Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ ПЕРВОГО ПРЕЗИДЕНТА РОССИИ Б.Н. ЕЛЬЦИНА
(УрФУ имени первого Президента России Б.Н. Ельцина)
Институт радиоэлектроники и информационных технологий — РТФ

ОТЧЁТ

по лабораторной работе №2

по дисциплине «Методы и инструменты анализа больших данных»

Преподаватель _____ (дата) _____ (подпись) С.Г. Мирвода

Студент _____ (дата) _____ (подпись) А.М. Белоусов

Студент _____ (дата) _____ (подпись) А.В. Жиденко

Группа: РИМ-201211

Екатеринбург 2021

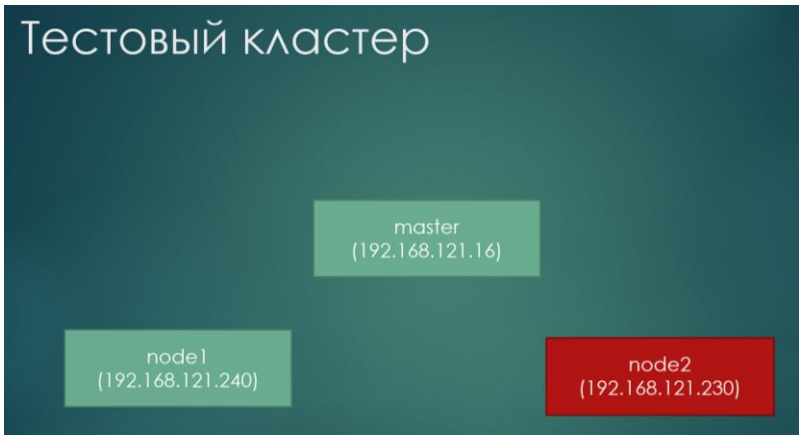
Цель работы: знакомство с распределённой файловой системой HDFS.

Задание 0

Задача: подготовить полигон.

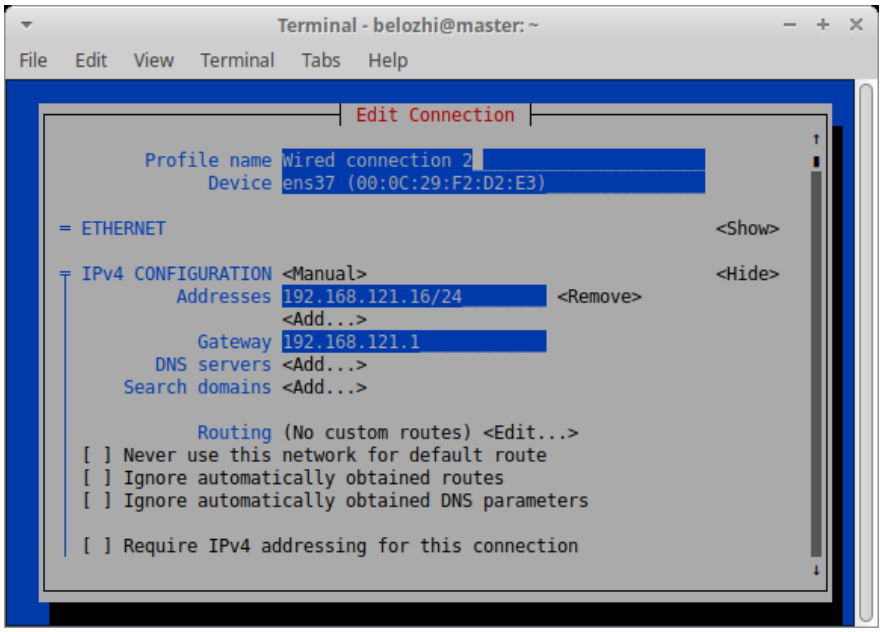
1. Установить и настроить кластер HDFS согласно инструкции Cluster и примеру

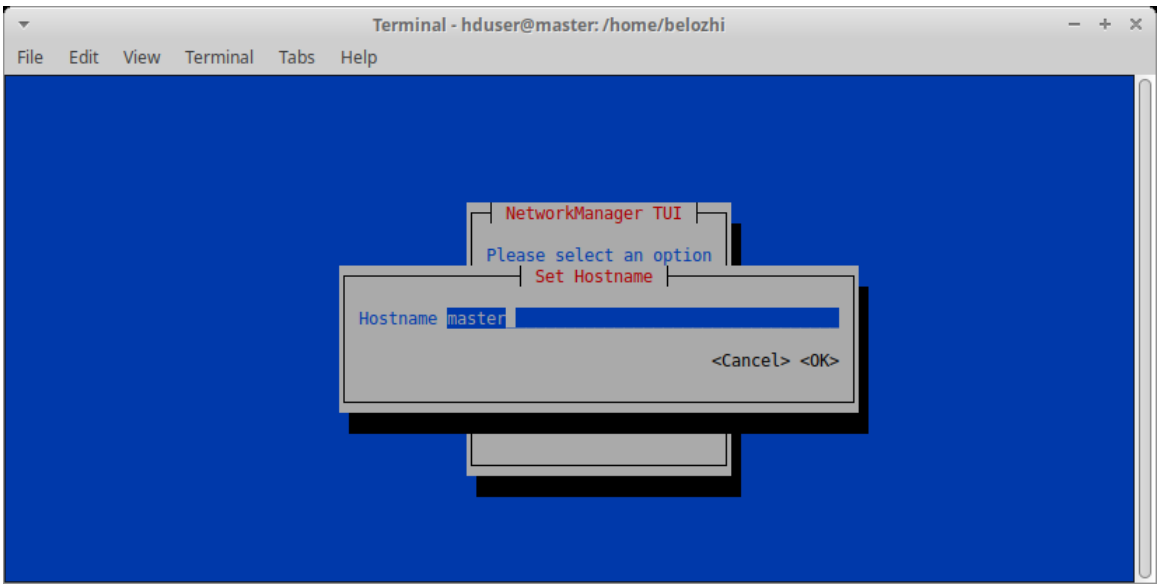
Настройка ip адреса согласно схеме задания

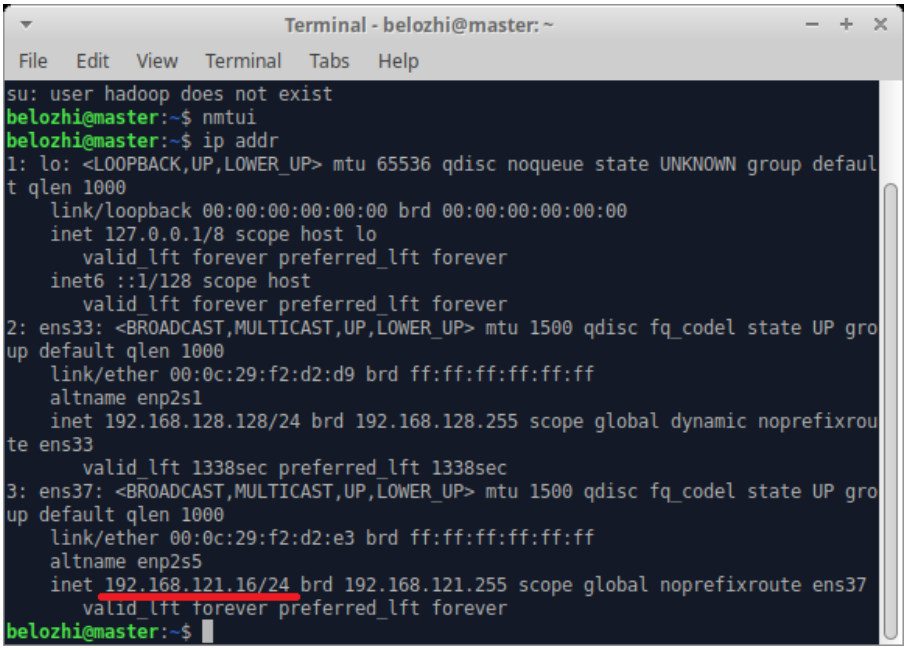


Настроим имя и ip-адрес ВМ Master с помощью утилиты nmtui.

