Elastic MapReduce

Data Migration

Product Documentation

Contents

Data Migration Getting Started

Last updated：2019-07-26 17:53:19

EMR currently allows you to migrate your data in two ways: 1. Migrating your data to COS; 2. Copying EMR cluster data to your self-built HDFS using DistCP. This method requires your self-built cluster is connect with the EMR cluster network.

Using COS to Migrate Data

Last updated：2019-07-26 17:53:25

- Migrating non-HDFS files

  If your source file is a non-HDFS file, upload it to COS via COS Console or API, and then analyze it in the EMR cluster.

- Migrating HDFS files

  i. Get the COS migration tool

  Click here to download the migration tool. For more migration tools, see here.

  ii. Configuring the tool

  All configuration files are stored in the conf directory of the tool directory. Copy the core-site.xml file of the HDFS cluster to be synced to conf, which contains the configuration information of the NameNode. Edit the configuration file cos_info.conf by including your appid, bucket, region, and key information.

  iii. Command parameter descriptions

```
-ak <ak> the cos secret id
-appid,--appid <appid> the cos appid
-bucket,--bucket <bucket_name> the cos bucket name
-cos_info_file,--cos_info_file <arg> the cos user info config default is ./conf/cos_info.conf
-cos_path,--cos_path <cos_path> the absolute cos folder path
-h,--help print help message
-hdfs_conf_file,--hdfs_conf_file <arg> the hdfs info config default is ./conf/core-site.xml
-hdfs_path,--hdfs_path <hdfs_path> the hdfs path
-region,--region <region> the cos region. legal value cn-south, cn-east, cn-north, sg
-sk <sk> the cos secret key
-skip_if_len_match,--skip_if_len_match skip upload if hadoop file length match cos
```

  iv. Executing data migration

```
# All operations must be performed in the tool directory. If both configuration files and command line parameters are set, the latter will prevail
```

```
./hdfs_to_cos_cmd -h

# Copy from HDFS to COS (The file will be overwritten if it already exists in COS)
./hdfs_to_cos_cmd --hdfs_path=/tmp/hive --cos_path=/hdfs/20170224/

# Copy from HDFS to COS, and if a file to be copied is of the same length as a file in COS, then i
t is skipped (this is suitable for repeated copy)
# Only the length is checked here, as the overheads would be very high if the digests of files in
Hadoop are to be calculated
./hdfs_to_cos_cmd --hdfs_path=/tmp/hive --cos_path=/hdfs/20170224/ -skip_if_len_match

# Set parameters completely through the command line
./hdfs_to_cos_cmd -appid 1252xxxxxx -ak
AKIDVt55xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx -sk
KS08jDVbVElxxxxxxxxxxxxxxxxxxxxxxxxxxxx -bucket test -cos_path /hdfs
-hdfs_path /data/data -region cn-south -hdfs_conf_file
/home/hadoop/hadoop-2.8.1/etc/hadoop/core-site.xml
```

- After the command is verified and run, a log will be outputted as shown below:

```
[Folder Operation Result : [ 53(sum)/ 53(ok) / 0(fail)]
[File Operation Result: [22(sum)/ 22(ok) / 0(fail) / 0(skip)]
[Used Time: 3 s]
```

As shown in the figure above, sum indicates the total number of files to be migrated; ok the number of files successfully migrated; fail the number of files failed to be migrated; skip the number of files skipped because they have the same length as the files of the same name in the destination after the skip_if_len_match parameter is added. You can also log in to the COS Console to see whether the data has been migrated correctly.

