Sqoop Troubleshooting

Date published: 2023-05-16
Date modified:

https://docs.cloudera.com/

Legal Notice

Contents

Merge process stops during Sqoop incremental imports

During Sqoop incremental import operations, if the target directory is located outside of Hadoop Distributed File System (HDFS), such as in Amazon S3 or Azure Blob Storage, the merge phase of the import process does not take effect.

Condition

Sqoop, by default, creates temporary directories within HDFS. However, you must be aware of certain considerations in choosing the target directory location while working with Sqoop's incremental import modes. By default, Sqoop operates seamlessly when the target directory resides within HDFS. However, the merge phase of the import process does not take effect outside the box if the target directory is located outside of HDFS.

Cause

During an import operation, Sqoop generally imports data to a target directory. If this target directory is a non-HDFS location, the merge process tries to acquire the temporary directory required for the merge on the same non-HDFS file system. Since Sqoop creates the temporary directory in HDFS by default, the merge process checks if the temporary directory exists in the target directory's file system and when it does not find it, the merge process simply stops.

Solution

If the target directory is present outside of HDFS, you must modify the default path of the temporary directory by adding the --temporary-rootdir Sqoop option and pointing to a path on the same file where the target directory is located. By aligning the temporary directory path with the file system of the target directory, Sqoop can effectively complete the import process.

Example:

Include the --temporary-rootdir Sqoop option as shown below:

```
sqoop-import --connect jdbc:mysql://.../transaction --username [***USER NAME
***] --table [***TABLE NAME***] --password [***PASSWORD***] --target-dir abf
s://foo@bar/targetdir -m 1 --temporary-rootdir abfs://foo@bar/_sqoop
```

Sqoop Hive import stops when HS2 does not use Kerberos authentication

Learn how to resolve the issue related to Sqoop Hive imports when either LDAP authentication or no authentication mechanism is enabled for the cluster.

Condition

When running Sqoop commands to import data into Hive from either the CLI or Oozie, the import job stops after the Sqoop import is done and while trying to connect to HiveServer (HS2) through JDBC. The following log is displayed and you will notice that the job stops on the last line:

```
23/07/24 18:10:17 INFO hive.HiveImport: Loading uploaded data into Hive
23/07/24 18:10:17 INFO hive.HiveImport: Collecting environment variables whi
ch need to be preserved for beeline invocation
...
23/07/24 18:10:20 INFO hive.HiveImport: SLF4J: Actual binding is of type
[org.apache.logging.slf4j.Log4jLoggerFactory]
```

```
23/07/24 18:10:21 INFO hive.HiveImport: Connecting to jdbc:hive2://HOSTNAME/
default;serviceDiscoveryMode=zooKeeper;ssl=true;sslTrustStore=/var/lib/clo
udera-scm-agent/agent-cert/cm-auto-global_truststore.jks;trustStorePassword=
changeit;zooKeeperNamespace=hiveserver2
```

Cause

This issue occurs when Kerberos is not used in the JDBC connection string, which Sqoop uses to connect to HS2. The issue affects unsecure clusters and clusters where LDAP authentication is enabled, and the beeline-site.xml configuration file does not use Kerberos authentication.

The underlying issue is that Beeline prompts for the username and password for a successful connection and since the Sqoop Hive import is a non-interactive session, you are unable to provide the credentials and therefore the import job stops.

Solution

Procedure

Perform one of the following steps to resolve this issue:

If...	Then...
No authentication is enabled for the cluster	Include the --hs2-url option in the Sqoop import command and provide the JDBC connection string. `--hs2-url <HS2 JDBC string>` This allows for a successful connection without prompting for the credentials.

LDAP authentication is enabled for the cluster	Include the --hs2-user and --hs2-password options in the Sqoop import command and provide the credentials.
	`--hs2-user <username>`
	`--hs2-password <password>`