Cloudera Runtime 7.2.6

Apache Oozie Reference

Date published: 2021-01-12
Date modified:

https://docs.cloudera.com/

Legal Notice

Contents

Submit Oozie Jobs in Data Engineering Cluster

You can submit Apache Hive jobs or queries through JDBC, ODBC or Oozie in a Data Engineering (DE) cluster. Automated batch jobs require a non-interactive mechanism. You can use Cron, CDP CML, Amazon Web Services (AWS) Lambda or Fargate, Azure Functions, and many other mechanisms for automated batch jobs. Apache Oozie has built-in integration with the DEs cluster and is a better mechanism for these purposes.

This document explains how to create and use a CDP machine user account to submit Oozie jobs to a DE cluster through Knox Oozie endpoint.

Pre-requisites

- You must be a CDP user with a PowerUser role. For information about user roles and cluster access, see Enabling admin and user access to environments.
- You must have a CDP environment running. For information about the CDP environment, see Accessing an environment. For information about CDP CLI, see CDP CLI .

Steps

1. Create a CDP machine user account
2. Set the workload password for the CDP machine user
3. Give the CDP machine user the EnvironmentUser role for your CDP environment
4. Perform synchronization of users
5. Create a DE cluster
6. Get the Oozie endpoint from the gateway tab in the DE cluster details page
7. Setup Oozie artifacts
8. Configure IDBroker mappings
9. Submit the Oozie job

- Submit the Oozie job to the Knox endpoint
- Submit the Oozie job using the CLI

You can also see an example of how to submit a Spark job using Oozie.

Note: In the above procedure, the first six steps are performed by you with the PowerUser role of a CDP tenant. The seventh and eighth steps are performed by the machine user.

1. Create a machine user.

   Create a machine user named *job-submitter* and assign *srv_job-submitter* to the workloadUsername parameter.

```
$ cdp iam create-machine-user --machine-user-name job-submitter

{
    "machineUser": {
        "machineUserName": "job-submitter",
        "crn": "crn:altus:iam:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9ed
bb73d:machineUser:job-submitter/f622d289-56a1-4cc3-a14e-9c42b7d4988f",
        "creationDate": "2020-04-24T22:19:13.889000+00:00",
        "workloadUsername": "srv_job-submitter"
    }
}
```

2. Set the workload password for the machine user.

   Set the workload password for the machine user. The following example sets the password as *Password@123*. Use the *actor crn* from Step 1.

   ```
   $ cdp iam set-workload-password --password 'Password@123' --actor-crn 'c
   rn:altus:iam:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:machineUser:
   job-submitter/f622d289-56a1-4cc3-a14e-9c42b7d4988f'
   ```

3. Assign the machine user the EnvironmentUser role for your CDP environment. The Resource role crn does not change. Resource crn is the Environment crn.

   ```
   $ cdp iam assign-machine-user-resource-role --machine-user-name job-subm
   itter --resource-crn 'crn:cdp:environments:us-west-1:9d74eee4-1cad-45d7-
   b645-7ccf9edbb73d:environment:d84ca408-3dbd-4849-946f-c0c194f15429' --re
   source-role-crn 'crn:altus:iam:us-west-1:altus:resourceRole:EnvironmentU
   ser'
   ```

4. Perform synchronization of users.

   Synchronize the machine user to be part of the environment users. You must use the Environment name as the Environment crn in the below command.

   ```
   cdp environments sync-all-users --environment-name 'crn:cdp:environments
   :us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:environment:d84ca408-3db
   d-4849-946f-c0c194f15429'
   ```

5. Create a DE cluster.

   ```
   cdp datahub create-aws-cluster --cluster-name 'gk1de' --environment-name
    'spark-ql5h6p' --cluster-definition-name '7.2.6 - Data Engineering for
   AWS'
   ```

6. Get the Oozie endpoint.

   ```
   cdp datahub describe-cluster --cluster-name gk1de | jq '.cluster.endpoin
   ts.endpoints[] | select( .serviceName == "OOZIE_SERVER" ) | .serviceUrl'
   ```

   ```
   "https://gk1de-master3.gk1priv.a465-9q4k.cloudera.site/gk1de/cdp-proxy-ap
   i/oozie/"
   ```

7. Setup Oozie artifacts.

   Set up the hive.hql and workflow.xml files in AWS S3.

   - hive.hql contains the hive query you want to execute periodically
   - workflow.xml contains a boilerplate template to execute hive.hql
   - The parameters are controlled by job.xml submitted in Step 9.

   The value of the credential name="aws_auth" type="filesystem property is required for the Oozie logs to be saved in cloud storage.

