Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики управления и технологий

Мареев Георгий Александрович БД-241м

**Практическая работа 1.1 Создание и управление базой данных на HDFS**

**Вариант 13**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Руководитель дисциплины:

Босенко Т.М., доцент департамента

информатики, управления и технологий,

кандидат технических наук

Москва

2025

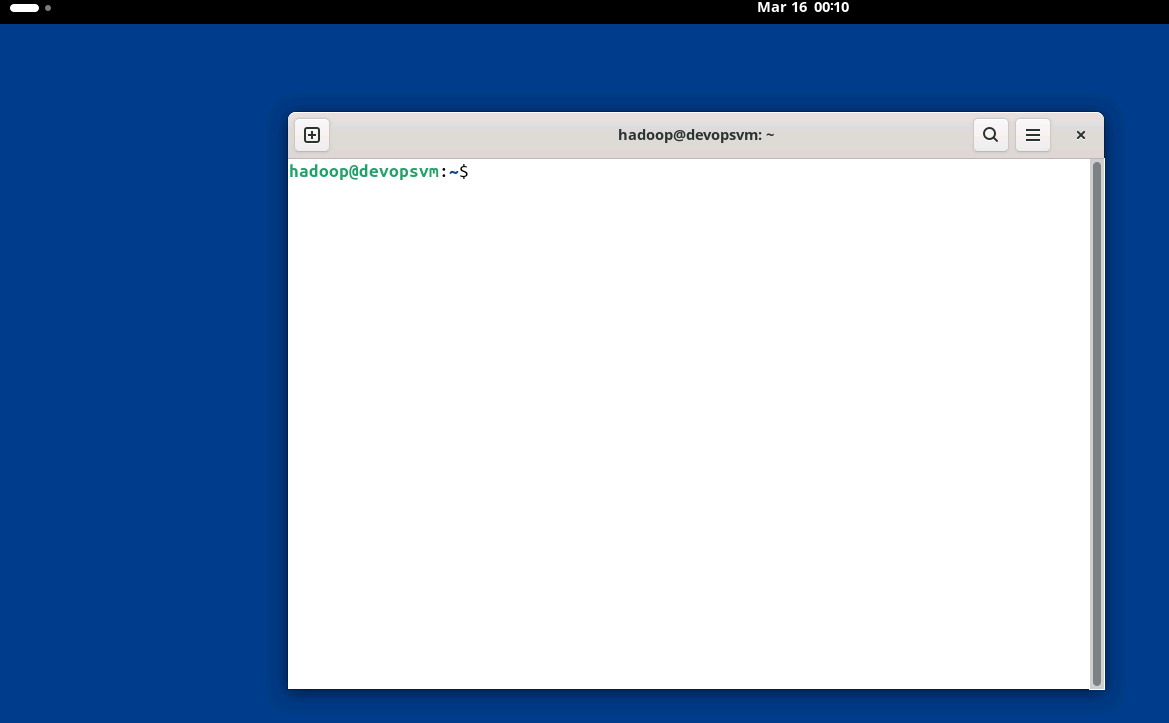
**Введение**

*Цель*: изучить основные операции и функциональные возможности системы, что позволит понять принципы работы с данными и распределенными вычислениями.

