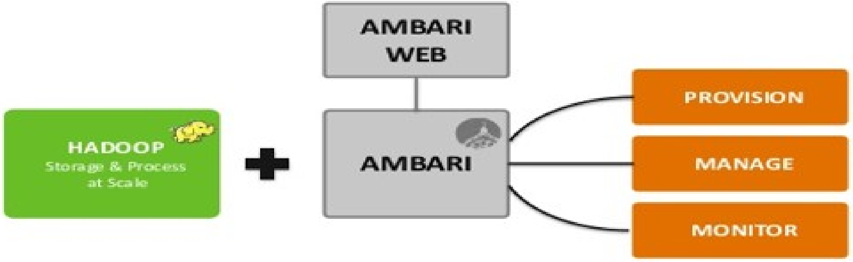
AMBARI

- ✓ What is Ambari?
- ✓ What is Hadoop?
- ✓ Types of managing tools
- ✓ Architecture of Ambari
- ✓ How to install Ambari?
- ✓ Setting up of Hadoop cluster

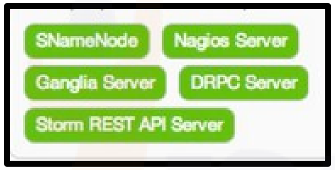
What is Ambari?

Apache Ambari is a tool for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari consists of a set of RESTful APIs and a browser-based management interface.

Ambari operations:

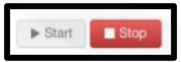


Provision:



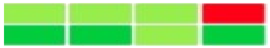
- Virtual, physical and cloud Environments.
- Deploy 10s,100s,1000s of Hadoop servers

Manage:-



- Advance configuration & host Controls.
- Single point for Host controls.

Monitor:-



- ✓ Pre-configuration metrics and alerts.
- ✓ Single pane of glass for Hadoop & system status.

What is Hadoop?

- Hadoop is a large-scale and distributed data storage and processing infrastructure using clusters of commodity hosts networked together.

- Monitoring and managing such complex distributed systems is a non-trivial task.

- To help you manage the complexity, Apache Ambari collects a wide range of information from the cluster's nodes and services and presents it to you in an easy-to-read and use, centralized web interface, Ambari Web.

What is Hadoop?

- Ambari Web displays information such as service-specific summaries, graphs, and alerts.

- You use Ambari Web to create and manage your HDP cluster and to perform basic operational tasks such as starting and stopping services, adding hosts to your cluster, and updating service configurations.

- You also can use Ambari Web to perform administrative tasks for your cluster such as enabling Kerberos security and performing Stack upgrades.

What are the types of managing tools?

Types of Managing Tools:

- Web Based data colloction → Nutch, Solr, Gora, Hbase
- Mapreduce Programming→ Fair and Capacity schedulers, Oozie
- Moving data→ Hadoop Commands, sqoop, flume, storm
- Monitoring→ Hue, Nagios ,Ganglia
- Analysis with sql → Impala, hive, spark
- ETL→ Pentaho, Talend
- Reporting--> Splunk, Talend

One liner management tool explanation:-

Nutch:-

Apache Nutch is a highly extensible and scalable open source web crawler software project.

Gora:-

The Apache Gora open source framework provides an in-memory data model and persistence for big data. Gora supports persisting to column stores, key value stores, document stores and RDBMSs, and analyzing the data with extensive Apache Hadoop MapReduce.

Solr:-

Apache solr is a standalone full text search server with Apache Lucene at the backend, User for Web application for text search, A Wrapper around Apache Lucene, Written at Cnet, now at Apache.

Hbase:-

HBase is a NoSQL databases which experienced a tremendous increase in popularity during recent years.

One liner management tool explanation:-

Oozi:-

Provides workflow management and coordination of those workflows, Manages Directed Acyclic graph of Actions.

Sqoop:-

sqoop is a tool to transfer data between Hadoop and relational database, Transform data in Hadoop with Mapreduce or hive, Export data back in to RDB.

Nagios:-

This is a system and network monitoring tool.

Hue:-

Hue is a lightweight Web server that help you to use hadoop directly from your browser. Hue is just a view on top of any Hadoop distribution and can be installed on any machine.

One liner management tool explanation:-

Imapala:-

General purpose sql query Engine, works both for analytical and transactional/Single row workloads.

Talend:-

Eclipse-base visual programming Editor, generates executable java code. Talend to bring an open source integration tool for easily connecting Apache Hadoop to hundreds of data systems without having to write code.

Now, let us talk about Minimum System Requirements.

Minimum System Requirements:-

Hardware Recommendations:-

There is no single set of hardware recommendations for installing Hadoop.

Operating Systems Requirements:-

Red Hat Enterprise Linux (RHEL) v5.x or 6.x (64-bit)

- CentOS v5.x or 6.x (64-bit)
- SUSE Linux Enterprise Server (SLES) 11, SP1 (64-bit)

Browser Requirements:-

The Ambari Install Wizard runs as a browser-based Web app. You must have a machine capable of running a graphical browser to use this tool.

The supported browsers are:

- Mac OS X_10.6 or later
  - Firefox latest stable release

Minimum System Requirements:-

- Safari latest stable release
- Google Chrome latest stable release

- Linux_RHEL, CentOS
  - Firefox latest stable release
  - Google Chrome latest stable release
- Windows Vista, 7
  - Internet Explorer 9.0 and higher
  - Firefox latest stable release
  - Safari latest stable release
  - Google Chrome latest stable release

Software Requirements:-

On each of your hosts:

- yum
- rpm
- scp
- curl
- wget
- pdsh

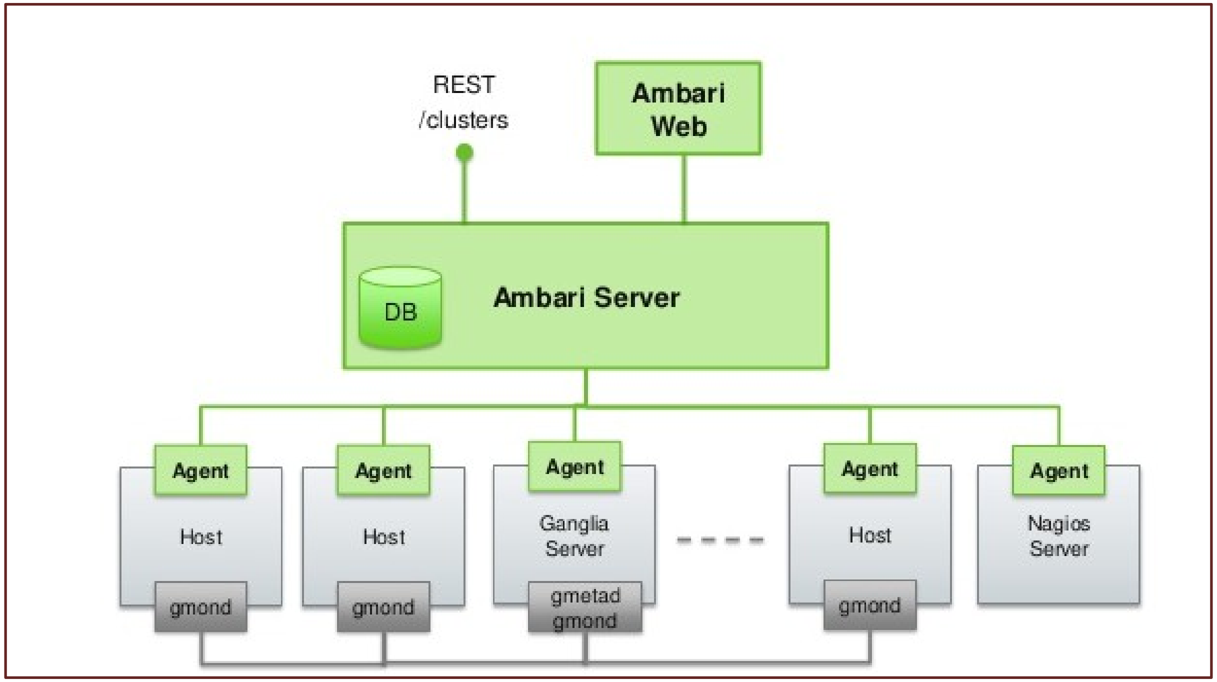
- Database Requirements:-

Hive or HCatalog requires a MySQL database for its use. You can choose to use a current instance or Let the Ambari install wizard create one for you.

Let us know about the architecture of Ambari.

Ambari Architecture:-

Ambari system Architecture:-



Ambari have two components. They are:

- Ambari server –

Master process which communicates with Ambari agents installed on each node participating in the cluster. This has postgres database instance which is used to maintain all cluster related metadata.

Ambari Architecture:-

Ambari Agent–

These are acting agents for Ambari on each node. Every agent periodically sends his own health status along with different metrics, installed services status and many more things. According master decides on next action and conveys back to the agent to act.

Ambari Installation:

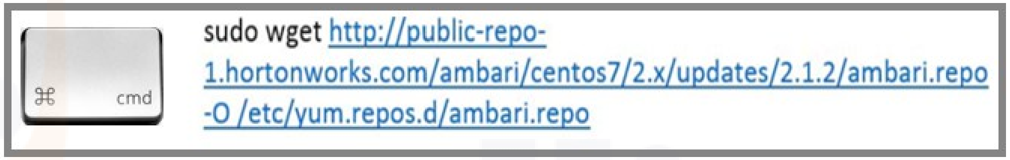
- Ambari installation is easy a task of few commands.
- We will cover Ambari installation and cluster setup.
- We are assumed of having 4 nodes. Node1, Node2, Node3 and Node4. And we are picking Node1 as our Ambari server.
- These are installation steps on the RHEL based system, for debian and other systems steps will vary little.

Let us see how to install Ambari

Installation of Ambari:-

From Ambari server node (Node 1 as we decided)

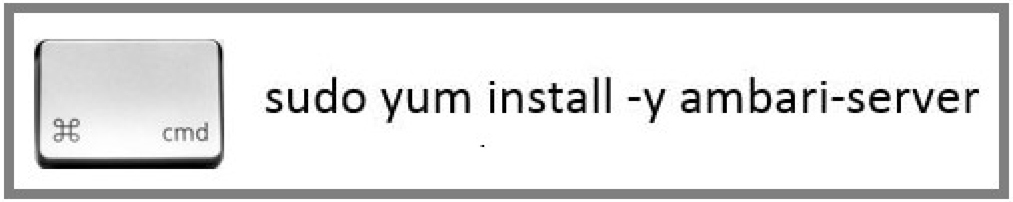
i. **Download Ambari public repo**



```
sudo wget http://public-repo-1.hortonworks.com/ambari/centos7/2.x/updates/2.1.2/ambari.repo -O /etc/yum.repos.d/ambari.repo
```

This command will add Hortonworks Ambari repository into yum which is a default package manager for RHEL systems.

ii. **Install Ambari RPMS**



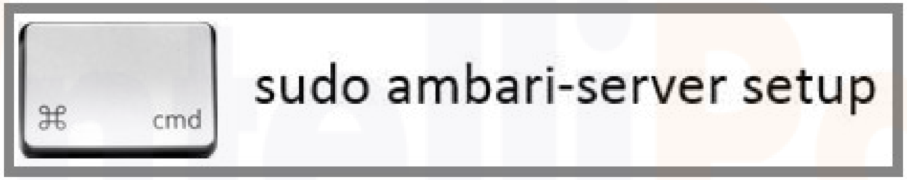
```
sudo yum install -y ambari-server
```

Installation of Ambari:-

iii. Configuring Ambari server

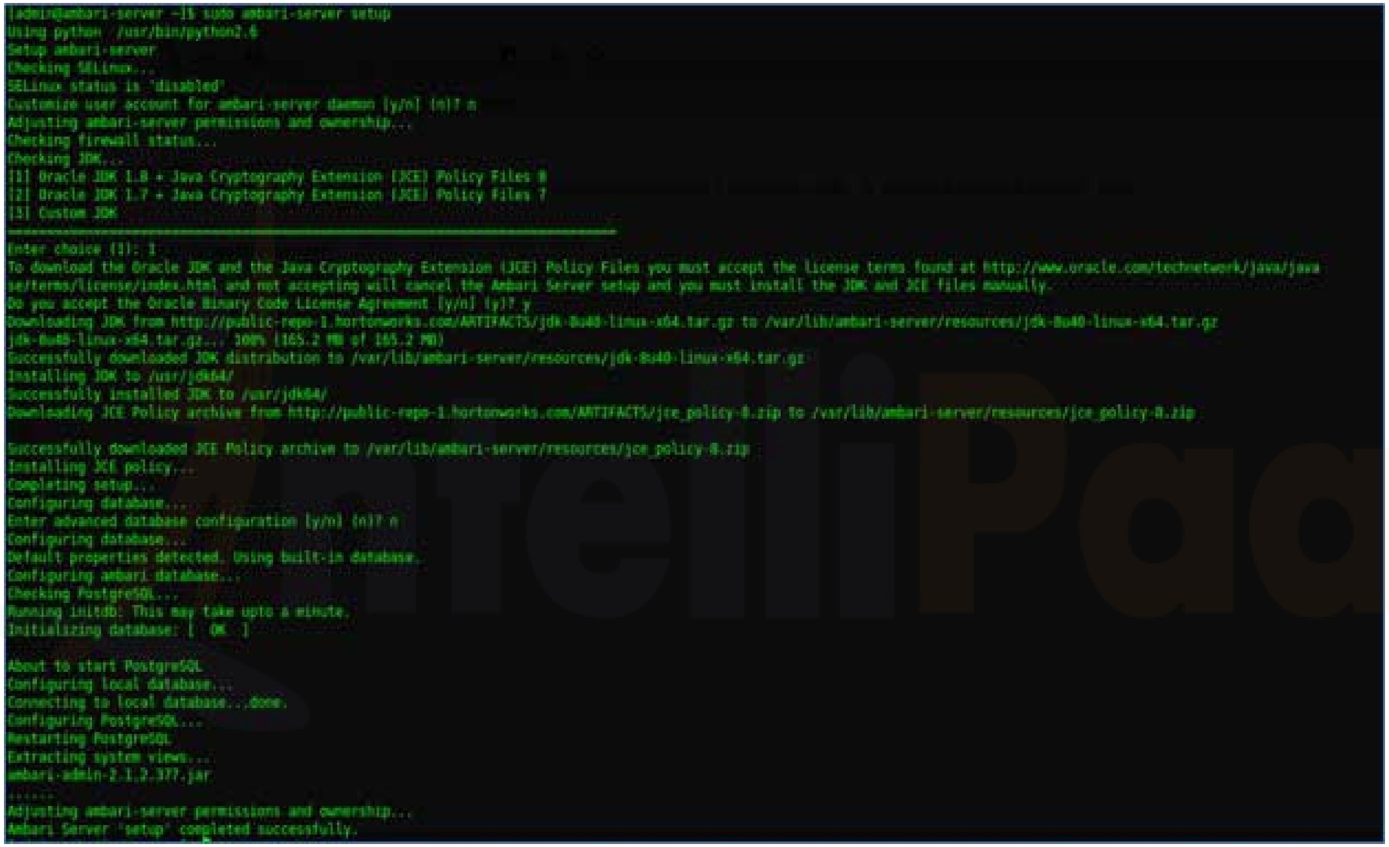
The next thing to do after Ambari installation is to configure Ambari and set it up to provision the cluster.

