How to set up Yandex Data Proc + Neoflex Datagram

1) Before you begin you need to decide whether to use existing Data Proc Cluster or create a new one. The minimal requirements for the Data Proc Cluster are:
   a. At least one master node and one compute node.
   b. Disk size - 20Gb.
   c. User with the permission to copy datagram libraries to the master HDHS system.
   d. Applications installed: hdfs, hive, livy, mapreduce, oozie, spark, sqoop, tez, yarn.

2) Create a Virtual Machine from the image.
   If you would like to create a new Data Proc Cluster for Datagram the Virtual Machine must have service account with the permissions to create the underlying resources, especially Data Proc cluster, Data Proc cluster service account with role **mdb.dataproc.agent.**

3) Login to the datagram instance.

4) If you would like to use existing Data Proc Cluster for Datagram, please change the appropriate parameters in the beginning section of /opt/datagram/create_cluster.sh script. Especially
   a. node_user - the user name who is used to ssh to the cluster nodes. The user must have appropriate permissions to copy datagram libraries to the cluster file system.
   b. key_file - ssh private key file for "node_user"
   c. master_node - IP or FQDN of the master node of the cluster
   d. compute_nodes - list of IP addresses or FQDNs of compute nodes of the cluster
   e. data_nodes - list of IP addresses or FQDNs of data nodes of the cluster

5) If you would like to create a new cluster for datagram, you can, optionally, change the default configuration parameters for the Data Proc Cluster in the beginning section of /opt/datagram/create_cluster.sh script. You may set
   a. Cluster name
   b. Service account name. The service account is the account which will be associated with Data Proc Cluster VMs.
   c. Master node configuration: size, disk, name.
   d. Compute nodes configuration: size, disk, name, number.
   e. Data nodes configuration: size, disk, name, number.

6) Run
   /opt/datagram/create_cluster.sh create
   in case you are going to create a new cluster

   /opt/datagram/create_cluster.sh deploy
   in case you would like to use the existing Data Proc Cluster

   /opt/datagram/create_cluster.sh destroy
   in case you would like to remove the created Data Proc Cluster.