StreamSets
Deployment Guide

Version 1.0

Table of Contents

1 INTRODUCTION 3
2 INTENDED AUDIENCE 3
3 DEPLOYMENT STEPS 3
3.1 Step 1. Install the StreamSets Custom Service Descriptor 3
3.2 Step 2. Distribute and Activate the StreamSets Parcel 4
3.3 Step 3: Configure the StreamSets Service 4
3.4 Step 4: Configuring Data Collector with Cloudera Manager 5

1 Introduction

This document provides guidance relating to StreamSets data collection deployment in EDL. You can install Data Collector and start it manually or run it as a service. If you use Cloudera Manager, you can install and administer Data Collector through Cloudera Manager. Data collection requires 32768 file descriptors set in the node in which it is installed. HTTP port for the UI is 18630. The HTTPS port is 18636. On a CM managed CDH cluster, StreamSets needs to be introduced as a custom service descriptor.

2 Intended Audience

CDH administrators

3 Deployment steps

3.1 Step 1. Install the StreamSets Custom Service Descriptor

Install the StreamSets custom service descriptor file (CSD), and then restart Cloudera Manager.

1. Use the following URL to download the CSD from the StreamSets website: https://streamsets.com/opensource.

Or, you can use the GNU Wget program to download the CSD from the command line by running the following commands:

```
export VERSION="3.8.0"
```

```
wget https://archives.streamsets.com/datacollector/$VERSION/csd/STREAMSETS-$VERSION.jar
```

2. Copy the Data Collector CSD file to the Local Descriptor Repository Path. By default, the path is /opt/cloudera/csd.

To verify the path to use, in Cloudera Manager, click Administration > Settings. In the navigation panel, select the Custom Service Descriptors category. Place the CSD file in the path configured for Local Descriptor Repository Path.

3. Set the file ownership to cloudera-scm:cloudera-scm with permission 644.

For example:

```
chown cloudera-scm:cloudera-scm /opt/cloudera/csd/STREAMSETS*.jar
```

```
chmod 644 /opt/cloudera/csd/STREAMSETS*.jar
```

4. Use one of the following commands to restart Cloudera Manager Server:

For Ubuntu 14.04, CentOS 6, Red Hat Enterprise Linux 6, or Oracle Linux 6:

```
service cloudera-scm-server restart
```

For Ubuntu 16.04, CentOS 7, Red Hat Enterprise Linux 7, or Oracle Linux 7:

```
systemctl restart cloudera-scm-server
```

5. In Cloudera Manager, to restart the Cloudera Management Service, click Home > Status. To the right of Cloudera Management Service, click the Menu icon and select Restart.

3.2 Step 2. Distribute and Activate the StreamSets Parcel

After you add the StreamSets repository to Cloudera Manager, you can download, distribute, and activate the StreamSets parcel across the cluster.

Note: The StreamSets parcel repository is added to Cloudera Manager during the installation of the CSD. However, if installing the parcel before the CSD, the StreamSets parcel repository URL is located at: https://archives.streamsets.com/index.html. Download the correct parcel type for the operating system that you use.

When working with multiple clusters, perform the following steps for each cluster.

1. To view the list of available parcels, in the menu bar, click the Parcels icon.

The StreamSets parcel displays in the list of available parcels. If it doesn't display, click Check for New Parcels.

2. To download the StreamSets parcel to the local repository, click Download.

After the parcel is downloaded, the Download button becomes the Distribute button.

3. To distribute the StreamSets parcel to the cluster, click Distribute.

After distribution, the Distribute button becomes the Activate button.

4. To activate the StreamSets parcel, click Activate.

3.3 Step 3: Configure the StreamSets Service

When you configure the service, you assign Data Collector to the hosts where you want it to run.

To run Data Collector in cluster streaming mode, colocate Data Collector on a node with the Spark Gateway role. To run Data Collector in cluster batch mode, colocate Data Collector on a node with the YARN Gateway role.

To write to HDFS, colocate Data Collector on a node with the HDFS Gateway role. Similarly, to write to HBase or Hive, colocate Data Collector on nodes with the HBase or Hive Gateway roles, respectively.

When working with multiple clusters, perform the following steps for each cluster.

1. In Cloudera Manager, click the menu for the cluster you want to use, then click Add a Service.
2. In the Service Types list, select StreamSets, then click Continue.

3. To select the hosts where you want to install StreamSets, on the Customize Role Assignments for StreamSets page, click Select Hosts to open a list of available hosts.
4. Select one or more hosts, then click OK. Click Continue.

The Review Changes page displays the Data and Resource directories for the Data Collector.

5. Optionally change the directories, then click Continue.

The First Run Command page displays status updates as Cloudera Manager starts Data Collector on the selected hosts.

