

HPCC Systems® Instant Cloud para AWS

Equipe de documentação de Boca Raton



HPCC Systems® Instant Cloud para AWS

Equipe de documentação de Boca Raton

Copyright © 2022 HPCC Systems®. All rights reserved

Sua opinião e comentários sobre este documento são muito bem-vindos e podem ser enviados por e-mail para <docfeedback@hpccsystems.com>

Inclua a frase **Feedback sobre documentação** na linha de assunto e indique o nome do documento, o número das páginas e número da versão atual no corpo da mensagem.

LexisNexis e o logotipo Knowledge Burst são marcas comerciais registradas da Reed Elsevier Properties Inc., usadas sob licença.

HPCC Systems é uma marca comercial registrada da LexisNexis Risk Data Management Inc.

Amazon Web Services, AWS, Amazon EC2, EC2, Amazon Elastic Compute Cloud, Amazon S3, Amazon Simple Storage Service, são marcas comerciais, marcas registradas ou identidades visuais da AWS nos EUA e/ou em outros países.

Os demais produtos, logotipos e serviços podem ser marcas comerciais ou registradas de suas respectivas empresas.

Todos os nomes e dados de exemplo usados neste manual são fictícios. Qualquer semelhança com pessoas reais, vivas ou mortas, é mera coincidência.

2022 Version 8.4.64-1

Introdução	4
Pré-requisitos e Premissas	5
Utilizando o Instant Cloud Launch	6
Antes de Começar	6
Login	7
Iniciar um Novo cluster HPCC	8
Executando ECL	18
Executando ECL no seu cluster HPCC Systems	18
Mais exemplos ECL	25
Exemplo ECL: Anagram1	25
Anagram2	28
Manipulação dos Dados (Processamento dos Dados)	34
Utilizando o S3 buckets	34
Próximos passos	37

Introdução

Este guia fornece informações e orientação sobre como executar a plataforma HPCC Systems® dentro do Amazon Web Services (AWS) Elastic Cloud (EC2) usando a página Instant Cloud para AWS .

Isso permite **instanciar** e executar clusters do HPCC Systems de diferentes tamanhos em tempo real.

O procedimento é útil para:


- Validação de conceito
- Experimentação
- Aprendizagem
- Usar a plataforma HPCC Systems sem estar sujeito a custos administrativos e de hardware.
- Crie e use um cluster do HPCC Systems imediatamente sem precisar comprar ou instalar um novo hardware.

É possível criar um cluster pequeno para tarefas menores ou clusters maiores para tarefas grandes. Essa flexibilidade permite combinar custo e poder de processamento para a tarefa em questão.

Instanciar nós EC2 temporários permite “locar” a capacidade computacional sem assumir compromissos em longo prazo. Dessa forma, você paga por utilização em vez de desembolsar um valor fixo alto logo no início.

Lembre-se de que você deve encerrar quaisquer instâncias desnecessárias para evitar pagar por um tempo de computação desnecessário. Todos os custos da AWS são de sua total responsabilidade.

O Instant Cloud está sendo atualizado para fornecer serviços gerenciados como Amazon Elastic Map Reduce (EMR). Integração S3, elasticidade, backup e recuperação são recursos em consideração.

	Sugerimos a leitura completa deste documento antes de começar.
---	---

Pré-requisitos e Premissas

Você vai precisar de:

- Uma conta Amazon Web Services com EC2 habilitado
- Uma estação de trabalho com conexão à Internet para acessar o Amazon Web Services; A estação de trabalho pode ser um:
 - *PC (Computador) Windows ou*
- Um navegador de Internet (Firefox, Internet Explorer, ou Chrome)

Opcionalmente é desejável ter:

- Uma ferramenta SSH, como PuTTY
- Uma ferramenta de geração e conversão de chaves, como PuTTYGen
- Uma ferramenta de cópia de segurança (tais como WinSCP)
- Familiaridade com navegação em sistemas de arquivo Linux



Para obter instruções detalhadas da Amazon sobre PuTTY/pcsp/PuTTYGen, acesse:

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

Utilizando o Instant Cloud Launch

Antes de Começar

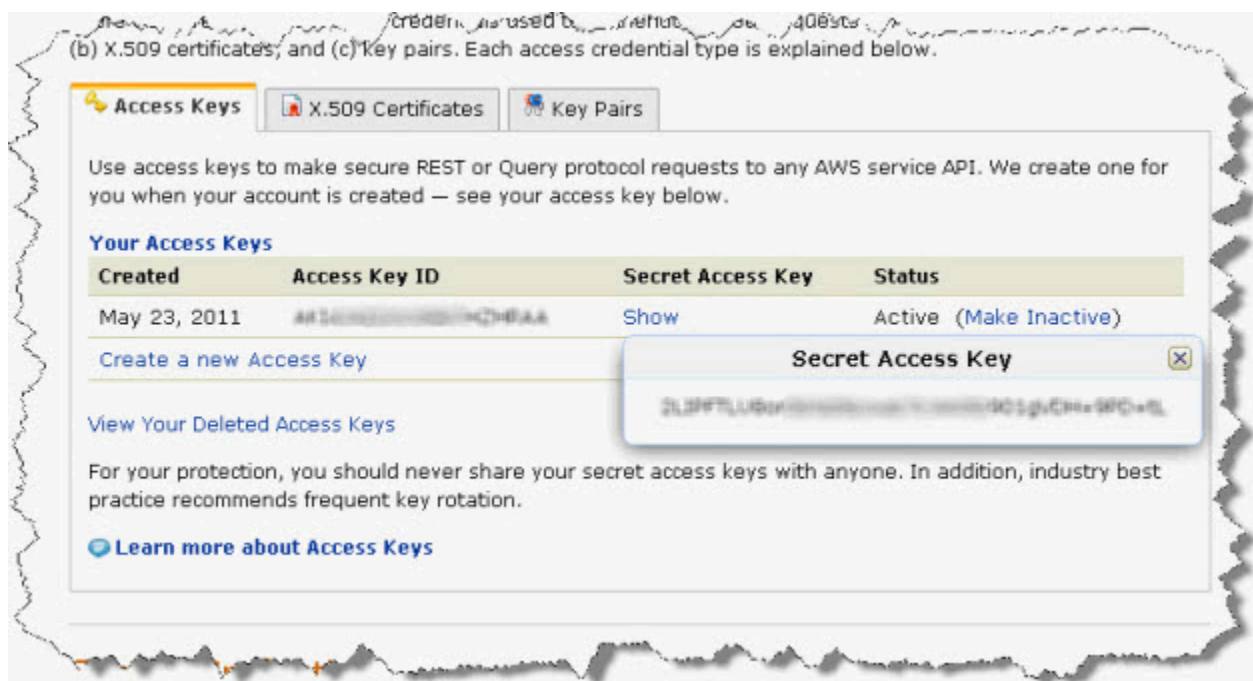
Nesta seção, reuniremos algumas informações necessárias antes de começar. Isso inclui:

- Sua ID de chave de acesso e a Chave de acesso secreta do AWS
- O tamanho do cluster desejado.

Encontre sua Amazon Access Key ID e Secret Access Key

1. Acesse **aws.amazon.com** e faça o login, se necessário.
2. Selecione **Account**.
3. Selecione **Credenciais**.
4. Na página, procure a seção **Access Credentials**.
5. Anote sua **Access Key ID** e sua **Secret Access Key**

Figure 1. Credenciais



Partes desta imagem foram intencionalmente desfocadas

Access Key ID	
Secret Access Key	

Login

1. Abra o navegador e acesse <https://aws.hpccsystems.com/>.

Caso não tenha feito o login, o link do **Login** será exibido no topo da página. Se a tela exibir o link **Logout**, isso significa que você já está logado no sistema.

2. Especifique sua **Access Key ID** e **Secret Access Key**. Estas informações nunca serão armazenadas em nossos sistemas.

Caso estas informações não estejam acessíveis, clique no link abaixo de **Can't find your Access Key ID?** para ir à seção de AWS Management Console.

Figure 2. Login

Getting Started | Comments | FAQs | HPCC AWS Forum | AWS Management Console

HPCC Systems

HPCC Systems® Instant Cloud for AWS Beta

● Login ● AWS Activities ● Resources ● Code Samples

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
AKIAI7J7ETQ3UCMB6W1Q

Secret access key
.....

