Hadoop as a Service

VMware vCloud Automation Center & Big Data Extension

Table of Contents

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

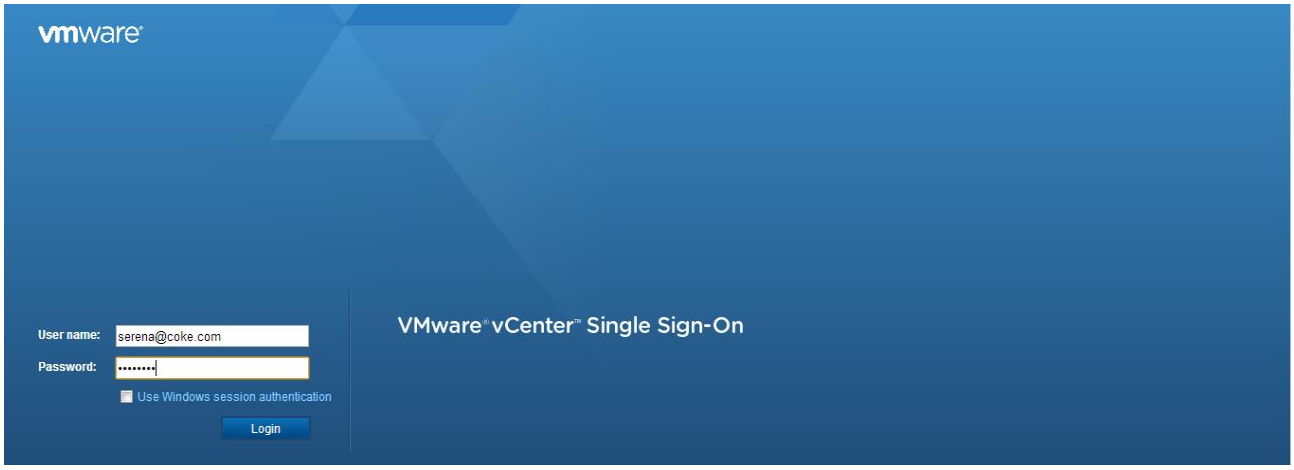
1. Install and configure vCO 6.0
2. Install and Configure vCAC 6.0
3. Install and configure VMware vCO 6.0 with vCAC 6.0
4. Big Data Extensions configuration with vSphere 5.5

3. Set up

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

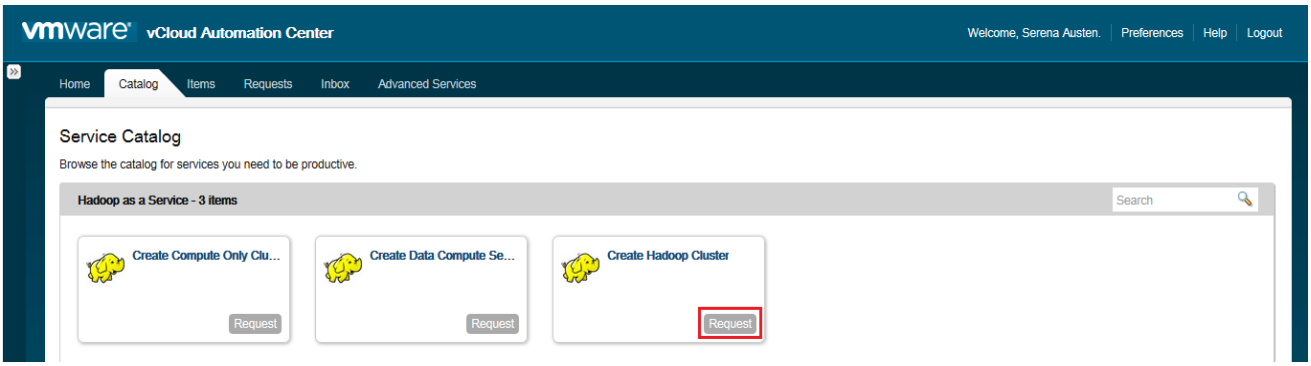
3.1 Request the Service as a Catalog Consumer

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) or tenant user (serena@coke.com/password)

