

# **Hadoop as a Service**

**VMware vCloud Automation Center & Big Data Extension**



# Table of Contents

- 1. Introduction..... 2
  - 1.1 How it works..... 2
- 2. System Pre-requisites..... 2
- 3. Set up..... 2
  - 3.1 Request the Service as a Catalog Consumer ..... 3
  - 3.2 View the status of the application deployment..... 6
  - 3.3 Launching the hadoop service..... 8
- 4. Summary ..... 10
- 5. Benefits..... 10

# 1. Introduction

VMware vCloud Automation Center is an innovative self-service provisioning and lifecycle management solution that simplifies and automates deployments of infrastructure, multi-tier applications, desktop... and now any kind of IT service! It provides a secure portal where authorized administrators, developers or business users can request new IT services as well as manage specific cloud and IT resources based on their roles and privileges.

BDE enables customers to run clustered, scale-out Hadoop applications through vSphere, delivering all the benefits of virtualization to Hadoop users. BDE delivers increased agility through an easy to use interface, elastic scaling through the separation of compute and storage resources, and increased reliability and security by leveraging proven vSphere technology.

Hadoop is designed to run on a large cluster of commodity servers and to scale to hundreds or thousands of nodes. Each disk, server, network link, and even rack within the cluster is assumed to be unreliable. This assumption allows the use of the least expensive cluster components consistent with delivering sufficient performance, including the use of unprotected local storage (JBODs).

Hadoop as a service runs on top of the Big Data Extensions allows you to automate the deployment and management of Apache Hadoop and HBase on virtual environments such as vSphere.

## 1.1 How it works

BDE is a downloaded virtual appliance integrated as a plug-in to VMware vCenter Server™. BDE requires that you have vSphere 5.0 or later license and an Enterprise or Enterprise Plus license to leverage vSphere HA / FT. The Serengeti virtual appliance runs on top of vSphere and includes two virtual machines: Serengeti Management Server and Hadoop.Template. The Serengeti Management Server allows users to setup the infrastructure for the cluster, including virtual machine creation and cloning the Hadoop Template VM. Once the Serengeti Management Server creates the nodes in the cluster, the Hadoop distribution software is injected into the newly created Hadoop Template virtual machines. Master node and Slave node roles are assigned to virtual machines and then the appropriate Hadoop service is started. Users can then configure / re-configure their cluster on the fly through vCenter.

## 2. System Pre-requisites

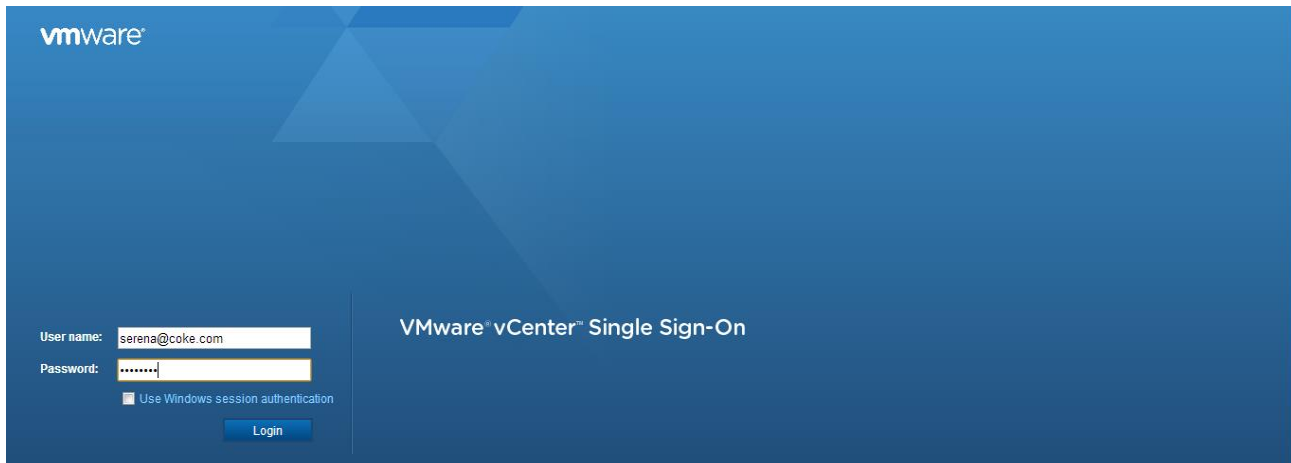
1. Install and configure vCO 6.0
2. Install and Configure vCAC 6.0
3. Install and configure VMware vCO 6.0 with vCAC 6.0
4. Big Data Extensions configuration with vSphere 5.5