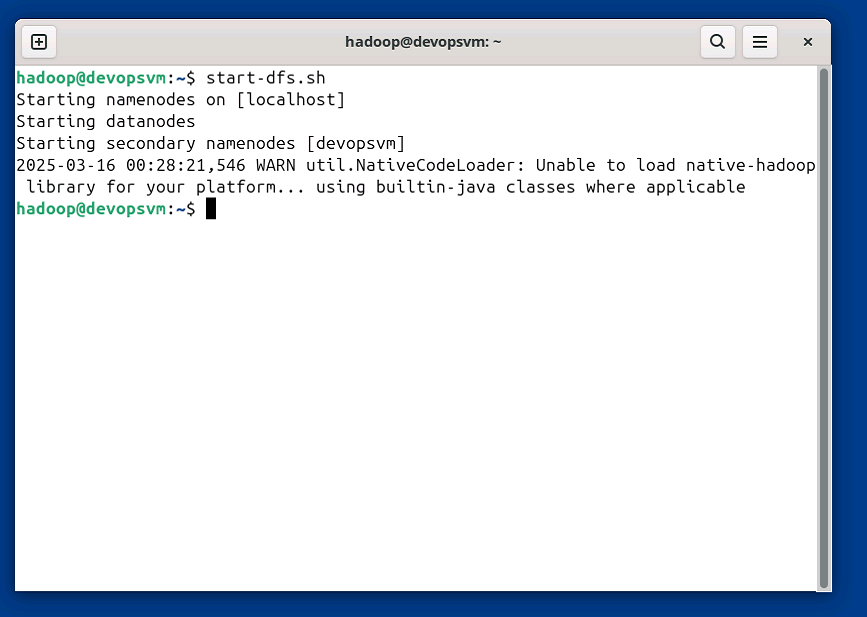
**Основная часть**

Шаг 1. Запуск Hadoop

Подключение в виртуальной машине через пользователя Hadoop

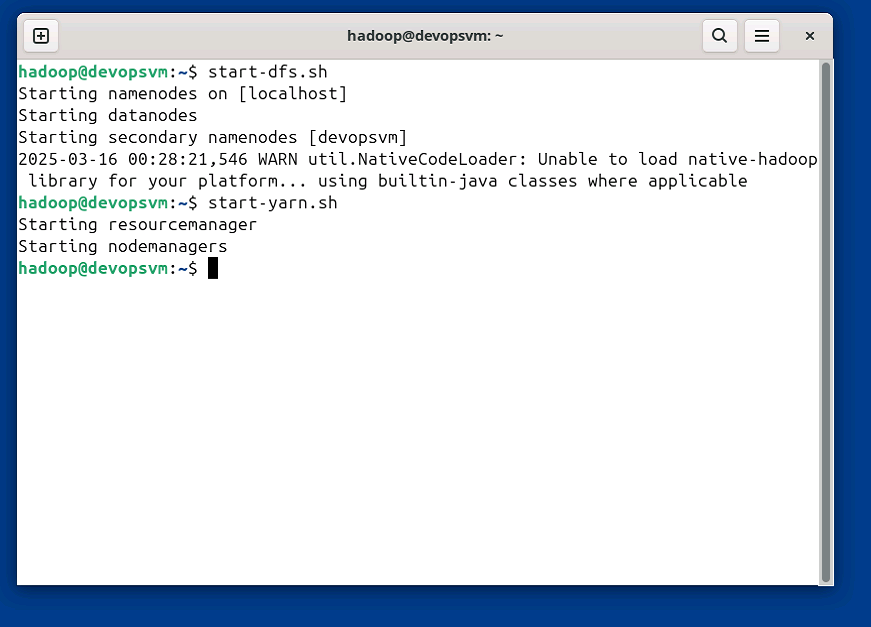


Запуск HDFS



start-dfs.sh

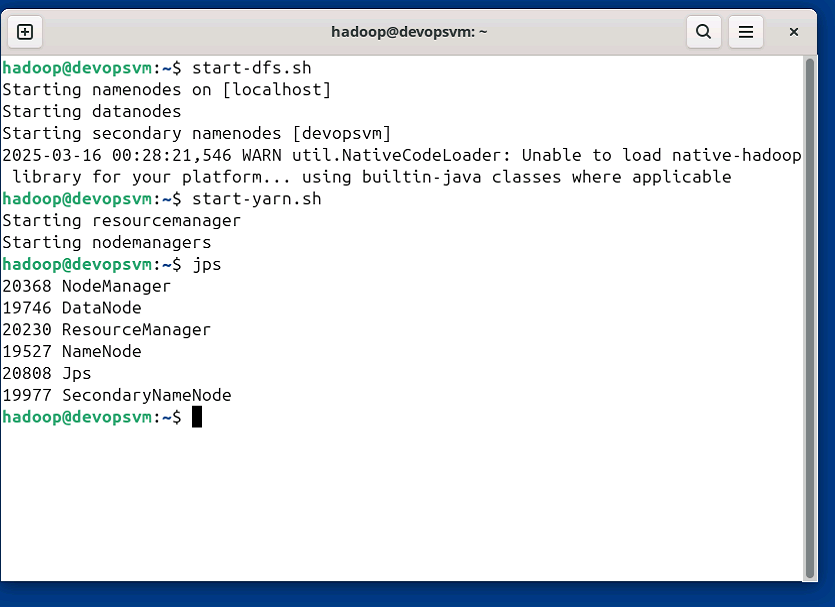
Запуск YARN



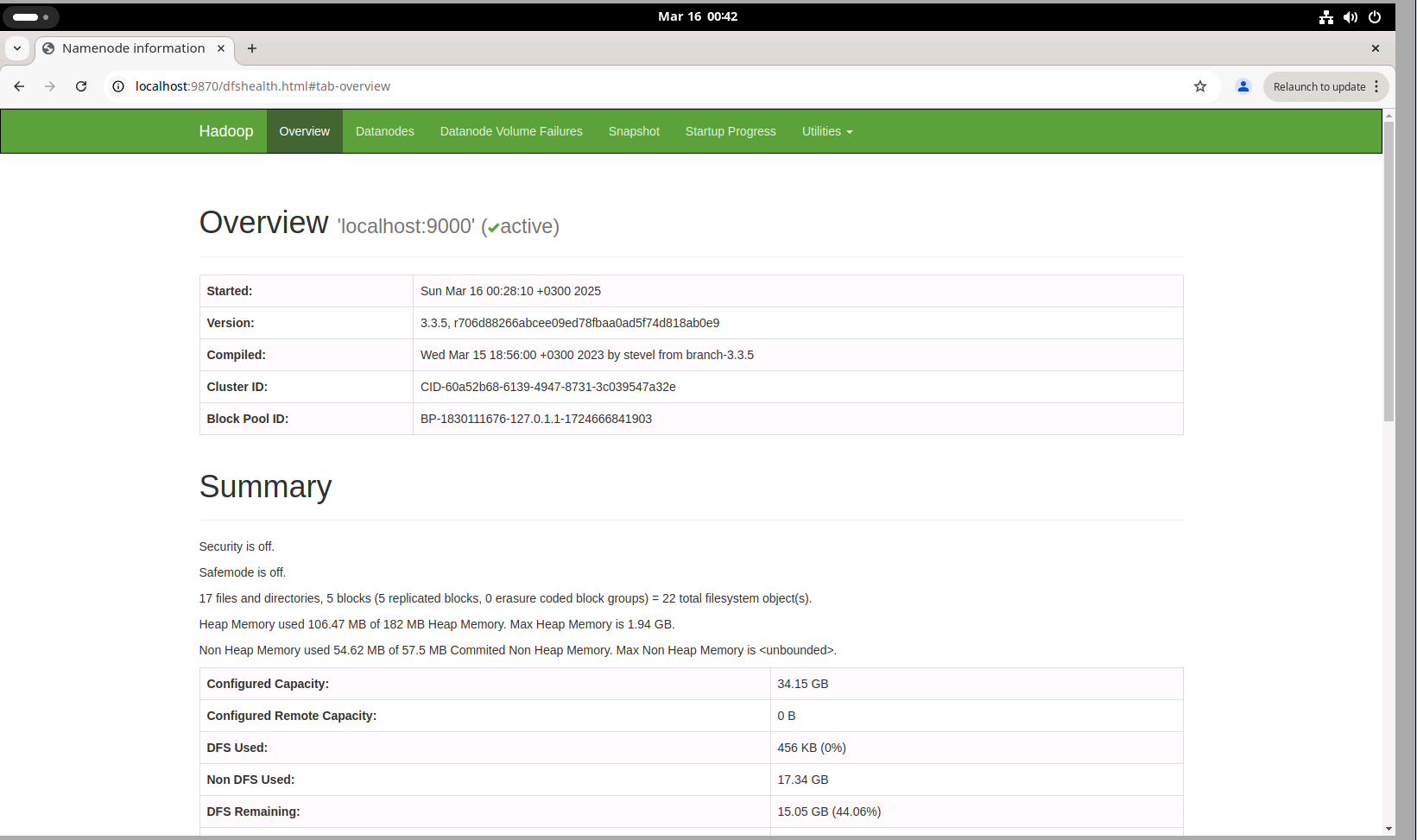
start-yarn.sh

Шаг 2. Проверка запущенных служб

jps

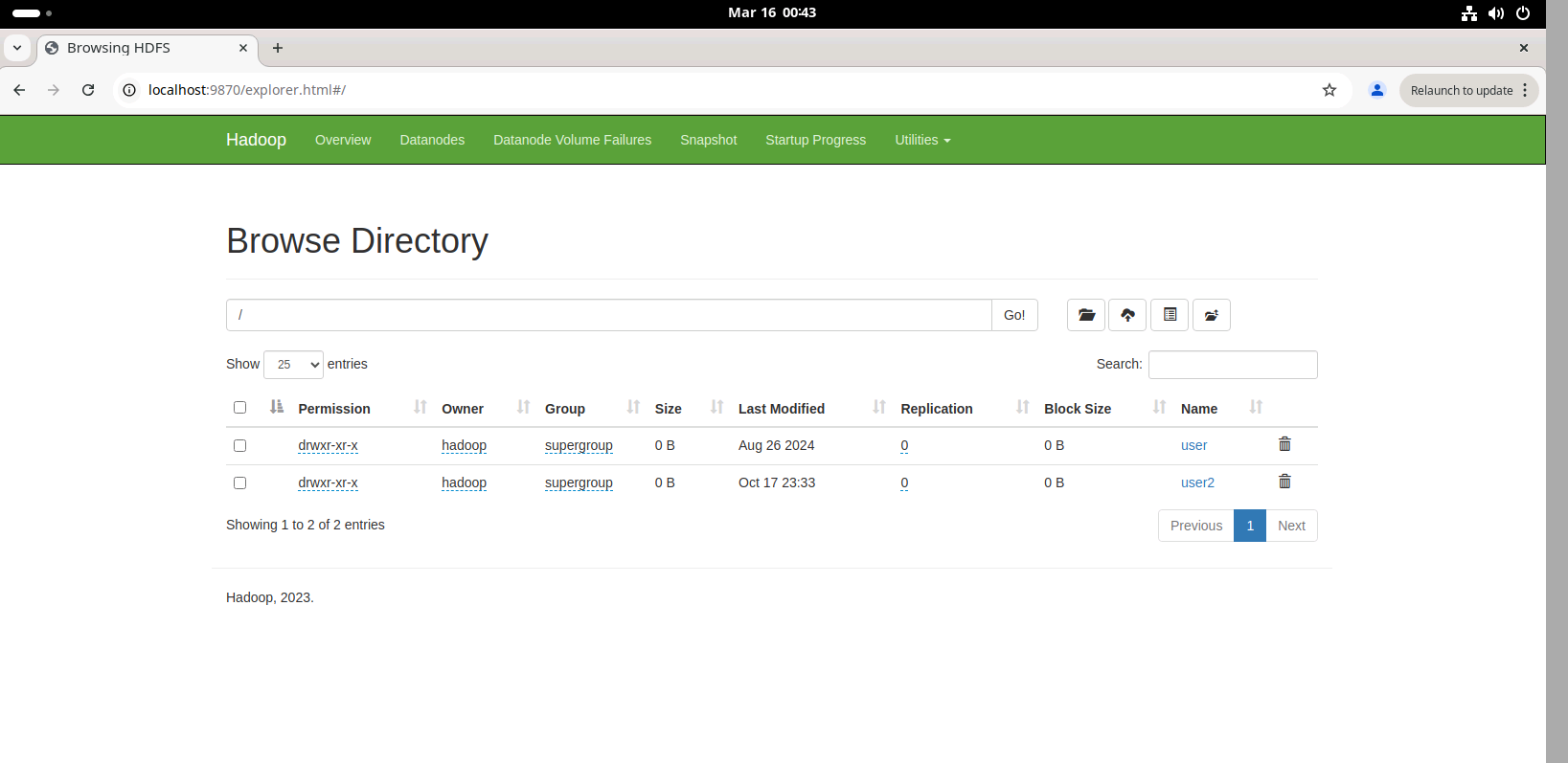


