



# HDInsight Administration and Security

# HDInsight – What is it?

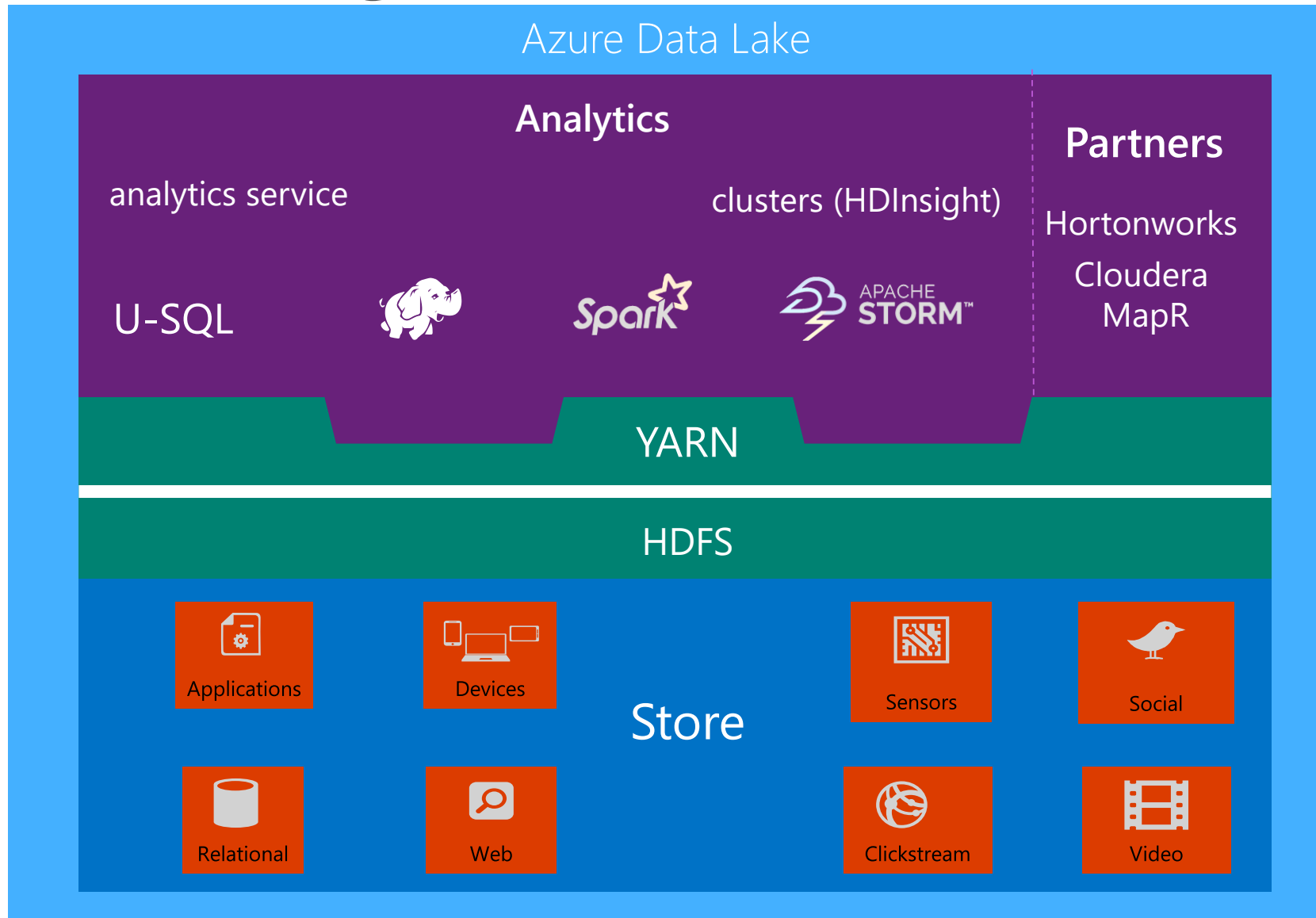
A standard Apache Hadoop distribution offered as a managed service on Microsoft Azure

- ❖ Based on the Hortonworks Data Platform (HDP)
- ❖ Provisioned as clusters on Azure. Clusters can run on Windows or Linux Servers.
- ❖ Offers a capacity-on-demand, pay-as-you-go pricing model
- ❖ Integrates with:
  - Azure Blob Storage and Azure Data Lake Store for the Hadoop File System (HDFS)
  - Azure Portal for management and administration
  - Visual Studio for application development tooling

In addition to the core, HDInsight supports the Hadoop Ecosystem

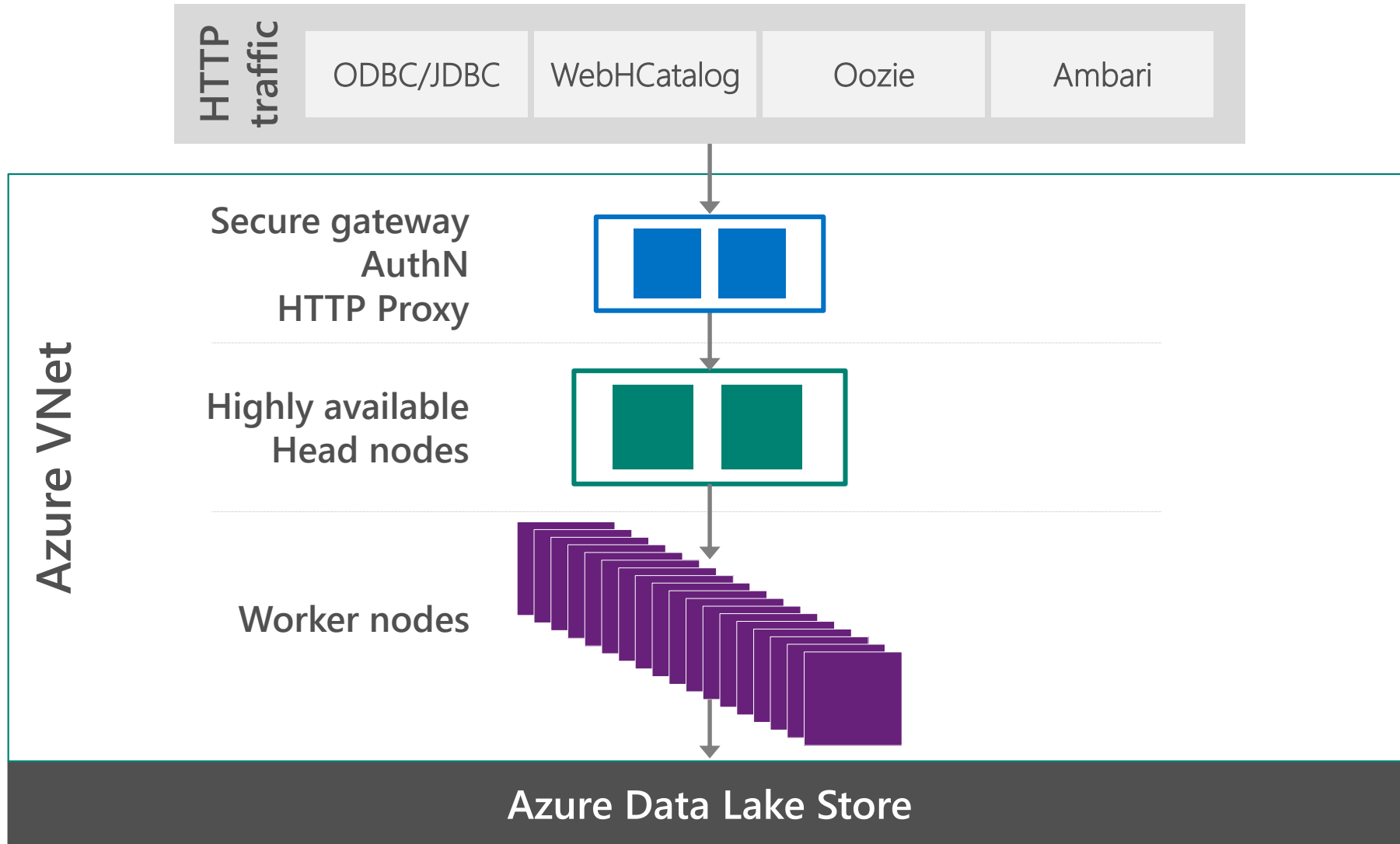


# HDI Insight: How/Where it fits?



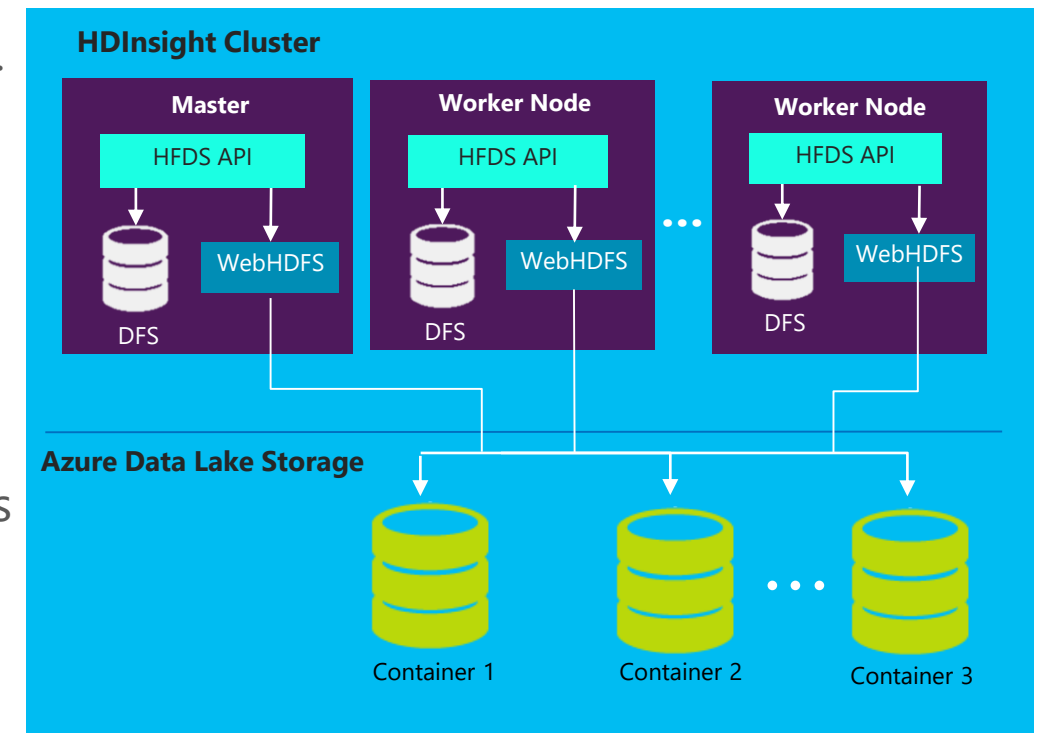
- ✓ Integrated analytics and storage
- ✓ Fully Managed
- ✓ Easy to use – “dial for scale”
- ✓ Proven at scale
- ✓ Analyze data of any size, shape or speed
- ✓ Open-standards based

# HDInsight cluster architecture

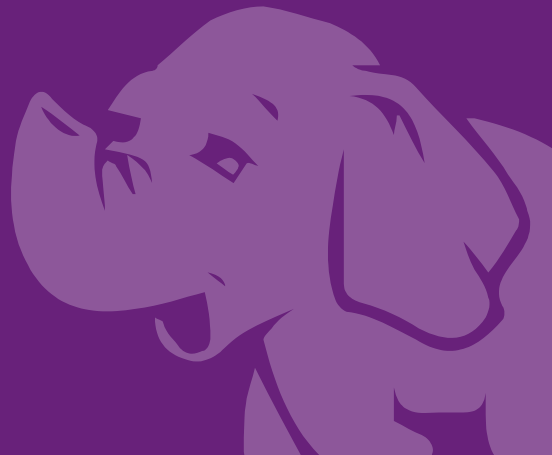


# Benefits of Hadoop as a Azure Service






- ❖ HDInsight clusters can be configured in just a few minutes
- ❖ Through the HDFS interface, HDInsight can operate directly on data stored in Azure Blob Storage or Azure Data Lake Store. A separate and dedicated HDFS cluster is not required. Benefits include:
  - Data reuse and sharing
  - Elastic scale-out
  - Lower data storage costs
  - Protection against data loss
- ❖ High Performance: The high-speed flat networks in Azure datacenters provide fast access between the virtual machines in the cluster and Azure Blob Storage or Azure Data Lake Store so data movement is very efficient
- ❖ Commission and decommission HDInsight clusters at will
- ❖ Visualize the data stored in Hive tables using Excel and the set of add-ins, such as *Power Query*, *Power View*, *PowerPivot*, and *Power Map*.



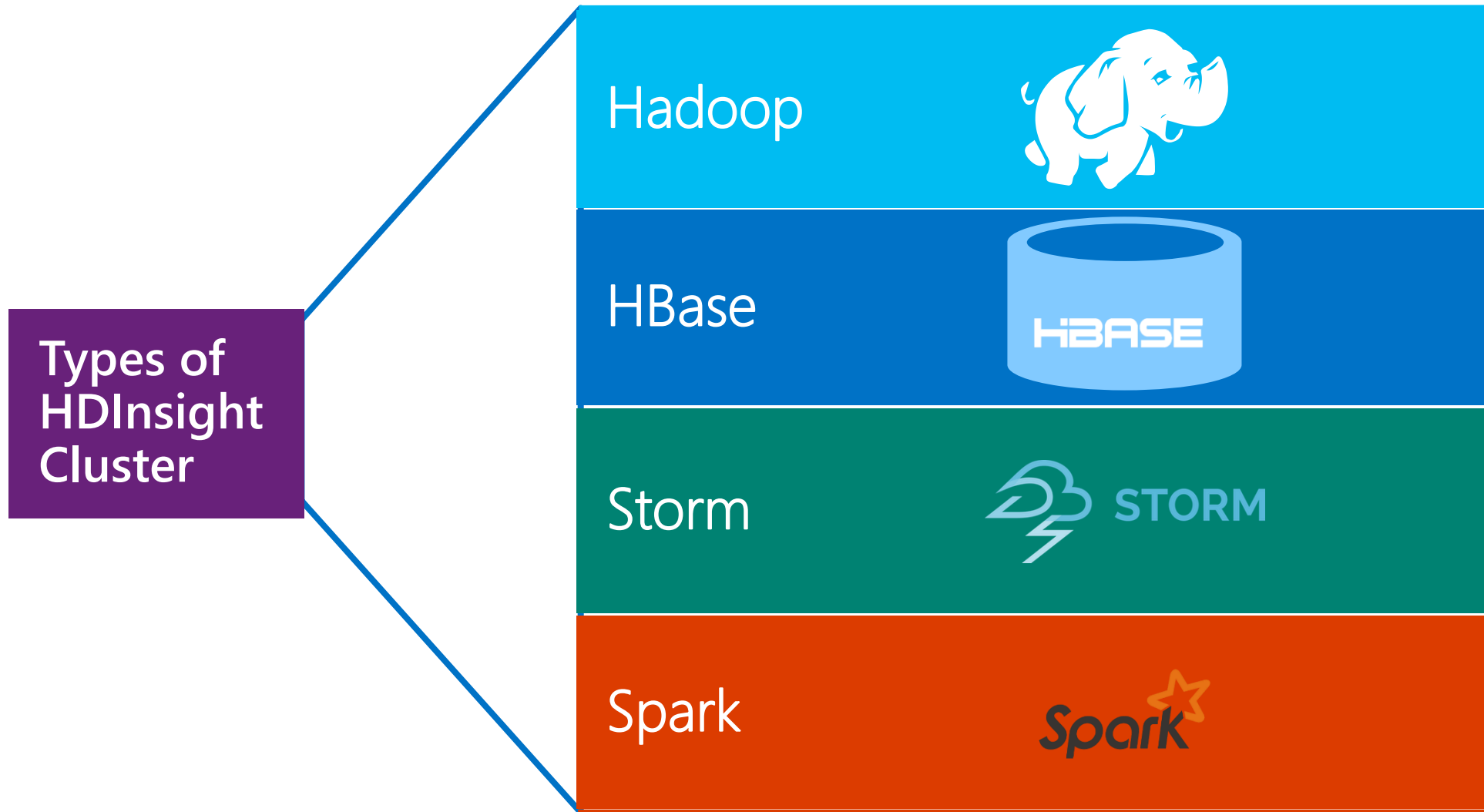
# HDI Insight on Linux: Administration Overview



## Agenda

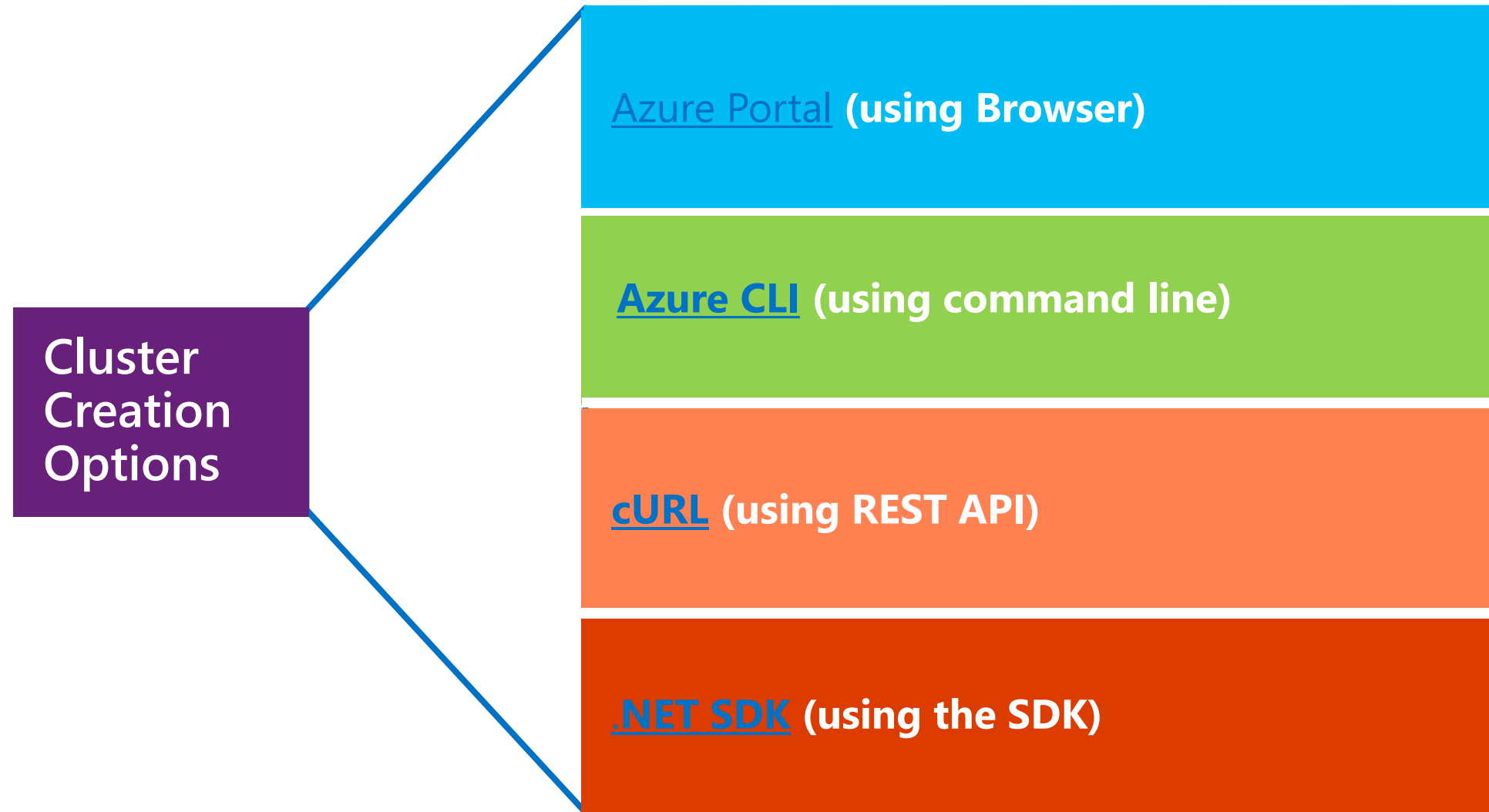
-  Creating HDI Clusters
  -  Script Actions
-  Audit Logs
-  HDI Configuration
-  Ambari Web UI

# Cluster Types Overview





# Ways to create HDInsight (Linux) Clusters





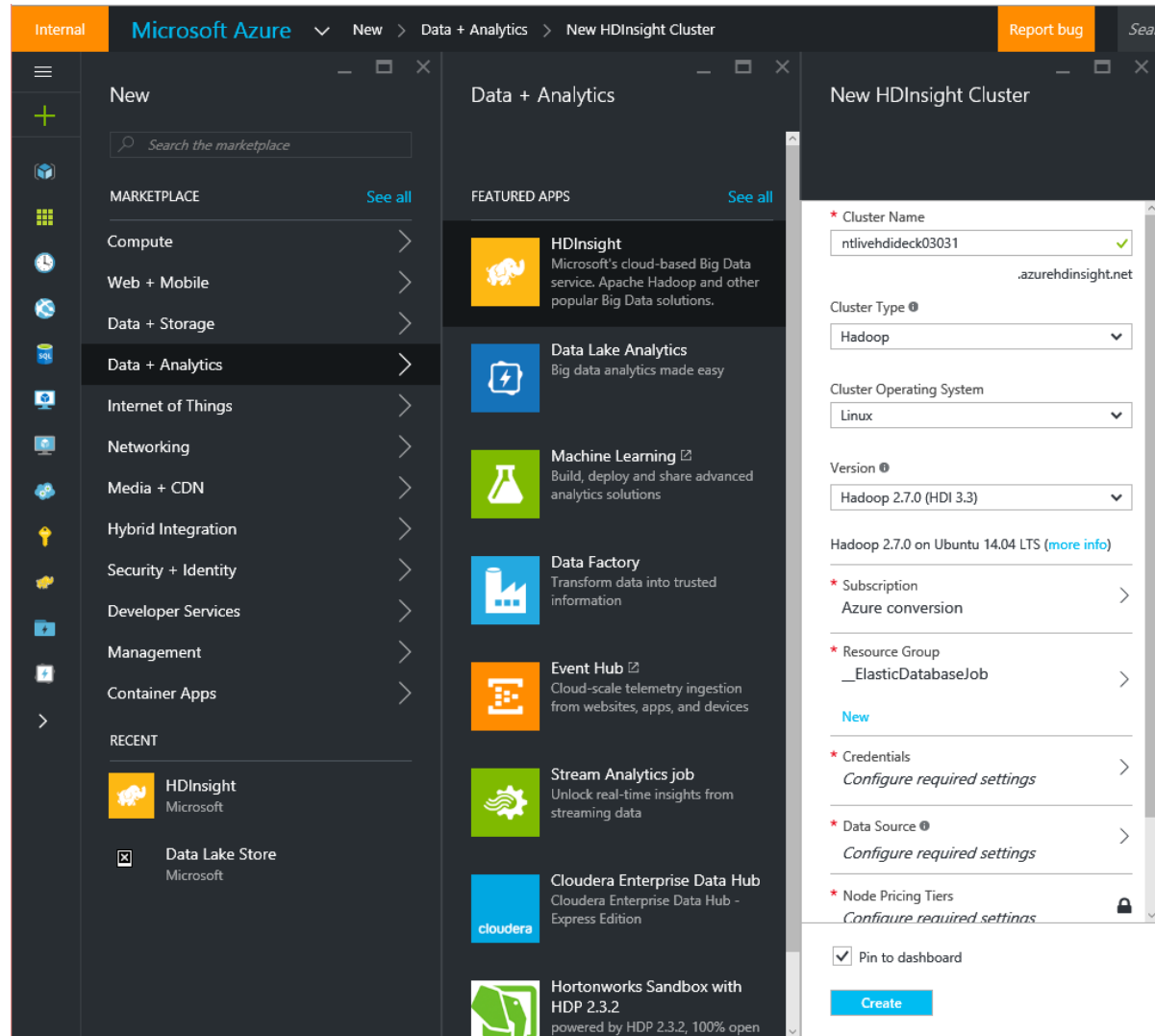
# Creating a HDInsight Cluster via the Portal

# Azure Portal

Azure Portal provides a guided wizard to create HDInsight clusters.

The key parameters to specify include:

- Type of Hadoop cluster
- OS (Linux or Windows)
- Hadoop Version
- Azure storage data source
- Number and size of nodes i.e. head nodes, worker nodes etc)
  - The actual types of nodes depends on cluster types
- Security credential for accessing web/REST APIs and for SSH
- Optional metadata store
- Azure Virtual Network
- Script Action for customization



# Step 1: Specify Cluster Type and OS

OS choices are:

- Windows
- Linux

Cluster type choices are:

- Hadoop
- HBase
- Storm
- Spark

The screenshot displays the 'New HDInsight Cluster' configuration page in the Microsoft Azure portal. The breadcrumb navigation at the top reads: 'Microsoft Azure > New > Data + Analytics > New HDInsight Cluster > Cluster Type configuration'. The left sidebar shows the 'New' button and various resource categories. The main configuration area is divided into two panels. The left panel lists required settings: 'Cluster Name' (samplehdi), 'Subscription' (Pay-As-You-Go), 'Credentials', 'Data Source', 'Node Pricing Tiers', and 'Resource Group' (ADLA\_Benchmark). The right panel, titled 'Cluster Type configuration', shows 'Cluster Type' set to 'Hadoop' and 'Operating System' set to 'Linux'. A tooltip is visible over the 'Operating System' dropdown, listing four options: Hadoop (Terabyte-scale processing), HBase (Fast and scalable NoSQL database), Storm (Reliably process infinite streams of data in real-time), and Spark (Fast data analytics and cluster computing using in-memory processing). Below the dropdown, pricing information is shown: '+ 0.00 USD/CORE/HOUR' for the standard tier and '+ 0.02 USD/CORE/HOUR' for the HDInsight tier.