## 3. Set up

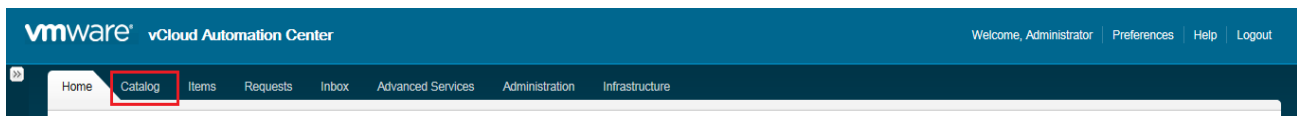
Follow the below steps to deploy the hadoop as a service in vCAC 6.0:

### 3.1 Request the Service as a Catalog Consumer

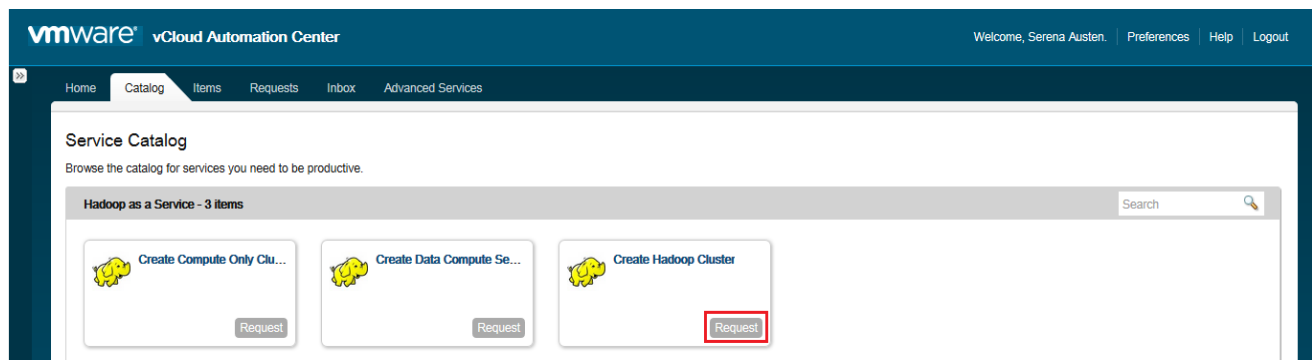
1. Log into vCAC Service Catalog (<https://wdc-auroravcf-gen-dhcp85.eng.vmware.com/shell-ui-app/org/qe/>) using SSO as a Catalog Consumer either tenant admin ([tony@coke.com](mailto:tony@coke.com)) or tenant user ([serena@coke.com](mailto:serena@coke.com)/password)



2. Click on the **Catalog** tab



3. There are list of catalog services. Click on the **Request** Button for “Create Hadoop Cluster”



4. Provide Description & Details and click **next**.

Home Catalog Items Requests Inbox Advanced Services

### New Request

Create Hadoop Cluster

Request Information Cluster Name Master Configuration (OPTIONAL) Worker Configuration (OPTIONAL) Client Configuration (OPTIONAL)

\* Description: hadoop\_cluster

\* Reasons: published from vCAC

< Back Next > Save Cancel

5. Enter the cluster name to be created and click **Next**

Home Catalog Items Requests Inbox Advanced Services

### New Request

Create Hadoop Cluster

Request Information Cluster Name Master Configuration (OPTIONAL) Worker Configuration (OPTIONAL) Client Configuration (OPTIONAL)

\* name: hadoop\_cluster

< Back Next > Save Cancel

6. Review all the Master Configuration. If required you can change the default value as per your requirement. Click **Next**.

Home Catalog Items Requests Inbox Advanced Services

### New Request

Create Hadoop Cluster

Request Information Cluster Name Master Configuration (OPTIONAL) Worker Configuration (OPTIONAL) Client Configuration (OPTIONAL)

\* Master CPU Number: 2

\* Master Memory Capacity MB: 4,096

\* Master Storage Size GB: 10

< Back Next > Save Cancel

7. Review all the Worker Configuration. If required you can change the default value as per your requirement. Click **Next**.