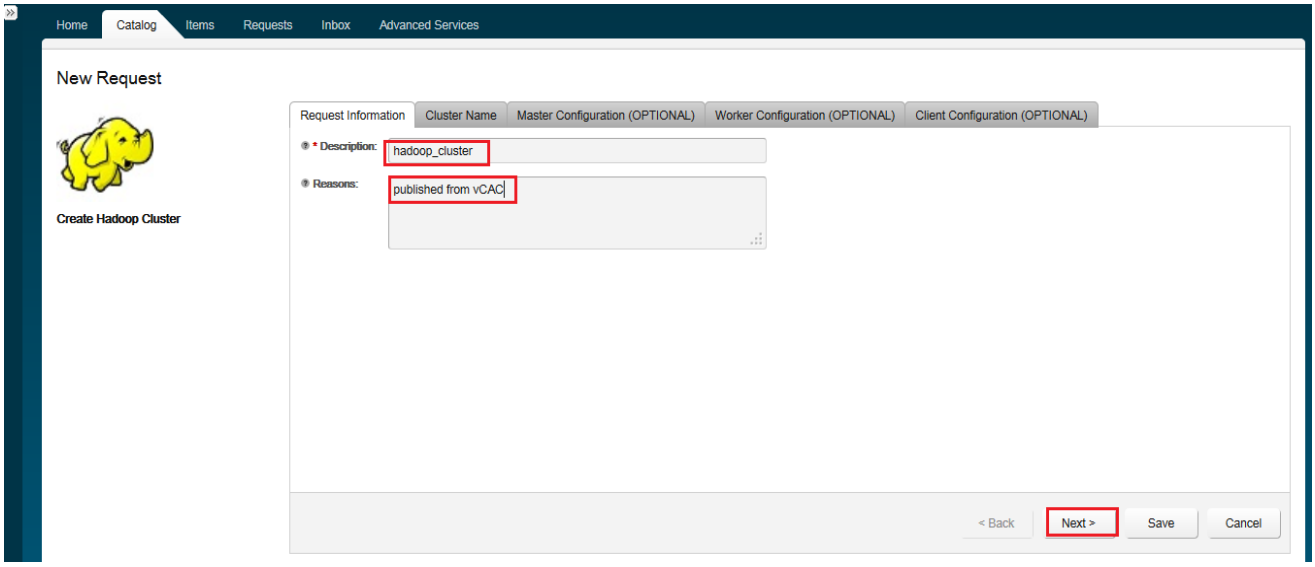


2. Click on the **Catalog** tab

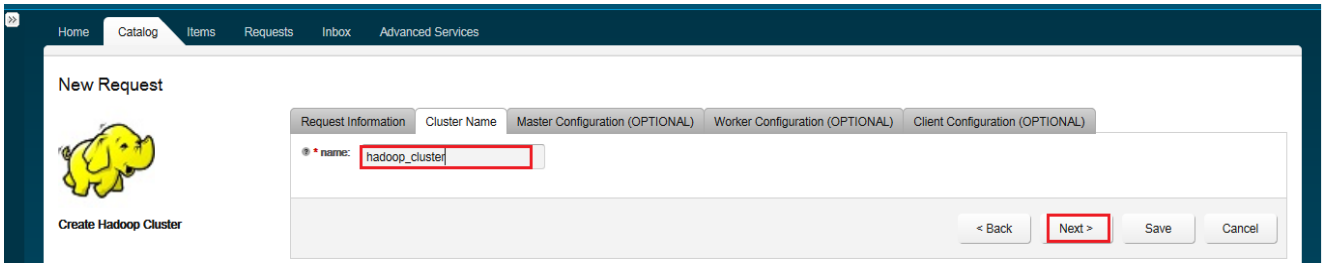
3. There are list of catalog services. Click on the **Request** Button for "**Create Hadoop Cluster**"



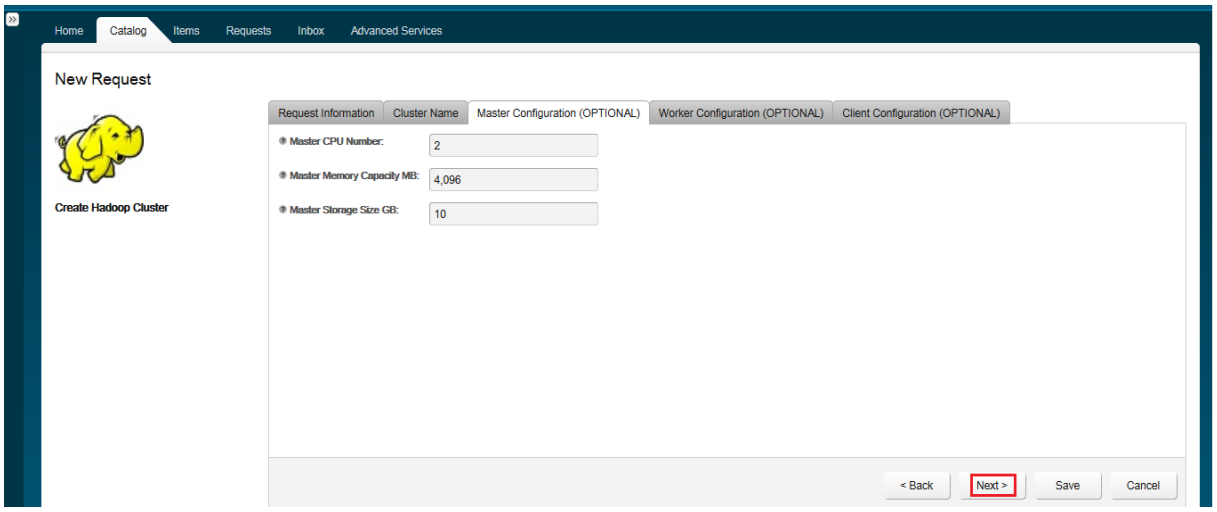
4. Provide Description & Details and click **next**.