Nmtui







Установка ssh

sudo apt-get update && sudo apt-get upgrade

sudo apt-get install ssh

Установка Java

$ sudo apt-get install openjdk-8-jdk

Создание отдельной учетной записи для запуска Hadoop

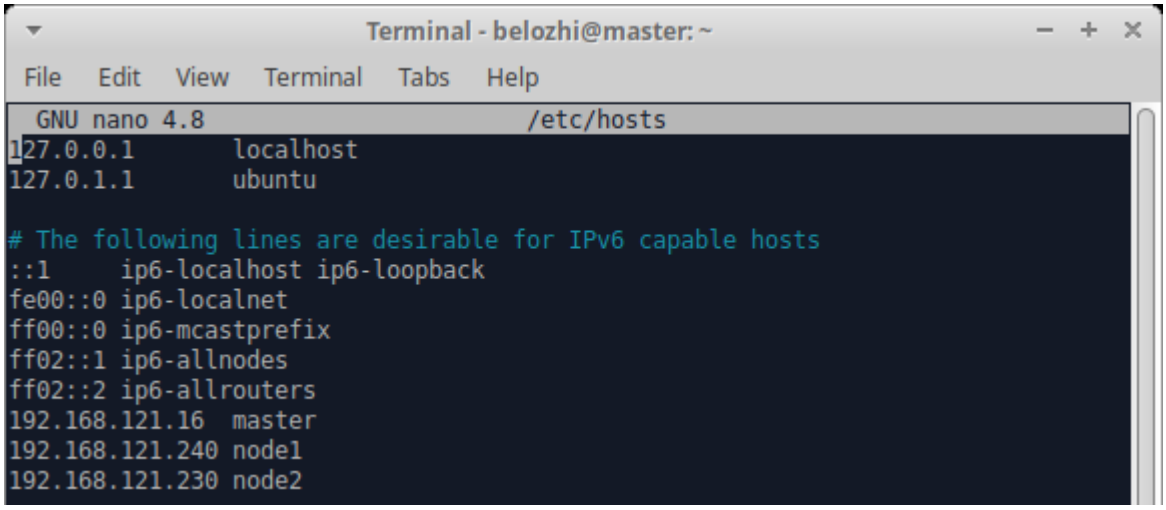
sudo addgroup hadoop

sudo adduser --ingroup hadoop hduser

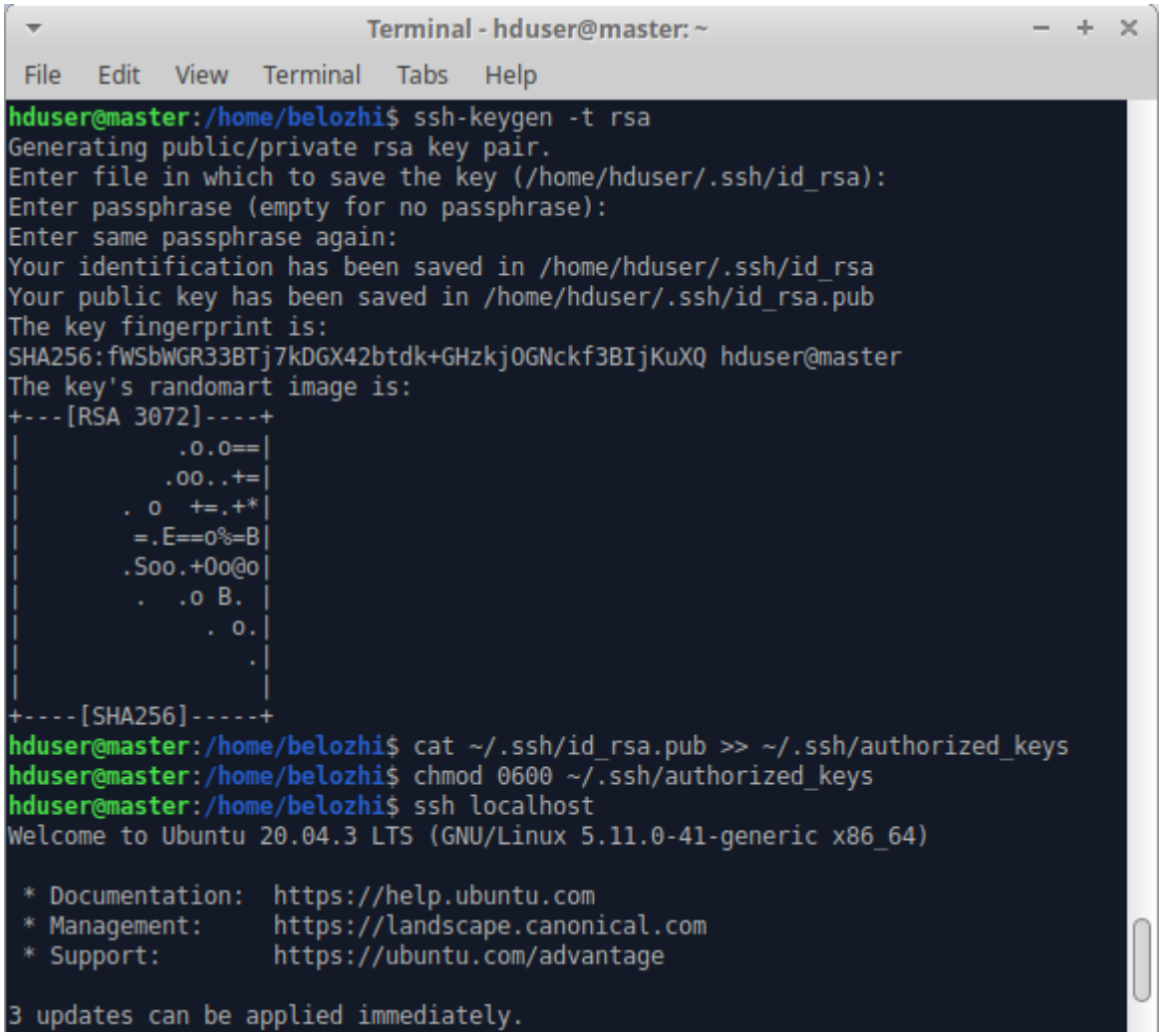
sudo usermod -aG sudo hduser

```
root@ubuntu:/home/belozhi# sudo adduser --ingroup hadoop hduser
Adding user `hduser' ...
Adding new user `hduser' (1001) with group `hadoop' ...
Creating home directory `/home/hduser' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
Sorry, passwords do not match.
passwd: Authentication token manipulation error
passwd: password unchanged
Try again? [y/N] y
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hduser
Enter the new value, or press ENTER for the default
        Full Name []: master
        Room Number []: 1
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] y
root@ubuntu:/home/belozhi# sudo usermod -aG sudo hduser
```

