Data Lake @ AWS

# Introduction

To start any POC/POI, there has to be one new platform.

Following components are needed for this platform.

1. Hadoop Cluster – Single or Multiple node
2. Edge node
3. Source Systems files sample (Either in the form of CSV of in the form of RDBMS data)
4. Security is not added here.
5. Otherwise needed few MIT Kerberos Instance. (Although this can be enabled on a single EC2 machine. )
6. One Data Store – Which is Oracle RDS instance in this case.
7. Node for Web server – Edge node can be used for the same.
8. For Framework
9. Framework code
10. Spring, Java As per need on Edge node
11. R installation in case of Edge node is used for R – Modelling.

# Initial Set up

# Setting up VPC

This is one time activity.

* 1. Create a new Key and save it for future use

## Setting up Sub net –

* Use wizard with one private, one public and
* Create Security Group
  + Setting up incoming for SSH, HTTP and HTTPS and
  + Setting up outgoing for ALL traffic

## Create one Oracle@RDS

This step will create one more sub net for RDS – for ex - VPC\_SN\_DB\_ABIM\_GG\_DEV

## Setting up NAT and others

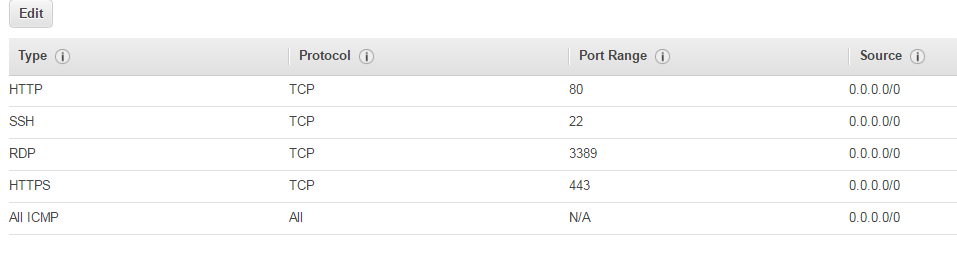
Note: every time NAT instance to be started and it should be assigned to one elastic IP.

Note: Initially Hadoop cluster node was made in private seub net but because that not can not be connected from outside for hadoop installation, this has been moved to publich sub net. Plan is to move this back to private subnet once Hadoop installation is done and environment has been stabilised.

(To move the instance from one subnet to another subnet , it has to be done after creating the AMI image for old machine and use that image for new creation, followed by creation, old machine should be terminated to save/reclaim the spaces. )

## Setting up security group

Need to maintain the following for the smooth ping.



## Setting up Edge node

Steps 🡪

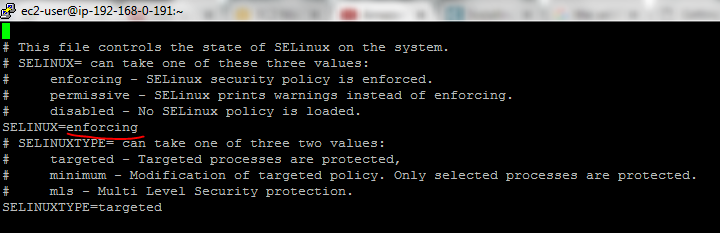
1. Create EC2 instance **in the public subnet**
2. Run sudo yum update for updating the latest patches and updates.
3. To disable SELinux ‘sudo vi /etc/selinux/config’
4. (Not needed )
   1. Run - ssh-keygen -t rsa
   2. Copy key to Hadoop machine after next step (Setting up Hadoop cluster node)
   3. ssh-copy-id [centos@10.0.1.52](mailto:centos@10.0.1.52)
5. instead to enable the connection
   1. first make sure the security group is allowed to have ICMP for ping
   2. Copy the content of the private key and append to id\_rsa file. For ex copy the content of the ABIM\_GG\_DEV\_08Dec.pem to /home/centos/.ssh/rd\_rsa file.
   3. This will enable ssh connection from edge node to hadoop cluster node.
   4. Use ssh – i </home/cenrtos/.ssh/id\_rsa> centos@10.0.1.152