Microsoft Azure > New > Data + Analytics > New HDInsight Cluster > Cluster Type configuration

New HDInsight Cluster

Cluster Type configuration

Learn about HDInsight and cluster versions. [Learn more](#)

Cluster Type: Hadoop

Operating System: Linux

Version: Hadoop 2.7.1 (HDI 3.4)

Cluster Tier (mo): STANDARD

Admin Management: + 0.00 USD/CORE/HOUR

Scalability: + 0.02 USD/CORE/HOUR

99.9% Uptime

Autonomous: + 0.00 USD/CORE/HOUR

for HDInsight: + 0.02 USD/CORE/HOUR

Hadoop: Terabyte-scale processing with Hadoop components like Hive (SQL in Hadoop), Pig, and Oozie.

HBase: Fast and scalable NoSQL database.

Storm: Reliably process infinite streams of data in real-time.

Spark: Fast data analytics and cluster computing using in-memory processing.

# Step 2 :Specify Version and Cluster Tier

Microsoft Azure

New > Data + Analytics > New HDInsight Cluster > Cluster Type configuration

### New HDInsight Cluster

Cluster Name: samplehdi ✓  
Subscription: Pay-As-You-Go  
Credentials: Configure required settings  
Data Source: Configure required settings  
Node Pricing Tiers: Please configure required settings  
Optional Configuration  
Resource Group: ADLA\_Benchmark

### Cluster Type configuration

Cluster Type: Hadoop  
Operating System: Linux Windows  
Version: Hadoop 2.7.1 (HDI 3.4)

### Cluster Tier (more info)

STANDARD	PREMIUM
<b>Administration</b> Manage, monitor, connect	<b>Administration</b> Manage, monitor, connect
<b>Scalability</b> On-demand node scaling	<b>Scalability</b> On-demand node scaling
<b>99.9% Uptime SLA</b>	<b>99.9% Uptime SLA</b>
<b>Automatic patching</b>	<b>Automatic patching</b>
<b>Microsoft R Server for HDInsight</b>	
+ 0.00 USD/CORE/HOUR	+ 0.02 USD/CORE/HOUR

Premium tier is available only with:

- Version 3.4
- For Hadoop and Spark

Other supported versions are 3.3 and 3.2

# Step 3: Specify SSH and Admin Credentials

Microsoft Azure

New > Data + Analytics > New HDInsight Cluster > Cluster Credentials

New

Resource groups

All resources

Recent

App Services

Virtual machines (classic)

Virtual machines

SQL databases

Cloud services (classic)

Subscriptions

Browse >

New HDInsight Cluster

Cluster Name  
samplehdi ✓  
.azurehdinsight.net

Subscription  
Pay-As-You-Go >

Select Cluster Type ⓘ  
Standard Hadoop on Linux (3.4) >

Credentials  
Configure required settings >

Data Source ⓘ  
Configure required settings >

Node Pricing Tiers  
Please configure required settings 🔒

Optional Configuration 🔒

Cluster Credentials

Create login and remote access credentials for the cluster.

Cluster Login Username ⓘ  
admin ✓

Cluster Login Password  
..... ✓

Confirm Password  
..... ✓

SSH Username ⓘ  
sshuser ✓

SSH Authentication Type  
PASSWORD PUBLIC KEY

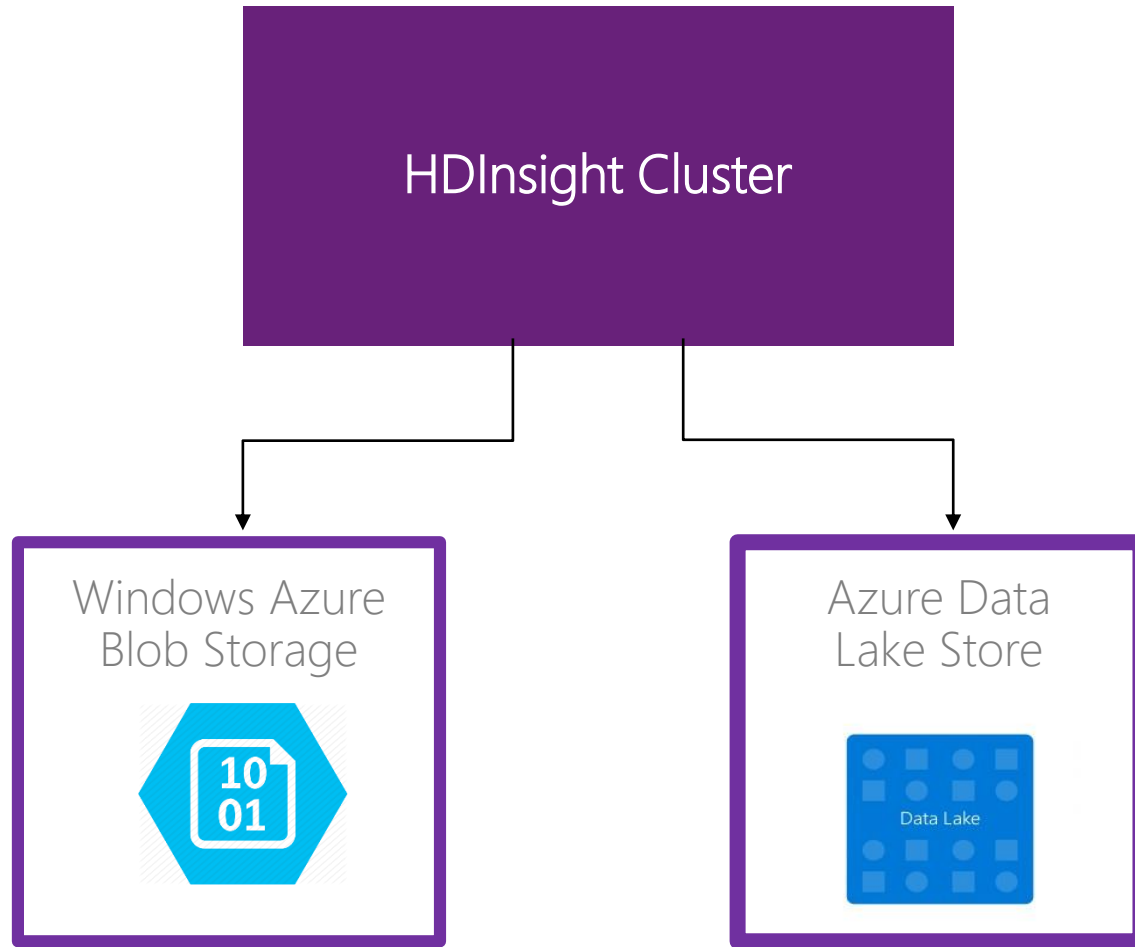
SSH Password  
..... ✓

Confirm Password  
..... ✓

Credentials to submit jobs to the cluster and the Ambari Dashboard

Credentials to remotely access the cluster

# Two storage options: WASB or ADLS

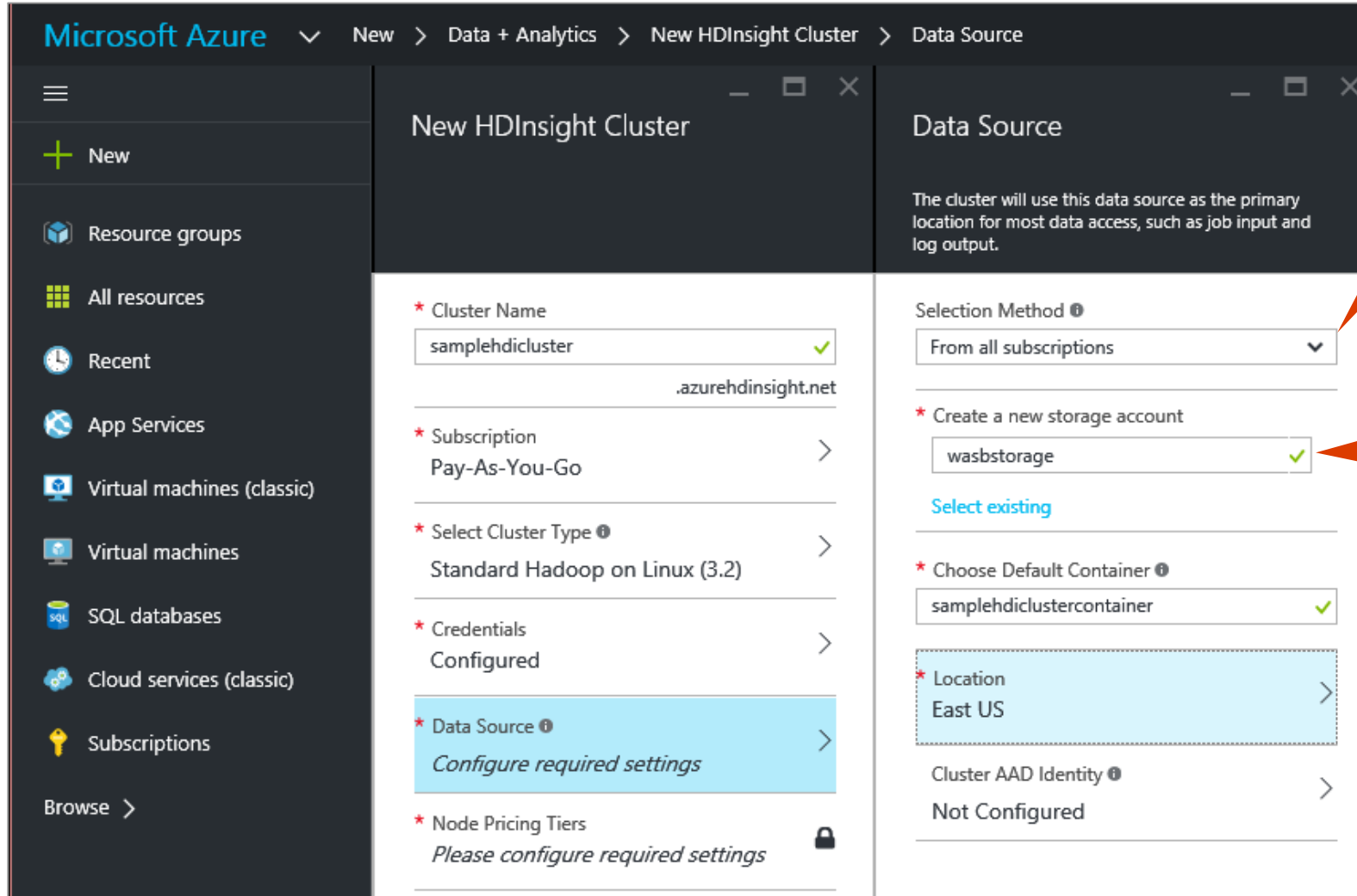


For **Hadoop Clusters**, ADLS can only be used with an additional storage account. The default is still WASB.

For **Storm clusters** ADLS can be used to write data from a Storm topology. Data Lake Store can also be used to store reference data that can then be read by a Storm topology.

**For HBase clusters** ADLS can be used as a default storage or additional storage—available only with HDI version 3.2

# Step 4(1): Specifying WASB for Storage



Microsoft Azure

New > Data + Analytics > New HDInsight Cluster > Data Source

New HDInsight Cluster

Data Source

The cluster will use this data source as the primary location for most data access, such as job input and log output.

\* Cluster Name  
samplehdicluster ✓  
.azurehdinsight.net

\* Subscription  
Pay-As-You-Go >

\* Select Cluster Type ⓘ  
Standard Hadoop on Linux (3.2) >

\* Credentials  
Configured >

\* Data Source ⓘ  
Configure required settings >

\* Node Pricing Tiers  
Please configure required settings 🔒

Selection Method ⓘ  
From all subscriptions ▼

\* Create a new storage account  
wasbstorage ✓  
[Select existing](#)

\* Choose Default Container ⓘ  
samplehdiclustercontainer ✓

\* Location  
East US >

Cluster AAD Identity ⓘ  
Not Configured >

Choose a storage account from all your subscriptions or specify the storage account name and access key

You can specify an existing storage account and container or have a new one created for you.



# Step 4(2): Specifying ADLS for Storage

**Step1:** Create a Service Principal (Azure Active Director ([AAD] Identity) that can represent the cluster

Data Source	Cluster AAD Identity	Create Service Principal
<p>The cluster will use this data source as the primary location for most data access, such as job input and log output.</p>	<p>This Azure Active Directory identity will represent the cluster. The cluster will use this identity to access your Data Lake Store accounts.</p>	<p>We will create a certificate, AD Application, and Service Principal for you. The certificate will be available once the cluster is provisioned.</p>
<p>Selection Method ⓘ</p> <p>From all subscriptions</p> <p>* Create a new storage account</p> <p>wasbstorage ✓</p> <p>Select existing</p> <p>* Choose Default Container ⓘ</p> <p>samplehdcclustercontainer ✓</p> <p>* Location</p> <p>East US</p> <p>Cluster AAD Identity ⓘ</p> <p>sampleprincipal1</p>	<p>Select AD Service Principal</p> <p>Use existing Create new</p> <p>* Service Principal ⓘ</p> <p>Not Configured</p> <p>Manage ADLS Access</p> <p>Service Principal Info:</p> <p>Keep this info if you want to recreate your cluster.</p> <p>Download Certificate</p> <p>Object ID:</p> <p>Not Configured</p> <p>Application ID:</p> <p>Not Configured</p>	<p>* Service Principal name</p> <p>sampleprincipal1 ✓</p> <p>Certificate start date</p> <p>2016-04-03</p> <p>Certificate expiration date</p> <p>2017-04-03</p> <p>* Certificate password</p> <p>..... ✓</p> <p>* Confirm password</p> <p>..... ✓</p> <p>Once you hit Create below, we will create a Azure AD Application and Service Principal on your behalf.</p>
Select	Select	Create

# Step 4(2): Specifying ADLS for Storage

**Step2:** Grant READ, WRITE and EXECUTE permissions to the Service Principal on the desired ADLS storage account.

Data Source

The cluster will use this data source as the primary location for most data access, such as job input and log output.

Selection Method ⓘ

From all subscriptions

\* Create a new storage account

wasbstorage

✓

Select existing

\* Choose Default Container ⓘ

samplehdiclustercontainer

✓

\* Location

East US

>

Cluster AAD Identity ⓘ

sampleprincipal1

>

Select

Cluster AAD Identity

This Azure Active Directory identity will represent the cluster. The cluster will use this identity to access your Data Lake Store accounts.

Select AD Service Principal

Use existing Create new

\* Service Principal ⓘ

sampleprincipal1

>

Upload Existing Certificate

Select a file

Certificate

Uploaded successfully

\* Certificate Password

••••••••

Manage ADLS Access

>

Service Principal Info:

Keep this info if you want to recreate your cluster.

Download Certificate

Select

Data Lake Store Root Folder Access

[Learn more](#)

i

Set the service principal's permissions on your Azure Data Lake Store (ADLS) accounts. You can change the root folder permissions of an ADLS account only if you are the owner of the root folder.

ADLS accounts whose root folders you own

DATA LAKE STORE	READ	WRITE	EXECUTE
adlsfordatagentesting	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
sampleadlsstorage1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
snapadlsforhdiinsight	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
snapadlsforhdiinsight2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Other ADLS accounts

DATA LAKE STORE	READ	WRITE	EXECUTE
tpchadls	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Save Permissions

# Step 5: Specify Cluster Configuration

Specify the number of worker nodes and the VM instance type for worker and head nodes

New HDInsight Cluster

\* Cluster Name

samplehdicluster

.azurehdinsight.net

\* Subscription

Pay-As-You-Go

\* Select Cluster Type

Standard Hadoop on Linux (3.2)

\* Credentials

Configured

\* Data Source

wasbstorage (East US)

\* Node Pricing Tiers

Please configure required settings

Optional Configuration

\* Resource Group

ADLA\_Benchmark

☐ Pin to dashboard

Create

Node Pricing Tiers

To learn more, visit our pricing page. [Learn more](#)

Number of Worker nodes

4

Worker Nodes Pricing Tier

D3 (4 nodes, 16 cores)

Head Node Pricing Tier

D3 (2 nodes, 8 cores)

WORKER NODES

0.62 x 4 = 2.49

HEAD NODES

0.62 x 2 = 1.24

TOTAL COST

3.73

USD/HOUR (ESTIMATED)

24 of 60 cores would be used in East US.

This price estimate does not include storage costs, network egress costs, or subscription discounts.

Questions? [Contact billing support.](#)

Note: Clusters with more than 32 Worker nodes require a Head node size with at least 8 cores and 14 GB RAM.

Select

Choose your pricing tier

Browse the available pricing tiers and their features. [Learn more](#)

Recommended | [View all](#)

D3 Optimized

4 Cores

14 GB RAM

8 Disks

200 GB Local SSD

0.62

USD/HOUR (ESTIMATED)

D4 Optimized

8 Cores

28 GB RAM

16 Disks

400 GB Local SSD

1.24

USD/HOUR (ESTIMATED)

D12 Optimized

4 Cores

28 GB RAM

8 Disks

200 GB Local SSD

0.76

USD/HOUR (ESTIMATED)

Select

# Step 6: Optional Configurations

Optionally you can configure:

- Virtual Network
- External Metastores
- Script Actions
- Linked Storage Accounts

The screenshot displays the 'Optional Configuration' page in the Microsoft Azure portal for a new HDInsight cluster. The breadcrumb navigation at the top reads: Microsoft Azure > New > Data + Analytics > New HDInsight Cluster > Optional Configuration. The left-hand navigation pane includes links for 'New', 'Resource groups', 'All resources', 'Recent', 'App Services', 'Virtual machines (classic)', 'Virtual machines', 'SQL databases', 'Cloud services (classic)', and 'Subscriptions'. The main content area is divided into two panels. The left panel, titled 'New HDInsight Cluster', contains the following configuration items: 'Cluster Name' (samplehdcluster, with a green checkmark and .azurehdinsight.net domain), 'Subscription' (Pay-As-You-Go), 'Select Cluster Type' (Standard Hadoop on Linux (3.2)), 'Credentials' (Configured), 'Data Source' (wasbstorage (East US 2)), 'Node Pricing Tiers' (D3/D3), 'Optional Configuration' (highlighted with a blue bar and a right arrow), and 'Resource Group' (ADLA\_Benchmark). At the bottom of this panel is a 'Pin to dashboard' checkbox and a 'Create' button. The right panel, titled 'Optional Configuration', lists four optional settings, all marked as 'Not Configured' with a right arrow: 'Virtual Network', 'External Metastores' (with a lock icon), 'Script Actions', and 'Linked Storage Accounts'. A 'Select' button is located at the bottom of this panel.

# Cluster Creation

Provisioning and configuring the cluster according to specification can take between 5 and 15 minutes.

The screenshot shows the HDInsight cluster management interface for a cluster named 'samplehdccluster'. The top navigation bar includes links for Settings, Dashboard, Secure Shell, Scale Cluster, Delete, and Move. A blue banner with an information icon and the text 'In Progress...' is prominently displayed. Below this, the 'Essentials' section provides key cluster details in a two-column layout:

Resource group	URL
<a href="#">ADLA_Benchmark</a>	<a href="https://samplehdccluster.azurehdinsight.net">samplehdccluster.azurehdinsight.net</a>
Status	Cluster Type
State Azure VM Configuration	Standard Hadoop on Linux
Location	Head Node, Worker Nodes
East US 2	D3 (x2), D3 (x4)
Subscription name	Learn more
<a href="#">Pay-As-You-Go</a>	<a href="#">Documentation</a>
Subscription id	Getting Started
bc2d3f0d-ae1d-4f7c-af1e-b2e6932bed5f	<a href="#">Quickstart</a>

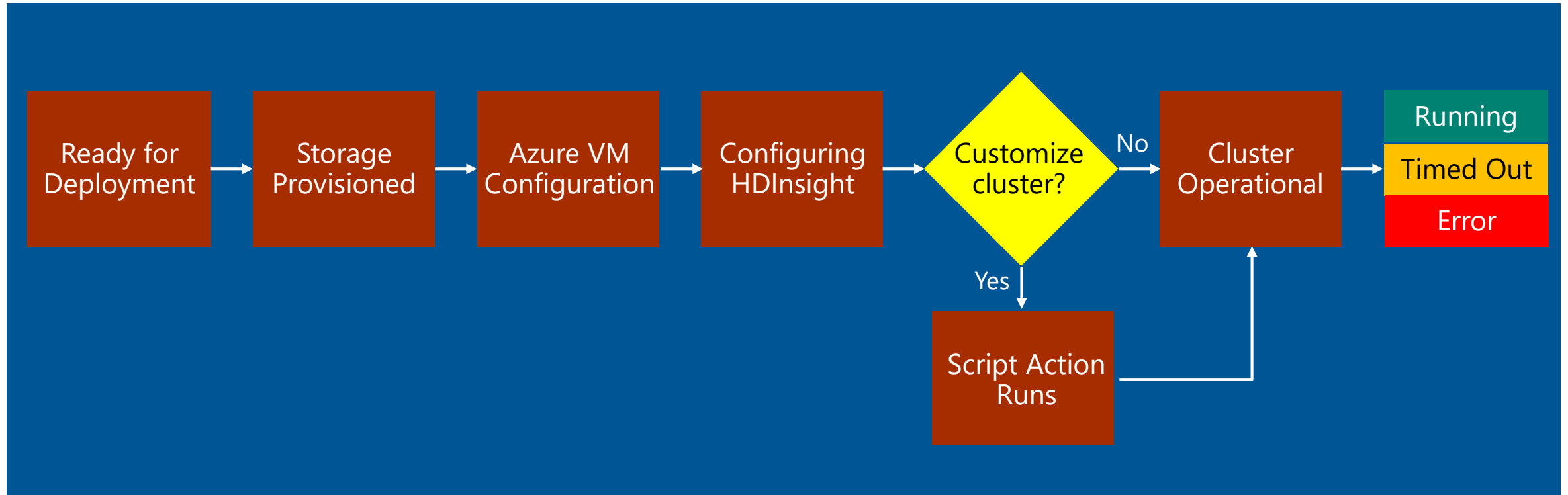
An 'All settings' button is located at the bottom right of the Essentials section. The right-hand sidebar contains a 'Settings' panel with the following categories and links:

- CONFIGURATION**
  - Cluster Login
  - Scale Cluster
  - Secure Shell
  - HDInsight Partner
  - External Metastores
- GENERAL**
  - Script Actions
  - Apps
- PROPERTIES**
  - Properties
  - Azure Storage Keys
  - Cluster AAD Identity
- RESOURCE MANAGEMENT**
  - Users
  - Tags

# Script Actions

# Customize with Script Actions

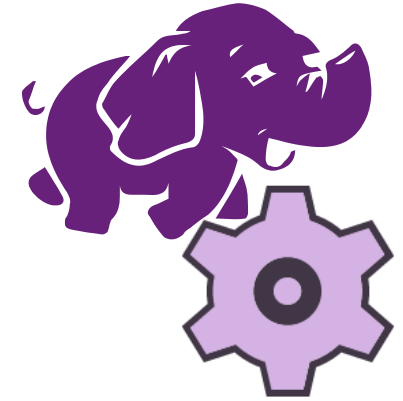
Script Actions enable clusters to be customized during creation using custom scripts: Clusters configuration can be changed or additional software installed.