- FAQs
  - Please ensure that the configuration information is correct, including appID, key, bucket, and region. Ensure that the server time is the same as Beijing time (one-minute difference is acceptable. If the difference is too large, please re-set your server time).
  - Please ensure that the server for the copy program is accessible to DateNode. The NameNode uses a public IP address and can be accessed, but the DateNode where the obtained block is located uses a private IP address and cannot be accessed; therefore, it is recommended that the copy program be placed in a Hadoop node for execution, so that both the NameNode and DateNode can be accessed.
  - In case of a permissions issue, use the current account to download a file with the Hadoop command, check whether everything is correct, and then use the synchronization tool to sync the

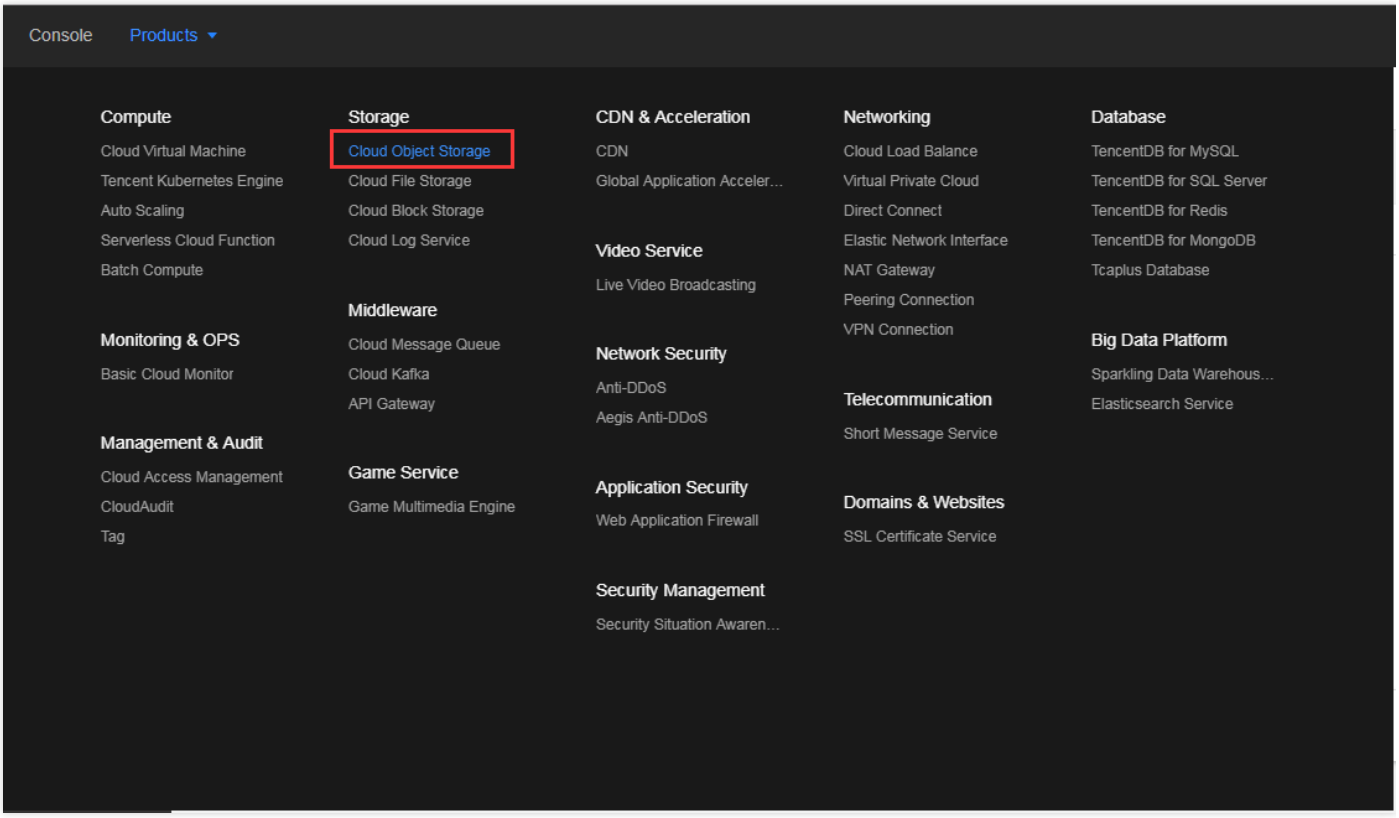
data in Hadoop.