## Setting up Hadoop cluster node

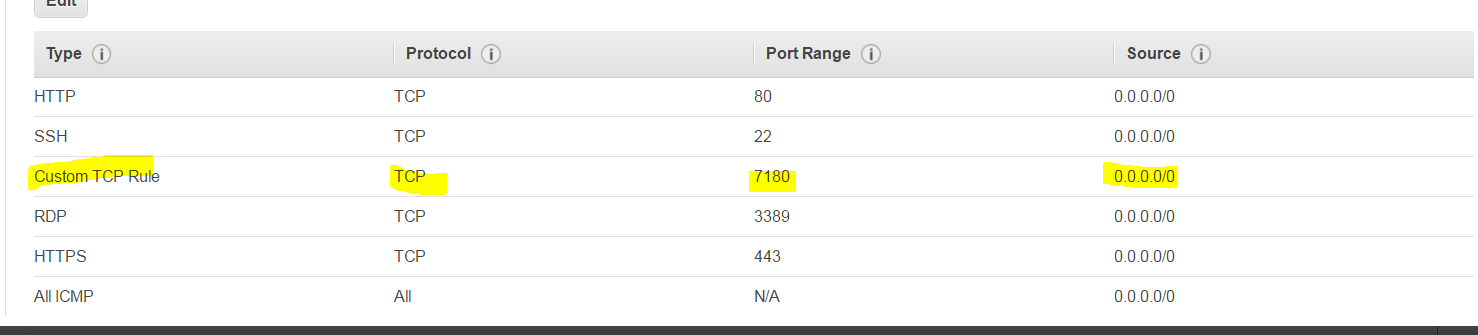
Steps 🡪

1. Create EC2 instance **in the public subnet (Known issue: if you create this in PVT subnet , connectivity between edge node and cluster node if not working, some setting is needed in the route table but this point is still open.) This is fixed now. The only change was to allowe ICMP traffic via security group.**

NOTE: But again due to some issue of the NAT instance set up internet from the private subnet was not working so as a temporary measure , its route table has been mapped to allow using the internet gateway. This was done at the subnet@ route table level.

1. Run sudo yum update for updating the latest patches and updates.
2. If ‘wget’ command is not working, we need to install it using ‘sudo yum install wget’.
3. wget <http://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin>
4. use ls –ltr to check this downloaded file and its permissions
5. Use this command to grant execute permission on cloudera-manager-installer.bin
   1. chmod +x cloudera-manager-installer.bin
6. To disable SELinux ‘sudo vi /etc/selinux/config’
   * 1. 
     2. In above change ‘enforcing’ to ‘disabled’
     3. *To make this change effective we need to ‘reboot’ the ec2 instace.*
7. After performing step 7 and rebooting EC2 instance, execute the cloudera manager installer using ‘sudo ./cloudera-manager-installer.bin’

## Checking hadoop cluster node - services

1. Use sudo service cloudera-scm-server status and sudo service cloudera-scm-server-db status to check the service status.
2. Also following entry is needed in the security group – for cloudera admin UI to work
3. 
4. Now after restart of the box please start the service again using following:
   1. sudo service cloudera-scm-server-db start
   2. sudo service cloudera-scm-server start

## Setting up Dev Machine and enable it for remote access

After Creating new EC2 machine, do the following -  This process is still not tested.   
  