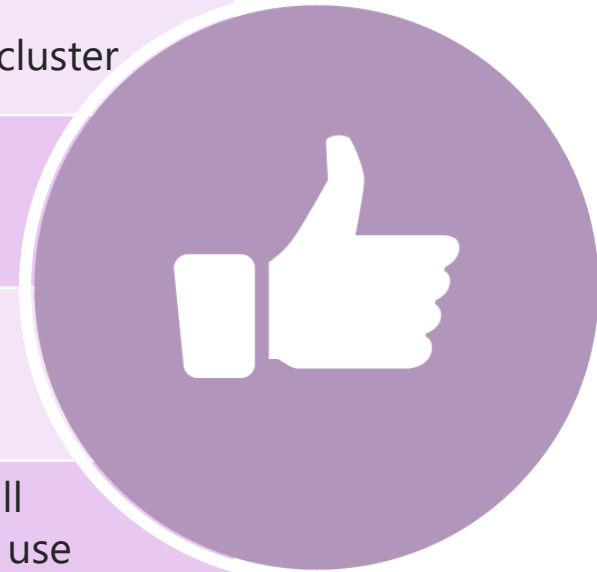
# Script Actions: Key concepts

- Script actions are Bash scripts that run when HDInsight is being configured.
- Scripts run in parallel on all the specified nodes in the cluster.
  - A script can be ran on the head nodes, the worker nodes, or both.
- Script actions must complete within 60 minutes, or they will timeout
- Each cluster can accept multiple script actions that are invoked in the order in which they are specified.
- **Script Action scripts can be used from:**
  - **The Azure Portal**
  - **Azure PowerShell**
  - **The HDInsight .NET SDK**



# Script Action: Best Practices

Target the right Hadoop version	Different versions of HDInsight have different versions of Hadoop services and components installed
Provide stable links to script resources:	All of the scripts and resources used by the script should remain available throughout the lifetime of the cluster
Use pre-compiled resources:	To minimize the time it takes to run the scripts
Ensure script idempotency	As nodes of an HDInsight cluster will be re-imaged during the cluster lifetime
Configure the custom components to use Azure Blob storage	On a cluster re-image, the HDFS file system gets formatted and all data that is stored there will be lost. Change the configuration to use Azure Blob storage (WASB) instead
Write information to STDOUT and STDERR	So the information is logged, and can be viewed after the cluster has been provisioned by using the Ambari web UI



# Provided Scripts

HDInsight provides Script Action scripts to install additional software

Software	Script
Hue	<a href="https://hdiconfigactions.blob.core.windows.net/linuxhueconfigactionv01/install-hue-uber-v01.sh">https://hdiconfigactions.blob.core.windows.net/linuxhueconfigactionv01/install-hue-uber-v01.sh</a> [See <a href="#">Install and use Hue on HDInsight clusters</a> ]
Spark	<a href="https://hdiconfigactions.blob.core.windows.net/linuxsparkconfigactionv02/spark-installer-v02.sh">https://hdiconfigactions.blob.core.windows.net/linuxsparkconfigactionv02/spark-installer-v02.sh</a> [See <a href="#">Install and use Spark on HDInsight clusters</a> ]
R	<a href="https://hdiconfigactions.blob.core.windows.net/linuxrconfigactionv01/r-installer-v01.sh">https://hdiconfigactions.blob.core.windows.net/linuxrconfigactionv01/r-installer-v01.sh</a> [See <a href="#">Install and use R on HDInsight clusters</a> ]
Solr	<a href="https://hdiconfigactions.blob.core.windows.net/linuxsolrconfigactionv01/solr-installer-v01.sh">https://hdiconfigactions.blob.core.windows.net/linuxsolrconfigactionv01/solr-installer-v01.sh</a> [See <a href="#">Install and use Solr on HDInsight clusters</a> ]
Giraph	<a href="https://hdiconfigactions.blob.core.windows.net/linuxgiraphconfigactionv01/giraph-installer-v01.sh">https://hdiconfigactions.blob.core.windows.net/linuxgiraphconfigactionv01/giraph-installer-v01.sh</a> [See <a href="#">Install and use Giraph on HDInsight clusters</a> ]
Hive libraries	<a href="https://hdiconfigactions.blob.core.windows.net/linuxsetupcustomhivelibsv01/setup-customhivelibsv01.sh">https://hdiconfigactions.blob.core.windows.net/linuxsetupcustomhivelibsv01/setup-customhivelibsv01.sh</a> [See <a href="#">Add Hive libraries on HDInsight clusters</a> ]

# Configuring Hadoop

# Core Hadoop Configuration files

Administrators configure settings for HDFS, Yarn and MapReduce (and other services) through these files

File Name	File Format	File Purpose
core-site.xml	Hadoop configuration XML	Hadoop core configuration settings that can be used by HDFS, YARN MapReduce and others
hdfs-site.xml	Hadoop configuration XML	HDFS configuration settings (NameNode and DataNode)
yarn-site.xml	Hadoop configuration XML	YARN configuration setting
Mapred-site.xml	Hadoop configuration XML	MapReduce configuration settings
Hadoop-env.sh	Bash script	Environment variables used by various Hadoop scripts and programs
log4j.properties	Java properties	System log file configuration settings
Hadoop-metrics2.properties	Java properties	Metrics publishing configuration settings.

Note: These files also define what should be recorded to the log files and how to process those log files.  
*Many of these settings can be configured using the Ambari Web UI (details in later slides)*

# Configuration Precedence

The actual configuration for any job running on a cluster is derived from a combination of sources including the default configuration, the per-cluster or per-node configuration, and the per-job configuration.

## Default Configuration

hadoop-common.jar  
hadoop-hdfs.jar  
Hadoop-mapreduce-client-core.jar  
Hadoop-yarn-common.jar

### JAR files contain (example)

Core-default.xml  
Hdfs-default.xml  
Mapred-default.xml  
Yarn-default.xml

Inherits  
from,  
extends,  
overrides



## Per-Cluster Configuration

Core-site.xml  
Hdfs-site.xml  
Mapred-site.xml  
Yarn-site.xml

Inherits  
from,  
extends,  
overrides



## Per-Job Configuration

#yarn jar -D prop=value  
...

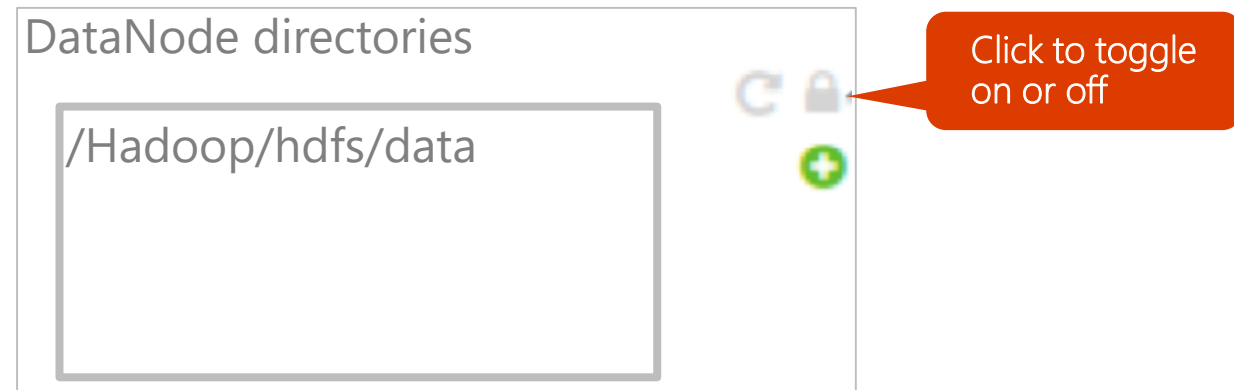
*Note: Cluster nodes with different hardware configurations commonly need different \*-site.xml files*

# Configuration: Final Properties

To prevent user applications from overriding a configuration property value, an administrator can declare the property value as ***final***.

- User applications may specify their own configuration settings when they are submitted to a cluster. In some cases, a user could choose a configuration setting that unfairly consumes a resource and negatively effects the performance other user applications.
- To prevent this, an administrator can declare a configuration property value as final. This prevents any user application from overriding a property's value.

Either the Ambari Web UI or a command-line editor can be used to make property settings final.



```
<property>
  <name> dfs.datanode.data.dir </name>
  <value> /hadoop/hdfs/data </value>
  <final> true </final>
</property>
```



# Configuration Management Options

Option	Description	Benefit
Ambari Web UI	Browser-based graphic user management interface	Ease of use, pre-built and ready-to-go
REST APIs: Ambari, WebHDFS, YARN etc	Use HTTP verbs (GET, PUT, POST, DELETE) management interface	Integration with other web-based management interfaces.
Manual Editing	Manually edit and distribute configuration files, manually restart services	No reliance on a GUI, no need to install Ambari. [Not compatible with Ambari management]
Command-line	Per-framework command-line management utilities	Scriptable, no reliance on a GUI

In an Ambari-managed cluster it is recommended to exclusively use Ambari—using other management method may cause conflicts.



# Monitoring and Managing Hadoop with Ambari Web UI

# Apache Ambari: What is it?

A 100% open source framework for provisioning, managing and monitoring Apache Hadoop clusters

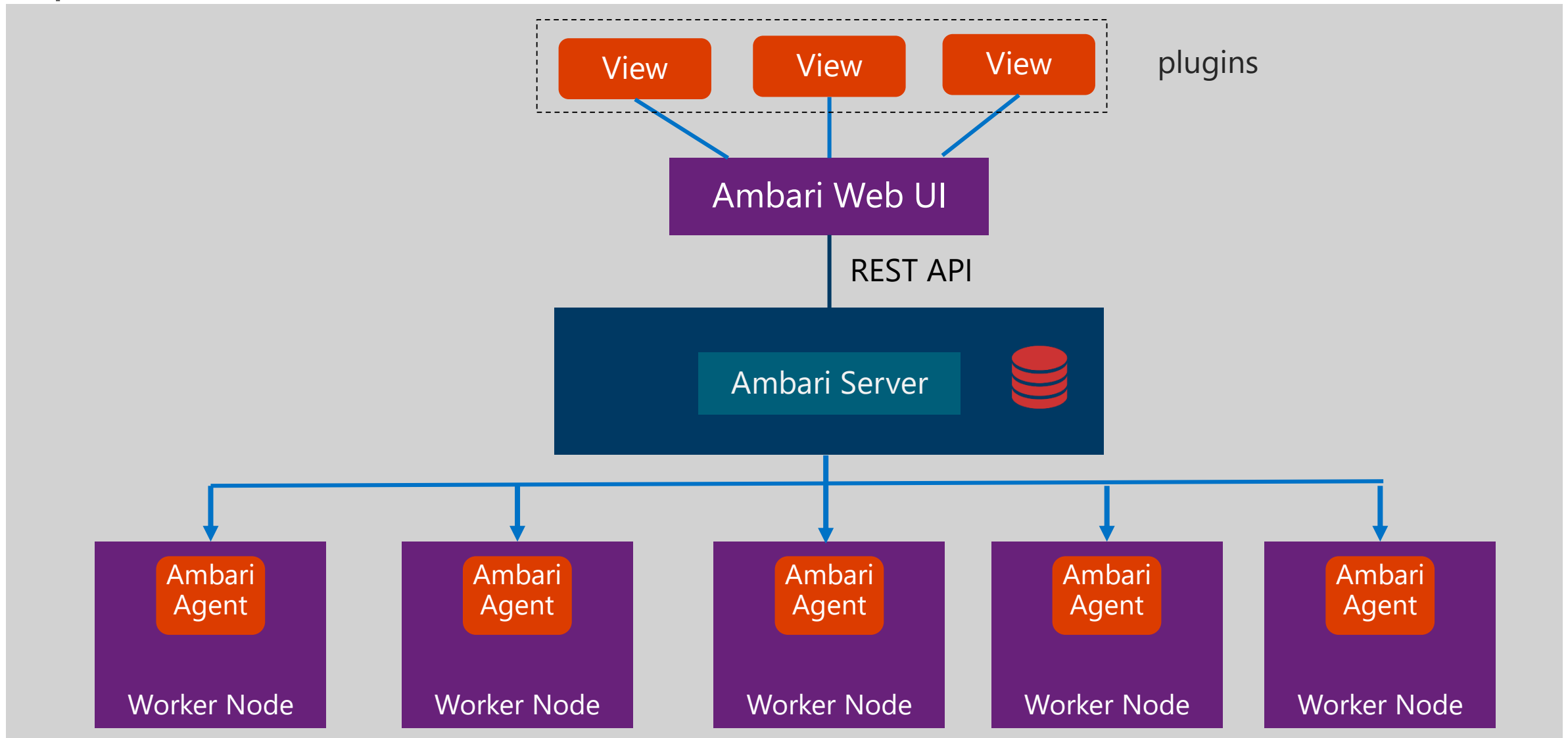
<b>Systems Administrators</b>	<b>Provisioning</b>	Provides step-by-step wizard for installing Hadoop services across any number of hosts
		Handles configuration of Hadoop services for the cluster
	<b>Managing</b>	Provides central management for starting, stopping, and reconfiguring Hadoop services across the entire cluster
	<b>Monitoring</b>	Provides dashboard for monitoring health and status of the Hadoop cluster
		Leverages <a href="#">Ambari Metrics System</a> for metrics collection
		Leverages <a href="#">Ambari Alert Framework</a> for system alerting and will notify you when your attention is needed (e.g., a node goes down, remaining disk space is low, etc)
<b>Application Developers and System Integrators</b>	Can easily integrate Hadoop provisioning, management, and monitoring capabilities to their own applications with the <a href="#">Ambari REST APIs</a> .	

# Ambari: Management Features Overview



- ❖ Interactive Wizard Driven cluster Installation
- ❖ Non-interactive API-driven cluster installation
- ❖ Granular control of cluster services start up and shut down
- ❖ Cluster service configuration management
- ❖ Dashboard cluster monitoring with alerts
- ❖ REST API for integration with other vendors
- ❖ Ambari Views for custom plug-in

# Apache Ambari: Architecture Overview



# Managing HDInsight with Amabari

# HDInsight and Ambari

The Ambari Web UI can be launched directly from the Azure HDInsight Portal

The screenshot displays the Azure HDInsight Portal interface for a cluster named 'ntlivefindemo0630'. The portal includes a top navigation bar with icons for Settings, Dashboard, Secure Shell, Scale Cluster, Delete, and Move. Below this, the 'Essentials' section provides details about the cluster, including its resource group, status (Running), location (South Central US), and subscription information. A 'Quick Links' section at the bottom offers shortcuts to 'Cluster Dashboards' and 'Ambari Views'. Overlaid on the right side of the portal is a Microsoft Edge browser window showing a login prompt. The prompt text reads: 'The server ntlivefindemo0630.azurehdinsight.net is asking for your user name and password. The server reports that it is from HDInsight.' Below the text is a user selection icon, a text input field containing 'admin', and a password input field represented by a series of dots. At the bottom right of the dialog are 'OK' and 'Cancel' buttons.

ntlivefindemo0630  
HDInsight Cluster

Settings Dashboard Secure Shell Scale Cluster Delete Move

Essentials ^

Resource group  
[ntfindemo0630](#)

Status  
Running

Location  
South Central US

Subscription name  
[Azure conversion](#)

Subscription id  
15c5cb6e-191a-40ea-9f69-08207a17fe97

URL  
[ntlivefindemo0630.a](#)

Cluster Type  
Standard Spark on L

Head Node, Worker Node  
D4 (x2), D4 (x2)

Learn more  
[Documentation](#)

Getting Started  
[Quickstart](#)

Quick Links

Cluster Dashboards

Ambari Views

Microsoft Edge

Microsoft Edge

The server ntlivefindemo0630.azurehdinsight.net is asking for your user name and password.  
The server reports that it is from HDInsight.

admin

OK Cancel



# Ambari Web UI Dashboard

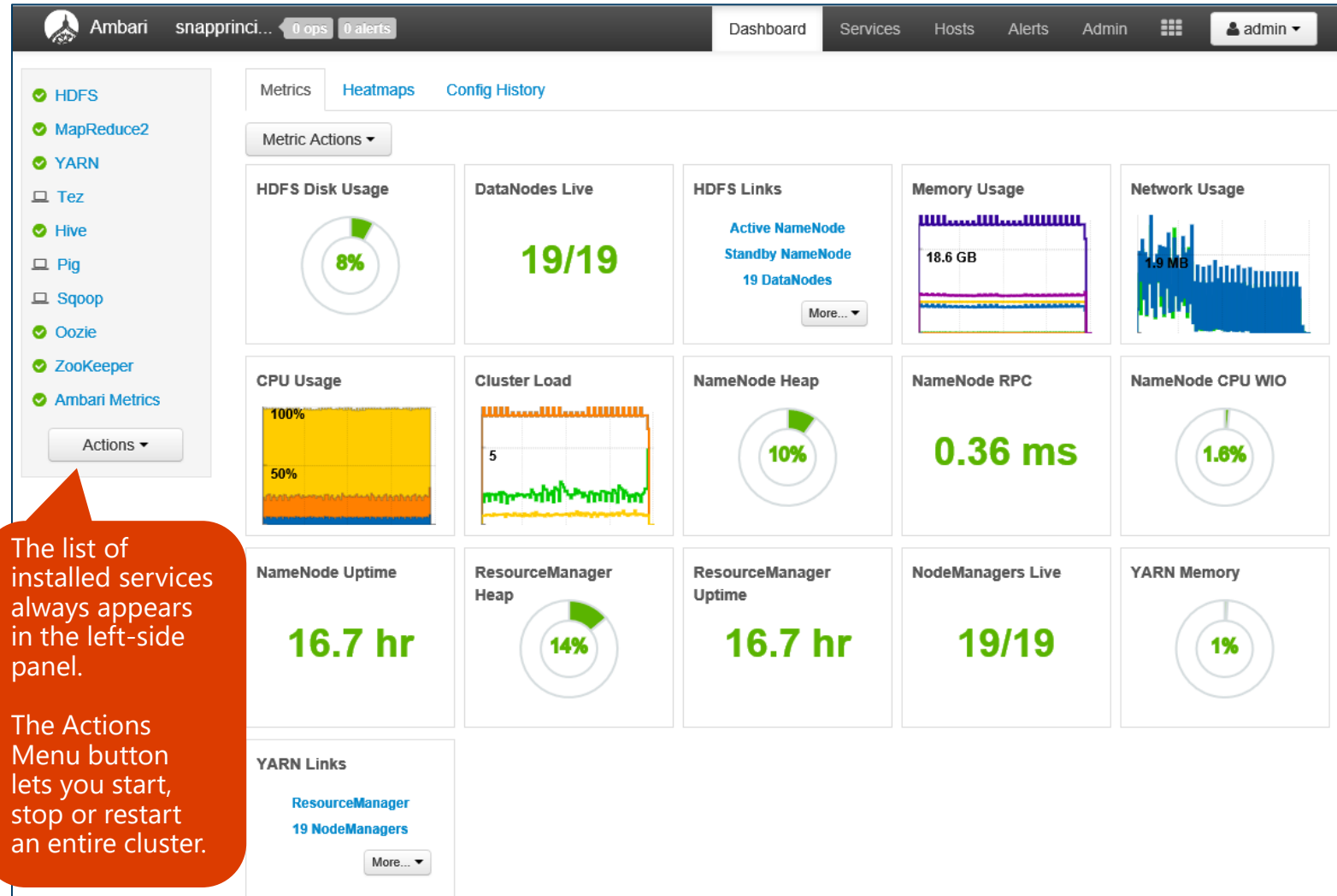
# Ambari Web UI: Dashboard Metrics

The Metrics tab of the Dashboard page displays cluster-level system metrics including:

- CPU Usage
- HDFS Disk Usage
- Memory Usage
- Network Usage
- ...

Dashboard enables you to understand the state of the cluster at-a-glance.

The dashboard look can be customized by adding and removing Widgets



# Ambari: Dashboard Metrics Drilldown

You can drilldown to get more details on

- CPU Usage
- Cluster Load
- Network Usage
- Memory Usage

The usage stats can be viewed over any custom period.

For other metrics you see additional info by hovering over the Widget.



# Ambari: Dashboard Widget Customization

For these Widgets:

- YARN Memory
- Node Managers
- Resource Managers
- NameNode CPU
- NameNode RPC
- NameNode Heap

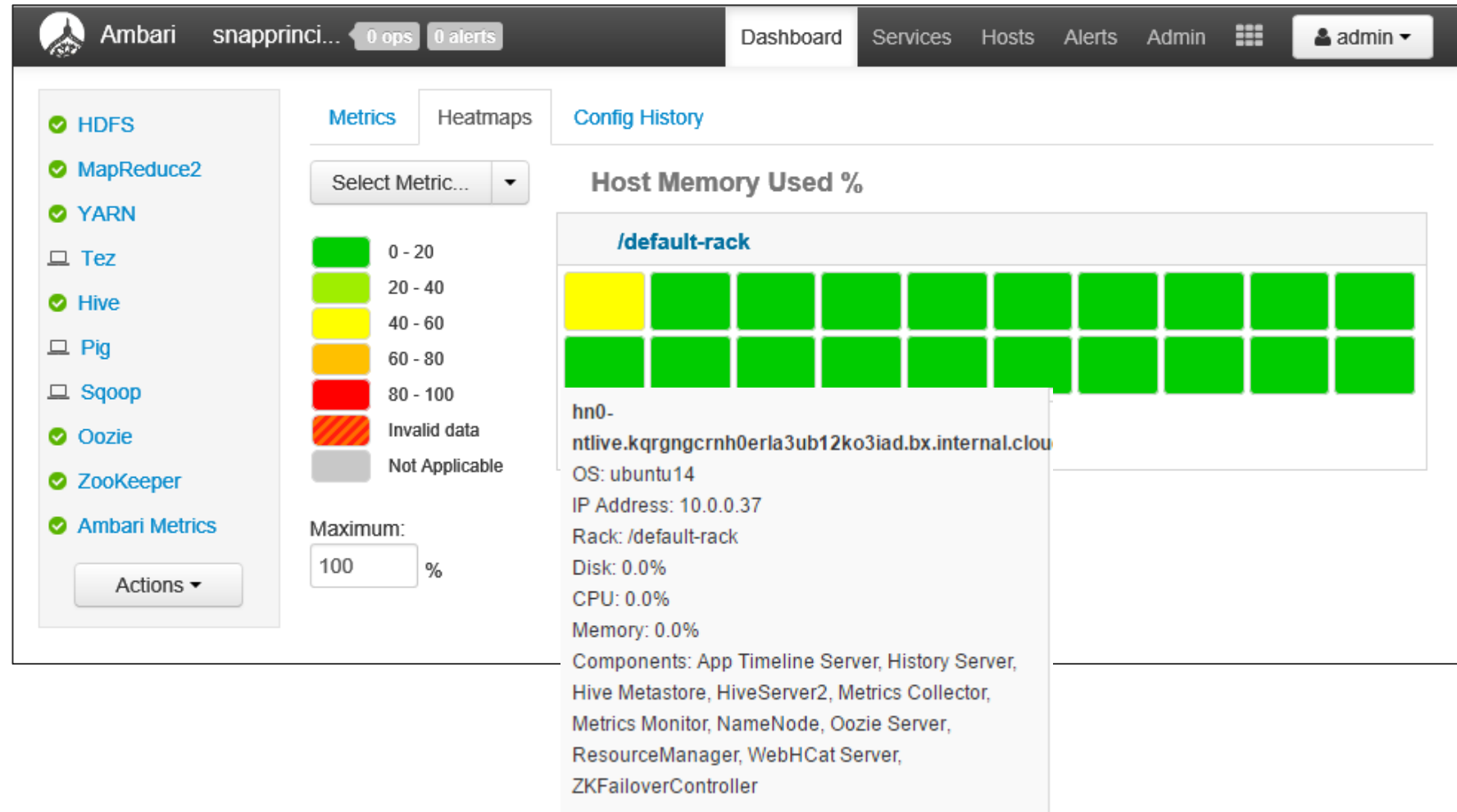
The color can be customized by configuring the % thresholds

The screenshot shows the Ambari dashboard interface. A modal dialog titled "Customize Widget" is open, allowing users to edit the percentage thresholds for a pie chart. The dialog includes a text box with the instruction: "Edit the percentage thresholds to change the color of current pie chart. Enter two numbers between 0 to 100". Below this, there is a horizontal slider bar with three segments: green (0-50%), orange (50-75%), and red (75-100%). The current values are 50 and 75, which are also displayed in input boxes. The background shows the Ambari dashboard with various widgets like "Network Usage" (4.7 MB), "NameNode CPU WIO" (n/a), and "YARN Memory".

# Ambari Dashboard: HeatMap

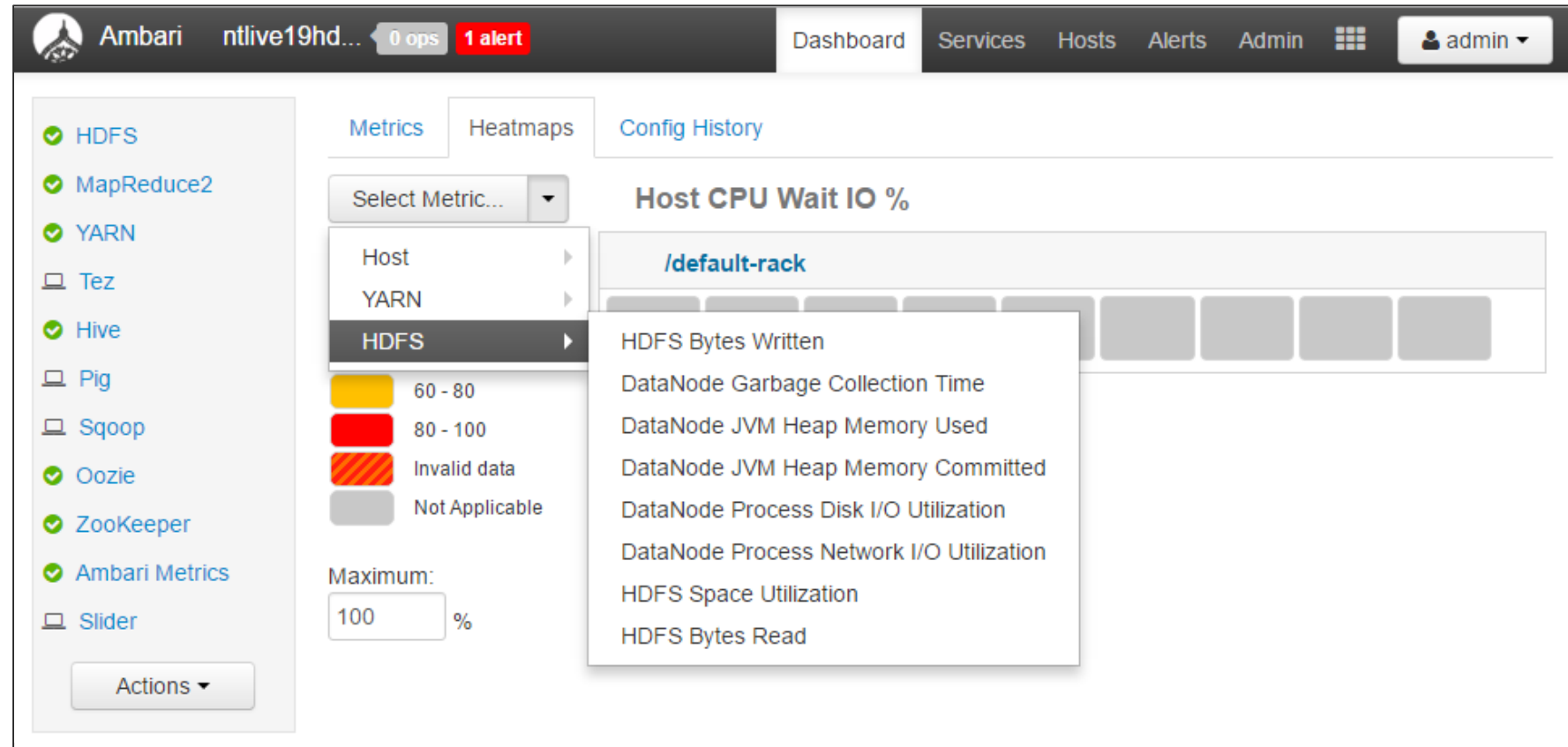
The Dashboard Heatmap view provides a color-coded view of each of the nodes in the cluster for selected metrics.

