Creating a Cluster on AWS

Date of Publish: 2019-02-06

https://docs.hortonworks.com/

Contents

Create a cluster on AWS

Use these steps to create a cluster with Cloudbreak.

If you experience problems during cluster creation, refer to Troubleshooting cluster creation.

Prerequisites

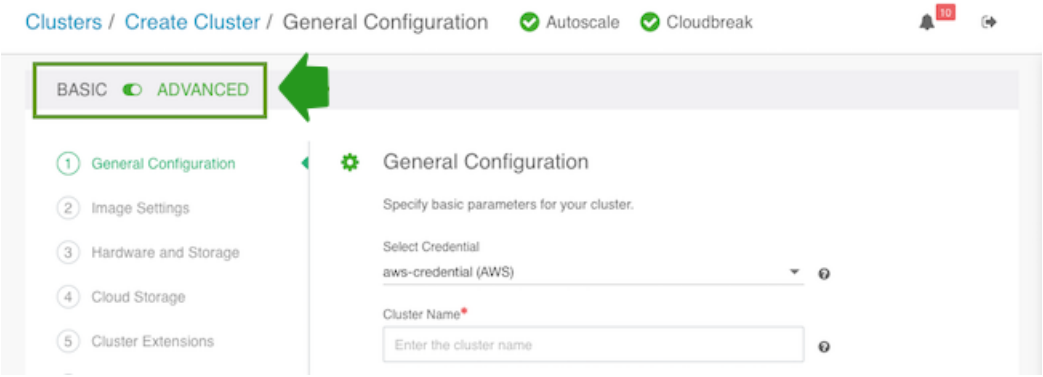
This prerequisite step is required for AWS GovCloud users only and is optional for other AWS users.

If you are planning to create clusters on AWS GovCloud, you must first prepare a custom image and register a custom image catalog with Cloudbreak. For instructions, refer to Custom images.

Steps

1. Log in to the Cloudbreak web UI.
2. Click the Create Cluster button and the Create Cluster wizard is displayed.

   By default, the Basic view is displayed. To view advanced options, click Advanced. To learn about advanced options, refer to Advanced cluster options.



3. On the General Configuration page, specify the following general parameters for your cluster:

Parameter	Description
Select Credential	Choose a previously created credential.
Cluster Name	Enter a name for your cluster. The name must be between 5 and 40 characters, must start with a letter, and must only include lowercase letters, numbers, and hyphens.
Region	Select the AWS region in which you would like to launch your cluster. For information on available AWS regions, refer to AWS regions and endpoints in AWS documentation.
Platform Version	Choose the HDP or HDF version to use for this cluster. Blueprints available for this platform version will be populated under “Cluster Type” below. If you selected the HDF platform, refer to Creating HDF clusters for HDF cluster configuration tips.
Cluster Type	Choose one of the default cluster configurations, or, if you have defined your own cluster configuration via Ambari blueprint, you can choose it here. For more information on default and custom blueprints, refer to Custom blueprints.
Flex Subscription	This option will appear if you have configured your deployment for a flex subscription.

4. On the Hardware and Storage page, for each host group provide the following information to define your cluster nodes and attached storage. To edit this section, click on the

When done editing, click on the

to save the changes.

Parameter	Description
Ambari Server	You must select one node for Ambari Server by clicking the  button. The “Instance Count” for that host group must be set to “1”. If you are using one of the default blueprints, this is set by default.
Instance Type	Select an instance type. For information about instance types on AWS refer to Amazon EC2 instance types in AWS documentation.
Instance Count	Enter the number of instances of a given type. Default is 1.
Storage Type	Select the volume type. The options are: (1) Magnetic (default), (2) General Purpose (SSD, (3) Throughput Optimized HDD. For more information about these options refer to Amazon EC2 Instance Store in AWS documentation.
Encryption	Under Encryption Key, you can select an existing encryption key. For more information, refer to EBS encryption.
Attached Volumes Per Instance	Enter the number of volumes attached per instance. Default is 1.
Volume Size	Enter the size in GB for each volume. Default is 100.
Root Volume Size	This option allows you to increase or decrease the root volume size. Default is 50 GB. This option is useful if your custom image requires more space than the default 50 GB.
Use Spot Instances	Check this option to use EC2 spot instances as your cluster nodes. Next, enter your bid price. The price that is pre-loaded in the form is the current on-demand price for your chosen EC2 instance type. For more information, refer to Use spot instances.

5. On the Network and Availability page, provide the following to specify the networking resources that will be used for your cluster:

Parameter	Description
Select Network	Select the virtual network in which you would like your cluster to be provisioned. You can select an existing network or create a new network.  **Note:** The Shared Networks option is only available for Google Cloud.
Select Subnet	Select the subnet in which you would like your cluster to be provisioned. If you are using a new network, create a new subnet. If you are using an existing network, select an existing subnet.
Subnet (CIDR)	If you selected to create a new subnet, you must define a valid CIDR for the subnet. Default is 10.0.0.0/16.

Note:

Cloudbreak uses public IP addresses when communicating with cluster nodes.

On AWS, you can configure it to use private IPs instead. For instructions, refer to Configure communication via private IPs on AWS in the Troubleshooting documentation.