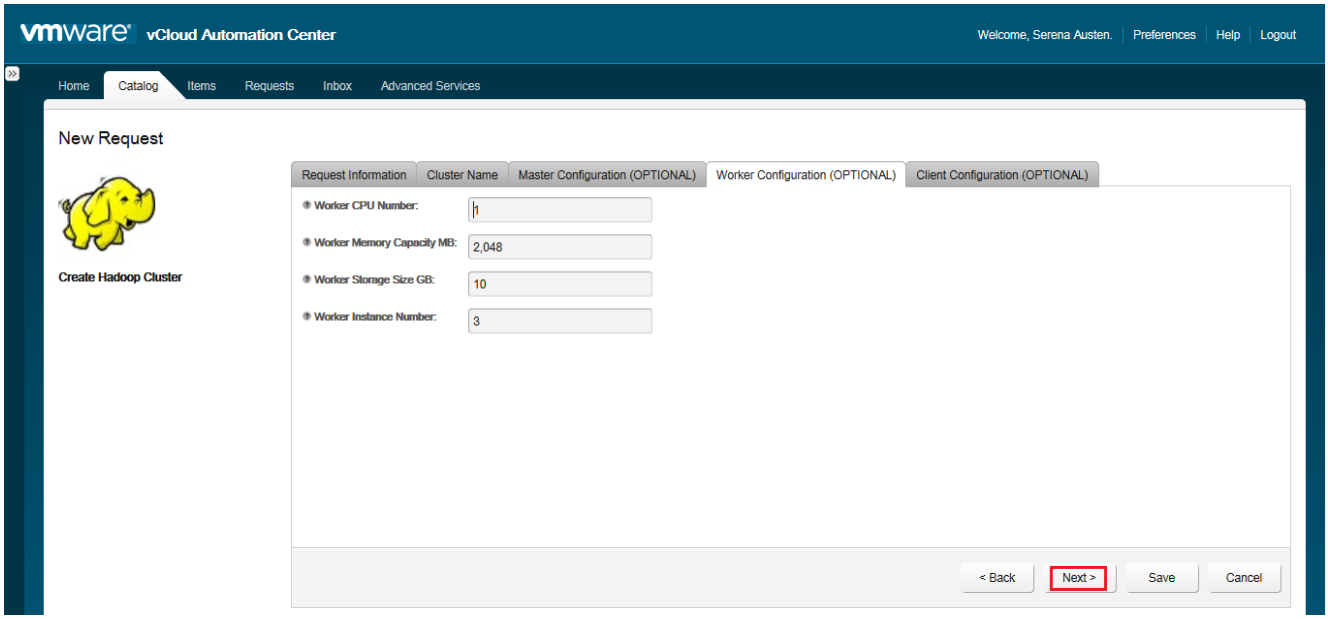
5. Enter the cluster name to be created and click **Next**



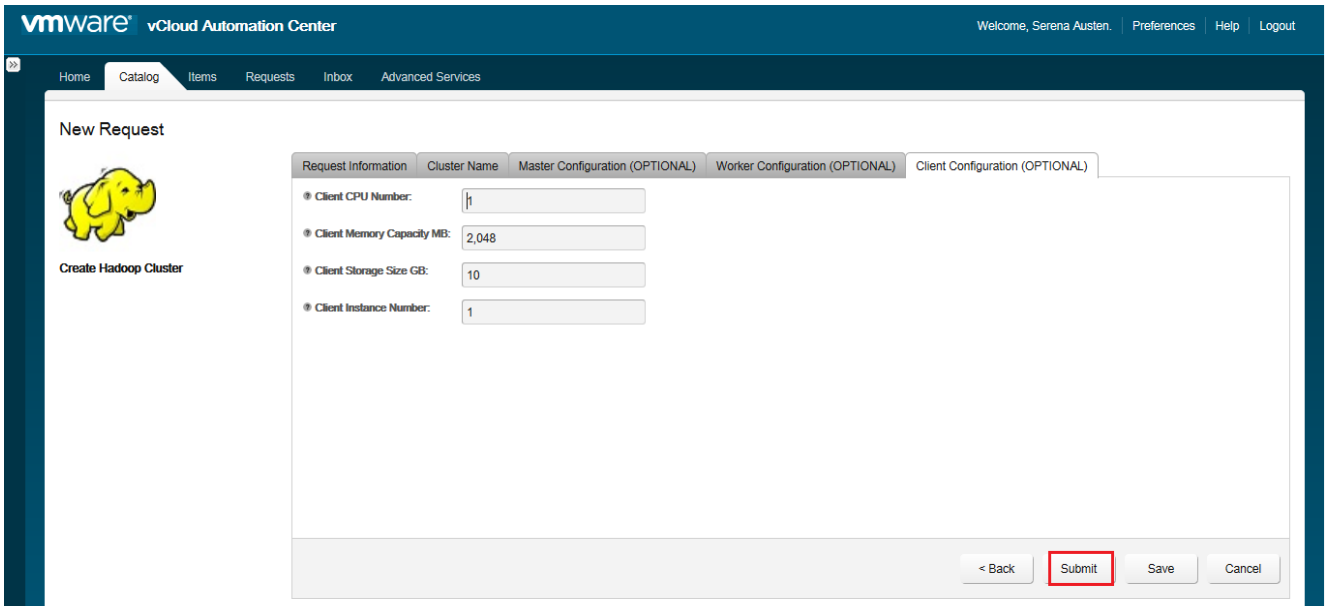
6. Review all the Master Configuration. If required you can change the default value as per your requirement. Click **Next**.