The screenshot shows the 'New Request' form in the VMware vCloud Automation Center. The 'Worker Configuration (OPTIONAL)' tab is selected. The form contains the following fields:

- Worker CPU Number: 1
- Worker Memory Capacity MB: 2,048
- Worker Storage Size GB: 10
- Worker Instance Number: 3

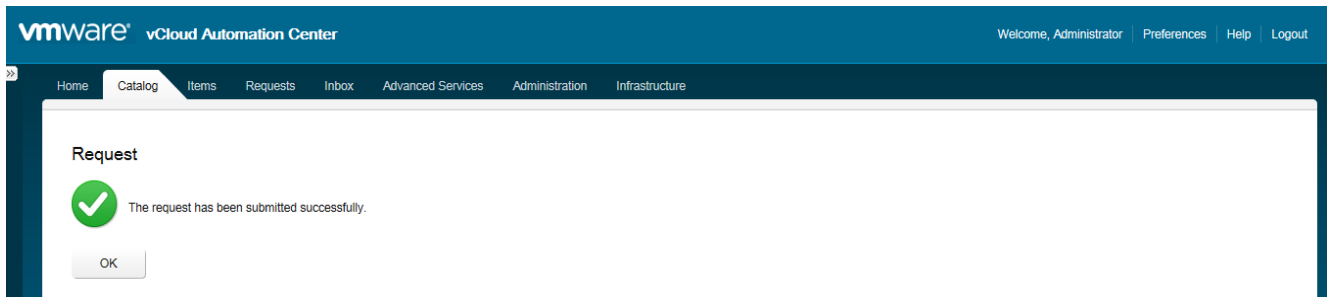
At the bottom right, there are four buttons: '< Back', 'Next >', 'Save', and 'Cancel'. The 'Next >' button is highlighted with a red rectangle.

8. Review all the Client Configuration. If required you can change the default value as per your requirement. Finally click **Submit**.

The screenshot shows the 'New Request' form in the VMware vCloud Automation Center. The 'Client Configuration (OPTIONAL)' tab is selected. The form contains the following fields:

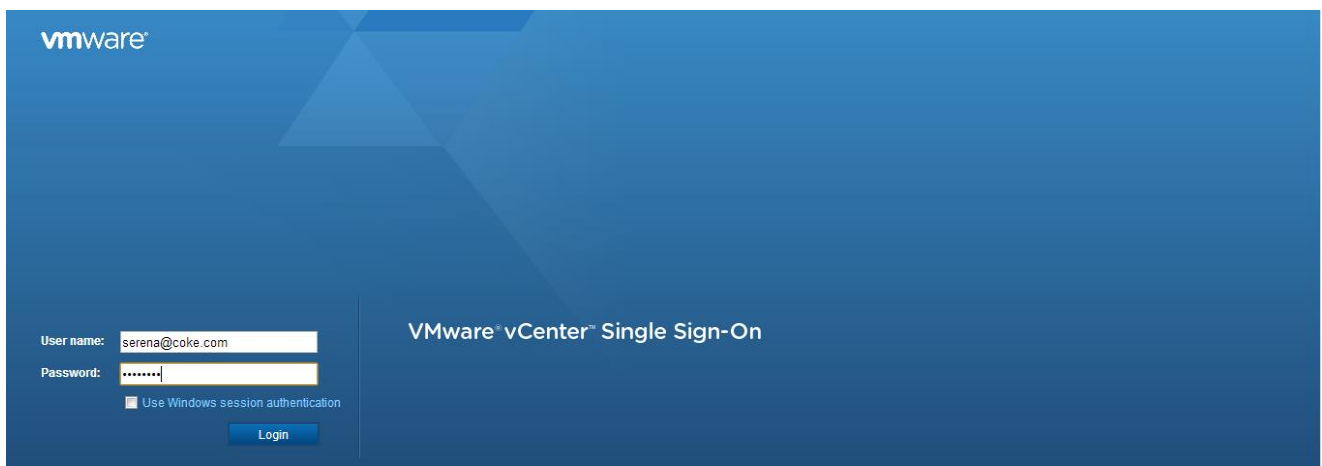
- Client CPU Number: 1
- Client Memory Capacity MB: 2,048
- Client Storage Size GB: 10
- Client Instance Number: 1

At the bottom right, there are four buttons: '< Back', 'Submit', 'Save', and 'Cancel'. The 'Submit' button is highlighted with a red rectangle.

