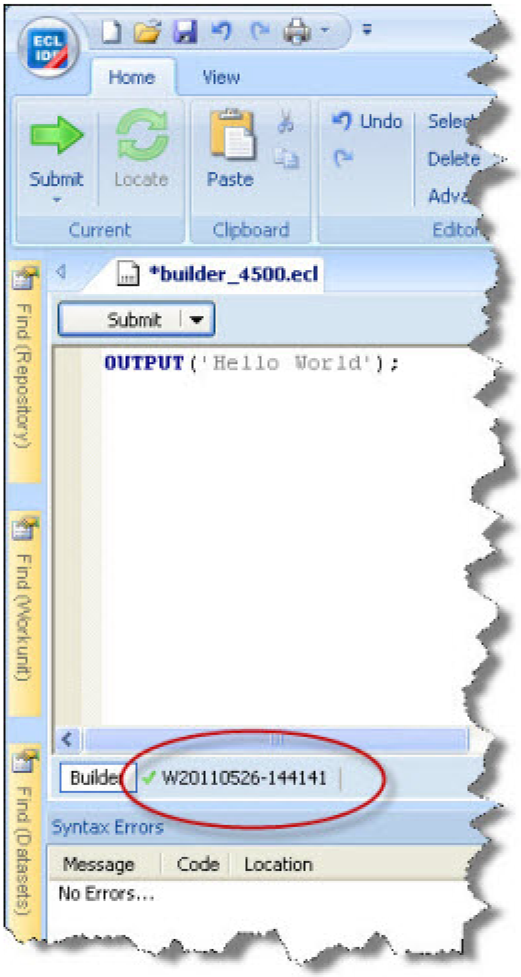
```
OUTPUT('Hello World');
```

No Errors...

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

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

Figure 17. Completed job



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

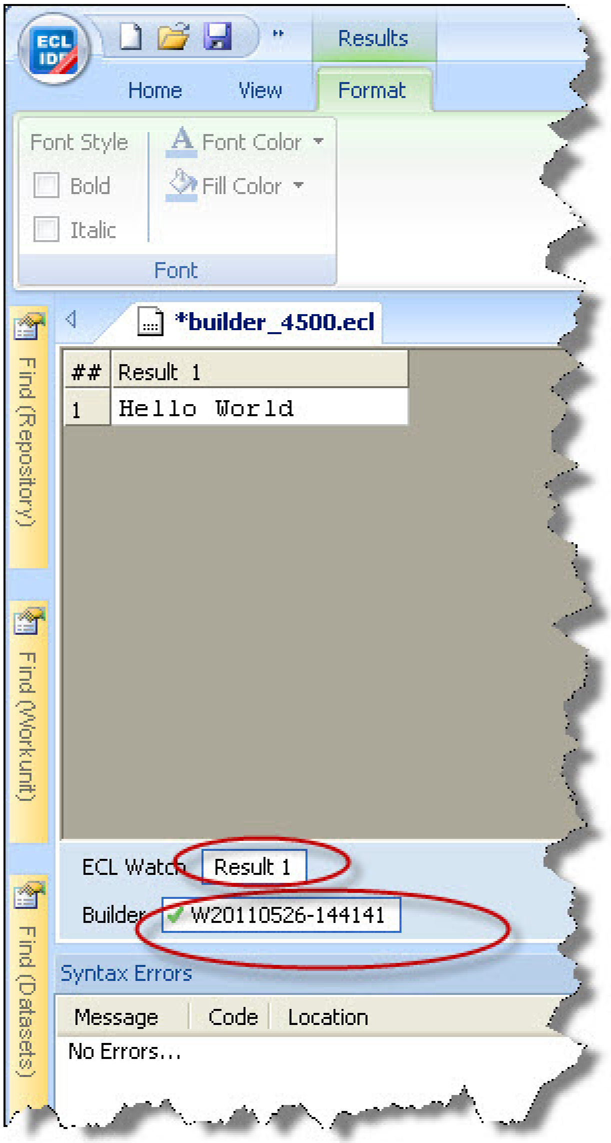
✓ W20110526-144141

Message	Code	Location
No Errors...		

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



##	Result 1
1	Hello World

Message	Code	Location
No Errors...		