- Files that already exist in COS are overwritten by default in case of repeated upload, unless you explicitly specify the -skip_if_len_match parameter, which indicates to skip files if they have the same length as the existing files.
- The COS path is always considered as a directory, and files that are eventually copied from HDFS will be stored in this directory.

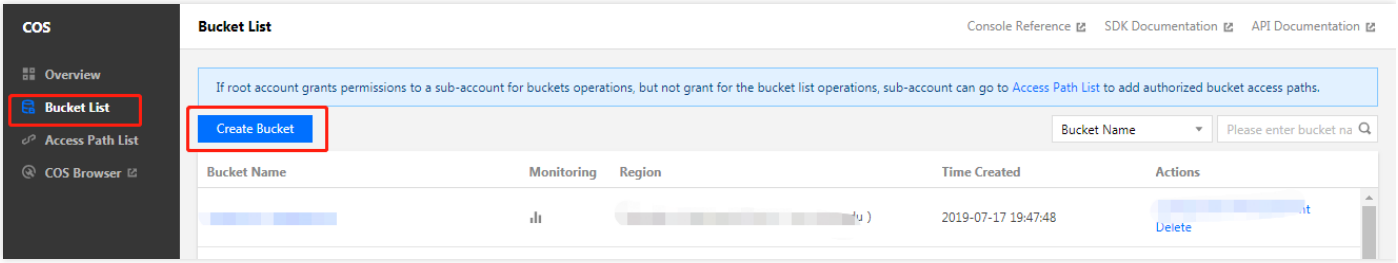
How to View COS Info

Last updated：2019-07-26 17:53:34

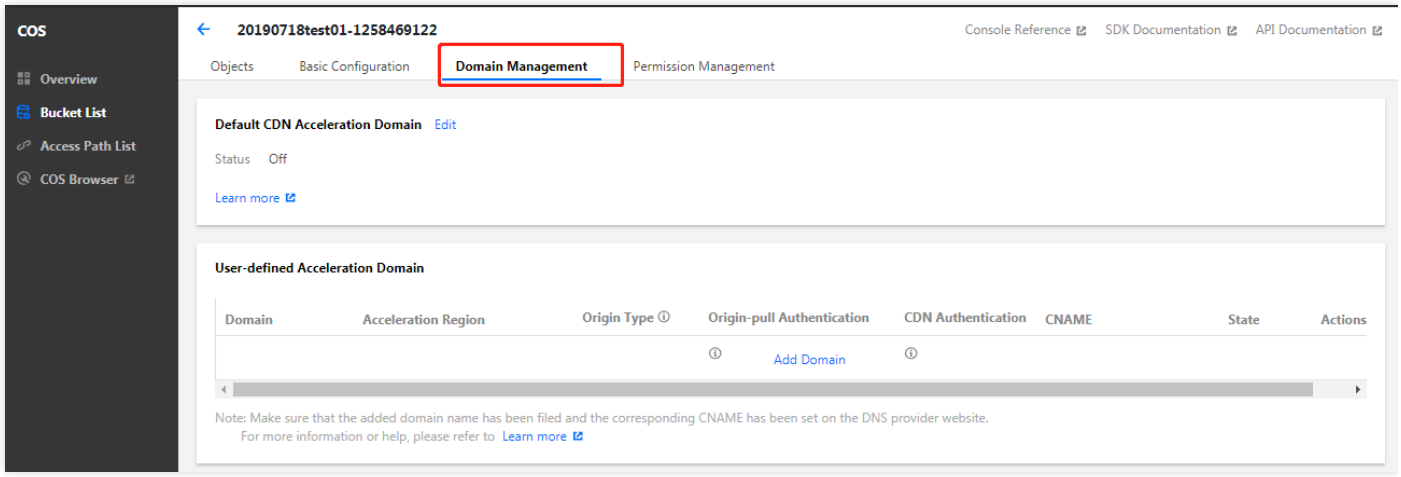
- Log in to your Tencent Cloud account, go to the console and select Cloud Object Storage