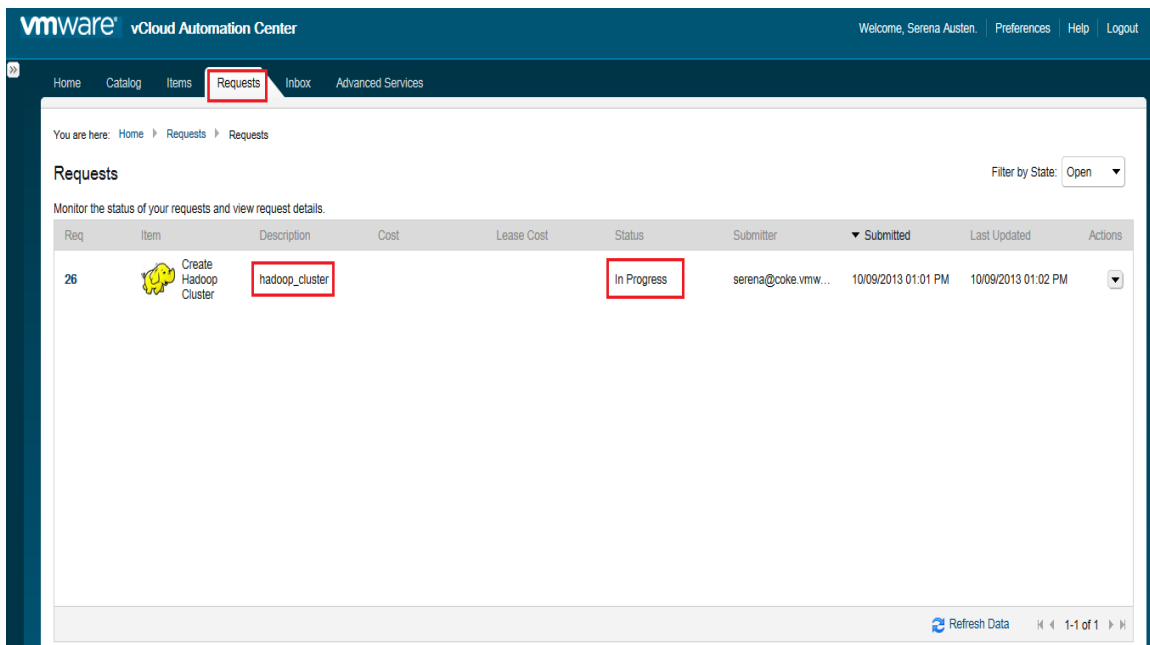


### 3.2 View the status of the application deployment

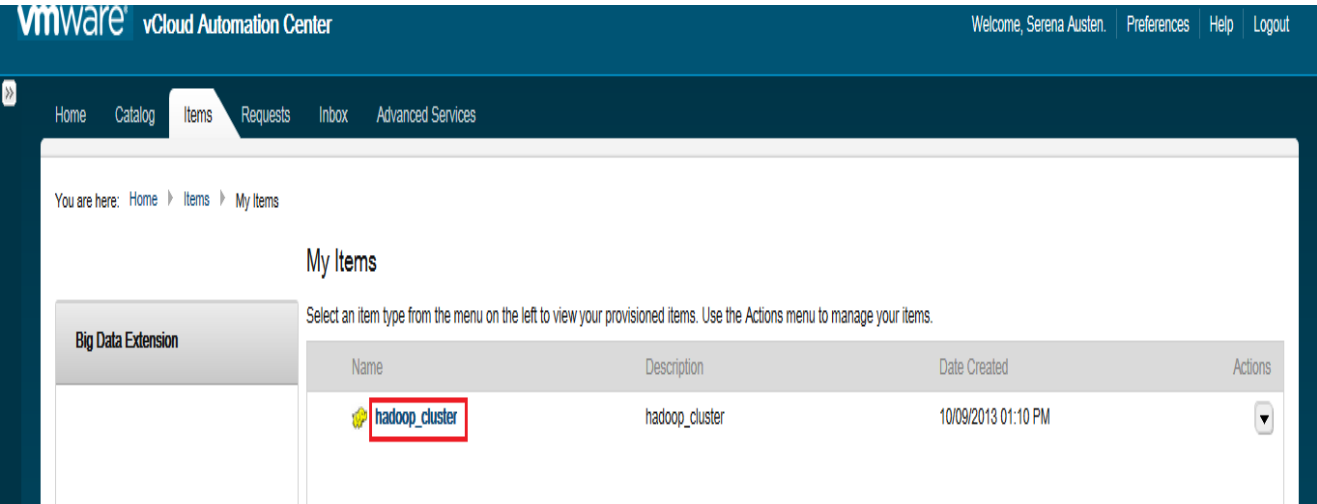
1. Log into vCAC Service Catalog (<https://wdc-auroravcf-gen-dhcp85.eng.vmware.com/shell-ui-app/org/qe/>) using SSO as a Catalog Consumer either tenant admin ([tony@coke.com](mailto:tony@coke.com)) or tenant user ([serena@coke.com](mailto:serena@coke.com)/password)



