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

Агафонов Антон Александрович БД-241м

## **Практическая работа 1. Введение в большие данные и их хранение. Инструменты обработки больших данных (Hadoop)**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Вариант 1

Москва

2025

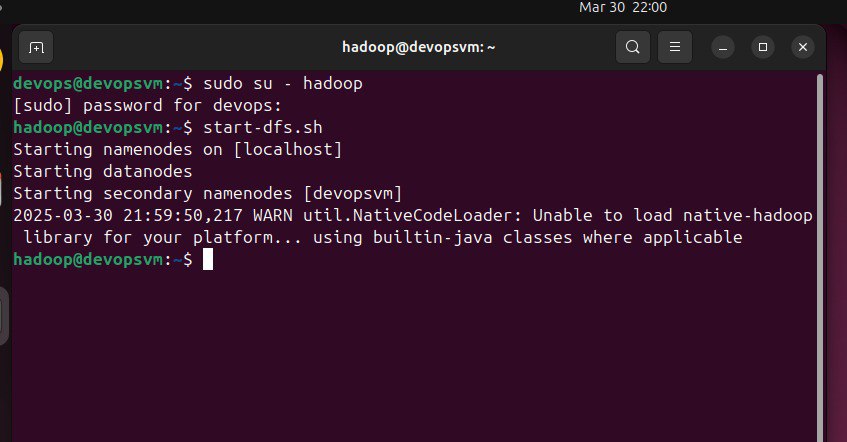
Цель

Изучить основные операции и функциональные возможности системы, что позволит понять принципы работы с данными и распределенными вычислениями.