   ```
   <workflow-app name="My Workflow" xmlns="uri:oozie:workflow:1.0">
     <credentials>
      <credential name="aws_auth" type="filesystem">
         <property>
           <name>filesystem.path</name>
           <value>s3a://gk1priv-cdp-bucket</value>
         </property>
       </credential>
       <credential name="hive2" type="hive2">
         <property>
           <name>hive2.jdbc.url</name>
   ```

```
        <value>${jdbcUrl}</value>
      </property>
      <property>
        <name>hive2.server.principal</name>
        <value>${hs2principal}</value>
      </property>
    </credential>
  </credentials>

    <start to="hive"/>

    <kill name="Kill">
        <message>Action failed, error message[${wf:errorMessage(wf:lastErr
orNode())}]</message>
    </kill>

    <action name="hive" cred="hive2,aws_auth">
        <hive2 xmlns="uri:oozie:hive2-action:1.0">
            <resource-manager>${resourceManager}</resource-manager>
            <name-node>${nameNode}</name-node>
            <jdbc-url>${jdbcUrl}</jdbc-url>
            <script>s3a://gk1priv-cdp-bucket/oozie/hive/hive.hql</script>
        </hive2>
        <ok to="End"/>
        <error to="Kill"/>
    </action>

    <end name="End"/>
</workflow-app>
```

In this example, the hive.hql file is setup at the *s3a://gk1priv-cdp-bucket/oozie/hive/hive.hql* location and this *workflow.xml* file is uploaded to *s3a://gk1priv-cdp-bucket/oozie/hive/*. The input and output S3 objects are in the same bucket. So, set the *s3a://gk1priv-cdp-bucket* in the credentials section of the bucket.

8. Configure IDBroker Mappings.
   a. Select your environment in Management Console > Environment.
   b. Click Actions > Manage Access > IDBroker Mappings.
   c. Click Edit.
   d. Add srv_job-submitter in User or Group.
   e. In Role, add the same role listed in the Data Access Role field.
   f. Click Save and Sync.
9. Submit the Oozie job

   You can submit the Oozie job using the knox endpoint or CLI.

   - Submit the Oozie job to the Knox endpoint.

     Submit the *srv_job-submitter* Oozie job. This is the workloadUserName from Step 1 and the password you have set in Step 2.

     For resourceManager, hs2principal, and jdbcUrl replace the hostname with the hostname in the Oozie endpoint. Specifically for the hs2principal the domain name is the upper cased sub-string (first label is not included) of the hostname in the Oozie endpoint.

     ```
     $ curl --request POST --location https://gk1de-master3.gk1priv.a465-9q4k
     .cloudera.site/gk1de/cdp-proxy-api/oozie/v1/jobs?action=start --header '
     Content-Type: application/xml;charset=UTF-8' -u 'srv_job-submitter:Passw
     ord@123' --data-raw '<?xml version="1.0" encoding="UTF-8"?>
     > <configuration>
     >     <property>
     >         <name>user.name</name>
     >         <value>srv_job-submitter</value>
     ```

```
>     </property>
>     <property>
>         <name>security_enabled</name>
>         <value>True</value>
>     </property>
>     <property>
>         <name>oozie.wf.application.path</name>
>         <value>s3a://gk1priv-cdp-bucket/oozie/hive</value>
>     </property>
>     <property>
>         <name>oozie.use.system.libpath</name>
>         <value>true</value>
>     </property>
>     <property>
>         <name>nameNode</name>
>         <value>will-be-replaced-by-knox</value>
>     </property>
>     <property>
>         <name>resourceManager</name>
>         <value>will-be-replaced-by-knox</value>
>     </property>
>     <property>
>         <name>jobTracker</name>
>         <value>will-be-replaced-by-knox</value>
>     </property>
>     <property>
>         <name>hs2principal</name>
>         <value>hive/gk1de-master3.gk1priv.a465-9q4k.cloudera.site@GK1
PRIV.A465-9Q4K.CLOUDERA.SITE</value>
>     </property>
>     <property>
>         <name>jdbcUrl</name>
>         <value>jdbc:hive2://gk1de-master3.gk1priv.a465-9q4k.clouder
a.site:2181/default;httpPath=cliservice;serviceDiscoveryMode=zooKeeper;s
sl=true;transportMode=http;zooKeeperNamespace=hiveserver2</value>
>     </property>
> </configuration>'
{"id":"0000002-200605195726045-oozie-oozi-W"}
```

