Hortonworks Data Platform

Deployment Guide for Azure IaaS

(March 1, 2016)

docs.cloudera.com

Hortonworks Data Platform: Deployment Guide for Azure IaaS

The Hortonworks Data Platform, powered by Apache Hadoop, is a massively scalable and 100% open source platform for storing, processing and analyzing large volumes of data. It is designed to deal with data from many sources and formats in a very quick, easy and cost-effective manner. The Hortonworks Data Platform consists of the essential set of Apache Hadoop projects including MapReduce, Hadoop Distributed File System (HDFS), HCatalog, Pig, Hive, HBase, ZooKeeper and Ambari. Hortonworks is the major contributor of code and patches to many of these projects. These projects have been integrated and tested as part of the Hortonworks Data Platform release process and installation and configuration tools have also been included.

Unlike other providers of platforms built using Apache Hadoop, Hortonworks contributes 100% of our code back to the Apache Software Foundation. The Hortonworks Data Platform is Apache-licensed and completely open source. We sell only expert technical support, training and partner-enablement services. All of our technology is, and will remain free and open source.

Please visit the Hortonworks Data Platform page for more information on Hortonworks technology. For more information on Hortonworks services, please visit either the Support or Training page. Feel free to Contact Us directly to discuss your specific needs.

Table of Contents

List of Tables

1. HDP Deployment Guide for Azure IaaS

Microsoft® Azure™ IaaS is an open, flexible, enterprise-level cloud computing platform that is used for building, deploying, and managing applications and services through a global network of Microsoft-managed data centers.

Microsoft has expanded Azure IaaS to include a marketplace with thousands of certified, open source, community software applications, developer services, and data, all pre-configured for Microsoft Azure.

You can deploy Hortonworks Data Platform (HDP) on Azure IaaS using the Azure Marketplace offerings designed for HDP.

1.1. HDP Azure Marketplace Offerings

There are two Hortonworks Data Platform (HDP) Azure marketplace offerings:

- Hortonworks Data Platform Standard
- Sandbox

This guide provides setup, deployment, and reference information for the Hortonworks Data Platform (HDP) Standard offering for Azure.

For information about deploying the Hortonworks Sandbox on Azure, see Deploying Hortonworks Sandbox on Microsoft Azure.

The following table shows the features of the two marketplace offerings for HDP.

Understanding the differences between the offerings helps you select the right method of deployment.

Table 1.1. HDP Azure Market Offerings

Azure Marketplace Offering	Targeted HDP Users	Description	Customization	Designed for Production Ready Environment
Hortonworks Data Platform (HDP) Standard	Advanced users. Experience with HDP and the ability to make cost/benefit decisions related to hardware.	Designed for a multi-node HDP cluster. You tell us how many nodes and the type of machines you want to use and we deploy it for you.	Cluster built based on your choice of machine type and number of nodes.	Yes
Sandbox	Beginning users. Little or no experience with Hadoop or HDP.	Single-node VM to get you started with HDP. Provides tutorials to explore various components of the HDP stack.	None	No

1.2. Before You Begin

Before you deploy an HDP multi-node cluster from the Azure Marketplace, familiarize yourself with the information in this section.

1.2.1. Hortonworks Data Platform Standard Specifications

This section discusses the specifications of the HDP Standard offering, including the Azure Resource Manager (ARM), Virtual Machines (VMs), Virtual Machine Image, storage, and security.

1.2.1.1. Azure Resource Manager (ARM)

The HDP Standard offering takes advantage of the Azure Resource Manager (ARM).

ARM allows you to deploy, manage, and monitor all of the resources as a collective group, as opposed to individually.

Once the HDP Standard offering is deployed, you can use the Azure Portal to easily identify and view information about the offering's components. You can search for a selected resource group or a resource group that was created during deployment.

ARM has template support that enables the offering to be launched outside of the Azure Marketplace. If you are an advanced user, you can modify existing templates for a more customized deployment. See for more information about templates, see the Azure Quickstart Templates

Note

The Azure Quickstart Templates are not currently available for the HDP Standard offering.