6. On the Gateway Configuration page, you can access gateway configuration options.

When creating a cluster, Cloudbreak installs and configures a gateway (powered by Apache Knox) to protect access to the cluster resources. By default, the gateway is enabled for Ambari; You can optionally enable it for other cluster services.

For more information, refer to Gateway configuration documentation.

7. On the Network Security Groups page, define security groups for each host group. You can either create new security groups and define their rules or reuse existing security groups:

   Note:

   Existing security groups are only available when an existing VPC is selected.

Option	Description
New Security Group	(Default) Creates a new security group with the rules that you defined: • A set of default rules is provided. You should review and adjust these default rules. If you do not make any modifications, default rules will be applied. • You may open ports by defining the CIDR, entering port range, selecting protocol and clicking +. • You may delete default or previously added rules using the delete icon. • If you don't want to use security group, remove the default rules.
Existing Security Groups	Allows you to select one or more existing security groups that exist in the selected region and network. To use an existing security group, select it from the dropdown and then click "Attach". Repeat these steps if you would like to use multiple security groups. This selection is disabled if no existing security groups are available in your chosen region and network.

   The default experience of creating network resources such as network, subnet and security group automatically is provided for convenience. We strongly recommend you review these options and for production cluster deployments leverage your existing network resources that you have defined and validated to meet your enterprise requirements. For more information, refer to Restrict inbound access to clusters.

8. On the Security page, provide the following parameters:

Parameter	Description
Cluster User	You can log in to the Ambari web UI using this username. By default, this is set to admin.
Password	You can log in to the Ambari web UI using this password.
Confirm Password	Confirm the password.
New SSH public key	Check this option to specify a new public key and then enter the public key. You will use the matching private key to access your cluster nodes via SSH.
Existing SSH public key	Select an existing public key. You will use the matching private key to access your cluster nodes via SSH. This is a default option as long as an existing SSH public key is available.

9. Click on Create Cluster to create a cluster.
10. You will be redirected to the Cloudbreak dashboard, and a new tile representing your cluster will appear at the top of the page.

Related Information

Troubleshooting cluster creation
Advanced cluster options
Guidelines for creating HDF clusters
Custom blueprints
Flex subscription
Gateway configuration
CIDR IP calculator
Default cluster security groups
Restrict inbound access to clusters

Custom images
AWS regions and endpoints (AWS)
Amazon EC2 instance types (AWS)
Amazon EC2 instance store (AWS)
EBS encryption
Use spot instances
Configure communication via private IPs on AWS

Default cluster configurations

Cloudbreak includes default cluster configurations (in the form of blueprints) and supports using your own custom cluster configurations (in the form of custom blueprints).

The following default cluster configurations are available:

HDP 3.1

Cluster type	Main services	Description	List of all services
Data Science	Spark 2, Zeppelin	Useful for data science with Spark 2 and Zeppelin.	HDFS, YARN, MapReduce2, Tez, Hive, Pig, Sqoop, ZooKeeper, Ambari Metrics, Spark 2, Zeppelin
EDW - Analytics	Hive 2 LLAP, Zeppelin	Useful for EDW analytics using Hive LLAP.	HDFS, YARN, MapReduce2, Tez, Hive 2 LLAP, Druid, Pig, ZooKeeper, Ambari Metrics, Spark 2

HDP 2.6

Cluster type	Main services	Description	List of all services
Data Science	Spark 2, Zeppelin	Useful for data science with Spark 2 and Zeppelin.	HDFS, YARN, MapReduce2, Tez, Hive, Pig, Sqoop, ZooKeeper, Ambari Metrics, Spark 2, Zeppelin
EDW - Analytics	Hive 2 LLAP, Zeppelin	Useful for EDW analytics using Hive LLAP.	HDFS, YARN, MapReduce2, Tez, Hive 2 LLAP, Druid, Pig, ZooKeeper, Ambari Metrics, Spark 2
EDW - ETL	Hive, Spark 2	Useful for ETL data processing with Hive and Spark 2.	HDFS, YARN, MapReduce2, Tez, Hive, Pig, ZooKeeper, Ambari Metrics, Spark 2

HDF 3.3

Cluster type	Main services	Description	List of all services
Flow Management	NiFi	Useful for flow management with NiFi.	NiFi, NiFi Registry, ZooKeeper, Ambari Metrics
Messaging Management	Kafka	Useful for messaging management with Kafka.	Kafka, ZooKeeper, Ambari Metrics

Cluster security groups

This section lists ports used by Cloudbreak-manged clusters.

The following tables lists the default and recommended cluster security group settings:

Note:

By default, when creating a cluster, a new network, subnet, and security groups are created automatically. The default experience of creating network resources such as network, subnet and security group automatically is provided for convenience. We strongly recommend that you review these options and for production cluster deployments leverage your existing network resources that you have defined and validated to meet your enterprise requirements.

Note:

Depending on the cluster components that you are planning to use, you may need to open additional ports required by these components.

