Lab 9.1: Use distcp to Copy Data from a Remote Cluster

|  |  |
| --- | --- |
| **Objective:** | To become familiar with how to copy data from one cluster to another. |
| **Successful Outcome:** | Data from a remote cluster is copied to your own cluster. |
| **Before You Begin:** | For this exercise use “node1” as your Remote-Cluster. |

1. Access Remote-Cluster
   1. Verify you can reach the NameNode of the Remote-Cluster by executing the ls command.

$ hadoop fs -ls hdfs://*node1*:8020/

You should see the / folder of the Remote-Cluster.

1. View the Remote Folder
   1. Use the ls command and view the contents of the remote directory that you are going to copy:

$ hadoop fs -ls hdfs://*node1*:8020/user/root

1. Copy the Data from Remote-Cluster
   1. First create a new directory in HDFS called **distcp\_target.**
   2. Execute the following command to copy a remote file into distcp\_target directory:

$ hadoop distcp   
hdfs://*Remote-Cluster*:8020/user/root/test\_data distcp\_target

* 1. View the contents of distcp\_target and verify test\_data file copied over to your cluster:

$ hadoop fs -ls distcp\_target

* 1. Copy one or many directories/files into distcp\_target:

$ hadoop distcp   
hdfs://*node1*:8020/user/root/wordcount

hdfs://node1:8020/user/root/constitution.txt

distcp\_target

* 1. View the contents of distcp\_target and verify the wordcount & constitution.txt file copied over to your cluster again.

1. Copy only new/updated files and directories using -update option.
   1. Check the timestamp for files in /user/root/wordcount directory. Delete **part-r-00000** file from **wordcount** directory.
   2. Now run following command with -update option.

$ hadoop distcp -update  
hdfs://*node1*:8020/user/root/wordcount

distcp\_target/wordcount

* 1. View the contents of distcp\_target and compare timestamp of all the files. You can see that the timestamp changed only for **part-r-00000** file and **wordcount** folder timestamp.

1. Copy data from a ‘Remote-Cluster’ running different version of Hadoop.
   1. Execute the following command to copy a remote file into distcp\_target.

$ hadoop distcp  
hftp://*node1*:50070/user/root/test/hdfs-audit.log

distcp\_target

* 1. View the contents of distcp\_target and verify test\_data file copied over to your cluster.

**RESULT**: You have learnt the steps to copy data from one cluster to another.