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:

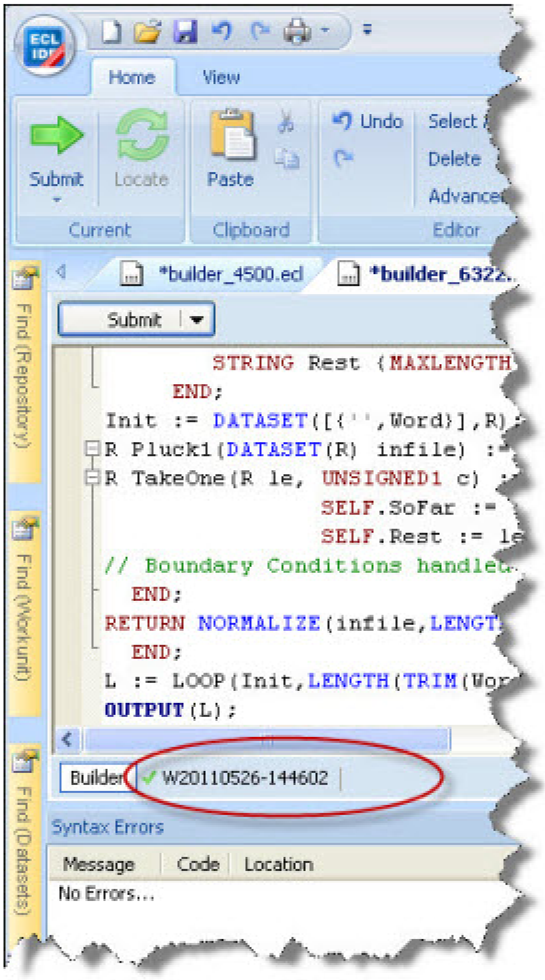
```
STRING Word := 'FRED' :STORED('Word');
R := RECORD
        STRING SoFar {MAXLENGTH(200)};
        STRING Rest {MAXLENGTH(200)};
     END;
Init := DATASET([{'',Word}],R);
R Pluck1(DATASET(R) infile) := FUNCTION
R TakeOne(R le, UNSIGNED1 c) := TRANSFORM
                SELF.SoFar := le.SoFar + le.Rest[c];
                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



