Hortonworks

Using The Hortonworks Virtual Sandbox

Powered By Apache Hadoop

Legal Notice

Contents

About the Hortonworks Virtual Sandbox

The Hortonworks Virtual Sandbox is a hosted environment that runs in an Amazon Web Services (AWS) EC2 environment. It provides a packaged environment for demonstration and trial use of the Apache Hadoop ecosystem on a single node (pseudo-distributed mode). For more details on the suite of components included in the Hortonworks Virtual Sandbox (including the Hortonworks Data Platform), see: About Hortonworks Data Platform.

The Virtual Sandbox is accessible as an Amazon Machine Image (AMI) and requires that you have an account with AWS.

An Amazon Machine Image (AMI) is a special type of pre-configured operating system and virtual application software which is used to create a virtual machine within the Amazon Elastic Compute Cloud (EC2).

Prerequisites

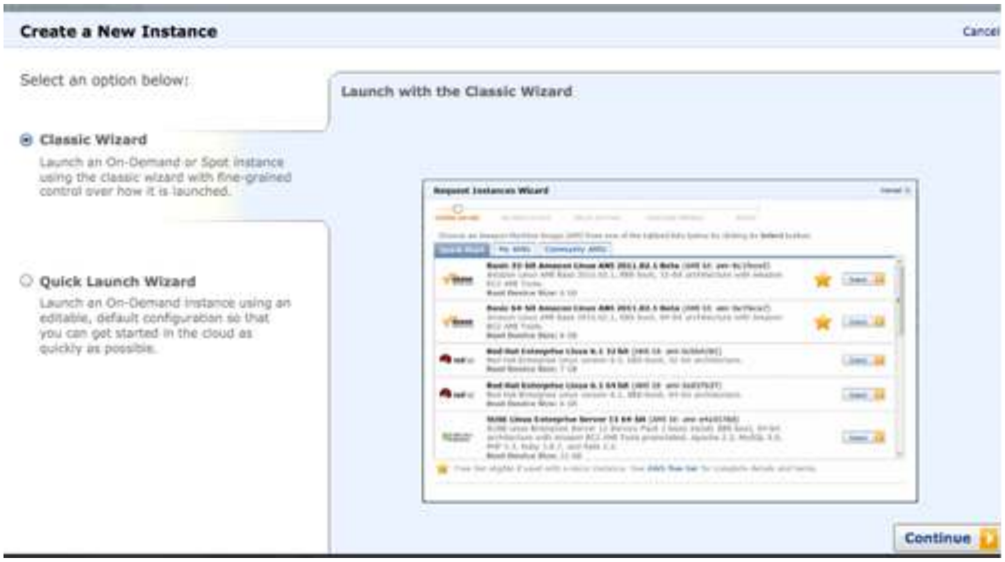
- Ensure that you have the Amazon Web Services (AWS) account ID (see: EC2 Getting Started Guide).
- To connect to the AWS instance from a Windows client machine, ensure that you install PuTTY on your local client machine. See: Installing PuTTY.
- To be able to execute import jobs for Apache Sqoop and/or execute Java Map/Reduce programs, you must install Java Development Kit (JDK) on the AWS instance.

Launching Hortonworks Virtual Sandbox

Step 1: Configure AWS instances.

Step 1-a: Log into the AWS Management Console and select the EC2 tab .

- On the left hand side, click on the 'EC2 Dashboard'.
- Under 'Getting Started' section, click the 'Launch Instance' button. You should now see the 'Create New Instance' wizard as shown below:



Create a New Instance

Cancel

Select an option below:

● Classic Wizard
Launch an On-Demand or Spot instance using the classic wizard with fine-grained control over how it is launched.

○ Quick Launch Wizard
Launch an On-Demand instance using an editable, default configuration so that you can get started in the cloud as quickly as possible.

Launch with the Classic Wizard

Request Instances Wizard

Basic 32-bit Amazon Linux AMI 2011.02.1 Beta

Basic 64-bit Amazon Linux AMI 2011.02.1 Beta

Red Hat Enterprise Linux 6.1 32 bit

Red Hat Enterprise Linux 6.1 64 bit

SUSE Linux Enterprise Server 11 64-bit

- Keep the default selection (‘Classic Wizard’) and click ‘Continue’.

Step 1-b: Select the Hortonworks Virtual Sandbox AMI.