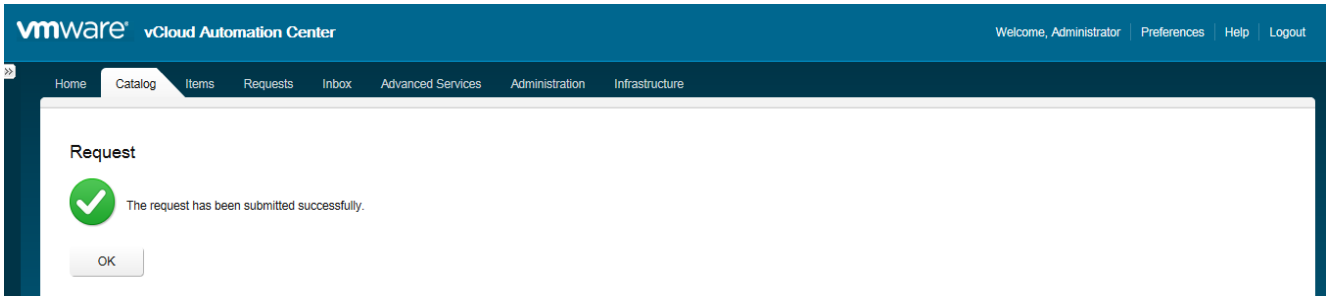


7. Review all the Worker Configuration. If required you can change the default value as per your requirement. Click **Next**.



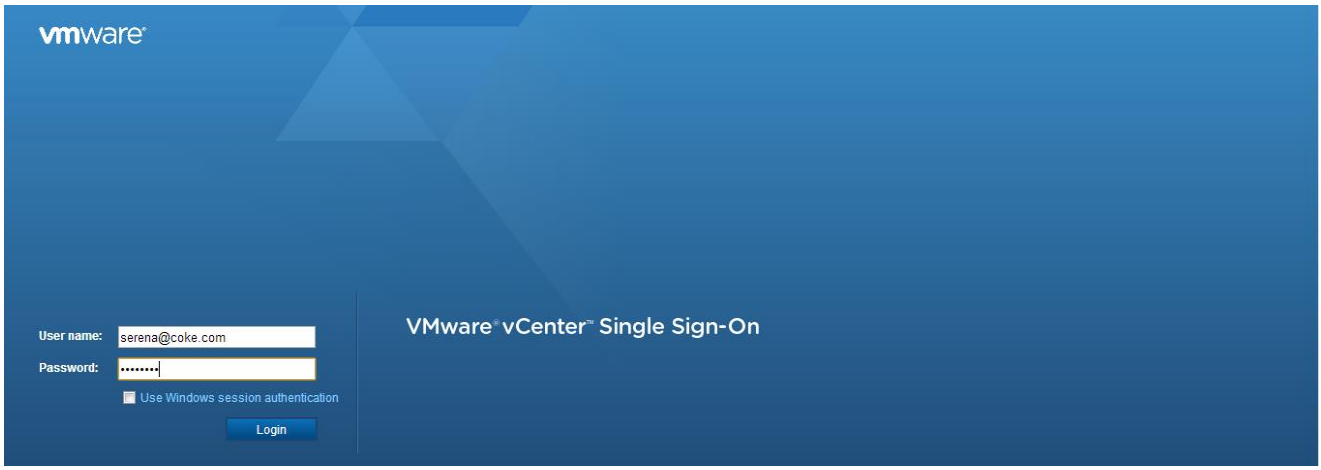
8. Review all the Client Configuration. If required you can change the default value as per your requirement. Finally click **Submit**.