```
            STRING Rest {MAXLENGTH
        END;
Init := DATASET([{'',Word}],R);
R Pluck1(DATASET(R) infile) :=
R TakeOne(R le, UNSIGNED1 c) :
                SELF.SoFar :=
                SELF.Rest := le
// Boundary Conditions handled
    END;
RETURN NORMALIZE(infile,LENGT
    END;
L := LOOP(Init,LENGTH(TRIM(Wor
OUTPUT(L);
```

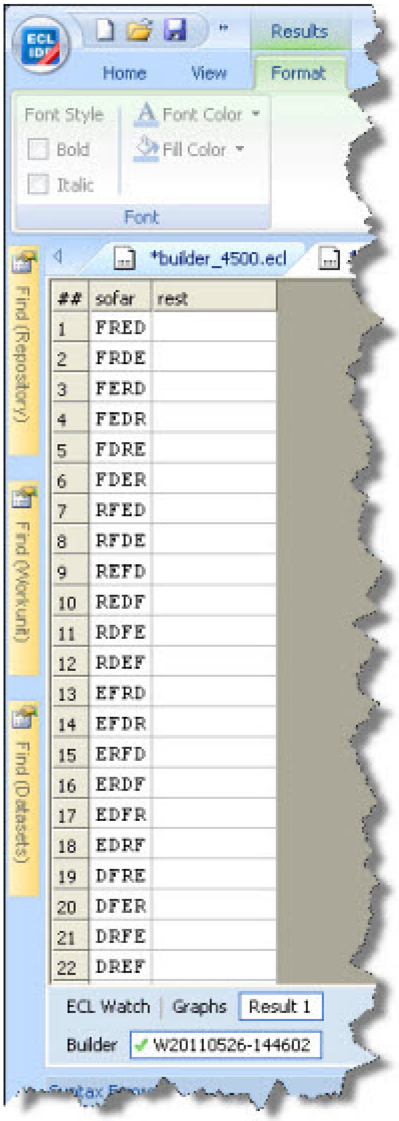
✓ W20110526-144602

Message	Code	Location
No Errors...		

The green check mark indicates successful completion.

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

Figure 20. Completed job output



##	sofar	rest
1	FRED	
2	FRDE	
3	FERD	
4	FEDR	
5	FDRE	
6	FDER	
7	RFED	
8	RFDE	
9	REFD	
10	REDF	
11	RDFE	
12	RDEF	
13	EFRD	
14	EFDR	
15	ERFD	
16	ERDF	
17	EDFR	
18	EDRF	
19	DFRE	
20	DFER	
21	DRFE	
22	DREF	

Builder ✓ W20110526-144602

Anagram2

In this example, we will download an open source data file of dictionary words, spray[1] 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).

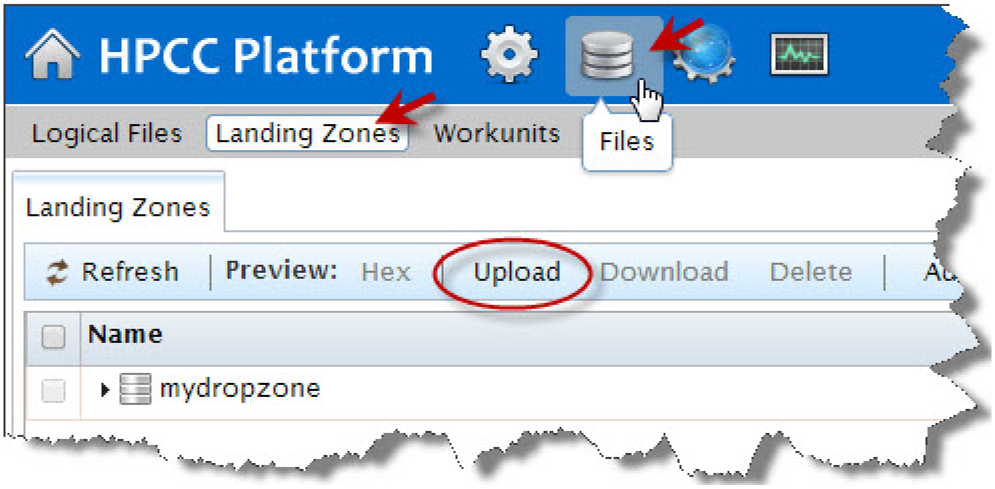
⚠	Your IP address could be different from the ones provided in the example images. Please use the IP address provided by **your** installation.

[1] 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.

2. From ECL Watch click on the **Files** icon, then click the **Landing Zones** link from the navigation sub-menu.

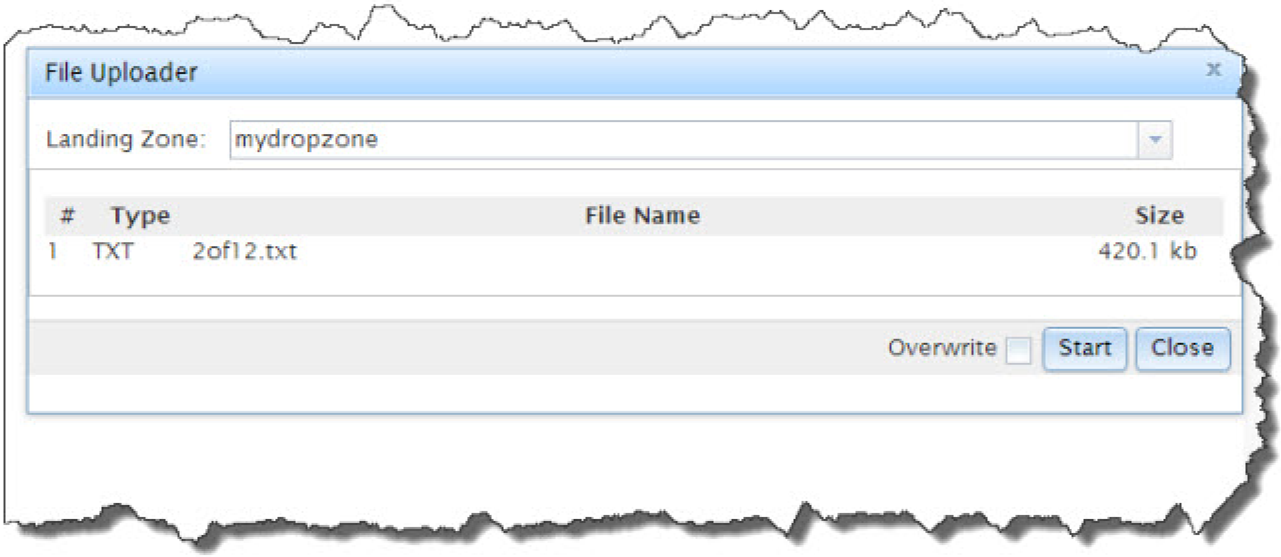
   Press the **Upload** action button.

Figure 21. Upload