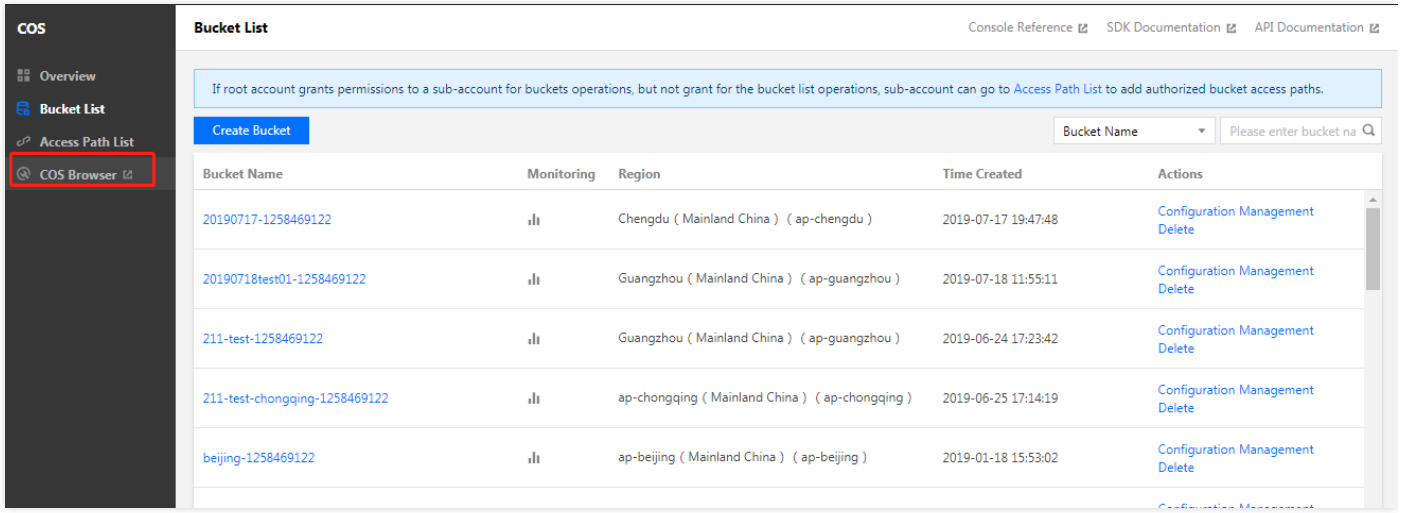
- Select a bucket



- View the bucket's domain name



- Query Secretkey



Using DistCp to Migrate Data
Brief Description

Last updated：2019-07-26 17:53:45

DistCp (distributed copy) is a tool used for large inter/intra-cluster copying. It uses MapReduce to effect its distribution, error handling and recovery, and reporting. It expands a list of files and directories into input to map tasks, each of which will copy a partition of the files specified in the source list. It is a file migration tool that comes with Hadoop.

Network Connectivity

Last updated：2019-08-12 20:04:18

Migrating Files in a Local Self-built HDFS to EMR

The migration of files in a local self-built HDFS to an EMR cluster requires a Direct Connect line for network connectivity. You can contact the developers for assistance.