External ports

Source	Target	Protocol	Port	Description
Cloudbreak	Ambari server	TCP	9443	• This port is used by Cloudbreak to maintain management control of the cluster. • The default security group opens 9443 from anywhere. You should limit this CIDR further to only allow access from the Cloudbreak host. This can be done by default by restricting inbound access from Cloudbreak to cluster.
*	All cluster hosts	TCP	22	• This is an optional port for end user SSH access to the hosts. • You should review and limit or remove this CIDR access.
*	Ambari server	TCP	8443	• This port is used to access the gateway (if configured). • You should review and limit this CIDR access. • If you do not configure the gateway, this port does not need to be opened. If you want access to any cluster resources, you must open this port explicitly on the security groups for their respective hosts.

Source	Target	Protocol	Port	Description
*	Ambari server	TCP	443	• This port is used to access Ambari directly. • If you are configuring the gateway, you should access Ambari through the gateway; In this case you do not need to open this port. • If you do not configure the gateway, to obtain access to Ambari, you can open this port on the security group for the respective host.

Internal ports

In addition to the ports described above, Cloudbreak uses certain ports for internal communication within the subnet. By default, Cloudbreak opens ports 0-65535 to the subnet's internal CIDR (such as 10.0.0.0/16). Use the following table to limit this CIDR:

Source	Target	Protocol	Port	Description
Salt-bootstrap	Gateway instance (Ambari server instance)	TCP	7070	Salt-bootstrap service launches and configures Saltstack.
Salt-master	All hosts in the cluster	TCP	4505, 4506	Salt-minions connect to the Salt-master(s).
Consul server	All hosts in the cluster	TCP, UDP	8300, 8301	Consul agents connect to the Consul server.
Consul agent (all hosts in the cluster)	All hosts in the cluster	TCP, UDP	8300, 8301	Consul agents connect to other Consul agents (Gossip protocol).
Prometheus node exporter	Gateway instance (Ambari server instance)	TCP	9100	Prometheus server scrapes metrics from the node exporters.
Ambari server	All hosts in the cluster	Refer to Default network port numbers for Ambari in Ambari documentation.		Ambari agents connect to the Ambari server.

When creating data lakes and their attached clusters, you must also open the following internal port:

Source	Target	Protocol	Port	Description
Data lake cluster	Clusters attached to the data lake	TCP	6080	Used for communication between the data lake and attached clusters.

Guidelines for creating HDF clusters

In general, the create cluster wizard offers prescriptive default settings that help you configure your HDF clusters properly; however, there are a few additional configuration requirements that you should be aware of.

For guidance, refer to the following documentation:

Creating HDF Flow Management clusters

When creating a Flow Management cluster from the default blueprint, make sure to follow these guidelines.

Network

On the Network Security Groups page, open the required ports:

- Open 9091 TCP port on the NiFi host group. This port is used by NiFi web UI; Without it, you will be unable to access the NiFi web UI.
- Open 61443 TCP port on the Services host group. This port is used by NiFi Registry.

LDAP and Kerberos

You must either use your existing LDAP or enable Kerberos.

- If using LDAP, you must first register it as an external authentication source in the Cloudbreak web UI.
- If using Kerberos, you can either use your own Kerberos (for production) or select for Cloudbreak to create a test KDC (for evaluation only).

Cluster user vs LDAP users

When creating an HDF cluster with LDAP, on the Security page:

- You must specify a Cluster User that is a valid user in the LDAP. This is a limitation with NiFi/NiFi Registry. Only one login provider can be configured for those components, and if an LDAP is supplied, then the login provider is set to LDAP; Consequently, this requires that the initial admin be in the given LDAP.
- While the passwords don't need to match between the cluster user and the same-named LDAP user, any other components that are used by that cluster user would require the password entered for the cluster user, not the same-named LDAP user.

Creating the NiFi Registry controller service in NiFi

When creating the NiFi Registry controller service in NiFi, the internal hostname must be used, e.g. https://ip-1-2-3-4.us-west-2.compute.internal:61443

Creating a custom HDF blueprint with NiFi

If you are creating a custom HDF blueprint which includes Apache NiFi, you must add the "nifi-bootstrap-env" configuration to the "configurations" section of the blueprint, exactly as it is done in this example: https://github.com/hortonworks/cloudbreak/blob/master/core/src/main/resources/defaults/blueprints/hdf32-flow-management.bp#L19-%23L21.

You can modify the value of nifi-bootstrap-env's "content" property, but the java.io.tmpdir arg is required to be populated with a directory that does not have noexec set on it, and the NiFi user must be able to write to that directory.

Scaling and autoscaling

Although Cloudbreak allows cluster scaling (including autoscaling), scaling is not supported by NiFi:

- Downscaling NiFi clusters is not supported - as it can result in data loss when a node is removed that has not yet processed all the data on that node.
- There is a known issue related to upscaling.