Following step will take care of this:-



```
sudo ambari-server setup
```

Installation of Ambari:-



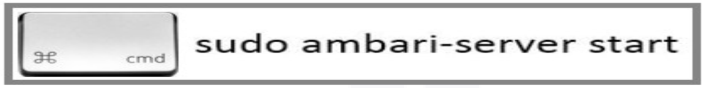
```
[admin@ambari-server ~]$ sudo ambari-server setup
Using python  /usr/bin/python2.6
Setup ambari-server
Checking SELinux...
SELinux status is 'disabled'
Customize user account for ambari-server daemon [y/n] (n)? n
Adjusting ambari-server permissions and ownership...
Checking firewall status...
Checking JDK...
[1] Oracle JDK 1.8 + Java Cryptography Extension (JCE) Policy Files 8
[2] Oracle JDK 1.7 + Java Cryptography Extension (JCE) Policy Files 7
[3] Custom JDK
==============================================================================
Enter choice (1): 1
To download the Oracle JDK and the Java Cryptography Extension (JCE) Policy Files you must accept the license terms found at http://www.oracle.com/technetwork/java/java
se/terms/license/index.html and not accepting will cancel the Ambari Server setup and you must install the JDK and JCE files manually.
Do you accept the Oracle Binary Code License Agreement [y/n] (y)? y
Downloading JDK from http://public-repo-1.hortonworks.com/ARTIFACTS/jdk-8u40-linux-x64.tar.gz to /var/lib/ambari-server/resources/jdk-8u40-linux-x64.tar.gz
jdk-8u40-linux-x64.tar.gz... 100% (165.2 MB of 165.2 MB)
Successfully downloaded JDK distribution to /var/lib/ambari-server/resources/jdk-8u40-linux-x64.tar.gz
Installing JDK to /usr/jdk64/
Successfully installed JDK to /usr/jdk64/
Downloading JCE Policy archive from http://public-repo-1.hortonworks.com/ARTIFACTS/jce_policy-8.zip to /var/lib/ambari-server/resources/jce_policy-8.zip

Successfully downloaded JCE Policy archive to /var/lib/ambari-server/resources/jce_policy-8.zip
Installing JCE policy...
Completing setup...
Configuring database...
Enter advanced database configuration [y/n] (n)? n
Configuring database...
Default properties detected. Using built-in database.
Configuring ambari database...
Checking PostgreSQL...
Running initdb: This may take upto a minute.
Initializing database: [  OK  ]

About to start PostgreSQL
Configuring local database...
Connecting to local database...done.
Configuring PostgreSQL...
Restarting PostgreSQL
Extracting system views...
ambari-admin-2.1.2.377.jar
......
Adjusting ambari-server permissions and ownership...
Ambari Server 'setup' completed successfully.
```

Wish to go with the default options which we do often use? Use-silent option.

Installation of Ambari:-

iv. Start the server and Login to web UI

Start the server with



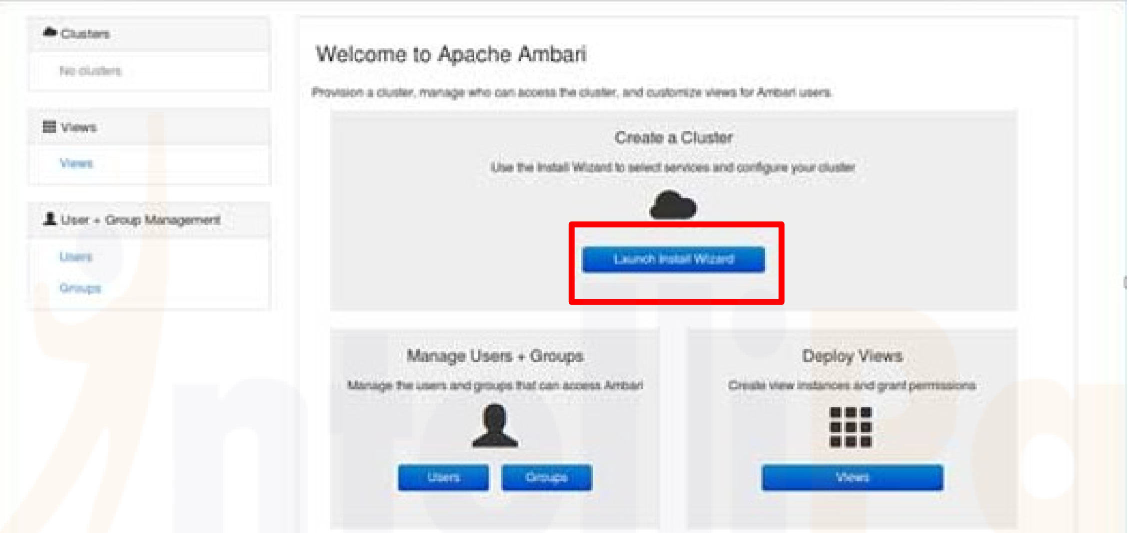
```
sudo ambari-server start
```

- Now we can access Ambari web UI (hosted on 8080 port).
- Login into Ambari with default username "admin" and default password "admin".

How to setup Hadoop Cluster?

Setting up Hadoop cluster:-

I. **Landing page:**



Welcome to Apache Ambari

Provision a cluster, manage who can access the cluster, and customize views for Ambari users.