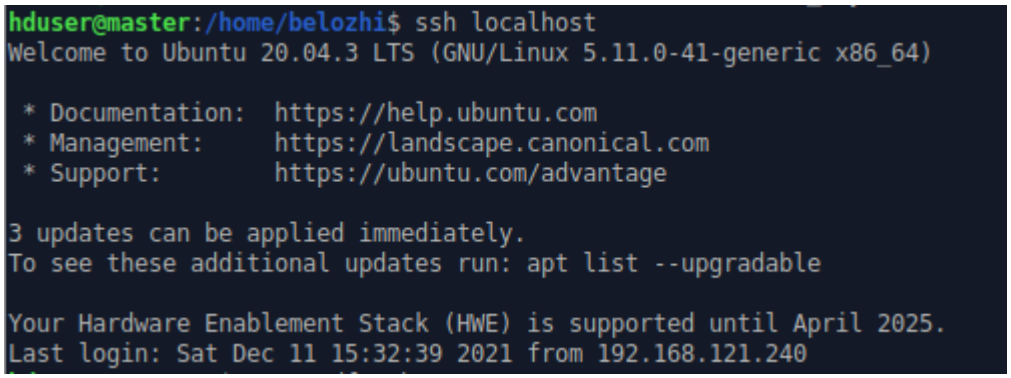
Редактирование файла /etc/hosts



Получение SSH ключей



Проверка подключения по ssh (*ssh localhost*)



Установка Apache Hadoop в Ubuntu 20.04

su - hduser

wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz

tar -xvzf hadoop-3.3.0.tar.gz

Затем переместим извлеченный каталог в :/usr/local/

sudo mv hadoop-3.3.0 /usr/local/hadoop

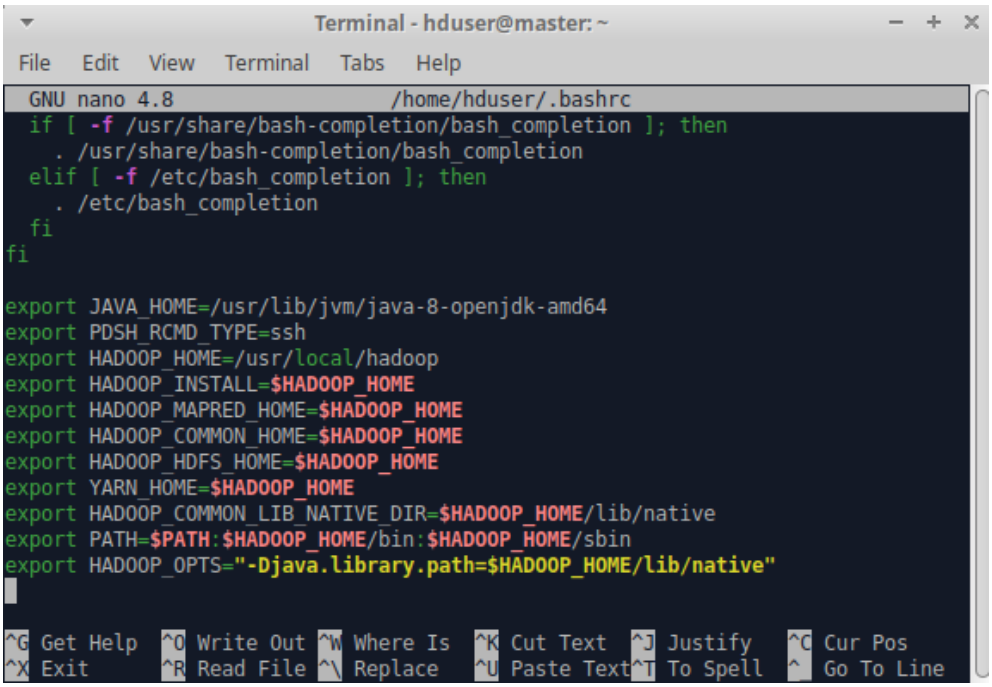
sudo mkdir /usr/local/hadoop/logs

chown -R hduser:hadoop /usr/local/hadoop

Настройте Apache Hadoop

Настройка переменных среды в файле ~/.bashrc

nano ~/.bashrc



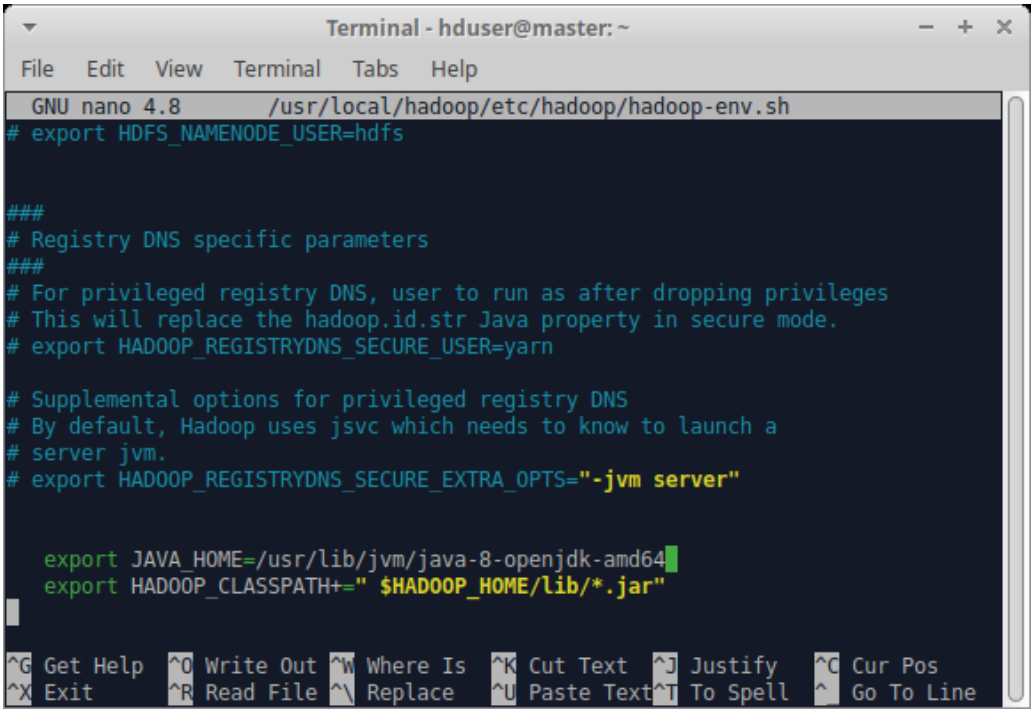
source ~/.bashrc

Затем нужно определить переменные среды Java, чтобы настроить параметры проекта, связанные с YARN, HDFS, MapReduce и Hadoop:hadoop-env.sh

sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh

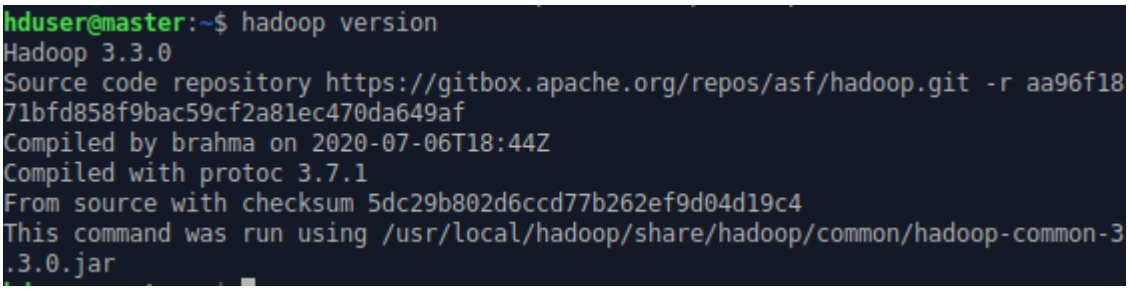
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export HADOOP_CLASSPATH+=" $HADOOP_HOME/lib/*.jar"