Troubleshooting HDF Flow Management cluster creation

NiFi returns the following error when the Flow Management cluster is set up with an LDAP:

Error:

```
Caused by: org.springframework.beans.factory.BeanCreationException:
 Error creating bean with name 'authorizer': FactoryBean
 threw exception on object creation; nested exception is
 org.apache.nifi.authorization.exception.AuthorizerCreationException:
 org.apache.nifi.authorization.exception.AuthorizerCreationException: Unable
 to locate initial admin admin to seed policies
        at
 org.springframework.beans.factory.support.FactoryBeanRegistrySupport.doGetObjectFromFac
        at
 org.springframework.beans.factory.support.FactoryBeanRegistrySupport.getObjectFromFacto
        at
 org.springframework.beans.factory.support.AbstractBeanFactory.getObjectForBeanInstance(
        at
 org.springframework.beans.factory.support.AbstractBeanFactory.doGetBean(AbstractBeanFac
        at
 org.springframework.beans.factory.support.AbstractBeanFactory.getBean(AbstractBeanFacto
        at
 org.springframework.beans.factory.support.BeanDefinitionValueResolver.resolveReference(
        ... 90 common frames omitted
Caused by:
 org.apache.nifi.authorization.exception.AuthorizerCreationException:
 org.apache.nifi.authorization.exception.AuthorizerCreationException: Unable
 to locate initial admin admin to seed policies
        at
 org.apache.nifi.authorization.FileAccessPolicyProvider.onConfigured(FileAccessPolicyPro
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at
 sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at
 sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at
 org.apache.nifi.authorization.AccessPolicyProviderInvocationHandler.invoke(AccessPolicy
        at com.sun.proxy.$Proxy72.onConfigured(Unknown Source)
        at
 org.apache.nifi.authorization.AuthorizerFactoryBean.getObject(AuthorizerFactoryBean.jav
        at
 org.springframework.beans.factory.support.FactoryBeanRegistrySupport.doGetObjectFromFac
        ... 95 common frames omitted
Caused by:
 org.apache.nifi.authorization.exception.AuthorizerCreationException: Unable
 to locate initial admin admin to seed policies
        at
 org.apache.nifi.authorization.FileAccessPolicyProvider.populateInitialAdmin(FileAccessP
        at
 org.apache.nifi.authorization.FileAccessPolicyProvider.load(FileAccessPolicyProvider.ja
        at
 org.apache.nifi.authorization.FileAccessPolicyProvider.onConfigured(FileAccessPolicyPro
        ... 103 common frames omitted
```

Cause:

When creating a HDF cluster with LDAP, on the Security page of the create cluster wizard you specified some Cluster User that is not a valid user in the LDAP.

Solution:

When creating a HDF cluster with LDAP, on the Security page of the create cluster wizard, specify a Cluster User that is a valid user in the LDAP.

Related Information

External authentication source for clusters

Known issues

Creating HDF Messaging Management clusters

When creating a Messaging Management cluster from the default blueprint, make sure to follow these guidelines.

External database for Schema Registry

If Schema Registry is included in your blueprint, you must:

1. Set up an external MySQL instance prior to cluster installation.

   Note:

   Only a MySQL database can be used. Other database types are not supported.
2. Access the instance by using MySQL Workbench or some other tool, and prepare it as described in Configuring SAM and Schema Registry Metadata Stores in MySQL.
3. Register the external database in Cloudbreak web UI under External Sources > Configure Databases prior to cluster installation. Make sure to select "Registry" under Type. For instructions, refer to Register an external database.
4. During cluster installation, navigate to the External Sources page of the advanced cluster wizard and attach the database.

Network

When creating a cluster, open 3000 TCP port on the Services host group for Grafana.

If Schema Registry is included, also open TCP ports 7788 and 7789 on the Services host group to ensure access to Schema Registry from the browser.

Scaling and autoscaling

Adding nodes to the host group including Kafka broker (via resizing or autoscaling) is not supported. Nevertheless, it can be achieved by using the following workaround: After scaling the host group, use Ambari to add Kafka broker to each newly added node.

Related Information

Configure Postgres to Allow Remote Connections (Hortonworks)

Configuring SAM and Schema Registry Metadata Stores in MySQL (Hortonworks)

Register an external database

Advanced cluster options

In the create cluster wizard, click on Advanced in the top-left corner to view the advanced cluster configuration options.

The following options are available:

Availability zone

The "Availability zone" option allows you to select an availability zone within the selected region. If you do not make a selection, an availability zone is selected automatically.

To access this option:

1. In create cluster wizard, click on Advanced.

2. Navigate to the General Configuration page.
3. Under Region, select a region.
4. Under Availability zone, select an availability zone within the selected region.

Enable lifetime management

The "Enable lifetime management" option allows you to have your cluster automatically terminated after a certain amount of time.

To access this option:

1. In create cluster wizard, click on Advanced.
2. Navigate to the General Configuration page.
3. If you would like your cluster to be automatically terminated after a specific amount of time (defined as “Time to Live” in minutes), check the Enable lifetime management option, and then specify Time to Live in minutes.

Note:

Your cluster will be terminated according to the setting specified; You will not be able to reverse this.

Tags

You can define tags that will be applied to your cluster-related resources (such as VMs) on your cloud provider account.

The tags added during cluster creation are displayed in your cloud account on the resources that Cloudbreak provisioned for your clusters. You can use tags to categorize your cloud resources by purpose, owner, and so on. Tags come in especially handy when you are using a corporate AWS account and you want to quickly identify which resources belong to your cluster(s). In fact, your corporate cloud account admin may require you to tag all the resources that you create, in particular resources, such as VMs, which incur charges.