Create a Cluster

Use the Install Wizard to select services and configure your cluster

Manage Users + Groups

Manage the users and groups that can access Ambari

Deploy Views

Create view instances and grant permissions

Click on "Launch Install Wizard" to start cluster setup

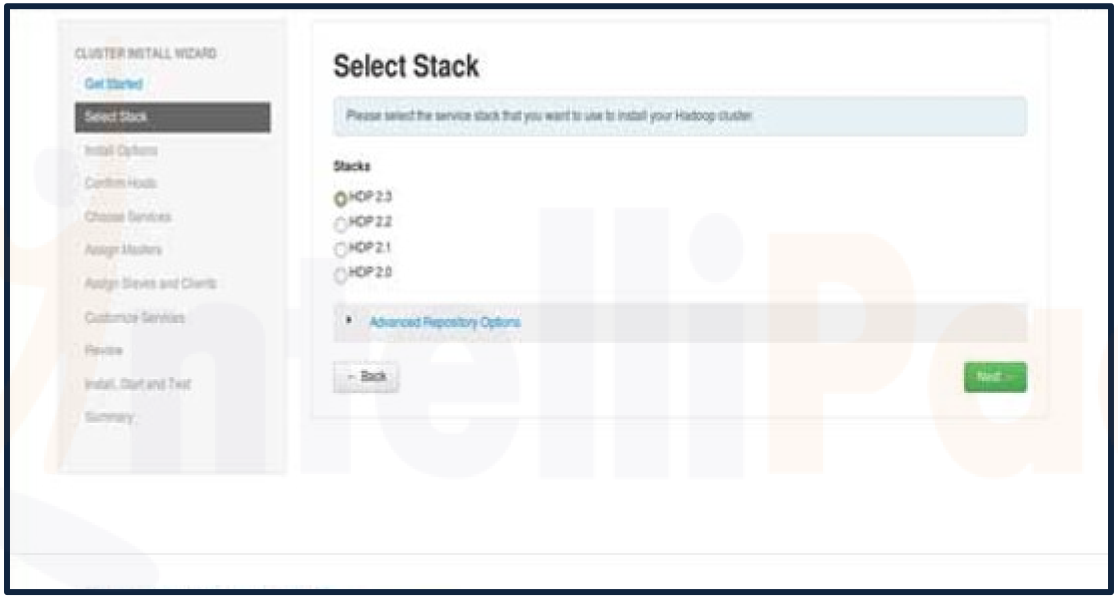
2. Cluster Name:

- Give your cluster a good name.

Note: This is just a simple name for cluster, it is not that significant, so don't worry about it and choose any name for it.

Setting up Hadoop cluster:-

3. Stack selection



Select Stack

Please select the service stack that you want to use to install your Hadoop cluster.

Stacks

● HDP 2.3
○ HDP 2.2
○ HDP 2.1
○ HDP 2.0

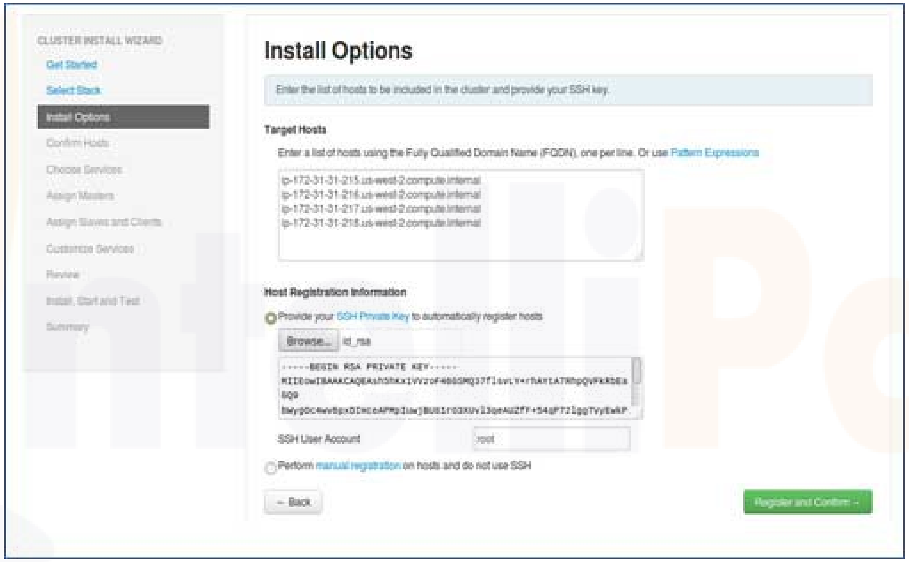
▸ Advanced Repository Options

- This page will list stacks available to install.
- Each stack is pre-packaged with Hadoop ecosystem component.
- These stacks are from Hortonworks. (We can install plain Hadoop too.)

Setting up Hadoop cluster:-

4. Hosts Entry and SSH key entry

Prior moving further this step we should have password less SSH setup for all the participating nodes.



Install Options

Enter the list of hosts to be included in the cluster and provide your SSH key.

Target Hosts

Enter a list of hosts using the Fully Qualified Domain Name (FQDN), one per line. Or use Pattern Expressions

```
ip-172-31-31-215.us-west-2.compute.internal
ip-172-31-31-216.us-west-2.compute.internal
ip-172-31-31-217.us-west-2.compute.internal
ip-172-31-31-218.us-west-2.compute.internal
```

Host Registration Information

● Provide your SSH Private Key to automatically register hosts

id_rsa

```
-----BEGIN RSA PRIVATE KEY-----
MIIEowIBAAKCAQEAshShKx1VVzoF4B8SMQ37flsvLY+rhAYtA7RhpQVFKRbEa
6Q9
bWygOc4wvBpxDIHc6APMpIuwjBUBiro3XUvl3qeAuZfY+54gFT2lggTVyEwKP
```