6. Click Continue, then click Finish.

3.4 Step 4: Configuring Data Collector with Cloudera Manager

1. When administering Data Collector with Cloudera Manager, configure all Data Collector configuration properties and environment variables through Cloudera Manager.
2. Manual changes to Data Collector configuration files can be overwritten by Cloudera Manager.
3. In Cloudera Manager, select the StreamSets service, then click Configuration.
4. The Configuration page displays Data Collector configuration properties.
5. On the Configuration page, in the navigation panel, you can select a category to configure related properties.

StreamSets LDAP properties for DEV:

```
ldap.authenticationMethod=simple
ldap.bindDn=CN=SA-ITS-DEV-EDL-AD,OU=Service Accounts,OU=NCSUS,DC=na,DC=jnj,DC=com
ldap.bindPassword=Cloudera@5
ldap.debug=true
ldap.forceBindingLogin=true
ldap.hostname=JNJ.COM
ldap.port=3269
ldap.roleBaseDn=OU=Groups,DC=jnj,DC=com
ldap.roleFilter=member={dn}
ldap.roleMemberAttribute=member
ldap.roleNameAttribute=cn
ldap.roleObjectClass=group
ldap.useLdaps=true
ldap.useStartTLS=false
ldap.userBaseDn=dc=JNJ,dc=COM
ldap.userFilter=sAMAccountName={user}
ldap.userIdAttribute=sAMAccountName
ldap.userObjectClass=person
ldap.userPasswordAttribute=
ldap.userRdnAttribute=uid
```

Other properties:

```
dpm.alias.name.enabled=false
dpm.base.url=
dpm.enabled=false
dpm.remote.control.events.recipient=jobrunner-app
dpm.remote.control.job.labels=all
dpm.remote.control.ping.frequency=5000
http.access.control.allow.headers=origin,content-type,accept,authorization,x-
requested-by,x-ss-user-auth-token,x-ss-rest-call
http.access.control.allow.methods=GET,POST,PUT,DELETE,OPTIONS,HEAD
http.access.control.allow.origin=*
http.authentication=form
http.authentication.ldap.role.mapping=ITS-APP-EDL-DEV-ADM-DVLP-USR:admin;ITS-APP-
EDL-DEV-PROJ59-DVLP-USR:manager;ITS-APP-EDL-DEV-PROJ59-DVLP-USR:creator;ITS-APP-
EDL-DEV-PROJ59-SUPP-USR:guest;
http.authentication.login.module=ldap
http.port=18630
http.realm.file.permission.check=false
http.session.max.inactive.interval=86400
https.cluster.keystore.password=changeit
https.cluster.keystore.path=/opt/cloudera/security/jks/truststore.jks
https.cluster.truststore.password=
https.cluster.truststore.path=
https.keystore.password=Cloudera@5
https.keystore.path=/opt/cloudera/security/jks/sdc.keystore
https.port=18636
https.truststore.password=
https.truststore.path=
kerberos.client.enabled=true
mail.smtp.auth=false
mail.smtp.host=localhost
mail.smtp.port=25
mail.smtp.starttls.enable=false
mail.smtps.auth=false
mail.smtps.host=localhost
mail.smtps.port=25
mail.transport.protocol=smtp
max.stage.private.classloaders=50
monitor.memory=false
pipeline.access.control.enabled=true
preview.maxBatchSize=10
preview.maxBatches=10
production.maxBatchSize=1000
production.maxErrorRecordsPerStage=100
production.maxPipelineErrors=100
runner.thread.pool.size=50
stage.conf_hadoop.always.impersonate.current.user=false
ui.enable.webSocket=true
ui.header.title=
ui.undo.limit=10
vault.ssl.enabled.protocols=TLSv1.2,TLSv1.3
vault.ssl.verify=false
xmail.from.address=sdc@edldev.jnj.com
xmail.username=
kerberos.client.keytab=streamsets.keytab
clouderaManager.managed=true
ui.local.help.base.url=/docs
kerberos.client.principal=sdc/itsusralsp07814.jnj.com@JNJ.COM
```

Also symlink the logs and the data folders to a larger volume:

Logs:

```
#-> ls -al /var/log | grep sdc

lrwxrwxrwx   1 root  root       14 Mar 21 21:25 sdc -> /apps/log/sdc/
```

Data:

```
#-> ls -al /var/lib | grep sdc

lrwxrwxrwx   1 root        root         17 Mar 21 17:12 sdc -> /data/data03/sdc/
```

Version	DD-Mmm-YYYY	Author	Change Summary	Major/Minor Change
1.0	22nd March2019	Paul Zubin	Initial release of document.	Major