1.2.1.2. Virtual Machines

The HDP Standard offering enables you to select various Azure Virtual Machines (VMs). See Azure Windows Virtual Machines Documentation for information regarding the types of Azure Virtual Machines.

Virtual Machines vary in number of cores, processor speeds, cache types, and the ability to be used with premium storage. When setting up the HDP Standard offering, you can selected an A-, D-, or DS-Series machine. The smallest non-high availability cluster size you can create with the offering is six machines (three masters nodes and three worker/slave nodes). When enabling High Availability (HA), you need five master nodes. For additional information see the Typical Hadoop Cluster section of the *Cluster Planning Guide*.

Note

The Azure Resource Manager's default number of cores is 20, and is not enough to deploy the HDP Standard offering. You need to request additional cores in advance. For more information regarding Azure limits see Azure subscription and service limitations, quotas, and constraints .

By default, the offering suggests to use D-Series machines for Masters and DS-Series for Slaves.

1.2.1.3. Virtual Machine Image

The HDP Standard offering utilizes a static image to deploy a multi-node HDP cluster for a consistent and quick deployment.

The HDP Standard offering is deployed with the OS, Java version, HDP version, and Apache components specified in the following table.

Note

These system specifications are part of the image used to deploy HDP on Azure IaaS, and are not configurable prior to deployment.

Table 1.2. HDP on Azure IaaS System Specifications

Specification	Supported Versions and Other Information
Operating system	CentOS Linux release 7.2.1511
Java version	• OpenJDK Version "1.8.0_71" • OpenJDK Runtime Environment (build 1.8.0_71-b15) • OpenJDK 64-Bit Server VM (build 25.71-b15, mixed mode)
HDP version	2.4.0 See HDP 2.4.0 Release Notes
Ambari version	2.2.1.0-161, Ambari 2.2.1.1 Release Notes
Supported HDP Components	All standard HDP components are installed, except for the following: • Apache Accumulo • Apache Mahout • Apache Ranger • HDP search Apache Accumulo, Apache Mahout, Apache Ranger, and HDP search can be installed post-deployment.
Supported Databases	• Ambari: postgresql-server-9.2.14-1 • Hive Metastore: mysql-community-server-5.6.30-2 • Oozie: derby database

1.2.2. Security

The HDP Standard offering provides the following security measures:

- Worker nodes do not have a public IP address to shield from public access.
- Only client and management nodes are publicly accessible from the Internet.

- The password for the Ambari UI is configured to a user specified value, not `admin`.
- The HDP Standard offering does not have Kerberos enabled at deployment. You can enable Kerberos and secure the cluster post-deployment. Refer to the Ambari Security Guide for more information.

Note

Though not recommended, you can set the exposure of machine ports in the network security section for the deployed cluster in from the Azure Portal.

1.2.3. Storage

The Azure cloud infrastructure supports the following storage types for big data:

- Azure Data Lake Store (ADLS)
- Windows Azure Blog Store (WABS)
- Virtual Hard Disks (VHD)

1.2.3.1. Azure Data Lake Store (ADLS)

Azure Data Lake Store (ADLS) is a hyper-scale repository for big data analytic workloads. For more information on ADLS, refer to Overview of Azure Data Lake Store.

The HDP Standard offering does not support ADLS.

1.2.3.2. Virtual Hard Disks (VHDs)

The HDP Standard offering uses Virtual Hard Disks (VHDs) for storage.

The HDP Standard offering uses 10 attached VHDs per VM, which gives the best balance of performance and storage density.

Depending on the type of VM that you use, it is possible to use Azure Premium Storage. DS-, DSv2-, or GS-Series Azure VMs can use Azure Premium Storage.

For additional information regarding Azure Premium storage, see Premium Storage: High-Performance Storage for Azure Virtual Machine Workloads.