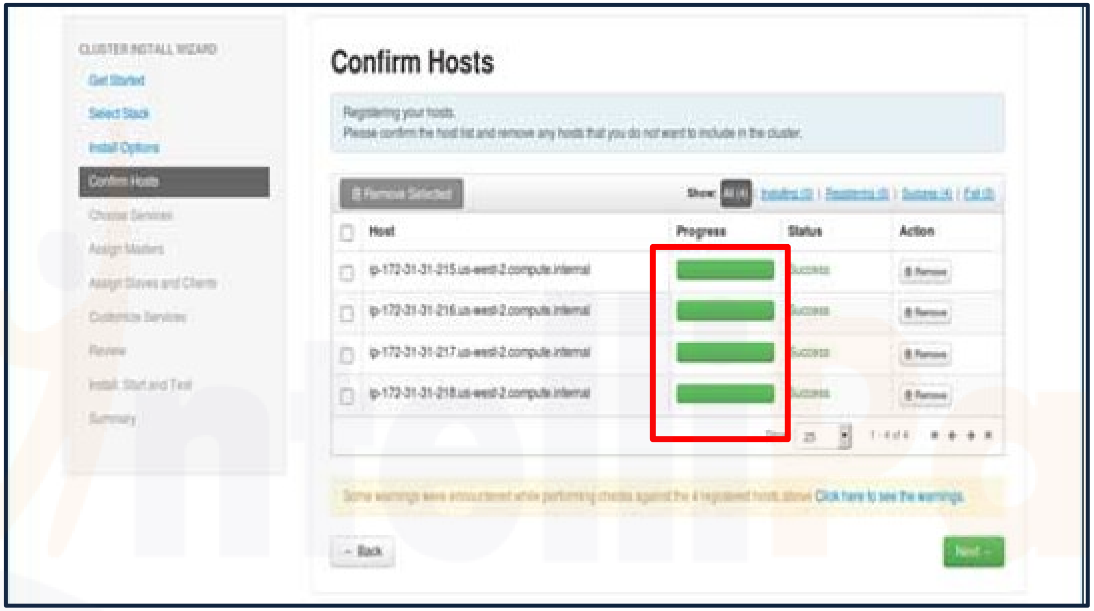
SSH User Account: root

○ Perform manual registration on hosts and do not use SSH

- Add the hostnames of the nodes, single entry on each line. [Add FQDN which can be obtained by hostname –f command].
- Select private key used while setting up password less SSH and username using which private key was created.

Setting up Hadoop cluster:-

5. Hosts registration status



CLUSTER INSTALL WIZARD

- Get Started
- Select Stack
- Install Options
- Confirm Hosts
- Choose Services
- Assign Masters
- Assign Slaves and Clients
- Customize Services
- Review
- Install, Start and Test
- Summary

Confirm Hosts

Registering your hosts.
Please confirm the host list and remove any hosts that you do not want to include in the cluster.

Remove Selected | Show: All (4) | Installing (0) | Registering (0) | Success (4) | Fail (0)

□	Host	Progress	Status	Action
□	ip-172-31-31-215.us-west-2.compute.internal	🟩	Success	Remove
□	ip-172-31-31-216.us-west-2.compute.internal	🟩	Success	Remove
□	ip-172-31-31-217.us-west-2.compute.internal	🟩	Success	Remove
□	ip-172-31-31-218.us-west-2.compute.internal	🟩	Success	Remove

Show: 25 1 - 4 of 4

Some warnings were encountered while performing checks against the 4 registered hosts above Click here to see the warnings.

← Back   Next →

- You can see some operations being performed; these operations include setting Ambari-agent on each node, creating basic setups on each nodes. Once we see ALL GREEN we are ready to move on. Sometimes this may take time as it installs few packages.

Sometime registering hosts fails, retry twice at least or Install Ambari-agent manually on each node.

Setting up Hadoop cluster:-

6. Choose services you wish to install:

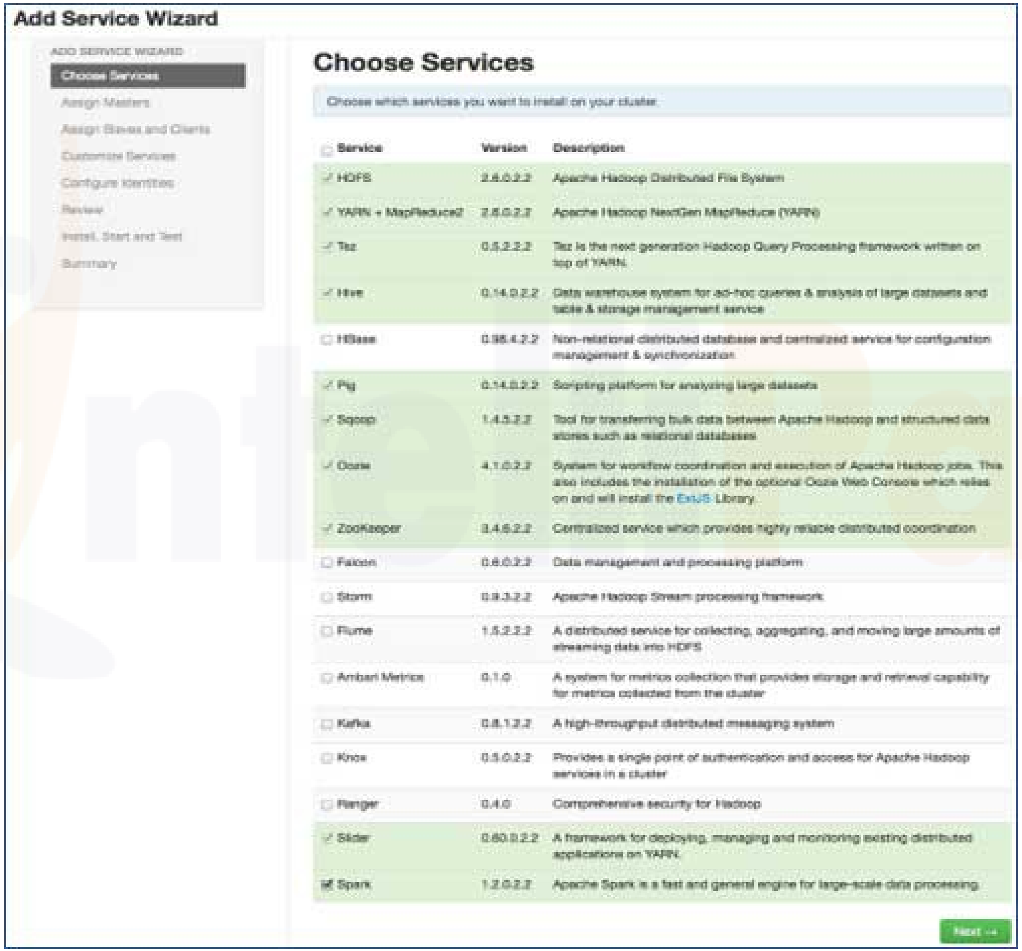
- As per selected stacks in step 3, we have number of services that we can install in the cluster. You can choose one you want.

- Ambari intelligently selects dependent services if you haven't selected it.

- For instance, you selected HBase but not Zookeeper it will prompt same and will add Zookeeper also to the cluster.