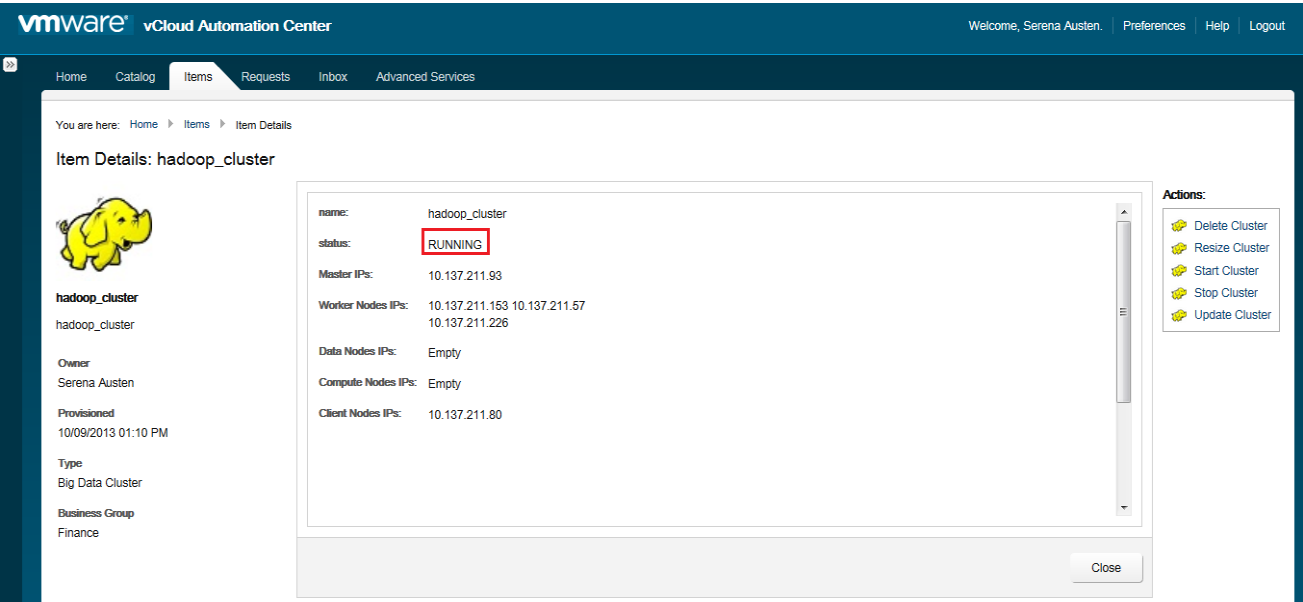
2. Click on the **Requests** tab.



- 3. Wait for a while to complete the deployment.
- 4. Click on the **Items** tab. And click on the **hadoop\_cluster**.



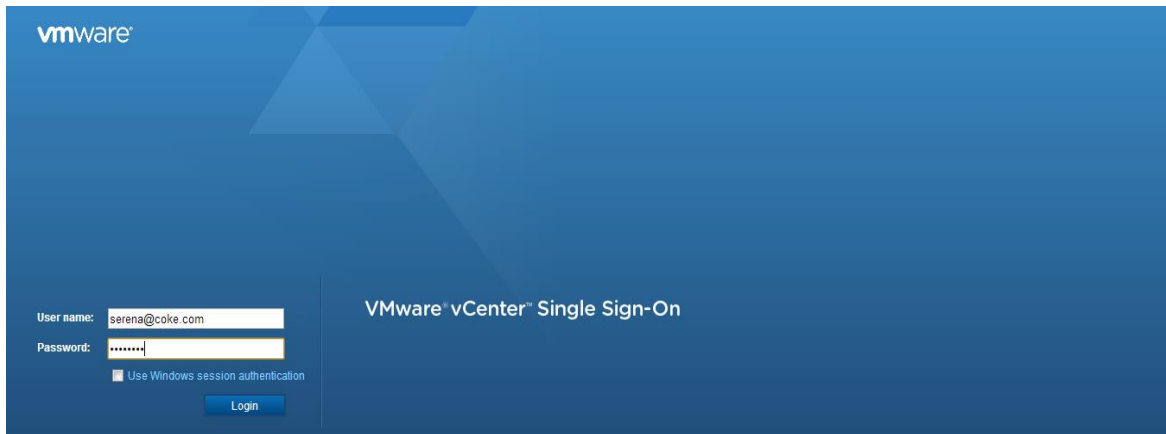
- 5. We can see the deployment status as **RUNNING**