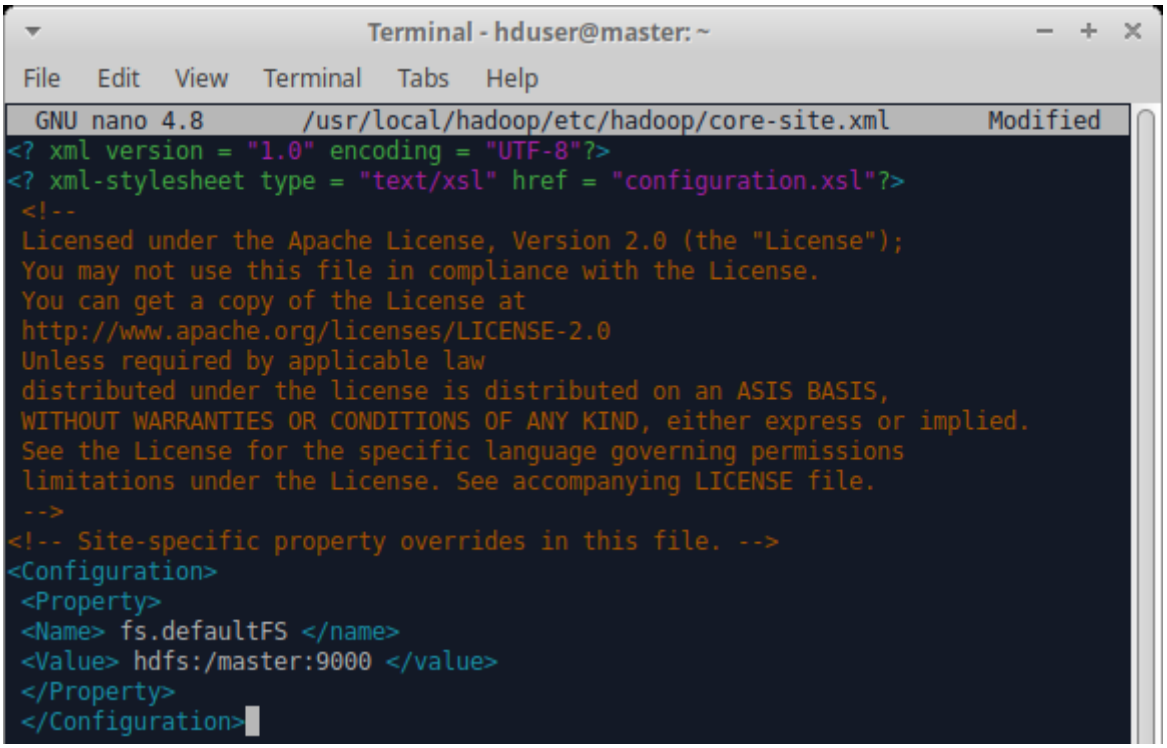
Проверка версии Hadoop

hadoop version

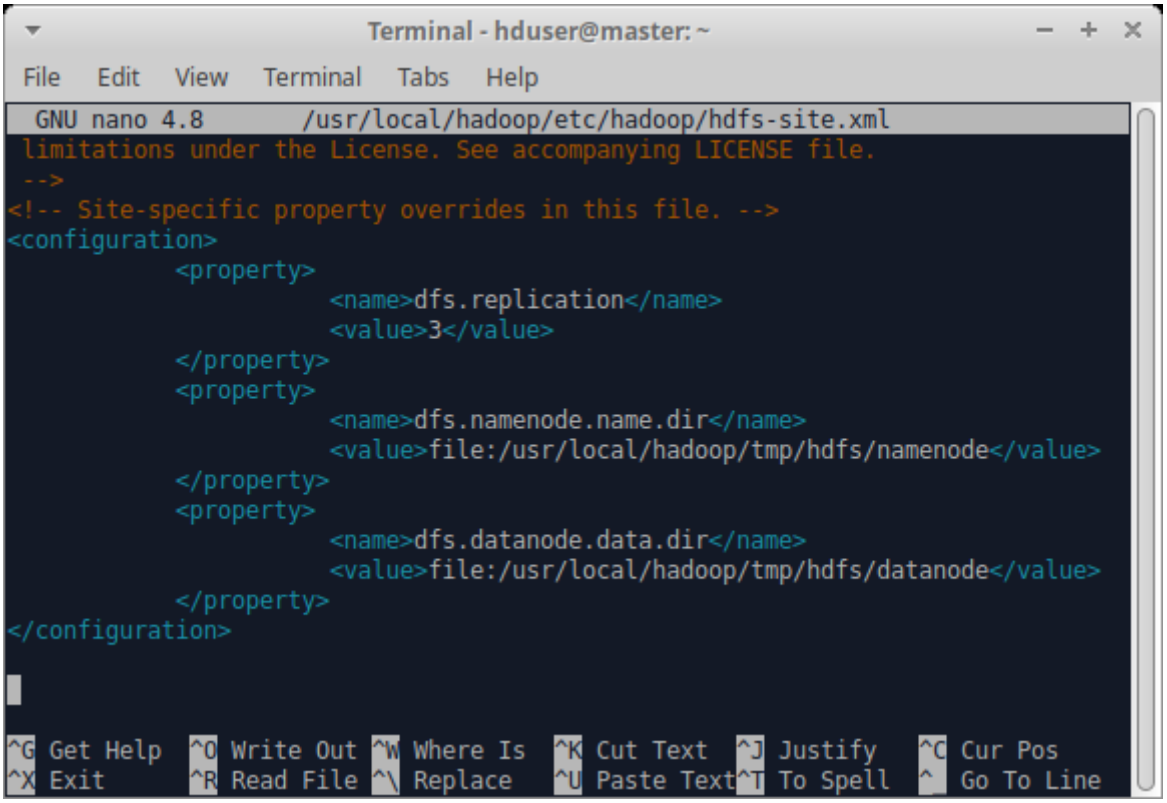


Настройка файлов

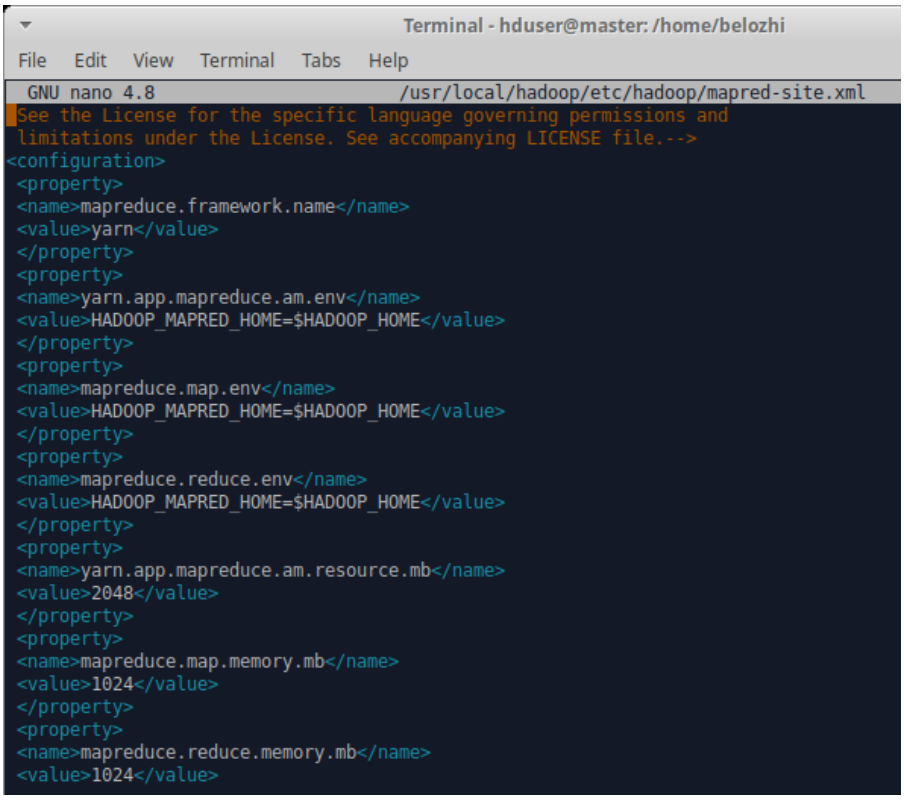
Настроить файл core-site.xml



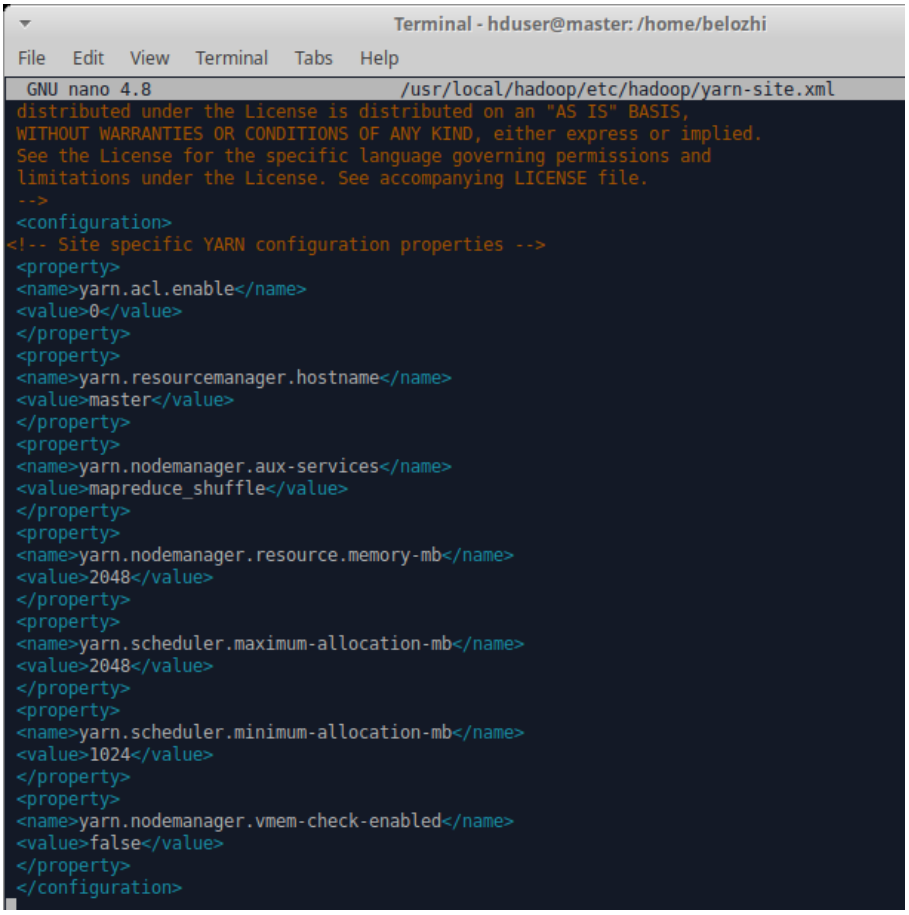
Настроить файл hdfs-site.xml



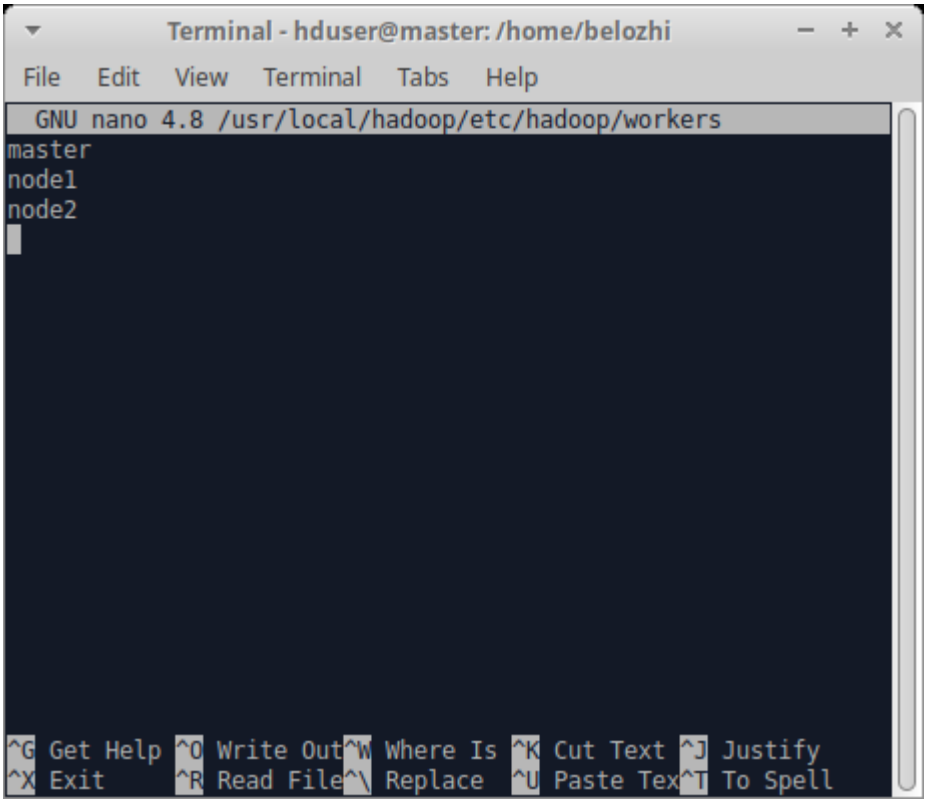
Настроить файл mapred-site.xml



Настроить файл yarn-site.xml



Добавим на узле master все рабочие узлы в файл etc/hadoop/workers.