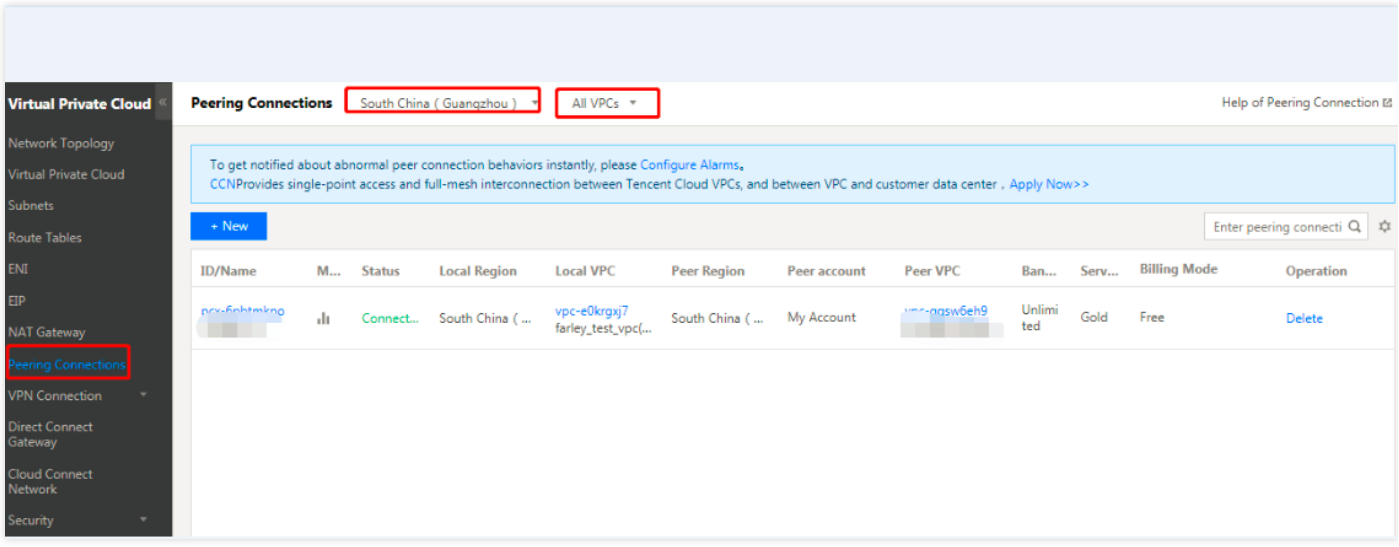
Migrating Files in a Self-built HDFS in CVM to EMR

- If the network where the CVM instance resides and the one where the EMR cluster resides are in the same VPC, the files can be transferred freely.
- Otherwise, a peering connection is required for network connectivity.

Using a Peering Connection

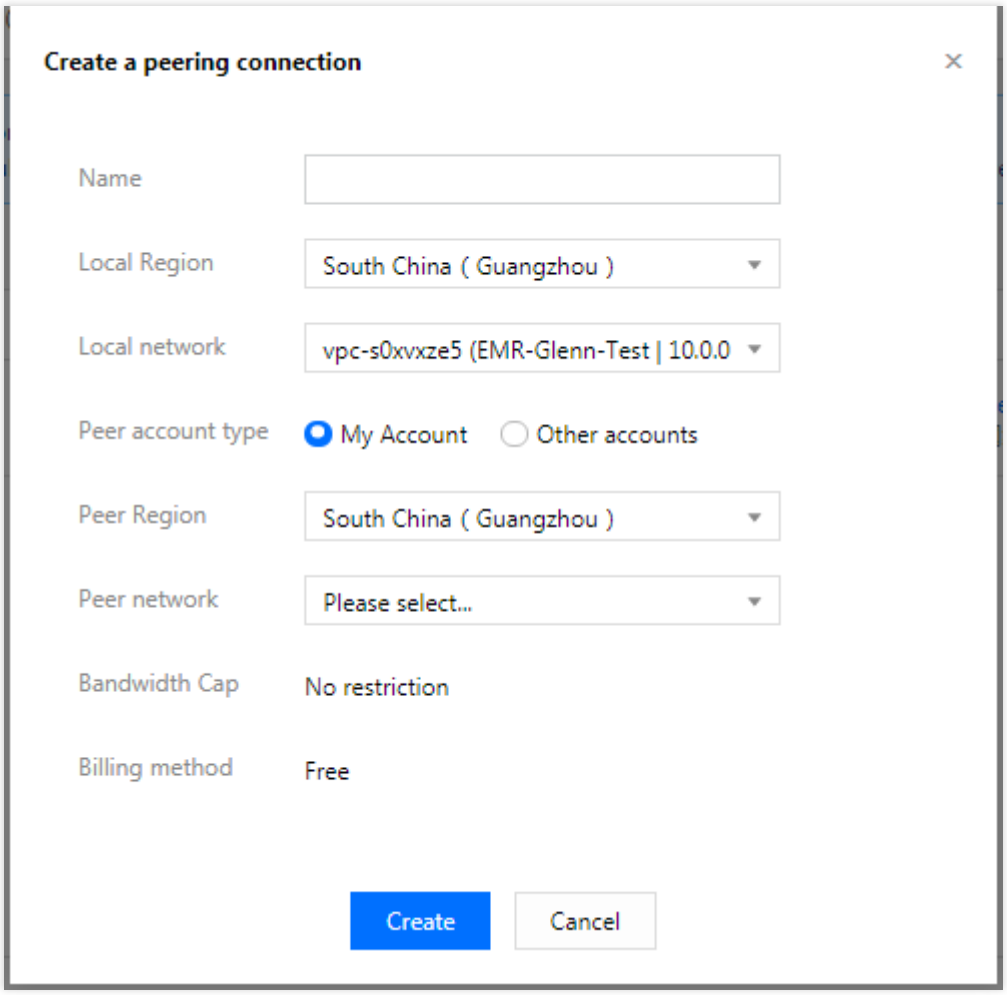
IP range 1: Subnet A 192.168.1.0/24 in VPC1 of Guangzhou.
IP range 2: Subnet B 10.0.1.0/24 in VPC2 of Beijing.