# Основная часть

# Запускаем hadoop

# Start-dfs.sh

****

# Рис.1 Запускаем файловую систему

**Start-yarn.sh**

Проверяем запущенные службы командой

**jps**

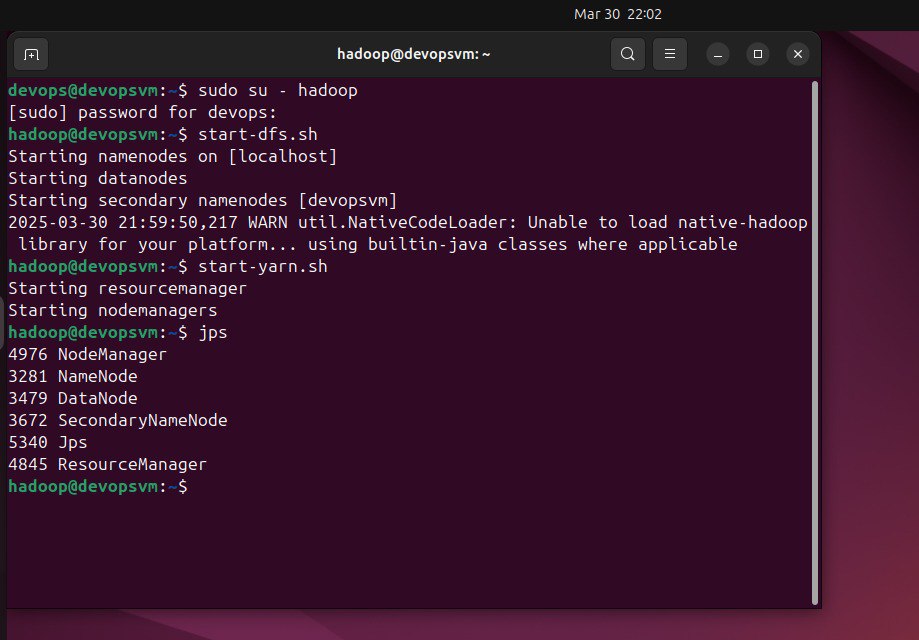
****

Рис.2 Проверка запущенных служб

Проверяем доступность запущенных систем

Переходим по ссылке для проверки запущен ли dfs по ссылке

**Localhost:9870/dfshealth.html#tab-overview**

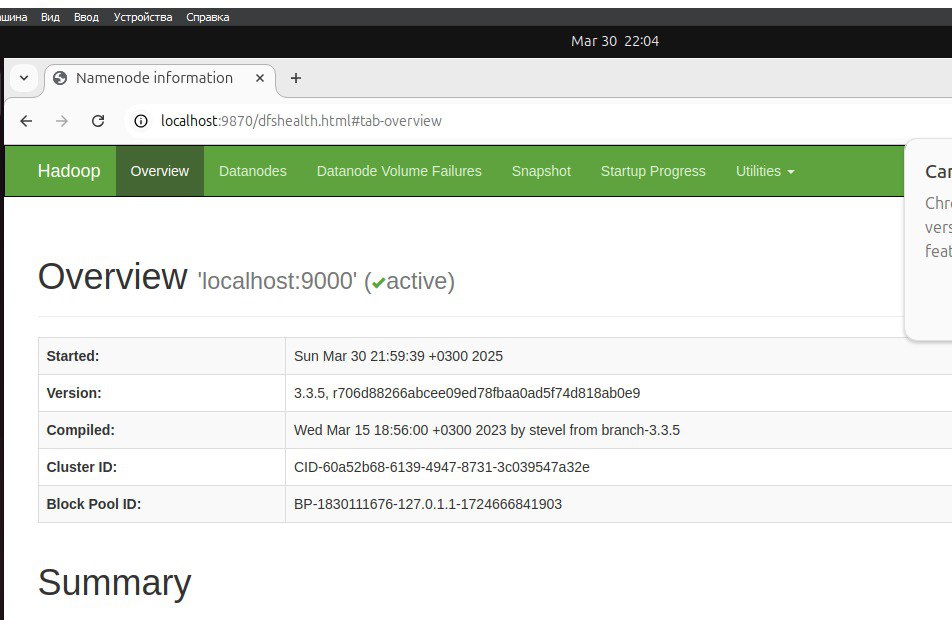
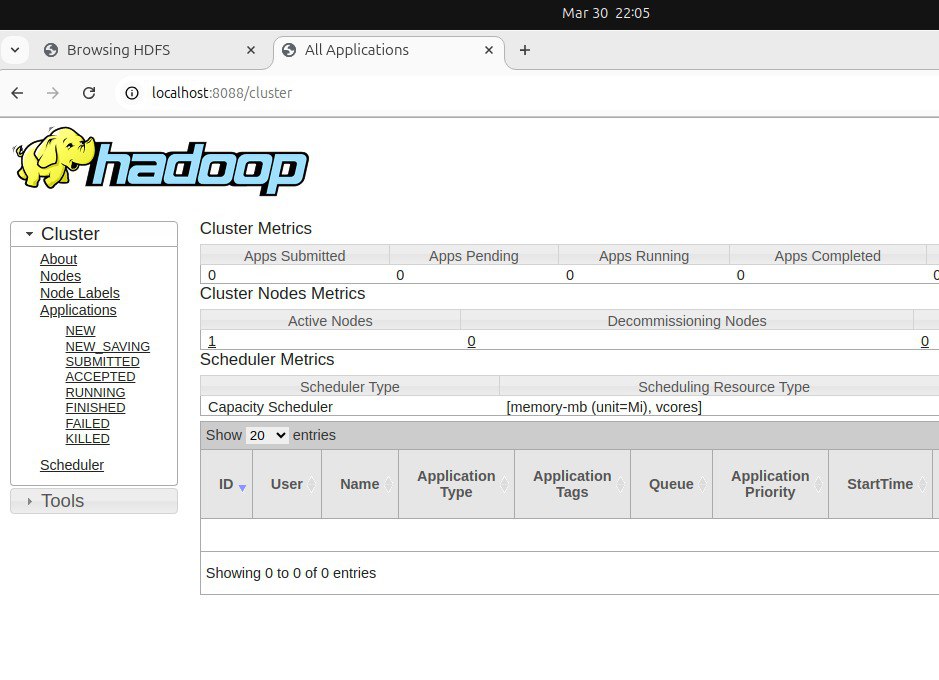
****

