|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |

**Институт информационных технологий (ИИТ)**

**Кафедра прикладной математики (ПМ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ**

по дисциплине «Методы анализа данных»

**Практическая работа № 1**

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы ИНБО-01-17 | *ИНБО-05-22, Новиков Денис Владиславович* | (подпись) | |
| Преподаватель | *Шкерин Алексей Леонидович* | (подпись) | |
| Отчет представлен | «11» сентября 2024 г. | |  | |

Москва 2024 г.

## Постановка задачи:

Изучить HDFS и работу с ней с помощью команд и веб-интерфейса, а также работу с YARN/MapReduce.

## Ход работы:

### Работа с HDFS

Чтобы начать использовать службу Hadoop, мы должны запустить демоны службы Hadoop. Мы делаем это как пользователь «hadoop» из нашей среды Linux. Изменяем пользователя на «hadoop» с помощью команды su на Рисунке 1.

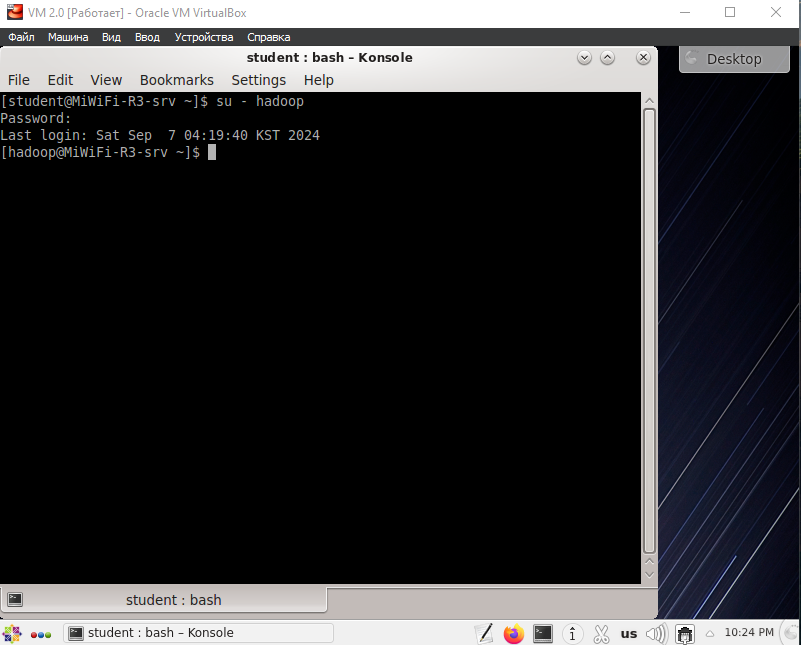


Рисунок 1 — Изменение пользователя

Меняем рабочий каталог на ~/hadoop/sbin и выполняем сценарий start-dfs.sh, чтобы запустить службы HDFS на Рисунке 2.

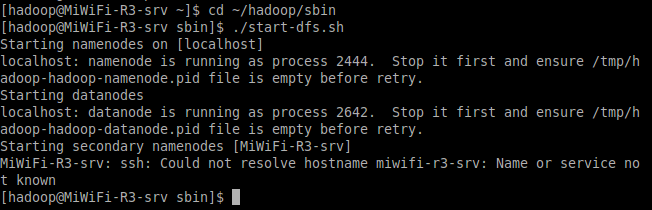


Рисунок 2 — Запуск HDFS

Выполняем сценарий start-yarn.sh, чтобы запустить службы YARN на Рисунке 3.

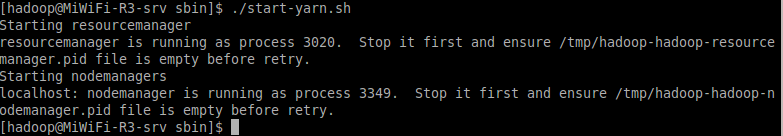


Рисунок 3 — Запуск HDFS

Выходим из сеанса Hadoop на Рисунке 4.



Рисунок 4 — Выход из пользователя hadoop

В терминале от имени пользователя student введем следующую команду на Рисунке 5.

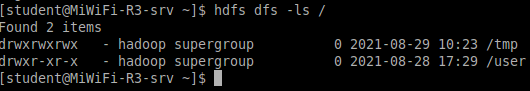


Рисунок 5 — Домашние каталоги пользователей в HDFS

Далее изучим домашние каталоги пользователей в HDFS на Рисунке 6.

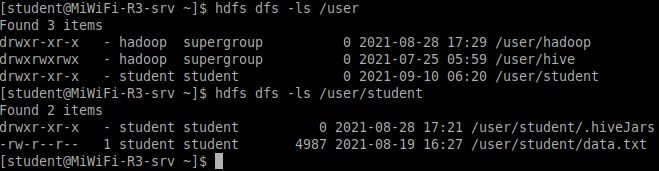


Рисунок 6 — Домашние каталоги пользователей в HDFS

Исследуем домашние каталоги пользователей для Linux на Рисунке 7.

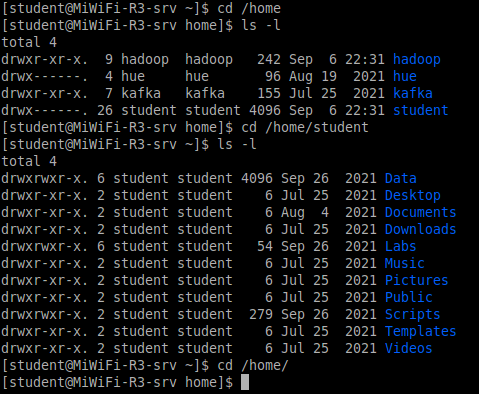


Рисунок 7 — Домашние каталоги пользователей Linux

Создаем пользователя Linux и домашний каталог. Для этого нам потребуются привилегии суперпользователя. Префикс команд sudo позволяет нам выполнять привилегированные команды. Мы можем использовать sudo, потому что студент является частью группы wheel, которой была предоставлена эта возможность. На Рисунке 8 представлен результат добавления новой учетной записи Linux.