Проверка доступности системы

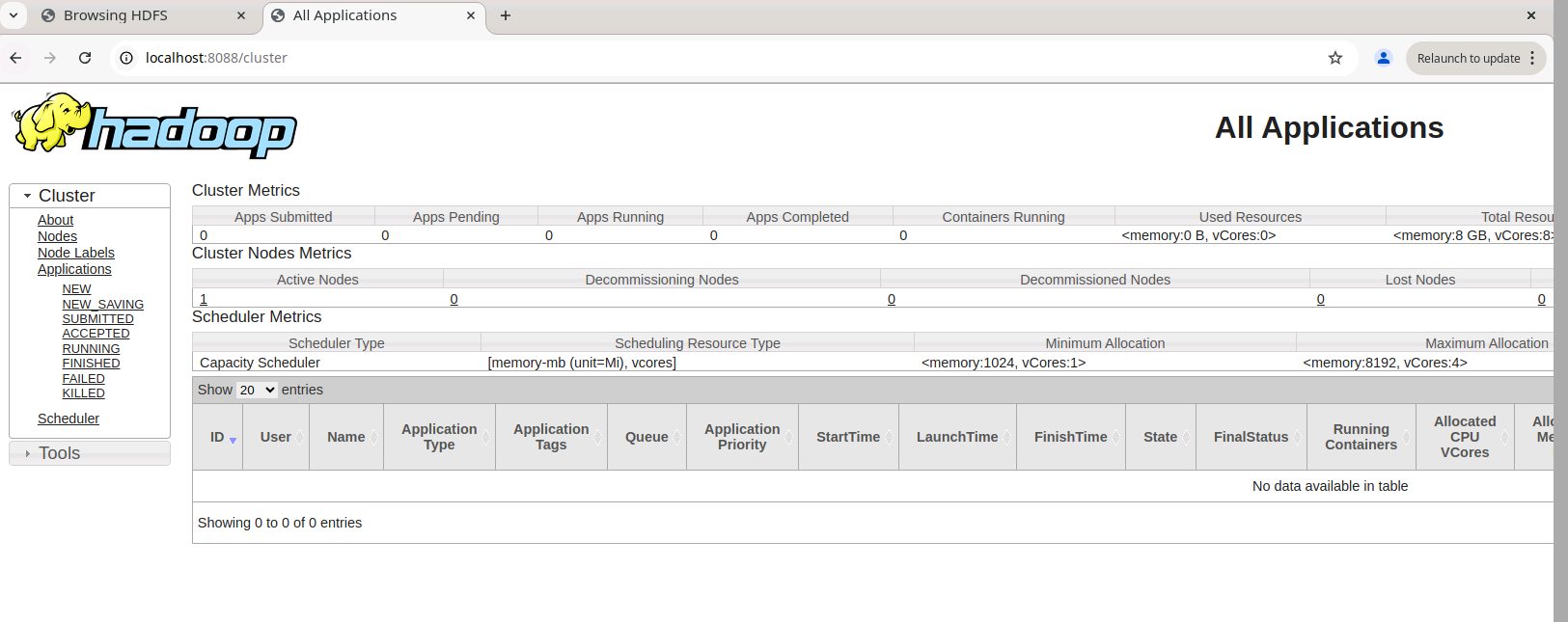


localhost:9870

Каталоги файловой системы



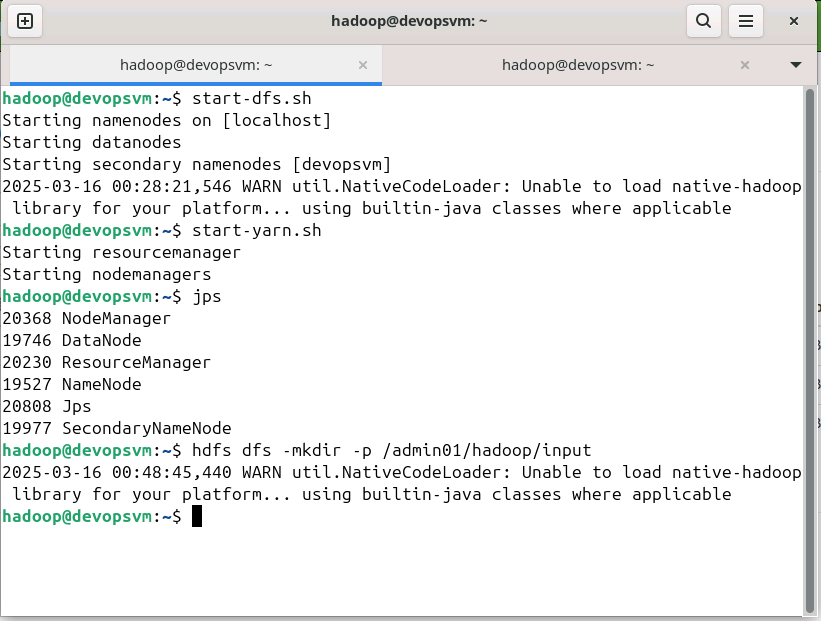
Проверка YARN

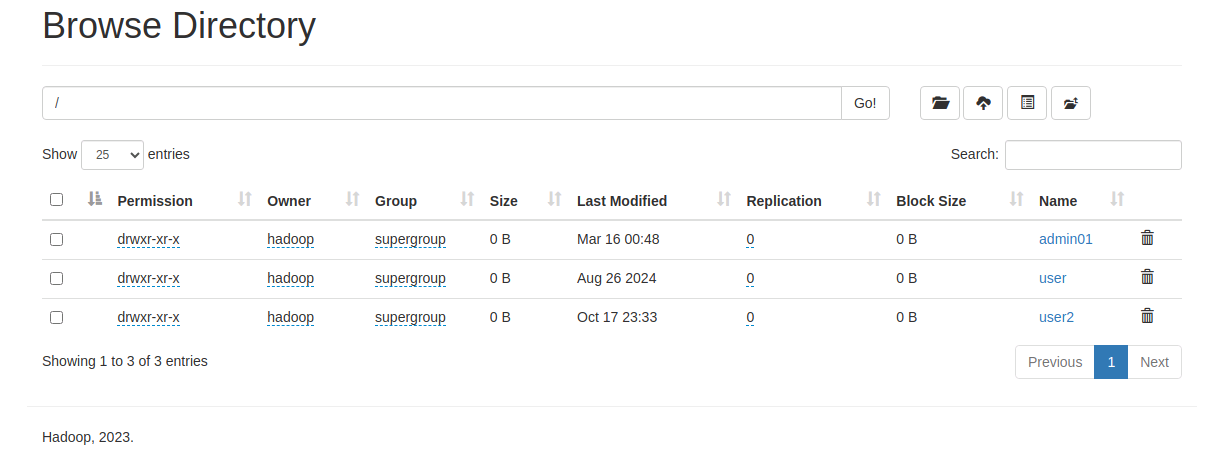


Шаг 3. Подготовка рабочего пространства

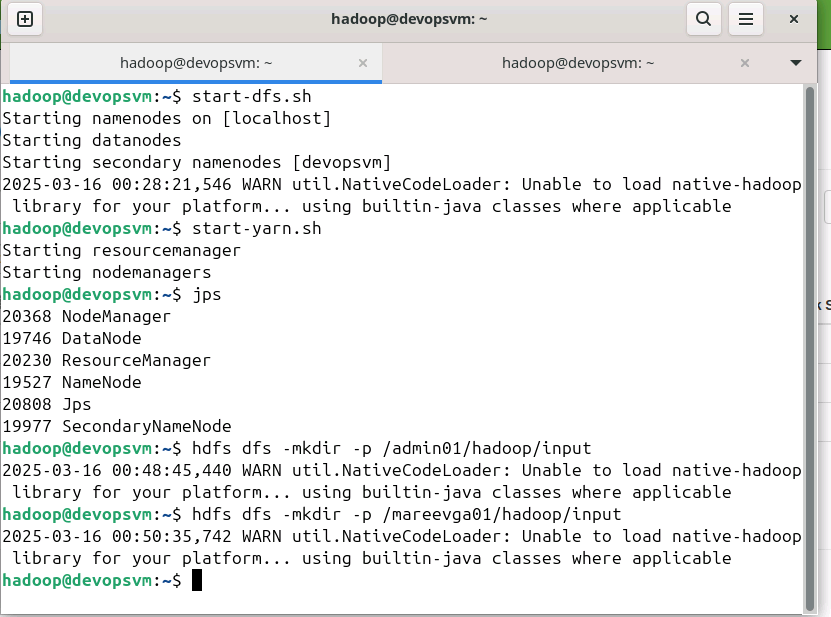
Создание директории в HDFS

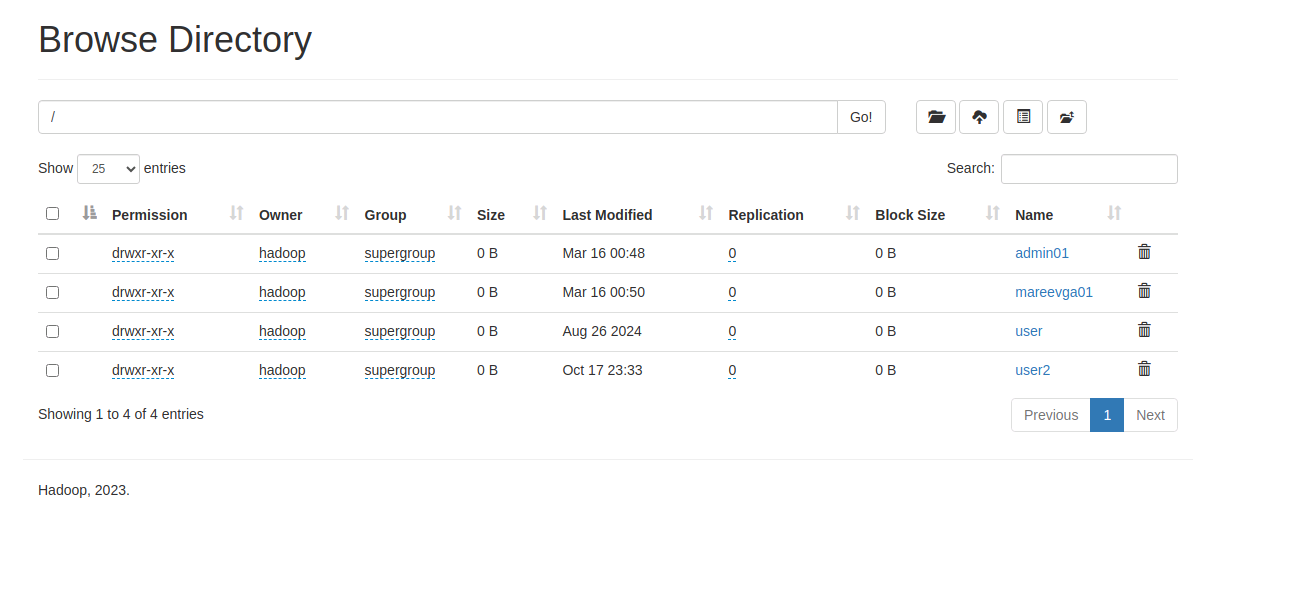
hdfs dfs -mkdir -p /admin01/hadoop/input