By default, the following tags are created:

Tag	Description
cb-version	Cloudbreak version
Owner	Your Cloudbreak admin email.
cb-account-name	Your automatically generated Cloudbreak account name stored in the identity server.
cb-user-name	Your Cloudbreak admin email.

You can optionally add additional tags. To add custom tags:

1. In create cluster wizard, click on Advanced.
2. Navigate to the General Configuration page.
3. Specify your tags in the Tags section by providing a key and value for each tag.

Note:

It is not possible to add tags via Cloudbreak after your cluster has been created. In this case, you can only add the tags manually via your cloud provider's interface.

To learn more about tags and their restrictions, refer to your cloud provider documentation.

Related Information

Tagging your Amazon EC2 resources

Add tags in Profile

Image settings

The options on the "Image Settings" page of the advanced create cluster wizard allow you to select custom image settings.

By default, Cloudbreak uses the prewarmed image from the image catalog provided with Cloudbreak. The following options allow you to customize image settings:

Choose image catalog

By default, Cloudbreak uses the image catalog provided with Cloudbreak. If you would like to use a custom image catalog instead of the default image catalog, you must first create and register it. For instructions, refer to Custom images.

Choose image type

By default, Cloudbreak uses the included prewarmed images with default Ambari and HDP/HDF version, but you can select a different prewarmed or base image to use for your cluster. Cloudbreak supports the following types of images for launching clusters:

Image type	Description	Default images provided	Support for custom images
Prewarmed Image	By default, Cloudbreak launches clusters from prewarmed images. Prewarmed images include the operating system as well as Ambari and HDP/HDF. The Ambari and HDP/HDF version used by prewarmed images cannot be customized.	Yes	No
Base Image	Base images include default configuration and default tooling. These images include the operating system but do not include Ambari or HDP/HDF software. If you would like to use an Ambari and HDP/HDF versions different than what the prewarmed image includes, you must select to use a base image.	Yes	Yes

If you would like to use an Ambari and HDP/HDF versions different than what the prewarmed image includes, you must select to use a base image (instead of a prewarmed image). For instructions on how to customize Ambari and/or HDP/HDF repos, refer to the documentation linked below

Choose image

This option allows you to select a different prewarmed or base image (default - if available - or from your custom image catalog).

Related Information

Custom images

Ambari repo specification

HDP/HDF repo specification

Ambari repo specification

You can specify a custom version of Ambari on the "Image Settings" page of the advanced create cluster wizard, under "Ambari Repo Specification".

By default, Cloudbreak uses the included prewarmed images with default Ambari and HDP/HDF version. If you would like to use a custom Ambari version:

1. In create cluster wizard, click on Advanced.
2. Navigate to the Image Settings page.
3. Under Choose image type, select to use a base image.

Note:

It is important that you select a base image. If you select a prewarmed image, you cannot customize the repo.

4. Provide the following information under Ambari repo specification:

Parameter	Description	Notes	Example
Version	Enter Ambari version.		2.6.2.2
Repo Url	Provide a URL to the Ambari version repo that you would like to use. You can obtain this URL from Ambari installation documentation.		http://public-repo-1.hortonworks.com/ambari/centos7/2.x/updates/2.6.2.2
Repo Gpg Key Url	Provide a URL to the repo GPG key. Each stable RPM package that is published by CentOS Project is signed with a GPG signature. By default, yum and the graphical update tools will verify these signatures and refuse to install any packages that are not signed, or have an incorrect signature.		http://public-repo-1.hortonworks.com/ambari/centos6/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins

Related Information

HDP/HDF repo specification

HDP/HDF repo specification

You can specify a custom version of HDP/HDF on the "Image Settings" page of the advanced create cluster wizard, under "HDP/HDF Repo Specification".

By default, Cloudbreak uses the included prewarmed images with default Ambari and HDP/HDF version. If you would like to use a custom HDP or HDF version:

Note:

Depending on the version that you would to use, you may need to provide a custom blueprint for that version.

1. In create cluster wizard, click on Advanced.
2. Navigate to the Image Settings page.
3. Under Choose image type, select to use a base image.

Note:

It is important that you select a base image. If you select a prewarmed image, you cannot customize the repo.

4. Provide the following information under HDP/HDF repo specification:

Parameter	Description	Notes	Example
Stack	This is populated by default based on the “Platform Version” parameter.		HDP