3. A dialog opens. **Browse** your local machine select the file to upload and then press the **Open** button.

Figure 22. File Uploader



Landing Zone: mydropzone

#	Type	File Name	Size
1	TXT	2of12.txt	420.1 kb

The file you selected should appear in the **File Name** field. The data file is named: **2of12.txt**.

4. Press the **Start** button 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.

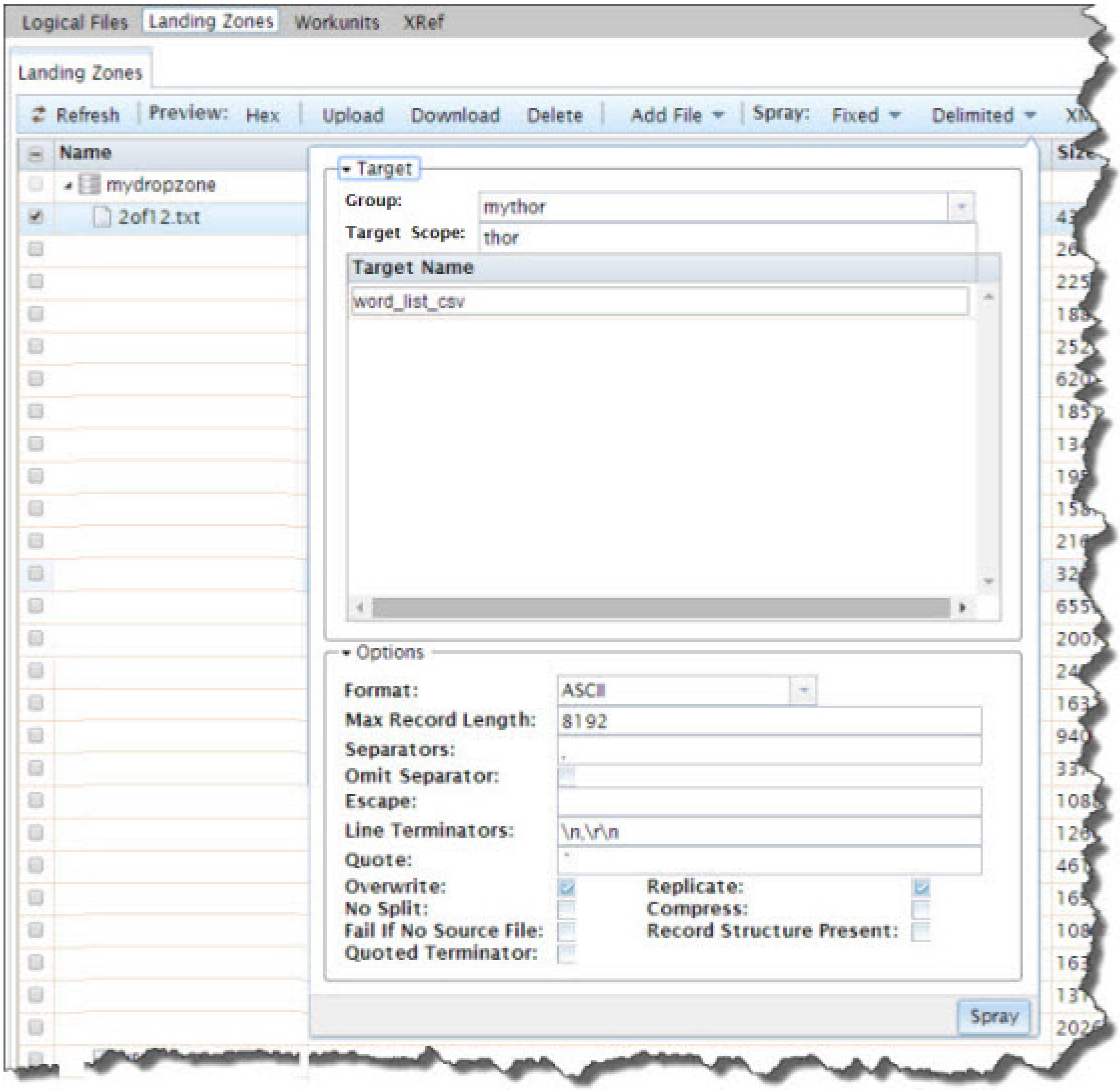
1. In a Web browser, connect to ECL Watch using http://**<PUBLIC_DNS>:8010** (where PUBLIC_DNS is the public DNS name of your ESP server).

2. Click on the **Files** icon, then click the **Landing Zones** link from the navigation sub-menu. Select the appropriate landing zone (if there are more than one landing zones). Click the arrow to the left of your landing zone to expand it.

3. Select the file from your drop zone by checking the box next to it.

4. Check the box next to 2of12.txt, then press the **Delimited** button.

Figure 23. Spray Delimited



Group: mythor
Target Scope: thor
Target Name: word_list_csv

Format: ASCII
Max Record Length: 8192
Separators: ,
Omit Separator: □
Escape:
Line Terminators: \n,\r\n
Quote: '
Overwrite: ☑
Replicate: ☑
No Split: □
Compress: □
Fail If No Source File: □
Record Structure Present: □
Quoted Terminator: □

The **DFU Spray Delimited** page displays.

5. Select mythor in the Target Group drop list.

6. Complete the Target Scope as *thor*.