hdfs dfs -mkdir -p /**mareevga01**/hadoop/input



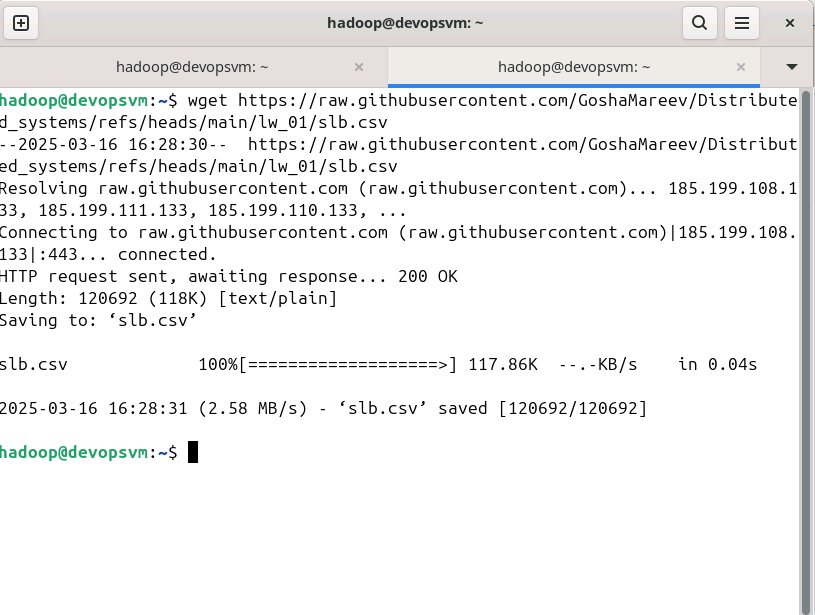


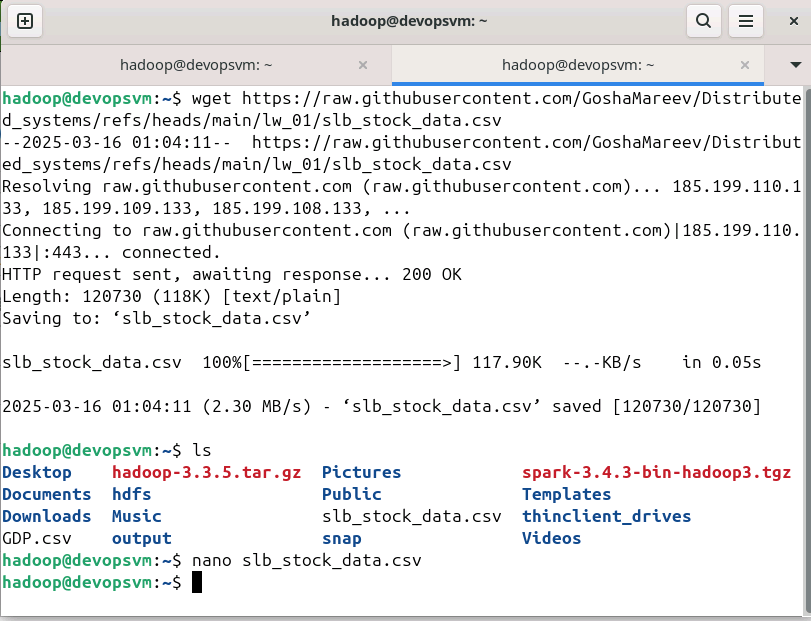
Шаг 4. Загрузка и подготовка данных

Скачивание файла с данными

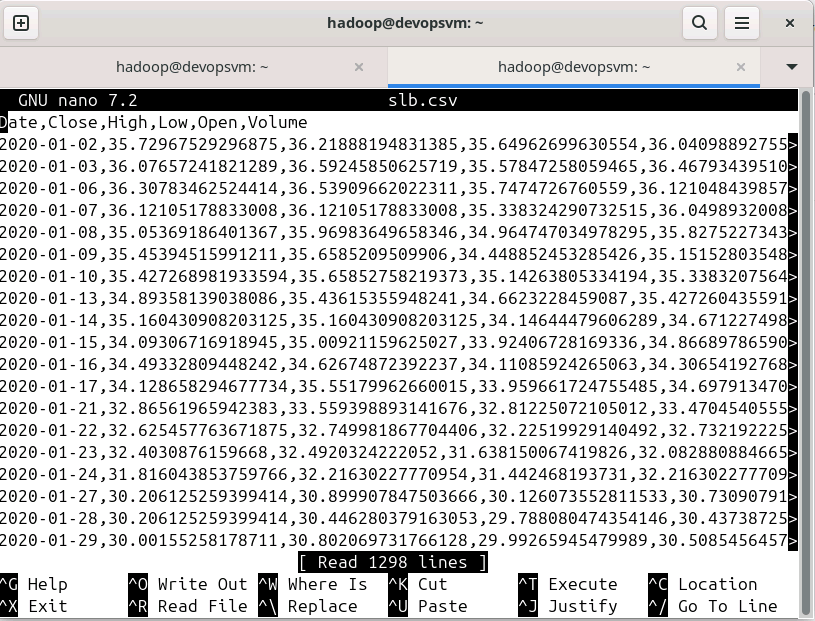
\*файл с данными был немного преобразован: сделан заголовок и изменено имя файла для простоты (все изменения отражены в репозитории)  
https://github.com/GoshaMareev/Distributed\_systems/tree/main/lw\_01

wget <https://raw.githubusercontent.com/GoshaMareev/Distributed_systems/refs/heads/main/lw_01/slb.csv>



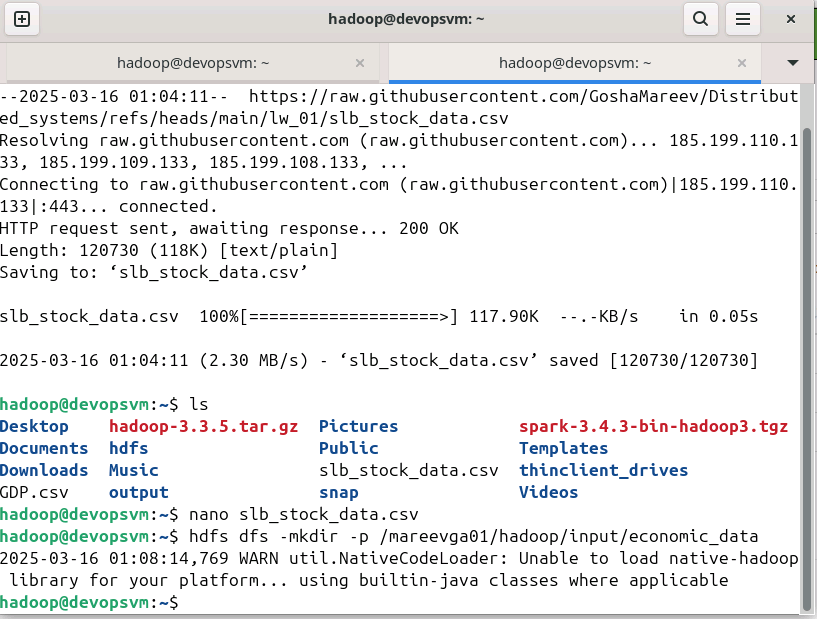


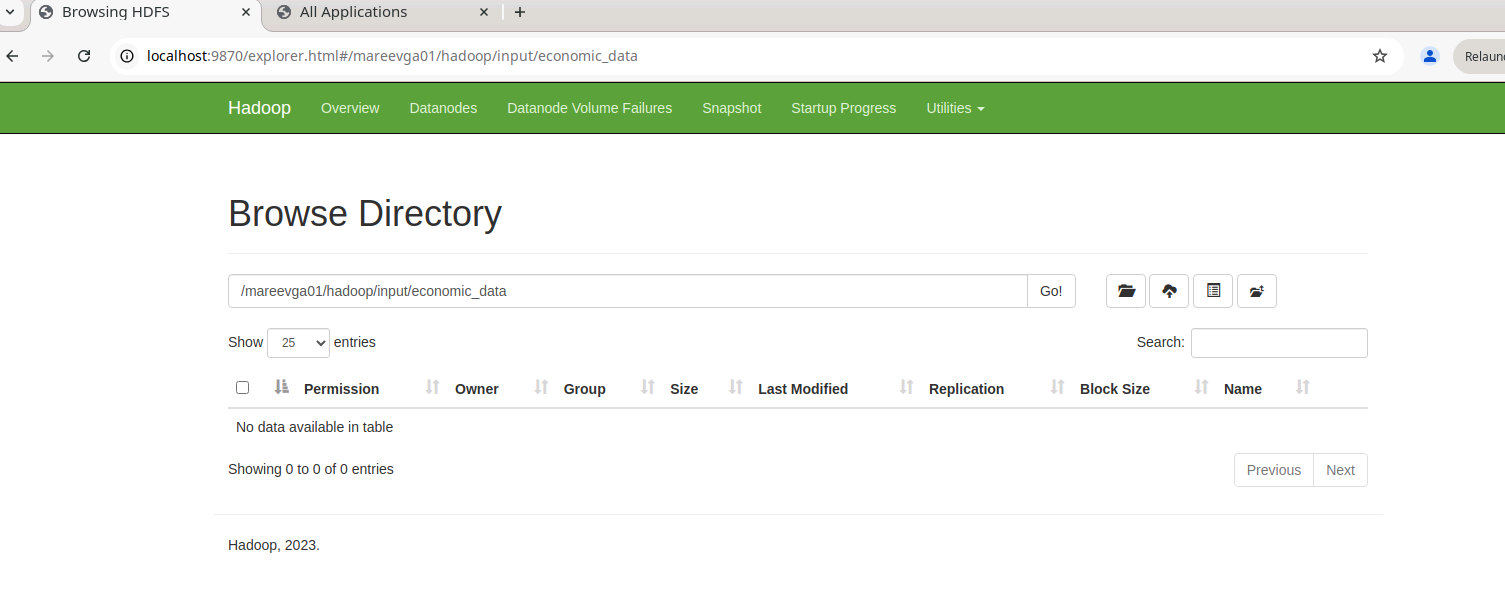
Nano slb.csv



Создание директории для экономических данных

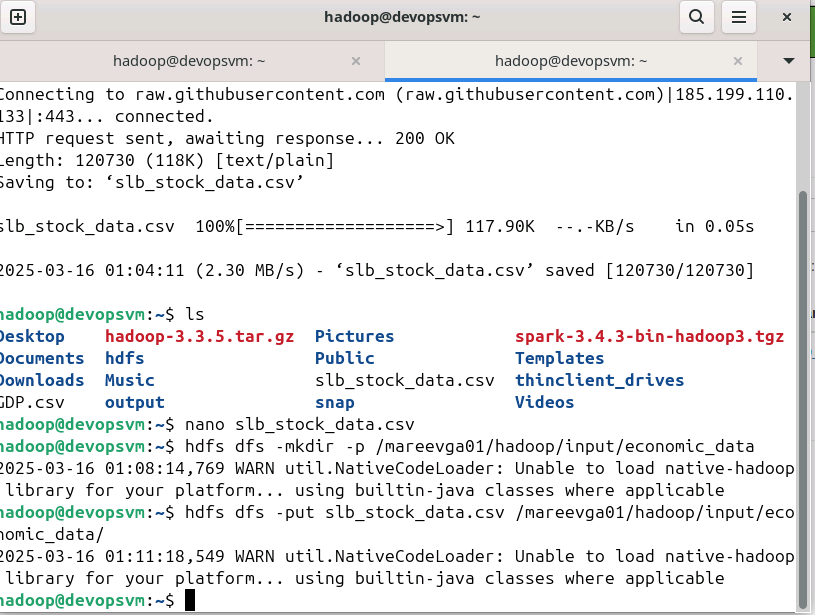
hdfs dfs -mkdir -p /mareevga01/hadoop/input/economic\_data