Track the id 0000002-200605195726045-oozie-oozi-W for result

```
curl --request GET --location https://gk1de-master3.gk1priv.a465-9q4k.cl
oudera.site/gk1de/cdp-proxy-api/oozie/v1/job/0000002-200605195726045-ooz
ie-oozi-W?show=info -u 'srv_job-submitter:Password@123'
{"appName":"My Workflow","externalId":null,"conf":"<configuration>\r\n
 <property>\r\n    <name>oozie.action.sharelib.for.hive</name>\r\n    <v
alue>hive</value>\r\n  </property>\r\n  <property>\r\n    <name>oozie.wf
.application.path</name>\r\n    <value>s3a://gk1priv-cdp-bucket/oozie/hi
ve</value>\r\n  </property>\r\n  <property>\r\n    <name>oozie.use.syste
m.libpath</name>\r\n    <value>true</value>\r\n  </property>\r\n  <prope
rty>\r\n    <name>security_enabled</name>\r\n    <value>True</value>\r\n
  </property>\r\n  <property>\r\n    <name>jdbcUrl</name>\r\n    <value>
jdbc:hive2://gk1de-master3.gk1priv.a465-9q4k.cloudera.site:2181/default;
httpPath=cliservice;serviceDiscoveryMode=zooKeeper;ssl=true;transportMod
e=http;zooKeeperNamespace=hiveserver2</value>\r\n  </property>\r\n  <pro
perty>\r\n    <name>hs2principal</name>\r\n    <value>hive/gk1de-master3
.gk1priv.a465-9q4k.cloudera.site@GK1PRIV.A465-9Q4K.CLOUDERA.SITE</value>
\r\n  </property>\r\n  <property>\r\n    <name>user.name</name>\r\n    <
value>srv_job-submitter</value>\r\n  </property>\r\n  <property>\r\n
<name>jobTracker</name>\r\n    <value>gk1de-master3.gk1priv.a465-9q4k.cl
oudera.site:8032</value>\r\n  </property>\r\n  <property>\r\n    <name>m
apreduce.job.user.name</name>\r\n    <value>srv_job-submitter</value>\r\
n  </property>\r\n  <property>\r\n    <name>resourceManager</name>\r\n
```

```
  <value>gk1de-master3.gk1priv.a465-9q4k.cloudera.site:8032</value>\r\n
 </property>\r\n  <property>\r\n    <name>nameNode</name>\r\n    <value>
hdfs://gk1de-master3.gk1priv.a465-9q4k.cloudera.site:8020</value>\r\n  <
/property>\r\n</configuration>","run":0,"acl":null,"appPath":"s3a://gk1p
riv-cdp-bucket/oozie/hive","parentId":null,"lastModTime":"Sat, 06 Jun 20
20 00:25:15 GMT","consoleUrl":"https://gk1de-master3.gk1priv.a465-9q4k.c
loudera.site:11443/oozie?job=0000002-200605195726045-oozie-oozi-W","crea
tedTime":"Sat, 06 Jun 2020 00:25:02 GMT","startTime":"Sat, 06 Jun 2020 0
0:25:02 GMT","toString":"Workflow id[0000002-200605195726045-oozie-oozi-
W] status[SUCCEEDED]","id":"0000002-200605195726045-oozie-oozi-W","endTi
me":"Sat, 06 Jun 2020 00:25:15 GMT","user":"srv_job-submitter","actions"
:[{"cred":null,"userRetryMax":0,"trackerUri":"-","data":null,"errorMessa
ge":null,"userRetryCount":0,"externalChildIDs":null,"externalId":"-","er
rorCode":null,"conf":"","type":":START:","transition":"hive","retries":0
,"consoleUrl":"-","stats":null,"userRetryInterval":10,"name":":start:","
startTime":"Sat, 06 Jun 2020 00:25:02 GMT","toString":"Action name[:star
t:] status[OK]","id":"0000002-200605195726045-oozie-oozi-W@:start:","end
Time":"Sat, 06 Jun 2020 00:25:02 GMT","externalStatus":"OK","status":"OK
"},{"cred":"hive2","userRetryMax":0,"trackerUri":"gk1de-master3.gk1priv.
a465-9q4k.cloudera.site:8032","data":null,"errorMessage":null,"userRetry
Count":0,"externalChildIDs":null,"externalId":"application_1591385267488
_0010","errorCode":null,"conf":"<hive2 xmlns=\"uri:oozie:hive2-action:0.
1\">\r\n  <job-tracker>gk1de-master3.gk1priv.a465-9q4k.cloudera.site:803
2</job-tracker>\r\n  <name-node>hdfs://gk1de-master3.gk1priv.a465-9q4k.c
loudera.site:8020</name-node>\r\n  <jdbc-url>jdbc:hive2://gk1de-master3.
gk1priv.a465-9q4k.cloudera.site:2181/default;httpPath=cliservice;service
DiscoveryMode=zooKeeper;ssl=true;transportMode=http;zooKeeperNamespace=h
iveserver2</jdbc-url>\r\n  <script>s3a://gk1priv-cdp-bucket/oozie/hive/h
ive.hql</script>\r\n  <configuration />\r\n</hive2>","type":"hive2","tra
nsition":"End","retries":0,"consoleUrl":"https://gk1de-master3.gk1priv.a
465-9q4k.cloudera.site:8090/proxy/application_1591385267488_0010/","stat
s":null,"userRetryInterval":10,"name":"hive","startTime":"Sat, 06 Jun 20
20 00:25:02 GMT","toString":"Action name[hive] status[OK]","id":"0000002
-200605195726045-oozie-oozi-W@hive","endTime":"Sat, 06 Jun 2020 00:25:15
 GMT","externalStatus":"SUCCEEDED","status":"OK"},{"cred":null,"userRetr
yMax":0,"trackerUri":"-","data":null,"errorMessage":null,"userRetryCount
":0,"externalChildIDs":null,"externalId":"-","errorCode":null,"conf":"",
"type":":END:","transition":null,"retries":0,"consoleUrl":"-","stats":nu
ll,"userRetryInterval":10,"name":"End","startTime":"Sat, 06 Jun 2020 00:
25:15 GMT","toString":"Action name[End] status[OK]","id":"0000002-200605
195726045-oozie-oozi-W@End","endTime":"Sat, 06 Jun 2020 00:25:15 GMT","e
xternalStatus":"OK","status":"OK"}],"status":"SUCCEEDED","group":null}
```