Setting up Hadoop cluster:-

6. Choose services you wish to install:



Choose Services

Choose which services you want to install on your cluster

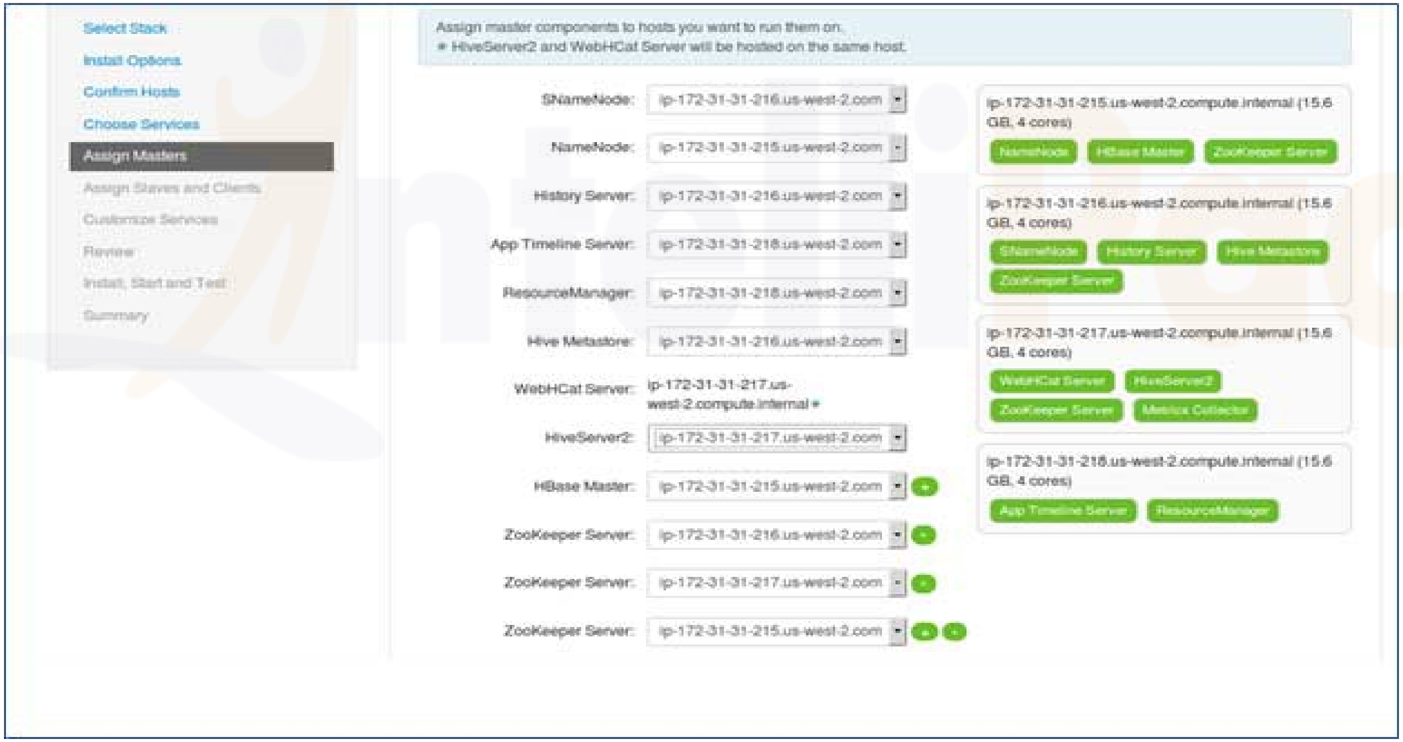
□ Service	Version	Description
✓ HDFS	2.6.0.2.2	Apache Hadoop Distributed File System
✓ YARN + MapReduce2	2.6.0.2.2	Apache Hadoop NextGen MapReduce (YARN)
✓ Tez	0.5.2.2.2	Tez is the next generation Hadoop Query Processing framework written on top of YARN.
✓ Hive	0.14.0.2.2	Data warehouse system for ad-hoc queries & analysis of large datasets and table & storage management service
□ HBase	0.98.4.2.2	Non-relational distributed database and centralized service for configuration management & synchronization
✓ Pig	0.14.0.2.2	Scripting platform for analyzing large datasets
✓ Sqoop	1.4.5.2.2	Tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases
✓ Oozie	4.1.0.2.2	System for workflow coordination and execution of Apache Hadoop jobs. This also includes the installation of the optional Oozie Web Console which relies on and will install the ExtJS Library.
✓ ZooKeeper	3.4.6.2.2	Centralized service which provides highly reliable distributed coordination
□ Falcon	0.6.0.2.2	Data management and processing platform
□ Storm	0.9.3.2.2	Apache Hadoop Stream processing framework
□ Flume	1.5.2.2.2	A distributed service for collecting, aggregating, and moving large amounts of streaming data into HDFS
□ Ambari Metrics	0.1.0	A system for metrics collection that provides storage and retrieval capability for metrics collected from the cluster
□ Kafka	0.8.1.2.2	A high-throughput distributed messaging system
□ Knox	0.5.0.2.2	Provides a single point of authentication and access for Apache Hadoop services in a cluster
□ Ranger	0.4.0	Comprehensive security for Hadoop
✓ Slider	0.60.0.2.2	A framework for deploying, managing and monitoring existing distributed applications on YARN.
☑ Spark	1.2.0.2.2	Apache Spark is a fast and general engine for large-scale data processing.

Setting up Hadoop cluster:-

7. Master services mapping with Nodes:

As you are aware of Hadoop ecosystem has tools which are based on master-slave architecture.

- ✓ In this step we will associate master processes with the node.



Assign master components to hosts you want to run them on.
* HiveServer2 and WebHCat Server will be hosted on the same host.

SNameNode: ip-172-31-31-216.us-west-2.com

NameNode: ip-172-31-31-215.us-west-2.com

History Server: ip-172-31-31-216.us-west-2.com

App Timeline Server: ip-172-31-31-218.us-west-2.com

ResourceManager: ip-172-31-31-218.us-west-2.com

Hive Metastore: ip-172-31-31-216.us-west-2.com

WebHCat Server: ip-172-31-31-217.us-west-2.compute.internal *

HiveServer2: ip-172-31-31-217.us-west-2.com

HBase Master: ip-172-31-31-215.us-west-2.com

ZooKeeper Server: ip-172-31-31-216.us-west-2.com

ZooKeeper Server: ip-172-31-31-217.us-west-2.com

ZooKeeper Server: ip-172-31-31-215.us-west-2.com

ip-172-31-31-215.us-west-2.compute.internal (15.6 GB, 4 cores)
NameNode HBase Master ZooKeeper Server

ip-172-31-31-216.us-west-2.compute.internal (15.6 GB, 4 cores)
SNameNode History Server Hive Metastore ZooKeeper Server

ip-172-31-31-217.us-west-2.compute.internal (15.6 GB, 4 cores)
WebHCat Server HiveServer2 ZooKeeper Server Metrics Collector

ip-172-31-31-218.us-west-2.compute.internal (15.6 GB, 4 cores)
App Timeline Server ResourceManager