Форматирование HDFS NameNode

su - hduser

hdfs namenode -format

Запуск кластера Hadoop

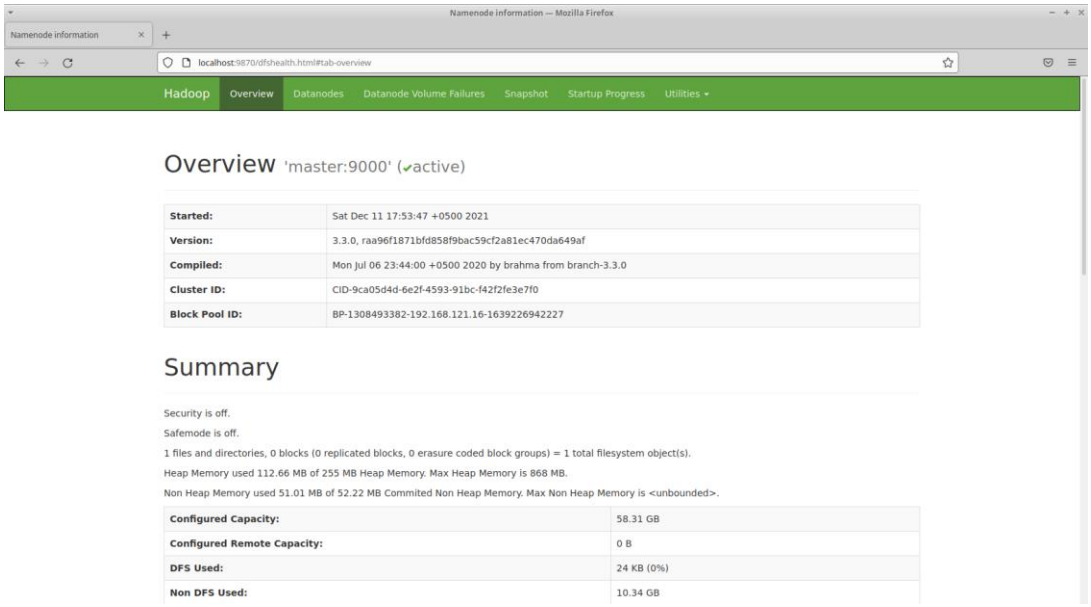
start-dfs.sh

```
hduser@master:~$ start-dfs.sh
Starting namenodes on [master]
Starting datanodes
Starting secondary namenodes [master]
```