* You are solely responsible for all AWS charges.

☒ I accept [Terms of Use](#) and agree with above

Login

Can't find your Access Key ID? [Click here](#)

Partes desta imagem foram intencionalmente desfocadas

3. Marque a caixa de seleção para aceitar os Termos de uso.
4. Pressione o botão **Login**.

A janela **View Clusters** será exibida. A janela mostrará todos os clusters iniciados. Aqui, você pode acessar o link para iniciar um novo cluster.

5. Clique no link **Launch Cluster** no topo da página.

A janela **Launch a New HPCC** será exibida.

Iniciar um Novo cluster HPCC

Nesta seção, iniciaremos um conjunto de máquinas Ubuntu 12.04 que serão usadas em sua plataforma Thor do HPCC Systems. A página Instant Cloud usa a entrada que você especificar para criar os clusters.

Ao pressionar o botão **Launch Cluster** , você poderá:

- Cria um nome de cluster exclusivo.
- Cria um Grupo de segurança com acesso às portas TCP e UDP habilitado.
- Cria um Par de chaves.
- Inicia o número de nós m1.large solicitados usando AMI (ami-e01698d0) fornecida.
- Reúne IPs privado e público.
- Instala os pacotes da plataforma HPCC Systems.
- Configura o Cluster Thor, o Cluster Roxie, e os nós de suporte exigidos.
- Cria o usuário interno (HPCC).
- Propaga o arquivo environment.xml para todos os nós.
- Inicializa todos os componentes.

Iniciar um Novo cluster Thor

1. Especifique a quantidade de nós Thor e Roxie a ser instanciada.

Figure 3. Iniciar um Novo cluster Thor

The screenshot shows the 'Launch A New HPCC Cluster' interface. At the top, it says 'HPCC Systems' and 'HPCC Systems® Instant Cloud for AWS Beta'. Below this is a navigation bar with links: 'Launch Cluster', 'View Clusters', 'AWS Activities', 'Resources', 'Code Samples', and 'Log Out'. The main section is titled 'Launch A New HPCC Cluster' with a warning: '*You are solely responsible for all AWS charges.' Below this is a form with the following fields: 'Region' (a dropdown menu showing 'Oregon'), 'Thor Nodes' (a text input with '6'), 'Roxie Nodes' (a text input with '3'), 'Support Nodes' (a text input with '1'), and 'Total Nodes*' (a text input with '10'). There is also a 'Snapshot IDs (optional)' text input field. At the bottom of the form is a 'Launch Cluster' button.

2. Opcionalmente, especifique a(s) ID(s) da imagem para anexar os dados à sua zona de entrada de arquivos.

Isso seria uma "imagem" salva anteriormente da armazenagem de dados da zona de entrada de arquivos.

3. Pressione o botão **Launch Cluster**.

A janela **Cluster Launch Log** será exibida. Essa janela mostra os detalhes durante a inicialização (ela é atualizada automaticamente durante a inicialização ou encerramento). Ela também exibe sua ID do Cluster (um identificador exclusivo) que pode ser útil para identificar o cluster quando mais de um cluster estiver sendo executado.

HPCC Systems

HPCC Systems® Instant Cloud for AWS Beta

[Launch Cluster](#) [View Clusters](#) [AWS Activities](#) [Resources](#) [Code Samples](#) [Log Out](#)

