Configure Hadoop gateway (client or edge node) on rhldmsdev001 to rhldmsdev004:

1. Download hadoop-3.3.5.tar.gz from Apache Hadoop website
2. Untar Hadoop-3.3.5.tar.gz at /ptoan/shared/hadoop (nfs shared mounted disk)
3. Configured core-site.xml, hdfs-site.xml, yarn-site.xml, Hadoop-env.sh, mapred-site.xml and log4j.properties to point to the namenode and datanode cluster.
4. No need to start any datanode or yarn services on the edge nodes as it is only a gateway to the cluster.

**core-site.xml:**

<?xml version="1.0" encoding="UTF-8"?>

<!--Autogenerated by Cloudera Manager-->

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://nameservice1</value>

</property>

<property>

<name>ipc.client.connection.maxidletime</name>

<value>30000</value>

</property>

<property>

<name>ipc.client.connect.max.retries</name>

<value>50</value>

</property>

<property>

<name>fs.trash.interval</name>

<value>1</value>

</property>

<property>

<name>io.compression.codecs</name>

<value>org.apache.hadoop.io.compress.DefaultCodec,org.apache.hadoop.io.compress.GzipCodec,org.apache.hadoop.io.compress.BZip2Codec,org.apache.hadoop.io.compress.DeflateCodec,org.apache.hadoop.io.compress.SnappyCodec,org.apache.hadoop.io.compress.Lz4Codec</value>

</property>

<property>

<name>hadoop.security.authentication</name>

<value>simple</value>

</property>

<property>

<name>hadoop.security.authorization</name>

<value>false</value>

</property>

<property>

<name>hadoop.proxyuser.oozie.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.oozie.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.HTTP.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.HTTP.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.hive.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.hive.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.hue.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.hue.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.httpfs.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.httpfs.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.knox.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.knox.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.livy.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.livy.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.impala.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.impala.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.hdfs.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.hdfs.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.yarn.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.yarn.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.phoenix.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.phoenix.groups</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.kudu.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.kudu.groups</name>

<value>\*</value>

</property>

<property>

<name>net.topology.script.file.name</name>

<value>/ptoan/shared/hadoop/hadoop-3.3.5/etc/hadoop/topology.py</value>

</property>

<property>

<name>io.file.buffer.size</name>

<value>65536</value>

</property>

<property>

<name>hadoop.ssl.enabled</name>

<value>false</value>

</property>

</configuration>

**hdfs-site.xml:**

<?xml version="1.0" encoding="UTF-8"?>

<!--Autogenerated by Cloudera Manager-->

<configuration>

<property>

<name>dfs.nameservices</name>

<value>nameservice1</value>

</property>

<property>

<name>dfs.client.failover.proxy.provider.nameservice1</name>

<value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider</value>

</property>

<property>

<name>dfs.ha.automatic-failover.enabled.nameservice1</name>

<value>true</value>

</property>

<property>

<name>ha.zookeeper.quorum</name> <value>rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181</value>

</property>

<property>

<name>dfs.ha.namenodes.nameservice1</name>

<value>namenode1546438949,namenode1546439187</value>

</property>

<property>

<name>dfs.namenode.rpc-address.nameservice1.namenode1546438949</name>

<value>rhldatdms22003.fairisaac.com:8020</value>

</property>

<property>

<name>dfs.namenode.servicerpc-address.nameservice1.namenode1546438949</name>

<value>rhldatdms22003.fairisaac.com:8022</value>

</property>

<property>

<name>dfs.namenode.http-address.nameservice1.namenode1546438949</name>

<value>rhldatdms22003.fairisaac.com:9870</value>

</property>

<property>

<name>dfs.namenode.https-address.nameservice1.namenode1546438949</name>

<value>rhldatdms22003.fairisaac.com:9871</value>

</property>

<property>

<name>dfs.namenode.rpc-address.nameservice1.namenode1546439187</name>

<value>rhldatdms22004.fairisaac.com:8020</value>

</property>

<property>

<name>dfs.namenode.servicerpc-address.nameservice1.namenode1546439187</name>

<value>rhldatdms22004.fairisaac.com:8022</value>

</property>

<property>

<name>dfs.namenode.http-address.nameservice1.namenode1546439187</name>

<value>rhldatdms22004.fairisaac.com:9870</value>

</property>

<property>

<name>dfs.namenode.https-address.nameservice1.namenode1546439187</name>

<value>rhldatdms22004.fairisaac.com:9871</value>

</property>

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

<property>

<name>dfs.blocksize</name>

<value>134217728</value>

</property>

<property>

<name>dfs.client.use.datanode.hostname</name>

<value>false</value>

</property>

<property>

<name>fs.permissions.umask-mode</name>

<value>022</value>

</property>

<property>

<name>dfs.client.block.write.locateFollowingBlock.retries</name>

<value>7</value>

</property>

<property>

<name>dfs.client.read.shortcircuit.skip.checksum</name>

<value>false</value>

</property>

<property>

<name>dfs.client.domain.socket.data.traffic</name>

<value>false</value>

</property>

<property>

<name>dfs.datanode.hdfs-blocks-metadata.enabled</name>

<value>true</value>

</property>

</configuration>

**yarn-site.xml:**

<?xml version="1.0" encoding="UTF-8"?>

<!--Autogenerated by Cloudera Manager-->

<configuration>

<property>

<name>yarn.acl.enable</name>

<value>true</value>

</property>

<property>

<name>yarn.admin.acl</name>

<value>yarn</value>

</property>

<property>

<name>yarn.log-aggregation.file-formats</name>

<value>IFile,TFile</value>

</property>

<property>

<name>yarn.log-aggregation.file-controller.TFile.class</name>

<value>org.apache.hadoop.yarn.logaggregation.filecontroller.tfile.LogAggregationTFileController</value>

</property>

<property>

<name>yarn.log-aggregation.TFile.remote-app-log-dir-suffix</name>

<value></value>

</property>

<property>

<name>yarn.log-aggregation.file-controller.IFile.class</name>

<value>org.apache.hadoop.yarn.logaggregation.filecontroller.ifile.LogAggregationIndexedFileController</value>

</property>

<property>

<name>yarn.node-labels.enabled</name>

<value>true</value>

</property>

<property>

<name>yarn.resourcemanager.ha.enabled</name>

<value>true</value>

</property>

<property>

<name>yarn.resourcemanager.ha.automatic-failover.enabled</name>

<value>true</value>

</property>

<property>

<name>yarn.resourcemanager.ha.automatic-failover.embedded</name>

<value>true</value>

</property>

<property>

<name>yarn.resourcemanager.recovery.enabled</name>

<value>true</value>

</property>

<property>

<name>yarn.resourcemanager.zk-address</name> <value>rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181</value>

</property>

<property>

<name>yarn.resourcemanager.store.class</name>

<value>org.apache.hadoop.yarn.server.resourcemanager.recovery.ZKRMStateStore</value>

</property>

<property>

<name>yarn.client.failover-sleep-base-ms</name>

<value>100</value>

</property>

<property>

<name>yarn.client.failover-sleep-max-ms</name>

<value>2000</value>

</property>

<property>

<name>yarn.resourcemanager.cluster-id</name>

<value>yarnRM</value>

</property>

<property>

<name>yarn.resourcemanager.address.rm1546449790</name>

<value>rhldatdms22003.fairisaac.com:8032</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address.rm1546449790</name>

<value>rhldatdms22003.fairisaac.com:8030</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.address.rm1546449790</name>

<value>rhldatdms22003.fairisaac.com:8031</value>

</property>

<property>

<name>yarn.resourcemanager.admin.address.rm1546449790</name>

<value>rhldatdms22003.fairisaac.com:8033</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.address.rm1546449790</name>

