

Как установить Datagram, работающим на кластере Data Proc в Яндекс.Облако

- 1) Изначально нужно определиться будет ли создаваться новый Data Proc кластер или будет использоваться уже существующий. Минимальные требования для Data Proc кластера это:
 - a. Один главный подкластер с управляющим хостом, подкластер для хранения данных и подкластер для вычислений.
 - b. Размер диска - 20Гб.
 - c. Пользователь с правами, которые позволяют скопировать Datagram библиотеки на HDFS систему на мастер главного подкластера.
 - d. Установлены приложения: hdfs, hive, livy, mapreduce, oozie, spark, sqoop, tez, yarn.
- 2) Создайте новую виртуальную машину из образа
Если вы собираетесь создать новый Data Proc кластер, у виртуальной машины должен быть сервисный аккаунт с правами, необходимыми для создания Data Proc кластера и сервисного аккаунта с ролью **mdb.dataproc.agent**
- 3) Зайдите с помощью SSH на вновь созданную виртуальную машину.
- 4) Если вы собираетесь использовать существующий Data Proc кластер, измените параметры в конфигурационной части скрипта запуска (**/opt/datagram/create_cluster.sh**).
Параметры, которые необходимо установить
 - a. `node_user`
имя пользователя для подключения к виртуальным машинам кластеры. Пользователь должен обладать достаточными правами для копирования Datagram библиотек в файловую систему кластера. .
 - b. `key_file`
ssh ключ для входа в систему
 - c. `master_node` - IP адрес или FQDN управляющего хоста главного под-кластера
 - d. `compute_nodes` - список IP адресов или FQDN'ов для под-кластеров с ролью "Compute"
 - e. `data_nodes` - список IP адресов или FQDN'ов для под-кластеров с ролью "Data"
- 5) Если вы собираетесь создать новый Data Proc кластер, вы можете изменить параметры кластера по умолчанию в конфигурационной части скрипта запуска (**/opt/datagram/create_cluster.sh**). Вы можете установить
 - a. Имя кластера
 - b. Имя сервисного аккаунта.
 - c. Конфигурационные параметры управляющего хоста главного под-кластера, такие как размер виртуальной машины, размер диска виртуальной машины, имя виртуальной машины.
 - d. Конфигурационные параметры под-кластеров с ролью "Compute" - размер виртуальной машины, размер диска виртуальной машины, имя виртуальной машины количество виртуальных машин
 - e. Конфигурационные параметры под-кластеров с ролью "Data" - размер виртуальной машины, размер диска виртуальной машины, имя виртуальной машины количество виртуальных машин

6) В терминале запустите
`/opt/datagram/create_cluster.sh create`
для варианта создания нового кластера

`/opt/datagram/create_cluster.sh deploy`
для варианта использования уже существующего кластера

`/opt/datagram/create_cluster.sh destroy`
для удаления созданного кластера.