HPCC Cluster Launch Log - Hpcc-XUWT

Status: **Ready**

```
2012/11/13 12:37:19 - 50.112.221.221: Starting mydatafsrv.... [ OK ]
2012/11/13 12:37:19 - 50.112.221.221: Starting myecdscheduler.... [ 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 mydatafsrv.... [ OK ]
2012/11/13 12:37:19 - 50.112.229.41: Starting myecdscheduler.... [ 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 mydatafsrv.... [ OK ]
2012/11/13 12:37:19 - 54.245.6.128: Starting myecdscheduler.... [ OK ]
2012/11/13 12:37:19 - 54.245.6.128: Start node is complete.
2012/11/13 12:37:19 - 54.245.6.128: Starting mydatafsrv.... [ OK ]
2012/11/13 12:37:19 - 54.245.6.128: Starting myecdscheduler.... [ 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 mydatafsrv.... [ OK ]
2012/11/13 12:37:20 - 50.112.236.154: Starting myecdscheduler.... [ OK ]
2012/11/13 12:37:20 - 50.112.236.154: Start node is complete.
2012/11/13 12:37:20 - 50.112.236.154: Starting mydatafsrv.... [ OK ]
2012/11/13 12:37:20 - 50.112.236.154: Starting myecdscheduler.... [ 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 mydatafsrv.... [ OK ]
2012/11/13 12:37:20 - 50.112.200.120: Starting myecdscheduler.... [ OK ]
2012/11/13 12:37:20 - 50.112.200.120: Start node is complete.
2012/11/13 12:37:20 - 50.112.200.120: Starting mydatafsrv.... [ OK ]
2012/11/13 12:37:20 - 50.112.200.120: Starting myecdscheduler.... [ OK ]
2012/11/13 12:37:20 - 50.112.200.120: Start node is complete.
2012/11/13 12:37:24 - 50.112.228.178: Starting mydatafsrv.... [ OK ]
2012/11/13 12:37:24 - 50.112.228.178: Starting myecdscheduler.... [ OK ]
2012/11/13 12:37:24 - 50.112.228.178: Start node is complete.
2012/11/13 12:37:24 - 50.112.228.178: Starting mydatafsrv.... [ OK ]
2012/11/13 12:37:24 - 50.112.228.178: Starting myecdscheduler.... [ OK ]
2012/11/13 12:37:24 - 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. Clique no link [View Clusters](#) para ver os clusters que estão sendo executados.

Essa lista possui links para a página do ECL Watch, para a página Iniciar Log, para o arquivo de configuração do cluster, para a lista de IPs, e para a chave SSH.

Também possui um link que permite **Encerrar** a instanciação do cluster.

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	EIP 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

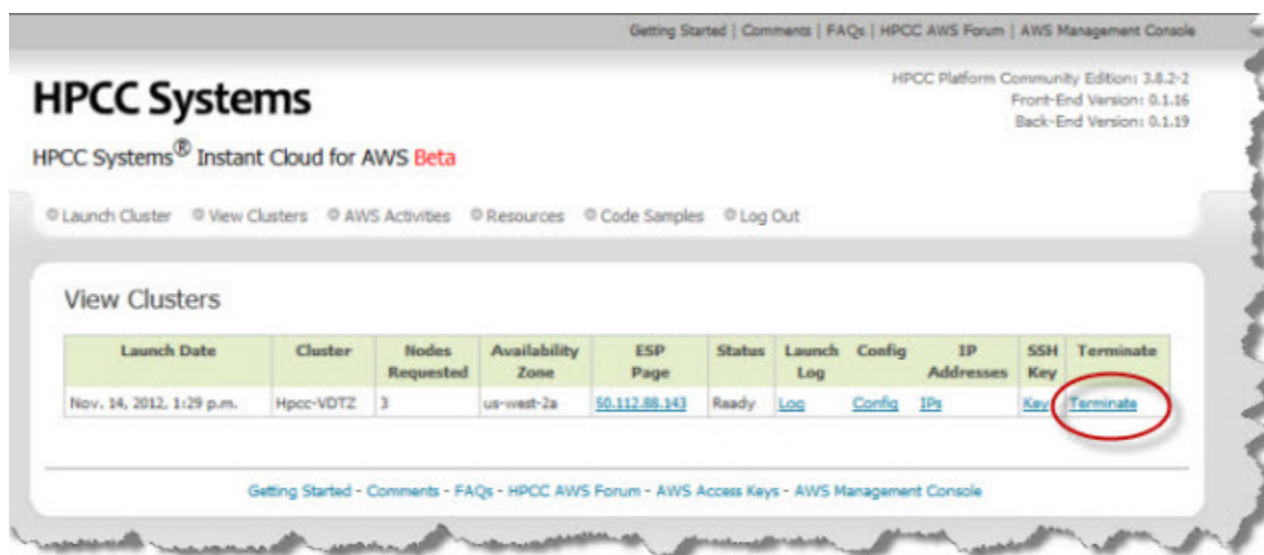
Encerrando as instâncias