start-yarn.sh

```
hduser@master:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

Проверка доступа к Apache Hadoop



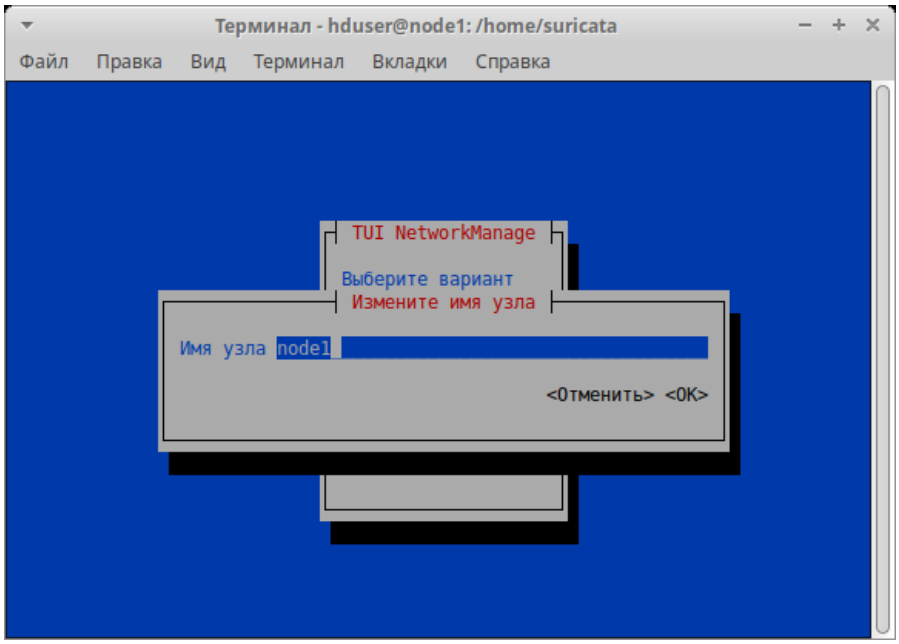
Задание 1

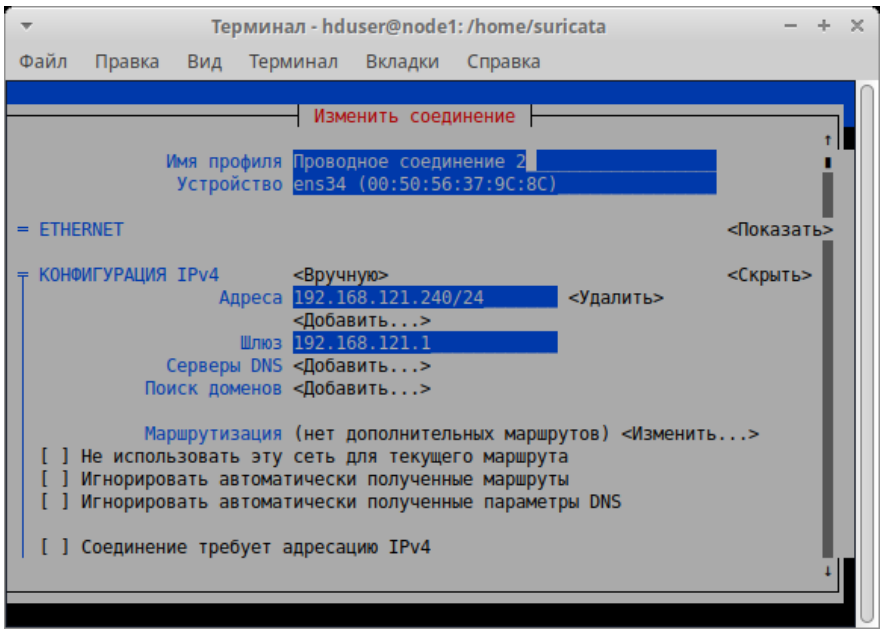
Задача: познакомиться с процессом добавления новых узлов в существующий кластер.

1. Воспользовавшись знаниями из Задания 0 и схемой кластера из презентации (Лабораторная №2), установите и настройте DataNode на узлы кластера node1 и node2.

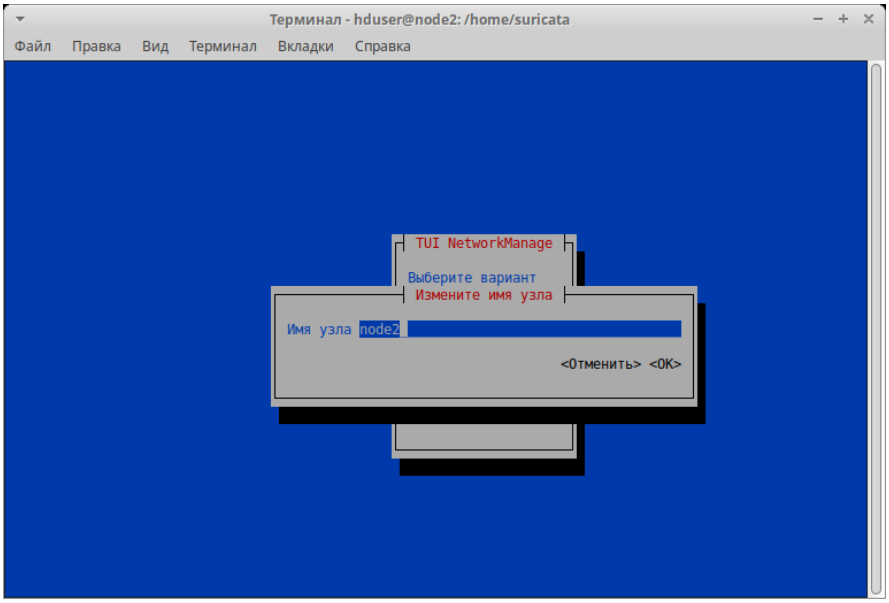
Настройка сетевых имён и ip-адресов на ВМ node1 и node2.

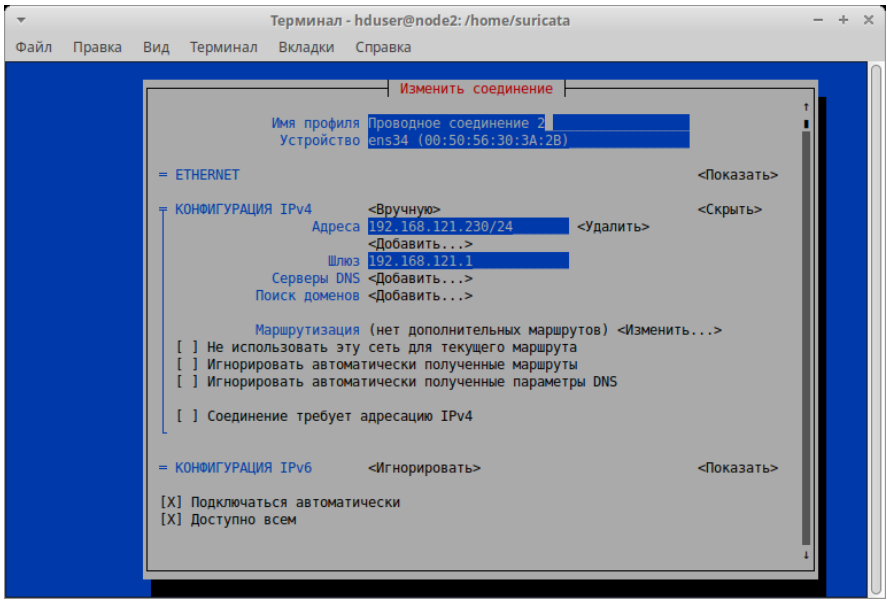
node1




Терминал - hduser@node1: /home/suricata
Файл Правка Вид Терминал Вкладки Справка
Изменить соединение
Имя профиля Проводное соединение 2
Устройство ens34 (00:50:56:37:9C:8C)
= ETHERNET <Показать>
= КОНФИГУРАЦИЯ IPv4 <Вручную> <Скрыть>
Адреса 192.168.121.240/24 <Удалить>
<Добавить...>
Шлюз 192.168.121.1
Серверы DNS <Добавить...>
Поиск доменов <Добавить...>
Маршрутизация (нет дополнительных маршрутов) <Изменить...>
[ ] Не использовать эту сеть для текущего маршрута
[ ] Игнорировать автоматически полученные маршруты
[ ] Игнорировать автоматически полученные параметры DNS
[ ] Соединение требует адресацию IPv4

node2


Терминал - hduser@node2: /home/suricata
Файл Правка Вид Терминал Вкладки Справка
TUI NetworkManage
Выберите вариант
Измените имя узла
Имя узла node2
<Отменить> <OK>


Терминал - hduser@node2: /home/suricata
Файл Правка Вид Терминал Вкладки Справка
Изменить соединение
Имя профиля Проводное соединение 2
Устройство ens34 (00:50:56:30:3A:2B)
= ETHERNET <Показать>
= КОНФИГУРАЦИЯ IPv4 <Вручную> <Скрыть>
Адреса 192.168.121.230/24 <Удалить>
<Добавить...>
Шлюз 192.168.121.1
Серверы DNS <Добавить...>
Поиск доменов <Добавить...>
Маршрутизация (нет дополнительных маршрутов) <Изменить...>
[ ] Не использовать эту сеть для текущего маршрута
[ ] Игнорировать автоматически полученные маршруты
[ ] Игнорировать автоматически полученные параметры DNS