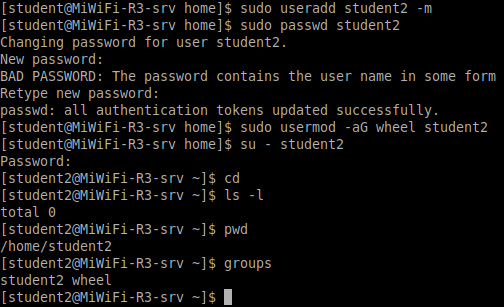


Рисунок 8 — Добавление нового пользователя Linux

Создадим домашний каталог HDFS для пользователя student2. Оставим терминал student2 открытым и откроем другой терминал. Нам нужно будет вводить команды от имени пользователя Hadoop, чтобы внести изменения в HDFS. Обычные пользователи могут вносить изменения только в свой домашний каталог HDFS. Однако им могут быть предоставлены дополнительные привилегии с помощью политик безопасности. Мы создадим новый каталог в /user/student2. Далее мы изменим владельца (chown) этого каталога на student2. На Рисунке 9 представлен результат выполнения указанных команд.

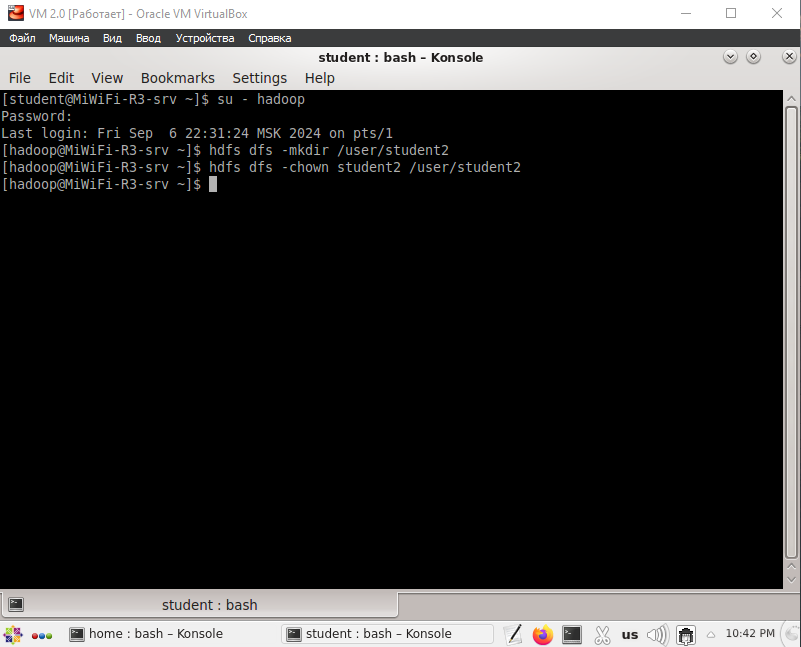


Рисунок 9 — Создание домашнего каталога HDFS для пользователя student2

Теперь вернемся к терминалу для student2 и убедимся, что для этого пользователя создан новый домашний каталог HDFS на Рисунке 10.

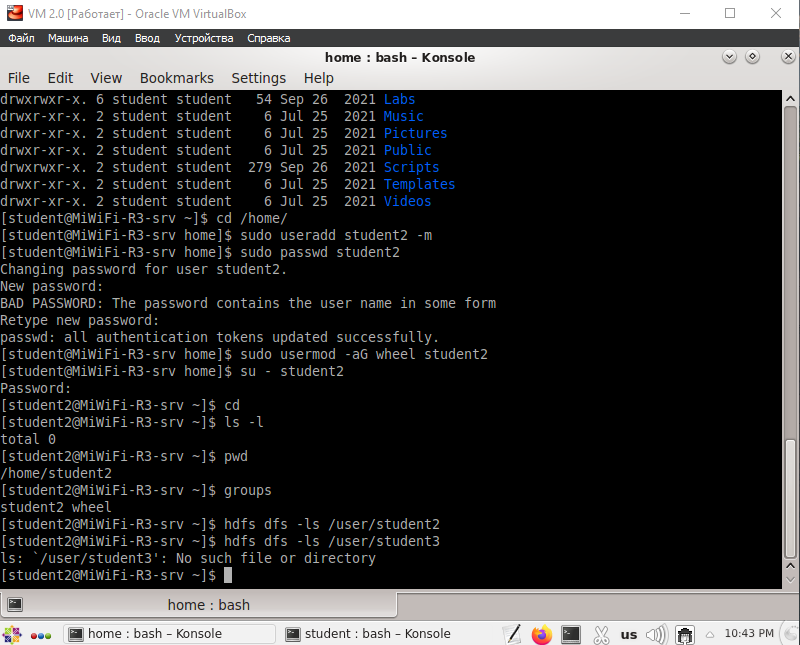


Рисунок 10 — Добавление нового пользователя Linux

Создаем подкаталог в домашнем каталоге студента HDFS и называем его «MRtest» на Рисунке 11.

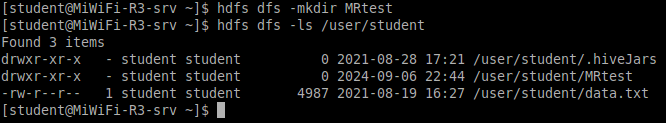


Рисунок 11 — Добавление нового пользователя Linux

Исследуем книгу, расположенную в подкаталоге вашего домашнего каталога на Рисунке 12 и 13.



Рисунок 12 — Исследование книги

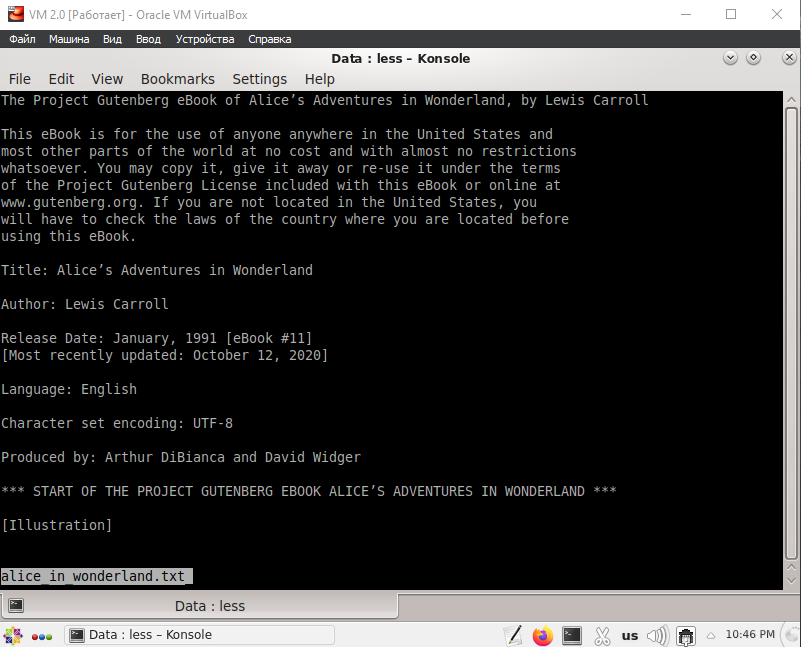


Рисунок 13 — Исследование книги

Поместим файл книги в каталог MRtest, а также на убедимся, что файл был скопирован в место назначения HDFS на Рисунке 14.