Hovering over each of the nodes pops ups additional information



# Ambari Dashboard: HeatMap

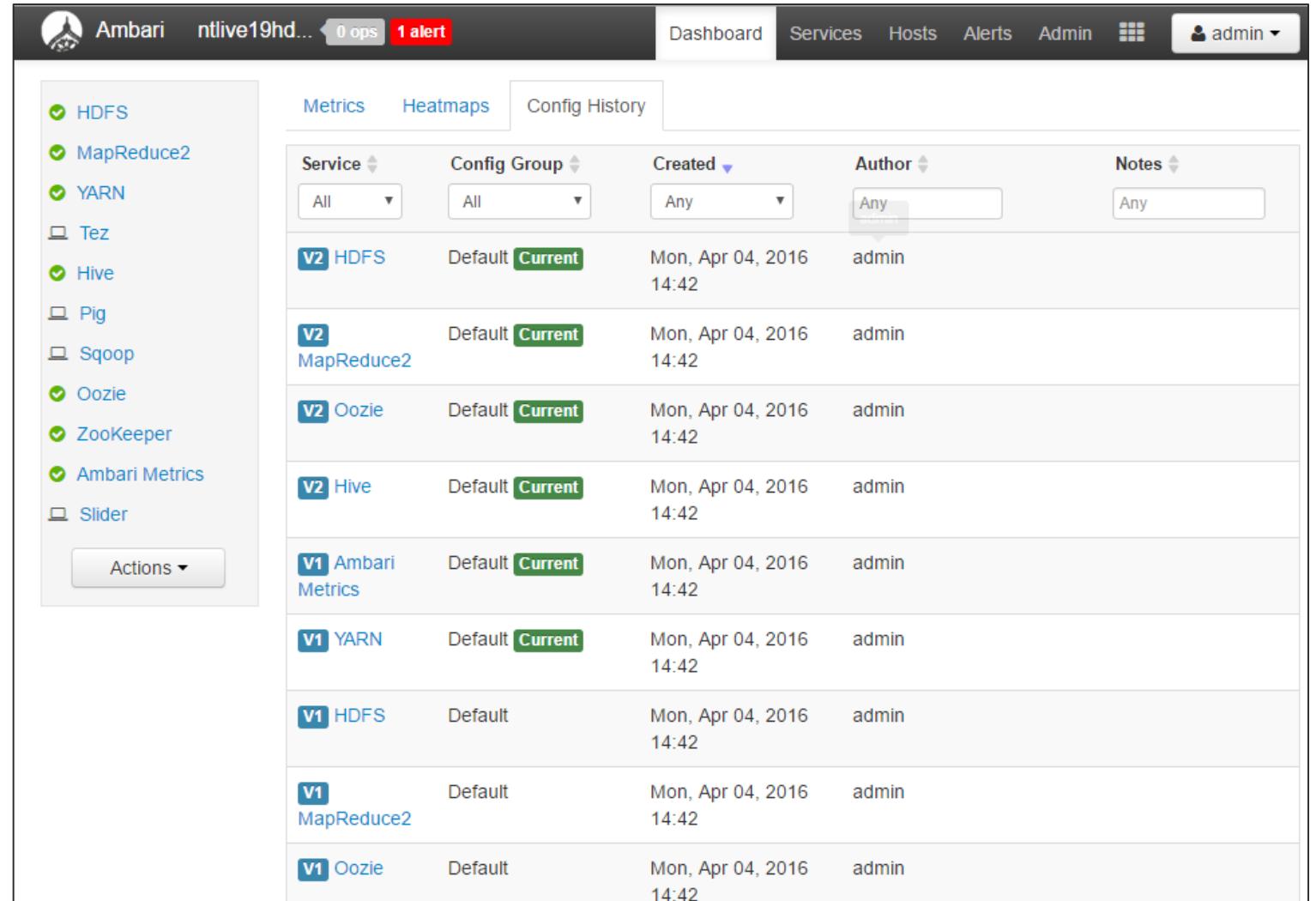
You can choose to show the Heatmap for 'Host', 'Yarn' and 'HDFS'. Each has an number of associated metrics for which the heatmap can be displayed.



# Ambari Dashboard: 'Config History'

The Dashboard 'Config History' view displays the list of the configuration changes made, along with 'when' and 'who' details.

Additional config history details can be seen by drilling down into the specific services—this can also be seen from the 'Services' view or by clicking on the Services links on the left of the page



Service	Config Group	Created	Author	Notes
V2 HDFS	Default <b>Current</b>	Mon, Apr 04, 2016 14:42	admin	
V2 MapReduce2	Default <b>Current</b>	Mon, Apr 04, 2016 14:42	admin	
V2 Oozie	Default <b>Current</b>	Mon, Apr 04, 2016 14:42	admin	
V2 Hive	Default <b>Current</b>	Mon, Apr 04, 2016 14:42	admin	
V1 Ambari Metrics	Default <b>Current</b>	Mon, Apr 04, 2016 14:42	admin	
V1 YARN	Default <b>Current</b>	Mon, Apr 04, 2016 14:42	admin	
V1 HDFS	Default	Mon, Apr 04, 2016 14:42	admin	
V1 MapReduce2	Default	Mon, Apr 04, 2016 14:42	admin	
V1 Oozie	Default	Mon, Apr 04, 2016 14:42	admin	

# Ambari UI: Alerts



# Ambari UI: Alerts

Ambari Web UI display any critical or Warning alerts at the top the page.

Clicking on the alert, pops up the list of alerts and current status

Clicking on 'Go to Alerts' definition display the complete list of alerts

The screenshot shows the Ambari Web UI interface. At the top, a navigation bar includes 'Ambari', a user profile 'admin', and tabs for 'Dashboard', 'Services', 'Hosts', 'Alerts', and 'Admin'. A red badge indicates '1 alert'. The 'Alerts' tab is active. A modal window titled '1 Critical or Warning Alerts' is displayed in the center. It contains a table with the following data:

Service / Host	Alert Definition Name	Status
Ambari	<a href="#">Ambari Server Alerts</a>	<b>CRIT</b> for 2 hours

Below the table, it states 'There are 15 stale alerts from 15 ho...'. A 'Go to Alerts Definitions' button is highlighted with a red box. The background shows a list of alert definitions with columns for 'Alert Definition Name', 'Status', and 'State'. The first few entries are 'Secondary NameNode', 'NFS Gateway Process', 'NameNode Last Check', and 'HDFS Upgrade Finalize', all with 'OK' status and 'Enabled' state. The bottom of the page shows '50 of 50 definitions showing - clear filters' and pagination controls.

# Ambari: List of Alerts

Ambari alerts are classified as:

- Critical
- OK
- Unknown
- None

You can drill down into specific alerts for details.

You can set the 'Check Interval' for the alerts by editing the alert.

The screenshot shows the Ambari web interface. The top navigation bar includes 'Dashboard', 'Services', 'Hosts', 'Alerts' (selected), and 'Admin'. A user dropdown shows 'admin'. A notification bar indicates '0 ops' and '1 alert'. The main content area is titled 'Host Disk Usage' with a 'Back' link and an 'OK (24)' status badge. The 'Configuration' section has an 'Edit' button highlighted with a red box. The description states: 'This host-level alert is triggered if the amount of disk space used goes above specific thresholds. The default threshold values are 50% for WARNING and 80% for CRITICAL'. The 'Check Interval' is set to '1 Minute'. On the right, the alert details are: State: Enabled, Service: Ambari, Component: Ambari Agent, Type: SCRIPT, Groups: AMBARI Default, Last: Mon, Apr 04, 2016, Changed: 21:49. The 'Instances' section contains a table with 4 columns: Service / Host, Status, 24-Hour, and Response.

Service / Host	Status	24-Hour	Response
Ambari / hn0- ntlive.kqrgngcrnh0erla3ub12ko3iad.bx.internal.cloudapp.net	OK for 9 hours	1	Capacity Used: [1.07%, 11.3 GB], Capacity Total: [1.1 T...
Ambari / hn1- ntlive.kqrgngcrnh0erla3ub12ko3iad.bx.internal.cloudapp.net	OK for 9 hours	1	Capacity Used: [1.11%, 11.8 GB], Capacity Total: [1.1 T...
Ambari / wn0- ntlive.kqrgngcrnh0erla3ub12ko3iad.bx.internal.cloudapp.net	OK for 2 hours	2	Capacity Used: [1.39%, 14.7 GB], Capacity Total: [1.1 ...

# Ambari UI: Services

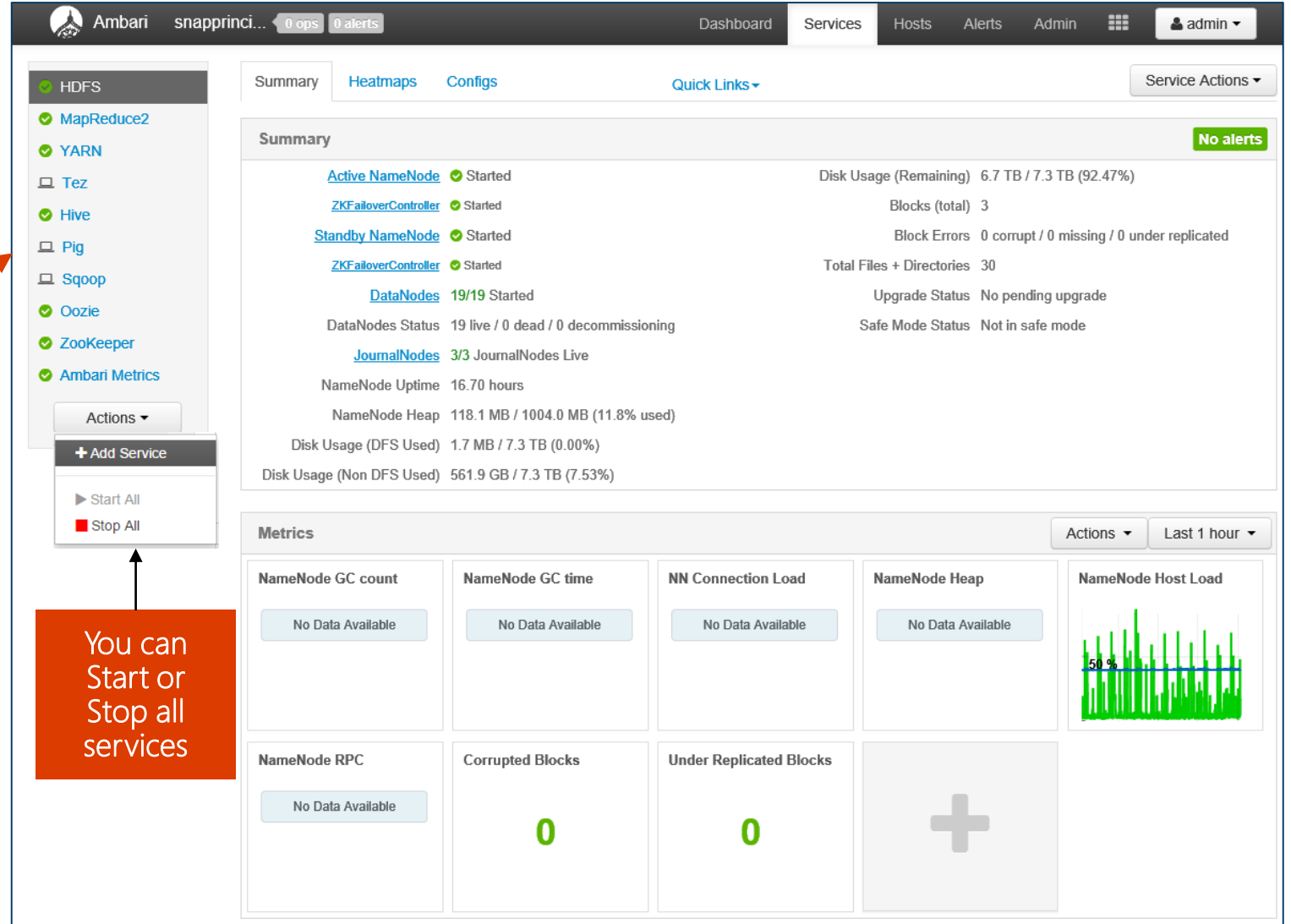
# Ambari Web UI: Services

The **Services** page provides quick insights into the status of the services running on the cluster.

In this case the list of services running on the cluster include: **HDFS, MapReduce2, YARN, Hive, Oozie and Zookeeper**

Icons indicate status or actions that should be taken.

Shown here the details for HDFS for the last 1 hour.



# Ambari Web UI: Service Actions

For each service there are a list of associated “**Service Actions**” to manage, monitor and configure the service.

As the Service Actions menu button is context-sensitive, the menu choices are different for each service.

The Service Actions for HDFS are shown here.

**Maintenance Mode** should be enabled when making cluster hardware or software changes. It suppresses Ambari alerts, warnings and status change indicators

The screenshot displays the Ambari Web UI interface. The top navigation bar includes the Ambari logo, the username 'snapprinci...', and status indicators for '0 ops' and '0 alerts'. The main navigation tabs are 'Dashboard', 'Services', 'Hosts', 'Alerts', and 'Admin'. The 'Services' tab is active, showing a list of services on the left: HDFS (selected), MapReduce2, YARN, Tez, Hive, Pig, Sqoop, Oozie, ZooKeeper, and Ambari Metrics. The 'Actions' button is visible below the service list. The main content area shows the 'Summary' tab for the HDFS service. It includes a 'Summary' section with various status indicators: 'Active NameNode' (Started), 'ZKFailoverController' (Started), 'Standby NameNode' (Started), 'ZKFailoverController' (Started), 'DataNodes' (19/19 Started), 'DataNodes Status' (19 live / 0 dead / 0 decommissioning), 'JournalNodes' (3/3 JournalNodes Live), 'NameNode Uptime' (16.70 hours), 'NameNode Heap' (118.1 MB / 1004.0 MB (11.8% used)), 'Disk Usage (DFS Used)' (1.7 MB / 7.3 TB (0.00%)), and 'Disk Usage (Non DFS Used)' (561.9 GB / 7.3 TB (7.53%)). Below the summary is a 'Metrics' section with several widgets: 'NameNode GC count', 'NameNode GC time', 'NN Connection Load', 'NameNode RPC', 'Corrupted Blocks' (0), and 'Under Replicated Blocks' (0). A 'Service Actions' dropdown menu is open on the right, listing various actions: 'Start', 'Stop', 'Restart All', 'Restart DataNodes', 'Restart JournalNodes', 'Restart ZKFailoverControllers', 'Move NameNode', 'Run Service Check', 'Turn On Maintenance Mode' (highlighted with a red box), 'Rebalance HDFS', and 'Download Client Configs'. A small line graph is visible in the bottom right corner of the metrics section.

# Ambari Web UI: Services (YARN)

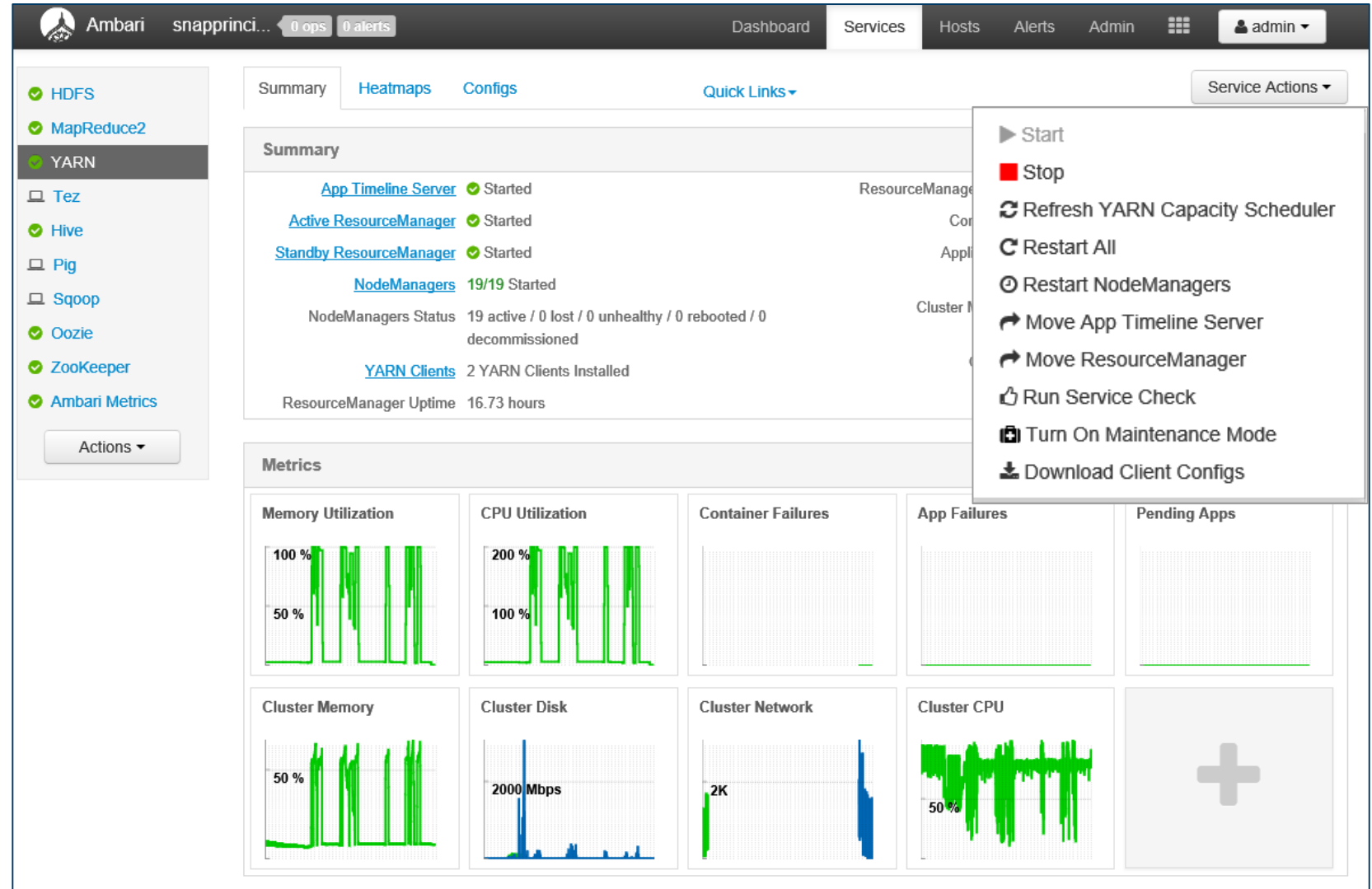
The **Services** page details  
for YARN for the last 1 hour

... and the last 12 hours



# Ambari: YARN "Service Actions"

Here are the list of actions that can be taken with YARN.



The screenshot shows the Ambari web interface with the YARN service selected. The 'Service Actions' dropdown menu is open, displaying the following options:

- ▶ Start
- Stop
- ↻ Refresh YARN Capacity Scheduler
- ↻ Restart All
- ⌂ Restart NodeManagers
- ↻ Move App Timeline Server
- ↻ Move ResourceManager
- 🔍 Run Service Check
- 🛑 Turn On Maintenance Mode
- 📄 Download Client Configs

The background interface shows the YARN service status as 'Started' and includes a 'Metrics' section with charts for Memory Utilization, CPU Utilization, Container Failures, App Failures, Pending Apps, Cluster Memory, Cluster Disk, Cluster Network, and Cluster CPU.

# Ambari Web UI: Hive Service Actions

This is the Hive Services page with the list of associated actions.

The screenshot displays the Ambari Web UI interface. The top navigation bar includes the Ambari logo, the cluster name 'n1live19hd...', and status indicators for '0 ops' and '1 alert'. Navigation tabs for 'Dashboard', 'Services', 'Hosts', 'Alerts', and 'Admin' are present, along with a user profile dropdown for 'admin'. On the left sidebar, a list of services is shown with 'Hive' selected. The main content area is divided into 'Summary' and 'Configs' tabs, with 'Summary' active. It lists the status of various Hive components: 'Hive Metastore' (Started), 'HiveServer2' (Started), 'WebHCat Server' (Started), 'HCat Client' (1 installed), and 'Hive Clients' (6 installed). A 'Service Actions' dropdown menu is open on the right, listing actions such as Start, Stop, Restart All, Move Hive Metastore, Move HiveServer2, Move WebHCat Server, Run Service Check, Turn On Maintenance Mode, Add Hive Metastore, Add HiveServer2, and Download Client Configs.

Ambari n1live19hd... 0 ops 1 alert Dashboard Services Hosts Alerts Admin admin

Summary Configs Service Actions

Summary

- [Hive Metastore](#) ✓ Started
- [Hive Metastore](#) ✓ Started
- [HiveServer2](#) ✓ Started
- [HiveServer2](#) ✓ Started
- [WebHCat Server](#) ✓ Started
- [WebHCat Server](#) ✓ Started
- [HCat Client](#) 1 HCat Client Installed
- [Hive Clients](#) 6 Hive Clients Installed

Service Actions

- ▶ Start
- Stop
- ↻ Restart All
- ↻ Move Hive Metastore
- ↻ Move HiveServer2
- ↻ Move WebHCat Server
- 🔍 Run Service Check
- 🔧 Turn On Maintenance Mode
- ➕ Add Hive Metastore
- ➕ Add HiveServer2
- 📄 Download Client Configs ▶



# Ambari Web UI: Hosts

# Ambari Web UI: Hosts

The Hosts page provides system-level metrics for each node in the cluster including.

Clicking on the components link, provides more details on the list of components running on the node.

The screenshot shows the Ambari Web UI interface. The top navigation bar includes links for Dashboard, Services, Hosts (selected), Alerts, and Admin. The user is logged in as 'admin'. The main content area displays a table of hosts with columns for Name, IP Address, Rack, Cores, RAM, Disk Usage, Load Avg, Versions, and Components. A modal window titled 'Components' is open, showing the components for the host 'hn0-snappr.rvofooxckjyu3c1i1gfwalgq0c.cx.internal.cloudapp.net'. The components listed are History Server, Hive Client, Hive Metastore, HiveServer2, MapReduce2 Client, Metrics Collector, and Metrics Monitor. The modal also has an 'OK' button.