Note

With Azure Premium Storage, you pay for all storage capacity up front. With non-premium storage, you only pay for the storage you use.

1.2.3.3. Windows Data Storage Blob (WASB)

If data is stored as Windows Data Storage Blob (WASB), it persists in the cloud and can be accessed after an HDP cluster has been de-provisioned.

WASB is an extension of the Hadoop Distributed File System (HDFS) APIs and allows multiple HDP clusters on Azure access to a central store.

The HDP Standard offering enables you to configure WASB after deployment.

Note

WASB is not supported with WebHDFS.

Ambari relies heavily on WebHDFS to access the filesystem. Because of this, WASB does not support Ambari Views.

For additional information regarding WASB, refer to:

- Understanding WASB and Hadoop Storage in Azure
- Why WASB Makes Hadoop on Azure So Very Cool
- Hadoop Azure Support: Azure Blob Storage

1.3. Gathering Your Deployment Information

Deployment of the HDP Standard offering requires that you provide information about the type of cluster you want to configure.

Before you deploy the HDP Standard offering, be sure to gather the cluster layout information specified in the following table:

Table 1.3. Cluster Layout Information for Your Hortonworks Data Platform Standard

Information	Description
Cluster name	Name of the HDP cluster you want to create. The cluster name must be between 3 and 24 characters, and can contain only numbers and lower case letters. Once created, you can search the Azure Portal dashboard for this cluster name.
Cluster admin Username	User name for the virtual machine user login and the Ambari administrator.
Authentication Type	Type of authentication to use. The following authentication types are supported: • Password authentication. Authentication password must be between six and 72 characters, and have three of the following requirements: • One lower case character • One upper case character • One number • One special character • OpenSSH public key. Generate an OpenSSH public key with ssh-keygen or PuttyGen, depending on your OS.
Ambari Password	Password to authenticate the Ambari administrator. This password can be the same as the authentication password, if you have selected password authentication. The same length and character requirements apply.
Do you want to create a High Availability (HA) cluster?	High Availability (HA) clusters run the storage and processing services in redundant mode, enabling your tasks to complete, even if a service they depend on has a failure.

Information	Description
	HA is recommended for production clusters. HA clusters are deployed with five master nodes, and non-HA clusters are deployed with three master nodes.
Master Node Type	In a Hadoop cluster, a master node oversees storage, processing, and management services. You can select from the following, which all have size, strength, speed, and price implications: • A-Series - A4, A7, A10 • D-Series - DS4, DS13, DS14, D13_V2, D14_V2, D4_V2, D4, D13 • G-Series - G4, G5, GS3, GS4, GS5 For more information regarding machine pricing on Azure, see Virtual Machines Pricing.
Number of Worker Nodes	In a Hadoop cluster, worker nodes make up the majority of virtual machines and perform data storage and processing jobs. When you are creating an HDP cluster in Azure, you can select either three or five worker nodes.
Worker Node Type	The choices of machines here are the same as the Master Node Type section.

1.4. Deploying Hortonworks Data Platform Standard

Prior to deployment, ensure that you have completed the following tasks:

- Created an Azure account.

  If you do not have one you can create one from the Azure Home page.

- Prepared the information requested in Gathering Your Deployment Information.

Perform the following steps to deploy the Azure Marketplace Hortonworks Data Platform Standard:

1. Go to the Azure Marketplace and use the search tool to search for Hortonworks.
2. Click the **Hortonworks Data Platform** icon to select it.
3. Click **Deploy** and then **Create**.

   If you are not already logged into Azure, the Login screen appears, and you are prompted to enter your credentials.

   The *Basics* page appears.
4. On the *Basics* page, provide some information about yourself, and select an existing resource group or create a new one.

   Click **OK**.

   The *Cluster Layout* page appears.
5. Provide the information you collected in the Gathering Your Deployment Information section.

   Click **OK**.

The *Summary* page displays.