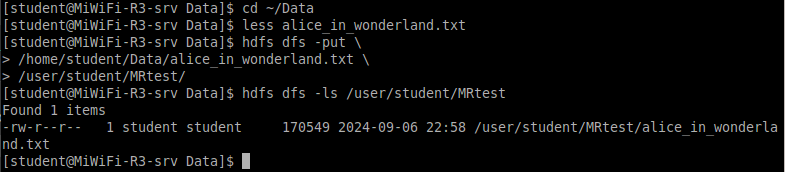


Рисунок 14 — Исследование книги

Исследуем HDFS с помощью веб-интерфейса. Откроем веб-браузер Firefox, выбрав на панели внизу слева. Откройте веб-интерфейс Namenode по адресу http://localhost:50070 на Рисунке 15.

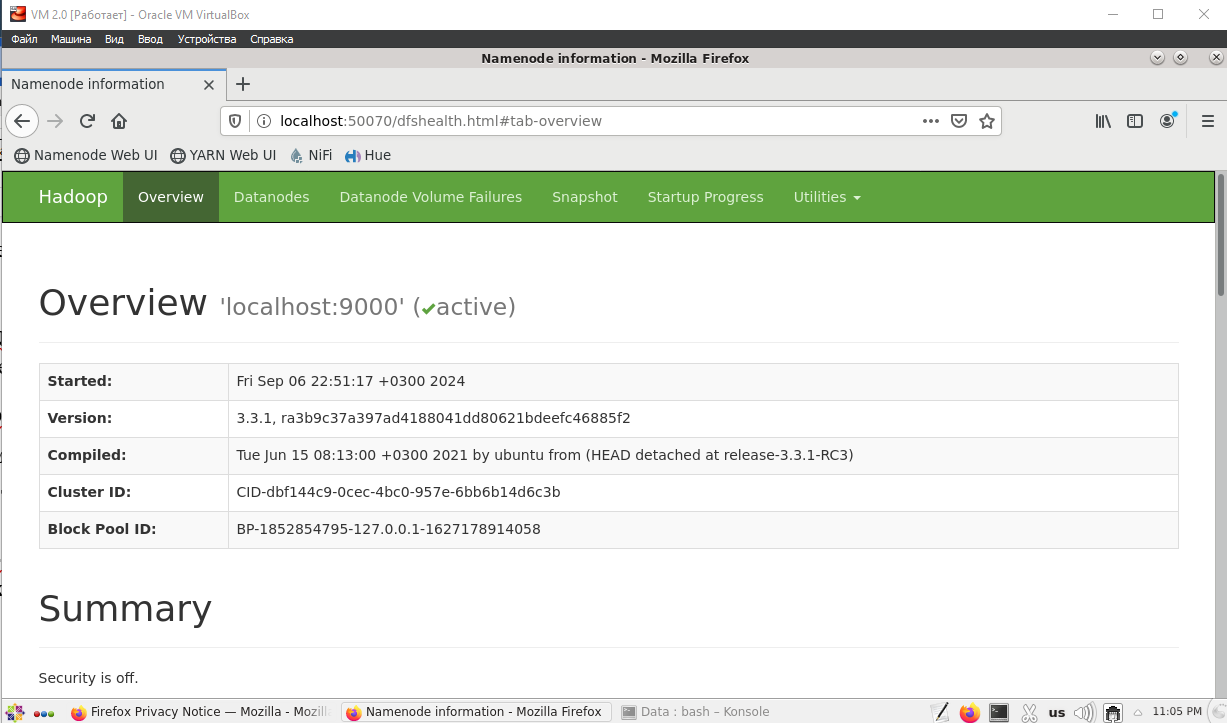


Рисунок 15 — Вкладка Overview

Изучим вкладку «Datanodes» («Узлы данных») на Рисунке 16.

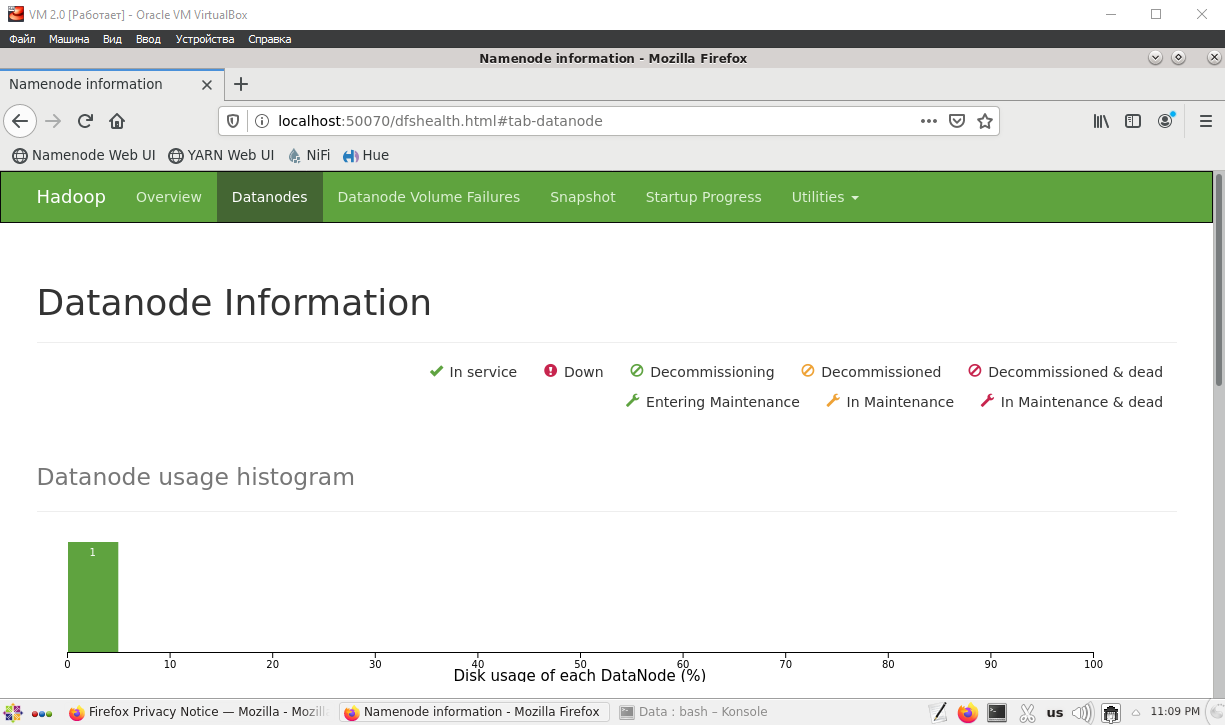


Рисунок 16 — Вкладка Datanodes

Изучим вкладку «Datanode Volume Failures» («Сбои тома данных») на Рисунке 17.