<value>rhldatdms22003.fairisaac.com:8088</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.https.address.rm1546449790</name>

<value>rhldatdms22003.fairisaac.com:8090</value>

</property>

<property>

<name>yarn.resourcemanager.address.rm1546449814</name>

<value>rhldatdms22004.fairisaac.com:8032</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address.rm1546449814</name>

<value>rhldatdms22004.fairisaac.com:8030</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.address.rm1546449814</name>

<value>rhldatdms22004.fairisaac.com:8031</value>

</property>

<property>

<name>yarn.resourcemanager.admin.address.rm1546449814</name>

<value>rhldatdms22004.fairisaac.com:8033</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.address.rm1546449814</name>

<value>rhldatdms22004.fairisaac.com:8088</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.https.address.rm1546449814</name>

<value>rhldatdms22004.fairisaac.com:8090</value>

</property>

<property>

<name>yarn.resourcemanager.ha.rm-ids</name>

<value>rm1546449790,rm1546449814</value>

</property>

<property>

<name>yarn.resourcemanager.client.thread-count</name>

<value>50</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.client.thread-count</name>

<value>50</value>

</property>

<property>

<name>yarn.resourcemanager.admin.client.thread-count</name>

<value>1</value>

</property>

<property>

<name>yarn.scheduler.minimum-allocation-mb</name>

<value>1024</value>

</property>

<property>

<name>yarn.scheduler.increment-allocation-mb</name>

<value>512</value>

</property>

<property>

<name>yarn.scheduler.maximum-allocation-mb</name>

<value>327680</value>

</property>

<property>

<name>yarn.scheduler.minimum-allocation-vcores</name>

<value>1</value>

</property>

<property>

<name>yarn.scheduler.increment-allocation-vcores</name>

<value>1</value>

</property>

<property>

<name>yarn.scheduler.maximum-allocation-vcores</name>

<value>160</value>

</property>

<property>

<name>yarn.resourcemanager.amliveliness-monitor.interval-ms</name>

<value>1000</value>

</property>

<property>

<name>yarn.am.liveness-monitor.expiry-interval-ms</name>

<value>600000</value>

</property>

<property>

<name>yarn.resourcemanager.am.max-attempts</name>

<value>2</value>

</property>

<property>

<name>yarn.resourcemanager.container.liveness-monitor.interval-ms</name>

<value>600000</value>

</property>

<property>

<name>yarn.resourcemanager.nm.liveness-monitor.interval-ms</name>

<value>1000</value>

</property>

<property>

<name>yarn.nm.liveness-monitor.expiry-interval-ms</name>

<value>600000</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.client.thread-count</name>

<value>50</value>

</property>

<property>

<name>yarn.application.classpath</name> <value>$HADOOP\_CLIENT\_CONF\_DIR,$HADOOP\_COMMON\_HOME/\*,$HADOOP\_COMMON\_HOME/lib/\*,$HADOOP\_HDFS\_HOME/\*,$HADOOP\_HDFS\_HOME/lib/\*,$HADOOP\_YARN\_HOME/\*,$HADOOP\_YARN\_HOME/lib/\*,$HADOOP\_MAPRED\_HOME/\*,$HADOOP\_MAPRED\_HOME/lib/\*</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.class</name>

<value>org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler</value>

</property>

<property>

<name>yarn.scheduler.capacity.resource-calculator</name>

<value>org.apache.hadoop.yarn.util.resource.DominantResourceCalculator</value>

</property>

<property>

<name>yarn.resourcemanager.max-completed-applications</name>

<value>10000</value>

</property>

<property>

<name>yarn.nodemanager.remote-app-log-dir</name>

<value>/tmp/logs</value>

</property>

<property>

<name>yarn.nodemanager.remote-app-log-dir-suffix</name>

<value>logs</value>

</property>

<property>

<name>yarn.nodemanager.vmem-check-enabled</name>

<value>false</value>

</property>

<property>

<name>yarn.resourcemanager.placement-constraints.handler</name>

<value>scheduler</value>

</property>

<property>

<name>hadoop.registry.zk.quorum</name> <value>rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181</value>

</property>

<property>

<name>yarn.service.framework.path</name>

<value>/user/yarn/services/service-framework/7.1.7/service-dep.tar.gz</value>

</property>

<property>

<name>yarn.service.classpath</name>

<value>$HADOOP\_CLIENT\_CONF\_DIR</value>

</property>

</configuration>

**mapred-site.xml:**

<?xml version="1.0" encoding="UTF-8"?>

<!--Autogenerated by Cloudera Manager-->

<configuration>

<property>

<name>mapreduce.job.split.metainfo.maxsize</name>

<value>10000000</value>

</property>

<property>

<name>mapreduce.job.counters.max</name>

<value>120</value>

</property>

<property>

<name>mapreduce.job.counters.groups.max</name>

<value>50</value>

</property>

<property>

<name>mapreduce.output.fileoutputformat.compress</name>

<value>false</value>

</property>

<property>

<name>mapreduce.output.fileoutputformat.compress.type</name>

<value>BLOCK</value>

</property>

<property>

<name>mapreduce.output.fileoutputformat.compress.codec</name>

<value>org.apache.hadoop.io.compress.DefaultCodec</value>

</property>

<property>

<name>mapreduce.map.output.compress.codec</name>

<value>org.apache.hadoop.io.compress.SnappyCodec</value>

</property>

<property>

<name>mapreduce.map.output.compress</name>

<value>true</value>

</property>

<property>

<name>zlib.compress.level</name>

<value>DEFAULT\_COMPRESSION</value>

</property>

<property>

<name>mapreduce.task.io.sort.factor</name>

<value>64</value>

</property>

<property>

<name>mapreduce.map.sort.spill.percent</name>

<value>0.8</value>

</property>

<property>

<name>mapreduce.reduce.shuffle.parallelcopies</name>

<value>10</value>

</property>

<property>

<name>mapreduce.task.timeout</name>

<value>600000</value>

</property>

<property>

<name>mapreduce.client.submit.file.replication</name>

<value>3</value>

</property>

<property>

<name>mapreduce.job.reduces</name>

<value>400</value>

</property>

<property>

<name>mapreduce.task.io.sort.mb</name>

<value>256</value>

</property>

<property>

<name>mapreduce.map.speculative</name>

<value>false</value>

</property>

<property>

<name>mapreduce.reduce.speculative</name>

<value>false</value>

</property>

<property>

<name>mapreduce.job.reduce.slowstart.completedmaps</name>

<value>0.8</value>

</property>

<property>

<name>mapreduce.jobhistory.address</name>

<value>rhldatdms22003.fairisaac.com:10020</value>

</property>

<property>

<name>mapreduce.jobhistory.webapp.address</name>

<value>rhldatdms22003.fairisaac.com:19888</value>

</property>

<property>

<name>mapreduce.jobhistory.webapp.https.address</name>

<value>rhldatdms22003.fairisaac.com:19890</value>

</property>

<property>

<name>mapreduce.jobhistory.admin.address</name>

<value>rhldatdms22003.fairisaac.com:10033</value>

</property>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

<property>

<name>yarn.app.mapreduce.am.staging-dir</name>

<value>/user</value>

</property>

<property>

<name>mapreduce.am.max-attempts</name>

<value>2</value>

</property>

<property>

<name>yarn.app.mapreduce.am.resource.mb</name>

<value>1024</value>

</property>

<property>

<name>yarn.app.mapreduce.am.resource.cpu-vcores</name>

<value>1</value>

</property>

<property>

<name>mapreduce.job.ubertask.enable</name>

<value>false</value>

</property>

<property>

<name>yarn.app.mapreduce.am.command-opts</name>

<value>-Djava.net.preferIPv4Stack=true -Xmx825955249</value>

</property>

<property>

<name>mapreduce.map.java.opts</name>

<value>-Djava.net.preferIPv4Stack=true</value>

</property>

<property>

<name>mapreduce.reduce.java.opts</name>

<value>-Djava.net.preferIPv4Stack=true</value>

</property>

<property>

<name>yarn.app.mapreduce.am.admin.user.env</name>

<value>LD\_LIBRARY\_PATH=$HADOOP\_COMMON\_HOME/lib/native:$JAVA\_LIBRARY\_PATH</value>