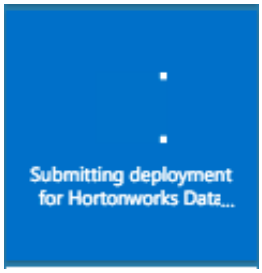
6. Review the information on the *Summary*.

   If you need to edit any of your information, click on the section that you want to make changes on and edit it.

   Click **OK**.

7. Click **Create**.

   The following displays on your Dashboard in the Azure Portal:



8. When the deployment is complete, the Azure home page displays

   At this point, the HDP Standard offering is up and running.

1.5. Post Installation Tasks

After you have deployed HDP, you can perform a few post-installation tasks to configure HDP according to your operational objectives.

1.5.1. Accessing Ambari

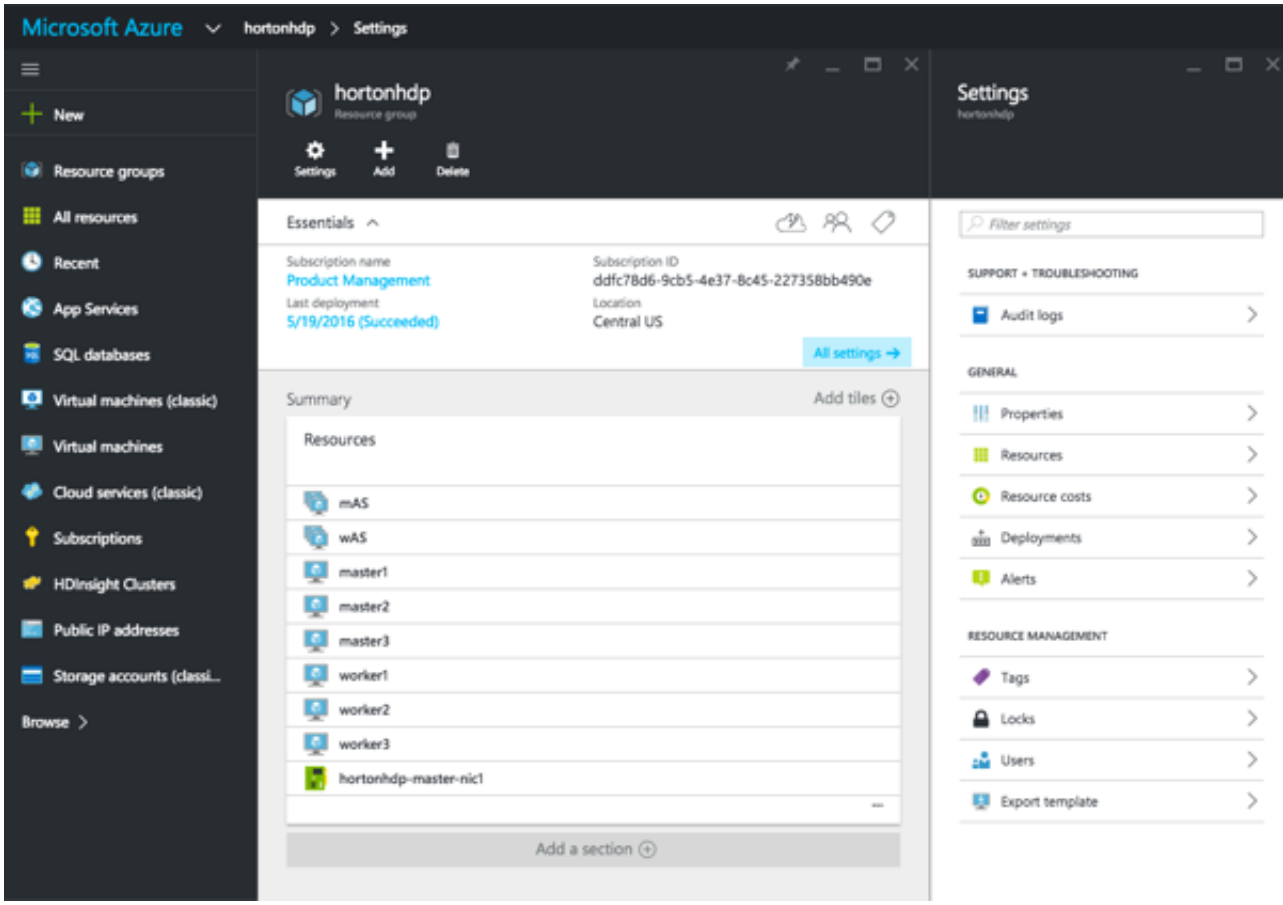
You can use Ambari to provision, manage, and monitor HDP.