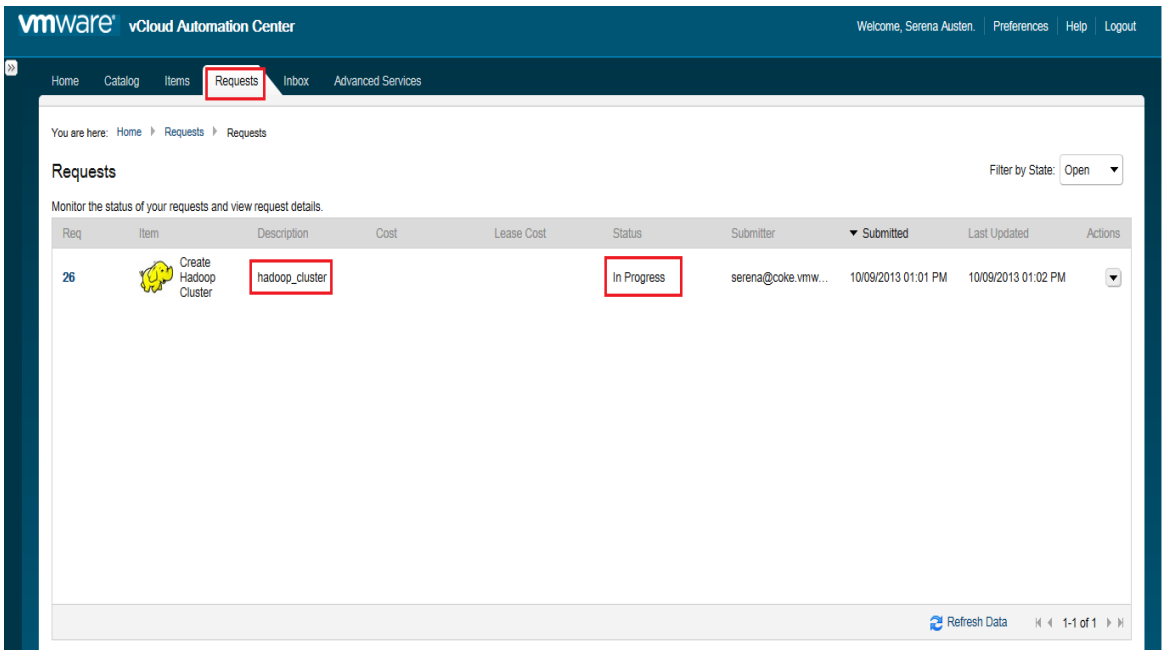


3.2 View the status of the application deployment

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) or tenant user (serena@coke.com/password)