Se precisar salvar seus dados, primeiro será necessário fazer o despray (consolidar dados dos nós) e salvar do seu cluster antes de desligar. Informações adicionais sobre o processamento de dados em uma plataforma HPCC Systems estão disponíveis no manual *Data Handling* (Processamento de dados). Consulte a seção “Próximos passos” para obter informações sobre como fazer o download de outros manuais.

Para encerrar seu cluster:

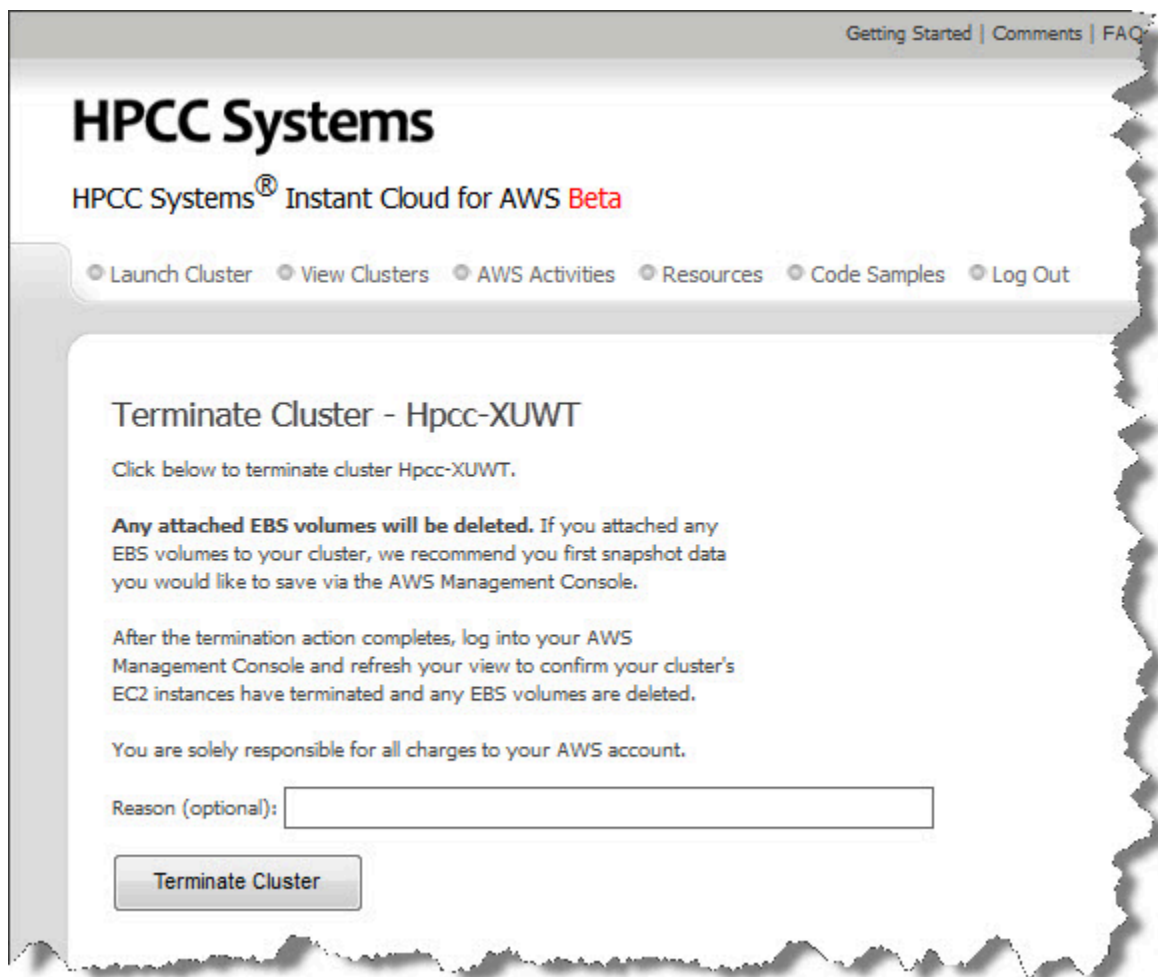
1. Abra a página **View Clusters** usando o link localizado no topo da página.

Figure 6. Clusters em Execução



2. Clique no link [Terminate](#) próximo ao cluster que você deseja fechar.

Figure 7. Terminate Cluster



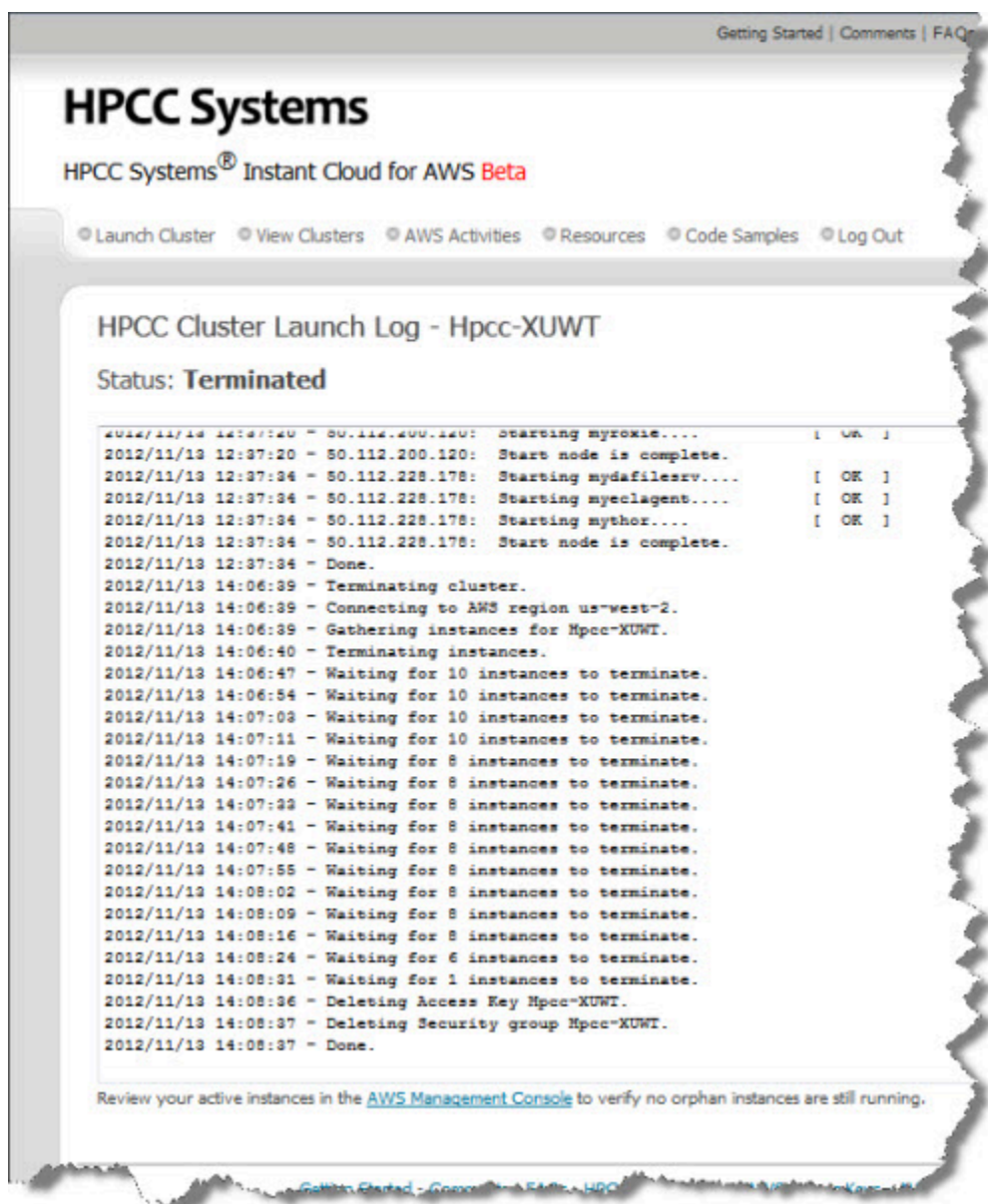
The screenshot shows the HPCC Systems web interface. At the top, there's a navigation bar with links: Getting Started | Comments | FAQ. Below this is the HPCC Systems logo and the text "HPCC Systems® Instant Cloud for AWS Beta". A secondary navigation bar contains links: Launch Cluster, View Clusters, AWS Activities, Resources, Code Samples, and Log Out. The main content area is titled "Terminate Cluster - Hpcc-XUWT". It contains the following text: "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.", and a text input field labeled "Reason (optional):". At the bottom of the form is a button labeled "Terminate Cluster".

3. Pressione o botão **Terminate Cluster** pressione o botão Encerrar Cluster e confirme quando solicitado.

A página **Cluster Launch Log** será exibida, mostrando a atividade durante o encerramento.

4. Aguarde até que o Log de inicialização do cluster diga **Status: Terminated**.

Figure 8. Cluster Encerrado



5. Como alternativa, acesse o Console de Gerenciamento do AWS (AWS management console) para confirmar que suas instâncias foram encerradas de forma adequada.

Todos os custos de sua conta AWS são de sua total responsabilidade.

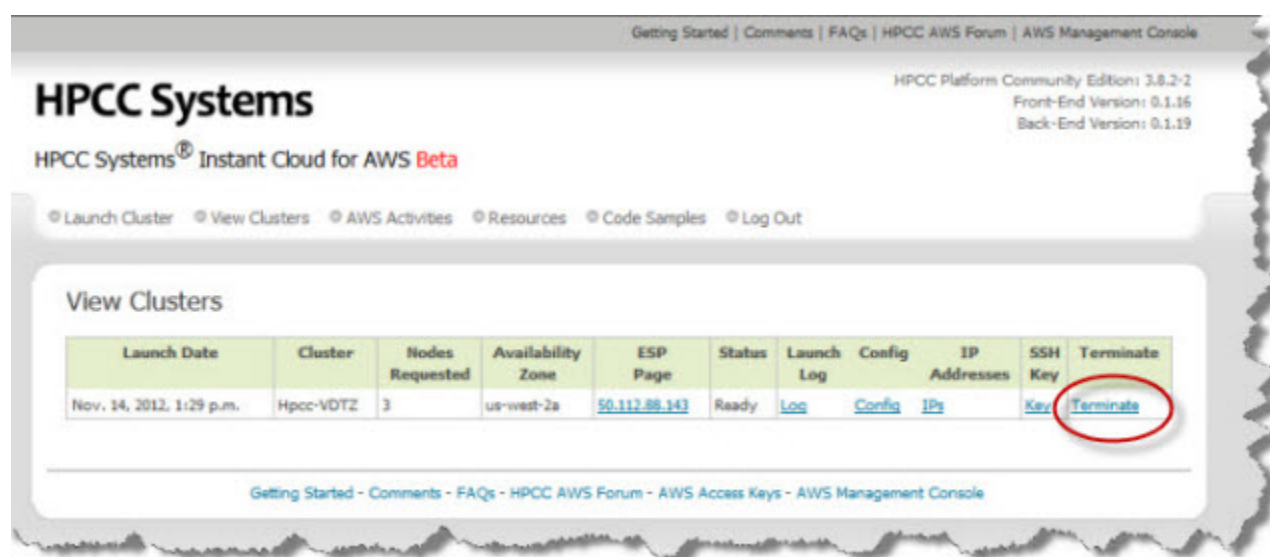
Outras Tarefas

Visualizar Clusters

Através da página **View Clusters** você pode acessar a Data/Hora de inicialização de cada cluster, a ID do cluster, o número de nós, a Zona, a página do ECL Watch, o Status, o Log de inicialização, o Arquivo de configuração, os endereços IP, e a chave SSH.

A página também fornece um link para encerrar um cluster com um único clique.

Figure 9. Clusters em Execução

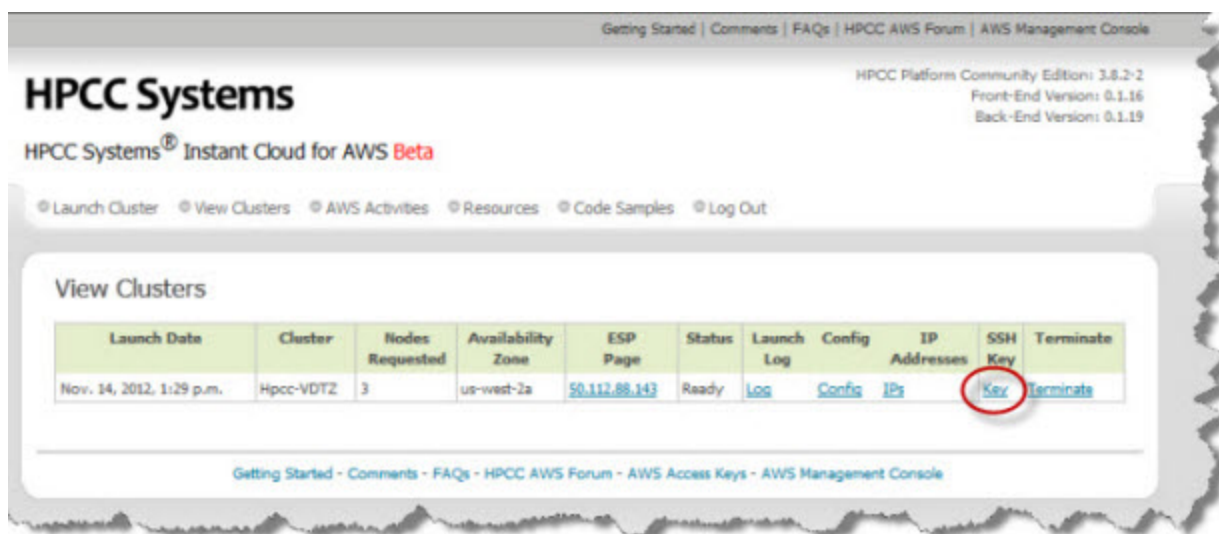


Gerenciar chaves SSH

A página de gerenciamento da Chave SSH permite baixar a chave SSH do seu cluster (arquivo .PEM) que será usada para autenticar a seção SSH, como por exemplo uma seção de console que utiliza PuTTY. A página também mostra como remover a chave do One-Click System.