You use Ambari to complete most post-installation tasks, so your first post-installation task is to launch the Ambari interface to access your new HDP cluster.

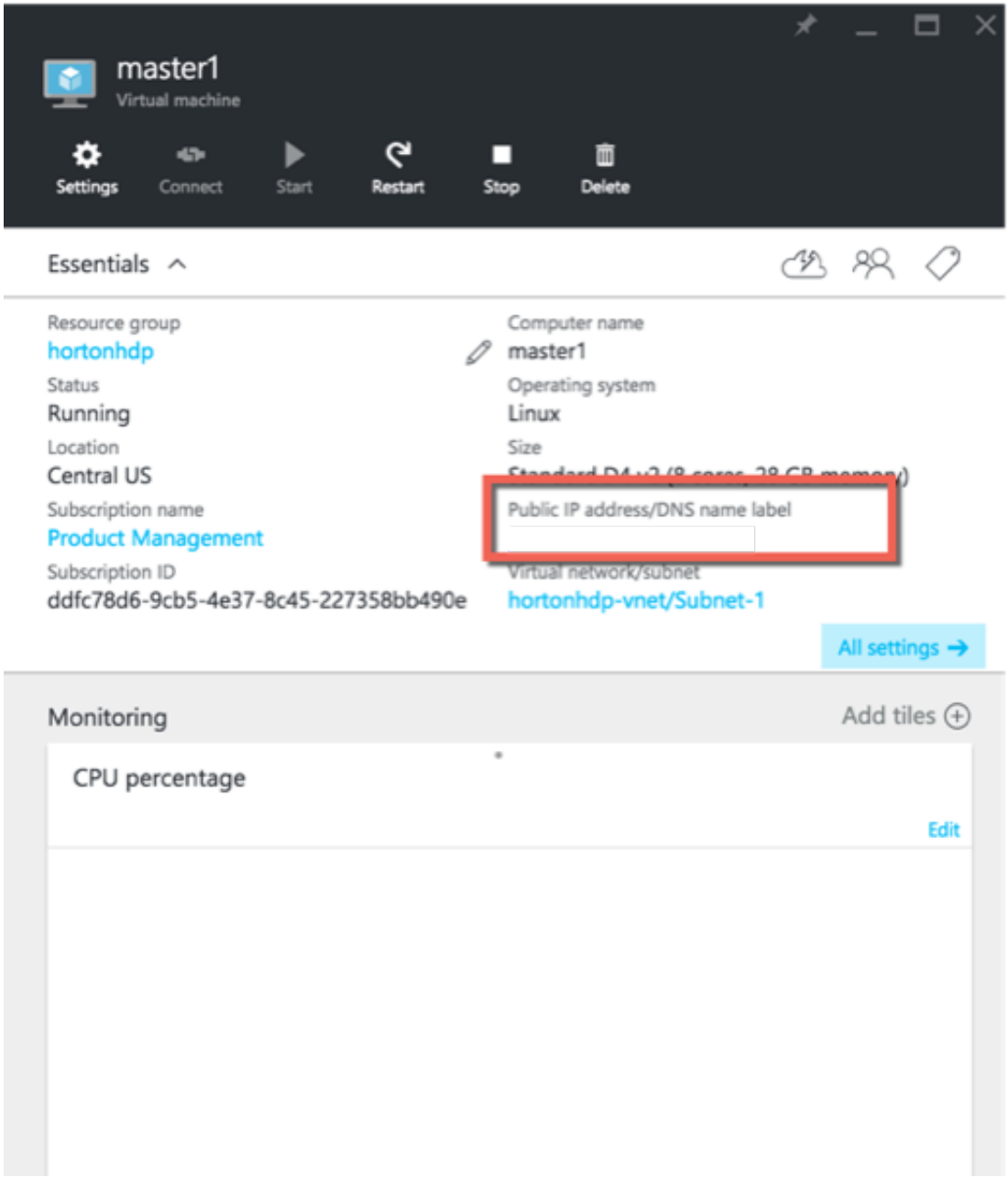
1. From the Azure Portal Dashboard, click the tile for the resource you just deployed. For example, if you named your HDP cluster `hortonhdp`, click the following tile:

2. From the cluster_name resource group blade that displays the virtual machines you have created, click on `master1` on which Ambari is deployed.

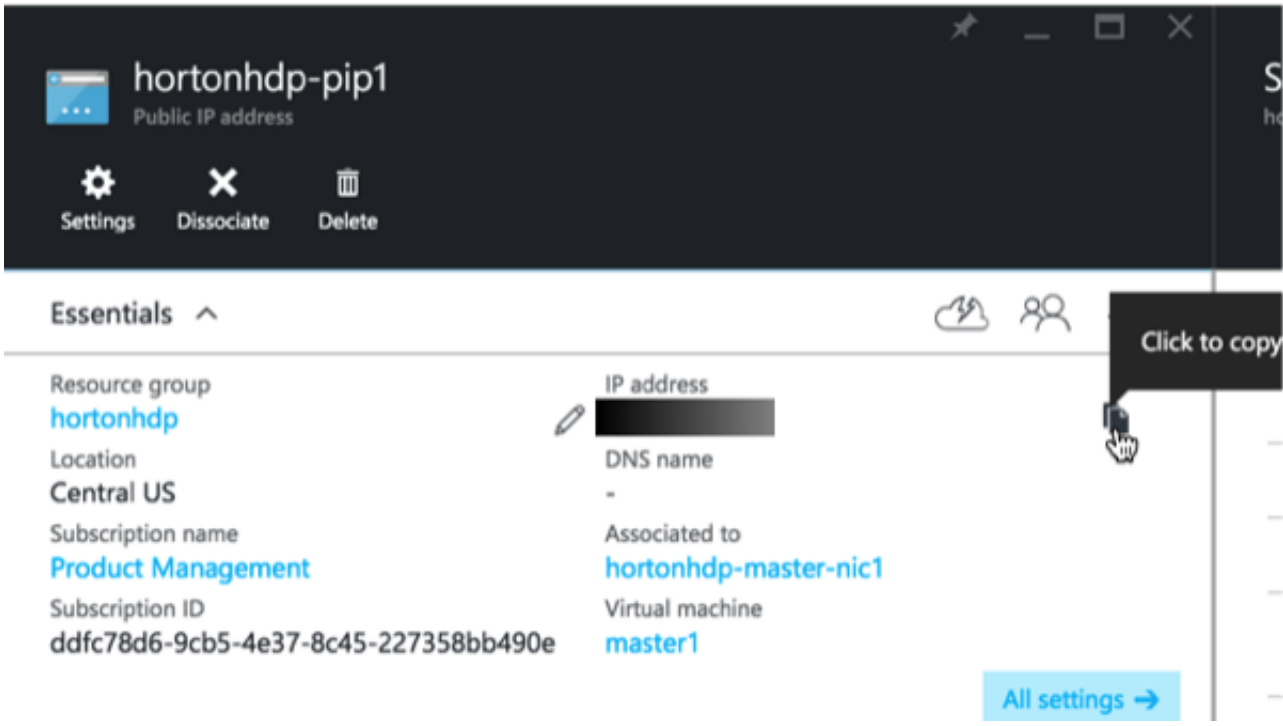
   For example, in the following screenshot, you have created a 3-master, 3-worker node cluster, and Ambari running on the master1 node.