Name	IP Address	Rack	Cores	RAM	Disk Usage	Load Avg	Versions	Components
Any	Any	Any	Any	Any	Any	Any	Filter	Filter
<input type="checkbox"/> <input checked="" type="checkbox"/> hn0-snappr.rvofooxckjyu3...								20 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> hn1-snappr.rvofooxckjyu3...								14 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn0-snappr.rvofooxckjyu...								7 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn1-snappr.rvofooxckjyu...								7 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn10-snappr.rvofooxckjy...								7 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn11-snappr.rvofooxckjy...								7 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn12-snappr.rvofooxckjy...								7 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn13-snappr.rvofooxckjy...								7 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn14-snappr.rvofooxckjy...								7 Components
<input type="checkbox"/> <input checked="" type="checkbox"/> wn15-snappr.rvofooxckjy...								7 Components

10 of 24 hosts showing - clear filters

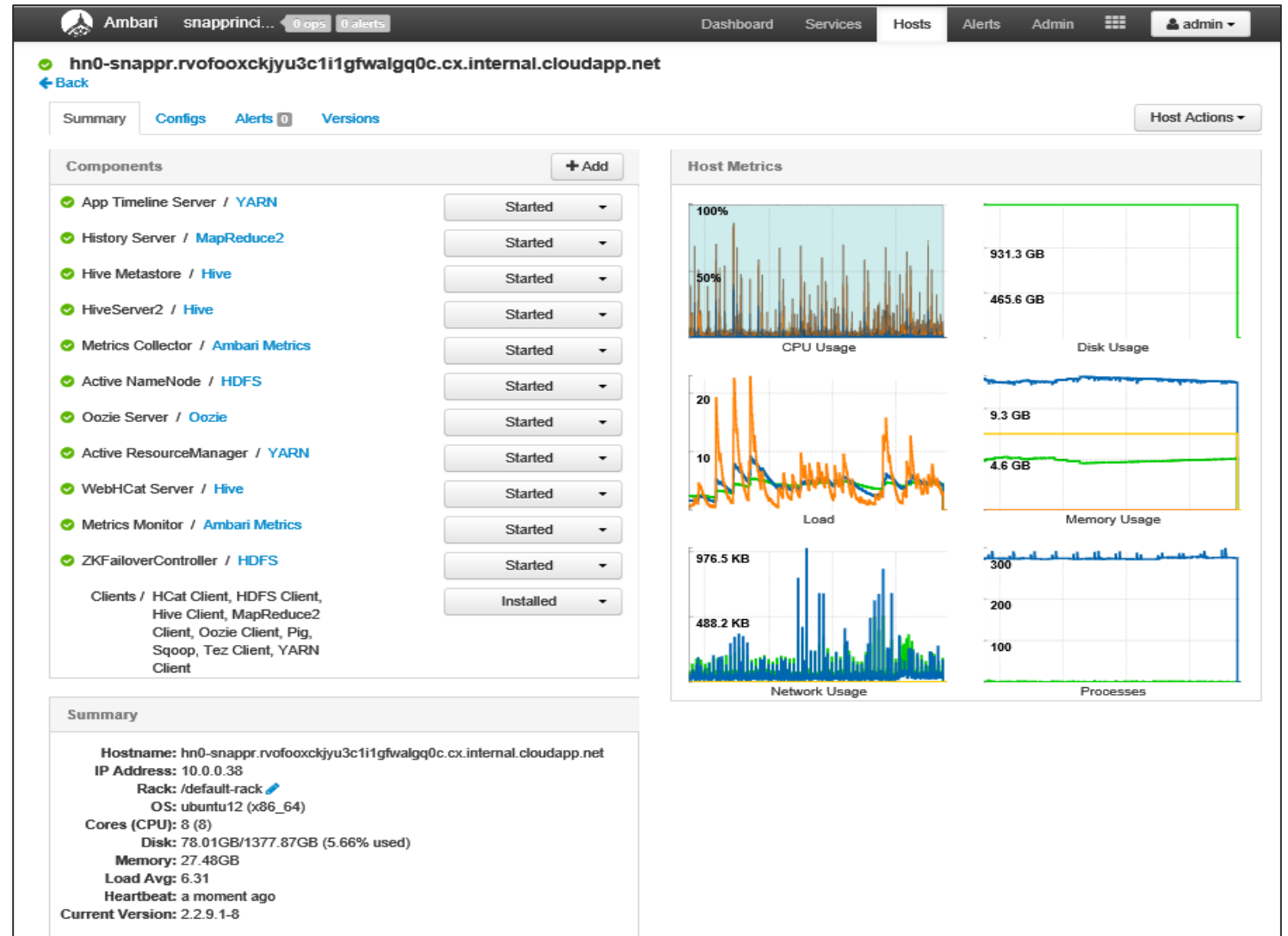
# Amber Web UI: Hosts Drilldown

You can drilldown into the details of any of the nodes in the cluster.

At a glance you can see the charts for CPU, Memory and Network usage.

The summary system-configuration information is also displayed.

You can see—and change—the status of each of the components running on the node.



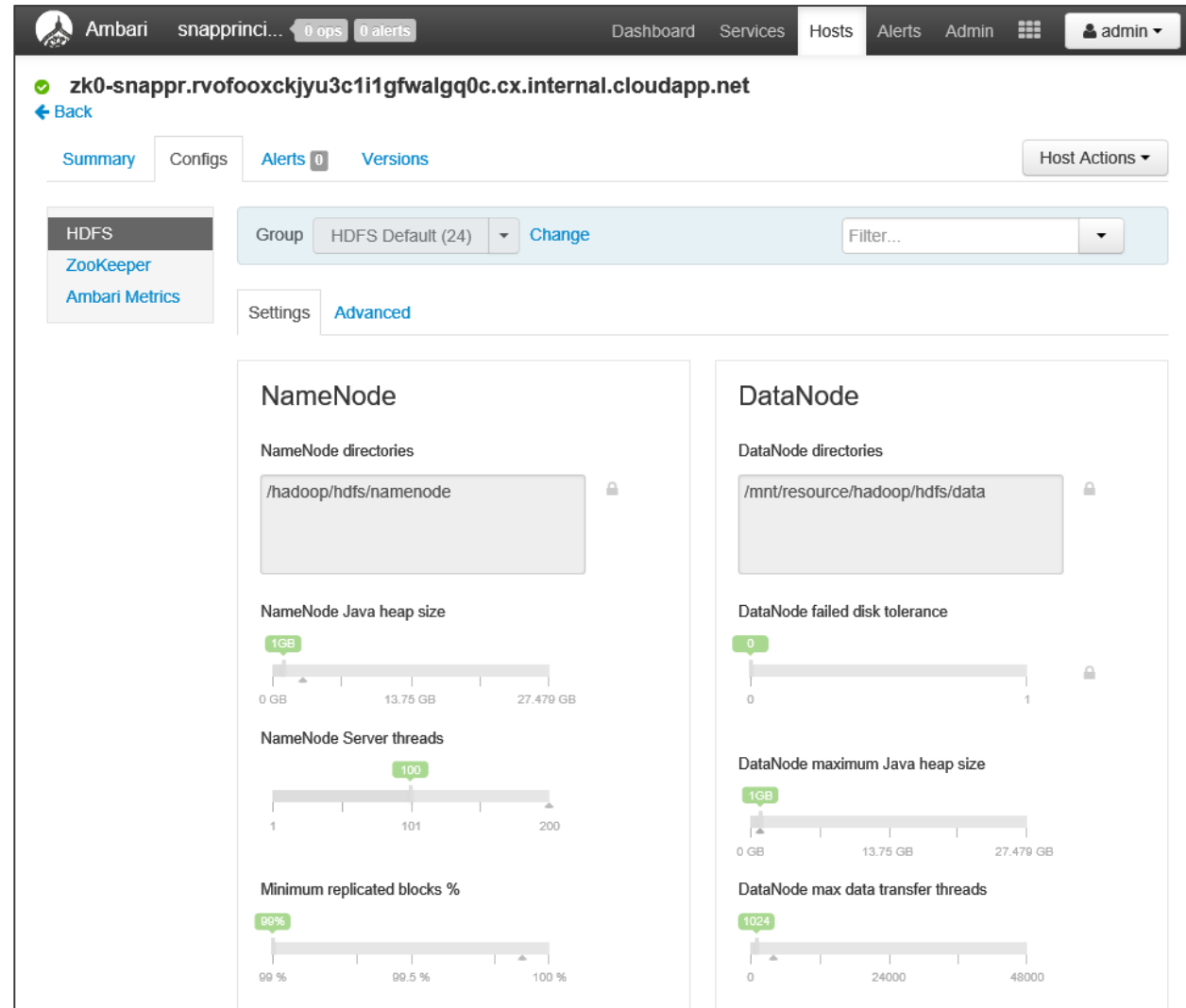
# Ambari Web UI: Host Component Drilldown

On the hosts page you can drill down into the details about any of the components running on the node.

This shows key metrics about the **HDFS component** running on the this node.

You can **configure HDFS parameters** such as:

- NameNode Java heap size
- NameNode Server threads
- Minimum replicated blocks %
- DataNode failed disk tolerance
- DataNode max Java heap size
- DataNode max data transfer threads



# Ambari Web UI: User Views

# Ambari: Capacity Scheduler View

The [YARN Capacity Scheduler](#) allows Hadoop to be shared among multiple independent tenants while providing guaranteed capacity and predictable SLAs.

The Capacity Scheduler divides resources through use of **YARN queues**, which are sized based on the relative allocations given to various tenants.

The **Capacity Scheduler View** lets you create and modify **YARN queues** and see their distribution at-a-glance.

The UI enforces configuration rules, highlights invalid conditions.

With the Capacity Scheduler View you can:

- Partition Hadoop resources among tenants.
- Define, view and modify queue definitions.
- Establish fine-grained control on who can run jobs in queues.


The screenshot displays the Ambari web interface for the Capacity Scheduler. The top navigation bar includes the Ambari logo, user 'snapprinci...', and status indicators for '0 ops' and '0 alerts'. The main content area is divided into several sections:

- Queue List:** A table showing three queues: 'root (100%)', 'default (95%)', and 'joblauncher (5%)'. The 'default' queue is highlighted in blue and has a green checkmark.
- Scheduler:** A section with a green checkmark, containing configuration fields for 'Maximum Applications' (10000), 'Maximum AM Resource' (33 %), 'Node Locality Delay' (0), 'Calculator' (org.apache.hadoop.yarn.util.resourc), 'Queue Mappings', and 'Queue Mappings Override' (Disabled).
- default Queue Details:** A detailed view for the 'default' queue, showing 'Capacity' (95 %) and 'Max Capacity' (100 %) with sliders. It also includes an 'Enable node labels' checkbox and a 'Show Peer Level Queues' link.
- Access Control and Status:** A section with 'State' (Running/Stopped), 'Administer Queue' (Anyone/Custom), and 'Submit Applications' (Anyone/Custom).
- Resources:** A section with configuration fields for 'User Limit Factor' (10), 'Minimum User Limit' (100 %), 'Maximum Applications' (Inherited), 'Maximum AM Resource' (Inhe %), and 'Ordering policy' (dropdown).
- Versions:** A section at the bottom showing 'v1' as the 'Current' version, with 'INITIAL' and 'load' buttons.

# Ambari: Tez View

The Tez View lists all the DAGs (currently executing and historical) over a time period.

You can drill down into specific DAG to see more details

 Ambari snapprinci... 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin

admin

All DAGs

Last refreshed at 28 Mar 2016 10:12:49 Refresh

Dag Name Id Submitter Status Application ID

Search... Search... Search... All Search...

First 1 2 Rows 10

Settings

Dag Name	Id	Submitter	Status	Start Time	End Time	Duration	Applicat
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	🔄 RUNNING...	28 Mar 2016 08:27:10	Not Available!	Not Available!	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	❌ FAILED	28 Mar 2016 08:16:49	28 Mar 2016 08:16:49	330 ms	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 08:02:15	28 Mar 2016 08:15:39	803 secs	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 07:53:52	28 Mar 2016 08:01:23	451 secs	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 07:52:55	28 Mar 2016 07:53:15	20 secs	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 07:38:38	28 Mar 2016 07:52:13	814 secs	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 06:14:06	28 Mar 2016 07:29:07	1.25 hours	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 04:05:35	28 Mar 2016 06:09:55	2.07 hours	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 03:43:04	28 Mar 2016 04:04:54	21.83 mins	<a href="#">applicatio</a>
<a href="#">snapssh_201603281...</a>	dag_1459125085625...	snapssh	✅ SUCCEEDED	28 Mar 2016 03:12:41	28 Mar 2016 03:41:54	29.22 mins	<a href="#">applicatio</a>

# Ambari: Tez View (DAG Details)

The Graphical View lets you visualize the DAG execution flow graphically. You can get more details about any vertex by clicking on it.

The screenshot shows the Ambari interface for viewing Tez DAG details. The top navigation bar includes 'Ambari', 'snapprinci...', '0 ops', '0 alerts', and links to 'Dashboard', 'Services', 'Hosts', 'Alerts', and 'Admin'. The user is logged in as 'admin'. The breadcrumb path is 'All DAGs / DAG [ snapssh\_20160328150202\_0b6115a6-c3f8-44bb-998d-6e30fa54b7c4:1 ]'. The 'Graphical View' tab is selected, showing a DAG flow: 'lineitem' (green rounded rectangle) → '929 1' (blue rounded rectangle) → 'R 169 cer 2' (blue rounded rectangle) → 'out\_Reducer..' (red rounded rectangle). A tooltip for 'Reducer 2' is displayed over the 'R 169 cer 2' vertex, showing the following details:

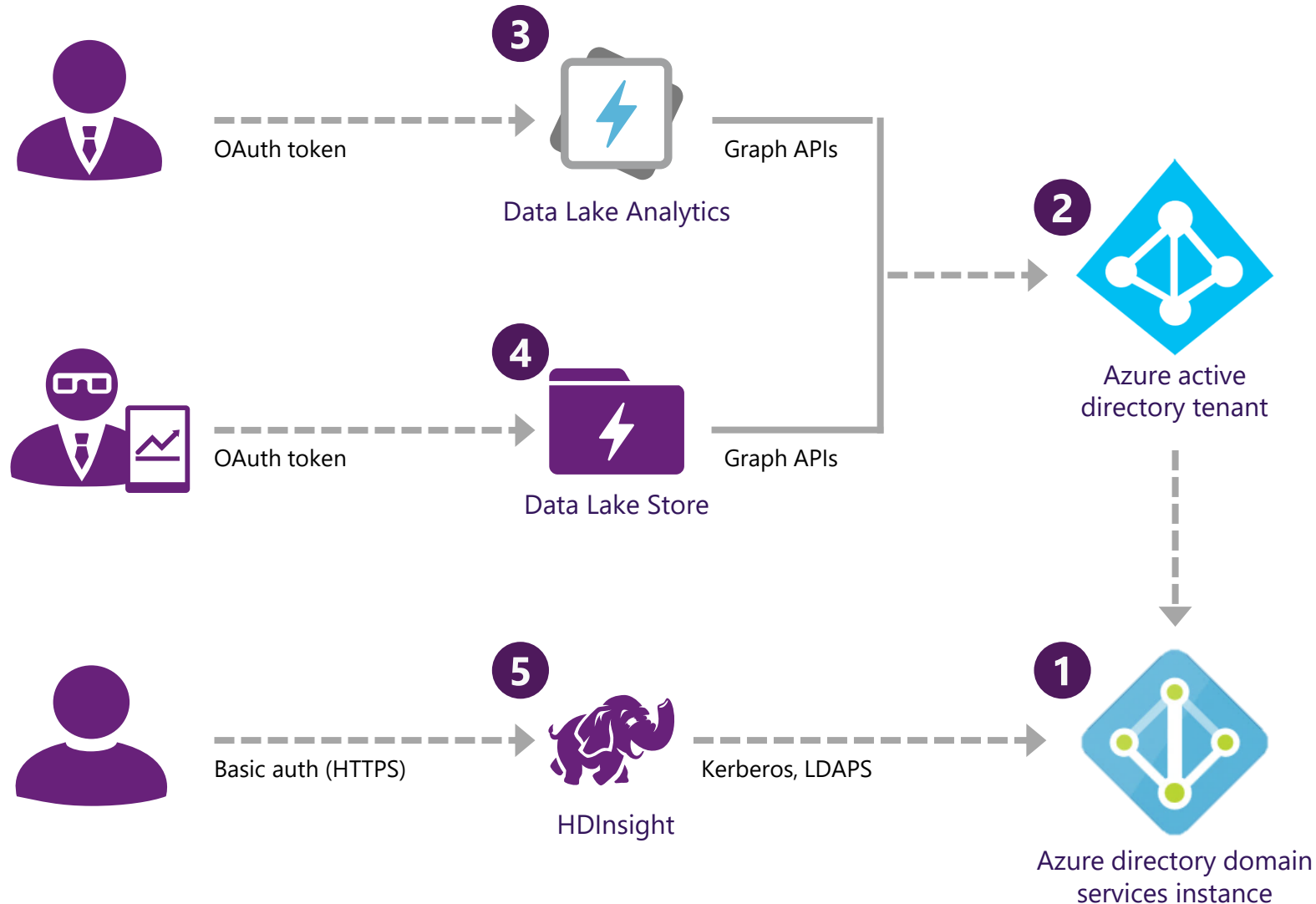
Reducer 2	
Vertex Name	Reducer 2
Vertex ID	vertex_1459125085625_0024_1_01
Progress	1
Start Time	28 Mar 2016 08:02:17
End Time	28 Mar 2016 08:15:39
Duration	801 secs
First Task Start Time	28 Mar 2016 08:10:17
Tasks	169
Processor Class	ReduceTezProcessor



# HDInsight on Linux: Security Overview

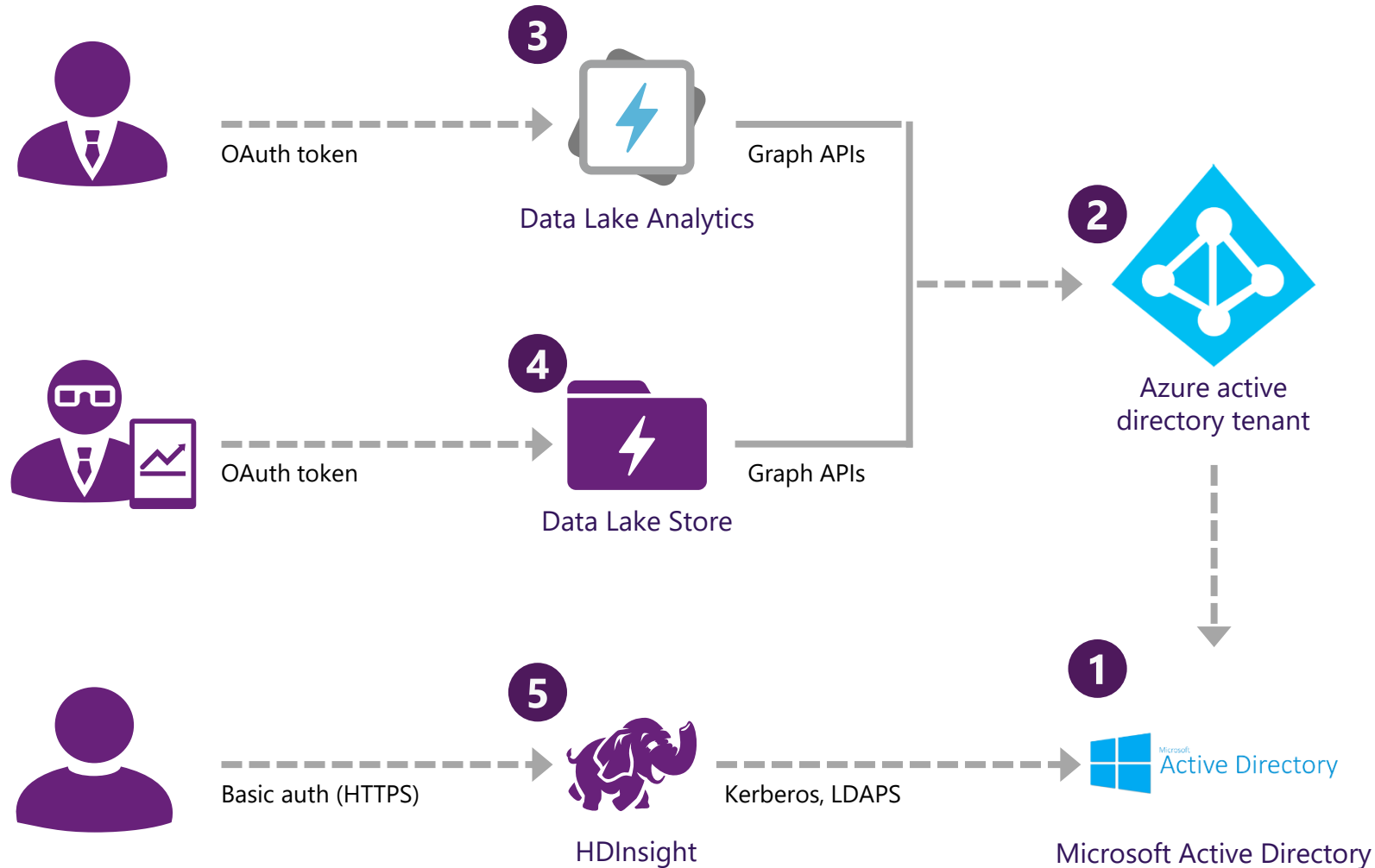


# Leveraging Azure Active Directory



- 1 Create ADDS instance in separate VNET
- 2 Add users to AAD Tenant
- 3 Add users to ADLA RBAC roles
- 4 Add users to ADLS RBAC roles & file system ACLs
- 5 Join HDInsight cluster to ADDS instance

# Incorporating Enterprise Active Directory



- 1 Add users to AD domain
- 2 Sync user info from enterprise AD to AAD
- 3 Add users to ADLA RBAC roles
- 4 Add users to ADLS RBAC roles & file system ACLs
- 5 Join HDInsight cluster to enterprise AD over Express Route

# Active Directory Domain Services

 [USERS](#) [GROUPS](#) [APPLICATIONS](#) [DOMAINS](#) [DIRECTORY INTEGRATION](#) [CONFIGURE](#) [REPORTS](#) [LICENSES](#)

domain services PREVIEW

ENABLE DOMAIN SERVICES FOR THIS DIRECTORY

YES

NO

Users will not be able to login to the domain using their credentials until you [enable password synchronization](#).

DNS DOMAIN NAME OF DOMAIN SERVICES

adnitya.onmicrosoft.com

CONNECT DOMAIN SERVICES TO THIS VIRTUAL NETWORK

aaddsvnet | aaddsvnetsubnet(10.0.0.0/20) | West US | subscri...

IP ADDRESS

10.0.0.4; 10.0.0.5

SECURE LDAP (LDAPS)

Configure certificate ...

SECURE LDAP CERTIFICATE

Thumbprint: 5F04A08F19535F7AD0EEA6DF7EE5D9A396ED4626  
Certificate expires: Sat, 31 Dec 2039 23:59:59 GMT

ENABLE SECURE LDAP ACCESS OVER THE INTERNET

YES

NO

EXTERNAL IP ADDRESS FOR LDAPS ACCESS

40.78.63.186

## Why ADDS?

ADDS enables Hadoop services that use Kerberos & LDAPS to continue to work without changes.

DNS name required during cluster provisioning

ADDS can only be added to a V1 VNET. Bridging required to ARM VNET hosting HDI.

## Note:

Tools to ease Kerberized HDI cluster provisioning are being developed. ETA: Q4 CY2016.

# Controlling user access to data



Admin



Dev



Finance

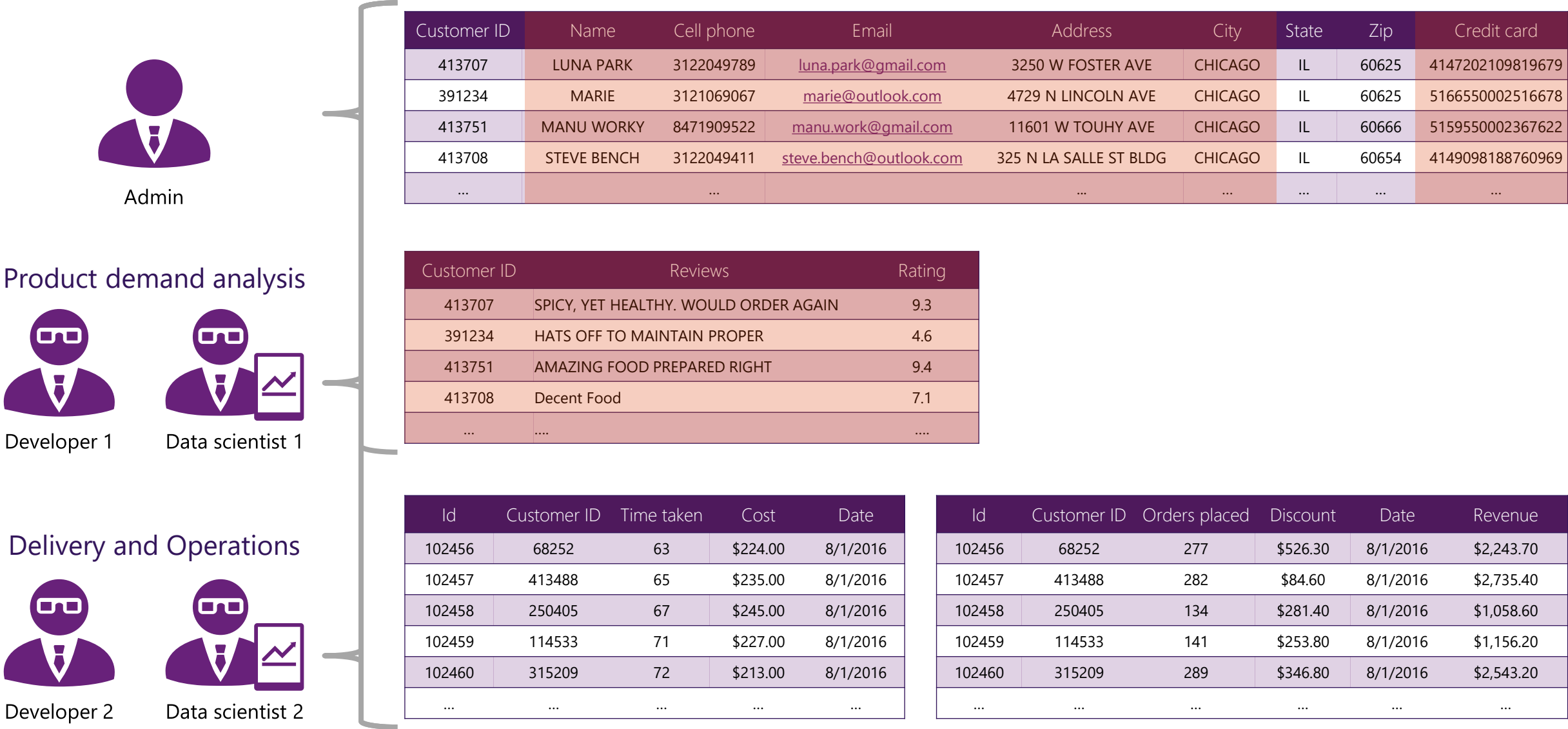
Customer ID	Name	Cell phone	Email	Address	City	State	Zip	Credit card
413707	LUNA PARK	3122049789	<a href="mailto:luna.park@gmail.com">luna.park@gmail.com</a>	3250 W FOSTER AVE	CHICAGO	IL	60625	4147202109819679
391234	MARIE	3121069067	<a href="mailto:marie@outlook.com">marie@outlook.com</a>	4729 N LINCOLN AVE	CHICAGO	IL	60625	5166550002516678
413751	MANU WORKY	8471909522	<a href="mailto:manu.work@gmail.com">manu.work@gmail.com</a>	11601 W TOUHY AVE	CHICAGO	IL	60666	5159550002367622
413708	STEVE BENCH	3122049411	<a href="mailto:steve.bench@outlook.com">steve.bench@outlook.com</a>	325 N LA SALLE ST BLDG	CHICAGO	IL	60654	4149098188760969
...		...		...	...	...	...	...