- Select the ‘Community AMIs’ tab Community AMIs .
- In the “Viewing:” pull-down menu, make sure it is set to either “All Images” or “Public Images”.
- Using the Search input box, enter “hortonworks” and press the return key.
- On the AWS Management Console, paste the AMI ID in the search box and click enter.
- Select the **Hortonworks-HDP-1.1.0.15-Single-Node** AMI that is found. Note: the AMI returned is specific for your EC2 Region. Please refer to the FAQ below – ‘Where can I find AMIs for a specific AWS region?’ for information on using an AMI from a different Region.

Step 1-c: Select the instance type.

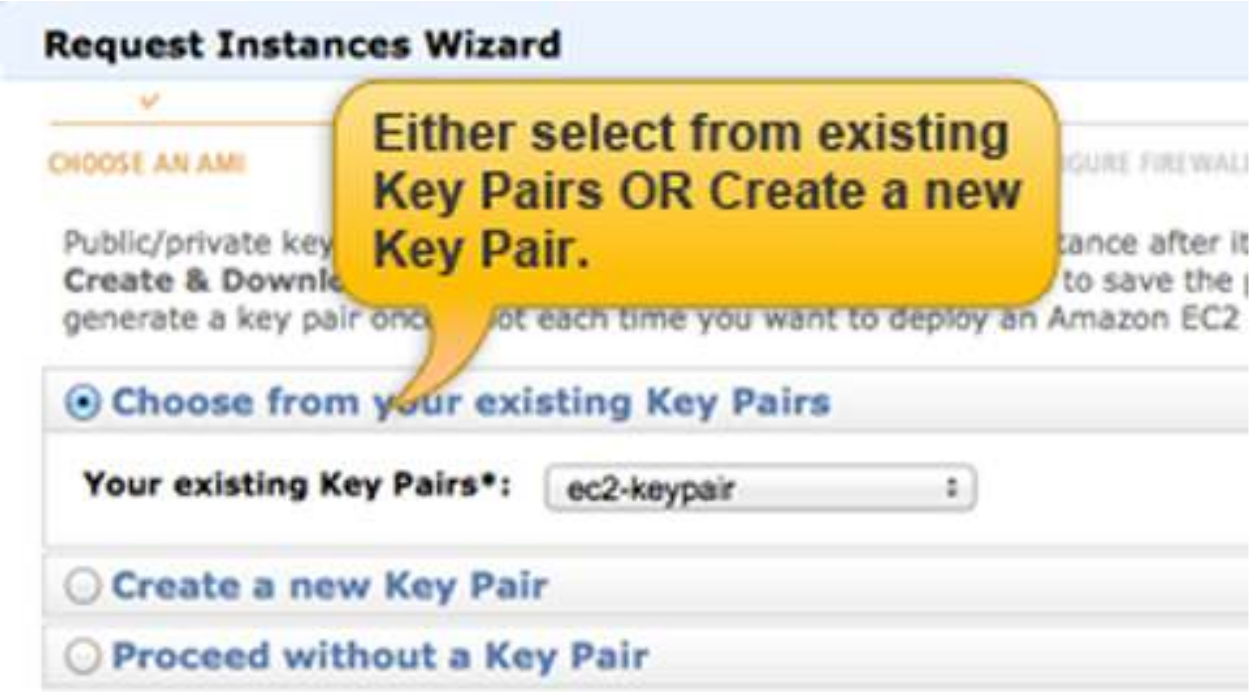
- Keep the default value for the ‘Number of Instances’.
- From the ‘Instance Type’ drop-down, select **Large (m1.large)** and click ‘Continue’.
- Keep the default ‘Advanced Instance Options’ and click ‘Continue’.
- Enter a name for the instance and click ‘Continue’.

Step 1-d: Select the EC2 Key Pair.

You can use one of the following options:

Option I: Select from existing Key Pair.

- If you already have existing EC2 Key Pairs, select the Key Pair you would like to assign to the instances. The following screen shot illustrates the Key Pair selection step:



Request Instances Wizard

Either select from existing Key Pairs OR Create a new Key Pair.

● Choose from your existing Key Pairs

Your existing Key Pairs*: ec2-keypair

○ Create a new Key Pair

○ Proceed without a Key Pair

OR

Option II: Alternatively, use the following instructions:

- Select ‘Create a new Key Pair’ and provide a name for your Key Pair.
- Click ‘Create & Download your Key Pair’.
  This will download the Key Pair file (for example: ec2-keypair) on your local client machine. Ensure that you remember this location.
- Click ‘Continue’.

Step 2: Configure AWS security groups.

You can use one of the following options:

Option I: Select existing AWS security group.