3. Click the **Public IP address** of the machine running the Ambari Server.



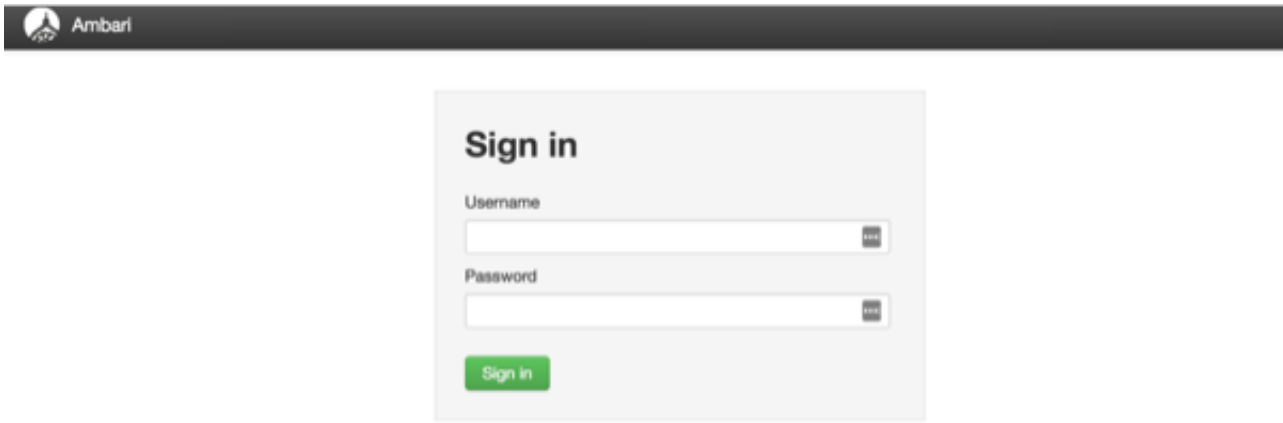
4. Copy the public IP address by clicking the **Click to copy** icon.



5. To launch Ambari, paste this IP address into your browser, followed by a colon (:) and the default Ambari port, `8080`.

   For example: `http://16.94.156.240:8080`

6. Login into Ambari with the user name `admin`, and the Ambari password you created during setup.



7. The Ambari Welcome screen appears.

Refer to the Ambari User's Guide for information about Ambari.

1.5.2. Enabling High Availability (HA)

The Hortonworks Data Platform Standard has the option of having a High Availability HDP cluster. High Availability is ideal for a production ready environment. The following components are Highly Available when you select “Yes” from the cluster layout at Azure Portal.

- HDFS (Namenodes)
- YARN (Resource managers)
- Hive (metastore, hiveserver2, webhcat)

Once the cluster is deployed you can also use the Ambari administration interface to enable HA post-installation, and on other services. For more information, see Managing Service High Availability in the *Ambari User's Guide*.

1.5.3. Scaling Your Cluster

You might need to scale your HDP cluster on Azure when space becomes fully utilized or you require more processing for your workloads.

Ambari allows you to scale your cluster, provided that the machines are pre-configured. Provisioning new machines requires multiple steps on the Azure infrastructure. Contact Hortonworks Professional Services for assistance.

1.5.4. Moving Your Data Out of Hortonworks Data Platform Standard

The HDP Azure Marketplace offerings do not support Windows Azure Storage Blob (WASB) or Azure Data Lake Store (ADLS).

If you decide you no longer need to use the cluster and want to keep the data, you must transfer your data outside of the cluster. Contact Hortonworks Professional Services for assistance.

1.6. Upgrading HDP

Once your HDP cluster is deployed using Azure IaaS, you can treat it like any other Ambari-managed cluster.

To upgrade to a more recent version of HDP, see the Ambari Upgrade Guide.