StreamSets DataCollector

on

Oracle BDCE

Authors: Mohsen Alam; Randall Lunetto

Instructor: Jongwook Woo

Date: 01/28/2021

Lab Tutorial

Mohsen Alam (malam@calstatela.edu)

Randall Lunetto (rlunett@calstatela.edu)

01/28/2021

Install and Run

StreamSets Data Collector

GUI on Oracle BDCE

Objectives

In this hands-on lab, you will learn how to:

- Download StreamSets Data Collector tarball file to your local desktop
- Upload the tarball file into Oracle BDCE
- Install the StreamSets Data Collector in Oracle BDCE
- Run the StreamSets Data Collector GUI from the Oracle BDCE via port forwarding

Platform Spec

- Windows 10
- CPU Speed: 2.38 GHz
- # of CPU cores: 1
- # of nodes: 1
- Total Memory Size: 16 GB

Step 1: Download StreamSets Tarball file

In this step, you will download the StreamSets data collector file

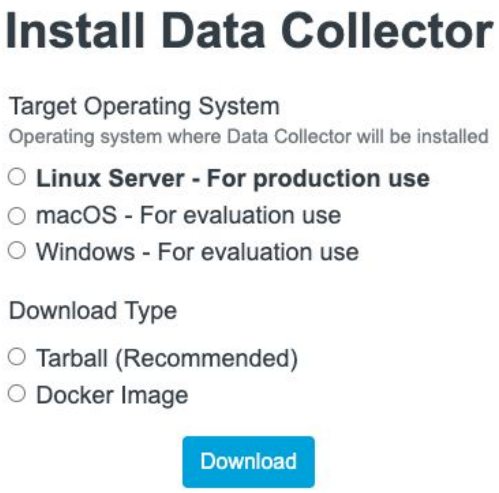
1. After you've logged in to your StreamSets account, go to: https://accounts.streamsets.com/install/select/data-collector
2. Select the following -

   For Target Operation system:

   Select Linux Server - For production use

   For Download Type:

   Select Tarball (Recommended)
3. Click the Download button and wait for the file to download to your desktop



Install Data Collector

Target Operating System

Operating system where Data Collector will be installed

○ **Linux Server - For production use**

○ macOS - For evaluation use

○ Windows - For evaluation use

Download Type

○ Tarball (Recommended)

○ Docker Image

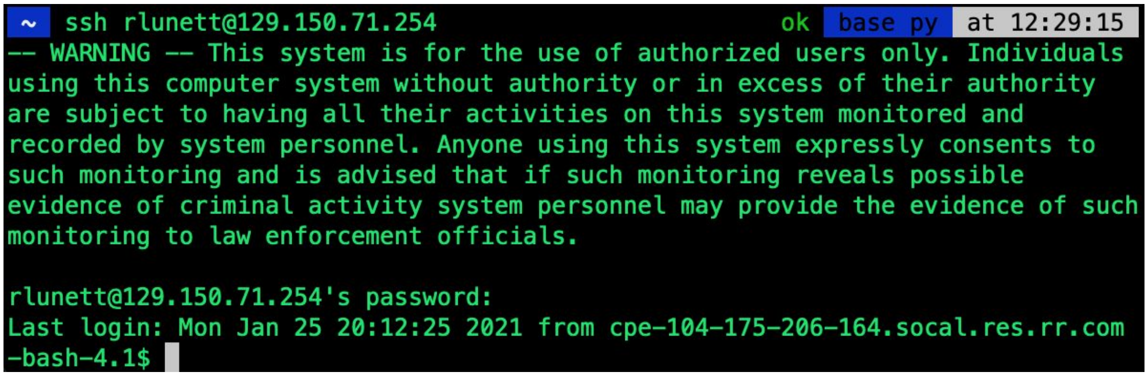
Step 2: Upload the Tarball file to your Oracle BDCE File System

In this step you will connect to your Oracle BDCE server from the terminal and upload the tarball file from your local desktop to your Oracle BDCE file system to complete the installation.