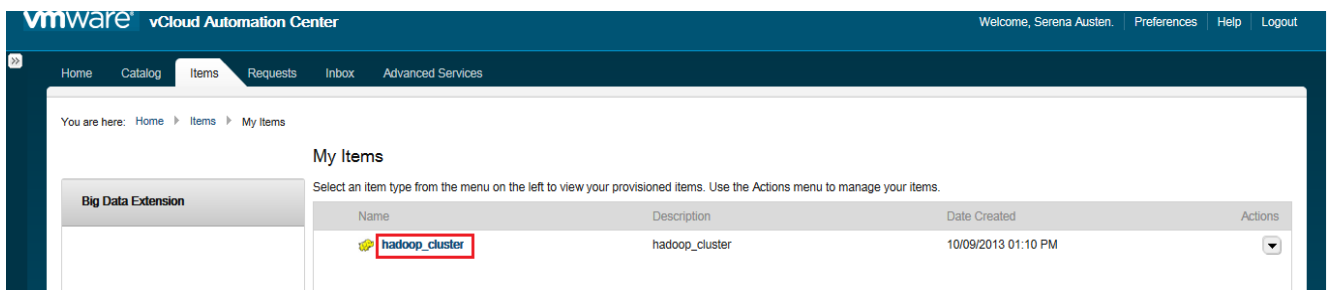


### 3.3 Launching the hadoop service

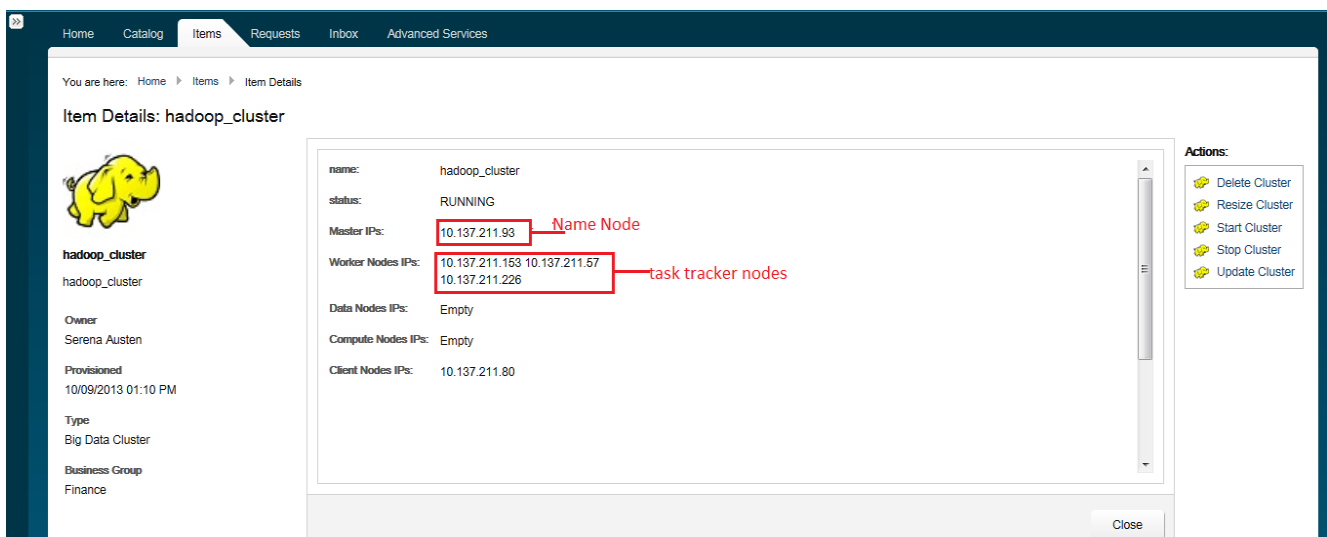
1. Log into vCAC Service Catalog (<https://wdc-auroravcf-gen-dhcp85.eng.vmware.com/shell-ui-app/org/qe/>) using SSO as a Catalog Consumer either tenant admin ([tony@coke.com](mailto:tony@coke.com)) or tenant user ([serena@coke.com](mailto:serena@coke.com)/password)