1. Log in to the Tencent Cloud Console and select **Products** > **Networking** > **Virtual Private Cloud** in the navigation bar.
2. In the VPC Console, go to the "Peering Connections" page, select the region "Guangzhou" in the list at the top, select "VPC1", and click **Create**.



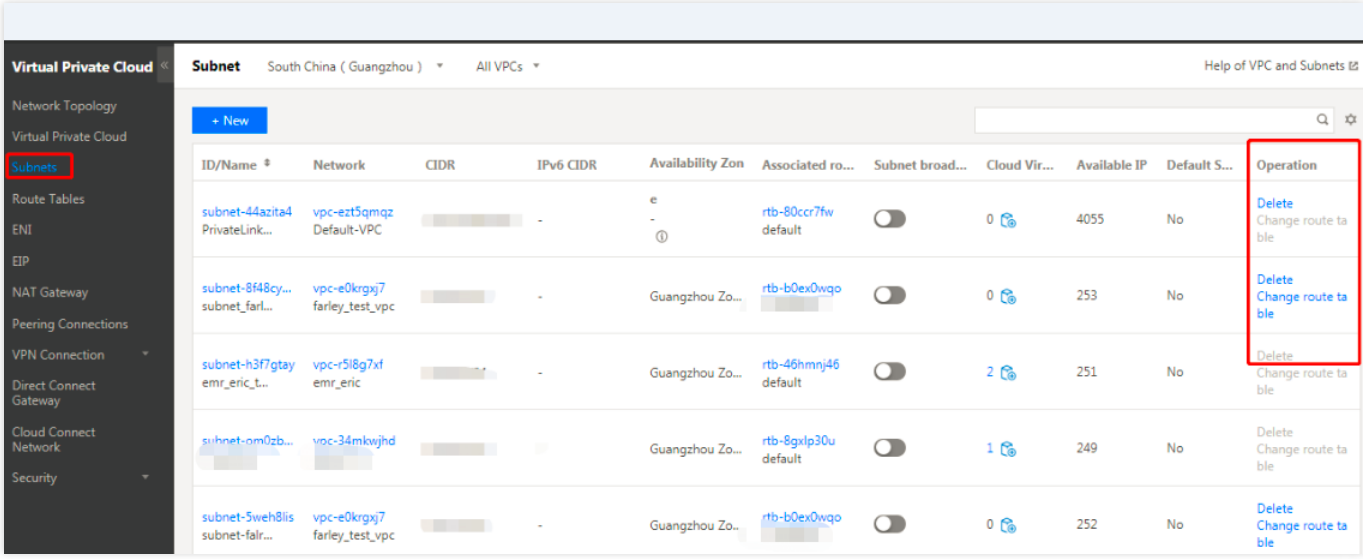
3. Go to the peering connection creation page.