</property>

<property>

<name>mapreduce.map.memory.mb</name>

<value>1024</value>

</property>

<property>

<name>mapreduce.map.cpu.vcores</name>

<value>1</value>

</property>

<property>

<name>mapreduce.reduce.memory.mb</name>

<value>1024</value>

</property>

<property>

<name>mapreduce.reduce.cpu.vcores</name>

<value>1</value>

</property>

<property>

<name>mapreduce.job.heap.memory-mb.ratio</name>

<value>0.8</value>

</property>

<property>

<name>mapreduce.application.classpath</name>

<value>$HADOOP\_CLIENT\_CONF\_DIR,$PWD/mr-framework/\*,$MR2\_CLASSPATH</value>

</property>

<property>

<name>mapreduce.application.framework.path</name>

<value>hdfs://nameservice1//user/yarn/mapreduce/mr-framework/3.1.1.7.1.7.0-551-mr-framework.tar.gz#mr-framework</value>

</property>

<property>

<name>mapreduce.jobhistory.jhist.format</name>

<value>binary</value>

</property>

<property>

<name>mapreduce.admin.user.env</name>

<value>LD\_LIBRARY\_PATH=$HADOOP\_COMMON\_HOME/lib/native:$JAVA\_LIBRARY\_PATH</value>

</property>

<property>

<name>mapreduce.job.redacted-properties</name> <value>fs.s3a.access.key,fs.s3a.secret.key,fs.adl.oauth2.credential,dfs.adls.oauth2.credential,fs.azure.account.oauth2.client.secret</value>

</property>

<property>

<name>mapreduce.job.acl-view-job</name>

<value> </value>

</property>

<property>

<name>mapreduce.job.acl-modify-job</name>

<value> </value>

</property>

<property>

<name>mapreduce.cluster.acls.enabled</name>

<value>false</value>

</property>

<property>

<name>mapreduce.shuffle.ssl.enabled</name>

<value>false</value>

</property>

</configuration>

**hadoop-env.sh:**

# Prepend/Append plugin parcel classpaths

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.362.b08-1.el7\_9.x86\_64/jre

export HADOOP\_HOME=/ptoan/shared/hadoop/hadoop-3.3.5

export HADOOP\_COMMON\_HOME=/ptoan/shared/hadoop/hadoop-3.3.5

export LD\_LIBRARY\_PATH=$LA\_LIBRARY\_PATH:/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

if [ "$HADOOP\_USER\_CLASSPATH\_FIRST" = 'true' ]; then

# HADOOP\_CLASSPATH={{HADOOP\_CLASSPATH\_APPEND}}

:

else

# HADOOP\_CLASSPATH=

:

fi

# JAVA\_LIBRARY\_PATH=

export YARN\_ROOT\_LOGGER="WARN,console"

export HADOOP\_MAPRED\_HOME=/ptoan/shared/hadoop/hadoop-3.3.5 🡨 this command needed for hive mr

export HADOOP\_OPTS="-Dhadoop.root.logger=WARN $HADOOP\_OPTS"

export YARN\_OPTS="-Xmx825955249 -Djava.net.preferIPv4Stack=true -Dhadoop.root.logger=${YARN\_ROOT\_LOGGER:-WARN,console} -Dyarn.root.logger=${YARN\_ROOT\_LOGGER:-WARN,console} $YARN\_OPTS"

export HADOOP\_CLIENT\_OPTS="-Djava.net.preferIPv4Stack=true $HADOOP\_CLIENT\_OPTS"

**log4j.properties:**

hadoop.root.logger=WARN,console 🡨 added this

yarn.root.logger=WARN,console 🡨 added this; also changed everything to WARN

**topology.map:**

[network\_topology]

rhldatdms22003.fairisaac.com=/default

10.109.15.13=/default

rhldatdms22004.fairisaac.com=/default

10.109.15.14=/default

rhlmodhdp22501.fairisaac.com=/default

10.109.15.11=/default

rhlmodhdp22502.fairisaac.com=/default

10.109.15.2=/default

rhlmodhdp22503.fairisaac.com=/default

10.109.15.3=/default

rhlmodhdp22504.fairisaac.com=/default

10.109.15.4=/default

rhlmodhdp22505.fairisaac.com=/default

10.109.15.5=/default

rhlmodhdp22506.fairisaac.com=/default

10.109.15.6=/default

rhlmodhdp22507.fairisaac.com=/default

10.109.15.7=/default

rhlmodhdp22508.fairisaac.com=/default

10.109.15.8=/default

rhlmodhdp22509.fairisaac.com=/default

10.109.15.9=/default

rhlmodhdp22510.fairisaac.com=/default

10.109.15.10=/default

**topology.py:**

#!/usr/bin/env python

#

# Copyright (c) 2021 Cloudera, Inc. All rights reserved.

#

'''

This script is provided by CMF for hadoop to determine network/rack topology.

It is automatically generated and could be replaced at any time. Any changes

made to it will be lost when this happens.

'''

import sys, os

# Support python2 and python3

try:

import ConfigParser

except:

import configparser as ConfigParser

DATA\_FILE\_NAME = '/etc/spark/conf.cloudera.spark\_on\_yarn/yarn-conf/topology.map'

DEFAULT\_RACK = '/default'

SECTION\_NAME = "network\_topology"

if 'CMF\_CONF\_DIR' in DATA\_FILE\_NAME:

# variable was not substituted. Use this file's dir

DATA\_FILE\_NAME = os.path.dirname(os.path.abspath(\_\_file\_\_)) + "/topology.map"

class TopologyScript():

def load\_rack\_map(self):

try:

#RACK\_MAP contains both host name vs rack and ip vs rack mappings

mappings = ConfigParser.ConfigParser()

mappings.read(DATA\_FILE\_NAME)

return dict(mappings.items(SECTION\_NAME))

except ConfigParser.NoSectionError:

return {}

def get\_racks(self, rack\_map, args):

if len(args) == 1:

return DEFAULT\_RACK

else:

return " ".join([self.lookup\_by\_hostname\_or\_ip(input\_argument, rack\_map) for input\_argument in args[1:]],)

def lookup\_by\_hostname\_or\_ip(self, hostname\_or\_ip, rack\_map):

#try looking up by hostname

rack = rack\_map.get(hostname\_or\_ip)

if rack is not None:

return rack

#try looking up by ip

rack = rack\_map.get(self.extract\_ip(hostname\_or\_ip))

#try by localhost since hadoop could be passing in 127.0.0.1 which might not be mapped

return rack if rack is not None else rack\_map.get("localhost.localdomain", DEFAULT\_RACK)

#strips out port and slashes in case hadoop passes in something like 127.0.0.1/127.0.0.1:50010

def extract\_ip(self, container\_string):

return container\_string.split("/")[0].split(":")[0]

def execute(self, args):

rack\_map = self.load\_rack\_map()

rack = self.get\_racks(rack\_map, args)

print(rack)

if \_\_name\_\_ == "\_\_main\_\_":

TopologyScript().execute(sys.argv)

A screenshot of a computer screen

Description automatically generated with medium confidence

Spark-3.0.1-bin-without-hadoop:

**spark-defaults.conf:**

spark.authenticate=false

spark.driver.log.dfsDir=/user/spark/driverLogs

spark.driver.log.persistToDfs.enabled=true

spark.dynamicAllocation.enabled=true

spark.dynamicAllocation.executorIdleTimeout=60

spark.dynamicAllocation.minExecutors=0

spark.dynamicAllocation.schedulerBacklogTimeout=1

spark.dynamicAllocation.shuffleTracking.enabled=true

spark.shuffle.tracking.enabled=true

spark.eventLog.enabled=true

spark.io.encryption.enabled=false

spark.lineage.enabled=true

spark.network.crypto.enabled=false

spark.serializer=org.apache.spark.serializer.KryoSerializer

spark.shuffle.service.enabled=false

spark.shuffle.service.port=7337

spark.ui.enabled=true

spark.ui.killEnabled=true

spark.master=yarn

spark.submit.deployMode=client

spark.eventLog.dir=hdfs://nameservice1/user/spark/applicationHistory

spark.yarn.historyServer.address=http://rhldatdms22003.fairisaac.com:18088

spark.yarn.jars=local:/ptoan/shared/opt/spark/spark-3.0.1-bin-without-hadoop/jars/\*,local:/ptoan/shared/jyc\_temp/\*

spark.driver.extraClassPath=file:///ptoan/shared/opt/spark/spark-3.0.1-bin-without-hadoop/jars/\*,file:///ptoan/shared/jyc\_temp/\*

spark.executor.extraClassPath=file:///ptoan/shared/opt/spark/spark-3.0.1-bin-without-hadoop/jars/\*,file:///ptoan/shared/jyc\_temp/\*

spark.driver.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.executor.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.yarn.am.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.yarn.config.gatewayPath=/ptoan/shared/hadoop

spark.yarn.config.replacementPath={{HADOOP\_HOME}}/../../..

spark.yarn.historyServer.allowTracking=true

spark.yarn.appMasterEnv.MKL\_NUM\_THREADS=1

spark.executorEnv.MKL\_NUM\_THREADS=1

spark.yarn.appMasterEnv.OPENBLAS\_NUM\_THREADS=1

spark.executorEnv.OPENBLAS\_NUM\_THREADS=1

spark.hadoop.fs.s3a.committer.name=directory

**spark-env.sh:**

#!/usr/bin/env bash

SELF="$(cd $(dirname $BASH\_SOURCE) && pwd)"

if [ -z "$SPARK\_CONF\_DIR" ]; then

export SPARK\_CONF\_DIR="$SELF"

fi

SPARK\_HOME=/ptoan/shared/hadoop/spark-3.0.1-bin-without-hadoop

export SPARK\_HOME

export SPARK\_CONF\_DIR=$SPARK\_HOME/conf

export HADOOP\_HOME=/ptoan/shared/hadoop/hadoop-3.3.5

export HADOOP\_COMMON\_HOME="$HADOOP\_HOME"

export YARN\_CONF\_DIR=$SPARK\_HOME/conf/yarn-conf

MASTER="yarn-cluster"

if [ -n "$HADOOP\_HOME" ]; then

LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH:${HADOOP\_HOME}/lib/native

fi

SPARK\_EXTRA\_LIB\_PATH=""

if [ -n "$SPARK\_EXTRA\_LIB\_PATH" ]; then

LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH:$SPARK\_EXTRA\_LIB\_PATH

fi

export LD\_LIBRARY\_PATH

HADOOP\_CONF\_DIR=${HADOOP\_CONF\_DIR:-$SPARK\_CONF\_DIR/yarn-conf}

HIVE\_CONF\_DIR=${HIVE\_CONF\_DIR:-$SPARK\_CONF\_DIR}

if [ -d "$HIVE\_CONF\_DIR" ]; then

HADOOP\_CONF\_DIR="$HADOOP\_CONF\_DIR:$HIVE\_CONF\_DIR"

fi

export HADOOP\_CONF\_DIR

# https://spark.apache.org/docs/3.0.1/hadoop-provided.html

export SPARK\_DIST\_CLASSPATH=$(hadoop classpath)

# Force single-threaded BLAS (CDH-58082).

export MKL\_NUM\_THREADS=${MKL\_NUM\_THREADS:-1}

export OPENBLAS\_NUM\_THREADS=${OPENBLAS\_NUM\_THREADS:-1}

**classpath.txt:**

/ptoan/shared/hadoop/hadoop-3.3.5/share/hadoop/client

------------------------------------------------------------------------------------------------------------------------------

**spark-3.3.0-bin-hadoop3: (somehow spark 3.3.0 without hadoop does not work)**

spark-defaults.conf:

spark.authenticate=false

spark.driver.log.dfsDir=/user/spark/driver3Logs

spark.driver.log.persistToDfs.enabled=true

spark.dynamicAllocation.enabled=true

spark.dynamicAllocation.executorIdleTimeout=60

spark.dynamicAllocation.minExecutors=0

spark.dynamicAllocation.schedulerBacklogTimeout=1

spark.dynamicAllocation.shuffleTracking.enabled=true

spark.shuffle.tracking.enabled=true

spark.eventLog.enabled=true

spark.io.encryption.enabled=false

spark.lineage.enabled=true

spark.network.crypto.enabled=false

spark.serializer=org.apache.spark.serializer.KryoSerializer

spark.shuffle.service.enabled=false

spark.shuffle.service.port=7337

spark.ui.enabled=true

spark.ui.killEnabled=true

spark.master=yarn

spark.submit.deployMode=client

spark.eventLog.dir=hdfs://nameservice1/user/spark/spark3ApplicationHistory

spark.yarn.historyServer.address=http://rhlmodhdp22506.fairisaac.com:18088

spark.yarn.populateHadoopClasspath=true 🡨 added this because this spark contains hadoop runtime libraries already and want yarn to prevent jar conflict; For with-hadoop Spark distribution, since it contains a built-in Hadoop runtime already, by default, when a job is submitted to Hadoop Yarn cluster, to prevent jar conflict, **it will not populate Yarn’s classpath into Spark**. To override this behavior, set spark.yarn.populateHadoopClasspath to true.

#spark.yarn.jars=local:/ptoan/shared/hadoop/spark-3.3.0-bin-hadoop3/jars/\*.jar 🡨 had to take this line out because spark-shell keeps giving errors saying that it cannot find or load main class org.apache.spark.deploy.yarn.ExecutorLauncher

spark.driver.extraClassPath=file:///ptoan/shared/hadoop/spark-3.3.0-bin-hadoop3/jars/\*.jar

spark.executor.extraClassPath=file:///ptoan/shared/hadoop/spark-3.3.0-bin-hadoop3/jars/\*.jar

spark.driver.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.executor.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.yarn.am.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.yarn.config.gatewayPath=/ptoan/shared/hadoop

spark.yarn.config.replacementPath={{HADOOP\_HOME}}/../../..

spark.yarn.historyServer.allowTracking=true

spark.yarn.appMasterEnv.MKL\_NUM\_THREADS=1

spark.executorEnv.MKL\_NUM\_THREADS=1

spark.yarn.appMasterEnv.OPENBLAS\_NUM\_THREADS=1

spark.executorEnv.OPENBLAS\_NUM\_THREADS=1

spark.hadoop.fs.s3a.committer.name=directory

spark-env.sh:

SELF="$(cd $(dirname $BASH\_SOURCE) && pwd)"

if [ -z "$SPARK\_CONF\_DIR" ]; then

export SPARK\_CONF\_DIR="$SELF"

fi

SPARK\_HOME=/ptoan/shared/hadoop/spark-3.3.0-bin-hadoop3

export SPARK\_HOME

export SPARK\_CONF\_DIR=$SPARK\_HOME/conf

export HADOOP\_HOME=/ptoan/shared/hadoop/hadoop-3.3.5

export HADOOP\_COMMON\_HOME="$HADOOP\_HOME"