Setting up Hadoop cluster:-

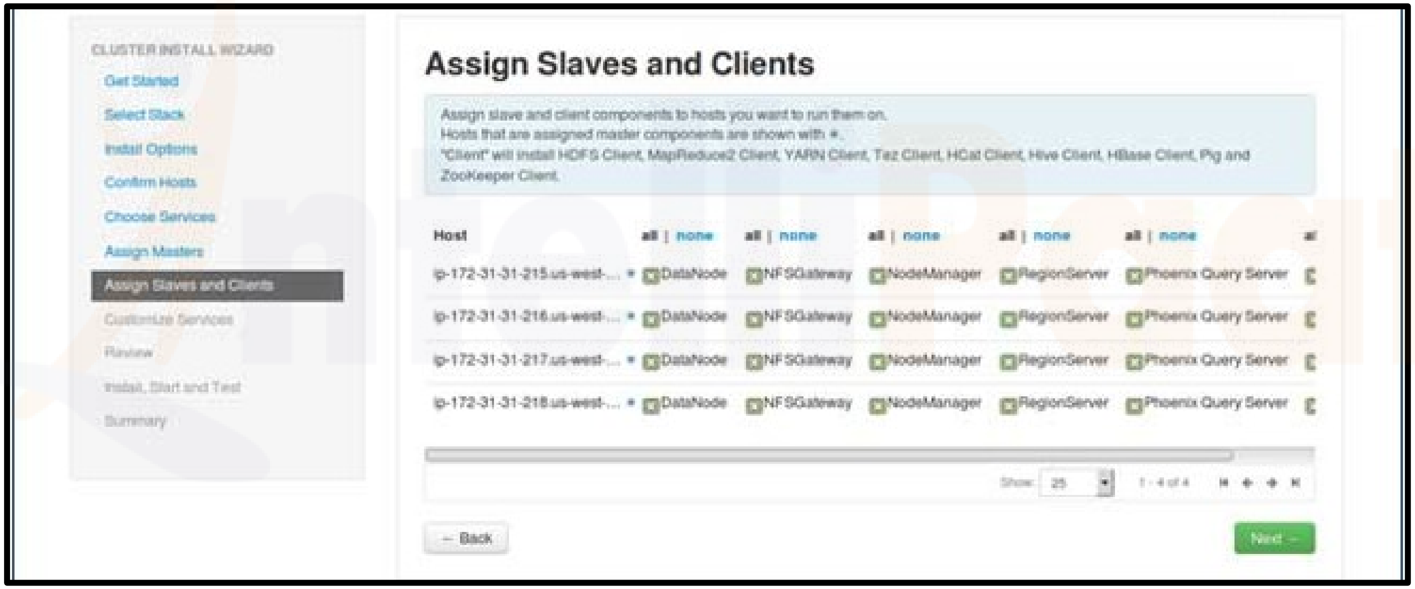
- ✓ Here make sure you properly balance your cluster.
- ✓ Also keep in mind primary and secondary services like Namenode and secondary Namenode are not on the same machine.

Always distribute services in such a way that you have only one master on one node. So that node failure does not disturb multiple services, keep good amount of main memory for optimum performance.

Setting up Hadoop cluster:-

8. Slaves mapping with Nodes:-

Similar to masters, map slave services on the nodes. In general, all the nodes will have slave process running at least for Datanodes and Nodemanagers.



Assign Slaves and Clients

Assign slave and client components to hosts you want to run them on.
Hosts that are assigned master components are shown with ✱.
"Client" will install HDFS Client, MapReduce2 Client, YARN Client, Tez Client, HCat Client, Hive Client, HBase Client, Pig and ZooKeeper Client.

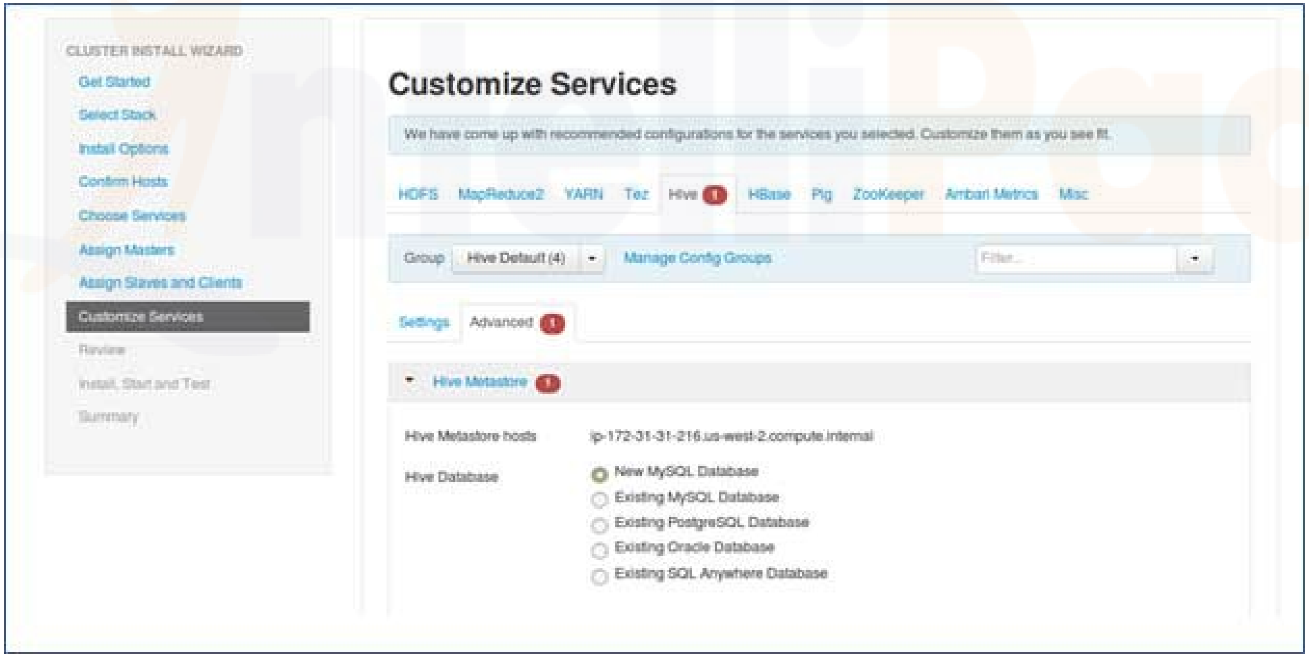
Host	all | none	all | none	all | none	all | none	all | none	al
ip-172-31-31-215.us-west-... ✱	☑ DataNode	☑ NFSGateway	☑ NodeManager	☑ RegionServer	☑ Phoenix Query Server	☑
ip-172-31-31-216.us-west-... ✱	☑ DataNode	☑ NFSGateway	☑ NodeManager	☑ RegionServer	☑ Phoenix Query Server	☑
ip-172-31-31-217.us-west-... ✱	☑ DataNode	☑ NFSGateway	☑ NodeManager	☑ RegionServer	☑ Phoenix Query Server	☑
ip-172-31-31-218.us-west-... ✱	☑ DataNode	☑ NFSGateway	☑ NodeManager	☑ RegionServer	☑ Phoenix Query Server	☑

Setting up Hadoop cluster:-

9. Customize services

This is very important page for Administrators!

- Here you can configure properties for your cluster to make it most suited to your use cases.
- Also it will have some required properties like Hive metastore password (if hive is selected) etc. These will be pointed with Red error like symbols.



Customize Services

We have come up with recommended configurations for the services you selected. Customize them as you see fit.