Customer ID	Reviews	Rating
413707	SPICY, YET HEALTHY. WOULD ORDER AGAIN	9.3
391234	HATS OFF TO MAINTAIN PROPER	4.6
413751	AMAZING FOOD PREPARED RIGHT	9.4
413708	Decent Food	7.1
...	....	....

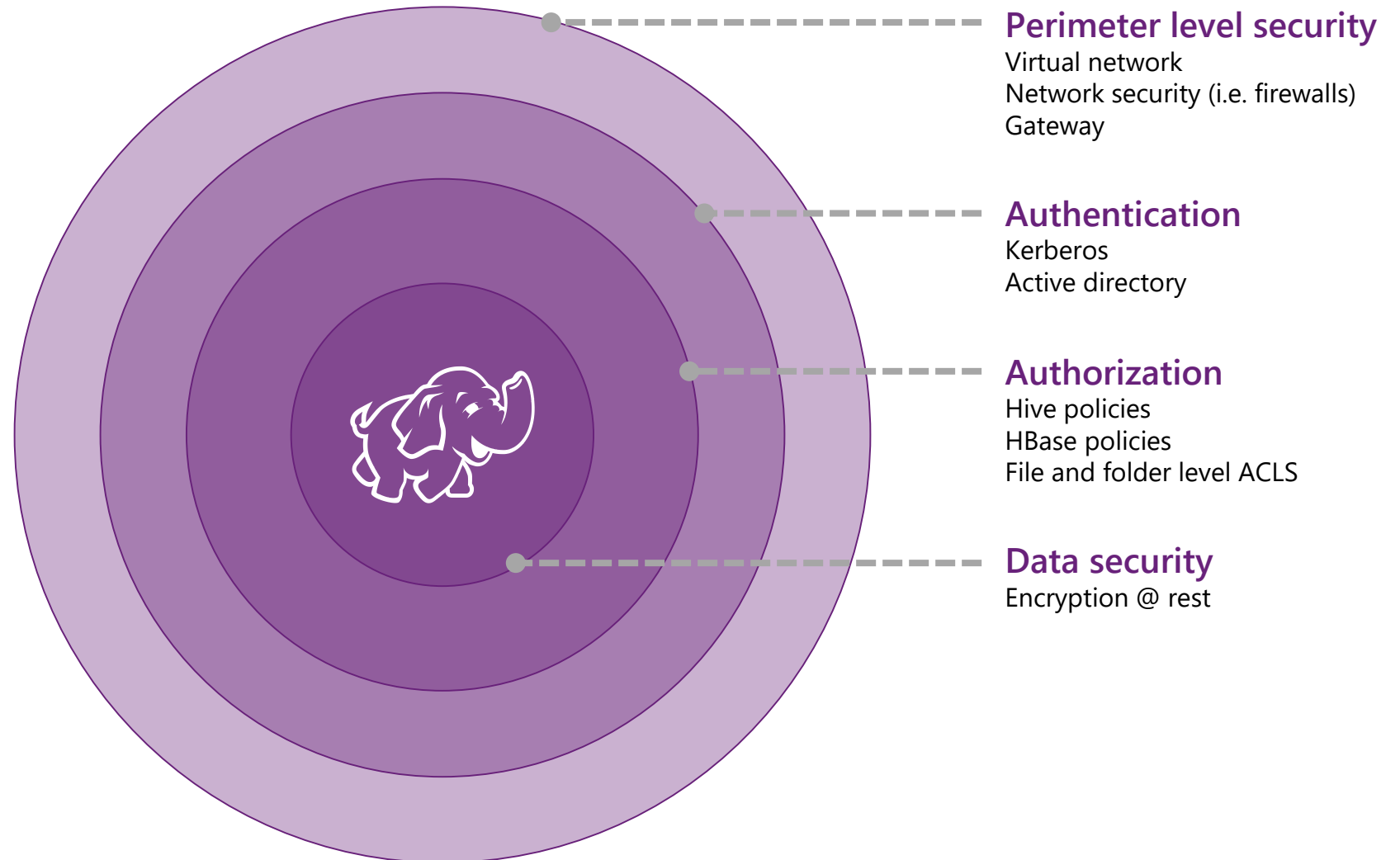
Id	Customer ID	Time taken	Cost	Date
102456	68252	63	\$224.00	8/1/2016
102457	413488	65	\$235.00	8/1/2016
102458	250405	67	\$245.00	8/1/2016
102459	114533	71	\$227.00	8/1/2016
102460	315209	72	\$213.00	8/1/2016
...	...	...	...	...

Id	Customer ID	Orders placed	Discount	Date	Revenue
102456	68252	277	\$526.30	8/1/2016	\$2,243.70
102457	413488	282	\$84.60	8/1/2016	\$2,735.40
102458	250405	134	\$281.40	8/1/2016	\$1,058.60
102459	114533	141	\$253.80	8/1/2016	\$1,156.20
102460	315209	289	\$346.80	8/1/2016	\$2,543.20
...	...	...	...	...	...

# Controlling user access to data

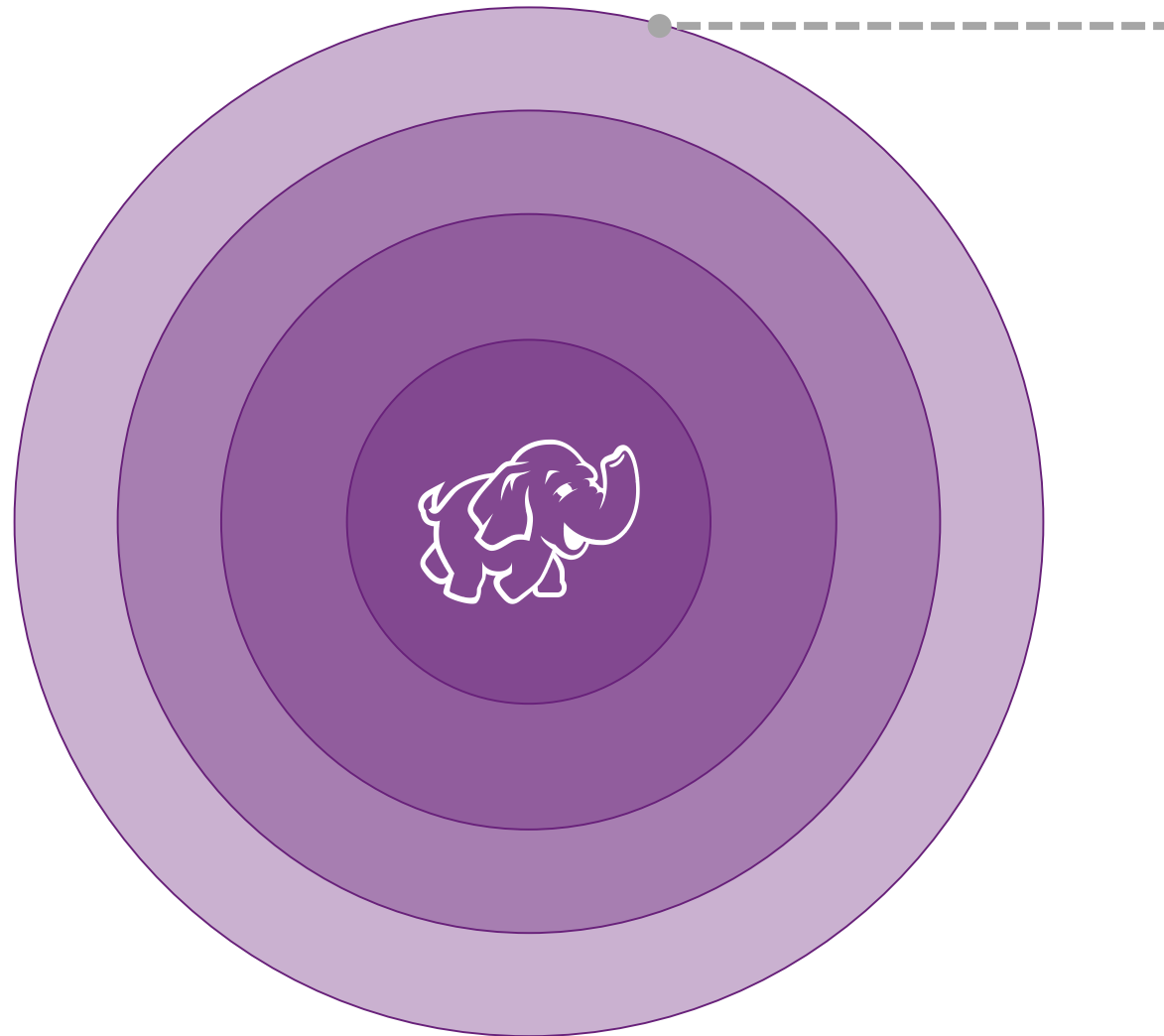


# HDInsight security – rings of defense



# Perimeter level security

Using virtual network and gateway service



## Perimeter level security

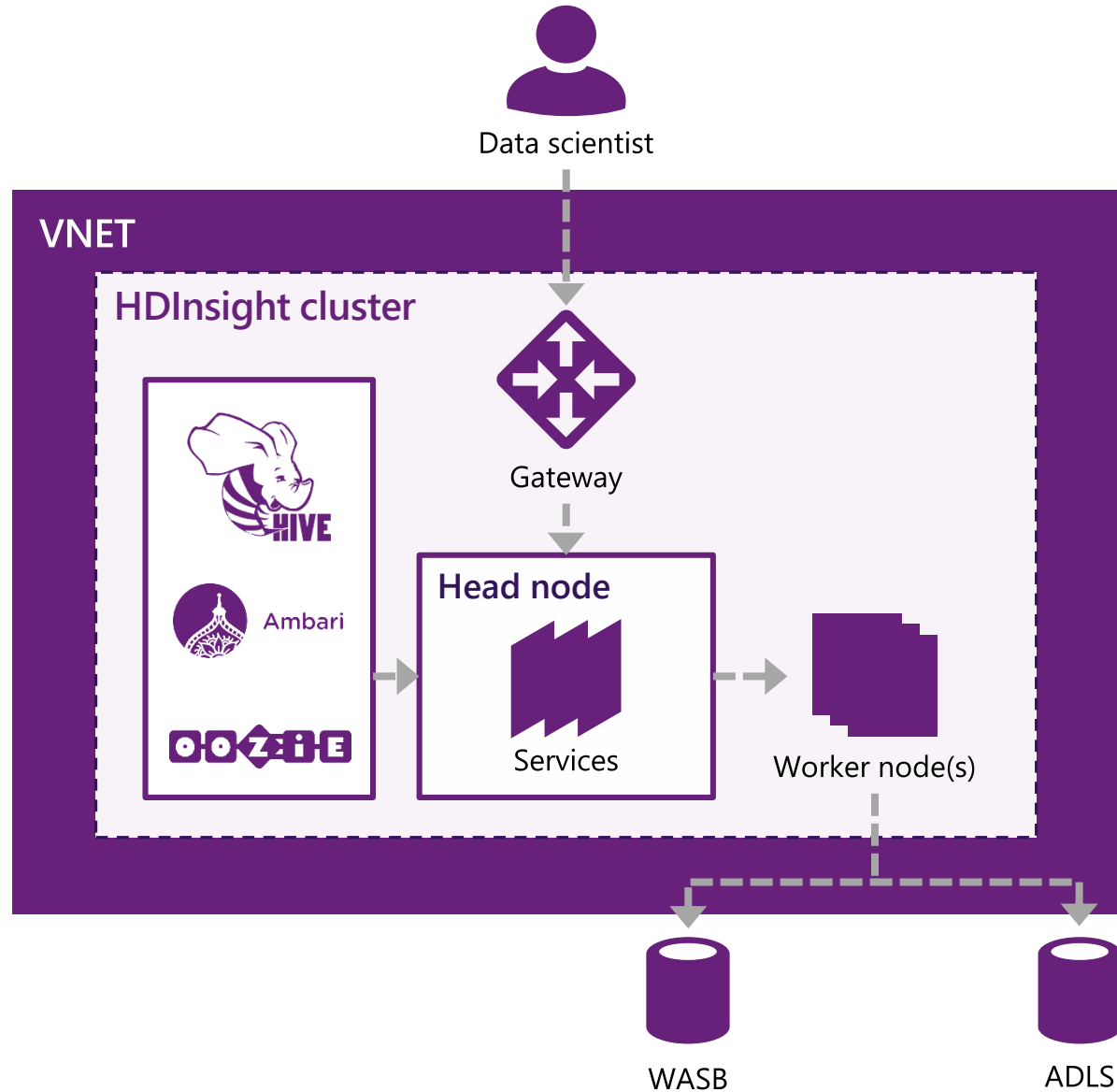
Virtual network

Network security (i.e. firewalls)

