Copy tables inter cluster

Top level steps

- 1. Export tables to files on the source cluster.
- 2.Distcp the files form source to destination cluster.
- 3.Import tables in destination cluster.

How to run?

- Copy the arguments.sh to ‘/home/sqoop/distcp/’ and ‘/home/hdsuser/horton/distcp/’ of the source and destination nodes of the source and destnation cluster.
- The script is doing all the operation from these 3 nodes only:
- azalvedledgdp02.p01eaedl.manulife.com[prod]
- azslvedledgdd01.d01saedl.manulife.com[VC]
- azalvedledgv01.p01eaedl.manulife.com[PV]
- Flow of the scripts:
- from_cluster_sqoop.sh[To be run from ‘sqoop’ user from source node] → from_cluster_hdsuser_to_cluster.sh[To be run from ‘hdsuser’ user from source node] → to_cluster_hdsuser.sh[To be run from ‘hdsuser’ user from destination node] → to_cluster_sqoop.sh[To be run from ‘sqoop’ user from destination node]

Step2-3 Explained

- **Step1:** After distcp we will have imported data in destination cluster.

Note: While Distcp throws error please modify the hdfs-site.xml in /etc/hadoop2/conf as per ‘pv_hdfs-site.xml’ or ’hdfs-site_dev.xml’

- **Step2**: Check if database exist in the To cluster; If not then create database as per the given location in ‘$create_database_path’. This is done in automated manner, but run it line by line for safety.[‘to_cluster_sqoop.sh’ line 19-43]

- **Step3**: Go to **step 4** and see if the table copy is failing, if yes do step3 and then step 4, else ok.

Step 3 explained:

Get the table details and drop the table and data from HDFS path corresponding to existing tables and delete the corresponding data[Since we are overwriting]. Sections[“to_cluster_sqoop.sh” line 47-75; though its automated please do it manually line by line for safety.]

Note: In PV cluster running import statement for 1st time will show error Table “**exists and contains data files”.** To mitigate that remove the data from HDFS path. The commands can be found for that table in “remove_existing_table_data.sh”.

Note: It can happen if data already exist so we have step3.

- **Step4**: In this step after generating import statement, please paste all the import commands from the import_$Table_number.sql generated in beeline prompt[“to_cluster_sqoop.sh” line 75-88; though its automated please do it manually line by line for safety.]

Arguments

- **Source cluster:**
- From_cluster="p01eaedl"
- Table_number="tables15.in"
- Distcp_script="distcp15.sh"
- sql_file_for_export="exp15"
- Beeline_URL_From_Cluster="beeline -u 'jdbc:hive2://azalvedlmstdp01.p01eaedl.manulife.com:2181,azalvedlmstdp02.p01eaedl.manulife.com:2181,azalvedlmstdp03.p01eaedl.manulife.com:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2-hive2'"
- SQOOP_principal_From_Cluster="sqoop@P01EAEDL.MANULIFE.COM"
- From_cluster_node="azalvedledgdp02.p01eaedl.manulife.com"
- sqoop_keytab_from="sqoop.headless.keytab“s
- **Destination Cluster:**
- SQOOP_principal_To_Cluster="sqoop-v01eaedl@P01EAEDL.MANULIFE.COM"
- sqoop_keytab_to="sqoop-v01eaedl.keytab"
- To_cluster_node="azalvedledgv01.p01eaedl.manulife.com"
- To_cluster="v01eaedl"
- Beeline_URL_To_Cluster="beeline -u 'jdbc:hive2://azalvedlmstv01.p01eaedl.manulife.com:2181,azalvedlmstv02.p01eaedl.manulife.com:2181,azalvedlnifv01.p01eaedl.manulife.com:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2'"
- To_cluster_node_pass=“****“
- **Other variables:**
- create_database_path="/asia/sg/prod/published/hive/“
- Export_path="/sg/tmp/export/"
- Import_path="/asia/sg/tmp/import/"

Import tables in dest cluster

- Prep steps:
- Check DB exists or not if not create db with proper location.[using same db name as src.]
- If table exists in db of dest cluster..:
  - Drop the existing table which we will copy. [Can we drop?]
  - Remove the data from external path if table is external
  - Is it overwrite of table or delta import ?
- Import the table using hive import command.
- Make sure the table and table data is in correct location.

To test tables on these DBs:

- To test tables on these DBs:
- sg_published_cas_db
- sg_published_ams_db
- sg_published_ccl_db
- 1st only few tables form each DB.
- Followed by all remaining tables.
- Schedule this script weekly.