Рис.3 проверка доступности систем

Проверяем запущен ли yarn по ссылке

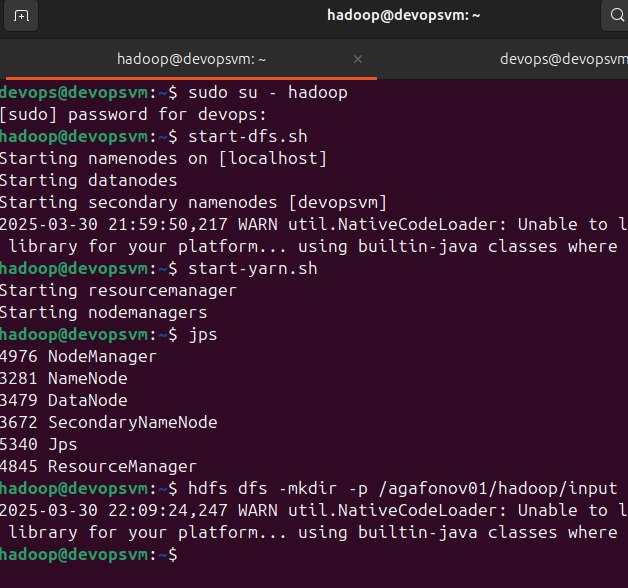
**Localhost:8088/cluster**



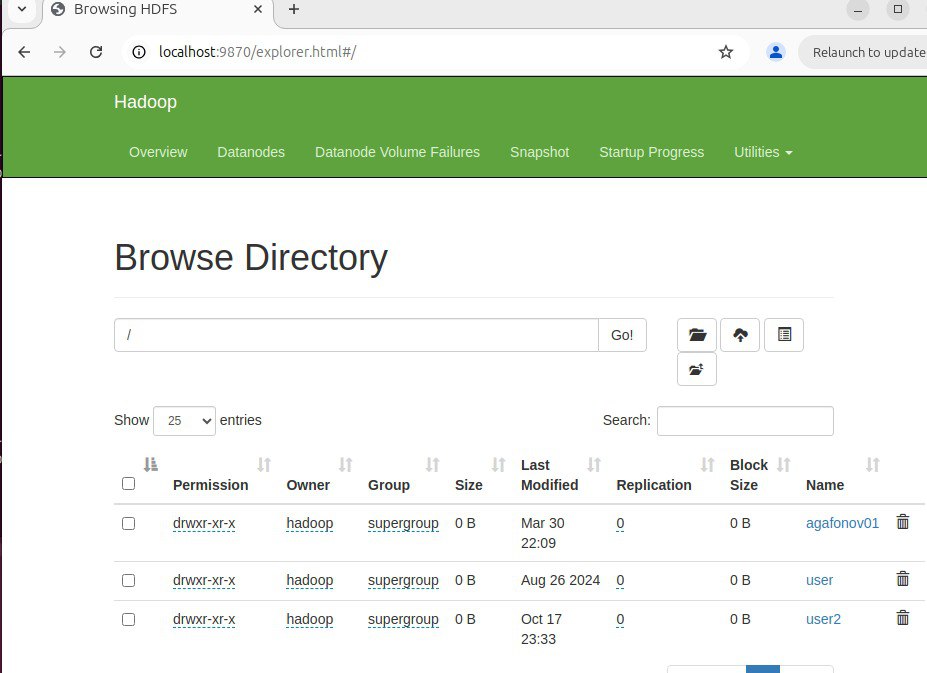
**Рис.4 Проверка работы yarn**

Создаем пользователя и каталог командой

**Hdfs dfs -mkdir -p /agafonov01/Hadoop/input**

****

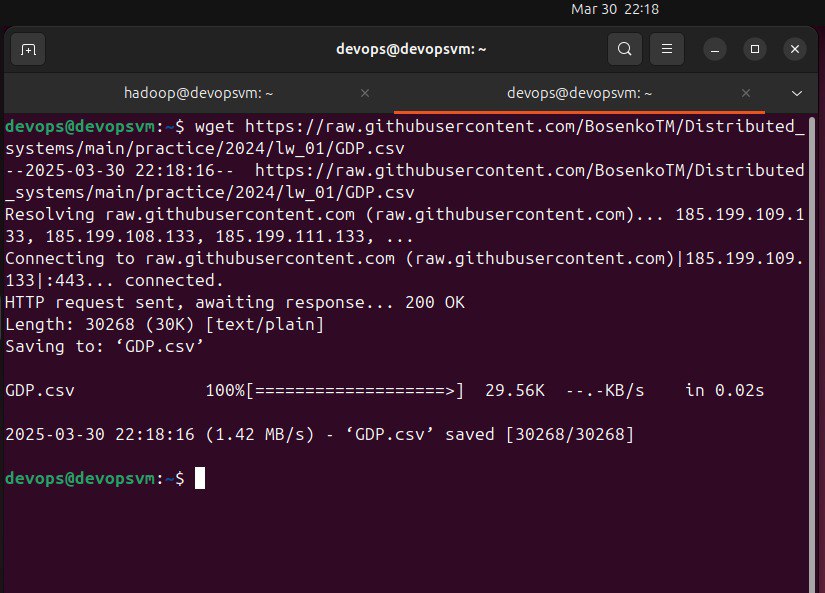
**Рис.5 Создание пользователя**

****

**Рис.6 Проверка созданного пользователя**

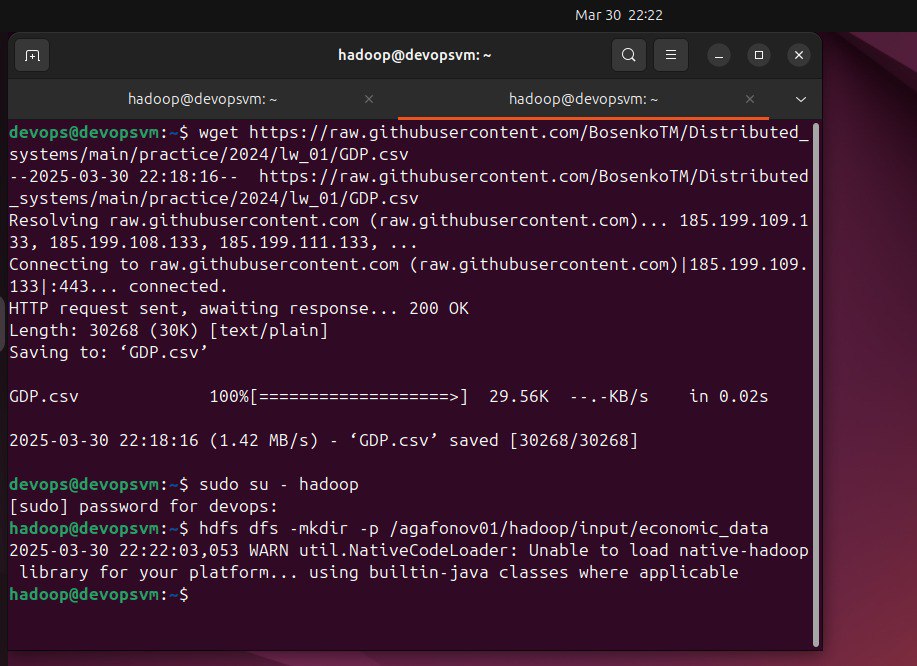