- If you have an existing AWS security group, select ‘Choose one or more of your existing security groups.’
- Ensure that your security group uses all the ports as specified in Mandatory Ports For Hortonworks Virtual Sandbox.

OR

Option II: Alternatively, use the following instructions:

- Select ‘Create a new Security Group’.
- Provide arbitrary names for the ‘Group Name’ and ‘Group Description’.
- Under the ‘Inbound Rule’ section, select **Custom TCP Rule** from the ‘Create a New Rule’ drop down.
- Provide the values for ‘Port Range’ and click ‘Add Rule’ button for each of the row provided in the following table:

TCP Port (Service)	Source
22 (SSH)	0.0.0.0/0
80 (HTTP)	0.0.0.0/0
8020	0.0.0.0/0
8080 (HTTP)	0.0.0.0/0
443 (HTTPS)	0.0.0.0/0
2181	0.0.0.0/0
8005	0.0.0.0/0
8010	0.0.0.0/0

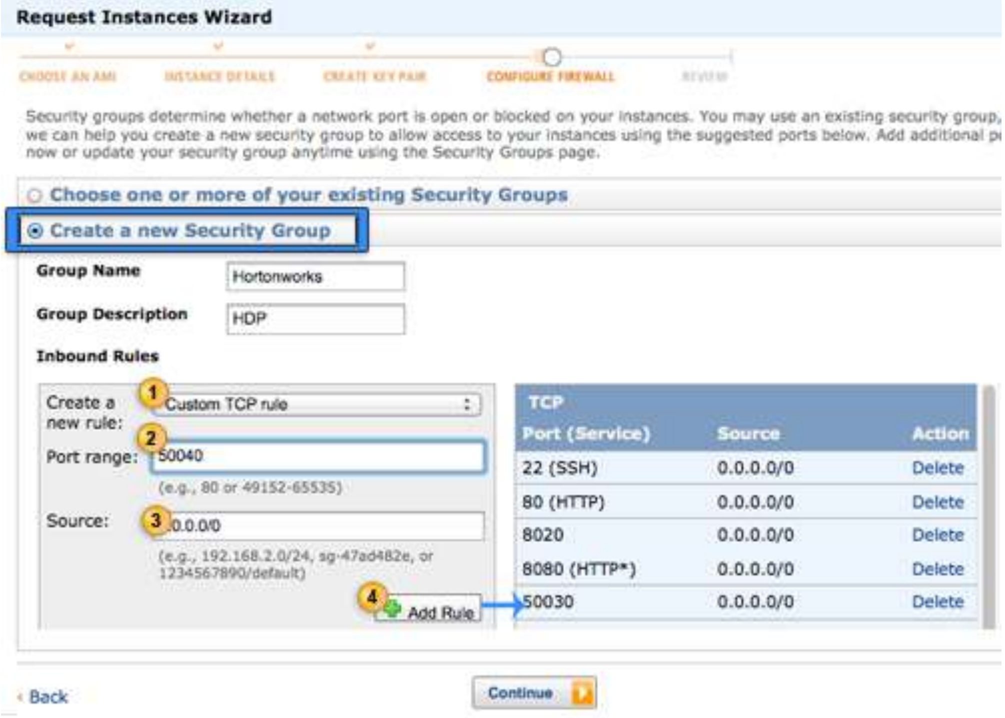
8649	0.0.0.0/0
8651	0.0.0.0/0
8660	0.0.0.0/0
8661	0.0.0.0/0
8662	0.0.0.0/0
8663	0.0.0.0/0
9933	0.0.0.0/0
11000	0.0.0.0/0
41652	0.0.0.0/0
50030	0.0.0.0/0
50040	0.0.0.0/0
50050	0.0.0.0/0
50060	0.0.0.0/0
50070	0.0.0.0/0
50090	0.0.0.0/0
50111	0.0.0.0/0
50300	0.0.0.0/0
51111	0.0.0.0/0
52686	0.0.0.0/0
60010	0.0.0.0/0

60020	0.0.0.0/0
60030	0.0.0.0/0

Table 1: Mandatory Ports For Hortonworks Virtual Sandbox

IMPORTANT: This step can make the AWS instance vulnerable and therefore it is strongly recommended that you should not load sensitive data in your AWS instance.

- You should now see the following result on your AWS Management Console:



Request Instances Wizard

Security groups determine whether a network port is open or blocked on your instances. You may use an existing security group, we can help you create a new security group to allow access to your instances using the suggested ports below. Add additional p now or update your security group anytime using the Security Groups page.