1. Open your local desktop terminal and connect to your Oracle BDCE server using this command:

   ```
   ssh [your_username]@129.150.71.254
   ```

   Note: the @IP Address is given to you by your instructor.

   ```
   ~  ssh rlunett@129.150.71.254                          ok  base py  at 12:29:15
   -- WARNING -- This system is for the use of authorized users only. Individuals
   using this computer system without authority or in excess of their authority
   are subject to having all their activities on this system monitored and
   recorded by system personnel. Anyone using this system expressly consents to
   such monitoring and is advised that if such monitoring reveals possible
   evidence of criminal activity system personnel may provide the evidence of such
   monitoring to law enforcement officials.

   rlunett@129.150.71.254's password:
   Last login: Mon Jan 25 20:12:25 2021 from cpe-104-175-206-164.socal.res.rr.com
   -bash-4.1$
   ```

2. Type in your user password - this should be the same as your username

   Note: you will not see your typed characters appear in the terminal.

3. Create a new directory called ‘streamsets’ using this command:

   ```
   mkdir streamsets
   ```

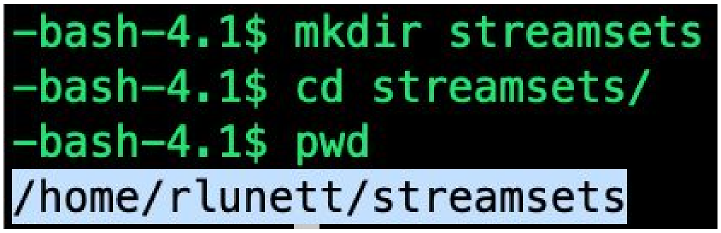
4. CD into the streamsets directory using this command:

   ```
   cd streamsets/
   ```

5. Get the path of this directory using this command:

   pwd

   then highlight and copy this path



```
-bash-4.1$ mkdir streamsets
-bash-4.1$ cd streamsets/
-bash-4.1$ pwd
/home/rlunett/streamsets
```

6. Open a new terminal (cntrl+t for new tab or cntrl+n for new window)
7. From this new terminal (your local desktop directory), you will upload the tarball file into your Oracle BDCE file system by using the SCP command:

   SCP Command Syntax:

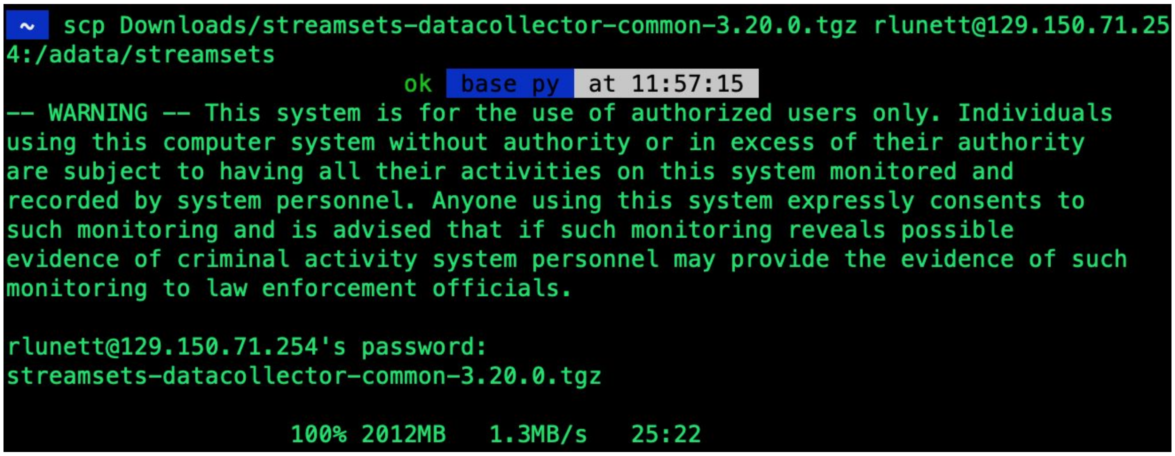
scp [PATH OF TAR FILE][YOUR_USERNAME]@129.150.71.254:[PASTE PATH OF STREAMSETS DIRECTORY]

   Type this command:

```
scp Downloads/streamsets-datacollector-common-3.20.0.tgz
rlunett@129.150.71.254:/home/rlunett/streamsets
```

   Note: This command is all one line. Be mindful of the space characters required.