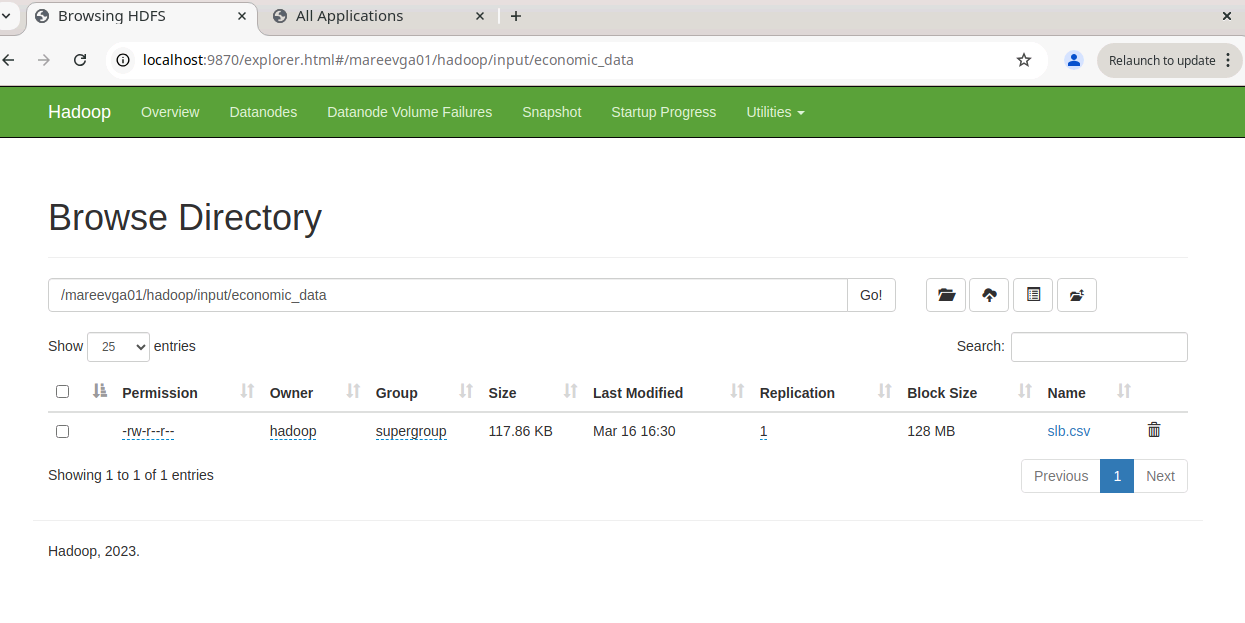




Загрузка данных в HDFS

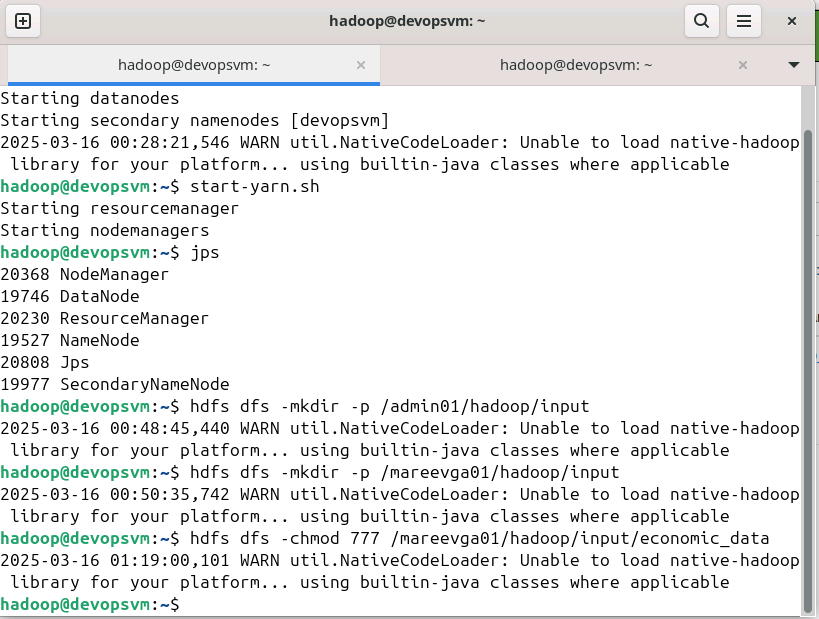
hdfs dfs -put slb.csv /mareevga01/hadoop/input/economic\_data/





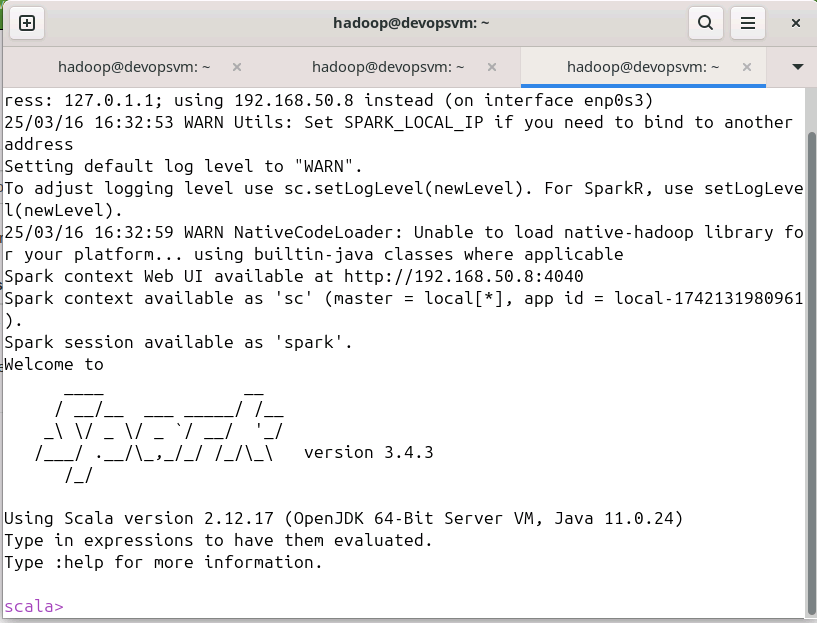
Установка прав доступа

hdfs dfs -chmod 777 /mareevga01/hadoop/input/economic\_data



Шаг 5. Обработка данных с помощью Spark

Spark-shell



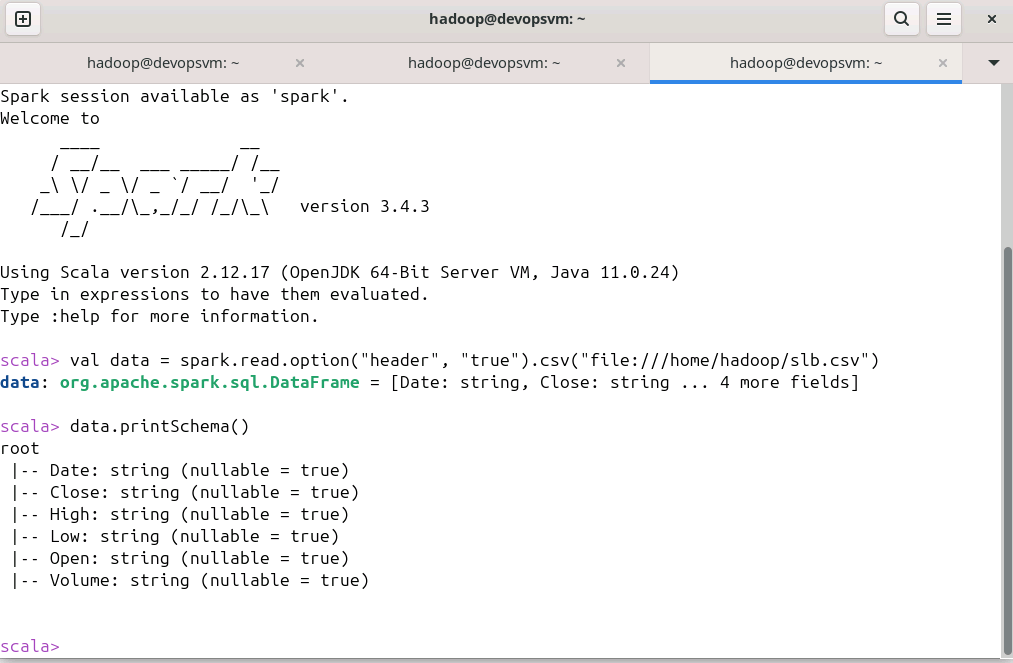
Spark Shell:

Загрузка данных из HDFS с правильным URI

val data = spark.read.option("header", "true").csv("file:///home/hadoop/slb.csv")

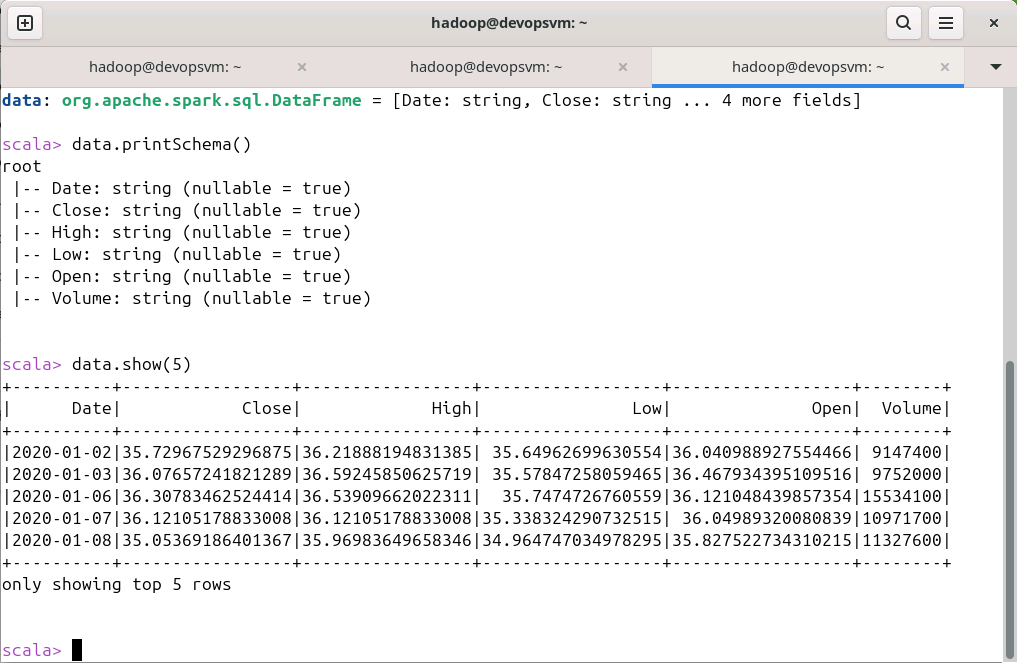
Проверка схемы данных

data.printSchema()



Первые 5 строк

Data.show(5)



Все столбцы имеют строковый (string) формат, для вычислений нужно будет преобразовать:

Data – в формат даты

Close, High, Low, Open, Volume в числовой формат

Вычисление среднего значения

Фильтрация данных за последние 3 года

Расчет медианной цены закрытия

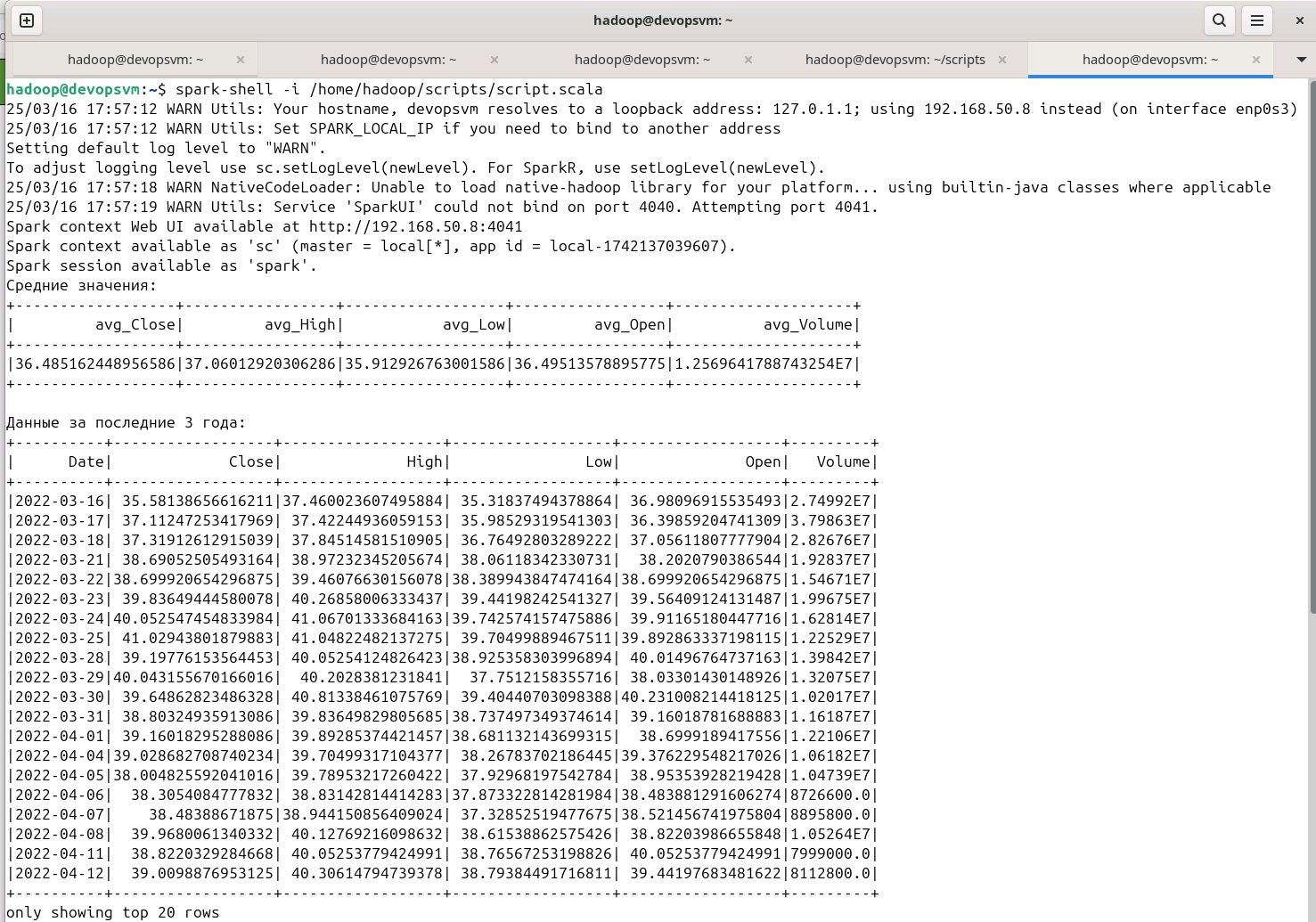
Группировка по кварталам

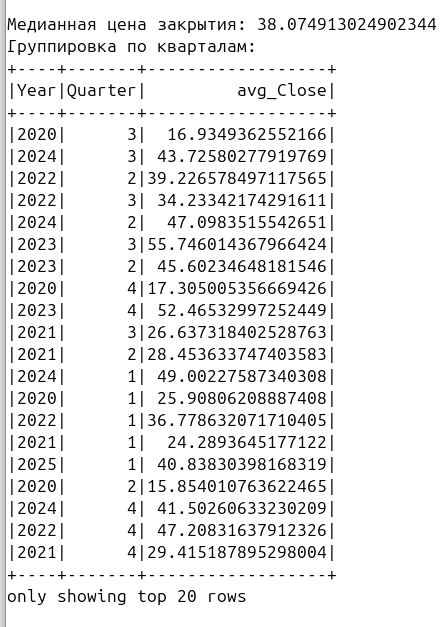
Сохранение результата в CSV файл

Для удобства была создана директория scripts/ и файл script.scala

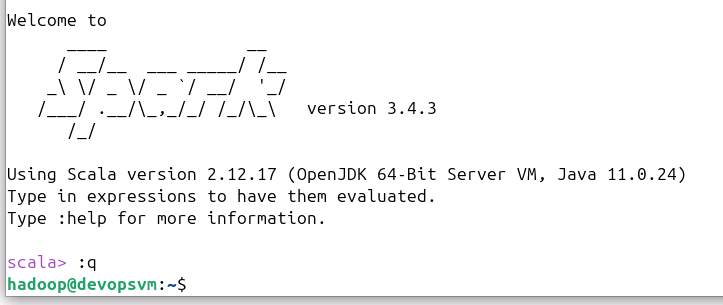
Создание директории и файла script.scala с кодом

Результат выполнения:





Выход из Spark-shell:



Шаг 6. Работа с результатами

сd /home/hadoop

Переход в директорию с результатами

**Проверка результатов**

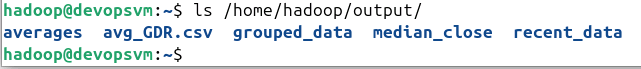
После выполнения скрипта результаты будут сохранены в указанных директориях:

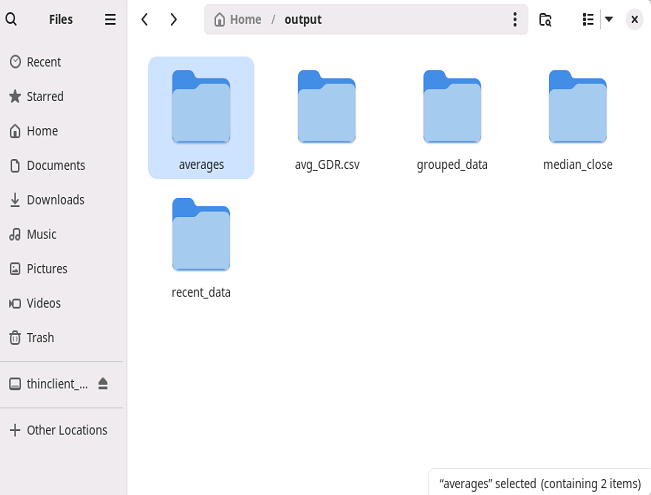
file:///home/hadoop/output/averages

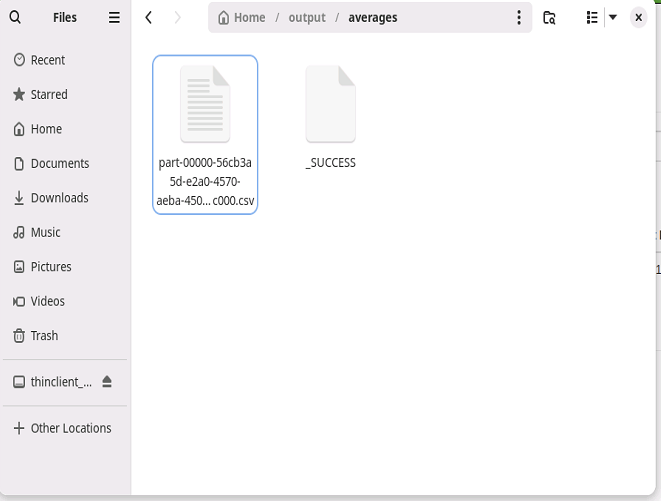
file:///home/hadoop/output/recent\_data

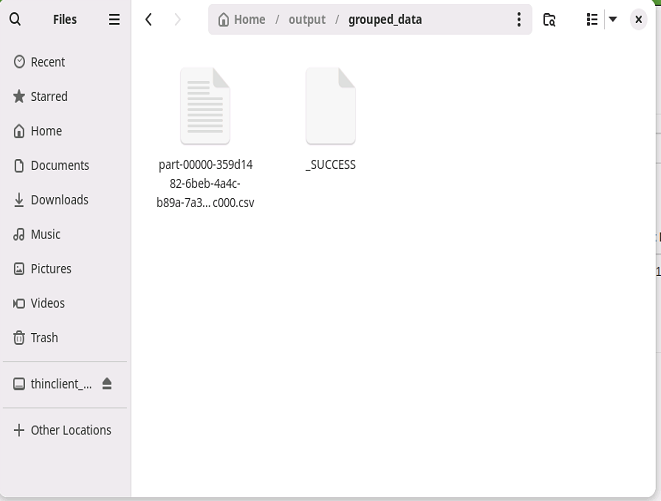
file:///home/hadoop/output/median\_close

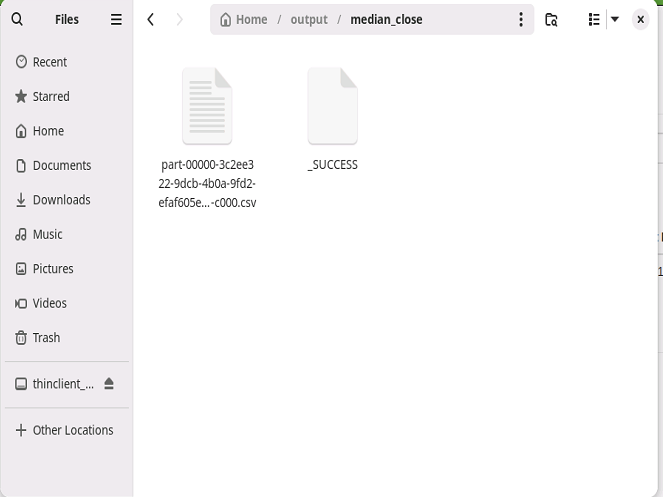
<file:///home/hadoop/output/grouped_data>

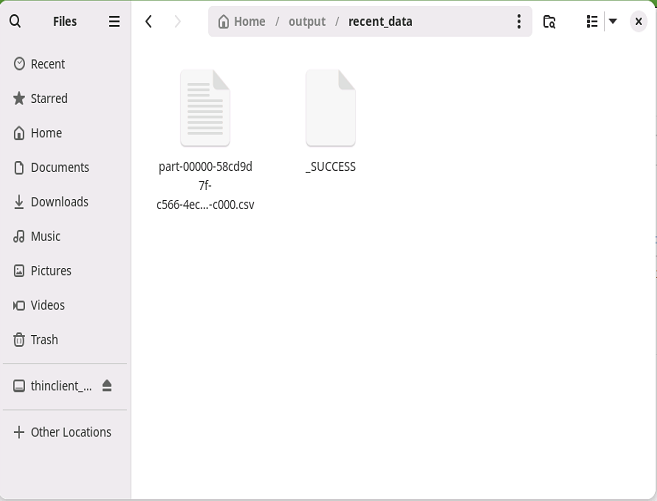




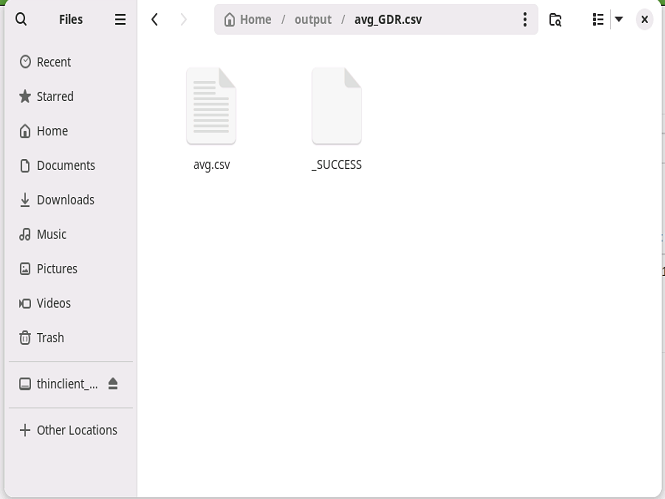








Файлы из примера



Переименование файлов с результатами

Averages:

mv part-00000-\*.csv avg.csv

grouped\_data:

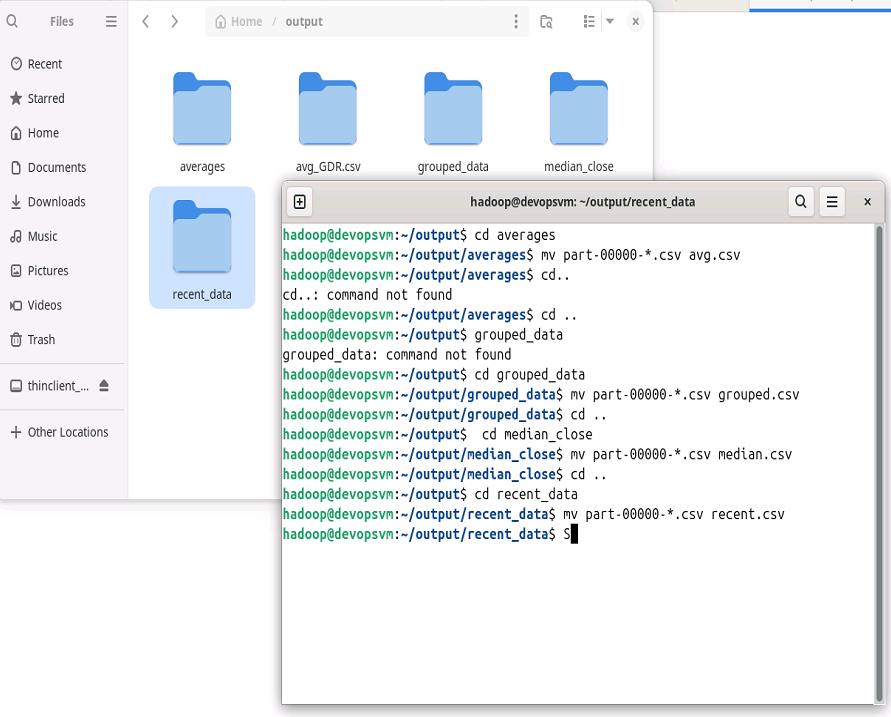
mv part-00000-\*.csv grouped.csv

median:

mv part-00000-\*.csv median.csv

recent:

mv part-00000-\*.csv recent.csv



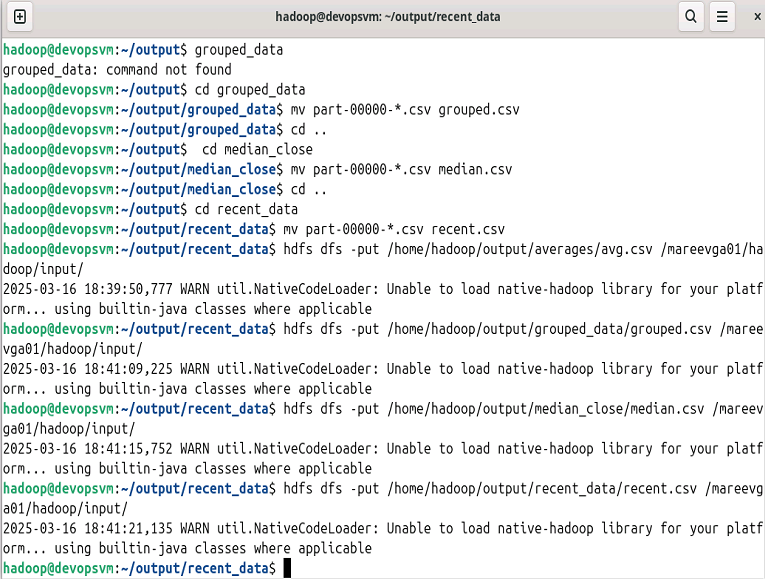
Загрузка результатов в HDFS

hdfs dfs -put /home/hadoop/output/averages/avg.csv /mareevga01/hadoop/input/

hdfs dfs -put /home/hadoop/output/grouped\_data/grouped.csv /mareevga01/hadoop/input/