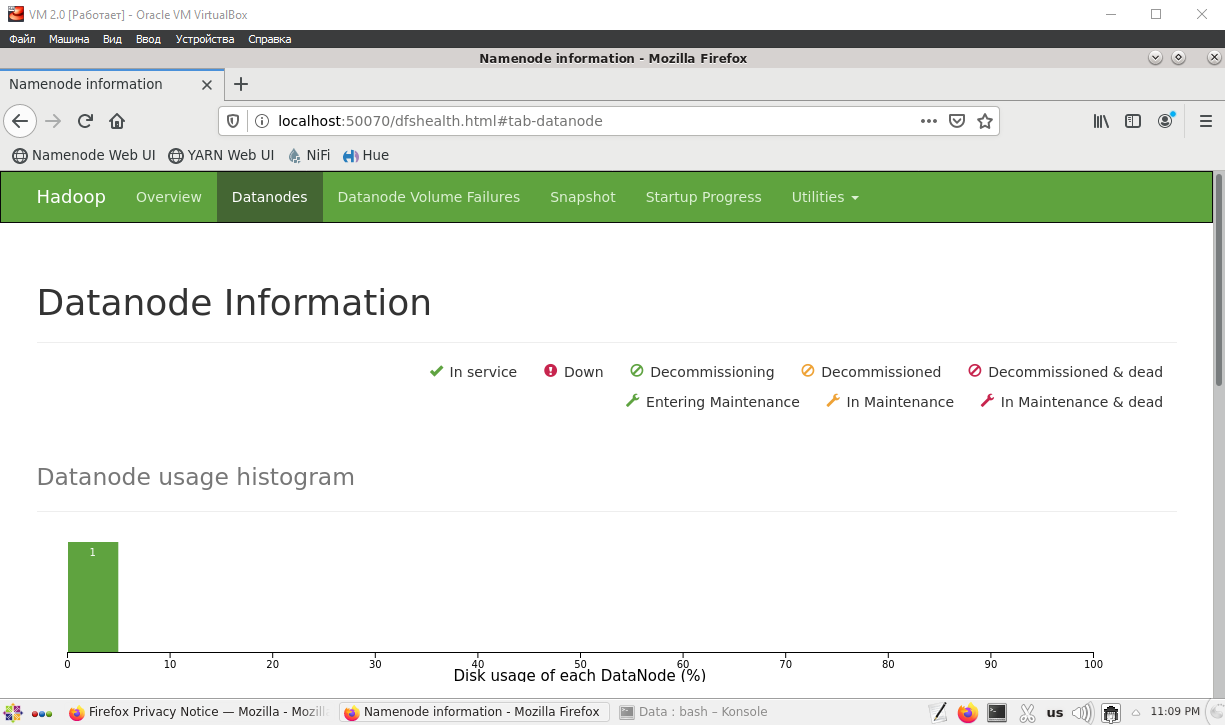


Рисунок 17 — Вкладка Datanode Volume Failures

Изучим вкладку «Snapshot» на Рисунке 18.

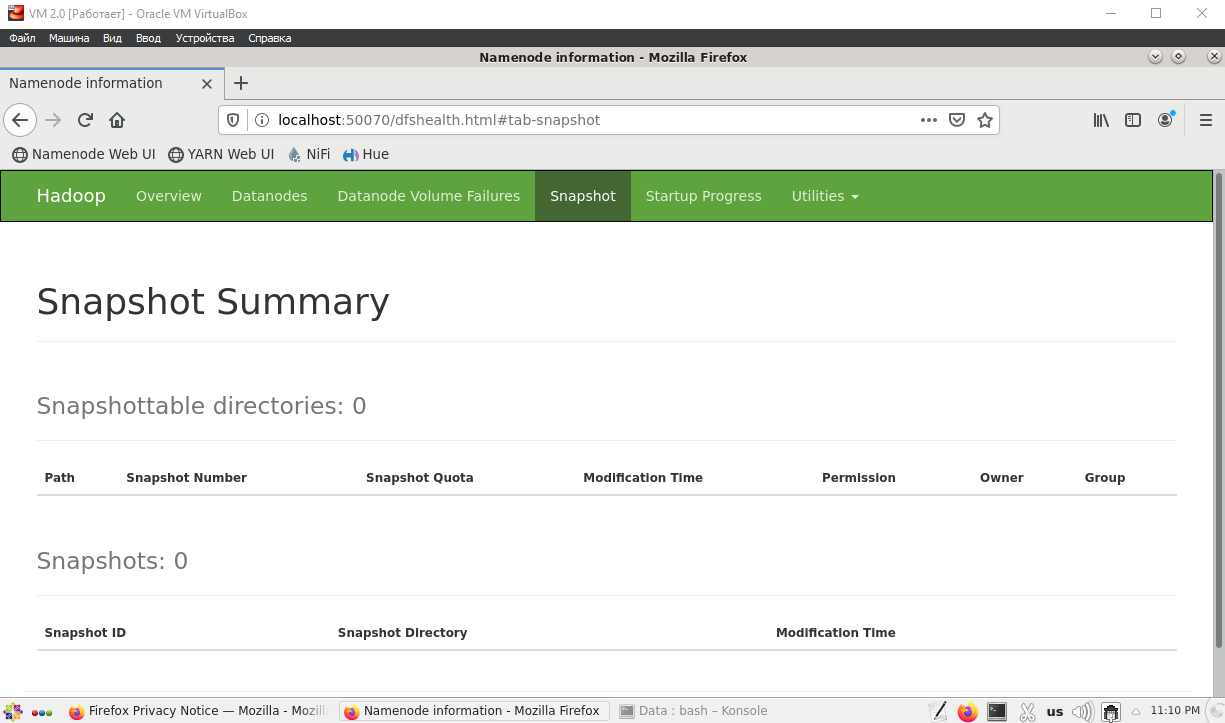


Рисунок 188 — Вкладка Snapshot

Изучим вкладку «Startup Progress» на Рисунке 19.