8. If you type the command successfully, you'll be prompted to type in your username and password to your Oracle BDCE file system and the upload process will begin, you will see a status bar of your upload progress:



```
 ~  scp Downloads/streamsets-datacollector-common-3.20.0.tgz rlunett@129.150.71.25
4:/adata/streamsets
                                             ok  base py  at 11:57:15
-- WARNING -- This system is for the use of authorized users only. Individuals
using this computer system without authority or in excess of their authority
are subject to having all their activities on this system monitored and
recorded by system personnel. Anyone using this system expressly consents to
such monitoring and is advised that if such monitoring reveals possible
evidence of criminal activity system personnel may provide the evidence of such
monitoring to law enforcement officials.

rlunett@129.150.71.254's password:
streamsets-datacollector-common-3.20.0.tgz

                    100% 2012MB   1.3MB/s   25:22
```

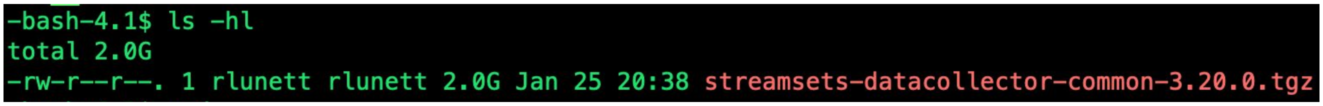
Note: It's a large file so this may take ~25 minutes.

Step 3: Install StreamSets Data by Extracting the Tarball file contents in Oracle BDCE

1. Go back into your other terminal (the Oracle BDCE file system), and verify if the tar file uploaded successfully typing this command:

   `ls -hl`

   If successful, you should see something like this:



```
-bash-4.1$ ls -hl
total 2.0G
-rw-r--r--. 1 rlunett rlunett 2.0G Jan 25 20:38 streamsets-datacollector-common-3.20.0.tgz
```

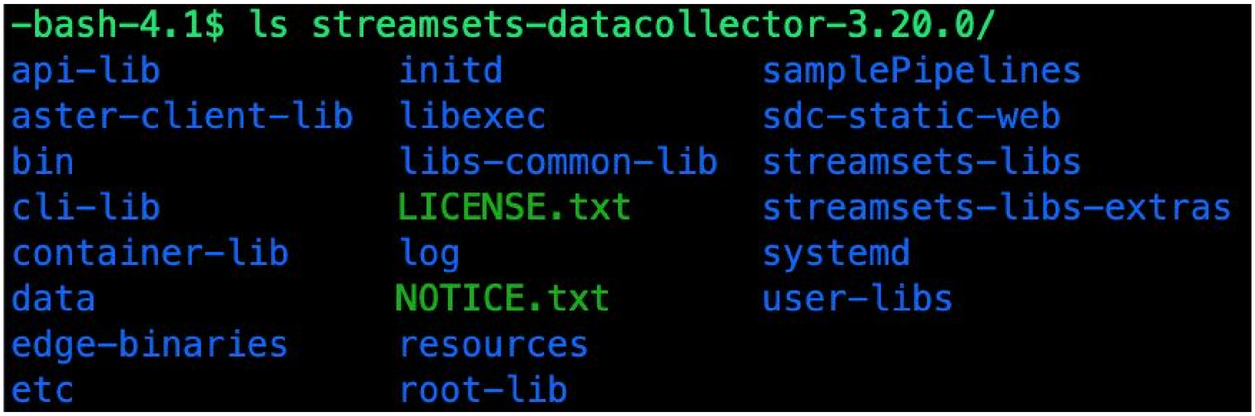
2. Now Unzip the tar file using this command:

   `tar -zxvf streamsets-datacollector-common-3.20.0.tgz`

3. Verify all of the file contents are unzipped successfully typing this command:

   `ls streamsets-datacollector-3.20.0/`

   You should see something like this:



```
-bash-4.1$ ls streamsets-datacollector-3.20.0/
api-lib            initd             samplePipelines
aster-client-lib   libexec           sdc-static-web
bin                libs-common-lib   streamsets-libs
cli-lib            LICENSE.txt       streamsets-libs-extras
container-lib      log               systemd
data               NOTICE.txt        user-libs
edge-binaries      resources
etc                root-lib
```

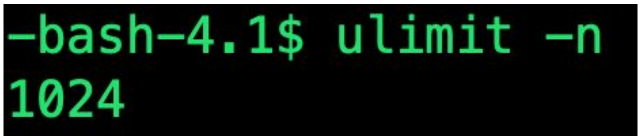
Step 4: Reconfigure the Open File Limit

The StreamSets Data Collector requires a reconfiguration of file descriptors since the default limit settings are too low. In this step, you will reconfigure the open file limit.

1. Check the default file descriptors limit of your Oracle BDCE by typing this command in your Oracle terminal:

   `ulimit -n`

You'll see some output like below:



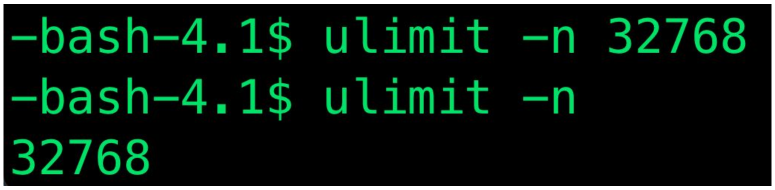
```
-bash-4.1$ ulimit -n
1024
```

2. if your terminal displays a number less than 32768, you need to reconfigure by typing:

   ```
   ulimit -n 32768
   ```

3. Verify the new limit is set with:

   ```
   ulimit -n
   ```



```
-bash-4.1$ ulimit -n 32768
-bash-4.1$ ulimit -n
32768
```

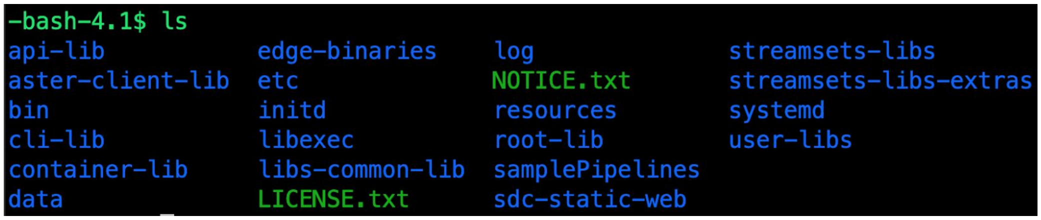
Step 5: Run StreamSets Data Collector GUI with Port Forwarding

In this step, you will run the GUI (graphical user interface) from your Oracle BDCE using port forwarding.

1. Go into your Oracle BDCE terminal and CD into your streamsets-datacollector-3.20.0 sub-directory by typing this command from your user directory:

   ```
   cd /streamsets/streamsets-datacollector-3.20.0
   ```

2. Confirm you're in the right directory by viewing the contents, you should see this:

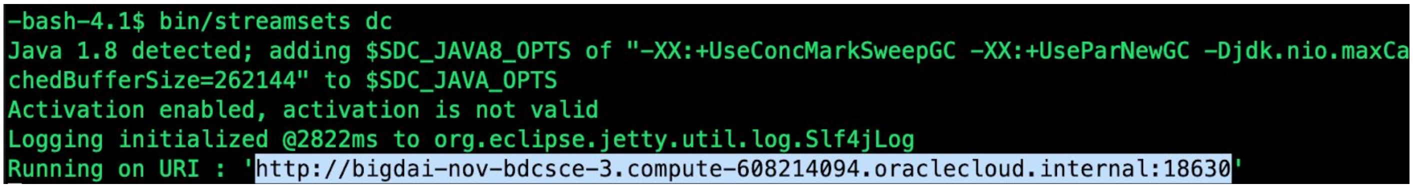


```
-bash-4.1$ ls
api-lib            edge-binaries    log              streamsets-libs
aster-client-lib   etc              NOTICE.txt       streamsets-libs-extras
bin                initd            resources        systemd
cli-lib            libexec          root-lib         user-libs
container-lib      libs-common-lib  samplePipelines
data               LICENSE.txt      sdc-static-web
```

3. Launch the StreamSets Data Collector by entering this command:

   ```
   bin/streamsets dc
   ```

4. Your terminal will output text stating "Running on URI:..." and a web address will be displayed -



```
-bash-4.1$ bin/streamsets dc
Java 1.8 detected; adding $SDC_JAVA8_OPTS of "-XX:+UseConcMarkSweepGC -XX:+UseParNewGC -Djdk.nio.maxCa
chedBufferSize=262144" to $SDC_JAVA_OPTS
Activation enabled, activation is not valid
Logging initialized @2822ms to org.eclipse.jetty.util.log.Slf4jLog
Running on URI : 'http://bigdai-nov-bdcsce-3.compute-608214094.oraclecloud.internal:18630'
```

Note: This output shows you where your StreamSets Data Collector GUI is running, on port "18630" in your Oracle BDCE server.

5. Open the GUI in your web browser from the terminal using port forwarding. Open your local desktop terminal and type this command:

```
ssh -N -L [RANDOM_NUMBER > 1023]:localhost:18630 [Your_Username]@129.150.71.254
```

Example: ssh -N -L 6547:localhost:18630 rlunett@129.150.71.254

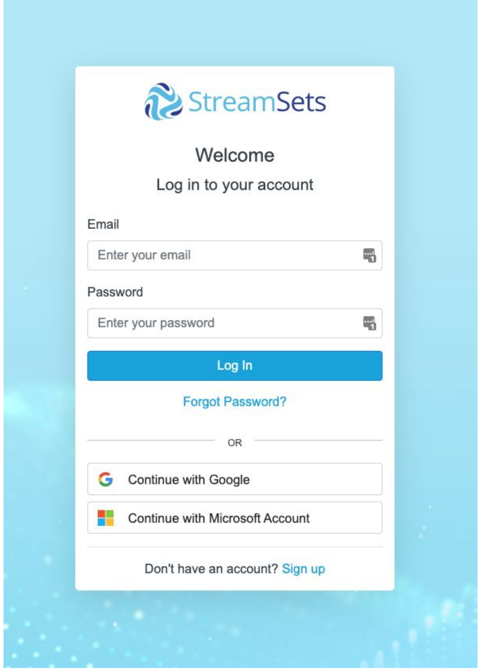
Note: the first port number '6540' is a random port number on localhost (any number above 1023 will work). The second port number '18630' is assigned from StreamSets.

6. Open any web browser and type into the web address bar:

   http://localhost:[PORT_NUMBER_YOU_CHOSE]/

   Example: http://localhost:6547/

7. You will be brought to your StreamSets Data Collector Log In screen and you will have the option to Link to Account - click the button that says ‘Link’.
8. Log in and begin building your pipeline.



This is the end of the tutorial.

References

1. https://github.com/mohsenualam/Twitter-in-StreamSets
2. https://streamsets.com/getting-started/download-install-data-collector/
3. https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Installation/InstallationAndConfig.html#concept_gbn_4lv_1r