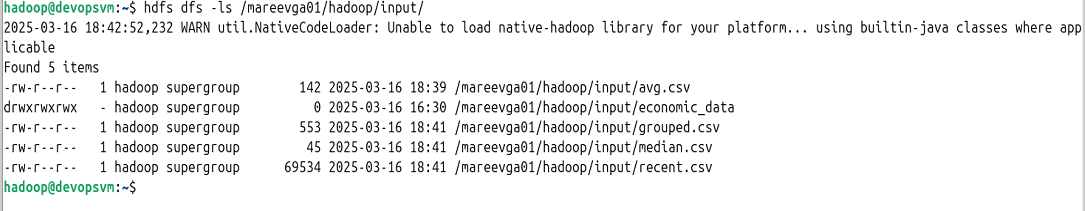
hdfs dfs -put /home/hadoop/output/median\_close/median.csv /mareevga01/hadoop/input/

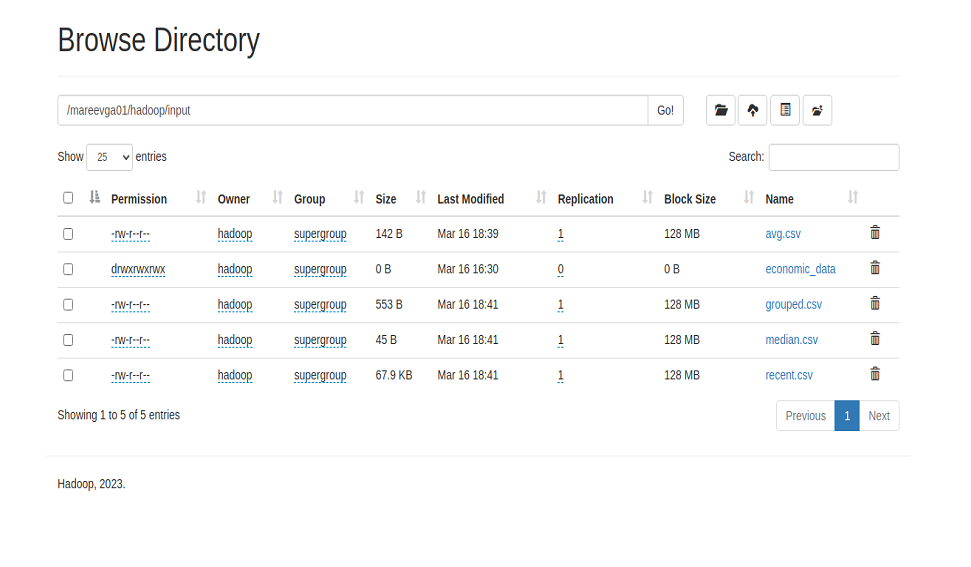
hdfs dfs -put /home/hadoop/output/recent\_data/recent.csv /mareevga01/hadoop/input/



Проверка загрузки

hdfs dfs -ls /mareevga01/hadoop/input/

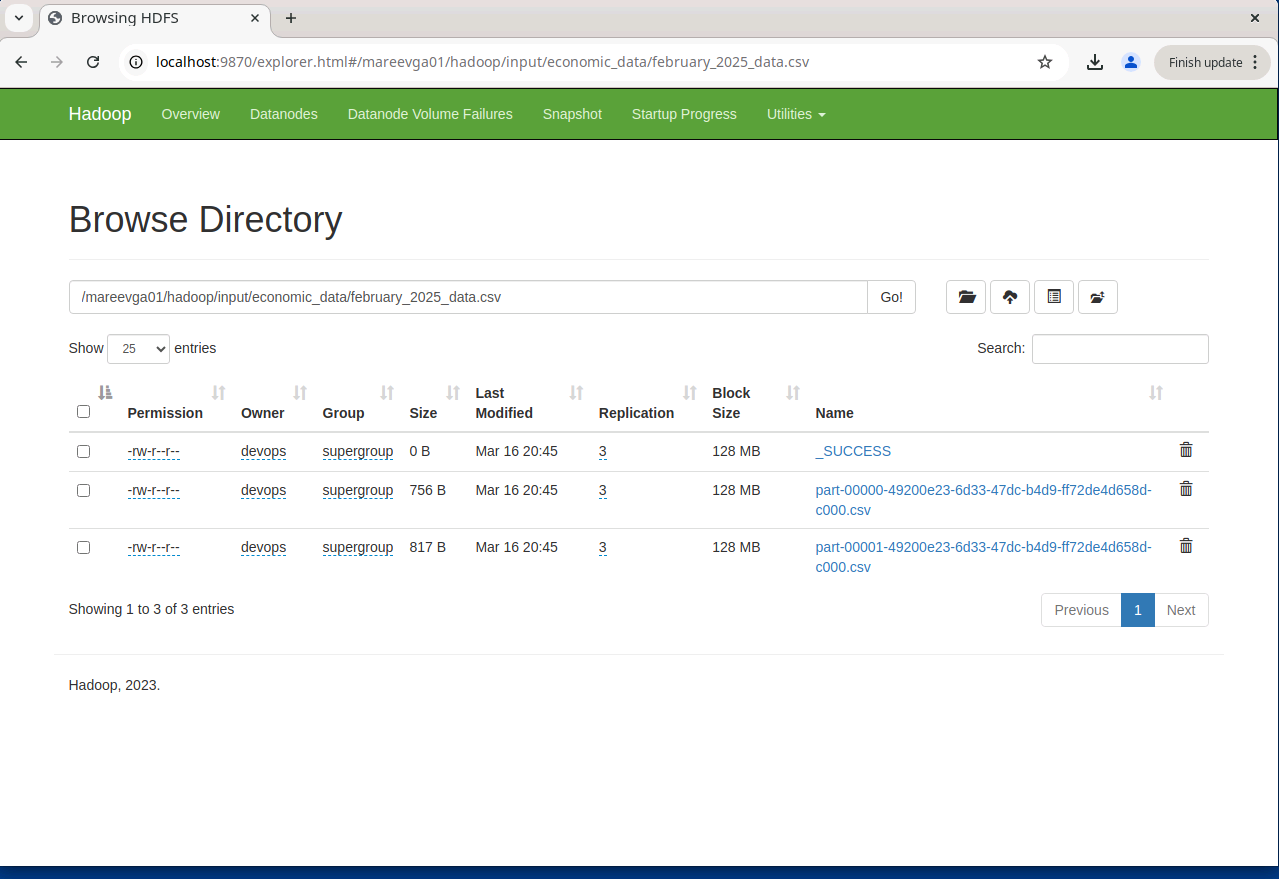




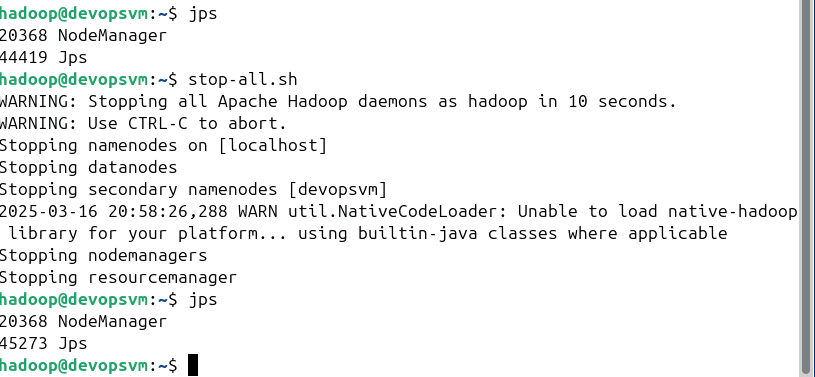
Работы в Jupyter(блокнот work\_with\_data\_2025 приложен):

****

Были выгружены данные за февраль 2025 г.

Результат записи 

Шаг 7. Завершение работы с Hadoop



**Индивидуальное задание. Вариант 13**

Поиск медианных значений и квартальная агрегация

* фильтрация данных за последние 3 года,
* расчет медианной цены закрытия,
* группировка по кварталам

*Исторические данные по акциям Мечела (MLTR)*

*https://www.kaggle.com/datasets/svtxvt/moscow-exchange-daily-pricedata), YahooFinance (*[*https://finance.yahoo.com/quote/MTL/history*](https://finance.yahoo.com/quote/MTL/history)*)*

Так как ссылки не работали, были скачаны данные по акциям компании Schlumberger Limited (SLB) с YahooFinance за период(с "2020-01-01" по "2025-03-01")

import yfinance as yf  
  
ticker = "SLB"  
  
data = yf.download(ticker, start="2020-01-01", end="2025-03-01")  
  
data.to\_csv("slb\_stock\_data.csv")  
  
  
print(data.head())

Результаты вычислений находятся в репозитории в /results:

avg.csv, grouped.csv, median.csv, recent.csv

**Заключение**

В результате изучения основных операций и функциональных возможностей системы удалось получить понимание принципов работы с данными и распределенными вычислениями. Были рассмотрены ключевые аспекты, такие как управление данными, выполнение операций чтения/записи, обработка больших объемов информации в распределенной среде, а также механизмы параллельных вычислений.