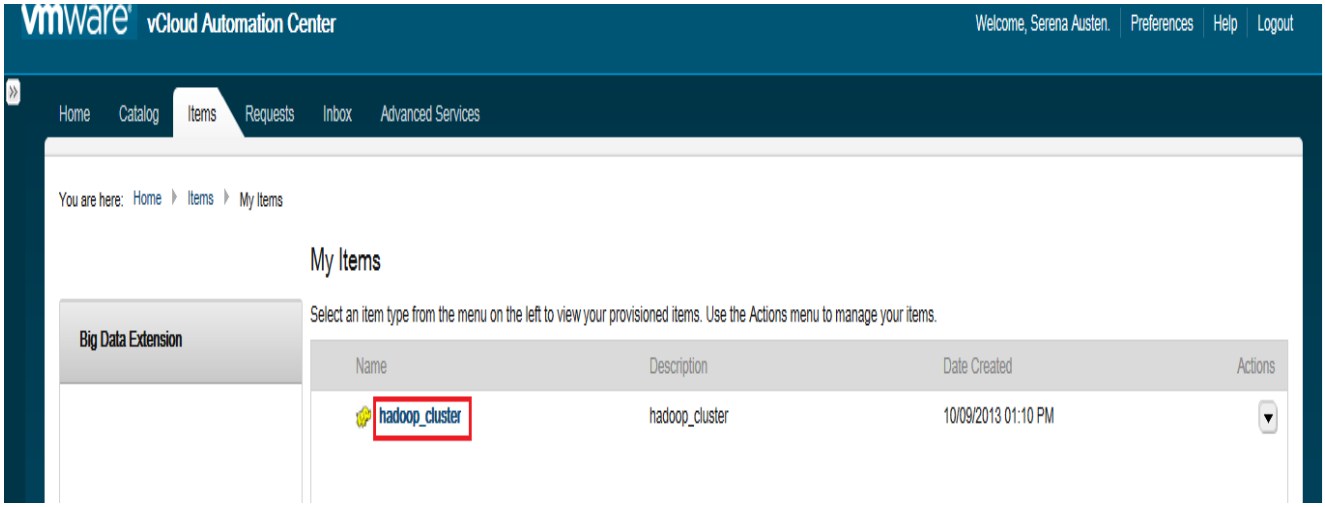


2. Click on the **Requests** tab.

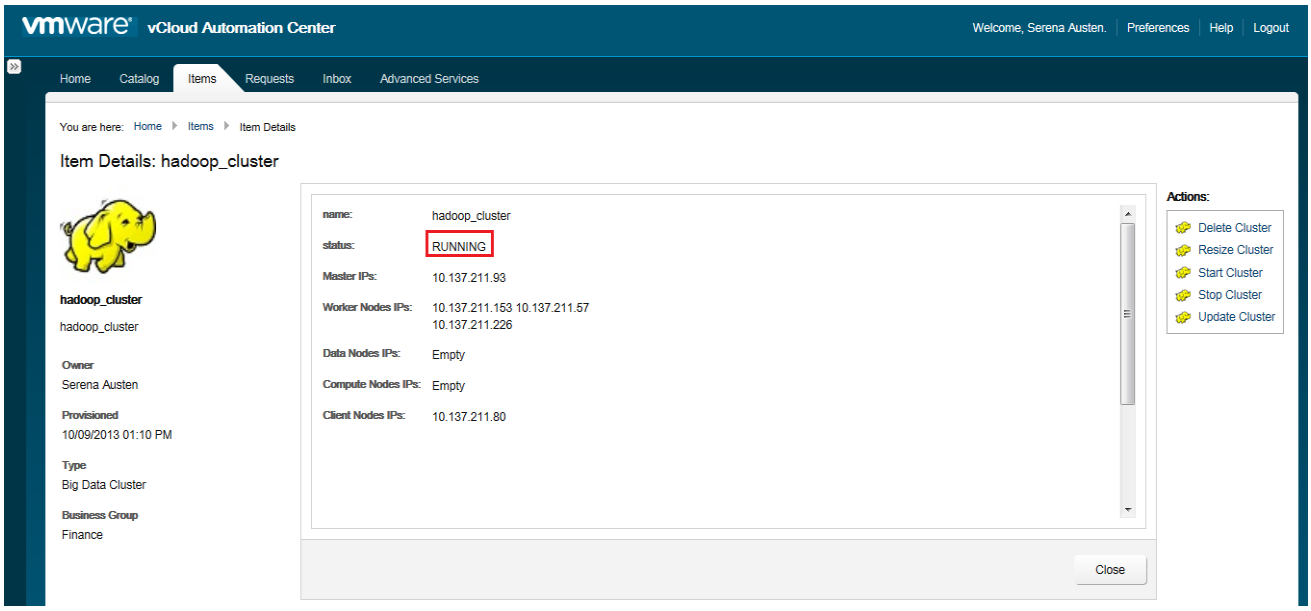


3. Wait for a while to complete the deployment.

4. Click on the **Items** tab. And click on the **hadoop_cluster**.

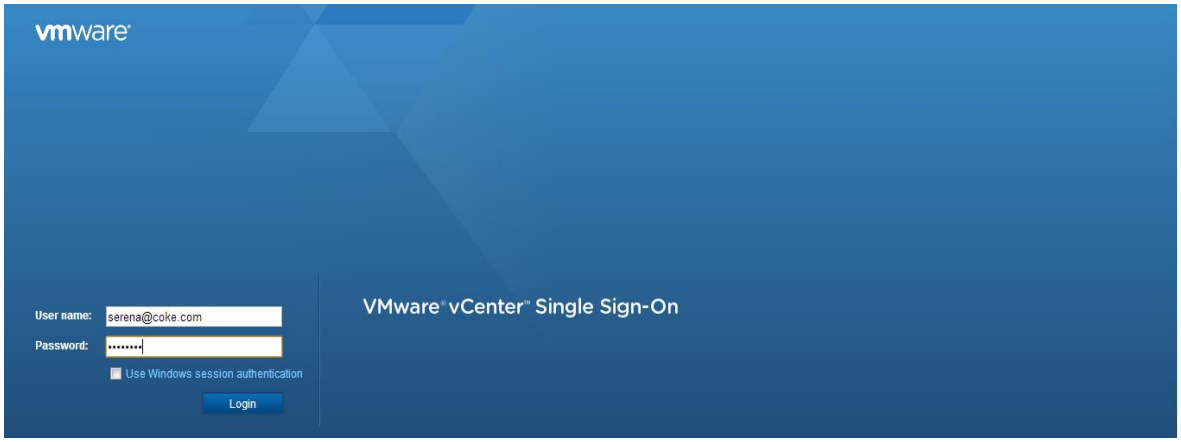


5. We can see the deployment status as RUNNING