- Enter the name of the peering connection as shown in box 1, for example, PeerConn.
- Enter the local region and network information as shown in box 2, for example, Guangzhou and VPC1.
- Enter the account of the opposite network as shown in box 3. If the two networks in Guangzhou and Beijing are under the same account, select **My account**; otherwise, select "Another account". It should be noted here that if both the local network and the opposite network are in the same region (e.g., Guangzhou), the communication is free of charge, and there is no need to select the upper limit for bandwidth as shown in box 5; otherwise, fees will be incurred and the upper limit for bandwidth can be set.
- Enter the opposite region and network as shown in box 4, for example, Beijing and VPC2.



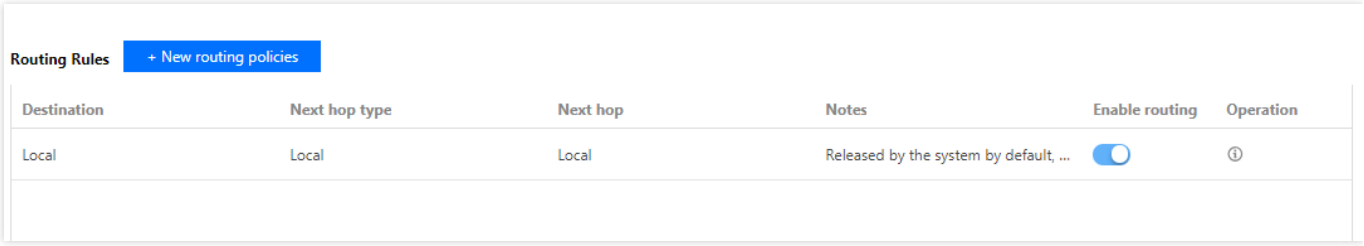
4. A peering connection between VPCs under the same account takes effect immediately after creation; otherwise, it can take effect only after the opposite account accepts it.
5. Configure the local and opposite route tables for the peering connection.

i. Go to the **Subnets** page in the VPC Console.



ii. Click "Associated Route Tables" (as shown in the red box above) of the subnet of the local network (e.g., subnet VPC1 in Guangzhou) to enter the route table details page.

iii. Click **Edit** to edit the routing policy. Enter the opposite CIDR (for example, the CIDR of VPC2 in Beijing is 10.0.0.1/24) for the destination, select **Peering connection** for the next hop type, and select the created peering connection (PeerConn) for the next hop.



iv. You've configured the route table from Guangzhou VPC1 to Beijing VPC2 in the previous steps. Now you need to repeat the steps above to configure the route table from Beijing VPC2 to Guangzhou VPC1.

v. After the route tables are configured, different VPC IP ranges can communicate to each other.

Copy

Last updated：2019-07-26 17:54:18

```
# Copy the specified folder from one cluster to another
hadoop distcp hdfs://nn1:9820/foo/bar hdfs://nn2:9820/bar/foo

# Copy the specified file
hadoop distcp hdfs://nn1:9820/foo/a hdfs://nn1:9820/foo/b hdfs://nn2:9820/bar/foo

# If too many files were specified, use -f parameter to separate them.
```

Must Knows

Last updated：2019-07-26 17:54:23

1. For the commands above, the source and destination versions must be the same.

2. You may fail to copy the source file if another client is writing data to it; you will fail to rewrite a file if it is being copied to the destination; you will fail to copy the source file if it was being moved, and you will see an error saying FileNotFoundException.