**Скачиваем данные**

**wget https://raw.githubusercontent.com/BosenkoTM/Distributed\_systems/main/practice/2024/l w\_01/GDP.csv**

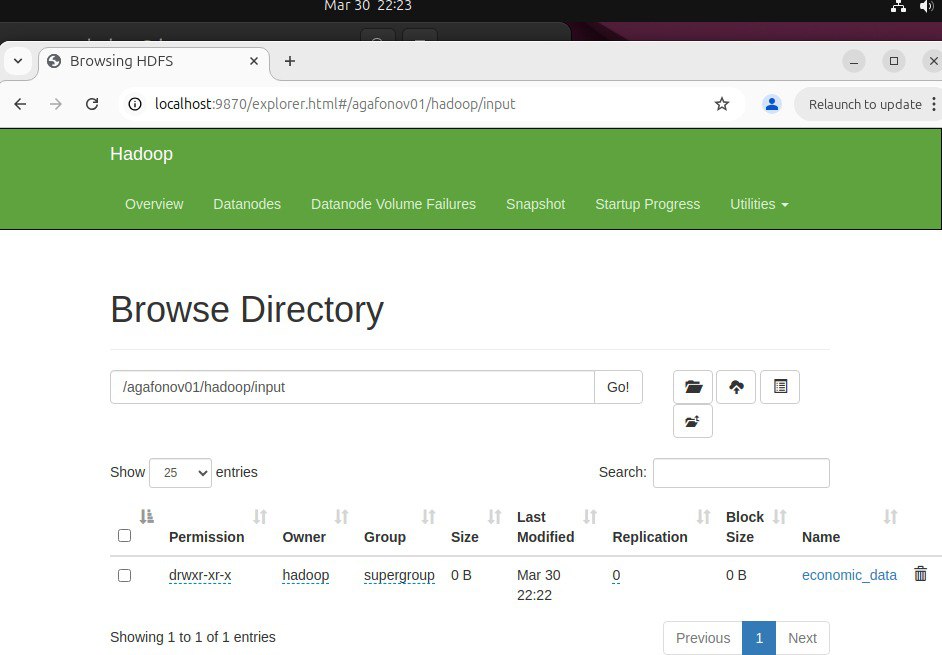
**Рис.7 Скачиваем данные**

Создаем каталог для данных

**Hdfs dfs -mkdir -p /agafonov01/Hadoop/input/economic\_data**

****

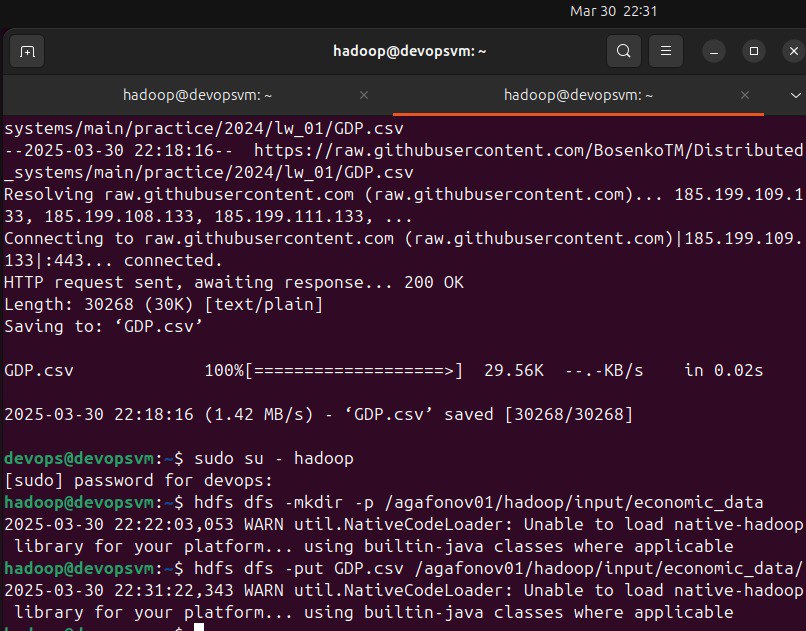
**Рис.8 Создание каталога economic\_data**

****

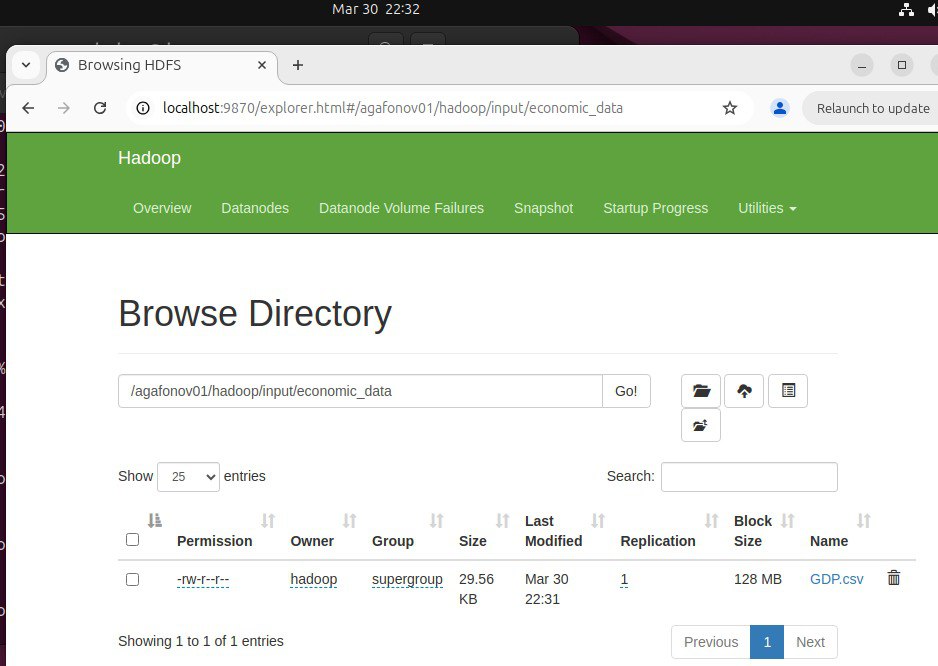
**Рис.9 Проверка созданного каталога economic\_data**