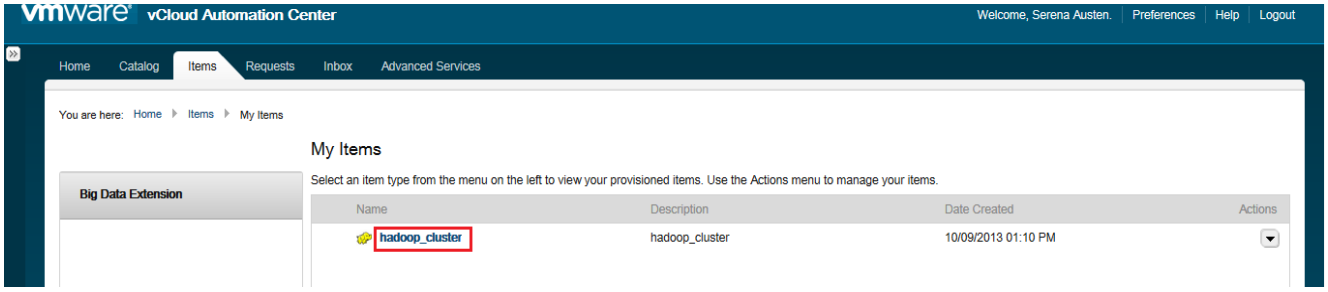


3.3 Launching the hadoop service

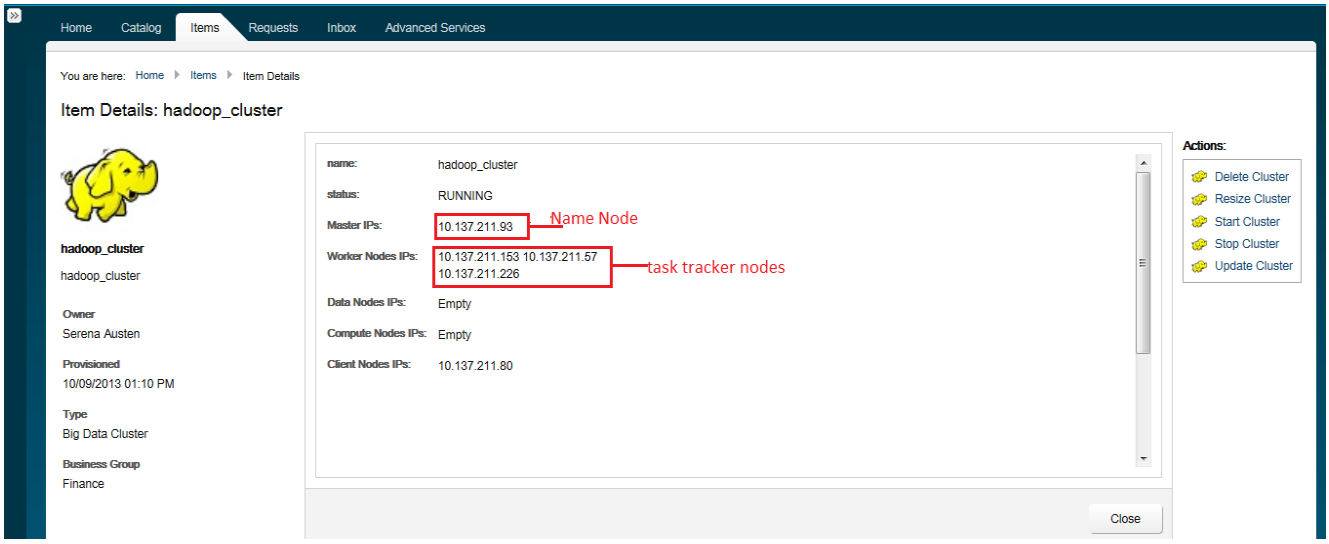
1. Log into vCAC Service Catalog (https://wdc-auroravcf-gen-dhcp85.eng.vmware.com/shell-ui-app/org/qe/) using SSO as a Catalog Consumer either tenant admin (tony@coke.com) or tenant user (serena@coke.com/password)