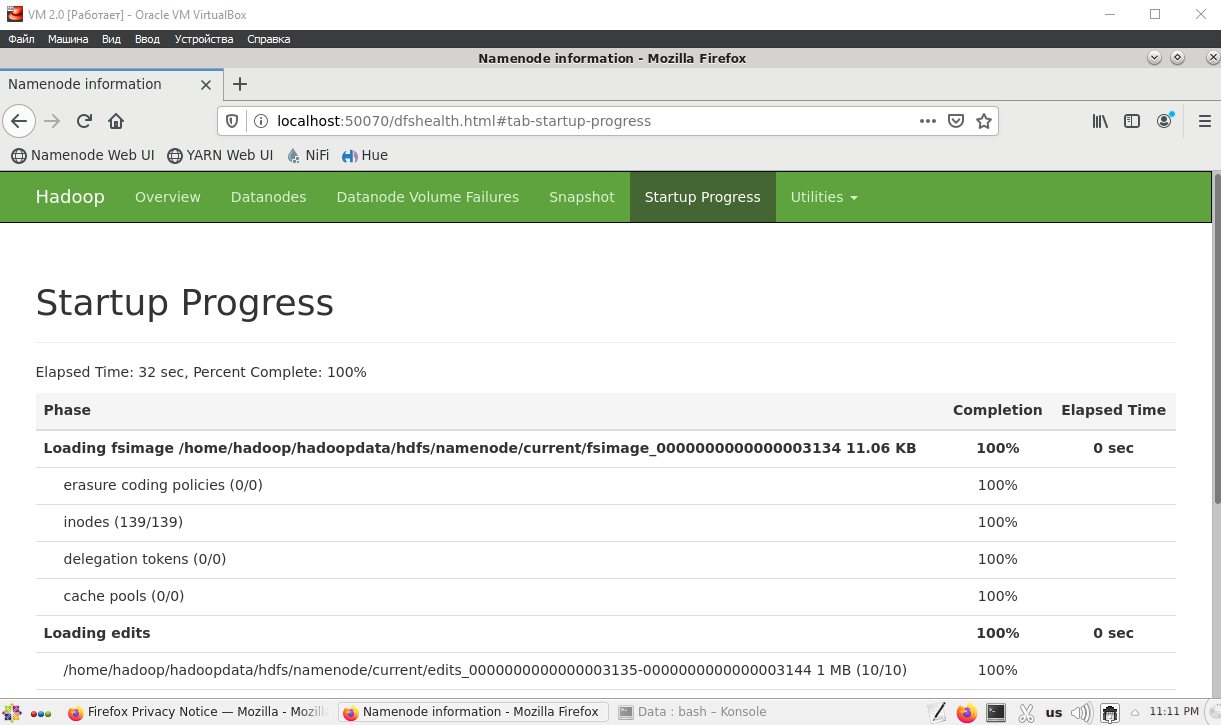


Рисунок 199 — Вкладка Startup Progress

В разделе Utilities («Утилиты») выберем файловую систему. Перейдем в каталог HDFS, куда мы ранее скопировали файл alice\_in\_wonderland.txt. на Рисунке 20.

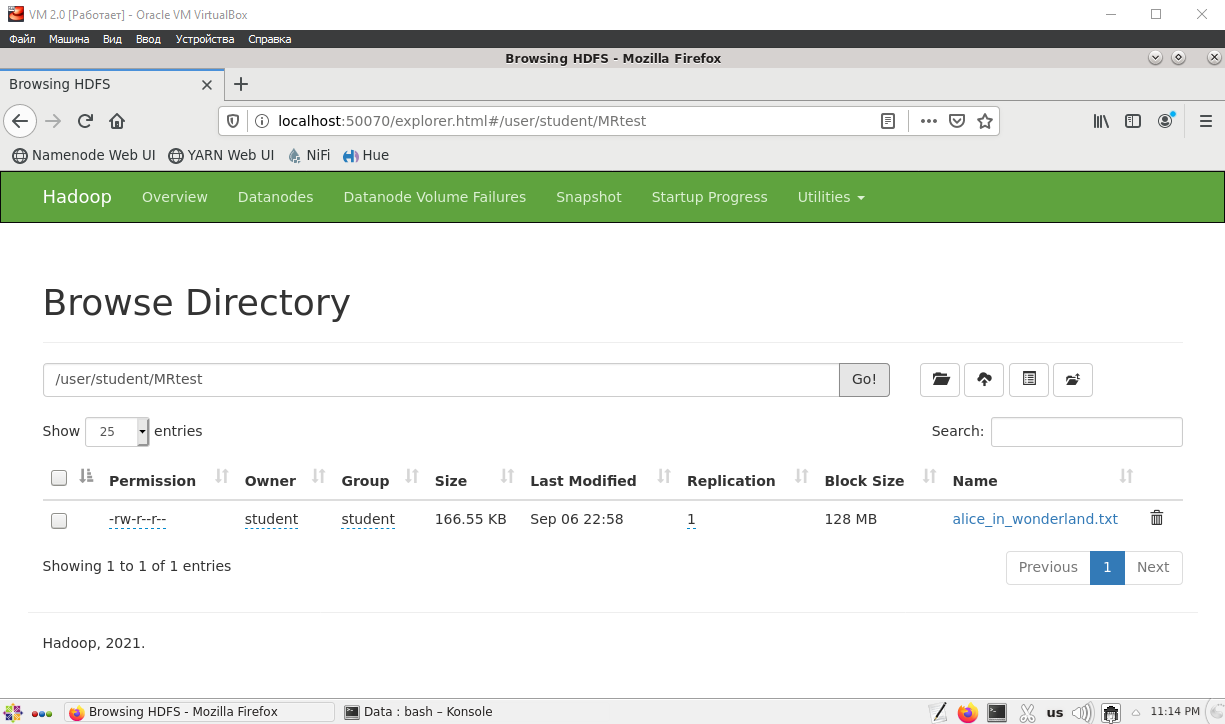


Рисунок 20 — Вкладка Browse Directory

Нажмем на файл, чтобы открыть всплывающее окно с информацией о блоке на Рисунке 21.