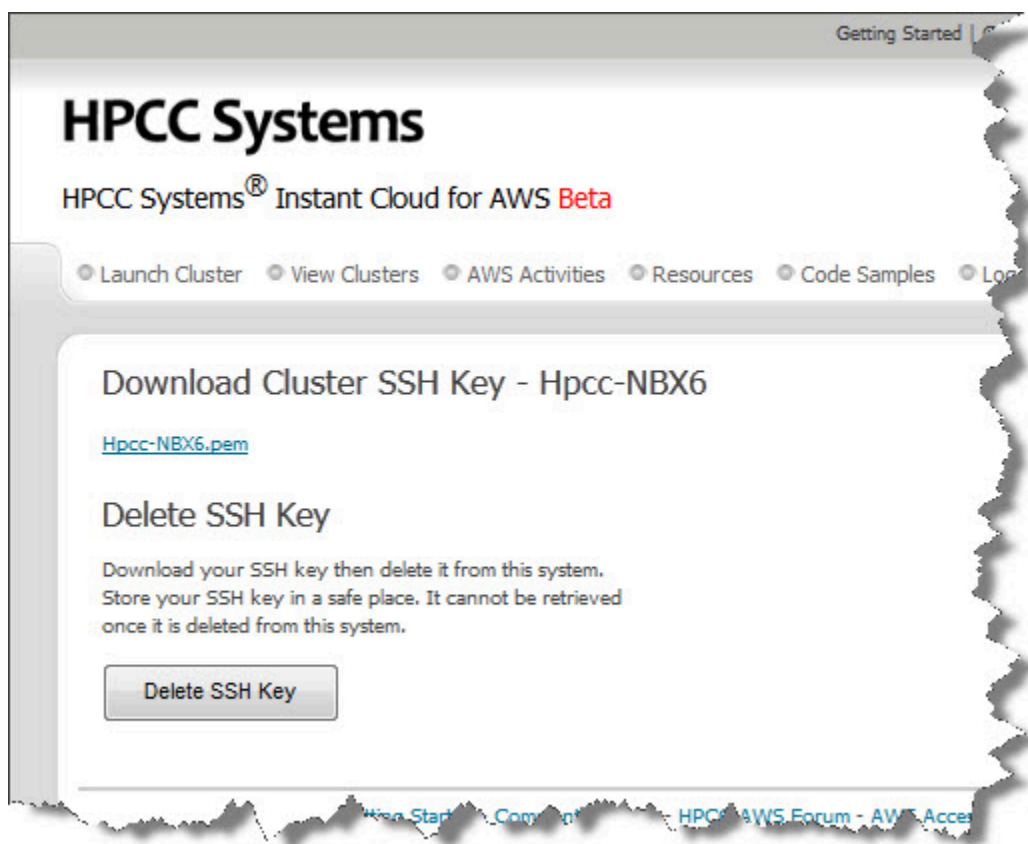
1. Abra a página **View Clusters** usando o link **View Clusters** localizado no topo da página.

Figure 10. Clusters em Execução



2. Clique no link [Key](#) próximo ao cluster.

Figure 11. Gerenciamento da chave



3. Clique no link [arquivo pem](#) para baixar a chave.

Este arquivo deve ser armazenado em um lugar seguro.

4. Pressione o botão **Delete SSH Key** para remover a chave SSH do One-Click system.

Observação: Este procedimento não remove as chaves do seu cluster em execução. Ele apenas remove as chaves do sistema Instant Cloud e impede downloads futuros da chave. Uma vez removida, a chave não pode ser recuperada.

Executando ECL

Executando ECL no seu cluster HPCC Systems

Agora que a plataforma está em execução, você pode criar e executar alguns ECL¹ Codifique usando o ECL IDE, o compilador ECL da linha de comando, ou a ferramenta ECLPlus.

Instalar o ECL IDE e HPCC Client Tools

O ECL IDE precisa ser instalado apenas uma vez. Pule esta seção caso ele já tenha sido instalado.

1. Em um navegador da Internet, conecte-se ao ECL Watch usando http://<PUBLIC_DNS>:8010 (onde PUBLIC_DNS é o nome do DNS público do seu servidor ESP).



Seu endereço IP poderá ser diferente dos endereços fornecidos nas imagens de exemplo. Favor usar o endereço IP do **seu** nó.

¹Enterprise Control Language (ECL) é uma linguagem de programação declarativa e centrada em dados usada para gerenciar todos os aspectos da junção, classificação e compilação de dados massivos que realmente diferenciam o HPCC (High Performance Computing Cluster) das demais tecnologias na sua capacidade de fornecer análise de dados flexíveis em escala massiva.

2. No menu ECL Watch Advanced, selecione o link **Additional Resources** .

Figure 12. Página ECL Watch Resource



Siga o link para a página do portal de download do HPCC System.

3. Clique no link **ECL IDE** . (ao lado direito da coluna Download, abaixo do título Free Community Edition)
4. Siga as instruções na página da Internet para instalar o ECL IDE.
5. Instale o ECL IDE, seguindo os prompts no programa de instalação. Após o ECL IDE ter sido instalado com sucesso, você pode prosseguir.

Executando o programa ECL do ECL IDE

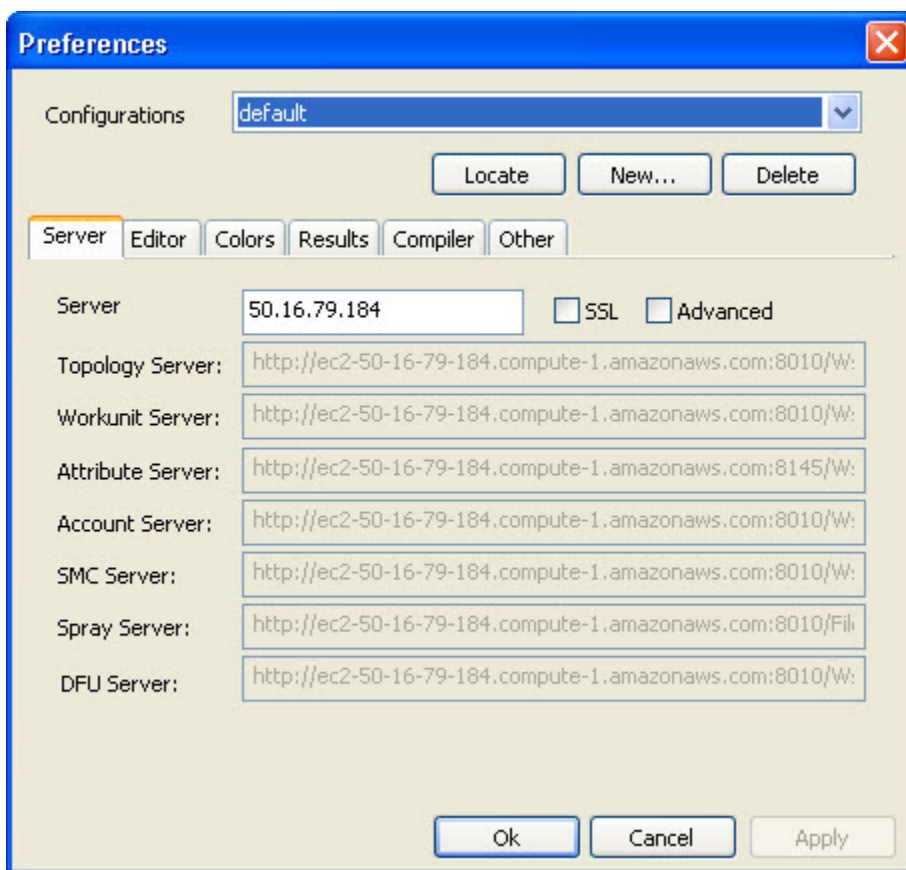
1. Abra o ECL IDE em sua estação de trabalho Windows, a partir do menu Iniciar. (**Start >> All Programs >> HPCCSystems >> ECL IDE**).