**Переносим данные в каталог**

**Hdfs dfs -put GDP.csv /agafonov01/Hadoop/input/economic\_data**



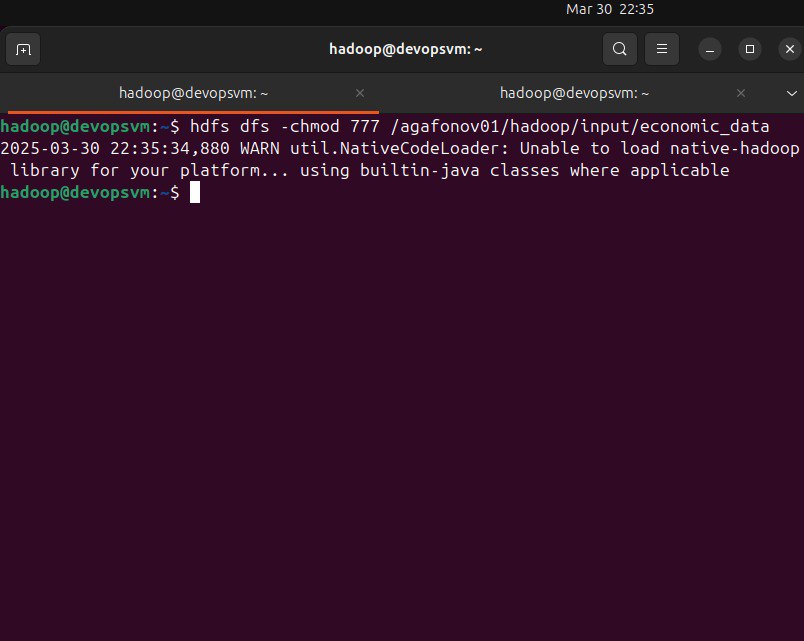
**Рис.10 Перенос данных в каталог**

****

**Рис.11 Проверка перенесенных данных**

Задаем права доступа

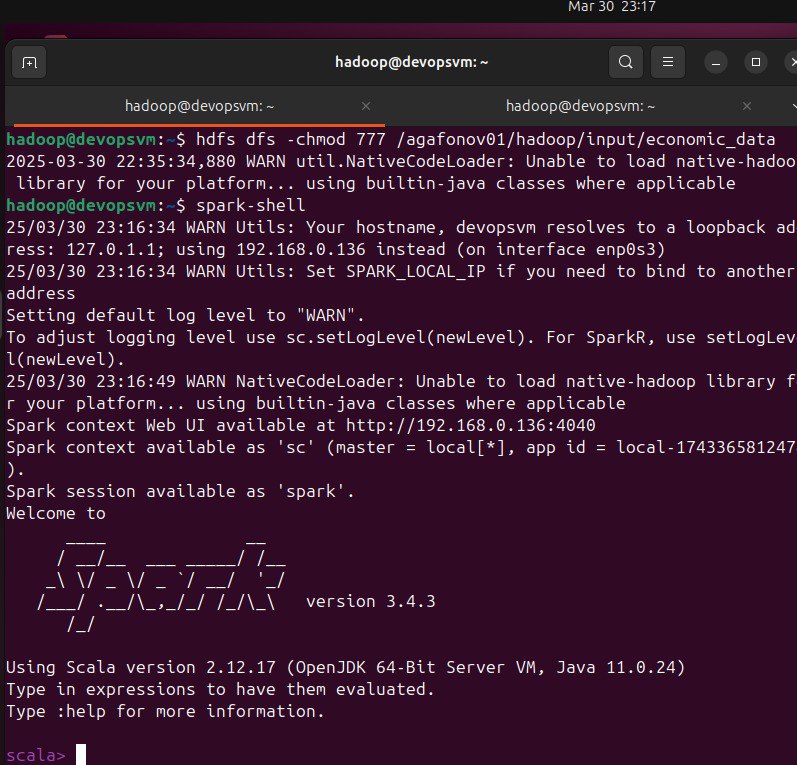
**hdfs dfs -chmod 777 /agafonov01/hadoop/input/economic\_data**



**Рис.12** Устанавливаем права доступа

**Обрабатываем данные при помощи Spark**

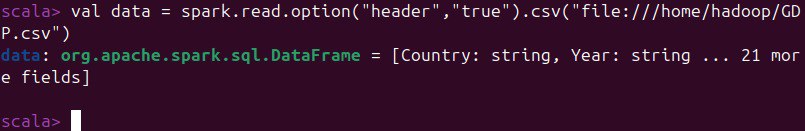
**Spark-shell**

****

**Рис.13** Запуск spark

Загружаем данных из hdfs

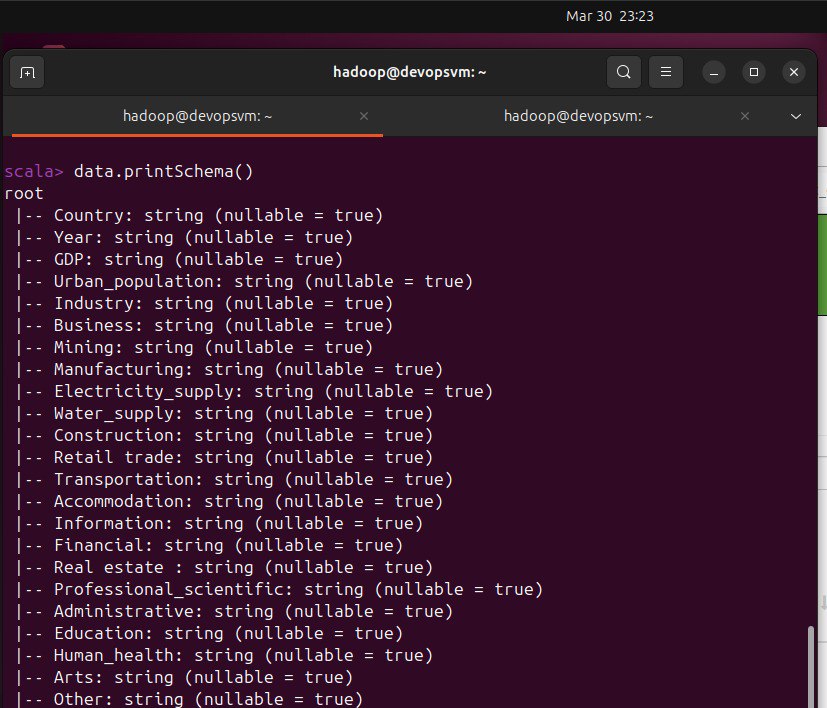
**val data = spark.read.option("header", "true").csv("file:///home/hadoop/GDP.csv")**