#export HADOOP\_CONF\_DIR=${HADOOP\_HOME}/etc/hadoop

#export YARN\_CONF\_DIR=$SPARK\_HOME/conf/yarn-conf

export YARN\_CONF\_DIR=${SPARK\_HOME}/conf/yarn-conf

#MASTER="yarn-cluster"

if [ -n "$HADOOP\_HOME" ]; then

LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH:${HADOOP\_HOME}/lib/native

fi

SPARK\_EXTRA\_LIB\_PATH=""

if [ -n "$SPARK\_EXTRA\_LIB\_PATH" ]; then

LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH:$SPARK\_EXTRA\_LIB\_PATH

fi

export LD\_LIBRARY\_PATH

HADOOP\_CONF\_DIR=${HADOOP\_CONF\_DIR:-$SPARK\_CONF\_DIR/yarn-conf}

HIVE\_CONF\_DIR=${HIVE\_CONF\_DIR:-$SPARK\_CONF\_DIR}

if [ -d "$HIVE\_CONF\_DIR" ]; then

HADOOP\_CONF\_DIR="$HADOOP\_CONF\_DIR:$HIVE\_CONF\_DIR"

fi

export HADOOP\_CONF\_DIR=/etc/spark/conf.cloudera.spark\_on\_yarn/yarn-conf:/etc/hive/conf:/etc/hadoop/conf.cloudera.yarn

# https://spark.apache.org/docs/3.0.1/hadoop-provided.html

export SPARK\_DIST\_CLASSPATH=$(/ptoan/shared/hadoop/hadoop-3.3.5/bin/hadoop classpath)

# Force single-threaded BLAS (CDH-58082).

export MKL\_NUM\_THREADS=${MKL\_NUM\_THREADS:-1}

export OPENBLAS\_NUM\_THREADS=${OPENBLAS\_NUM\_THREADS:-1}

----------------------------------------------------------------------------------------------------------------------------------

**spark-3.3.0-bin-without-hadoop:**

spark-defaults.conf:

spark.authenticate=false

spark.driver.log.dfsDir=/user/spark/driver3Logs

spark.driver.log.persistToDfs.enabled=true

spark.dynamicAllocation.enabled=true

spark.dynamicAllocation.executorIdleTimeout=60

spark.dynamicAllocation.minExecutors=0

spark.dynamicAllocation.schedulerBacklogTimeout=1

spark.dynamicAllocation.shuffleTracking.enabled=true

spark.shuffle.tracking.enabled=true

spark.eventLog.enabled=true

spark.io.encryption.enabled=false

spark.lineage.enabled=true

spark.network.crypto.enabled=false

spark.serializer=org.apache.spark.serializer.KryoSerializer

spark.shuffle.service.enabled=false

spark.shuffle.service.port=7337

spark.ui.enabled=true

spark.ui.killEnabled=true

spark.master=yarn

spark.submit.deployMode=client

spark.eventLog.dir=hdfs://nameservice1/user/spark/spark3ApplicationHistory

spark.yarn.historyServer.address=http://rhlmodhdp22506.fairisaac.com:18088

spark.yarn.populateHadoopClasspath=true 🡨 add this line

#spark.yarn.jars=local:/ptoan/shared/opt/spark/spark-3.3.0-bin-without-hadoop/jars/\*,local:/ptoan/shared/jyc\_temp\_spark330/\* 🡨 took this one out because got error that it could not find or load main class org.apache.spark.deploy.yarn.ExecuteLauncher or missing sl4j logging static binder

spark.driver.extraClassPath=file:///ptoan/shared/hadoop/spark-3.3.0-bin-without-hadoop/jars/\*,file:///ptoan/shared/jyc\_temp\_spark330/\*

spark.executor.extraClassPath=file:///ptoan/shared/hadoop/spark-3.3.0-bin-without-hadoop/jars/\*,file:///ptoan/shared/jyc\_temp\_spark330/\*

spark.driver.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.executor.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.yarn.am.extraLibraryPath=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native

spark.yarn.config.gatewayPath=/ptoan/shared/hadoop

spark.yarn.config.replacementPath={{HADOOP\_HOME}}/../../..

spark.yarn.historyServer.allowTracking=true

spark.yarn.appMasterEnv.MKL\_NUM\_THREADS=1

spark.executorEnv.MKL\_NUM\_THREADS=1

spark.yarn.appMasterEnv.OPENBLAS\_NUM\_THREADS=1

spark.executorEnv.OPENBLAS\_NUM\_THREADS=1

spark.hadoop.fs.s3a.committer.name=directory

spark-env.sh:

SELF="$(cd $(dirname $BASH\_SOURCE) && pwd)"

if [ -z "$SPARK\_CONF\_DIR" ]; then

export SPARK\_CONF\_DIR="$SELF"

fi

SPARK\_HOME=/ptoan/shared/hadoop/spark-3.3.0-bin-without-hadoop

export SPARK\_HOME

export SPARK\_CONF\_DIR=$SPARK\_HOME/conf

export HADOOP\_HOME=/ptoan/shared/hadoop/hadoop-3.3.5

export HADOOP\_COMMON\_HOME="$HADOOP\_HOME"

#export HADOOP\_CONF\_DIR=${HADOOP\_HOME}/etc/hadoop

#export YARN\_CONF\_DIR=$SPARK\_HOME/conf/yarn-conf

export YARN\_CONF\_DIR=${SPARK\_HOME}/conf/yarn-conf

#MASTER="yarn-cluster"

if [ -n "$HADOOP\_HOME" ]; then

LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH:${HADOOP\_HOME}/lib/native

fi

SPARK\_EXTRA\_LIB\_PATH=""

if [ -n "$SPARK\_EXTRA\_LIB\_PATH" ]; then

LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH:$SPARK\_EXTRA\_LIB\_PATH

fi

export LD\_LIBRARY\_PATH

HADOOP\_CONF\_DIR=${HADOOP\_CONF\_DIR:-$SPARK\_CONF\_DIR/yarn-conf}

HIVE\_CONF\_DIR=${HIVE\_CONF\_DIR:-$SPARK\_CONF\_DIR}

if [ -d "$HIVE\_CONF\_DIR" ]; then

HADOOP\_CONF\_DIR="$HADOOP\_CONF\_DIR:$HIVE\_CONF\_DIR"

fi

export HADOOP\_CONF\_DIR=/etc/spark/conf.cloudera.spark\_on\_yarn/yarn-conf:/etc/hive/conf:/etc/hadoop/conf.cloudera.yarn

# https://spark.apache.org/docs/3.0.1/hadoop-provided.html

export SPARK\_DIST\_CLASSPATH=$(/ptoan/shared/hadoop/hadoop-3.3.5/bin/hadoop classpath)

# Force single-threaded BLAS (CDH-58082).

export MKL\_NUM\_THREADS=${MKL\_NUM\_THREADS:-1}

export OPENBLAS\_NUM\_THREADS=${OPENBLAS\_NUM\_THREADS:-1}

A screen shot of a computer code

Description automatically generated with medium confidence

Recreate symbolic link for yarn-conf and hive-site.xml to point to /etc/hive/conf.cloudera.hive/hive-site.xml and /etc/spark/conf.cloudera.spark\_on\_yarn/yarn-conf

--------------------------------------------------------------------------------------------------------------------------------------

apache-hive-3.1.3-bin:

hive-site.xml:

<?xml version="1.0" encoding="UTF-8"?>

<configuration>

<property>

<name>hive.metastore.client.socket.timeout</name>

<value>300</value>

</property>

<property>

<name>hive.metastore.warehouse.dir</name>

<value>/warehouse/tablespace/managed/hive</value>

</property>

<property>

<name>hive.metastore.warehouse.external.dir</name>

<value>/warehouse/tablespace/external/hive</value>

</property>

<property>

<name>hive.server2.transport.mode</name>

<value>binary</value>

</property>

<property>

<name>hive.auto.convert.join</name>

<value>true</value>

</property>

<property>

<name>hive.auto.convert.join.noconditionaltask.size</name>

<value>52428800</value>

</property>

<property>

<name>hive.optimize.bucketmapjoin.sortedmerge</name>

<value>false</value>

</property>

<property>

<name>hive.smbjoin.cache.rows</name>

<value>10000</value>

</property>

<property>

<name>hive.server2.logging.operation.enabled</name>

<value>true</value>

</property>

<property>

<name>hive.server2.logging.operation.log.location</name>

<value>/var/log/hive/operation\_logs</value>

</property>

<property>

<name>mapred.reduce.tasks</name>

<value>-1</value>

</property>

<property>

<name>hive.exec.reducers.bytes.per.reducer</name>

<value>67108864</value>

</property>

<property>

<name>hive.exec.copyfile.maxsize</name>

<value>33554432</value>

</property>

<property>

<name>hive.exec.reducers.max</name>

<value>1009</value>

</property>

<property>

<name>hive.vectorized.groupby.checkinterval</name>

<value>4096</value>

</property>

<property>

<name>hive.vectorized.groupby.flush.percent</name>

<value>0.1</value>

</property>

<property>

<name>hive.compute.query.using.stats</name>

<value>true</value>

</property>

<property>

<name>hive.vectorized.execution.enabled</name>

<value>true</value>

</property>

<property>

<name>hive.vectorized.execution.reduce.enabled</name>

<value>true</value>

</property>

<property>

<name>hive.vectorized.use.vectorized.input.format</name>

<value>true</value>

</property>

<property>

<name>hive.vectorized.use.checked.expressions</name>

<value>true</value>

</property>

<property>

<name>hive.vectorized.use.vector.serde.deserialize</name>

<value>false</value>

</property>

<property>

<name>hive.vectorized.adaptor.usage.mode</name>

<value>chosen</value>

</property>

<property>

<name>hive.vectorized.input.format.excludes</name>

<value></value>

</property>

<property>

<name>hive.merge.mapfiles</name>

<value>true</value>

</property>

<property>

<name>hive.merge.mapredfiles</name>

<value>false</value>

</property>

<property>

<name>hive.cbo.enable</name>

<value>true</value>

</property>

<property>

<name>hive.fetch.task.conversion</name>

<value>more</value>

</property>

<property>

<name>hive.fetch.task.conversion.threshold</name>

<value>1073741824</value>

</property>

<property>

<name>hive.limit.pushdown.memory.usage</name>

<value>0.04</value>

</property>

<property>

<name>hive.merge.smallfiles.avgsize</name>

<value>16777216</value>

</property>

<property>

<name>hive.merge.size.per.task</name>

<value>268435456</value>

</property>

<property>

<name>hive.optimize.reducededuplication</name>

<value>true</value>

</property>

<property>

<name>hive.optimize.reducededuplication.min.reducer</name>

<value>4</value>

</property>

<property>

<name>hive.map.aggr</name>

<value>true</value>

</property>

<property>

<name>hive.map.aggr.hash.percentmemory</name>

<value>0.5</value>

</property>

<property>

<name>hive.exec.dynamic.partition</name>

<value>true</value>

</property>

<property>

<name>hive.optimize.sort.dynamic.partition</name>

<value>false</value>

</property>

<property>

<name>hive.metastore.execute.setugi</name>

<value>true</value>

</property>

<property>

<name>hive.metastore.dml.events</name>

<value>true</value>

</property>

<property>

<name>hive.support.concurrency</name>

<value>true</value>

</property>

<property>

<name>hive.zookeeper.quorum</name>

<value>rhlmodhdp22502.fairisaac.com,rhlmodhdp22503.fairisaac.com,rhlmodhdp22504.fairisaac.com</value>

</property>

<property>

<name>hive.zookeeper.client.port</name>

<value>2181</value>

</property>

<property>

<name>hive.cluster.delegation.token.store.class</name>

<value>org.apache.hadoop.hive.thrift.MemoryTokenStore</value>

</property>

<property>

<name>hive.metastore.transactional.event.listeners</name>

<value>org.apache.hive.hcatalog.listener.DbNotificationListener</value>

</property>

<property>

<name>hive.strict.checks.orderby.no.limit</name>

<value>false</value>

</property>

<property>

<name>hive.strict.checks.no.partition.filter</name>

<value>false</value>

</property>

<property>

<name>hive.strict.checks.type.safety</name>

<value>true</value>

</property>

<property>

<name>hive.strict.checks.cartesian.product</name>

<value>false</value>

</property>

<property>

<name>hive.strict.checks.bucketing</name>

<value>true</value>

</property>

<property>

<name>hive.server2.support.dynamic.service.discovery</name>

<value>true</value>

</property>

<property>

<name>hive.async.log.enabled</name>

<value>false</value>

</property>

<property>

<name>hive.repl.cm.enabled</name>

<value>false</value>

</property>

<property>

<name>hive.metastore.uris</name>

<value>thrift://rhldatdms22003.fairisaac.com:9083</value>

</property>

<property>

<name>hive.tez.container.size</name>

<value>16384</value>

</property>

<property>

<name>hive.tez.input.format</name>

<value>org.apache.hadoop.hive.ql.io.HiveInputFormat</value>

</property>

<property>

<name>hive.tez.java.opts</name>

<value>-server -Djava.net.preferIPv4Stack=true -Xmx3276m -XX:NewRatio=8 -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB -XX:+PrintGCDetails -verbose:gc</value>

</property>

<property>

<name>hive.tez.cpu.vcores</name>

<value>-1</value>

</property>

<property>

<name>hive.tez.log.level</name>

<value>WARN</value>

</property>

<property>

<name>hive.tez.auto.reducer.parallelism</name>

<value>true</value>

</property>

<property>

<name>hive.tez.min.partition.factor</name>

<value>0.25</value>

</property>

<property>

<name>hive.tez.max.partition.factor</name>

<value>2.0</value>

</property>

<property>

<name>hive.tez.dynamic.partition.pruning</name>

<value>true</value>

</property>

<property>

<name>hive.tez.dynamic.partition.pruning.max.event.size</name>

<value>1048576</value>

</property>

<property>

<name>hive.tez.dynamic.partition.pruning.max.data.size</name>

<value>104857600</value>

</property>

<property>

<name>hive.tez.smb.number.waves</name>

<value>0.5</value>

</property>

<property>

<name>hive.tez.input.generate.consistent.splits</name>

<value>true</value>

</property>

<property>

<name>hive.tez.exec.print.summary</name>

<value>true</value>

</property>

<property>

<name>hive.tez.cartesian-product.enabled</name>

<value>true</value>

</property>

<property>

<name>hive.tez.bucket.pruning</name>

<value>true</value>

</property>

<property>

<name>hive.server2.zookeeper.namespace</name>

<value>hiveserver2</value>

</property>

</configuration>

**hive-env.sh:**

# HIVE\_AUX\_JARS\_PATH={{HIVE\_AUX\_JARS\_PATH}}

# JAVA\_LIBRARY\_PATH={{JAVA\_LIBRARY\_PATH}}

#export HADOOP\_CONF\_DIR="${HADOOP\_CONF\_DIR:-"$( cd "$( dirname "${BASH\_SOURCE[0]}" )" && pwd )"}"