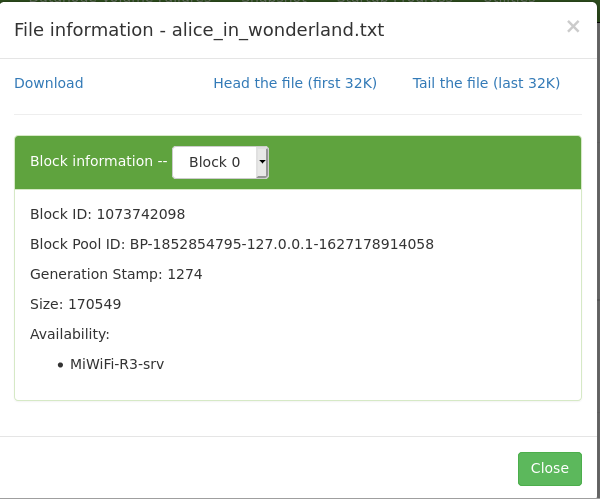


Рисунок 21 — Информация о файле alice\_in\_wonderland.txt

Откроем терминал и войдем в систему как пользователь hadoop. Используем команду Linux find для поиска Block Pool ID (идентификатор пула блоков) на Рисунке 22.

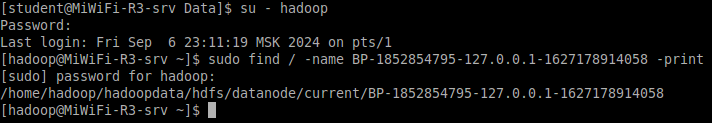


Рисунок 22 — Поиск Block Pool ID

Перейдем в каталог, где был найден Block Pool ID (идентификатор пула блоков), и изучим его на Рисунке 23.

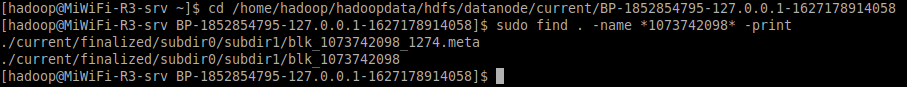


Рисунок 23 — Переход по Block Pool ID

Мы снова воспользуемся командой find для поиска блоков данных. На этот раз мы скажем find начать поиск с текущего каталога с точечной нотацией ( . ). Вы должны увидеть два найденных файла. Один из файлов будет содержать метаданные. Это бинарный файл, и вы не сможете увидеть его содержимое. Однако другой файл является текстовым, а фактическое содержимое, которое Hadoop сохранило для файла alice\_in\_wonderland.txt, сохраняется в HDFS. Используем команду less или cat для просмотра текстового файла на Рисунке 24.



Рисунок 24 — Просмотр текстового файла

### Работа с YARN/MapReduce

Мы будем использовать файл alice\_in\_wonderland.txt, который был сохранен в HDFS, для запуска программы MapReduce подсчета слов. Включим YARN historyserver. Откроем новый терминал как hadoop (используем команду su – hadoop ) и запустим historyserver на Рисунке 25.

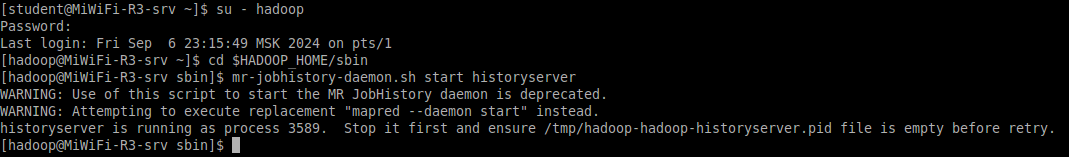


Рисунок 25 — Запуск historyserver

Запустим программу подсчета слов из jar примеров MapReduce. Откроем новый терминал как student. Перейдем в следующий каталог. Запустим файл hadoop-mapreduce-examples-3.3.1.jar с параметром подсчета слов. на Рисунке 26.

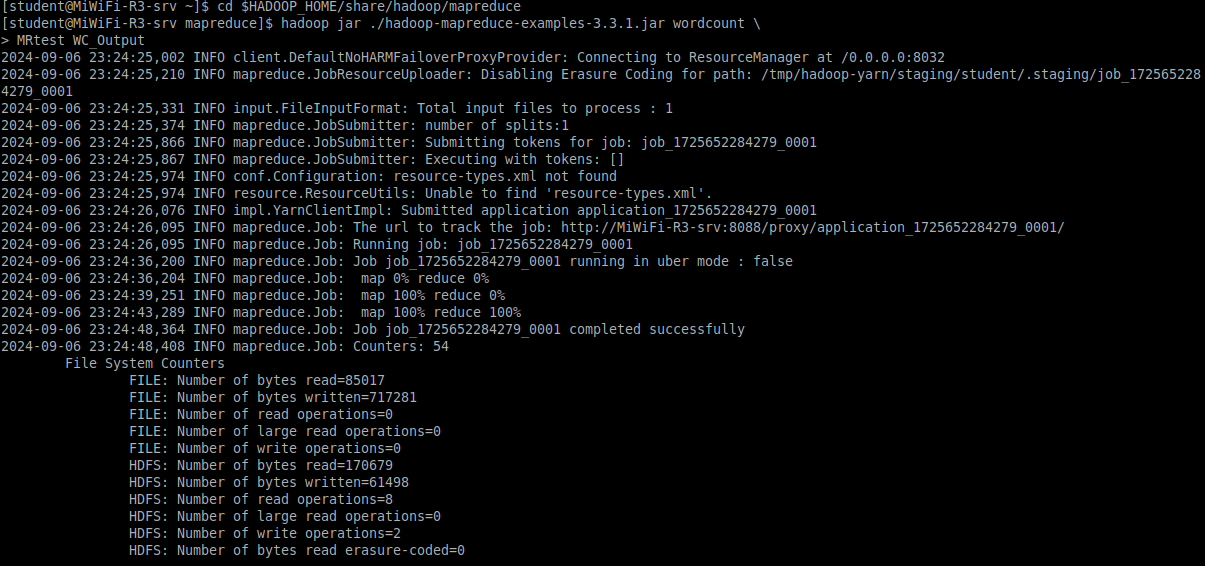


Рисунок 26 — Выполнение hadoop-mapreduce-examples-3.3.1.jar

Во время работы программы откроем Firefox и перейдем к веб-интерфейсу YARN по адресу http://localhost:8088. на Рисунке 27.

