

Инструкция по развертыванию и настройке кластера hdfs

Мои данные:

Узел для входа 176.109.91.28

Jump node 192.168.1.118

Name node 192.168.1.119

Data node-00 192.168.1.120

Data node-01 192.168.1.121

Открываем командную строку и выполняем в ней следующие команды:

ssh team@176.109.91.28 -- подключаемся к jump ноде

sudo apt install tmux -- устанавливаем менеджер терминалов tmux

tmux -- запускаем tmux

ssh-keygen -- генерируем ssh-ключ

cat .ssh/id_ed25519.pub >> .ssh/authorized_keys - добавляем наш ключ

Распространяем ssh-ключ на ноды:

scp .ssh/authorized_keys 192.168.1.119:/home/team/.ssh/

scp .ssh/authorized_keys 192.168.1.120:/home/team/.ssh/

scp .ssh/authorized_keys 192.168.1.121:/home/team/.ssh/

Через редактор меняем файл hosts, чтобы можно было обращаться к нодам по определенным именам:

sudo vim /etc/hosts

Файл hosts у jump node должен принять похожий вид:

```
# 127.0.0.1 localhost
```

```
127.0.0.1 jn
```

```
# 192.168.1.118 jn
```

```
192.168.1.119 nn
```

```
192.168.1.120 dn-00
```

```
192.168.1.121 dn-01
```

```
# The following lines are desirable for IPv6 capable hosts
```

```
# ::1    ip6-localhost ip6-loopback
```

```
# fe00::0 ip6-localnet
```

```
# ff00::0 ip6-mcastprefix
```

```
# ff02::1 ip6-allnodes
```

```
# ff02::2 ip6-allrouters
```

Сохраняем файл и выходим из редактора.

sudo vim /etc/hostname -- в этом файле меняем изначальное имя ноды на то, которое прописали в hosts

Сохраняем файл и выходим из редактора.

Добавляем пользователя, от имени которого будут выполняться сервисы hadoop:

sudo adduser hadoop

ssh 192.168.1.119 -- заходим на name node

sudo adduser hadoop -- добавляем пользователя

sudo vim /etc/hosts -- редактируем файл hosts для name node:

```
# 127.0.0.1 localhost
```

```
192.168.1.118 jn
```

```
192.168.1.119 nn
```

```
192.168.1.120 dn-00
```

```
192.168.1.121 dn-01
```

```
# The following lines are desirable for IPv6 capable hosts
```

```
# ::1    ip6-localhost ip6-loopback
```

```
# fe00::0 ip6-localnet
```

```
# ff00::0 ip6-mcastprefix
```

```
# ff02::1 ip6-allnodes  
# ff02::2 ip6-allrouters
```

Сохраняем файл и выходим из редактора.

sudo vim /etc/hostname -- в этом файле меняем изначальное имя ноды на то, которое прописали в hosts

Сохраняем файл и выходим из редактора.

exit

ssh 192.168.1.120 -- переключаемся на первую дата ноду

sudo adduser hadoop -- добавляем пользователя

sudo vim /etc/hosts -- редактируем файл hosts для data node-00:

```
# 127.0.0.1 localhost  
127.0.0.1 dn-00  
  
192.168.1.118 jn  
192.168.1.119 nn  
# 192.168.1.120 dn-00  
192.168.1.121 dn-01  
  
# The following lines are desirable for IPv6 capable hosts  
  
# ::1    ip6-localhost ip6-loopback  
# fe00::0 ip6-localnet  
# ff00::0 ip6-mcastprefix  
# ff02::1 ip6-allnodes  
# ff02::2 ip6-allrouters
```

Сохраняем файл и выходим из редактора.

sudo vim /etc/hostname -- в этом файле меняем изначальное имя ноды на то, которое прописали в hosts

Сохраняем файл и выходим из редактора.

exit

ssh 192.168.1.121 -- переключаемся на оставшуюся дата ноду

sudo adduser hadoop -- добавляем пользователя

sudo vim /etc/hosts -- редактируем файл hosts для data node-01:

```
# 127.0.0.1 localhost
127.0.0.1 dn-01

192.168.1.118 jn
192.168.1.119 nn
192.168.1.120 dn-00
# 192.168.1.121 dn-01

# The following lines are desirable for IPv6 capable hosts
# ::1    ip6-localhost ip6-loopback
# fe00::0 ip6-localnet
# ff00::0 ip6-mcastprefix
# ff02::1 ip6-allnodes
# ff02::2 ip6-allrouters
```

Сохраняем файл и выходим из редактора.

sudo vim /etc/hostname -- в этом файле меняем изначальное имя ноды на то, которое прописали в hosts

Сохраняем файл и выходим из редактора.