○ Choose one or more of your existing Security Groups

● Create a new Security Group

Group Name: Hortonworks

Group Description: HDP

Inbound Rules

1 Create a new rule: Custom TCP rule

2 Port range: 50040 (e.g., 80 or 49152-65535)

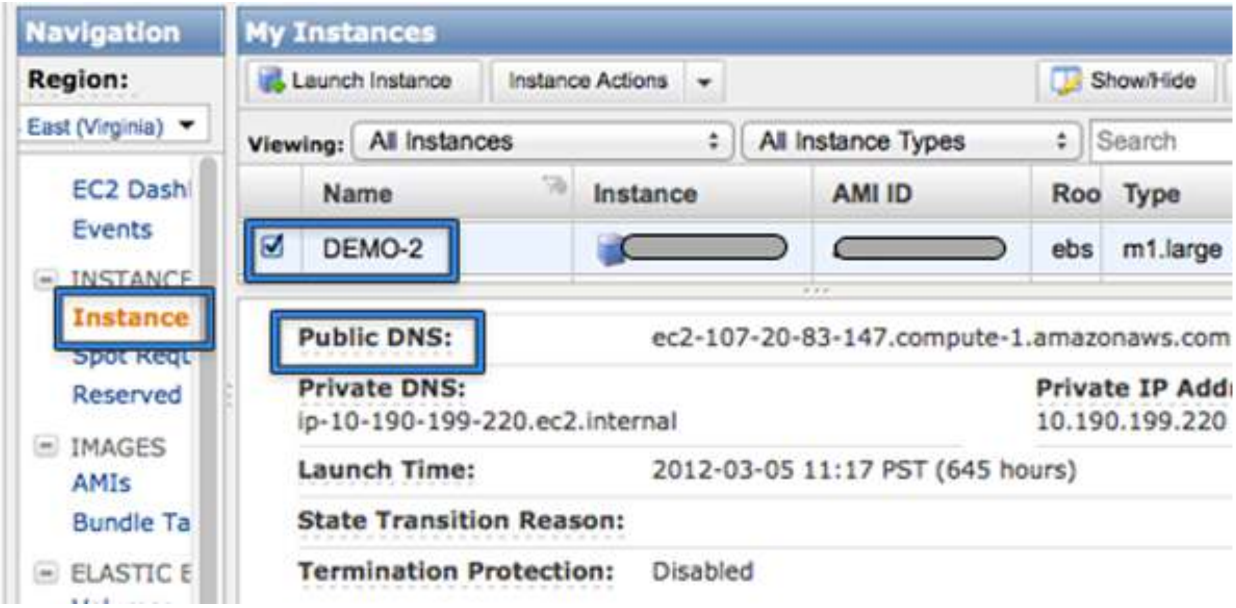
3 Source: 0.0.0.0/0 (e.g., 192.168.2.0/24, sg-47ad482e, or 1234567890/default)

4 Add Rule

TCP		
Port (Service)	Source	Action
22 (SSH)	0.0.0.0/0	Delete
80 (HTTP)	0.0.0.0/0	Delete
8020	0.0.0.0/0	Delete
8080 (HTTP*)	0.0.0.0/0	Delete
50030	0.0.0.0/0	Delete

Step 3: Review and launch.

- Review all of your settings. Ensure that your instance type is a Large (m1.Large) instance.
- Click 'Launch'.
  This should take two to three minutes to launch your instance.
- On your AWS management console, click on the 'Instances' link (on the left hand side navigation menu). This will open the 'My Instances' page on the right hand side.
- On the 'My Instances' page, scroll to your newly launched AWS instance.
  (This instance will have the name as provided in **Step 1-c**).
- Select this instance and copy the public DNS name (as shown in the screen shot below).
  (The public DNS name is also required for Step 4 and Step 6.)



	Name	Instance	AMI ID	Roo	Type
☑	DEMO-2	[redacted]	[redacted]	ebs	m1.large

Public DNS: ec2-107-20-83-147.compute-1.amazonaws.com

Private DNS: ip-10-190-199-220.ec2.internal

Private IP Addr 10.190.199.220

Launch Time: 2012-03-05 11:17 PST (645 hours)

State Transition Reason:

Termination Protection: Disabled

Step 4: Connect to your AWS instance using SSH.

Step 4-a: On the AWS Management Console, browse to the My Instances page. Right click on the row containing your instance and click ‘Connect’.