[ ] Соединение требует адресацию IPv4
= КОНФИГУРАЦИЯ IPv6 <Игнорировать> <Показать>
[X] Подключаться автоматически
[X] Доступно всем

Далее аналогичным образом (как в в Задании 0) выполняется:

- Настройка учетной записи hduser
- Получение SSH ключей
- Установка и настройка Hadoop

Копируем ключи SSH на подчиненные машины:

hduser@master: ~$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hduser@node1

и на вторую:

hduser@master: ~$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hduser@ node2

Последний шаг настройки - удаление старых каталогов HDFS.

Для этого на мастере и подчиненных узлах нужно удалить каталоги **/usr/local/hadoop/hadoop_tmp/**:

sudo rm -rf /usr/local/hadoop/hadoop_tmp/

Далее на **мастере** создать каталоги namenode и datanode, так как мастер-машина у нас одновременно мастер и подчиненная.

Команды:

sudo mkdir -p /usr/local/hadoop/hadoop_tmp/hdfs/namenode

sudo mkdir -p /usr/local/hadoop/hadoop_tmp/hdfs/datanode

Затем назначить владельцем каталога пользователя hduser группы hadooop:

sudo chown hduser:hadoop -R /usr/local/hadoop/hadoop_tmp/

На подчиненных машинах сделать тоже самое, но создать только каталог datanode:

sudo mkdir -p /usr/local/hadoop/hadoop_tmp/hdfs/datanode

sudo chown hduser:hadoop -R /usr/local/hadoop/hadoop_tmp/

Теперь все готово, отформатируем файловую систему, выполнив команду на Мастере:

hdfs namenode -format

Если нет ошибок, запустим hdfs и yarn с помощью скрипта из первой части, либо командами:

start-dfs.sh

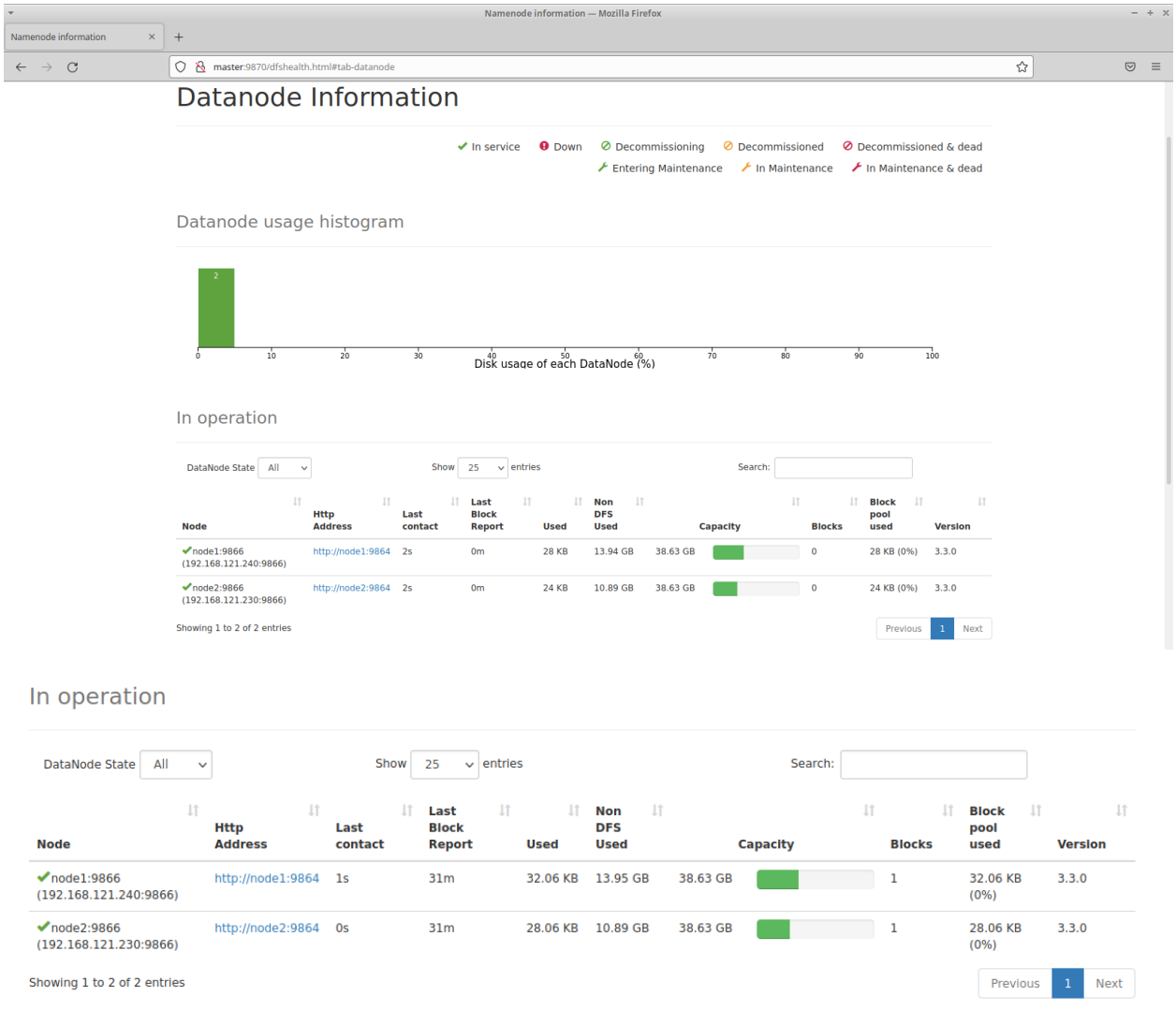
start-yarn.sh

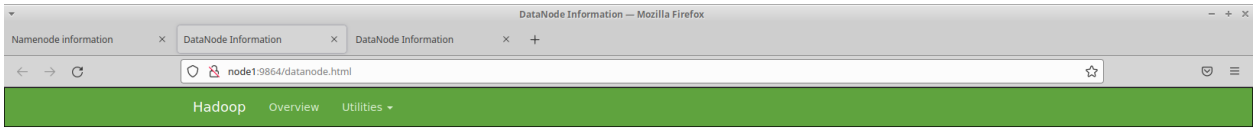
```
hduser@master:/home/belozhi$ start-dfs.sh
Starting namenodes on [master]
Starting datanodes
Starting secondary namenodes [master]
hduser@master:/home/belozhi$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
node2: ERROR: Cannot set priority of nodemanager process 2023
```

С помощью команды jps можно посмотреть запущенные java-процессы на мастере.