Parameter	Description	Notes	Example
Version	This information is populated by default based on the “Platform Version” parameter (which in turn is derived from available blueprints). This means that in order to change the value of this parameter, you must provide an appropriate blueprint.		2.6
OS	Operating system. This should be centos7 (Azure, GCP, OpenStack) or amazonlinux (AWS).		centos7
Repository Version	Enter repository version.		2.6.5.0-292
Version Definition File	Enter the URL of the VDF file.	Only required when using Ambari 2.6 or newer. In this case you do not need to provide Base URL, Utils Repo ID, and Utils Base URL.	http://public-repo-1.hortonworks.com/HDP/centos7/2.x/updates/2.6.5.0/HDP-2.6.5.0-292.xml
Stack Repo Id	This information is populated by default based on the “Version” parameter.	Only displayed for Ambari 2.5 and earlier.	HDP-2.6
Base Url	URL to the repo storing the desired stack version.	Only required for Ambari 2.5 and earlier.	http://public-repo-1.hortonworks.com/HDP/centos7/2.x/updates/2.6.1.0
Utils Repo Id	Identifier for the Utils Base URL repo.	Only required for Ambari 2.5 and earlier.	HDP-UTILS-1.1.0.22
Utils Base Url	URL to the repo storing utilities for the desired stack version.	Only required for Ambari 2.5 and earlier.	http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.22/repos/centos7
MPack Url	Provide MPack URL.	Only required for HDF.	http://public-repo-1.hortonworks.com/HDF/centos7/3.x/updates/3.2.0.0/tars/hdf_ambari_mp/hdf-ambari-mpack-3.2.0.0-520.tar.gz
Enable Ambari Server to download and install GPL Licensed LZO packages?	Use this option to enable LZO compression in your HDP/HDF cluster. LZO is a lossless data compression library that favors speed over compression ratio. Ambari does not install nor enable LZO compression libraries by default, and must be explicitly configured to do so. For more information, refer to Enabling LZO in Ambari documentation.	Optional and only available if using Ambari 2.6.1.0 or newer.	Disabled.

If you choose to use a base image with custom Ambari and/or HDP/HDF version, Cloudbreak validates the information entered. WhenCloudbreak detects that the information entered is incorrect, it displays a warning marked with the  sign. You should review all the warnings before proceeding and make sure that the information that you entered is correct. If you choose to proceed in spite of the warnings, check “Ignore repository warnings”.

Related Information

Ambari repo specification

Enabling LZO (Ambari)

Use spot instances

The "Use Spot Instances" option, available for each host group, allows you to use spot instances for all VMs on a given host group.

This is available for each host group on the Hardware and Storage page of the create cluster wizard. To use this option:

1. In create cluster wizard, navigate to the Hardware and Storage page.
2. For each host group, check the Use Spot Instances option option to use EC2 spot instances as your cluster nodes. Next, enter your bid price. The price that is pre-loaded in the form is the current on-demand price for your chosen EC2 instance type.

Note that:

- We recommend not using spot instances for any host group that includes Ambari server components.
- If you choose to use spot instances for a given host group when creating your cluster, any nodes that you add to that host group (during cluster creation or later) will be using spot instances. Any additional nodes will be requested at the same bid price that you entered when creating a cluster.
- If you decide not to use spot instances when creating your cluster, any nodes that you add to your host group (during cluster creation or later) will be using standard on-demand instances.
- Once someone outbids you, the spot instances are taken away, removing the nodes from the cluster.
- If spot instances are not available right away, creating a cluster will take longer than usual.

After creating a cluster, you can view your spot instance requests, including bid price, on the EC2 dashboard under INSTANCES > Spot Requests. For more information about spot instances, refer to AWS documentation.

Related Information

Creating a spot instance request (AWS)

Cloud storage

The options on the "Cloud storage" page of the advanced create cluster wizard allow you to configure access to Amazon S3.

If you would like to access S3 from your cluster, you must configure access through an instance profile as described in Configuring access to Amazon S3.

Related Information

Configuring access to Amazon S3

Recipes

The "Recipes" option allows you to select previously uploaded recipes (scripts that can be run pre or post cluster deployment) for each host group.

This option is available from the Cluster Extensions page of the advanced create cluster wizard. For more information, refer to Recipes.

Related Information

Recipes

Management packs

The "Management Packs" option allows you to select previously uploaded management packs.

This option is available on the Cluster Extensions page of the create cluster wizard. For more information, refer to Management packs.

Related Information

Management packs

Custom properties

The "Custom Properties "option allows you to set properties on a per-cluster basis by having them injected in the cluster blueprint.

Through this option, Cloudbreak allows you to create dynamic blueprints with property variables and set properties on a per-cluster basis by providing property values during cluster creation.

Setting custom properties

In order to set custom properties for a cluster you must:

1. Create a blueprint that includes property variables for the properties that you want to set.
2. When creating a cluster, select the blueprint and then specify the property values under Cluster Extensions > Custom Properties in the advanced view of the cluster wizard.

In the cluster creation phase, the property values in the blueprint will be replaced based on the input, picking up the parameter values that you provided.