Você pode criar um atalho em sua área de trabalho para acessar rapidamente o ECL IDE.

2. Na janela de Login , pressione o botão **Preferences** .
3. No controle de entrada **Server** , digite o IP público do ESP Server de seu servidor ESP) e pressione o botão **Ok** .

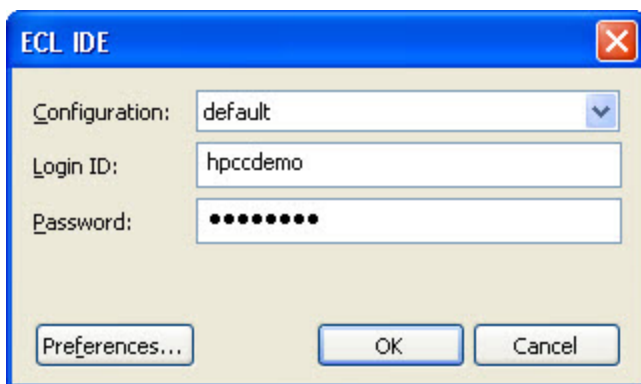
Figure 13. Janela de Login



4. Digite a **Login ID (ID do Login)** e **senha** fornecidas na caixa de diálogo Login.

Login ID	hpccdemo
Password	hpccdemo

Figure 14. Janela de Login



5. Abra uma nova **Janela do compilador** (CTRL+N) e escreva o seguinte código:

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

Isso também poderia ser escrito como:

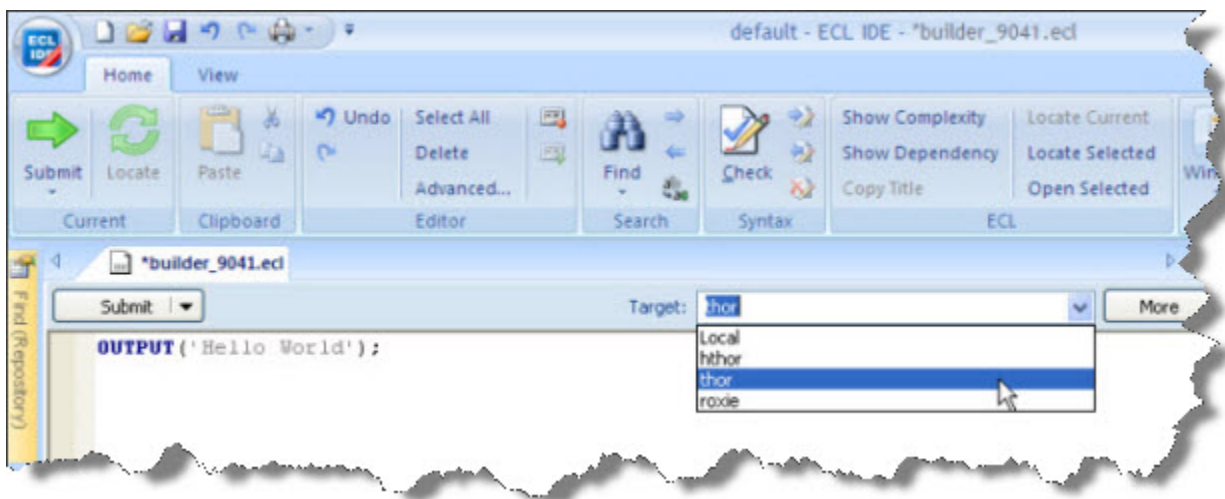
```
'Hello World';
```

Na segunda listagem de programa, a palavra-chave OUTPUT é ocultada. Isso é possível porque a linguagem é declarativa e a ação OUTPUT é implícita.

6. Selecione **thor** como seu cluster de destino.

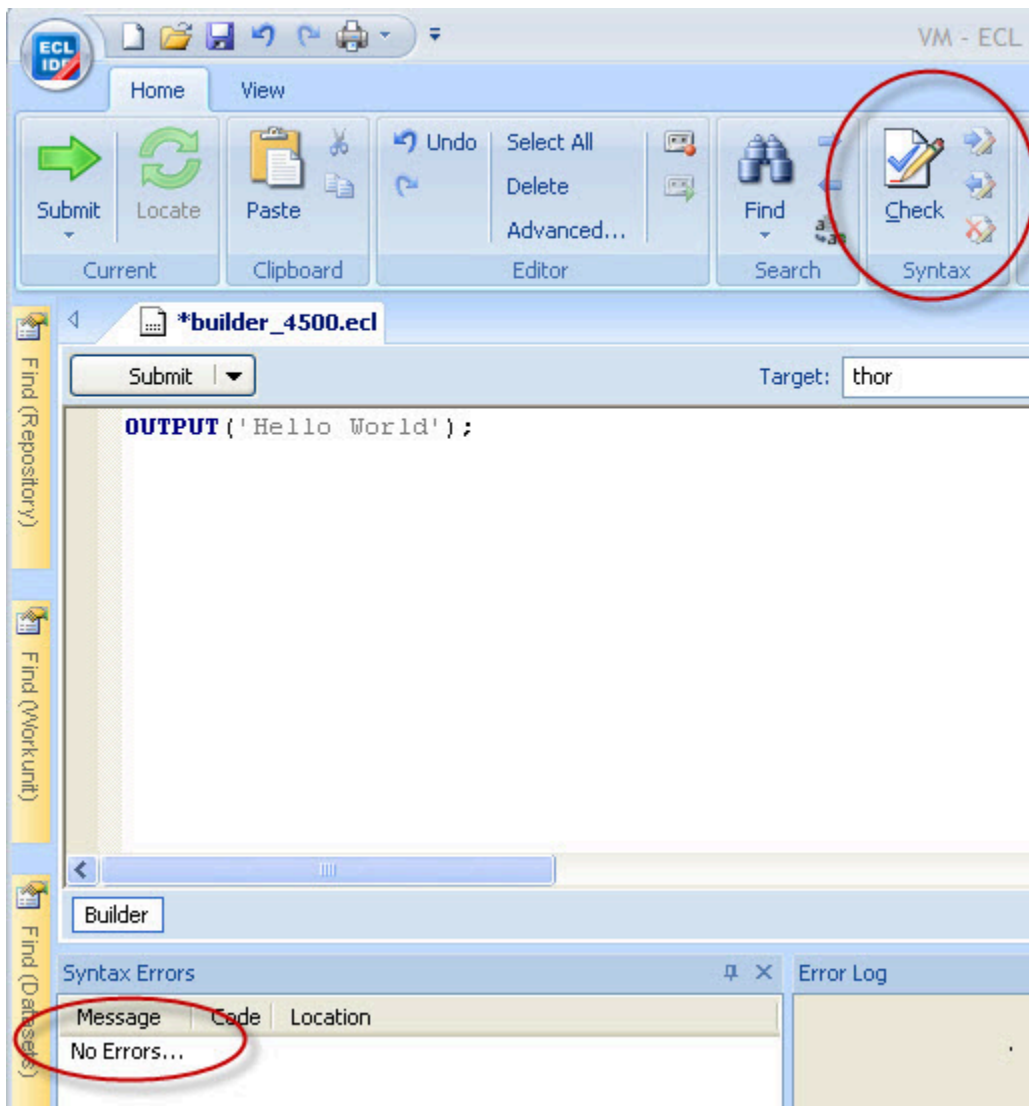
Thor é o componente da Refinaria de dados do seu HPCC. Trata-se de um cluster de computador massivamente paralelo baseado em disco, otimizado para classificar, manipular e transformar uma quantidade massiva de dados.

Figure 15. Selecionar destino



7. Pressione o botão de verificação de sintaxe localizado na barra de ferramentas principal (ou pressione F7).

Figure 16. Verificação de sintaxe



Uma verificação de sintaxe bem-sucedida exibe a mensagem “No errors...”.

8. Pressione o botão **Submit** (ou as teclas ctrl+enter).

Figure 17. Job Concluído



A marcação na cor verde indica uma conclusão bem-sucedida.

9. Clique na guia do número da workunit e, em seguida, na guia Result 1 para ver os resultados.

Figure 18. Resultado do job concluído



Mais exemplos ECL

Esta seção contém exemplos adicionais de ECL que podem ser usados em sua plataforma Thor do HPCC. Eles podem ser executados em um sistema de nó único ou em um cluster maior com vários nós.

Exemplo ECL: Anagrama1

Este exemplo pega uma STRING e gera todos os anagramas possíveis a partir dela. Este código serve de base para um segundo exemplo que analisa quais destas são palavras reais usando um arquivo de dados da lista de palavras.

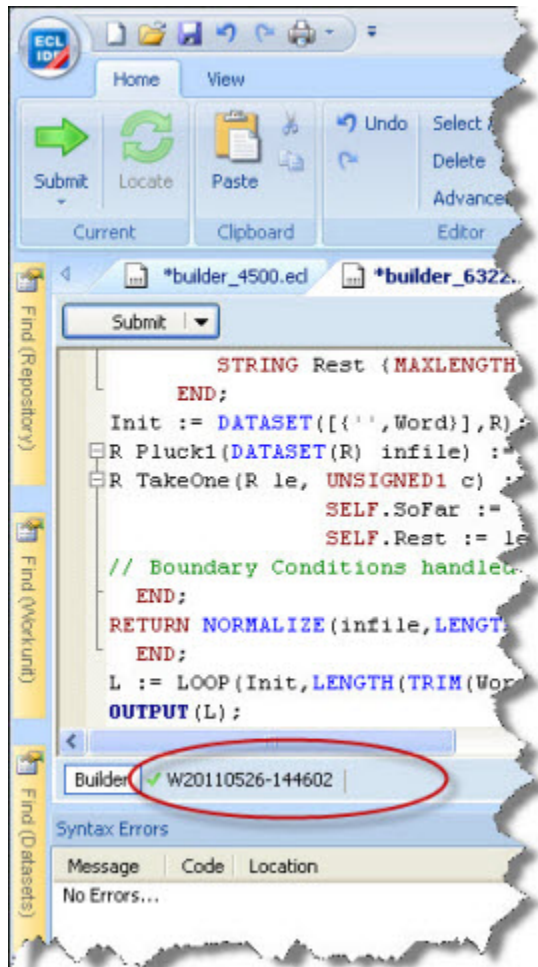
1. Abra o ECL IDE (**Start >> All Programs >> HPCC Systems >> ECL IDE**) e faça o login no HPCC.
2. Abra uma nova **Janela do compilador** (CTRL+N) e escreva o seguinte código:

```
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. Selecione **thor** como seu cluster de destino.
4. Pressione o botão de verificação de sintaxe localizado na barra de ferramentas principal (ou pressione F7)

5. Pressione o botão **Submit** (ou as teclas ctrl+enter).

Figure 19. Job Concluído



A marcação na cor verde indica uma conclusão bem-sucedida.

6. Clique na guia do número da workunit e, em seguida, na guia Result 1 para ver os resultados.

Figure 20. Resultado do job concluído

##	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	

Anagram2

Neste exemplo, vamos baixar um arquivo de dados de código público de palavras do dicionário, *spray*.¹ esse arquivo para nosso cluster Thor, e em seguida validar os anagramas em comparação com esse arquivo para determinar quais palavras são válidas. A etapa de validação usa um JOIN da lista de anagramas para o arquivo do dicionário. O uso de um índice e de um JOIN indexado seria mais eficiente, mas isso serve apenas como um simples exemplo.

Fazer o download da Lista de Palavras

Vamos fazer o download da lista de palavras em <http://wordlist.sourceforge.net/> Look for a link to the **2of12.txt** file on that page.

1. Faça o download o pacote *Official 12 Dicts* . Os arquivos estão disponíveis no formato tar.gz ou ZIP.
2. Extraia o arquivo **2of12.txt** para uma pasta em sua máquina local.

Carregar o arquivo de dicionário para sua Zona de Entrada de Arquivo

Nesta etapa, você copiará os arquivos de dados para um local onde eles possam ser distribuídos aos nós de seu cluster Thor do HPCC. Uma zona de entrada de arquivos é um local de armazenagem anexado ao seu HPCC. Ela possui um utilitário em execução para facilitar o spraying (processo de distribuir dados aos nós) para um cluster.

Para arquivos de dados menores, com tamanho máximo de 2GB, você pode usar o utilitário enviar/baixar arquivo no ECL Watch. Este arquivo de dados possui apenas 400 kb (aproximadamente).

Em seguida, você distribuirá (ou fará o spray) o dataset para todos os nós no cluster Thor do HPCC. O poder do HPCC vem da sua capacidade de atribuir vários processadores para trabalhar nas diferentes partes do arquivo de dados em paralelo. Até mesmo a versão, que possui apenas um nó único, os dados precisam ser distribuídos aos nós do cluster.

1. Em um navegador da Internet, conecte-se ao ECL Watch usando http://<PUBLIC_DNS>:8010 (onde PUBLIC_DNS é o nome do DNS público do seu servidor ESP).



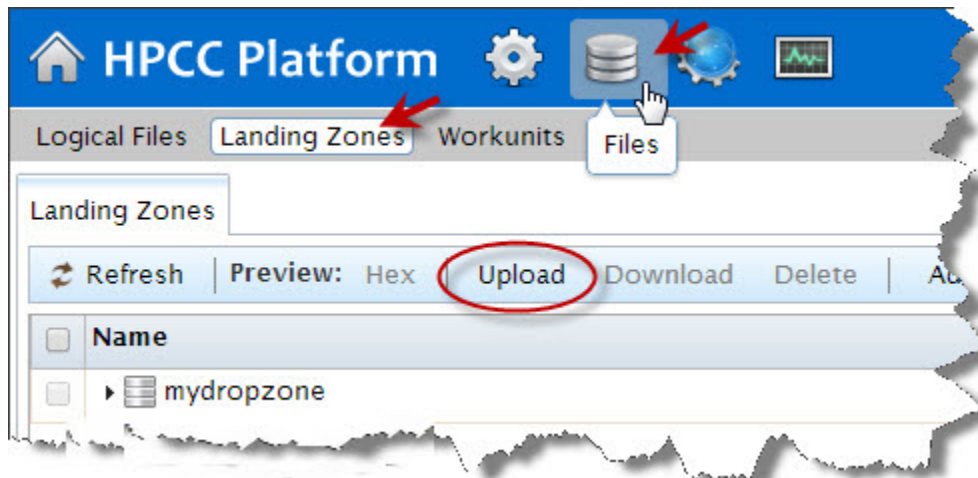
Seu endereço IP poderá ser diferente dos endereços fornecidos nas imagens de exemplo. Use o endereço IP fornecido pela **sua** instalação.

¹Um *spray* ou *importação* é a transferência de um arquivo de dados de um local (como a zona de entrada de arquivos) para um cluster da Refinaria de dados. O termo spray foi adotado devido à natureza da transferência dos arquivos – o arquivo é particionado entre todos os nós em um cluster.

2. No ECL Watch, clique no ícone **Files** e no link **Landing Zones** localizados no submenu de navegação.

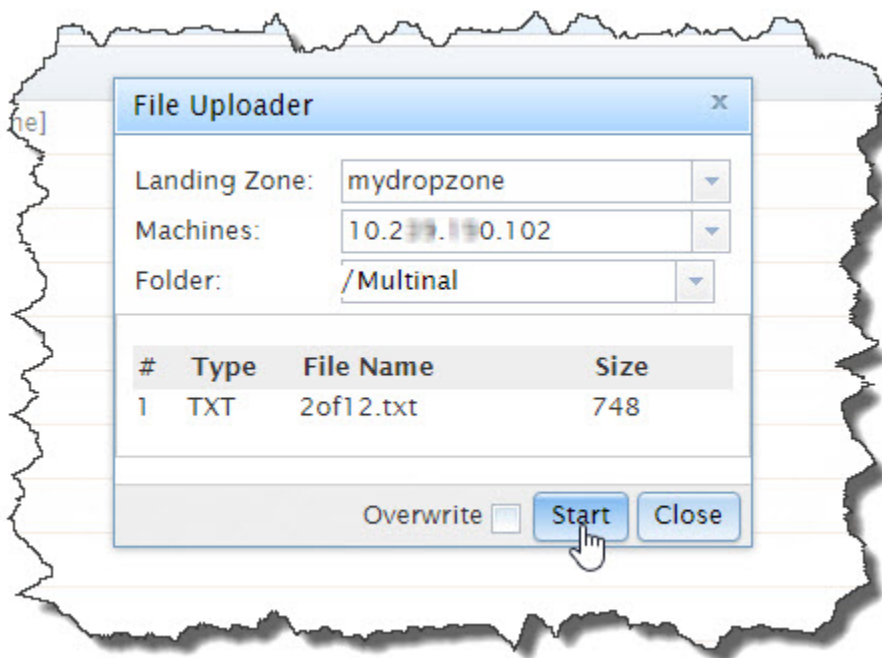
Pressione o botão de ação **Upload** .

Figure 21. Enviar



3. Uma caixa de diálogo será aberta. **Navegue** e selecione o arquivo a ser enviado e pressione o botão **Open** .

Figure 22. O arquivo selecionado deve aparecer no



campo **File Name** . O arquivo de dados possui o seguinte nome: **2of12.txt**..

4. Pressione o botão **Start** para concluir o envio do arquivo.

Spray do arquivo de dados para o seu *Thor Cluster*

Para usar o arquivo de dados em seu Thor do HPCC System, é preciso fazer o "spray" (distribuir) desse arquivo para todos os nós. O *spray* ou *importação* é a transferência de um arquivo de dados de um local (como a zona de entrada de arquivos) para diversas partes do arquivo ou nós em um cluster.

O arquivo distribuído passa a ter um *nome de arquivo lógico* como segue: **~thor::word_list_csv**. O sistema mantém uma lista de arquivos lógicos e as localizações dos arquivos físicos correspondentes das partes do arquivo.

1. Em um navegador da Internet, conecte-se ao ECL Watch usando http://<PUBLIC_DNS>:8010 (onde PUBLIC_DNS é o nome do DNS público do seu servidor ESP).
2. Clique no hiperlink **Arquivos** e no link **Zona de entrada de arquivos** localizados no submenu de navegação. Selecione a zona de entrada de arquivos apropriada (caso haja mais de uma zona de entrada de arquivos). Clique na seta à esquerda da sua zona de entrada de arquivos para expandir.
3. Selecione o arquivo na zona de entrada de arquivos marcando a caixa ao lado dele.