****

**Рис.14** Загрузка данных

Проверка полученной схемы данных

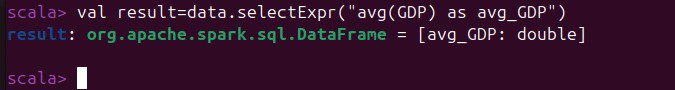
**data.printSchema()**



**Рис.15** Выводим схему

Вычисление среднего значения GDP

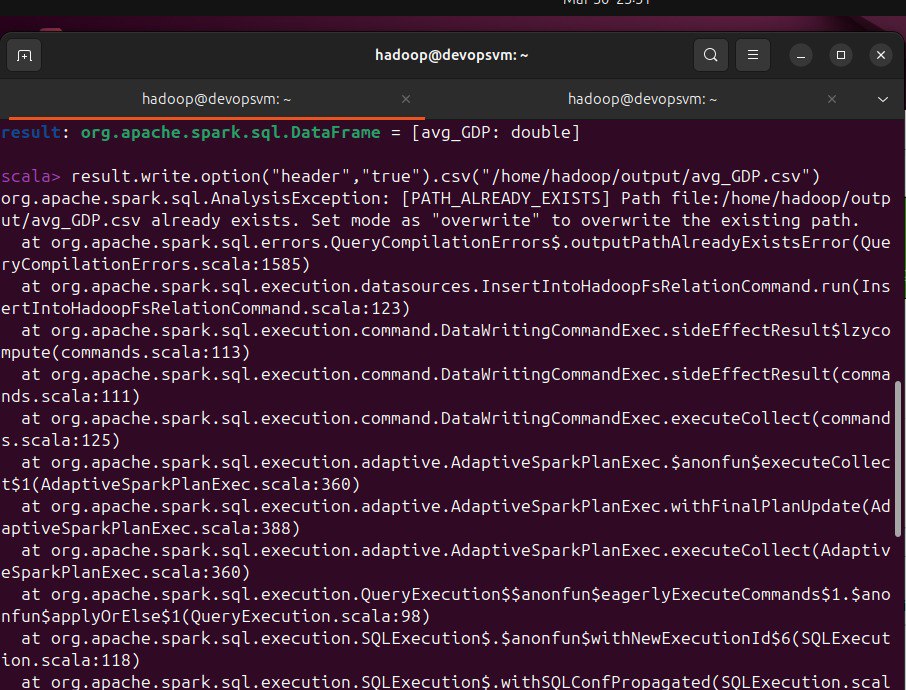
**val result = data.selectExpr("avg(GDP) as avg\_GDP")**

****

**Рис.16** Вычисление среднего значения GDP

Сохраняем результата в CSV файл

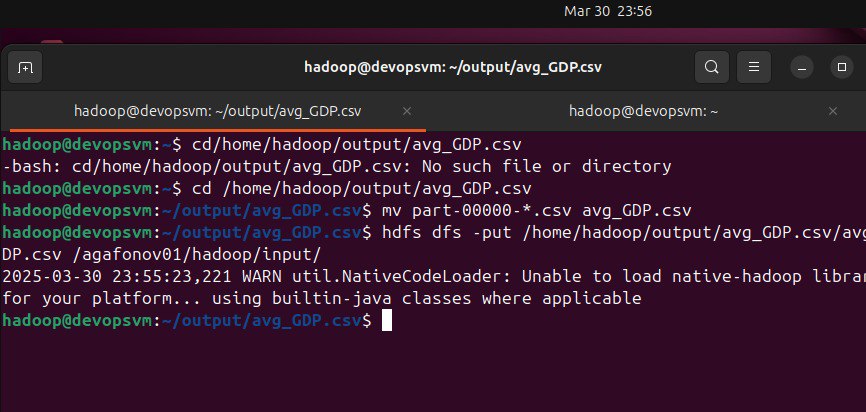
**result.write.option("header", "true").csv("/home/hadoop/output/avg\_GDP.csv")**