Теперь все ноды знают друг друга по именам.

Не завершаем сессию tmux, просто из нее выходим (Ctrl + b d)

sudo -i -u hadoop -- переключаемся на пользователя hadoop

tmux -- запускаем сессию tmux

Нужно скачать дистрибутив с hadoop:

wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0.tar.gz

Пока скачивается выходим из сессии (Ctrl + b d)

Ноды знают друг друга по именам, но, чтобы пользователь hadoop мог заходить на другие узлы от своего имени, надо проделать ту же операцию с ssh-ключами - генерацию ключа и распространение его на остальные ноды:

ssh-keygen

cat .ssh/id_ed25519.pub >> .ssh/authorized_keys

scp -r .ssh/ nn:/home/hadoop

scp -r .ssh/ dn-00:/home/hadoop

scp -r .ssh/ dn-01:/home/hadoop

Мы сделали ключ для пользователя hadoop и раскидали его по всем нодам.

Подключаемся обратно к сессиии tmux где скачивался дистрибутив. Разбрасаем его на все ноды:

scp hadoop-3.4.0.tar.gz nn:/home/hadoop

scp hadoop-3.4.0.tar.gz dn-00:/home/hadoop

scp hadoop-3.4.0.tar.gz dn-01:/home/hadoop

Начнем настраивать Hadoop:

Распаковываем архив на каждой ноде:

tar -xzf hadoop-3.4.0.tar.gz

ssh nn

tar -xzf hadoop-3.4.0.tar.gz

ssh dn-00

tar -xzf hadoop-3.4.0.tar.gz

ssh dn-01

tar -xzf hadoop-3.4.0.tar.gz

ssh jn

exit -- закрыли подключение к jn

exit -- закрыли подключение к dn-01

exit -- закрыли подключение к dn-00

exit -- закрыли подключение к nn

vim .profile -- редачим файл

Добавляем в него строки:

```
export HADOOP_HOME=/home/hadoop/hadoop-3.4.0
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

Сохраняем файл и выходим из редактора.

source .profile -- активируем изменения

Раскидываем файл на все ноды:

scp .profile dn-01:/home/hadoop

scp .profile dn-00:/home/hadoop

scp .profile nn:/home/hadoop

exit

Нужно поправить скрипт, который устанавливает переменные окружения, в него нужно добавить путь к джаве:

cd hadoop-3.4.0/etc/hadoop/

vim hadoop-env.sh

Добавляем в файл строчку:

```
JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

Сохраняем файл и выходим из редактора.

vim core-site.xml

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://nn:9000</value>
  </property>
</configuration>
```

vim hdfs-site.xml -- устанавливаем фактор репликации 3

```
<configuration>
  <property>
    <name>dfs.replication</name>
```

```
<value>3</value>

</property>

</configuration>
```

Сохраняем файл и выходим из редактора.

vim workers

```
# localhost
```

```
nn
```

```
dn-00
```

```
dn-01
```

Сохраняем файл и выходим из редактора.

Копируем файлы на все ноды:

```
scp hadoop-env.sh nn:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp hadoop-env.sh dn-00:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp hadoop-env.sh dn-01:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp core-site.xml nn:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp core-site.xml dn-00:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp core-site.xml dn-01:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp hdfs-site.xml nn:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp hdfs-site.xml dn-00:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp hdfs-site.xml dn-01:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp workers nn:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp workers dn-00:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

```
scp workers dn-01:/home/hadoop/hadoop-3.4.0/etc/hadoop
```

Запускаем Hadoop:

```
ssh nn -- переключаемся на Name node
```

cd hadoop-3.4.0/

bin/hdfs namenode -format -- создает и форматирует файловую систему

sbin/start-dfs.sh - запускаем кластер hadoop