4. Marque a caixa ao lado de 2of12.txt, e pressione o botão **Delimited**.

Figure 23. Spray delimitado

The screenshot shows the HPCC Systems DFU Spray Delimited configuration window. The window has a tabbed interface with 'Delimited' selected. It contains two main sections: 'Target' and 'Options'. The 'Target' section has fields for 'Group' (mythor), 'Queue' (dfusever_queue), 'Target Scope' (~thor), and a 'Target Name' list with 'word_list_csv' selected. The 'Options' section includes 'Format' (ASCII), 'Max Record Length' (8192), 'Separators' (\.), 'Omit Separator' (unchecked), 'Escape' (empty), 'Line Terminators' (\n,\r\n), 'Quote' (empty), 'Overwrite' (checked), 'Replicate' (checked), 'Compress' (unchecked), 'Record Structure Present' (unchecked), 'Fail If No Source File' (unchecked), 'Quoted Terminator' (unchecked), and 'Expire in (days)' (empty). A 'Spray' button is at the bottom right.

A página DFU Spray Delimited será exibida.

5. Selecione "mythor" na lista suspensa do Grupo Target.
6. Preencha o Target Scope como *thor*.

7. Preencha os demais parâmetros (caso ainda não tenham sido preenchidos).

- Máximo tamanho do registro 8192
- Separador \,
- Terminador de linhas \n,\r\n
- Aspas: '

8. Preencha o Target Name usando o restante do nome do arquivo lógico desejado: word_list_csv

9. Não se esqueça de marcar a **caixa** Overwrite.

Se disponível, certifique-se de que a caixa **Replicate** esteja marcada. (A opção replicar está disponível apenas em sistemas em que a replicação tenha sido ativada.)

10. Pressione o botão **Spray**.

A guia exibe a tarefa DFU onde é possível ver o progresso do spray (distribuição aos nós).

Executar o ECL no Thor

1. Abra uma nova **Janela do compilador** (CTRL+N) e escreva o seguinte código:

```
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. Selecione **thor** como seu cluster de destino.
3. Pressione o botão de verificação de sintaxe localizado na barra de ferramentas principal (ou pressione F7)
4. Pressione o botão **Submit**.
5. Quando o envio estiver concluído, selecione a guia Workunit e em seguida a guia Results.
6. Examine o resultado.

Manipulação dos Dados (Processamento dos Dados)

Esta seção explica o manuseio de dados em uma configuração AWS. Informações adicionais sobre o processamento de dados em uma plataforma HPCC Systems estão disponíveis no manual *Data Handling (Processamento dos Dados)* manual.

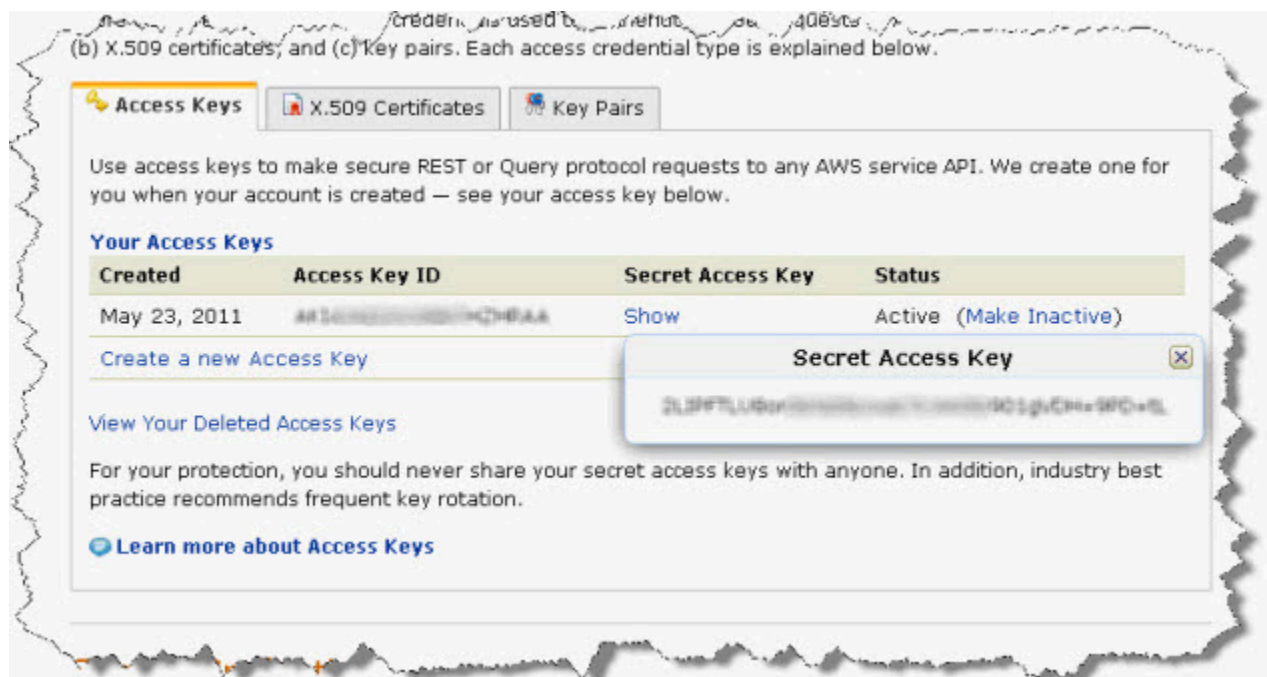
Utilizando o S3 buckets

Os buckets do S3 fornecem um meio de armazenamento constante dentro do Amazon Web Services. Você precisa configurar uma conta no AWS para ativar um conjunto de Chaves de acesso habilitadas e criar buckets do S3. Após ter criado e ativado seu conjunto de chaves de acesso, e criado um bucket do S3 exclusivo, ambos serão usados em futuras instâncias.

Encontre sua Amazon Access Key ID e Secret Access Key

1. Acesse **aws.amazon.com** e faça o login, se necessário.
2. Selecione **Account**.
3. Selecione **Credenciais**.
4. Na página, procure a seção **Access Credentials**.
5. Anote sua **Access Key ID** e sua **Secret Access Key**

Figure 24. Credenciais



Partes desta imagem foram intencionalmente desfocadas

Access Key ID	
Secret Access Key	

Instalar e configurar pacotes do S3 em seu nó da zona de entrada de arquivos

Para mover arquivos do – ou para – o armazenamento do S3, os pacotes do S3 devem estar instalados e configurados no nó da sua zona de entrada de arquivos.

1. Abrir a janela do console e conectar-se ao nó da Zona de entrada de arquivos (LZ)
2. Execute esses comandos:

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

3. Insira sua **Access Key**
4. Insira sua **Secret Access Key**
5. Deixe a senha criptografada em branco
6. Deixe o caminho para o programa GPG em branco
7. Responda à pergunta “Use HTTPS?”
 - Digite “não” para melhorar o desempenho
 - Digite “sim” se a privacidade de dados for uma preocupação para você.
8. Deixe o servidor proxy em branco
9. Insira **Yes** em Test Access
10. Insira **Yes** em Save Settings

Criando e utilizando S3 Buckets

Para armazenar dados no S3, é preciso criar um bucket exclusivo para todo o sistema s3. Uma vez criado, esse bucket existirá mesmo ao fechar as instâncias dos servidores.

Você pode fazer o despray (consolidar dados dos nós) de um arquivo do Thor para a zona de entrada de arquivos, e depois copiar para um bucket do S3 para um armazenamento mais duradouro. Mais tarde, você pode copiar os arquivos do bucket do S3 para a zona de entrada de arquivos e fazer o spray do arquivo para o cluster Thor. Informações adicionais sobre o processamento de dados em uma plataforma HPCC Systems estão disponíveis no manual *Processamento de dados*.

Criar um bucket

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

Listar um Buckets

```
s3cmd ls
```

Enviar um arquivo para o bucket

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

Obter um arquivo do bucket

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

Acesse <http://s3tools.org/s3cmd> para obter mais informações sobre como usar o s3cmd

Próximos passos

Para se familiarizar com o que o seu sistema é capaz de fazer, recomendamos realizar as seguintes etapas:

- **O Tutorial de dados do HPCC**
- **Exemplo** da teoria dos seis graus de separação (de Kevin bacon).
- Ler **Como usar o Gerenciador de Configurações** para aprender como configurar uma plataforma do HPCC usando a Visão avançada.
- Use suas novas habilidades para processar seu próprio dataset massivo!

O Portal do HPCC Systems ([HPCCSystems.com](https://hpccsystems.com)) também é um recurso valioso para obter mais informações, incluindo:

- Vídeos tutoriais
- Exemplos adicionais
- Informe técnico
- Documentação
- Fóruns de usuários