**Рис.17** сохранение результатов

Переходим в директорию с результатами

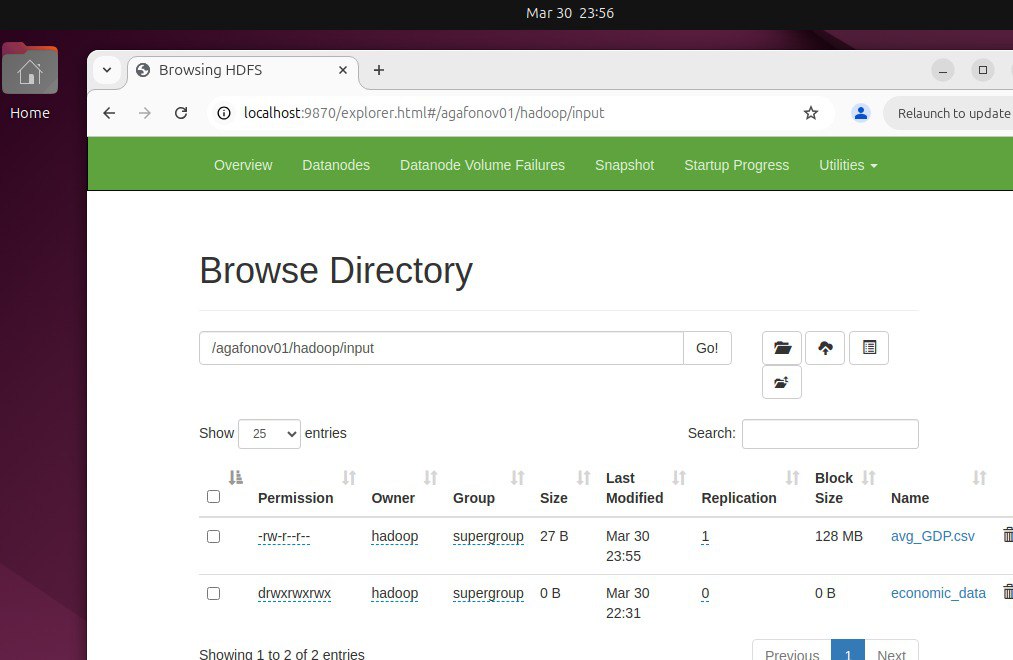
**cd /home/hadoop/output/avg\_GDP.csv**



**Рис.18 Проверяем полученный файл**

Загружаем полученный файл в HDFS

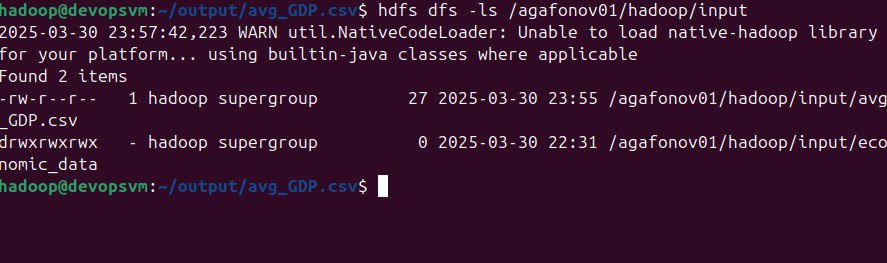
**hdfs dfs -put /home/hadoop/output/avg\_GDP.csv /agafonov01/hadoop/input/**

****

**Рис.19 Проверка загруженного файла в hdfs**

Проверка загрузки

**hdfs dfs -ls /user01/hadoop/input/**



**Рис.20 Проверка загрузки**

**Задание для самостоятельной работы**

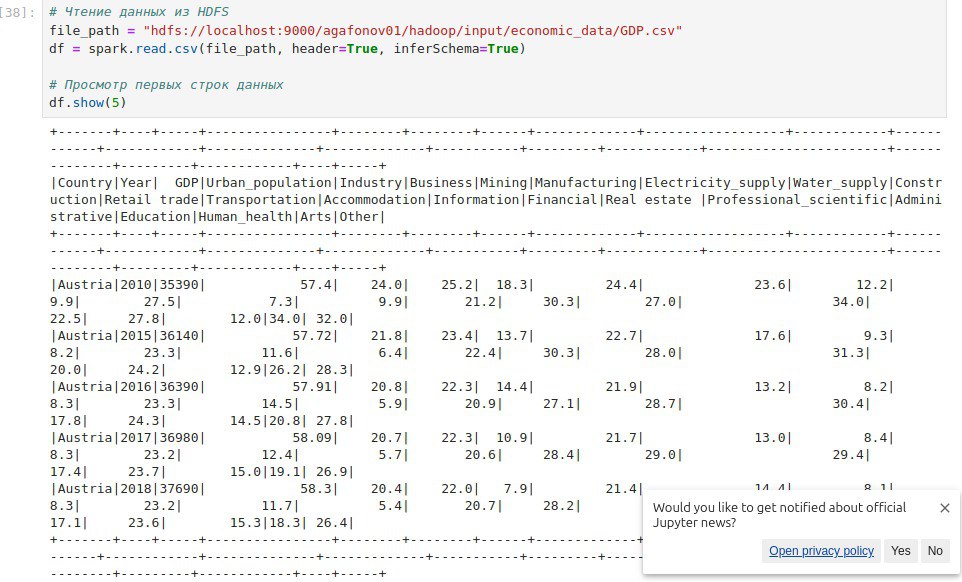
Подключиться к HDFS и убедиться, что файл доступен по пути hdfs://localhost:9000/agafonov01/hadoop/economic\_data/GDP.csv