- For Data Engineering HA clusters, you can use multiple ZooKeeper services with retries:

```
jdbc:hive2://imb6t-master0.gk1de.xcu2-8y8x.dev.cldr.work:2181,imb6t-mast
er1.gk1de.xcu2-8y8x.dev.cldr.work:2181,imb6t-masterx0.gk1de.xcu2-8y8x.de
v.cldr.work:2181/default;httpPath=cliservice;serviceDiscoveryMode=zooKee
per;ssl=true;transportMode=http;zooKeeperNamespace=hiveserver2;retries=5
```

- Submit the Oozie job using the CLI.

  You can also submit the Oozie job from inside the cluster using the CLI.

  a. Setup the job.properties file

  ```
  user.name=srv_job-submitter
  security_enabled=True
  oozie.wf.application.path=s3a://gk1priv-cdp-bucket/oozie/hive
  oozie.use.system.libpath=true
  nameNode=[Knox will replace this automatically]
  resourceManager=[Knox will replace this automatically]
  hs2principal=hive/gk1de-master3.gk1priv.a465-9q4k.cloudera.site@GK1
  PRIV.A465-9Q4K.CLOUDERA.SITE
  ```

```
jdbcUrl=jdbc:hive2://gk1de-master3.gk1priv.a465-9q4k.cloudera.site:
2181/default;httpPath=cliservice;serviceDiscoveryMode=zooKeeper;ssl=
true;transportMode=http;zooKeeperNamespace=hiveserver2
```

b. Run the Oozie command line. The workload username is *srv_job-submitter* and password is the same as that of the machine user.

```
oozie job -oozie {oozieKnoxUrl} -auth BASIC -username {workloadUser}
 -password {workloadPassword} -config job.properties -run
```

c. Check the status.

```
oozie job -oozie {oozieKnoxUrl} -auth BASIC -username {workloadUser}
 -password {workloadPassword} -info 0000002-200605195726045-oozie-oo
zi-W
```

Example - Spark submit using Oozie

```
<workflow-app name="My Workflow" xmlns="uri:oozie:workflow:1.0">
    <start to="spark-8c75"/>
    <kill name="Kill">
        <message>Action failed, error message[${wf:errorMessage(wf:lastErro
rNode())}]</message>
    </kill>
    <action name="spark-8c75">
        <spark xmlns="uri:oozie:spark-action:1.0">
           <resource-manager>${resourceManager}</resource-manager>
            <name-node>${nameNode}</name-node>
            <master>yarn</master>
            <mode>client</mode>
            <name></name>
              <class>org.apache.spark.examples.SparkPi</class>
            <jar>spark-examples_2.11-2.4.5.7.2.2.0-244.jar</jar>
            <file>s3a://gk1priv-cdp-bucket/spark-examples_2.11-2.4.5.7.2.2.
0-244.jar#spark-examples_2.11-2.4.5.7.2.2.0-244.jar</file>
        </spark>
        <ok to="End"/>
        <error to="Kill"/>
    </action>
    <end name="End"/>
</workflow-app>
```

To quickly get started with Oozie for other applications you could try the Hue Oozie editor. For more information about the Hue Oozie editor, see Hue Oozie Workflow Editor.