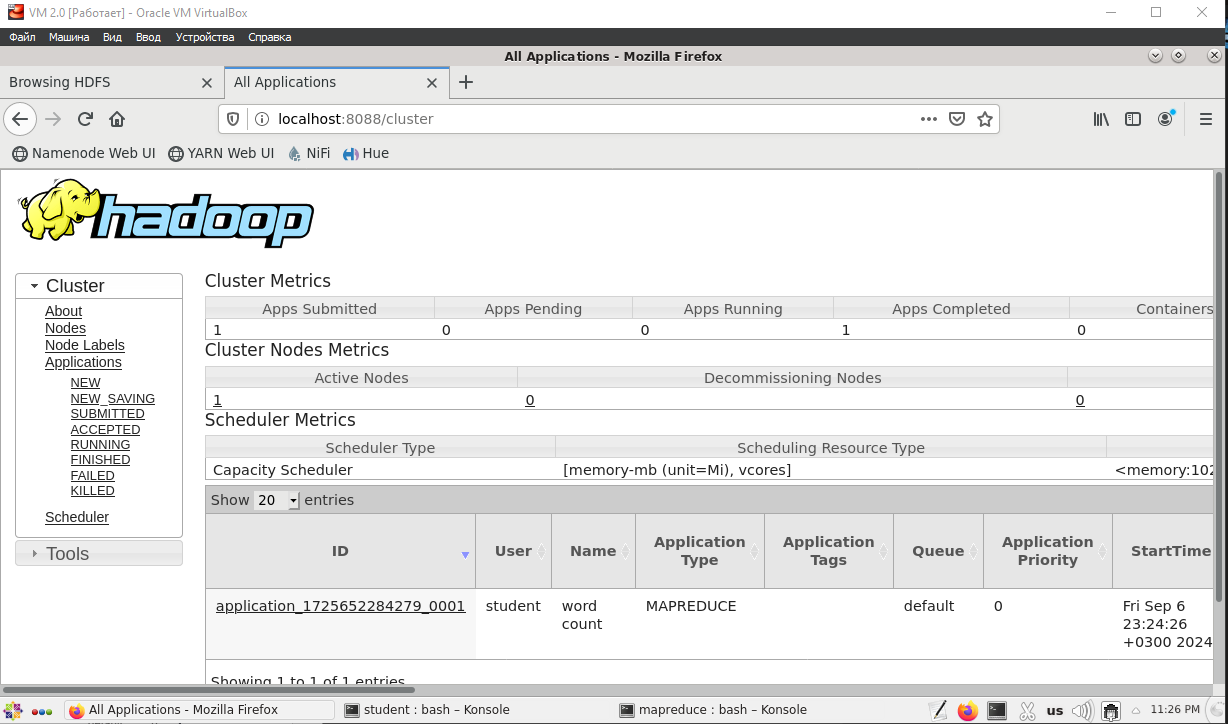


Рисунок 27 — Веб-интерфейс YARN

Изучаем Application Master на Рисунке 28.

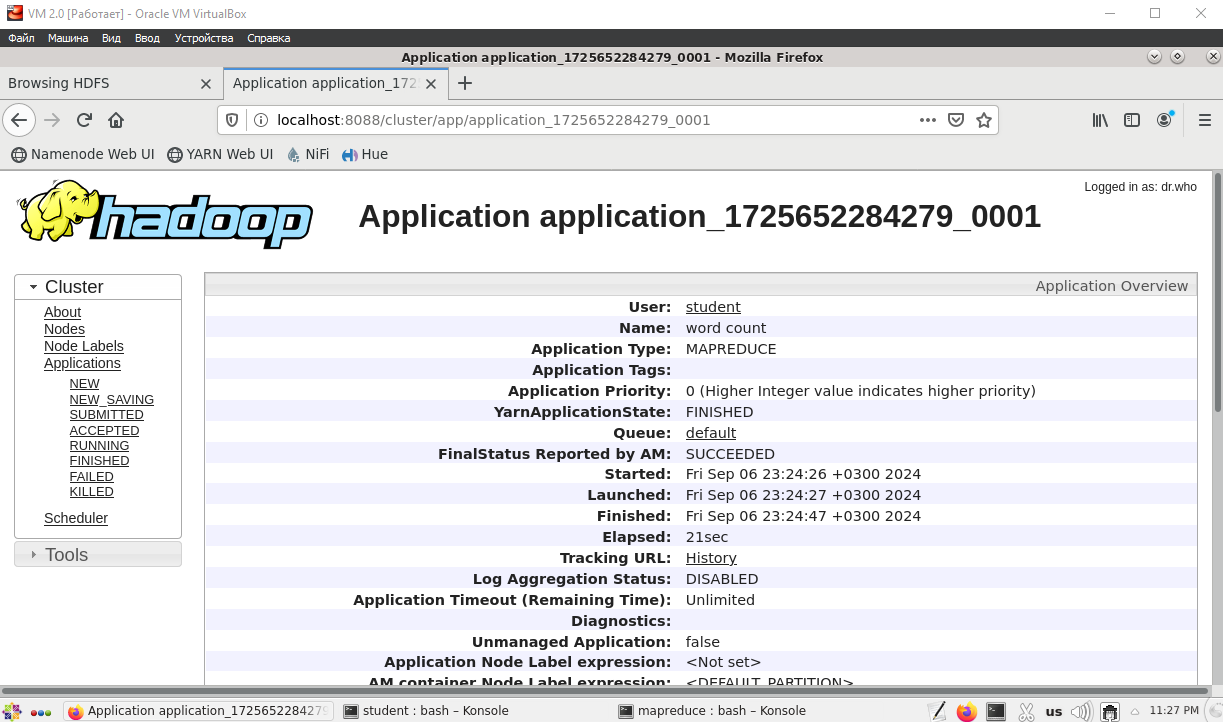


Рисунок 28 — Application Master

Результат выполнения hadoop-mapreduce-examples-3.3.1.jar представлен на Рисунке 29.