Запускаем jupyterlab и загружаем файлы из hdfs

**File\_path= “hdfs://localhost:9000/agafonov01/Hadoop/input/economic\_data/GDP.csv”**

**df = spark.read.csv(file\_path, header=True, inferShema =True )**

**df.show(5)**



**Индивидуальное задание**

**Вариант 1**

**https://raw.githubusercontent.com/BosenkoTM/Distributed\_systems/main/practice/2024/l w\_01/AAPL.csv**

Для начала создаем каталог в hdfs

**Hdfs dfs -mkdir -p /agafonov01/hadoop/input/vol**

****

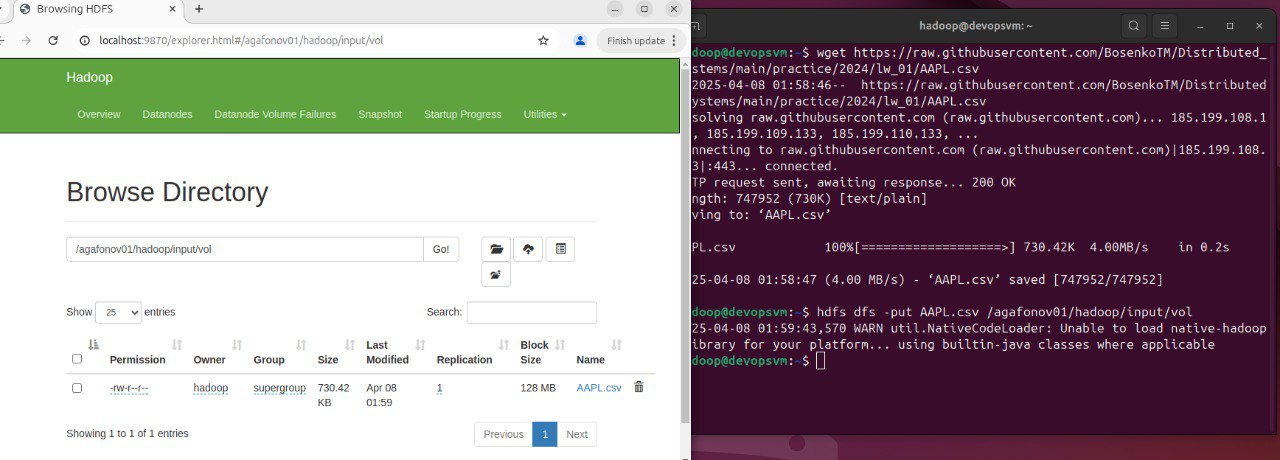
**Рис.21 Cоздание каталога surgutneftgaz в hdfs**

Переносим файл из пользователя devops в пользователя Hadoop

**sudo mv /home/devops/Desktop/AAPL.csv /home/hadoop**

Переносим данные в каталог

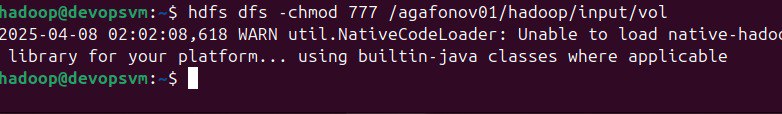
**Hdfs dfs -put AAPL.csv /agafonov01/Hadoop/input/vol**

****

**Рис.22 Переносим данные об акциях в hdfs**

Устанавливаем права доступа

**hdfs dfs -chmod 777 /agafonov01/hadoop/input/vol**



**Рис.23** Устанавливаем права доступа

Загружаем данные из hdfs

Import pandas as pd

Import matplotlib.pyplot as plt

From pyspart.sql import SparkSession

Spark = SparkSession.builder\

.appName(“Economic Data Analysis”)\

.config(“spark.hadoop.fs.defaultFS”,”hdfs://localhost:9000”)\

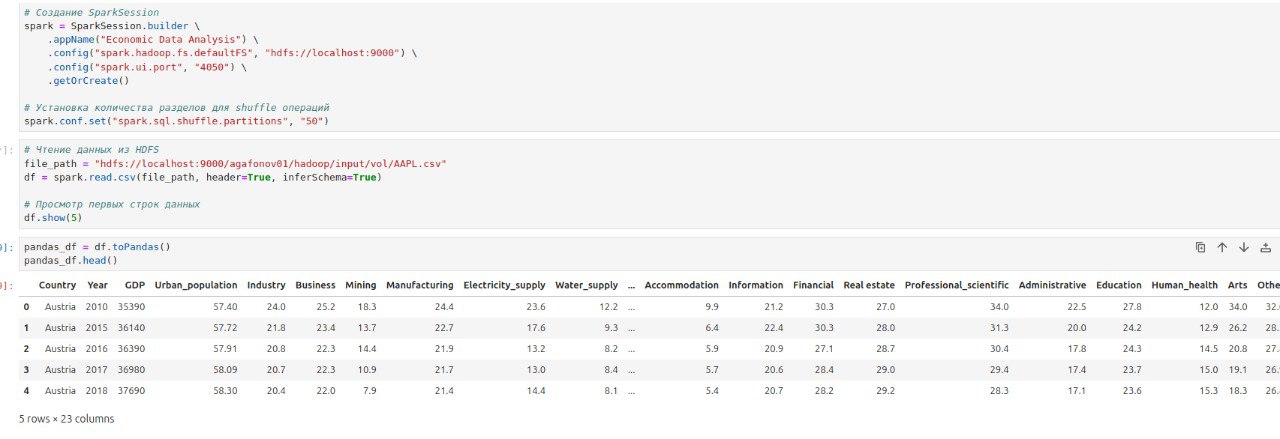
.get0Create()

Spark.conf.set(“spark.sql.shuffle.partition”, “50”)

File\_path = “hdfs://localhost:9000/agafonov01/Hadoop/input/vol/AAPL.csv

Df= spark.read.csv(file\_path, header = True, inferSchema =True)

Df.show(5)

****

# Заключение

В ходе проделанной лабораторной работы, были изучены основные операции и функциональные возможности системы, что позволило понять принципы работы с данными и распределенными вычислениями, также было выполнено задание по вариантам.