BDA Admin Guide

Current BDA version: v4.14 OL7 (no Kerberos)

Last updated by: Chujun (25/11/2019)

**How to read this document:**

* Title name – the last updated date
* Description: the command you should run

**Contents**

[User Administration 2](#_Toc52780830)

[Create cluster-wide user (HDFS/Hadoop/Software configuration) - 25/11/2019 2](#_Toc52780831)

[Set-up Hadoop Client for VM – 25/11/2019 2](#_Toc52780832)

[Put data into HDFS (e.g /user/data\_warehouse) from remote host – 25/11/2019 3](#_Toc52780833)

[Spark Application Resource Allocation 3](#_Toc52780834)

[Software Installation/Upgrade 5](#_Toc52780835)

[Install Anaconda/Python3.7 in selected nodes – 25/11/2019 5](#_Toc52780836)

[Installation Steps for Accumulo – 27/11/2019 5](#_Toc52780837)

[Installation for Screen – 27/11/2019 5](#_Toc52780838)

[Installation Steps for Zeppelin – 28/11/2019 5](#_Toc52780839)

# User Administration

## Create cluster-wide user (HDFS/Hadoop/Software configuration) - 25/11/2019

1. SSH into the first node of the cluster (10.10.28.10), with root account.
2. Create cluster-wide Linux user: dcli -C adduser <theusername>
3. Change user password: dcli -C “echo <thepasswordforuser> | passwd <theusername> --stdin”
4. Done.

Do the following steps only if the user needs HDFS/Hadoop access:

1. Add user into Hadoop and HDFS Group: dcli -C usermod -aG hadoop,hdfs <theusername>
2. Login to HUE Web UI: <https://10.10.28.13:8888/hue/editor/?type=hive>, with admin account
3. Click on the right tab (admin) > manage users > add users. Fill in the same user name and password as Linux, make sure the [Create home directory] is checked and add the user into [singtel\_engineers] group.
4. Done.

Do the following steps only if the user needs [software] folder\* access:

\*The user will be using the default python version 2.7.5 if the user is not in software group

\*Python3.7 and Anaconda is installed in node 8 and 9, only these 2 nodes are able to run python3.7

1. SSH into the first node of the cluster (10.10.28.10), with root account.
2. Add user into software group: dcli -C usermod -aG software <theusername>
3. Append the following text into "~/.bashrc": vi ~/.bashrc

# >>> conda initialize >>>

# !! Contents within this block are managed by 'conda init' !!

\_\_conda\_setup="$('/home/software/anaconda3/bin/conda' 'shell.bash' 'hook' 2> /dev/null)"

if [ $? -eq 0 ]; then

eval "$\_\_conda\_setup"

else

if [ -f "/home/software/anaconda3/etc/profile.d/conda.sh" ]; then

. "/home/software/anaconda3/etc/profile.d/conda.sh"

else

export PATH="/home/software/anaconda3/bin:$PATH"

fi

fi

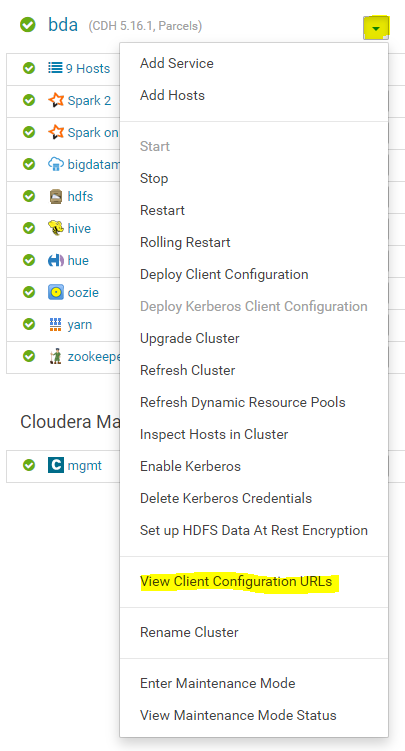
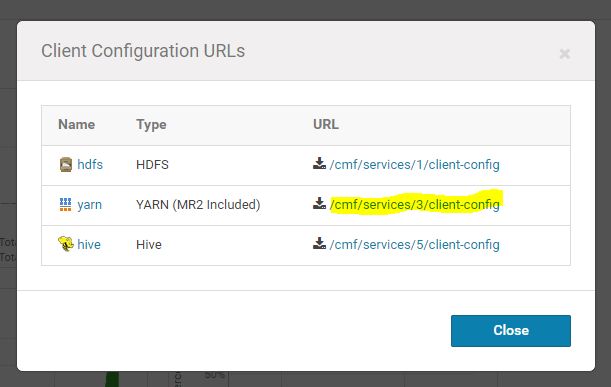
unset \_\_conda\_setup

# <<< conda initialize <<<

1. Update local environment variable: source ~/.bashrc
2. Run `python`, now the ‘python’ will be 3.7.
3. Done.

## Set-up Hadoop Client for VM – 25/11/2019

1. Download yarn client configurations from Cloudera manager

1. Unzip it and copy the entire [yarn-conf] folder into VM /etc/hadoop/, and rename the folder to [conf]
2. Use this command to test whether hdfs is working: hdfs dfs -ls /user
3. Done.