Gateway

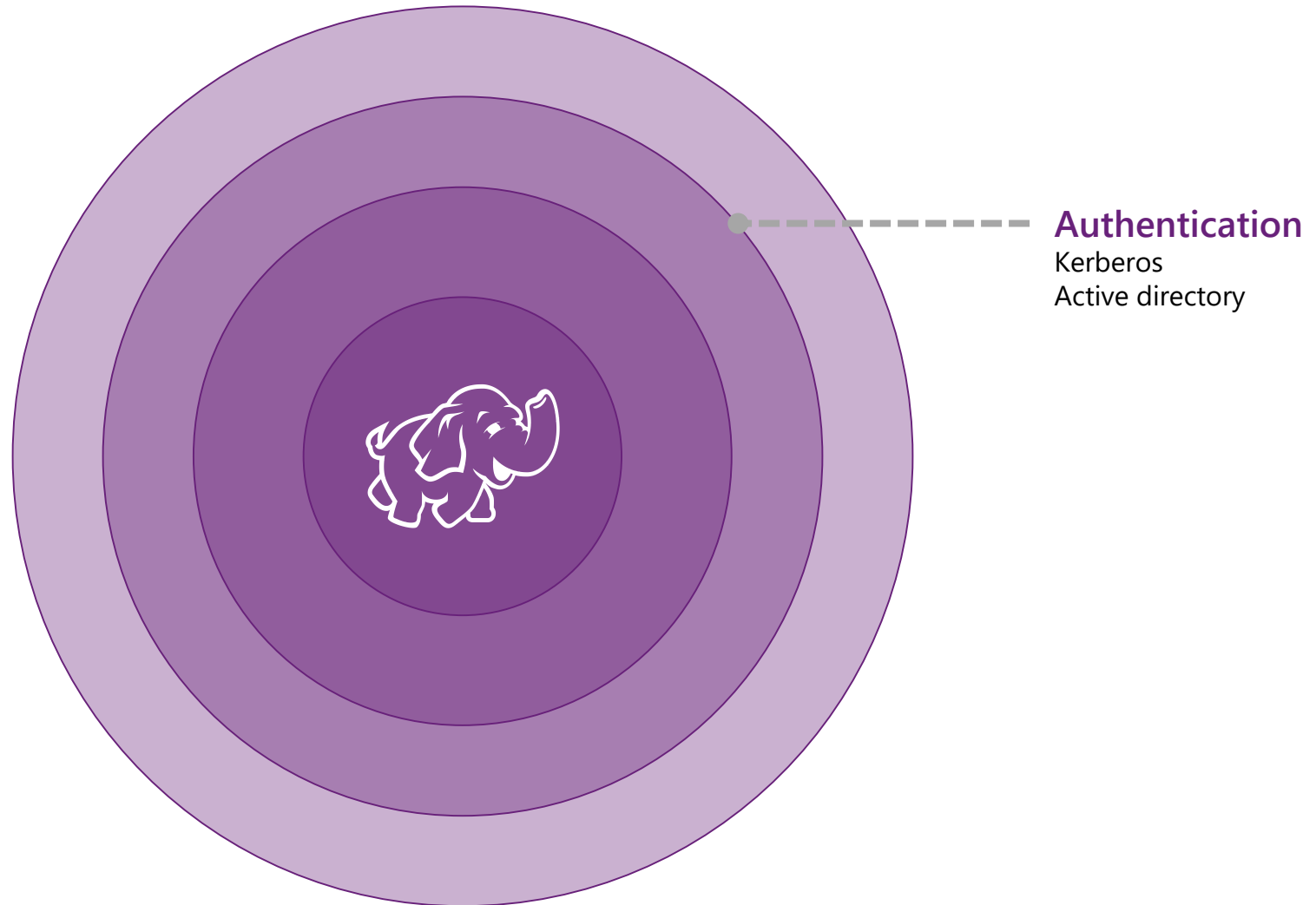


# Perimeter level security



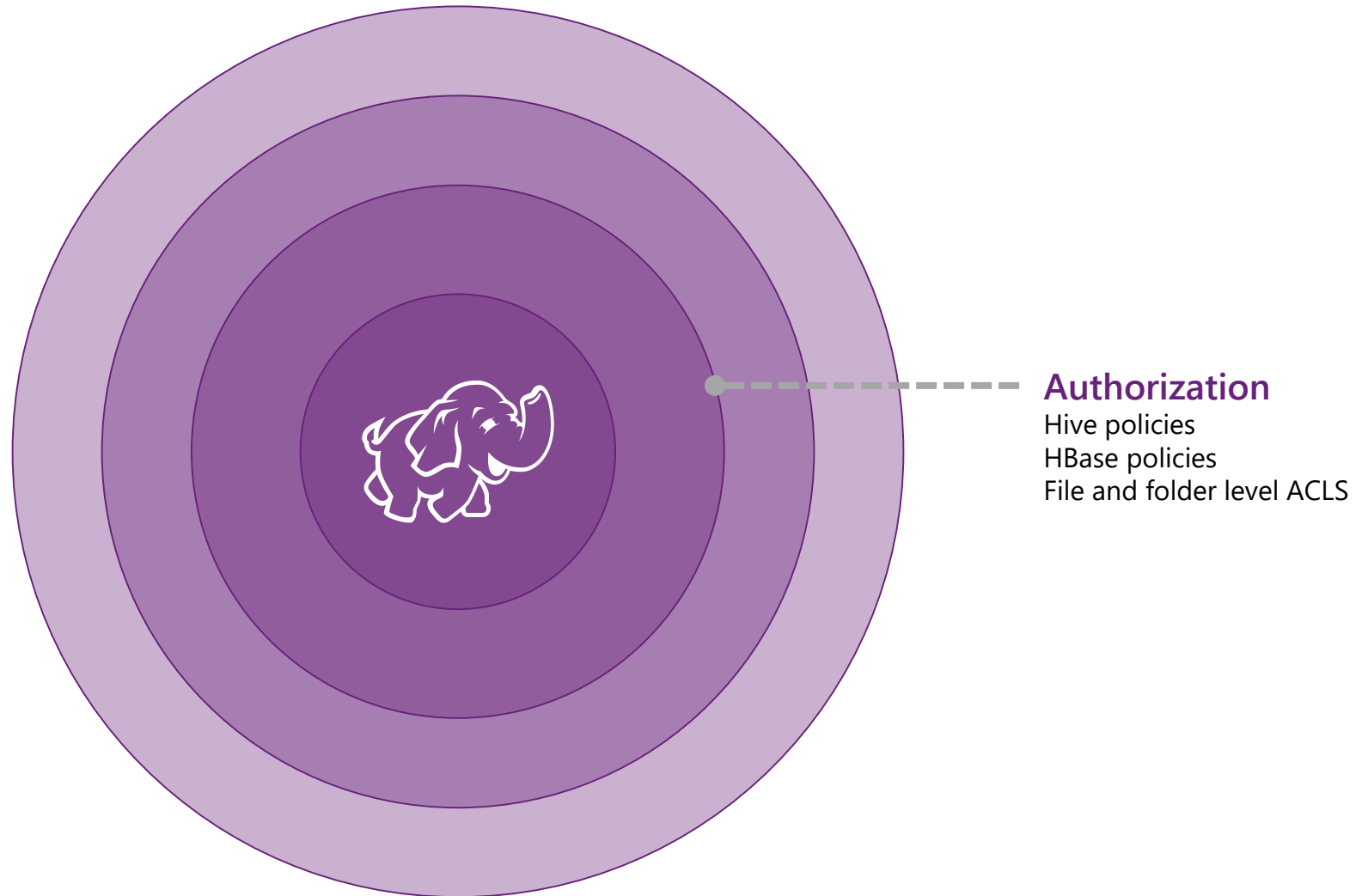
# Authentication

Integration with Azure Active Directory

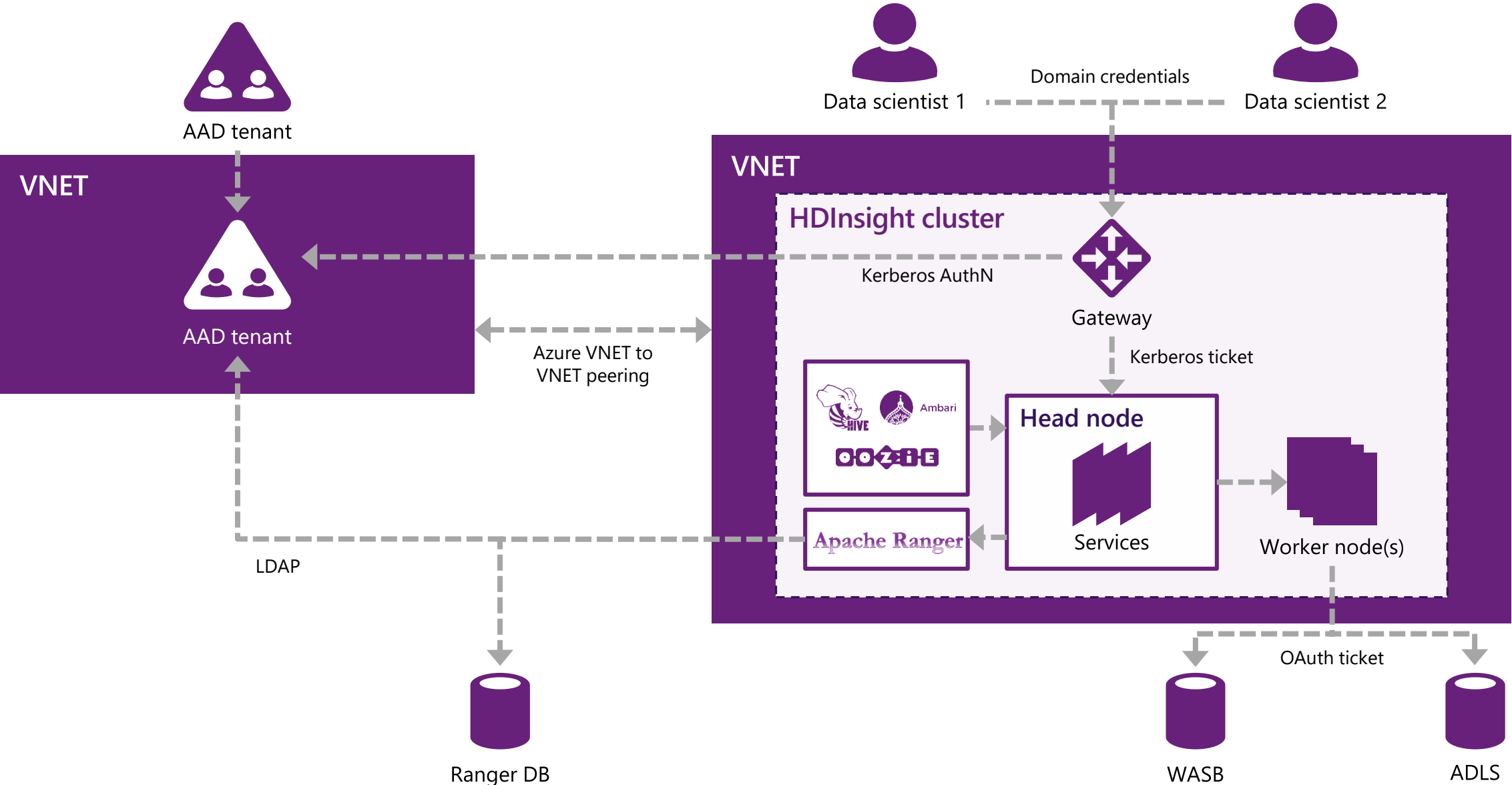


# Authorization

Application and data-level authorization



# Authorization



# Authorization

Secure Endpoints in HDInsight cluster

## Access to all users



HiveServer2



Ambari & Views

Apache Ranger

Ranger

## Access to only cluster admin



SSH



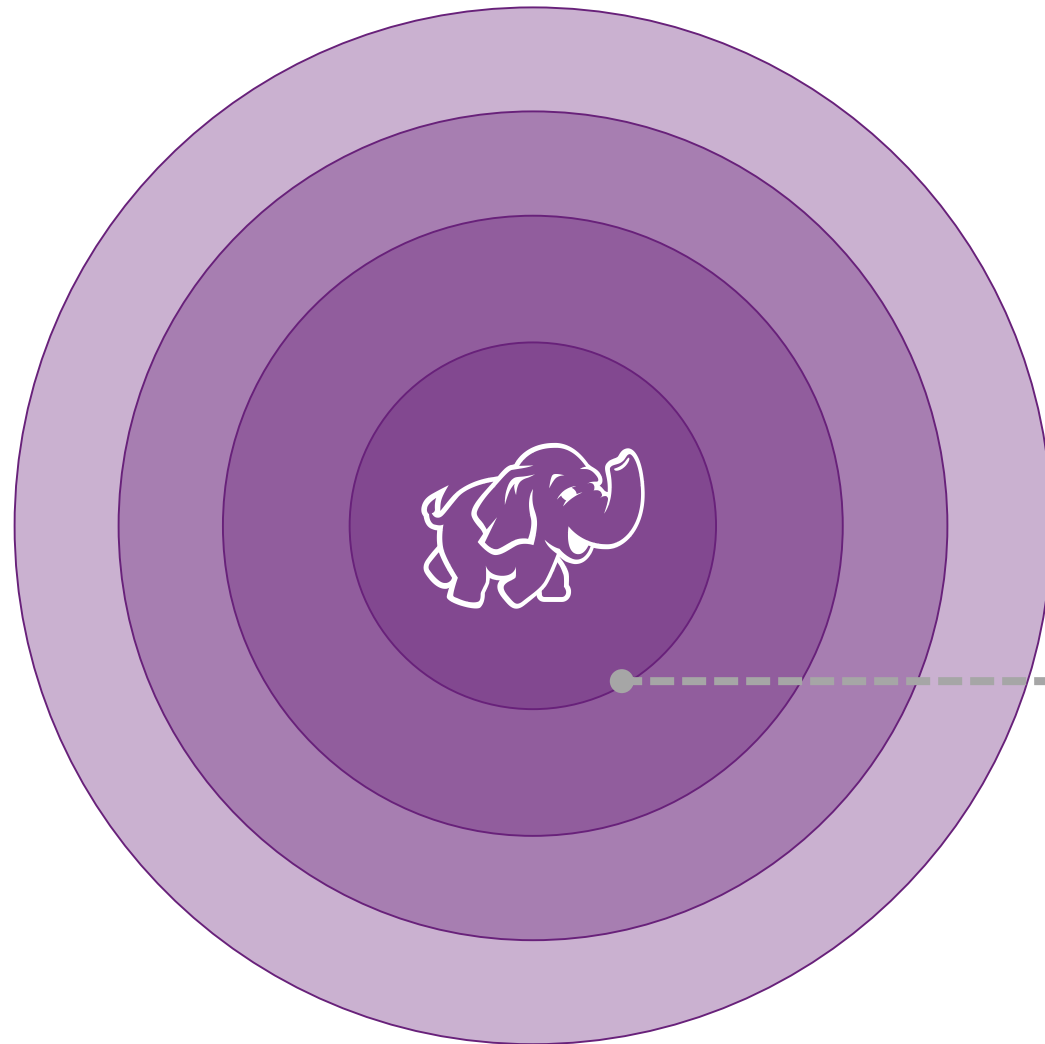
WebHCat



Oozie

# Data security

Transparent Server Side Encryption



**Data security**  
Encryption @ rest

# Transparent Server Side Encryption



Public Preview

## Azure Data Lake Storage

**ALWAYS ON** transparent encryption

All reads/writes are encrypted/decrypted

Service managed keys as well as Customer managed keys



General Availability

## Windows Azure Storage Blob

**ALWAYS ON** transparent encryption

All reads/writes are encrypted/decrypted

Service managed keys

# Data Lake Store

Role-based & POSIX Access control

## Azure roles for account management



### Owner

Can manage accounts and account settings  
Can manage users/access



### Contributor

Can manage accounts and account settings



### Reader

Can view account settings



### User access administrator

Can view account settings  
Can manage users/access



#### PRO TIP:

- Role assignment required to use Azure Portal
- Automated uploaders, downloaders do not need Role assignments

## POSIX acls for files and folders



### On new files/folders

Owner has full permissions – Read/Write/Execute  
Owner group has Read & Execute permissions  
Others have Read & Execute permissions



### Default ACL

Default ACLs (755) are propagated to child objects  
Default ACL can be configured in the UX



#### PRO TIP:

- Use Security groups when setting ACLs and for RBAC role assignments
- A single user to SG assignment will work for both ACL and RBAC.



# Encryption At Rest



## Always on

All newly created ADLS accounts will have encryption at rest enabled by default.

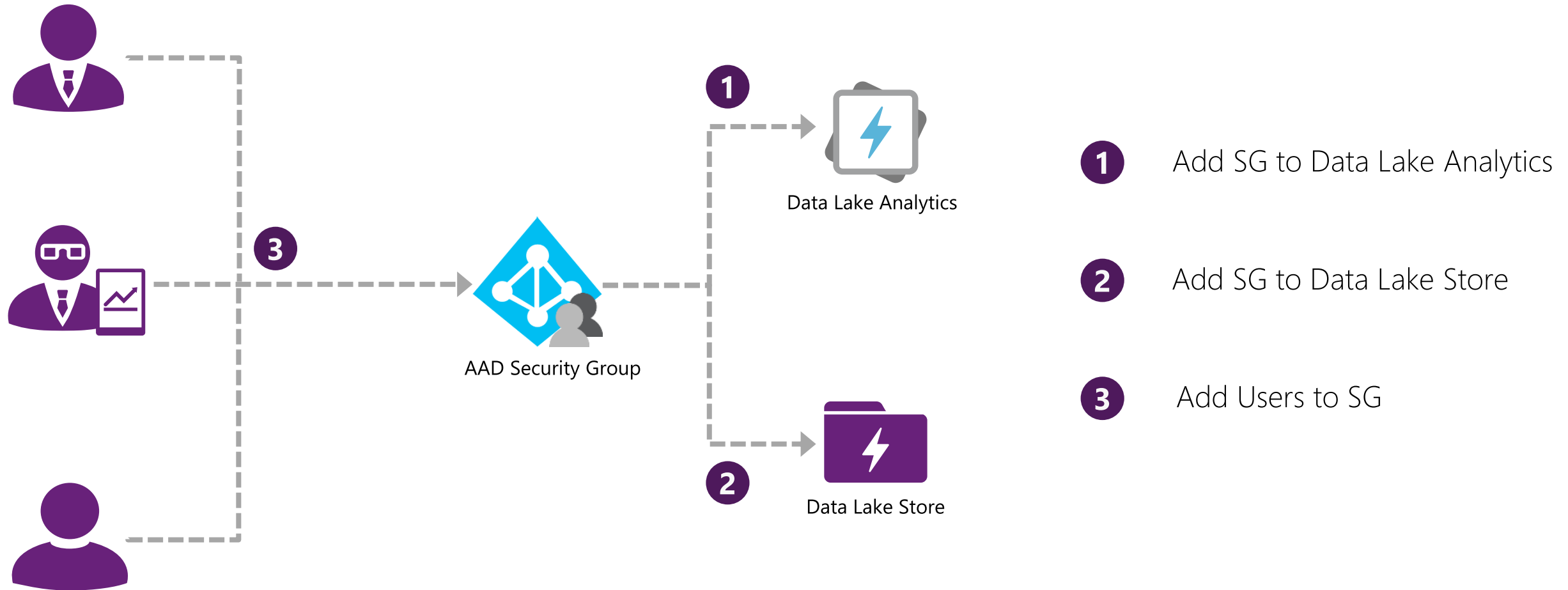


## Fully integrated with Azure Key Vault for encryption key management

Ask ADLS to provision and manage keys

Provision and manage their own keys in AKV

# Simplifying Access Control with Security Groups





# Get started today!

- For more information visit: <http://azure.microsoft.com/en-us/services/hdinsight/>







✓ HDFS

✓ MapReduce2

✓ YARN

📄 Tez

✓ Hive

📄 Pig

📄 Sqoop

✓ Oozie

✓ ZooKeeper

✓ Ambari Metrics

Actions ▾

Summary

Configs

Service Actions ▾

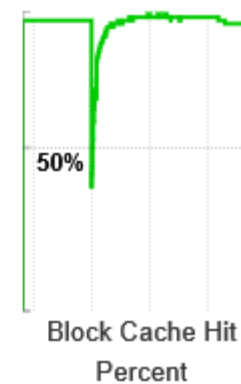
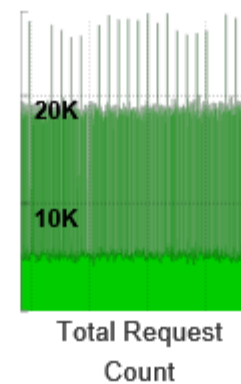
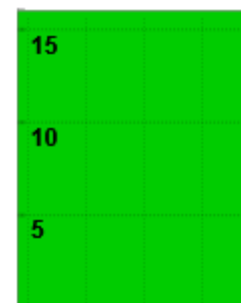
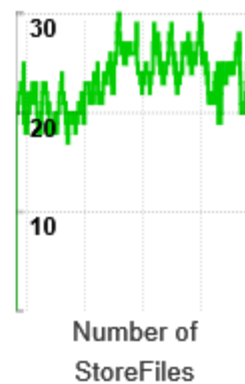
## Summary

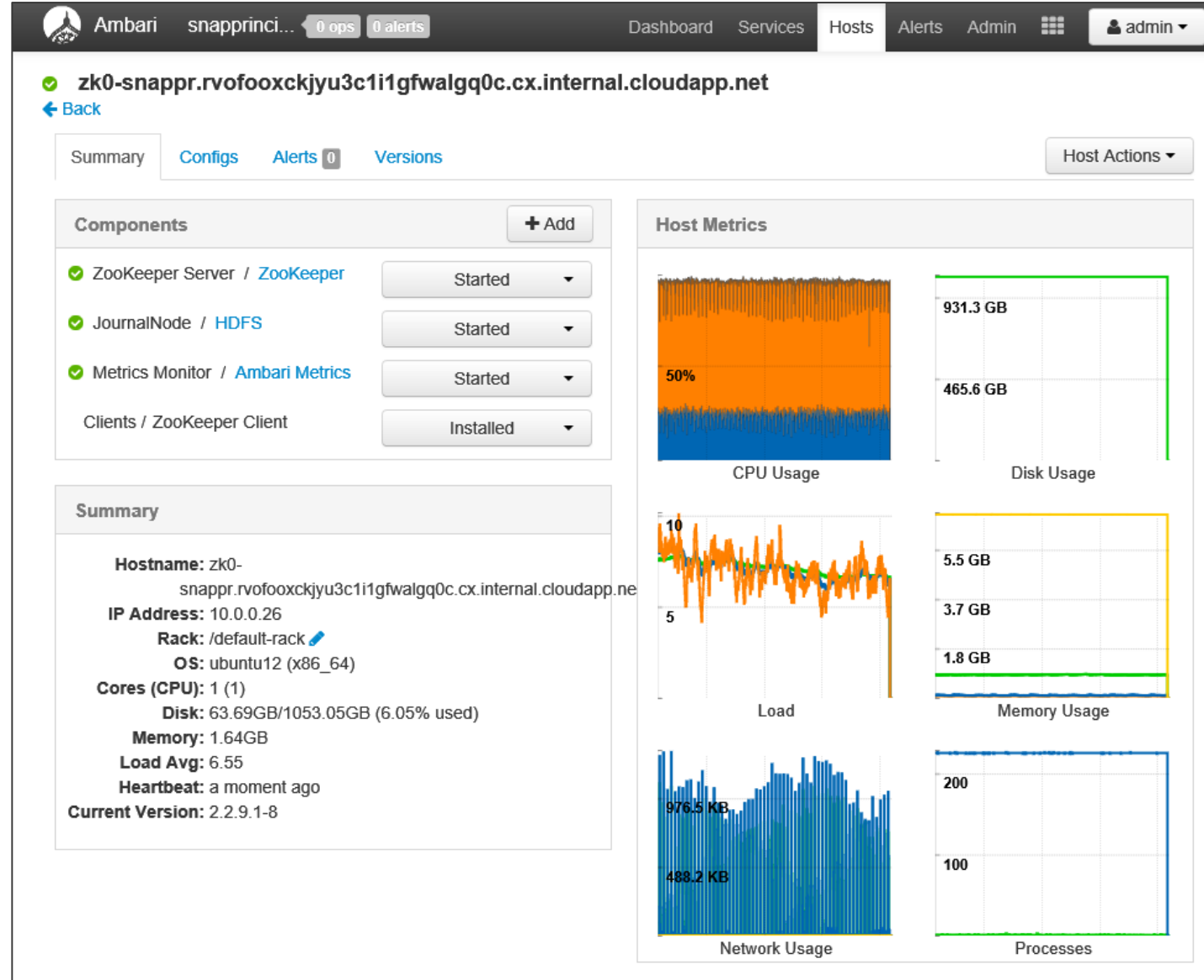
No alerts

[Metrics Collector](#) ✓ Started[Metrics Monitors](#) 24/24 Metrics Monitors Live

## Metrics

Last 24 hours ▾







## NodeManager Web UI

[← Back](#)

OK (19)

## Configuration

[Edit](#)

Description

This host-level alert is triggered if the NodeManager Web UI is unreachable.

Check Interval

5

Minute

Thresholds

OK

HTTP {0} response in {2..3f}s

WARNING

HTTP {0} response from {1} in {2..3f}s ({3})

CRITICAL

Connection failed to {1} ({3})

State: Enabled

Service: YARN

Component: NodeManager

Type: WEB

Groups: YARN Default

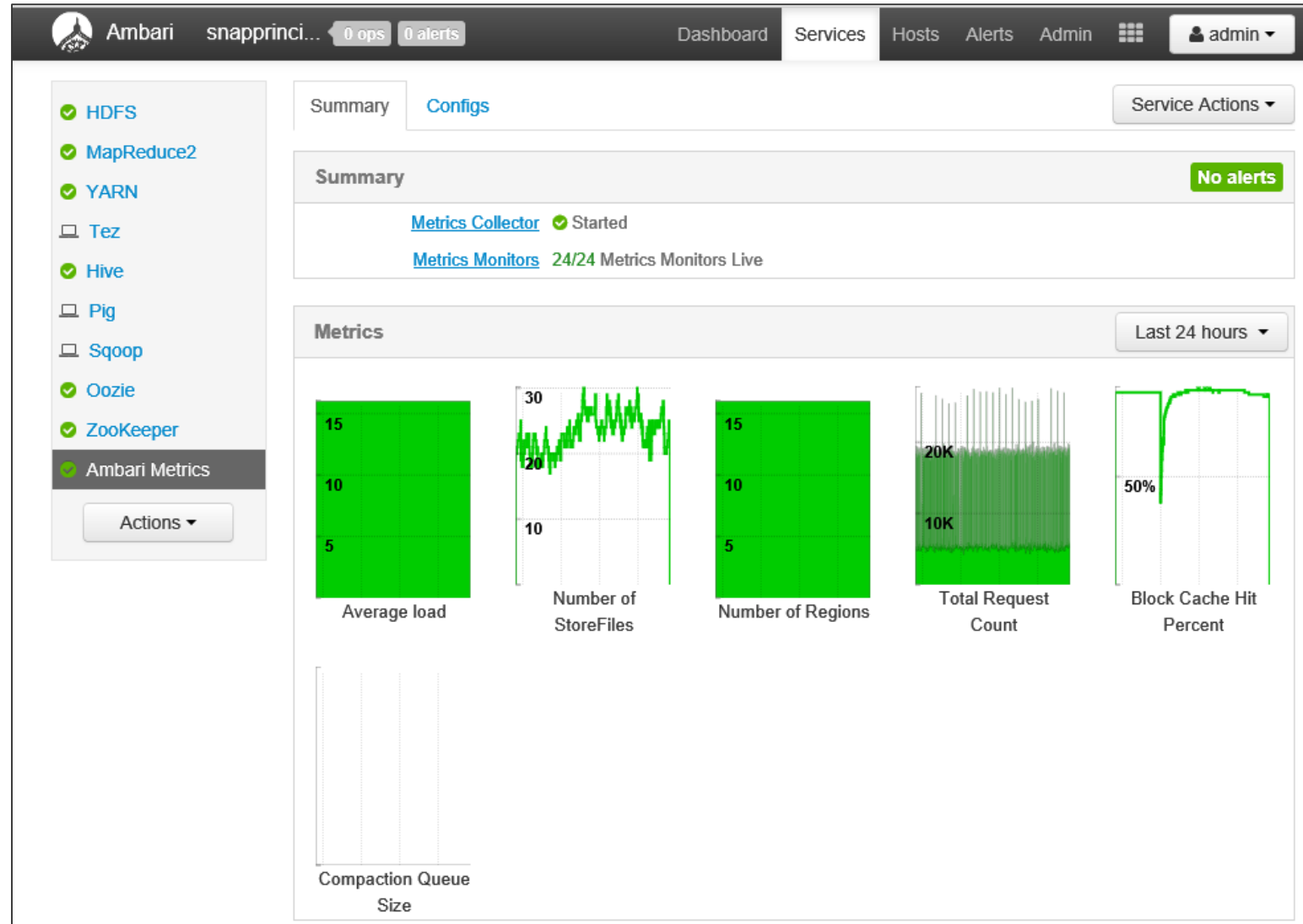
Last Changed: Fri, Mar 25, 2016 11:15

## Instances

Service / Host	Status	24-Hour	Response
<a href="#">YARN / wn0-snappr.rvofooxckjyu3c1i1gfwalgq0c.cx.internal.cloudapp.net</a>	for 3 days	0	HTTP 200 response in 0.000s
<a href="#">YARN / wn1-snappr.rvofooxckjyu3c1i1gfwalgq0c.cx.internal.cloudapp.net</a>	for 3 days	0	HTTP 200 response in 0.000s
<a href="#">YARN / wn10-snappr.rvofooxckjyu3c1i1gfwalgq0c.cx.internal.cloudapp.net</a>	for 3 days	0	HTTP 200 response in 0.000s



# Ambari Web UI: Services (Ambari Metrics)





# Creating a cluster with .NET SDK

# Creating a cluster with .NET SDK

Code to create HDI cluster:

- Linux OS
- Hadoop
- 15 worker nodes
- "EAST US 2" location

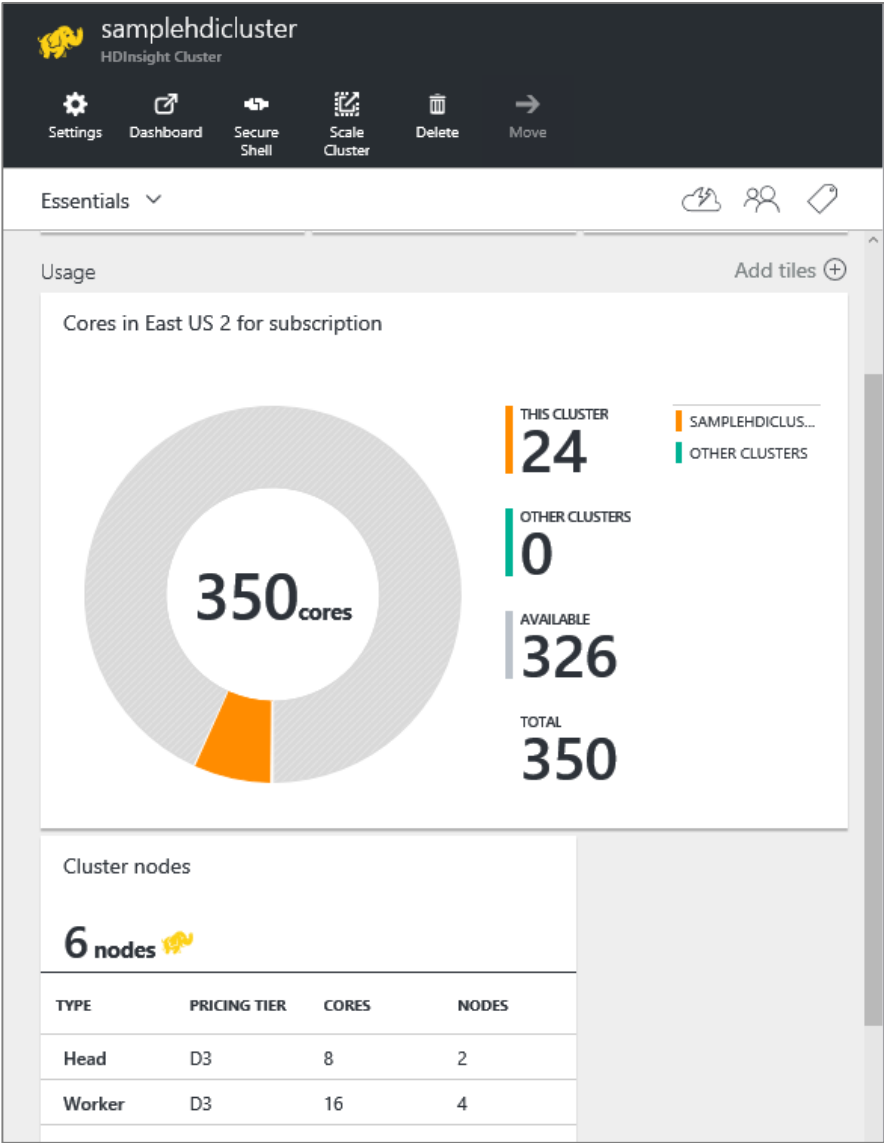
```
var tokenCreds = GetTokenCloudCredentials(); //See notes section for definition of this function
var subCloudCredentials = GetSubscriptionCloudCredentials (tokenCreds, "My Subscription ID");
var resourceManagementClient = new ResourceManagementClient(subCloudCredentials);
var rpResult = resourceManagementClient.Providers.Register("Microsoft.HDIInsight");
_hdiManagementClient = new HDInsightManagementClient (subCloudCredentials);
//specify the cluster configuration details
var parameters = new ClusterCreateParameters {
    ClusterSizeInNodes = 15,
    ClusterType = HDInsightClusterType.Hadoop,
    OSType = OSType.Linux,
    Version = "3.2",
    DefaultStorageAccountName = "mystorageaccount.blob.core.windows.net",
    DefaultStorageAccountKey = "my-storage-key",
    DefaultStorageContainer = "HDIInsightContainer",
    ClusterUserName = "admin",
    Password = "MyPassword",
    Location = "EAST US 2",
    SshUserName = "sshuser",
    SshPublicKey = @"----- BEGIN SSH2 PUBLIC KEY -----
mPCsJVGQLu6O1wqcxRqiKk7keYq8b
P5s30v6blljsLZYTnyReNUa5LtFw7eauGr
----- END SSH2 PUBLIC KEY -----";

};
//Now create the cluster
_hdiManagementClient.Clusters.Create("MyResourceGroup", "MySampleCluster", parameters);
```

# Cluster: Resource Usage Overview

The Azure Portal provides a report on:

- # of cores consumed by this cluster and other clusters
- # of cores available for additional clusters



# Cluster reconfiguration

After the cluster has been created, you can dynamically change (increase or decrease) the number of Worker nodes.

Note: The VM instance type *cannot* be changed.