Step 4-b: Connect to the AWS instance using SSH. You will need the Key Pair file downloaded in step 1-d and the Public DNS Name for your AWS instance (obtained in Step-3).

- **For UNIX:**

```
cd <Full_Path_To_Private_Key_Pair_File_On_Your_Client_Machine>
chmod 400 <Name_Of__Key_Pair_File>

ssh -i <Full_Path_To_Key_Pair_File> root@<Public_DNS
Name_For_AWS_Instance>
```

- **For Windows**, see: Connecting from a Windows Machine using PuTTY.

INFO: You can also copy this command from the AWS Management Console.

Step 5: Install Java JDK.

The Java JDK is ***required*** if you want to compile Java programs for use with Map/Reduce or import Sqoop jobs. By default, the Hortonworks Virtual Sandbox does not include the Java JDK.

To install JDK, use the instructions listed below:

- On your local client machine, point your browser to download the Oracle Java JDK version 6, update 31. Download 64-bit JDK installer binary file (with *.bin extension).
- From your local client machine, perform a secure copy of the downloaded JDK binary files to the AWS instance

```
cd <Path_To_Private_Key_Pair_File_On_Client_Machine>
scp -i <Private_Key_Pair_File_Name>
<Path_To_Java_JDK_Binary_File_On_Client_Machine>
root@<AWS_Public_DNS_Name>:.
```

RESULT: This command will copy the JDK installer binary to the root directory of your AWS instance.

- Connect to your AWS instance (see Step-4).
- Under the usr directory, create the `jdk64` directories:

```
cd /usr
mkdir jdk64
```

- Move the JDK installer binary files to the appropriate directories created above:

```
mv root/jdk-6u31-linux-x64.bin /usr/jdk64
```

- Ensure that the JDK installer binary has correct permissions.

```
chmod 700 jdk-6u31-linux-x64.bin
```

- Install JDK.

```
sh /usr/jdk64/jdk-6u31-linux-x64.bin
```

- Update `JAVA_HOME` and `PATH` variables in the `.bash_profile` file:

```
vi ~/.bash_profile

JAVA_HOME="/usr/jdk64/jdk1.6.0_31"

PATH=${JAVA_HOME}/bin:$PATH:$HOME/bin
```

- Update the value for the `JAVA_HOME` variable defined in the `hadoop-env.sh` file:

```
vi /etc/hadoop/hadoop-env.sh

JAVA_HOME="/usr/jdk64/jdk1.6.0_31"
```

Step 6: Verify if the HDP services are started successfully.

- Use an editor of your choice to view the `hdp-stack-start-<date>-<time>.log` file (located here: /`root`).
  This file provides a list of all the Hadoop services that have started successfully.
  For example, the following snapshot provides the output of the tail end of this log file:

```
****************          Java Process           ***************
2010 hdfs -Dproc_namenode
2313 hdfs -Dproc_secondarynamenode
2496 hdfs -Dproc_datanode
5765 mapred -Dproc_jobtracker
6026 mapred -Dproc_historyserver
6199 mapred -Dproc_tasktracker
6441 hcat -Dproc_jar
6638 oozie -Djava.util.logging.config.file=/var/lib/oozie/
oozieserver/conf/logging.properties
6713 502 -Dzookeeper.log.dir=/hdp/disk0/data/HDP/zk_log_dir
6827 hbase -XX:OnOutOfMemoryError=kill
7001 hbase -XX:OnOutOfMemoryError=kill
7227 2001 -Dproc_jar
```

- You can also use the `hdp-stack-syscheck-<date>-<time>.log` to verify if all the smoke tests passed successfully.
  The following provides a snapshot of a sample syscheck log file after the smoke tests are successfully completed (the smoke test completion typically takes at least 10-15 minutes):

```
Smoke Test Result
================================================================
Hadoop Smoke Test    : Pass
Pig Smoke Test       : Pass
Zookeeper Smoke Test : Pass
Hbase Smoke Test     : Pass
Hcat Smoke Test      : Pass
Templeton Smoke Test : Pass
Hive Smoke Test      : Pass
Sqoop Smoke Test     : Pass
Oozie Smoke Test     : Pass
=================================================================
```

Step 7: Enable access to the HDP Monitoring Dashboard from your local client machine.

Step 7-a: On your local machine, open command line utility and execute the following command. (The public DNS name for the AWS instance can be obtained from Step -3).

```
ping -c1 <Public_DNS_Name_For_AWS_Instance>
```

RESULT: This command will provide you the IP address for your AWS instance.