## Put data into HDFS (e.g /user/data\_warehouse) from remote host – 25/11/2019

1. Make sure the Hadoop client is installed, if not please refer to: [Set-up Hadoop Client](#_Set-up_Hadoop_Client)
2. Use the following command to put files/folders

HADOOP\_USER\_NAME=<hdfsUserName) hdfs dfs -put <localPath> <hdfsPath>

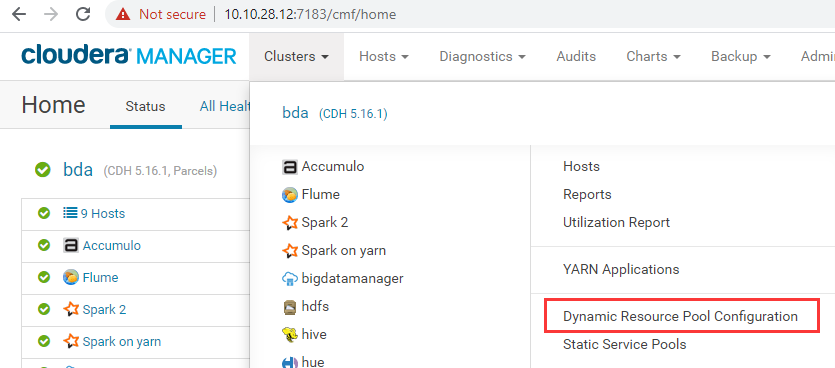
e.g.

HADOOP\_USER\_NAME=data\_warehouse hdfs dfs -put megapop\_netflow/ /user/data\_warehouse

1. Done.

## Spark Application Resource Allocation

1. Login into Cloudera Manager
2. Select “Dynamic Resource Pool Configuration” under “Cluster”



1. Configure it accordingly.

# Software Installation/Upgrade

## Install Anaconda/Python3.7 in selected nodes – 25/11/2019

1. (root) Create user called software -> this user is used for sharing software
2. (software) Download Anaconda Linux 64 bits: <https://docs.anaconda.com/anaconda/install/linux/>
3. (software) Put Anaconda3-2019.10-Linux-x86\_64.sh into the software home folder
4. (software) Install Anaconda
5. (root) Change "/home/software" permission to 770
6. Done.

## Installation Steps for Accumulo – 27/11/2019

Refer to <https://docs.cloudera.com/content/www/en-us/documentation/other/accumulo/1-7-2/PDF/Apache-Accumulo-Installation-Guide-1-7-2.pdf>

## Installation for Screen – 27/11/2019

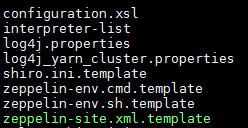
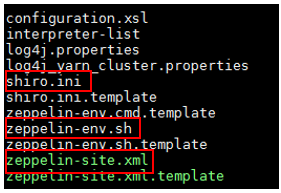
1. Download screen rpm from <https://centos.pkgs.org/7/centos-x86_64/screen-4.1.0-0.25.20120314git3c2946.el7.x86_64.rpm.html>
2. Put the rpm into node1, root@10.10.28.10:/tmp
3. Distribute the rpm to all the nodes:

dcli -f /tmp/screen-4.1.0-0.25.20120314git3c2946.el7.x86\_64.rpm -d /tmp

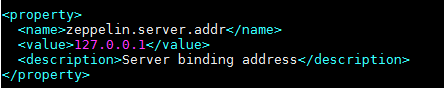
1. Make sure the rpm is successfully distributed: dcli ls /tmp | grep screen
2. Install screen on all the nodes: dcli rpm -i /tmp/screen-4.1.0-0.25.20120314git3c2946.el7.x86\_64.rpm
3. Check the installation of screen: dcli rpm -qa | grep screen

## Installation Steps for Zeppelin – 28/11/2019

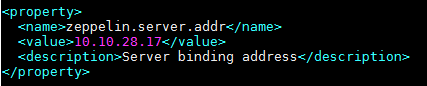
1. Download Zeppelin and unpack it into /home/software by using Linux user software
2. Go to configuration folder: cd /home/software/zeppelin-0.8.2-bin-all/conf
3. Use Linux copy command to copy these 3 files and remove the “template”

 TO 

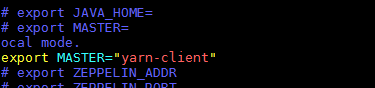
1. Update “zeppelin-site.xml” - change the IP address of zeppelin