```
hduser@master:/home/belozhi$ jps
15136 Jps
5205 NodeManager
4758 SecondaryNameNode
4971 ResourceManager
4285 NameNode
```

Проверка работы кластера в web-интерфейсе.





DataNode on node1:9866

Cluster ID:	CID-5ba5b554-c819-470b-b7ec-4c6734423567
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af

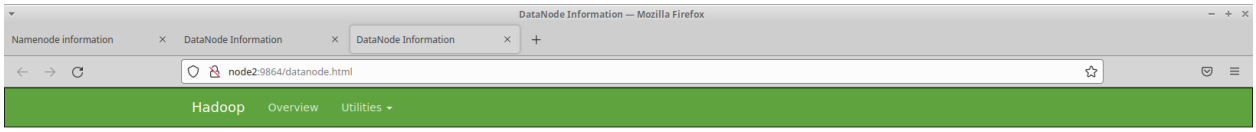
Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
master:9000	BP-939744533-192.168.121.16-1639248653348	RUNNING	0s	9 minutes	0 B (128 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/usr/local/hadoop/tmp/hdfs/datanode	DISK	28 KB	22.7 GB	0 B	0 B	0

Hadoop, 2020.



DataNode on node2:9866

Cluster ID:	CID-5ba5b554-c819-470b-b7ec-4c6734423567
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
master:9000	BP-939744533-192.168.121.16-1639248653348	RUNNING	0s	9 minutes	0 B (128 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/usr/local/hadoop/tmp/hdfs/datanode	DISK	24 KB	25.75 GB	0 B	0 B	0

Hadoop, 2020.

2. Подключитесь к NameNode с помощью ssh и выполните команду hdfs dfsadmin -report в отчёте вы должны увидеть число узлов (Live datanodes) и свой свежедобавленный узел.

```
hduser@master:/home/belozhi$ ssh node2
Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.11.0-41-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

17 updates can be applied immediately.
Чтобы просмотреть дополнительные обновления выполните: apt list --upgradable

Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Sat Dec 11 23:58:01 2021 from 192.168.121.16
```

```
hduser@node2:~$ hdfs dfsadmin -report
Configured Capacity: 82951274496 (77.25 GB)
Present Capacity: 52018753536 (48.45 GB)
DFS Remaining: 52018700288 (48.45 GB)
DFS Used: 53248 (52 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
```

```
-------------------------------------------------
Live datanodes (2):

Name: 192.168.121.230:9866 (node2)
Hostname: node2
Decommission Status : Normal
Configured Capacity: 41475637248 (38.63 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 11692052480 (10.89 GB)
DFS Remaining: 27646300160 (25.75 GB)
DFS Used%: 0.00%
DFS Remaining%: 66.66%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sun Dec 12 00:06:38 YEKT 2021
Last Block Report: Sun Dec 12 00:03:20 YEKT 2021
Num of Blocks: 0
```

```
Name: 192.168.121.240:9866 (node1)
Hostname: node1
Decommission Status : Normal
Configured Capacity: 41475637248 (38.63 GB)
DFS Used: 28672 (28 KB)
Non DFS Used: 14965948416 (13.94 GB)
DFS Remaining: 24372400128 (22.70 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.76%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sun Dec 12 00:06:37 YEKT 2021
Last Block Report: Sun Dec 12 00:03:19 YEKT 2021
Num of Blocks: 0
```

Задание 2

Задача: работа с HDFS с помощью консольных утилит.

1. Разделитесь на команды и с помощью утилиты hadoop создайте в кластере:

1.1 Папку с идентификатором своей команды

```
hduser@node2:~$ hadoop fs -mkdir hdfs://master:9000/Belousov_Zhidenko
mkdir: `hdfs://master:9000/Belousov_Zhidenko': File exists
```

```
hduser@node2:~$ hadoop fs -ls /
Found 1 items
drwxr-xr-x   - hduser supergroup          0 2021-12-12 00:17 /Belousov_Zhidenko
```

1.2 Внутрь папки положите файл с фамилиями участников команды

```
hduser@node2:~$ hadoop fs -ls /
Found 1 items
drwxr-xr-x   - hduser supergroup          0 2021-12-12 00:28 /Belousov_Zhidenko
hduser@node2:~$ hadoop fs -ls /Belousov_Zhidenko
Found 1 items
-rw-r--r--   3 hduser supergroup         52 2021-12-12 00:28 /Belousov_Zhidenko/BeloZhi.txt
```

```
hduser@node2:~$ hadoop fs -cat /Belousov_Zhidenko/BeloZhi.txt
1. Belousov Andrey
2. Zhidenko Alexander
RIM-201211
```

Посмотреть файлы в каталоге Hadoop можно также через web-интерфейс.