Hive Metastore hosts: ip-172-31-31-216.us-west-2.compute.internal

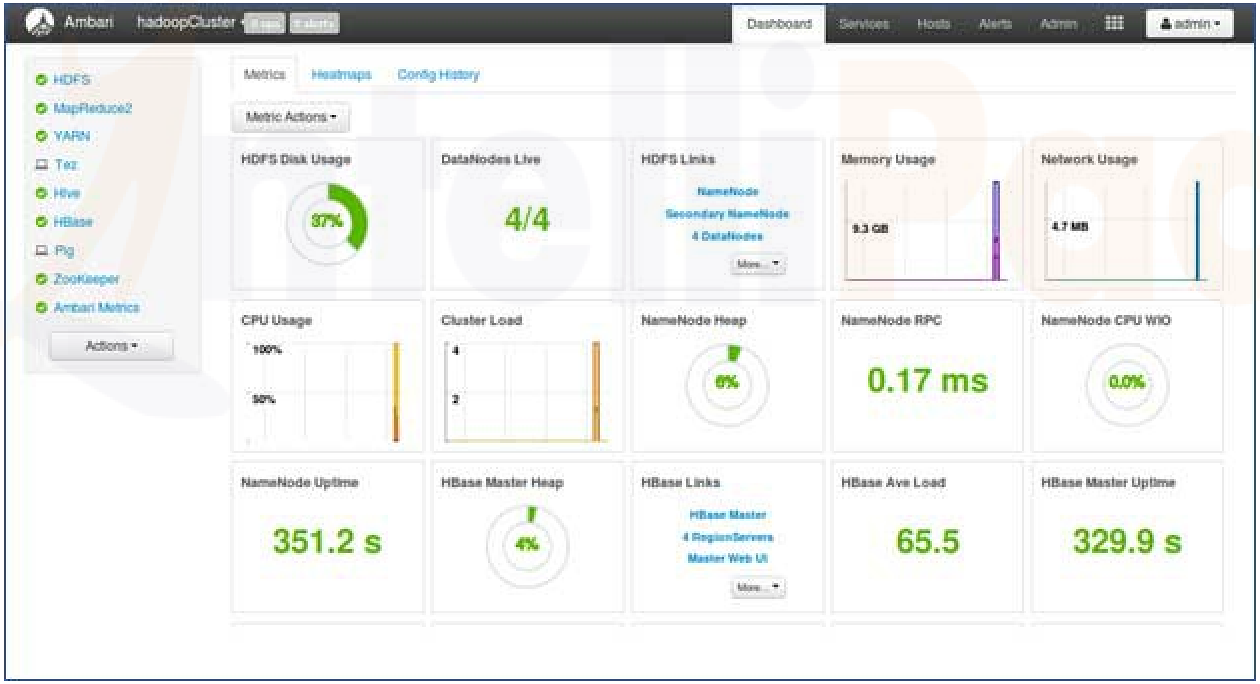
Hive Database:
● New MySQL Database
○ Existing MySQL Database
○ Existing PostgreSQL Database
○ Existing Oracle Database
○ Existing SQL Anywhere Database

Setting up Hadoop cluster:-

10. Review and start provisioning

Make sure you review the cluster configuration before launch as this will save from unknowingly set wrong configurations.

11. Launch and stay back until status becomes GREEN.



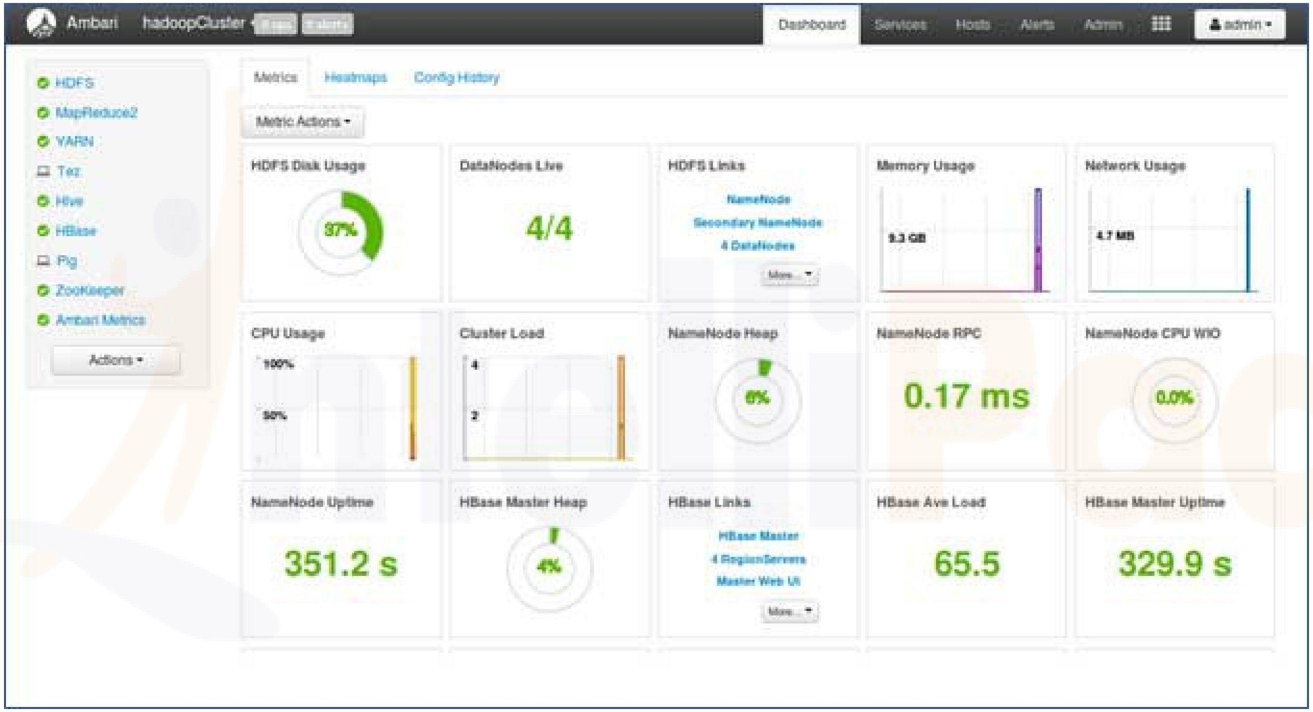
Setting up Hadoop cluster:-

10. Review and start provisioning:

Make sure you review the cluster configuration before launch as this will save from unknowingly set as wrong configurations.

Setting up Hadoop cluster:-

11. Launch and stay back until status becomes GREEN.



Wow! We have successfully Installed Hadoop and all the components on all the nodes of the cluster. Now we can get ourselves start playing with Hadoop.

Conclusion:

- ✓ We have now learned how to install Hadoop and its components on multi-node cluster
  using a simple web based tool called Apache Ambari.

- ✓ Apache Ambari provides us a simpler interface and saves lots of our efforts on installation, monitoring and management which would have be very tedious with so many components and their different installation steps and monitoring controls.

Thank You

Email us – support@intellipaat.com

Visit us - https://intellipaat.com