7. 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: '

8. Fill in the Target Name using the rest of the Logical File name desired: word_list_csv

9. Make sure the **Overwrite** box is checked.

   If available, make sure the **Replicate** box is checked. (The Replicate option is only available on systems where replication has been enabled.)

10. Press the **Spray** button.

    A tab displays the DFU Workunit where you can see the progress of the spray.

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, 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 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.

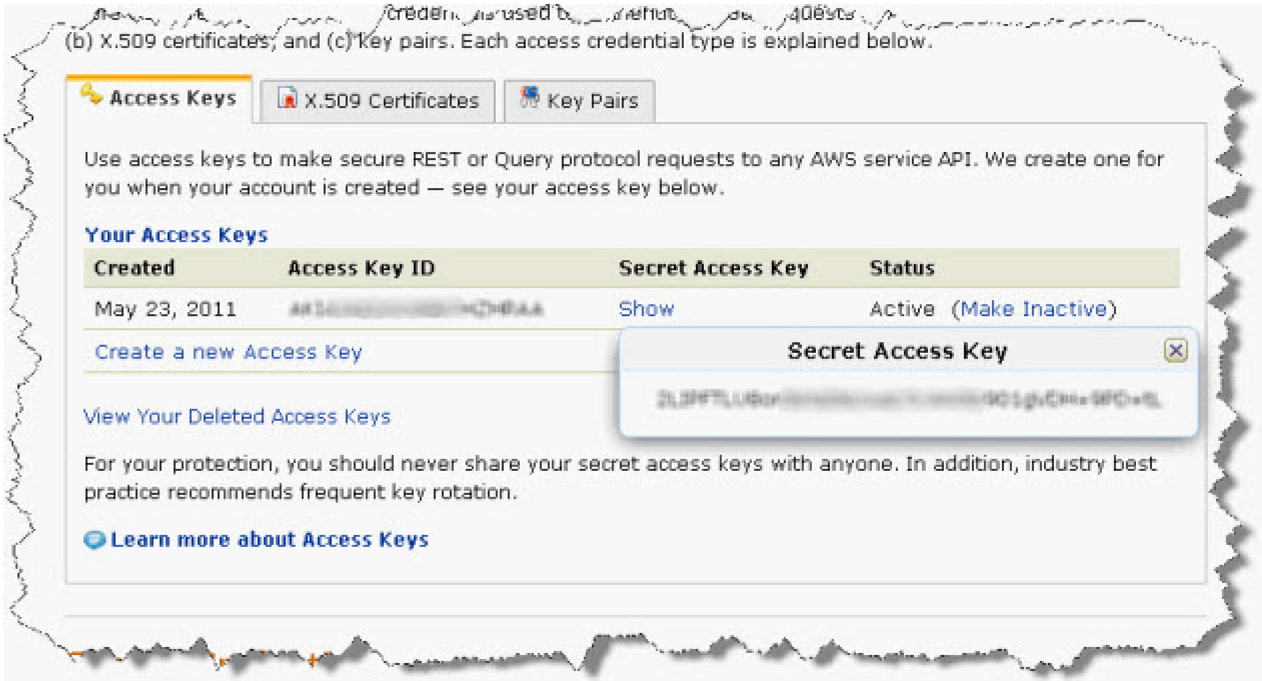
Using S3 buckets

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 24. Credentials



(b) X.509 certificates, and (c) Key pairs. Each access credential type is explained below.

Use access keys to make secure REST or Query protocol requests to any AWS service API. We create one for you when your account is created — see your access key below.

Your Access Keys

Created	Access Key ID	Secret Access Key	Status
May 23, 2011	[redacted]	Show	Active (Make Inactive)

Secret Access Key

[redacted]

For your protection, you should never share your secret access keys with anyone. In addition, industry best practice recommends frequent key rotation.

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

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