2. Click on the **Items** tab. And click on the **hadoop_cluster**.

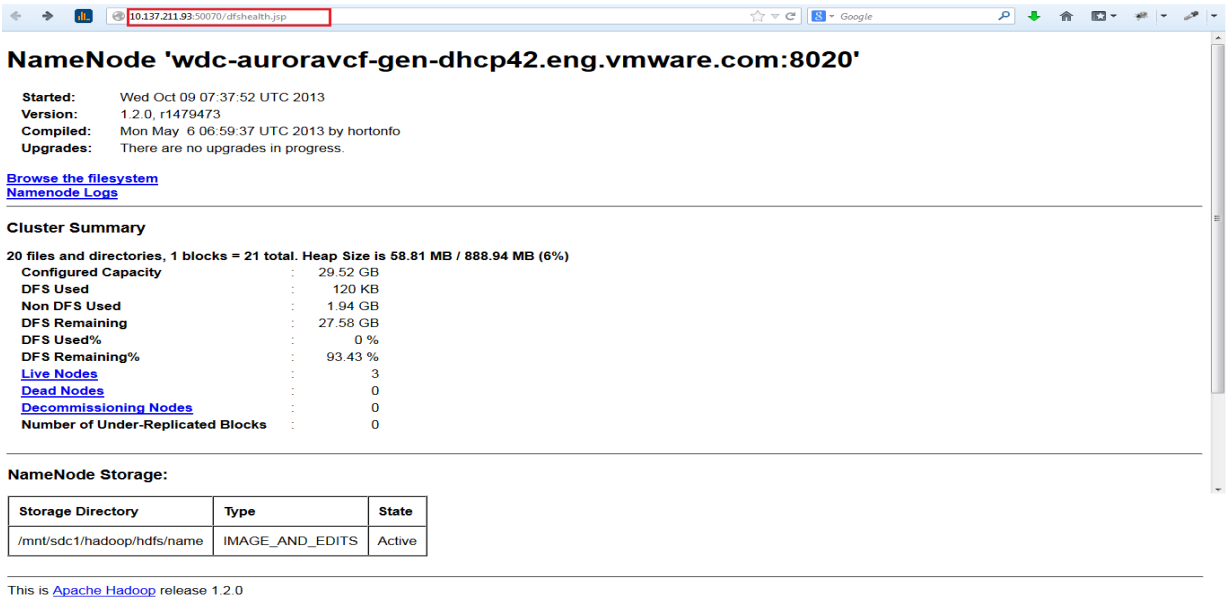


3. We can see the machine's IPs of name node and task tracker nodes

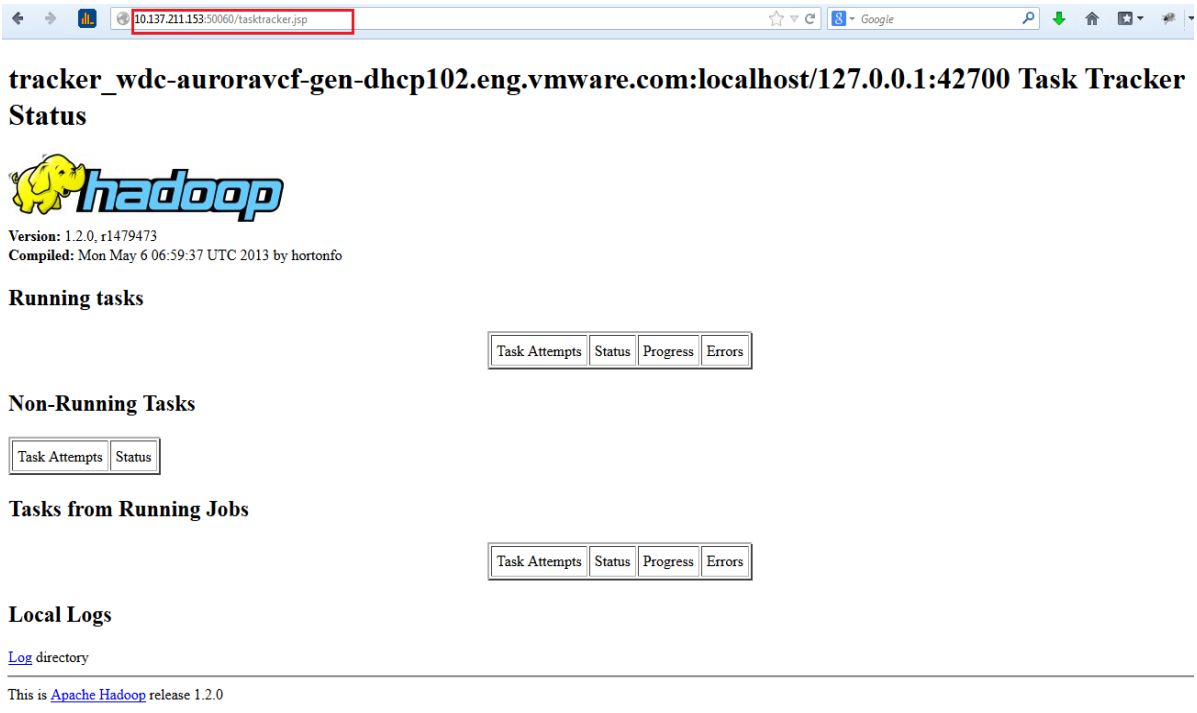


4. Open any browser of your choice and access the following links

a. http://<nameNode_ip>:50070 for the Name node
For example: http://10.137.211.93:50070



b. http://<taskTracket_ip>:50060 for task tracker.
For example: http://10.137.211.153:50060



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.