2. Click on the **Items** tab. And click on the **hadoop\_cluster**.

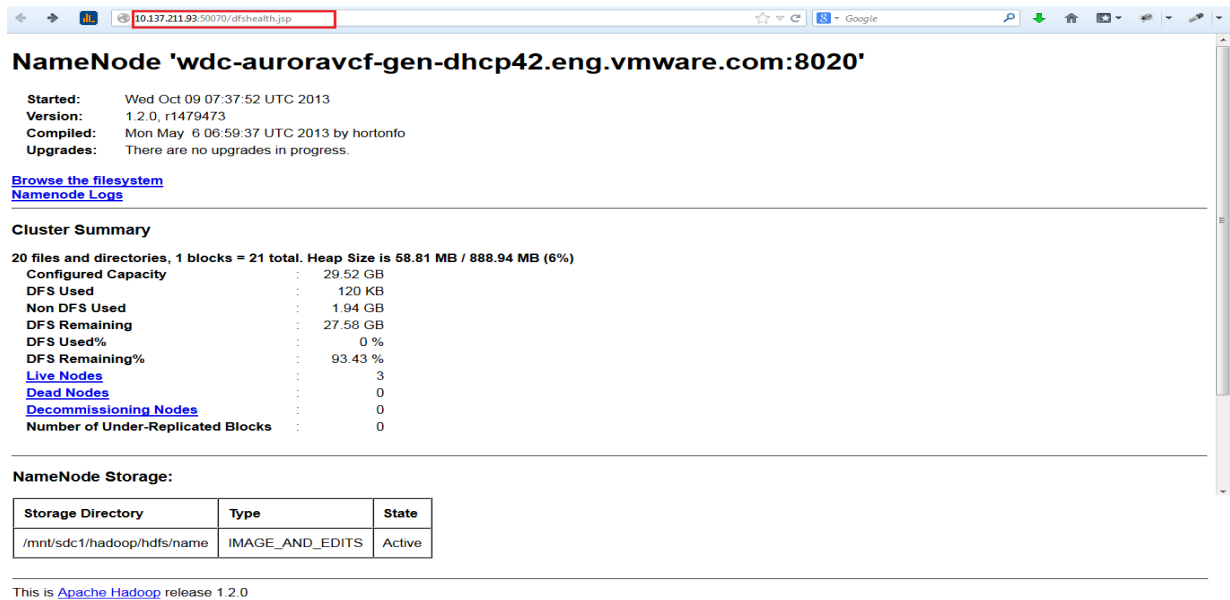


3. We can see the machine's IPs of name node and task tracker nodes



4. Open any browser of your choice and access the following links

- a. [http://<nameNode\\_ip>:50070](http://<nameNode_ip>:50070) for the Name node  
For example: <http://10.137.211.93:50070>



The screenshot shows the NameNode web interface for 'wdc-auroravcf-gen-dhcp42.eng.vmware.com:8020'. It displays metadata such as start time, version, and compilation date. A 'Cluster Summary' section provides details on file/directory counts, capacity, and node status. A 'NameNode Storage' table lists the storage directory, type, and state. The footer indicates it is an Apache Hadoop release 1.2.0.

**NameNode 'wdc-auroravcf-gen-dhcp42.eng.vmware.com:8020'**

Started: Wed Oct 09 07:37:52 UTC 2013  
Version: 1.2.0, r1479473  
Compiled: Mon May 6 06:59:37 UTC 2013 by hortonfo  
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)  
[Namenode Logs](#)

**Cluster Summary**

20 files and directories, 1 blocks = 21 total. Heap Size is 58.81 MB / 888.94 MB (6%)

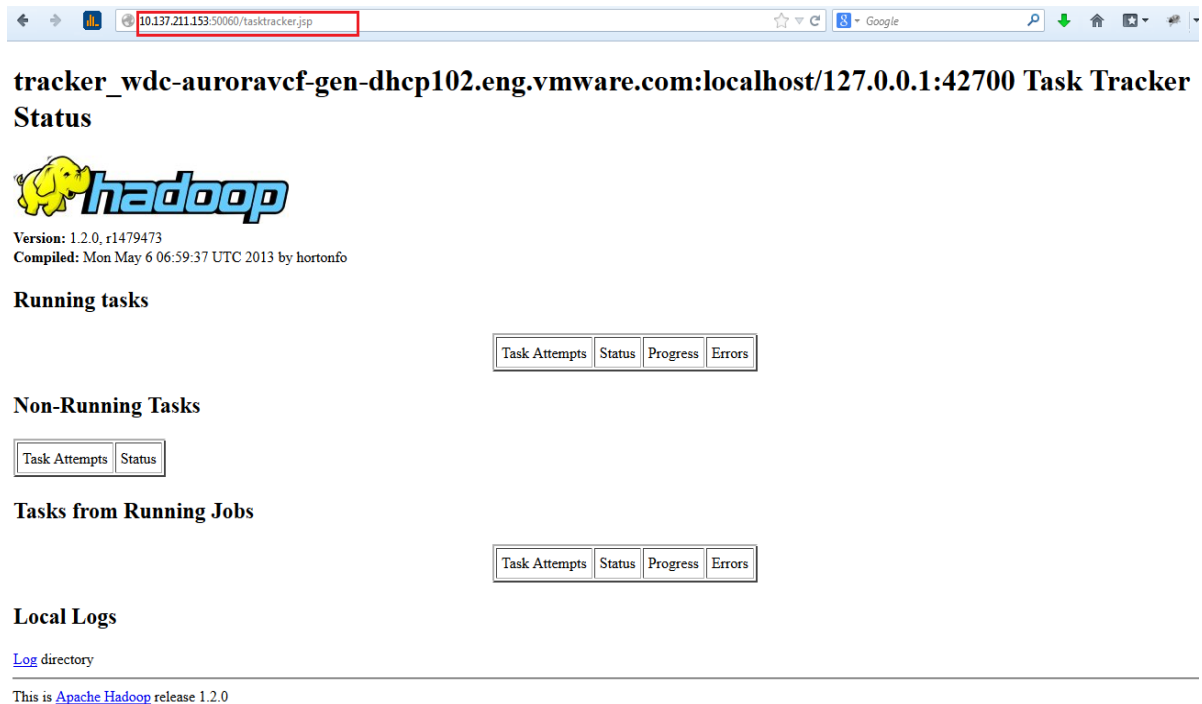
Configured Capacity	: 29.52 GB
DFS Used	: 120 KB
Non DFS Used	: 1.94 GB
DFS Remaining	: 27.58 GB
DFS Used%	: 0 %
DFS Remaining%	: 93.43 %
<a href="#">Live Nodes</a>	: 3
<a href="#">Dead Nodes</a>	: 0
<a href="#">Decommissioning Nodes</a>	: 0
Number of Under-Replicated Blocks	: 0