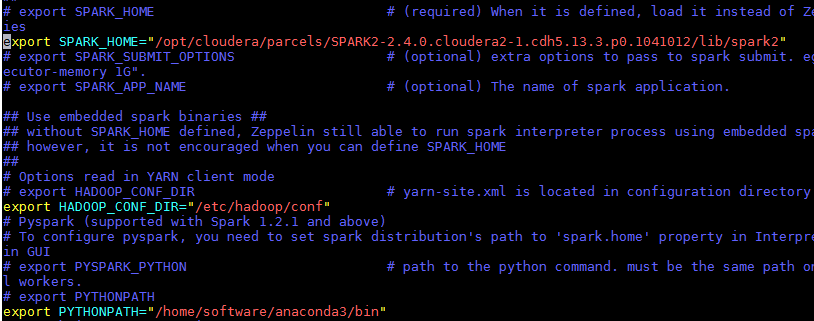


TO



1. Update “zeppelin-env.sh” – add 4 export variables

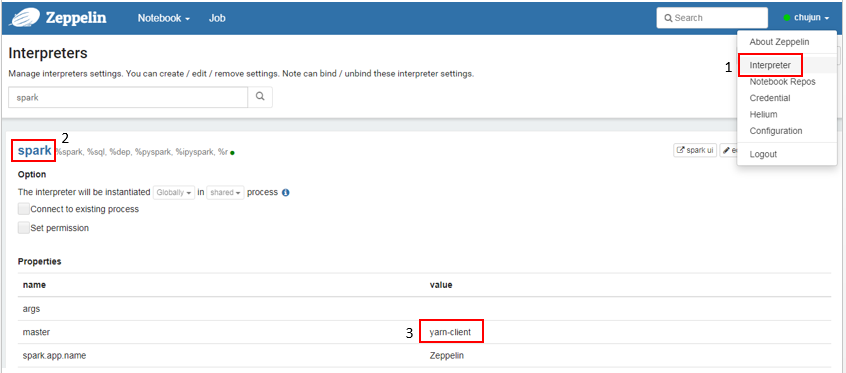




1. Update/add new user name password in the “shiro.ini”
2. Use the following command to start/stop/restart Zeppelin: e.g.

./zeppelin-0.8.2-bin-all/bin/zeppelin-daemon.sh start

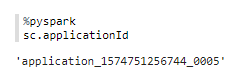
1. Access Zeppelin via browser with this URL -> 10.10.28.17:8080
2. Change the SPARK interpreter configuration: the value of master from “local[\*]” to “yarn-client”



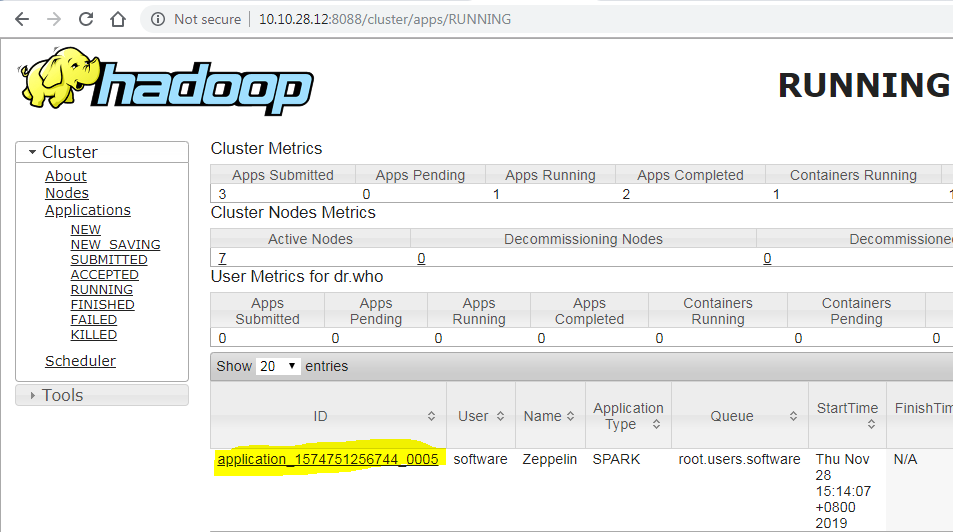
1. Save and restart the SPARK interpreter

To confirm the Zeppelin is using cluster mode not local resource:

1. Run “sc.applicationId” with spark interpreter. (can be %spark or %pyspark)



1. Access SPARK WebUI: <http://10.10.28.12:8088/cluster/apps/RUNNING>
2. Make sure the Application ID of the notebook is in the SPARK UI:



If the ID is not in the list that means the application isn’t submit to the cluster. It is using local resource.

1. Done.