export HADOOP\_CONF\_DIR=/ptoan/shared/hadoop/hadoop-3.3.5/etc/Hadoop

export HIVE\_AUX\_JARS\_PATH=/ptoan/shared/hadoop/apache-hive-3.1.3-bin/auxlib

## HIVE\_AUX\_JARS\_PATH=$([[ -d /ptoan/shared/hadoop/apache-hive-3.1.3-bin/auxlib ]] && sed "s: :,:g" <<< $(find /ptoan/shared/hadoop/apache-hive-3.1.3-bin/auxlib -name "\*.jar" 2> /dev/null)),$([[ -n $HIVE\_AUX\_JARS\_PATH ]] && echo $HIVE\_AUX\_JARS\_PATH,)$( ([[ ! '/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hive/lib/hive-hbase-handler-3.1.3000.7.1.7.1000-141.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-server.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-common.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-protocol.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-client.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-hadoop-compat.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-hadoop2-compat.jar' =~ HIVE\_HBASE\_JAR ]] && echo /opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hive/lib/hive-hbase-handler-3.1.3000.7.1.7.1000-141.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-server.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-common.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-protocol.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-client.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-hadoop-compat.jar,/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p1000.24102687/lib/hbase/hbase-hadoop2-compat.jar) || echo ${HBASE\_HIVE\_DEFAULT\_JAR:-}),$(find /usr/share/java/mysql-connector-java.jar 2> /dev/null),$(find /usr/share/cmf/lib/postgresql-\*jdbc\*.jar 2> /dev/null | tail -n 1),$(find /usr/share/java/oracle-connector-java.jar 2> /dev/null)

#export HIVE\_AUX\_JARS\_PATH=$(echo $HIVE\_AUX\_JARS\_PATH | sed 's/,,\*/,/g' | sed 's/^,//' | sed 's/,$//')

#export HIVE\_OPTS="${HIVE\_OPTS} --hiveconf hive.query.redaction.rules=${HIVE\_HOME}/conf/redaction-rules.json --hiveconf hive.exec.query.redactor.hooks=org.cloudera.hadoop.hive.ql.hooks.QueryRedactor"

export HADOOP\_CLIENT\_OPTS="-Xmx4294967296 -Djava.net.preferIPv4Stack=true $HADOOP\_CLIENT\_OPTS"

**beeline-site.xml:**

<?xml version="1.0" encoding="UTF-8"?>

<!--Autogenerated by Cloudera Manager-->

<configuration>

<property>

<name>beeline.hs2.jdbc.url.default</name>

<value>hive\_on\_tez</value>

</property>

<property>

<name>beeline.hs2.jdbc.url.hive\_on\_tez</name> <value>jdbc:hive2://rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2</value>

</property>

</configuration>

**redaction-rules.json:**

{

"version": 1,

"rules": [

{

"description": "Redact passwords from json files",

"trigger": "password",

"search": "\"password\"[ ]\*:[ ]\*\"[^\"]+\"",

"caseSensitive": false,

"replace": "\"password\": \"LOG-REDACTED\""

},

{

"description": "Redact password= and password:",

"trigger": "password",

"search": "password[:=][^ \"\\\\]+",

"caseSensitive": false,

"replace": "password=LOG-REDACTED"

},

{

"description": "Redact passwd= and passwd:",

"trigger": "passwd",

"search": "passwd[:=][^ \"\\\\]+",

"caseSensitive": false,

"replace": "passwd=LOG-REDACTED"

},

{

"description": "Redact pass= and pass:",

"trigger": "pass",

"search": "pass[:=][^ \"\\\\]+",

"caseSensitive": false,

"replace": "pass=LOG-REDACTED"

},

{

"description": "Redact PASSWORD,",

"trigger": "PASSWORD,",

"search": "PASSWORD,[^\"\\\\]+",

"caseSensitive": false,

"replace": "PASSWORD, LOG-REDACTED"

},

{

"description": "Redact secret= and secret:",

"trigger": "secret",

"search": "secret[:=][^ \"\\\\]+",

"caseSensitive": false,

"replace": "secret=LOG-REDACTED"

},

{

"description": "Credit Card numbers (with separator)",

"search": "\\d{4}[^\\w:]\\d{4}[^\\w:]\\d{4}[^\\w:]\\d{4}",

"caseSensitive": true,

"replace": "XXXX-XXXX-XXXX-XXXX"

},

{

"description": "Social Security numbers (with separator)",

"search": "\\d{3}[^\\w:]\\d{2}[^\\w:]\\d{4}",

"caseSensitive": true,

"replace": "XXX-XX-XXXX"

}

]

}

**log4j.properties:**

log4j.rootLogger=${hive.root.logger}

hive.root.logger=INFO,console

log4j.appender.console=org.apache.log4j.ConsoleAppender

log4j.appender.console.target=System.err

log4j.appender.console.layout=org.apache.log4j.PatternLayout

log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss} %p %c{2}: %m%n

---------------------------------------------------------------------------------

export JAVA\_HOME=/ptoan/dms\_staging/parquet/jdk1.8.0\_231/

export HADOOP\_HOME=/ptoan/shared/hadoop/hadoop-3.3.5

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export SPARK\_HOME=/ptoan/shared/hadoop/spark-3.0.1-bin-without-hadoop

export HIVE\_HOME=/ptoan/shared/hadoop/apache-hive-3.1.3-bin

export HBASE\_HOME=/ptoan/shared/hadoop/hbase-2.5.4

export PATH=$HBASE\_HOME/bin:$HIVE\_HOME/bin:$SPARK\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$PATH export HADOOP\_CLASSPATH=`hadoop classpath`

If using local tez and not Cloudera, then:

export HADOOP\_CLASSPATH=/ptoan/shared/hadoop/apache-tez-0.10.2-bin/\*:/ptoan/shared/hadoop/apache-tez-0.10.2-bin/lib/\*:$HADOOP\_CLASSPATH

if [ -d "/usr/lib/tez" ];  
then export HADOOP\_CLASSPATH=$HADOOP\_CLASSPATH:/usr/lib/tez/\*:/usr/lib/tez/lib/\*:/etc/tez/conf  
fi

<https://dwgeek.com/hiveserver2-beeline-command-line-shell-options-examples.html/>

set hive.execution.engine=tez;

set mapreduce.framework.name=yarn-tez;

bash-4.2$ beeline --hiveconf hive.tez.container.size=8192 --hiveconf tez.am.resource.memory.mb=8192 --hiveconf hive.tez.java.opts=-Xmx6554m

WARNING: Use "yarn jar" to launch YARN applications.

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/ptoan/shared/hadoop/apache-hive-3.1.3-bin/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/ptoan/shared/hadoop/hadoop-3.3.5/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Connecting to jdbc:hive2://rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181/default;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2

Enter username for jdbc:hive2://rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181/default: janecheng

Enter password for jdbc:hive2://rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181/default: \*\*\*\*\*\*\*\*\*\*

23/05/31 21:00:05 [main]: INFO jdbc.HiveConnection: Connected to rhlmodhdp22509.fairisaac.com:10000

Connected to: Apache Hive (version 3.1.3000.7.1.7.1000-141)

Driver: Hive JDBC (version 3.1.3)

Transaction isolation: TRANSACTION\_REPEATABLE\_READ

Beeline version 3.1.3 by Apache Hive

0: jdbc:hive2://rhlmodhdp22502.fairisaac.com:> **set hive.execution.engine=tez;**

No rows affected (0.048 seconds)

0: jdbc:hive2://rhlmodhdp22502.fairisaac.com:> **show tables from jane2;**

INFO : Compiling command(queryId=hive\_20230531210034\_2934e425-a3ac-4807-945e-55f5e1f451c3): show tables from jane2

INFO : Semantic Analysis Completed (retrial = false)

INFO : Created Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab\_name, type:string, comment:from deserializer)], properties:null)

INFO : Completed compiling command(queryId=hive\_20230531210034\_2934e425-a3ac-4807-945e-55f5e1f451c3); Time taken: 0.008 seconds

INFO : Executing command(queryId=hive\_20230531210034\_2934e425-a3ac-4807-945e-55f5e1f451c3): show tables from jane2