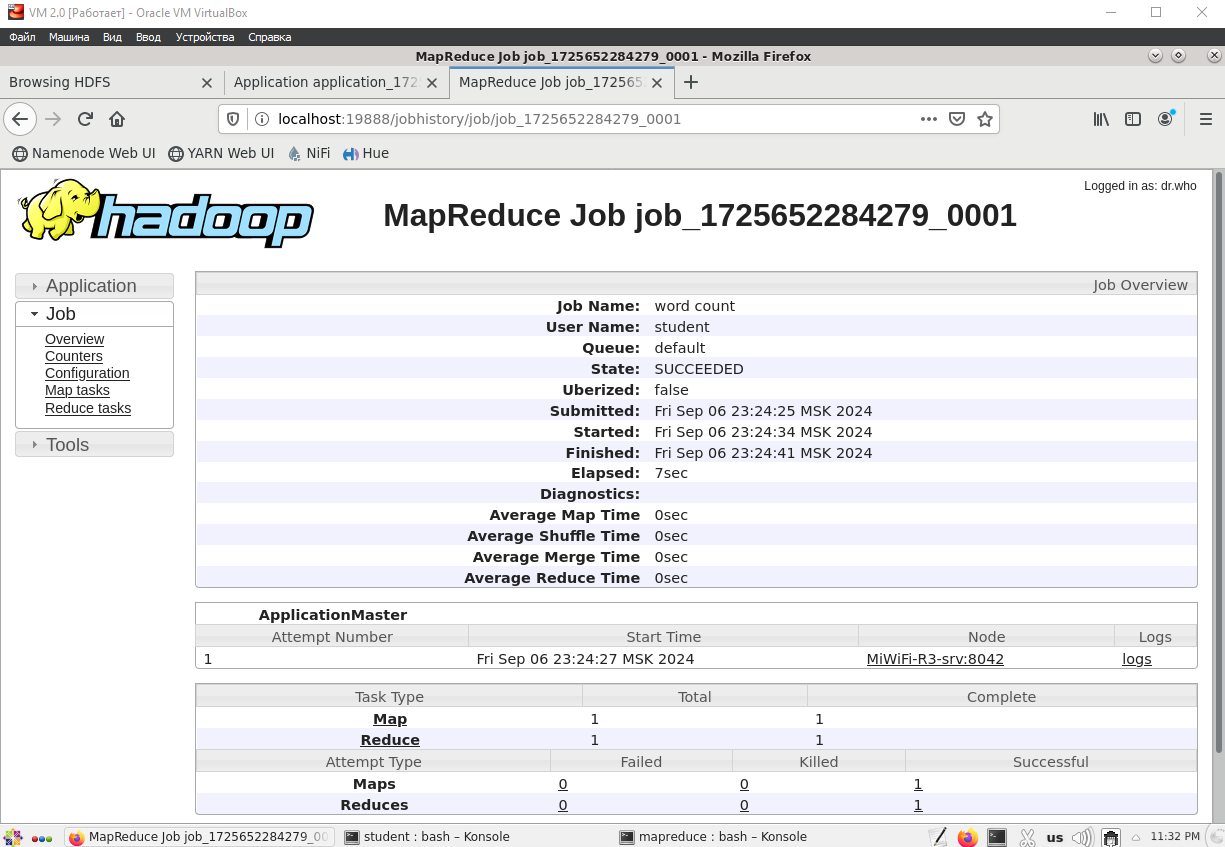


Рисунок 29 — Выполненное задание

Каждый узел показывает ресурсы контейнера, выделенные для задания на Рисунке 30.

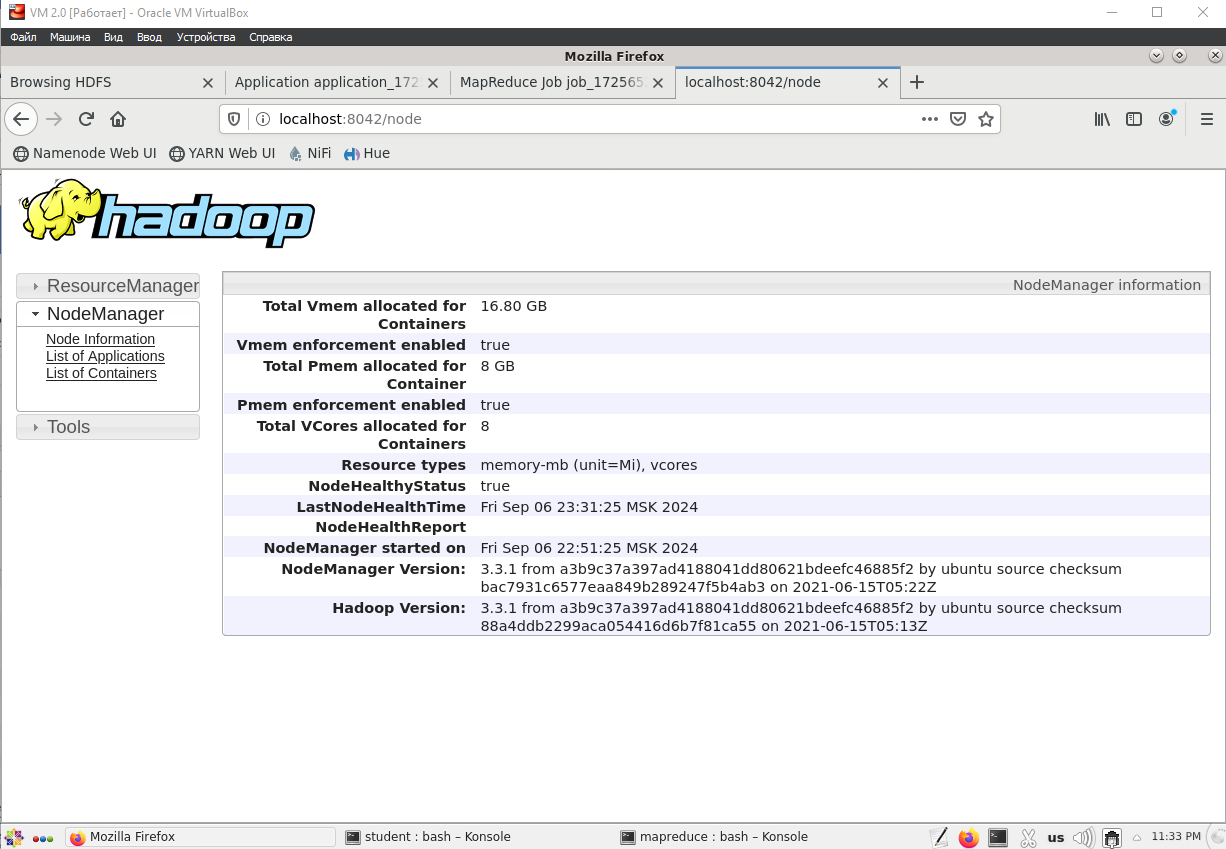


Рисунок 30 — Переход по Block Pool ID