Steps

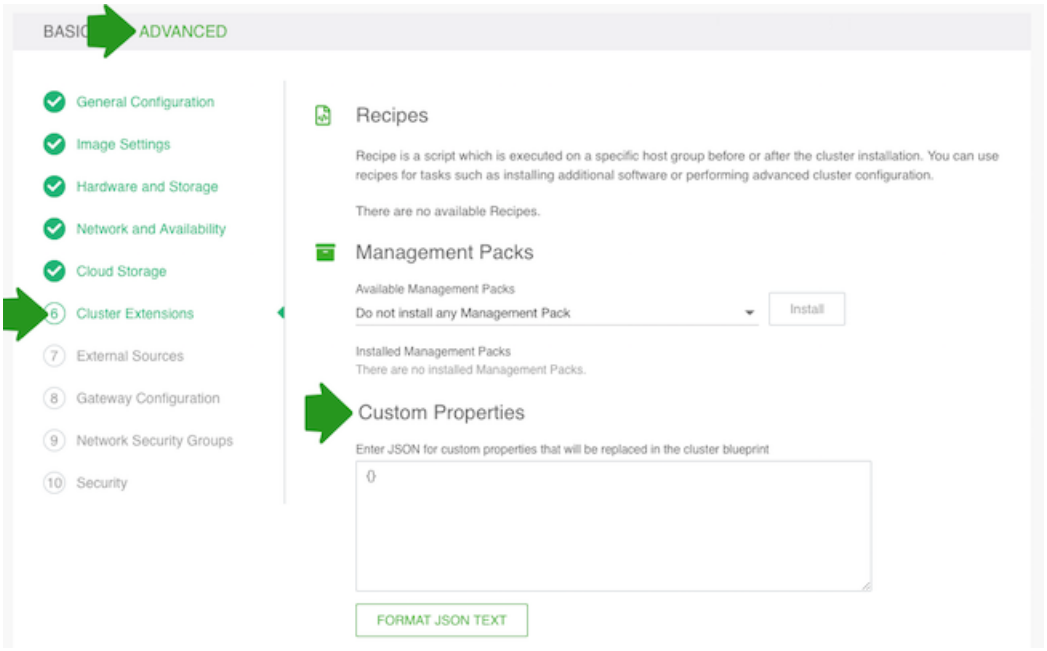
1. Prepare a blueprint which includes a template for the properties that you would like to set. Make sure to:
   - Include these templates in the “configurations” section of your blueprint.
   - Use the mustache format. Cloudbreak supports mustache kind of templating with {{{ }}} syntax.

   Example: This example provides a template for setting three properties:
   - fs.trash.interval
   - hadoop.tmp.dir
   - hive.exec.compress.output

```
...
{
  "core-site": {
    "fs.trash.interval": "{{{ fs.trash.interval }}}",
    "hadoop.tmp.dir": "{{{ my.tmp.directory }}}"
  }
},
{
  "hive-site": {
    "hive.exec.compress.output": "{{{ hive.exec.compress.output }}}"
  }
},
```

```
...
```

2. When creating a cluster:
   a. Under General Configuration > Cluster Type, select the blueprint prepared in the previous step.
   b. In the advanced view of the cluster wizard, under Cluster Extensions > Custom Properties, include a JSON file which defines the property values:



Example: The following JSON entry sets the values for the properties from the previous step:

```
{
  "fs.trash.interval": "4320",
  "hive.exec.compress.output": "true",
  "my.tmp.directory": "/hadoop/tmp"
 }
```

3. As a result, the values of hive.exec.compress.output, my.tmp.directory and fs.trash.interval will be replaced in the blueprint based on the input that you provided.

   Example: The property values will be replaced for the cluster as follows based on what was defined in the previous step:

```
...
{
  "core-site": {
    "fs.trash.interval": "4320",
    "hadoop.tmp.dir": "/hadoop/tmp"
  }
},
{
  "hive-site": {
    "hive.exec.compress.output": "true"
  }
},
```

External sources

The options on the "External Sources" page allow you to select existing external sources (databases, LDAP/AD, and proxy) to be used for a specific cluster.

To register external sources with Cloudbreak, refer to:

- External authentication source for clusters
- External database for cluster services
- Configure clusters to use a proxy

Related Information

External authentication source for clusters

External database for cluster services

Configure clusters to use a proxy

Ambari server master key

The "Ambari Server Master Key" option allows you to use an Ambari server master key.

The Ambari server master key is used to configure Ambari to encrypt database and Kerberos credentials that are retained by Ambari as part of the Ambari setup. To set this option:

1. In create cluster wizard, click on Advanced.
2. Navigate to the Security page.
3. Under Ambari Server Master Key, provide an Ambari server master key.

Enable Kerberos security

The "Enable Kerberos Security" option allows you to enable Kerberos security for your cluster.

This option is available on the Security page of the advanced create cluster wizard. For information about available Kerberos options, refer to Kerberos security.