Step 7-b: On your local client machine, open a command line utility and execute the following commands. Replace the `<ec2-public-IP-Address>` with the IP address obtained in Step 7-a.

- **For Unix (use root privileges):**

```
sudo vi /etc/hosts
<ec2-public-IP-Address> hortonworks-sandbox.localdomain
hortonworks-sandbox
```

- **For Windows (use Administrator privileges):**

```
notepad \WINDOWS\system32\drivers\etc\hosts
<ec2-public-IP-Address> hortonworks-sandbox.localdomain
hortonworks-sandbox
```

INFO: Use the `hosts` file to enable access to the AWS instance for use in your local machine. This step ensures that you can access the HDP Monitoring Dashboard from your local client machine. For more details, see: Using the Hosts file.

Step 7-c: Use the following URL to access the HDP Monitoring Dashboard:
http://hortonworks-sandbox.localdomain/hdp/dashboard/ui/home.html

Step 7-d: To use Nagios UI, click on the Nagios tab
Ensure that you use the following credentials: `nagiosadmin/admin`.

NOTE: The Hortonworks Virtual Sandbox provides a total disk space of 840 GB for your data operations. The data on the AWS instance store will survive only if your instance is rebooted. This data will ***not*** be persisted under the following circumstances:

- Failure of an underlying drive.
- Running an instance on degraded hardware.
- Stopping an Amazon EBS-backed instance.

- Terminating an instance.

For more details, see: Amazon EC2 Instance Storage.

Using Hortonworks Virtual Sandbox

The Hortonworks tutorials are located here: `/root/tutorial.` You can also go to the Virtual Sandbox page and execute these tutorials.

We are working hard to add more tutorials, so check back often.

You can also take the Hortonworks training, check the latest class schedule here.

Frequently Asked Questions (FAQs)

Q. I am unable to connect to the HDP Monitoring Dashboard.

This problem arises if the firewall settings for your system are enabled. Follow the instructions listed below to disable firewall settings for your AWS instance:

Step 1: Verify if the existing firewall settings for the AWS instance are disabled:

```
/etc/init.d/iptables status
```

- If this command does not display the following message, execute the next step to disable the existing firewall settings (stop iptables).

Step 1: Execute the following command to disable existing firewall settings.

```
/etc/init.d/iptables stop
```

Q. I am unable to see the metrics on the Ganglia UI.

This happens when your Ganglia server fails to start. Use the following steps to restart the Ganglia server:

- Connect to the AWS instance (using command line utility or PuTTY).
- Execute the following commands as a root user (`su -`):

```
/etc/init.d/hdp-gmetad restart
/etc/init.d/hdp-gmond restart
```

Q. How to execute the smoke tests to verify if my HDP components are working correctly?

Smoke tests are executed after the AWS instance is launched. These tests can be executed again using the following command:

```
/etc/init.d/hdp-stack syscheck
```

Q. How do I start or stop individual Hadoop services?

To start or stop all the Hadoop services, you can use the following auxiliary scripts provided with the Hortonworks Virtual Sandbox:

- To stop HDP services: `/etc/init.d/hdp-stack stop`
- To start HDP services: `/etc/init.d/hdp-stack start`

However, to start or stop the individual Hadoop services, you must use the instructions provided here: Controlling Hadoop Services Manually (Starting/Stopping).

Q. Where can I find AMIs for a specific AWS region?

To choose an AMI based on a specific AWS region, you can search for the following AMI IDs:

REGION	REGION NAME	END POINT URL	AMI ID
us-east-1	(US West (N. California))	ec2.us-east-1.amazonaws.com	ami-754bfa1c
us-west-2	(US West (Oregon))	ec2.us-west-2.amazonaws.com	ami-f6bc32c6
us-west-1	(US West (N. California))	ec2.us-west-1.amazonaws.com	ami-63bf9b26
eu-west-1	(EU West (Ireland))	ec2.eu-west-1.amazonaws.com	ami-6bafa91f
ap-southeast-1	(Asia Pacific (Singapore))	ec2.ap-southeast-1.amazonaws.com	ami-bc2868ee
ap-northeast-1	(Asia Pacific (Tokyo))	ec2.ap-northeast-1.amazonaws.com	ami-be922ebf
sa-east-1	(South America (Sao Paulo))	ec2.sa-east-1.amazonaws.com	ami-aed009b3

Known Issues

HBase Master user interface (UI) experiences issues while rendering on some platforms.
For more information, see: https://issues.apache.org/jira/browse/HBASE-5286