**NameNode Storage:**

Storage Directory	Type	State
/mnt/sdc1/hadoop/hdfs/name	IMAGE_AND_EDITS	Active


This is [Apache Hadoop](#) release 1.2.0

- b. [http://<taskTracker\\_ip>:50060](http://<taskTracker_ip>:50060) for task tracker.  
For example: <http://10.137.211.153:50060>



The screenshot shows the Task Tracker web interface for 'wdc-auroravcf-gen-dhcp102.eng.vmware.com:localhost/127.0.0.1:42700'. It features the Hadoop logo, version information, and sections for 'Running tasks', 'Non-Running Tasks', 'Tasks from Running Jobs', and 'Local Logs'. Each task section includes buttons for 'Task Attempts', 'Status', 'Progress', and 'Errors'.

**tracker\_wdc-auroravcf-gen-dhcp102.eng.vmware.com:localhost/127.0.0.1:42700 Task Tracker Status**



Version: 1.2.0, r1479473  
Compiled: Mon May 6 06:59:37 UTC 2013 by hortonfo

**Running tasks**

Task Attempts	Status	Progress	Errors
---------------	--------	----------	--------

**Non-Running Tasks**

Task Attempts	Status
---------------	--------

**Tasks from Running Jobs**

Task Attempts	Status	Progress	Errors
---------------	--------	----------	--------

**Local Logs**

[Log](#) directory

This is [Apache Hadoop](#) release 1.2.0

## 4. Summary

Hadoop as a service utilizes industry's leading Cloud-enabled development and process automation platform, VMware **vCenter Orchestrator** and Cloud-enabled self-service provisioning solution available with integration of the VMware **vCloud Automation Center**.

## 5. Benefits

### 5.1 Agility with Performance

- **Rapid Deployment** – Launch Hadoop clusters in minutes by leveraging customizable cluster templates in vCenter.
- **Operational Simplicity** – Proactively monitor the health of your Hadoop clusters as well as eliminate manual processes with intelligent automation.
- **Performance** – Performance benchmark testing shows on par performance when compared to physical deployments depending on configuration.

### 5.2 Multi-Tenancy and Elastic Scaling

- **True Multi-tenancy** – Separating data from compute allows for seamless scaling of the compute layer while keeping data persistent and safe. Users can run mixed workloads simultaneously on a single physical host.
- **Automated Resource Rebalancing** – Pre-specify ranges to elastically shrink and expand clusters. Your mission critical Hadoop jobs will automatically get the resources they need prioritized.
- **Reduce Hardware Costs** – Avoid the costs related to building and operating separate physical clusters with dedicated hardware. Pool compute and storage resources on a common virtual platform to increase hardware utilization.

### 5.3 Reliability and Security

- **VM-based Isolation** – Ensure you have reserve resources to meet your needs and run concurrent applications or Hadoop distributions. Provide privacy and data isolation between multiple users of your Hadoop cluster.
- **High Availability** – One-click failover protection against hardware and operating system failures will allow your Hadoop jobs to restart where they left off.