The screenshot shows the 'Scale Cluster' interface for an HDInsight cluster named 'samplehdicluster'. The top navigation bar includes links for Settings, Dashboard, Secure Shell, Scale Cluster, Delete, and Move. The main content area is divided into two columns. The left column, titled 'Essentials', displays cluster details: Resource group (ADLA\_Benchmark), Status (Running), Location (East US 2), Subscription name (Pay-As-You-Go), Subscription id (bc2d3f0d-ae1d-4f7c-af1e-b2e6932bed5f), URL (samplehdicluster.azurehdinsight.net), Cluster Type (Standard Hadoop on Linux), and Head Node/Worker Nodes (D3 (x2), D3 (x4)). The right column, titled 'Scale Cluster', shows the 'Number of Worker nodes' set to 4, with a red box highlighting this section. Below this, it shows 'Worker Nodes Pricing Tier' as D3 (4 nodes, 16 cores) and 'Head Node Pricing Tier' as D3 (2 nodes, 8 cores). A summary table at the bottom right calculates the total cost: WORKER NODES (0.62 x 4 = 2.49) and HEAD NODES (0.62 x 2 = 1.24), resulting in a TOTAL COST of 3.73 USD/HOUR (ESTIMATED). A note states that this estimate does not include storage costs, network egress costs, or subscription discounts. The bottom of the interface features 'Quick Links' for Cluster Dashboard, Ambari Views, and Scale Cluster (highlighted with a blue box), and a 'Usage' section with an 'Add tiles' button.

samplehdicluster  
HDInsight Cluster

Settings Dashboard Secure Shell Scale Cluster Delete Move

Essentials ^

Resource group [ADLA\\_Benchmark](#)  
Status **Running**  
Location **East US 2**  
Subscription name [Pay-As-You-Go](#)  
Subscription id **bc2d3f0d-ae1d-4f7c-af1e-b2e6932bed5f**

URL [samplehdicluster.azurehdinsight.net](#)  
Cluster Type **Standard Hadoop on Linux**  
Head Node, Worker Nodes **D3 (x2), D3 (x4)**  
Learn more [Documentation](#)  
Getting Started [Quickstart](#)

[All settings →](#)

Quick Links Add tiles +

Cluster Dashboard Ambari Views Scale Cluster

Usage Add tiles +

Scale Cluster  
samplehdicluster

Save

Number of Worker nodes 4 ✓

Worker Nodes Pricing Tier **D3 (4 nodes, 16 cores)**

Head Node Pricing Tier **D3 (2 nodes, 8 cores)**

WORKER NODES	0.62 x 4 = 2.49
HEAD NODES	0.62 x 2 = 1.24
TOTAL COST	<b>3.73</b>

USD/HOUR (ESTIMATED)


24 of 350 cores would be used in East US 2.







This price estimate does not include storage costs, network egress costs, or subscription discounts.




# Post-creation Actions

# Security: Role-based Access

New users can be added in the role of "Owner", "Contributor", Reader or "User Access Administrator"  
Users can be added or deleted at anytime

**samplehdicluster**  
HDInsight Cluster

 Settings  Dashboard  Secure Shell  Scale Cluster  Delete  Move

Essentials   

Resource group  
[ADLA\\_Benchmark](#)

Status  
**Running**

Location  
East US 2

Subscription name  
[Pay-As-You-Go](#)

Subscription id  
bc2d3f0d-ae1d-4f7c-af1e-b2e6932bed5f

URL  
[samplehdicluster.azurehdinsight.net](#)

Cluster Type  
**Standard Hadoop on Linux**




Head Node, Worker Nodes  
D3 (x2), D3 (x4)

Learn more  
[Documentation](#)

Getting Started  
[Quickstart](#)

[All settings →](#)

Quick Links Add tiles +

 Cluster Dashboard  Ambari Views  Scale Cluster

Users  
samplehdicluster





 Add  Roles

USER	ROLE	ACCESS
 [redacted]@outlook.com	Owner	Inherited ...
 Subscription admins ⓘ	Owner	Inherited ...

Roles  
samplehdicluster

NAME
 Owner ⓘ
 Contributor ⓘ
 Reader ⓘ
 User Access Administrator ⓘ

# Security: Roles and Privileges

Roles	
samplehdicluster	
NAME	
 Owner ⓘ	
 Contributor ⓘ	
 Reader ⓘ	
 User Access Administrator ⓘ	

Role	Privilege
Owner	Lets you manage everything
Contributor	Lets you manage everything except access to resources
Reader	Lets you view everything but not make changes
User Access Administrator	Lets you manage user access to Azure resources

# HDInsight Cluster Settings

The screenshot shows the 'samplehdicluster' settings page in the Azure Portal. The top navigation bar includes links for Settings, Dashboard, Secure Shell, Scale Cluster, Delete, and Move. The main content area is divided into two columns. The left column, titled 'Essentials', displays key cluster information: Resource group (ADLA\_Benchmark), Status (Running), Location (East US 2), Subscription name (Pay-As-You-Go), and Subscription id (bc2d3f0d-ae1d-4f7c-af1e-b2e6932bed5f). It also provides the URL (samplehdicluster.azurehdinsight.net), Cluster Type (Standard Hadoop on Linux), and Head Node/Worker Nodes (D3 (x2), D3 (x4)). A 'Quick Links' section offers shortcuts to the Cluster Dashboard, Ambari Views, and Scale Cluster. The right column, titled 'Settings', contains a 'Filter settings' search bar and a list of settings categories: SUPPORT + TROUBLESHOOTING (Audit logs), GETTING STARTED (Quick Start), CONFIGURATION (Cluster Login, Scale Cluster, Secure Shell, HDInsight Partner, External Metastores), GENERAL (Script Actions, Apps), PROPERTIES (Properties, Azure Storage Keys, Cluster AAD Identity), and RESOURCE MANAGEMENT (Users, Tags). A blue 'All settings' button is located at the bottom of the Essentials section.

The Azure Portal lets you view and change all these settings after the cluster has been created

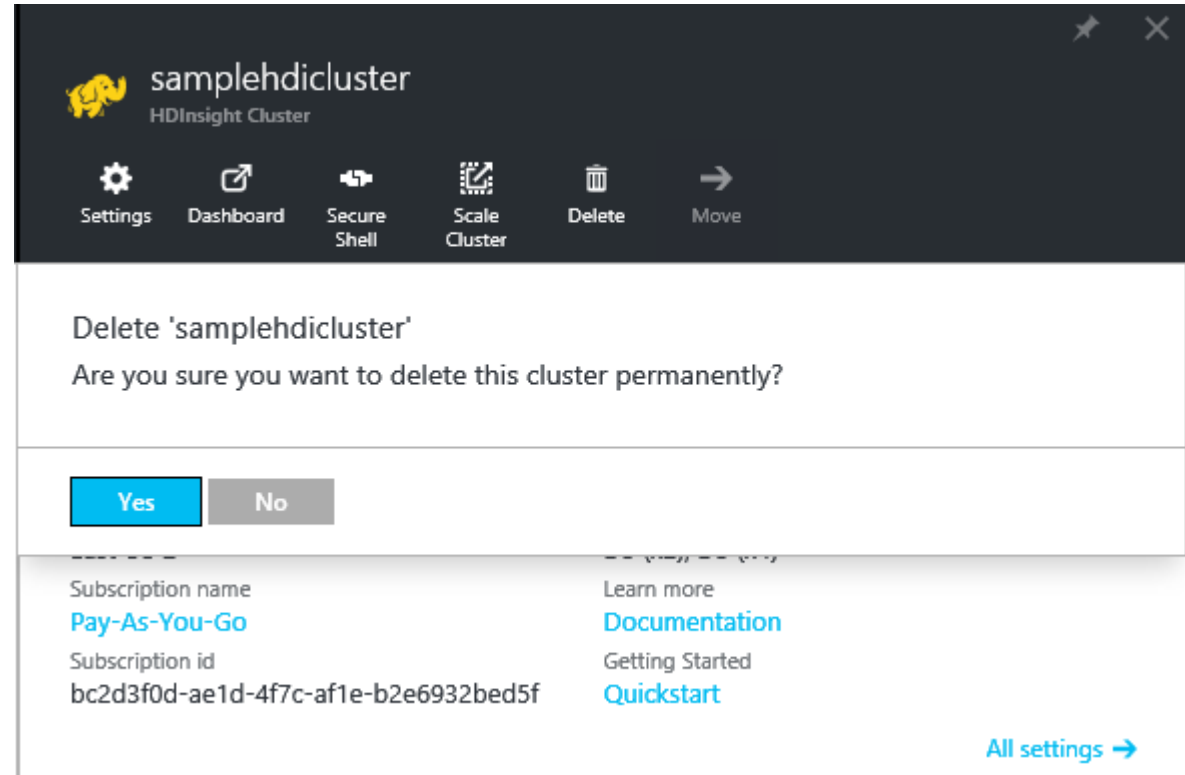


# Deleting a HDI Cluster

A running cluster can be deleted permanently freeing up the used cores.

Freed cores can be used to create a new cluster or expand an existing one.

Storage (WASB or ADLS) must be deleted separately



# Audit Logs

# Audit Logs

Audit Logs shows  
**Critical, Error, Warning**  
and **Informational**  
events

Audit logs can be  
archived into Azure  
storage or stream to  
Azure Event Hub

**Settings**  
ntlive19hdbm0531

**Events**

Filter Columns Hide chart Export

Filtered for past week  
by resource /subscriptions/15c5cb6e-191a-40ea-9f69-08207a17fe97/resourceGroups/\_ElasticData...  
event category = All, levels = All

1  
0.9  
0.8  
0.7  
0.6  
0.5  
0.4  
0.3  
0.2  
0.1  
0

10:29 AM

CRITICAL 0 ERROR 0 WARNING 0 INFORMATIONAL 1

Filter items ...

OPERATION	LEVEL	STATUS	RESOURCE	TI...
Write Clusters	Informational	Succeeded	...clusters/ntlive19h...	1 d ago

**Export Audit Logs (PREVIEW)**

Save Discard Reset

Archive your Audit logs to a storage account or stream them to an Azure Event Hub. Diagnostic data is billed at normal storage rates.

\* Subscription ⓘ  
Azure conversion

\* Regions ⓘ  
0 selected

\* STORAGE ACCOUNT  
Configure required settings

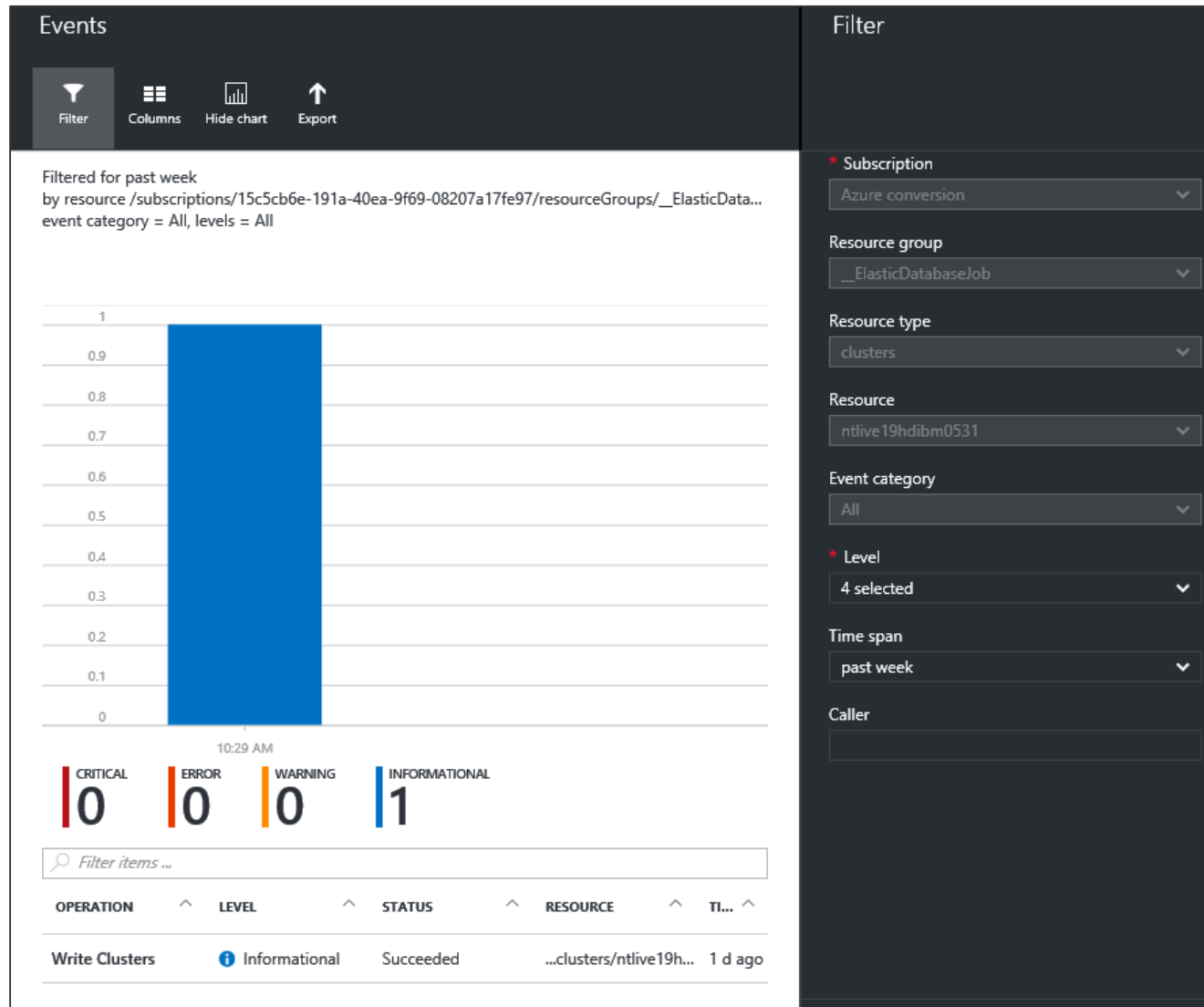
Retention (days) ⓘ  
0

AZURE EVENT HUB ⓘ  
Optionally configure Event Hub

# Filtering Audit Logs

Audit Logs entries can be filtered by:

- Time
- Type
- Level
- ...



\* Level

4 selected

- ☒ Critical
- ☒ Error
- ☒ Warning
- ☒ Informational

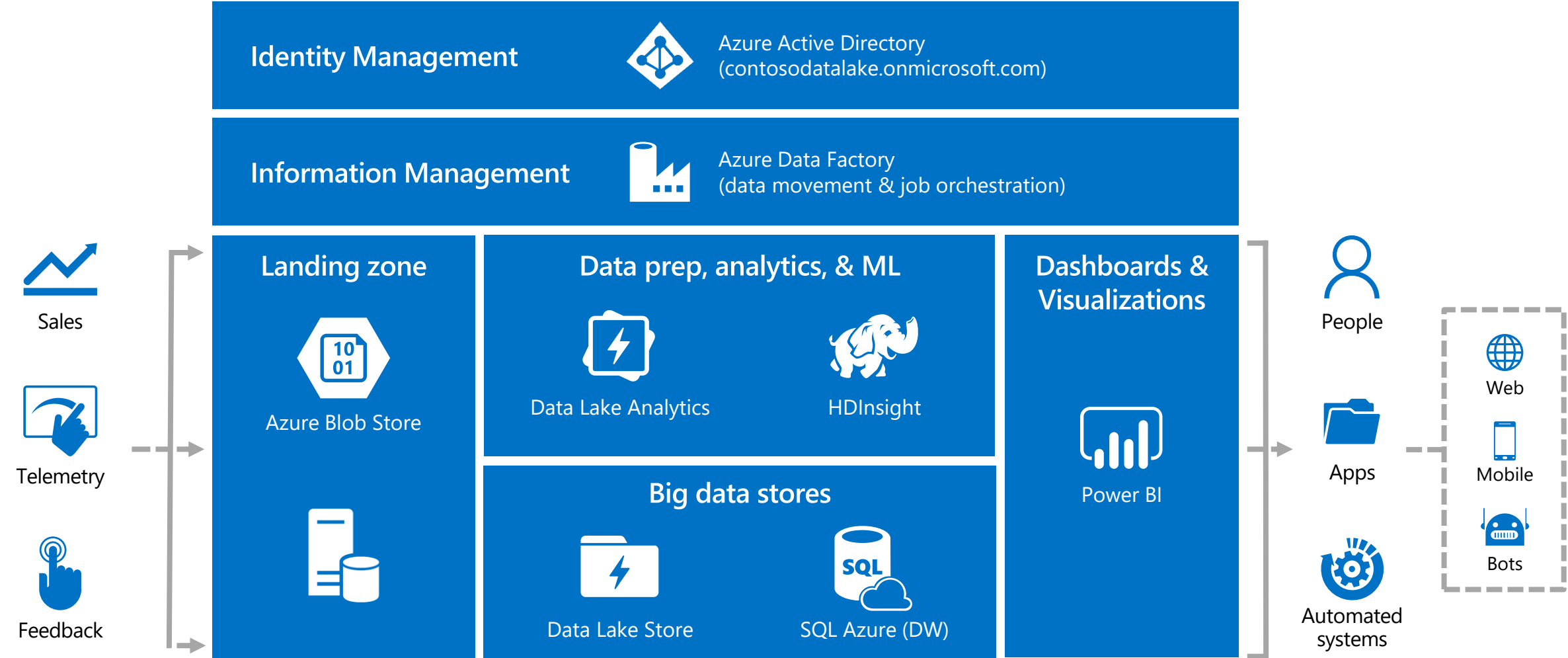
\* Level

4 selected

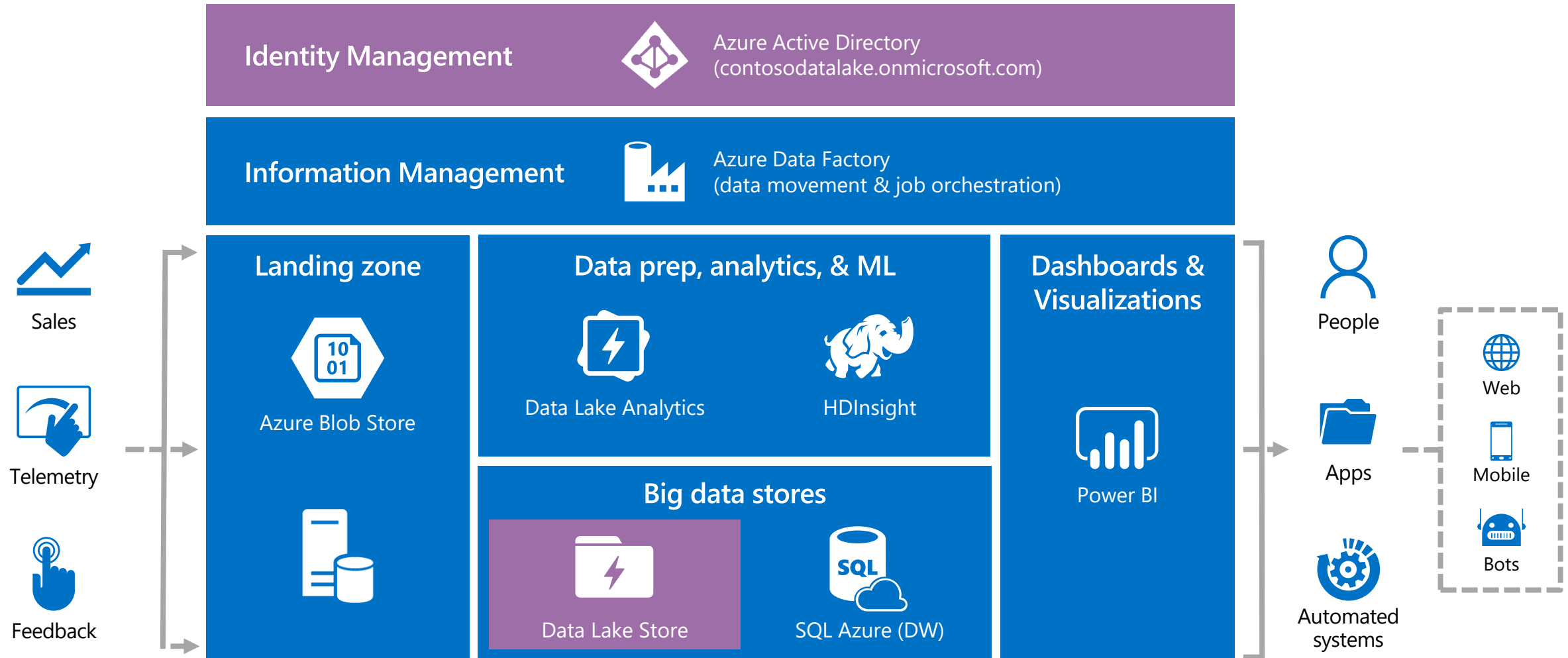
- past 1 hour
- past 24 hours
- past week
- custom

# Appendix

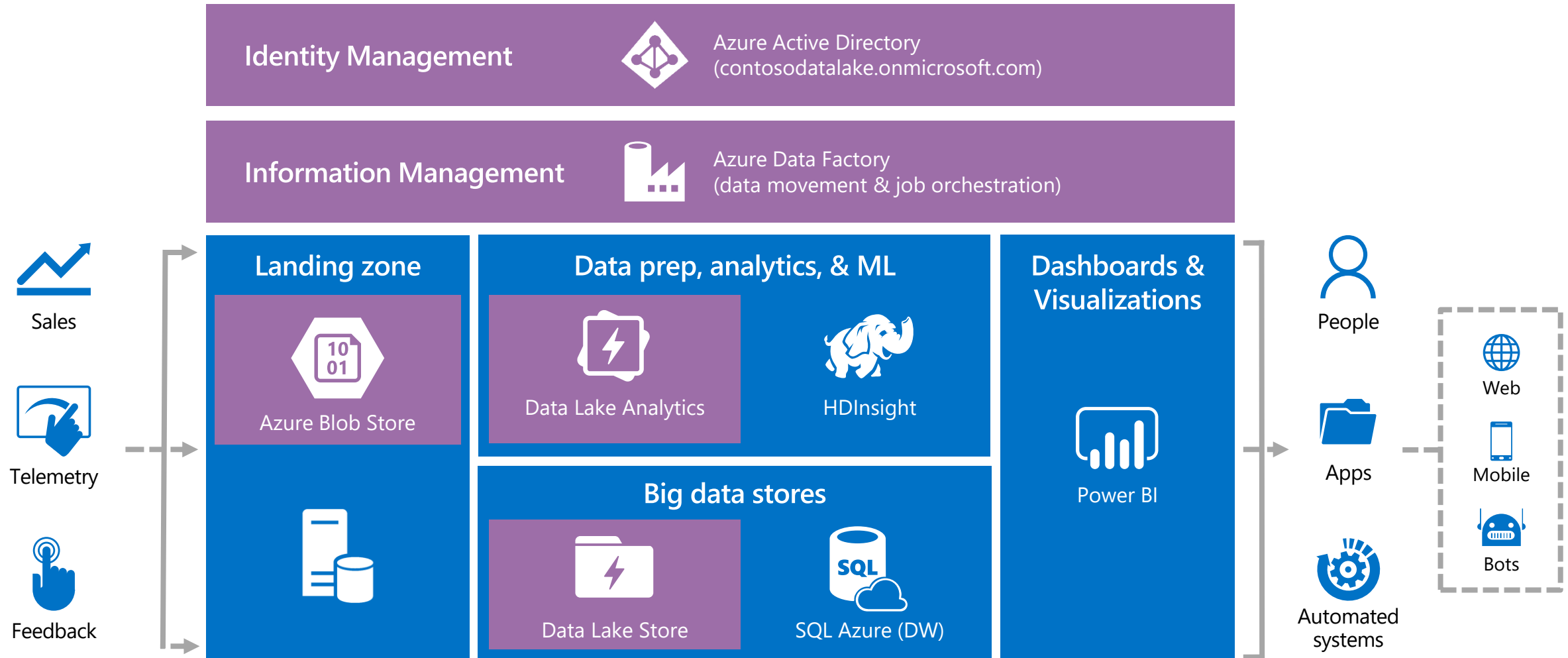
# Contoso big data pipeline



# Contoso big data pipeline



# Contoso big data pipeline





# Data Lake Analytics

## Role-based Access Control

### Account & job management



#### **Owner**

- Manage accounts and account settings
- Manage users/access
- Submit, monitor, and manage jobs



#### **Contributor**

- Manage accounts and account settings
- Submit, monitor, and manage jobs



#### **Reader**

- Monitor jobs



#### **User access administrator**

- Manage users/access
- Only monitor jobs



#### **ADL analytics developer:**

- Submit, monitor, and manage their own jobs

### Account & job management



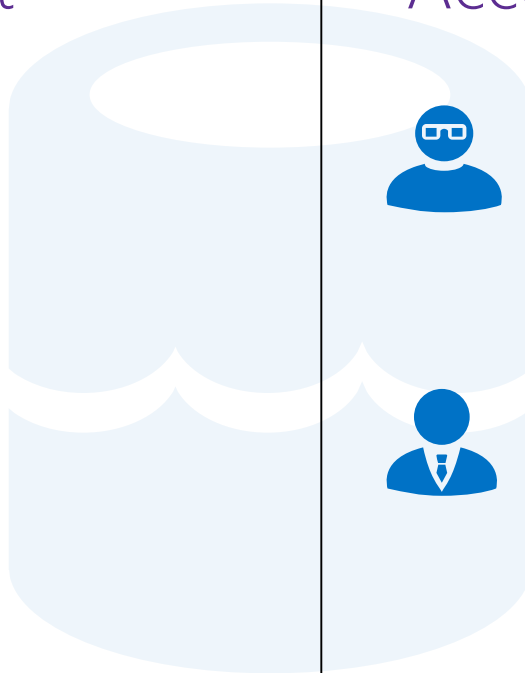
#### **Creator of a database**

- Owns all the objects within the database
- Read, write and delete objects
- Grant permissions to others



#### **Owner**

- Grant Read access to a database (incl. definitions)
- Enumerate objects in database
- Create & update objects within database



# Contoso big data pipeline