Related Information

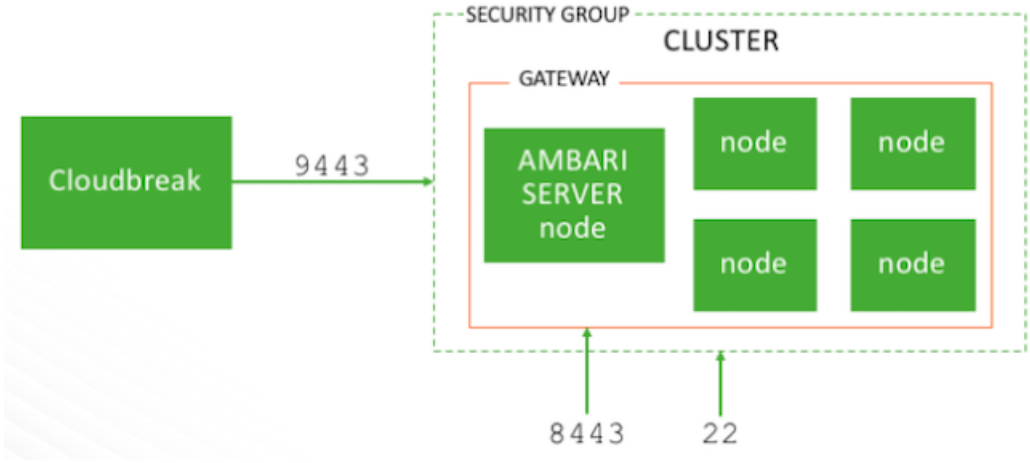
Kerberos security

Gateway configuration

The Gateway Configuration options allow you to change the settings of the gateway, which Cloudbreak installs and configures by default.

When creating a cluster, Cloudbreak installs and configures a gateway, powered by Apache Knox, to protect access to the cluster resources:

- This gateway is installed on the same host as the Ambari server.
- By default, transport layer security on the gateway endpoint is via a self-signed SSL certificate on port 8443. This is illustrated on the following diagram:



- The cluster resources that are accessible through the gateway are determined by the settings provided on the Gateway Configuration page of the basic create cluster wizard.
- By default, the gateway is deployed and Ambari is proxied through the gateway.
- The choice of cluster services to expose and proxy through the gateway depends on your blueprint. Cloudbreak analyzes your blueprint and provides a list of services that can be exposed through the gateway. You should review this list and select the services that should be proxied through the gateway.
- If you do not enable the gateway, or you do not expose Ambari (or any other service) through the gateway, you must configure access to those services on the security group on your own.
- Once a cluster is running, the gateway configuration selected during cluster create cannot be altered or reversed.

Services available via gateway

When gateway is configured. the cluster resources that are accessible through the gateway are determined by the settings provided on the Gateway Configuration page of the basic create cluster wizard.

The following cluster resources are available for access via the gateway endpoint:

Cluster resource	**URL**
Ambari	https://{gateway-host}:8443/{cluster-name}/{topology-name}/ambari/
Hive and Hive Server Interactive*	https://{gateway-host}:8443/{cluster-name}/{topology-name}/hive/
Job History Server	https://{gateway-host}:8443/{cluster-name}/{topology-name}/jobhistory/
Name Node	https://{gateway-host}:8443/{cluster-name}/{topology-name}/hdfs/
WebHDFS	https://{gateway-host}:8443/{cluster-name}/{topology-name}/webhdfs/
Resource Manager	https://{gateway-host}:8443/{cluster-name}/{topology-name}/yarn/
Spark History Server	https://{gateway-host}:8443/{cluster-name}/{topology-name}/sparkhistory/
Zeppelin	https://{gateway-host}:8443/{cluster-name}/{topology-name}/zeppelin/

* Refer to Access Hive via JDBC for more information.

The following applies:

- The gateway-host is the IP address or hostname of the Ambari server node.
- The cluster-name is the name of the cluster.

- The topology-name is the name of the gateway topology that you entered when creating the cluster. By default this is set to db-proxy.

Related Information

Access Hive via JDBC

Configure the gateway

When creating a cluster, you can configure the gateway on the Gateway Configuration page of the basic create cluster wizard.

Steps

1. In the create cluster wizard, navigate to the Gateway Configuration page.
2. Under Gateway Topology:
   - The gateway option is enabled by default.
   - The Name for your gateway is set to db-proxy by default. You can update it if you would like. This name is used in the URLs for the cluster resources.
   - Under Exposable Services, the choice of cluster services to expose and proxy through the gateway depends on your blueprint. Cloudbreak analyzes your blueprint and provides a list of services that can be exposed through the gateway. You should review this list and select the services that should be proxied through the gateway. By default, only Ambari is exposed through the gateway.
   - We recommend that you expose the following services through the gateway:
     - Analytics blueprint: Hive and Zeppelin
     - ETL blueprint: No additional services need to be exposed
     - Data science: Spark and Zeppelin
3. Under Exposable Services, use the dropdown to select services that should be exposed via the gateway. To expose a service, select it and click Expose. Select ALL to expose all.

Related Information

Services available via gateway

Configure single sign-on (TP)

When creating a cluster, if you selected to configure a gateway, on the advanced Gateway Configuration page of the advanced create cluster wizard, you can also configure the gateway to be the SSO identity provider.

Note:

This option is technical preview.

Prerequisites

You must have an existing authentication source (LDAP or AD) and register it with Cloudbreak, as described in Using an external authentication source for clusters.

Steps

1. In the create cluster wizard, select the advanced mode.
2. On the External Sources page, under Configure Authentication, select to attach a previously configured LDAP to the cluster.
3. On the Gateway Configuration page, under Single Sign On (SSO), click the toggle button to enable SSO.

Related Information

Eternal authentication source for clusters

Obtain gateway URLs

Once your cluster is running, you can obtain the Ambari URL and the URLs of the cluster services from the Gateway tab in the cluster details.

Note:

This tab is only available when gateway is enabled.

The URL structure is as described in Services available via gateway.

Related Information

Services available via gateway