  
[~~http://www.linuxtechi.com/install-configure-vnc-server-centos-7-rhel-7/~~](http://www.linuxtechi.com/install-configure-vnc-server-centos-7-rhel-7/) ~~Step 1:   
        sudo yum groupinstall "GNOME Desktop"   
Step 2:   
        sudo yum install tigervnc-server xorg-x11-fonts-Type1   
Step 3:   
        sudo cp /lib/systemd/system/vncserver@.service /etc/systemd/system/vncserver@:3.service   
Step 4:   
        sudo  vi /etc/systemd/system/vncserver@:3.service   
        change <USER> at two places with abimggnusr   
Step 5:   
        Set the Firewall Rule if firewall is enabled on your linux box ?   
  
Step 6:   
        Switch to the user (linuxtechi) and run vncserver command to set the password   
        sudo useradd <USER>   
        sudo passwd  <USER>   
abimggnusr@123   
  
step 7:   
        switch to user abimggnusr   
        run cmd vncserver~~   
  
  
<http://devopscube.com/how-to-setup-gui-for-amazon-ec2-rhel-7-instance/>   
  
sudo yum -y update   
sudo yum groupinstall -y "Server with GUI"   
sudo systemctl set-default graphical.target  
sudo systemctl default   
  
  
sudo rpm -Uvh <http://li.nux.ro/download/nux/dextop/el7/x86_64/nux-dextop-release-0-1.el7.nux.noarch.rpm>   
  
Setting Up XRDP   
  
sudo yum install -y xrdp tigervnc-server   
  
chcon --type=bin\_t /usr/sbin/xrdp  
chcon --type=bin\_t /usr/sbin/xrdp-sesman   
  
  
sudo systemctl start xrdp   
sudo systemctl enable xrdp   
  
sudo firewall-cmd --permanent --add-port=3389/tcp  
sudo firewall-cmd --reload   
  
sudo passwd ec2-user   
 <ec2-user@123>   
  
Further step was not needed and it worked without root password change.

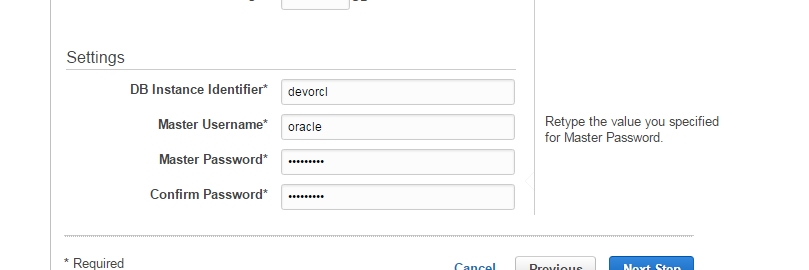
# Appendix

Also, below is the link for 'How to increase the size of EBS'

<http://www.n2ws.com/how-to-guides/how-to-increase-the-size-of-an-aws-ebs-cloud-volume-attached-to-a-linux-machine.html>

# Appendix

# RDS – Oracle \_ DB instance details



Password is oracle123 for the above instance created on 12 Dec

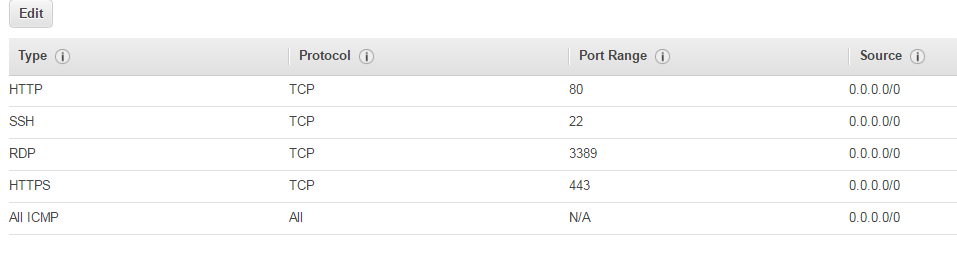
Adding following AMI

ami-d2c924b2 – ami ID

**CentOS 7 (x86\_64) - with Updates HVM**

CentOS Linux 7 x86\_64 HVM EBS 1602

Following is the security group



## Cloudera install – screenshots

Database passwords