INFO : Starting task [Stage-0:DDL] in serial mode

DEBUG : Task getting executed using mapred tag : hive\_20230531210034\_2934e425-a3ac-4807-945e-55f5e1f451c3,userid=janecheng

INFO : Completed executing command(queryId=hive\_20230531210034\_2934e425-a3ac-4807-945e-55f5e1f451c3); Time taken: 0.01 seconds

INFO : OK

DEBUG : Shutting down query show tables from jane2

+------------+

| tab\_name |

+------------+

| janetable |

+------------+

1 row selected (0.034 seconds)

0: jdbc:hive2://rhlmodhdp22502.fairisaac.com:> !quit

Closing: 0: jdbc:hive2://rhlmodhdp22502.fairisaac.com:2181,rhlmodhdp22503.fairisaac.com:2181,rhlmodhdp22504.fairisaac.com:2181/default;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2

hbase-2.5.4:

bash-4.2$ sudo ln -s /ptoan/shared/hadoop/hbase-2.5.4/bin/hbase /etc/alternatives/hbase

bash-4.2$ sudo ln -s /etc/alternatives/hbase /usr/bin/hbase

/ptoan/shared/hadoop/apache-tez-0.10.2-bin:

tez-site.xml:

<?xml version="1.0" encoding="UTF-8"?>

<configuration>

<property>

<name>tez.am.am-rm.heartbeat.interval-ms.max</name>

<value>250</value>

</property>

<property>

<name>tez.am.container.idle.release-timeout-max.millis</name>

<value>20000</value>

</property>

<property>

<name>tez.am.container.idle.release-timeout-min.millis</name>

<value>10000</value>

</property>

<property>

<name>tez.am.container.reuse.enabled</name>

<value>true</value>

</property>

<property>

<name>tez.am.container.reuse.locality.delay-allocation-millis</name>

<value>250</value>

</property>

<property>

<name>tez.am.container.reuse.non-local-fallback.enabled</name>

<value>false</value>

</property>

<property>

<name>tez.am.container.reuse.rack-fallback.enabled</name>

<value>true</value>

</property>

<property>

<name>tez.am.launch.cluster-default.cmd-opts</name>

<value>-server -Djava.net.preferIPv4Stack=true</value>

</property>

<property>

<name>tez.am.launch.cmd-opts</name>

<value>-XX:+PrintGCDetails -verbose:gc -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB -XX:+HeapDumpOnOutOfMemoryError -XX:HeapDumpPath=/tmp</value>

</property>

<property>

<name>tez.am.launch.env</name>

<value>LD\_LIBRARY\_PATH=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native</value>

</property>

<property>

<name>tez.am.log.level</name>

<value>WARN</value>

</property>

<property>

<name>tez.am.max.app.attempts</name>

<value>2</value>

</property>

<property>

<name>tez.am.maxtaskfailures.per.node</name>

<value>10</value>

</property>

<property>

<name>tez.am.resource.memory.mb</name>

<value>4096</value>

</property>

<property>

<name>tez.am.view-acls</name>

<value>\*</value>

</property>

<property>

<name>tez.cluster.additional.classpath.prefix</name>

<value>/ptoan/shared/hadoop/apache-tez-0.10.2-bin/lib/guava-27.0-jre.jar</value>

</property>

<property>

<name>tez.counters.max</name>

<value>10000</value>

</property>

<property>

<name>tez.counters.max.groups</name>

<value>3000</value>

</property>

<property>

<name>tez.generate.debug.artifacts</name>

<value>false</value>

</property>

<property>

<name>tez.grouping.max-size</name>

<value>1073741824</value>

</property>

<property>

<name>tez.grouping.min-size</name>

<value>16777216</value>

</property>

<property>

<name>tez.grouping.split-waves</name>

<value>1.7</value>

</property>

<property>

<name>tez.history.logging.proto-base-dir</name>

<value>/warehouse/tablespace/managed/hive/sys.db</value>

</property>

<property>

<name>tez.history.logging.timeline-cache-plugin.old-num-dags-per-group</name>

<value>5</value>

</property>

<property>

<name>tez.lib.uris</name>

<value>/user/tez/0.10.2/tez.tar.gz</value>

</property>

<property>

<name>tez.runtime.compress</name>

<value>true</value>

</property>

<property>

<name>tez.runtime.compress.codec</name>

<value>org.apache.hadoop.io.compress.SnappyCodec</value>

</property>

<property>

<name>tez.runtime.convert.user-payload.to.history-text</name>

<value>false</value>

</property>

<property>

<name>tez.runtime.io.sort.mb</name>

<value>2048</value>

</property>

<property>

<name>tez.runtime.optimize.local.fetch</name>

<value>true</value>

</property>

<property>

<name>tez.runtime.pipelined.sorter.sort.threads</name>

<value>2</value>

</property>

<property>

<name>tez.runtime.shuffle.fetch.buffer.percent</name>

<value>0.6</value>

</property>

<property>

<name>tez.runtime.shuffle.keep-alive.enabled</name>

<value>true</value>

</property>

<property>

<name>tez.runtime.shuffle.memory.limit.percent</name>

<value>0.25</value>

</property>

<property>

<name>tez.runtime.unordered.output.buffer.size-mb</name>

<value>409</value>

</property>

<property>

<name>tez.session.am.dag.submit.timeout.secs</name>

<value>300</value>

</property>

<property>

<name>tez.session.client.timeout.secs</name>

<value>-1</value>

</property>

<property>

<name>tez.shuffle-vertex-manager.max-src-fraction</name>

<value>0.4</value>

</property>

<property>

<name>tez.shuffle-vertex-manager.min-src-fraction</name>

<value>0.2</value>

</property>

<property>

<name>tez.staging-dir</name>

<value>/tmp/${user.name}/staging</value>

</property>

<property>

<name>tez.task.am.heartbeat.counter.interval-ms.max</name>

<value>4000</value>

</property>

<property>

<name>tez.task.generate.counters.per.io</name>

<value>true</value>

</property>

<property>

<name>tez.task.get-task.sleep.interval-ms.max</name>

<value>200</value>

</property>

<property>

<name>tez.task.launch.cluster-default.cmd-opts</name>

<value>-server -Djava.net.preferIPv4Stack=true</value>

</property>

<property>

<name>tez.task.launch.cmd-opts</name>

<value>-XX:+PrintGCDetails -verbose:gc -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB -XX:+HeapDumpOnOutOfMemoryError -XX:HeapDumpPath=/tmp</value>

</property>

<property>

<name>tez.task.launch.env</name>

<value>LD\_LIBRARY\_PATH=/ptoan/shared/hadoop/hadoop-3.3.5/lib/native</value>

</property>

<property>

<name>tez.task.max-events-per-heartbeat</name>

<value>500</value>

</property>

<property>

<name>tez.task.resource.memory.mb</name>

<value>32768</value>

</property>

<property>

<name>tez.use.cluster.hadoop-libs</name>

<value>false</value>

</property>

<property>

<name>yarn.timeline-service.enabled</name>

<value>false</value>

</property>

<property>

<name>tez.runtime.sorter.class</name>

<value>PIPELINED</value>

</property>

<property>

<name>tez.history.logging.service.class</name>

<value>org.apache.tez.dag.history.logging.proto.ProtoHistoryLoggingService</value>

</property>

</configuration>

TEZ on cloudera:

bash-4.2$ ls -lt /etc/tez/conf.cloudera.tez

-rw-r--r-- 1 root root 20 May 1 12:25 \_\_cloudera\_generation\_\_

-rw-r--r-- 1 root root 26 May 1 12:25 \_\_cloudera\_metadata\_\_

-rw-r--r-- 1 root root 6151 May 1 12:25 tez-site.xml