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

Кузьмина Дарья Юрьевна БД-241м

Инструменты хранения и анализа больших данных

**Лабораторная работа 1.1. Создание и управление базой данных на HDFS.**

**Введение в большие данные и их хранение. Инструменты**

**обработки больших данных (Hadoop)**

**Вариант 11**

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,

доктор экономических наук

Москва

2025

Содержание

[Введение 2](#_Toc192845116)

[Основная часть 2](#_Toc192845117)

[Заключение 12](#_Toc192845118)

## Введение

**Цель**

изучить основные операции и функциональные возможности системы, что позволит

понять принципы работы с данными и распределенными вычислениями.

**Задачи**

## Действия, которые требуется выполнить:

## 1. Подключение к Hadoop и загрузка данных.

## - Подключиться к HDFS и убедиться, что файл доступен по пути

## hdfs://localhost:9000/user01/hadoop/economic\_data/ВАШ\_ФАЙЛ.csv

## - Использовать PySpark или Pandas для загрузки данных из HDFS в DataFrame, который

## можно будет использовать для анализа.

## 2. Исследование и очистка данных.

## - Проверить структуру данных и типы столбцов (например, с помощью printSchema() для

## PySpark или describe() для Pandas).

## - Убедиться, что все данные корректны, и преобразовать необходимые столбцы в числовые

## форматы, если они изначально представлены в виде строк.

## - Проверить данные на наличие пропущенных или некорректных значений, удалить или

## заполнить такие значения в зависимости от ситуации.

## 3. Анализ данных.

## - Провести базовый статистический анализ данных:

## - Вычислить средние значения, медианы, минимумы и максимумы для экономических

## параметров.

## - Проанализировать и выявить тенденции.

## - Построить временные ряды, чтобы понять, как изменялась их экономика с течением

## времени.

## 4. Визуализация данных.

## - Построить графики (например, графики временных рядов).

## - Построить диаграммы для сравнения экономических показателей.

## 5. Сохранение и экспорт результатов.

## - Сохранить результаты анализа и визуализации в формате CSV или изображений.

## - Сохранить обработанные данные (например, данные только для отдельных стран) обратно в

## HDFS, чтобы другие команды могли использовать их для дальнейшего анализа.

## - Создать отчет, включающий ключевые выводы и визуализации, для представления

## результатов анализа заинтересованным сторонам.

## 6. Автоматизация процесса (опционально).

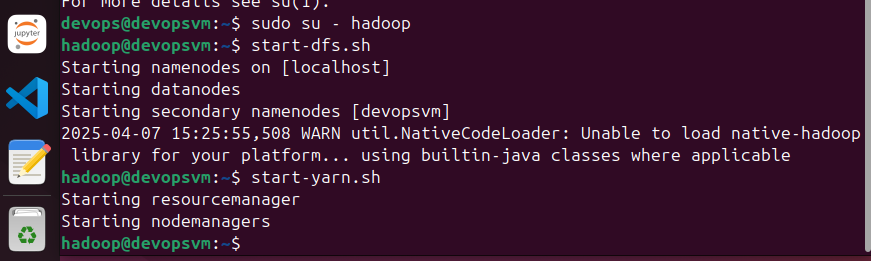
## - Создать скрипт или Jupyter Notebook, который автоматизирует процесс загрузки, анализа и

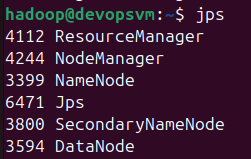
## визуализации данных для упрощения дальнейших исследований и повторного использования

## кода.

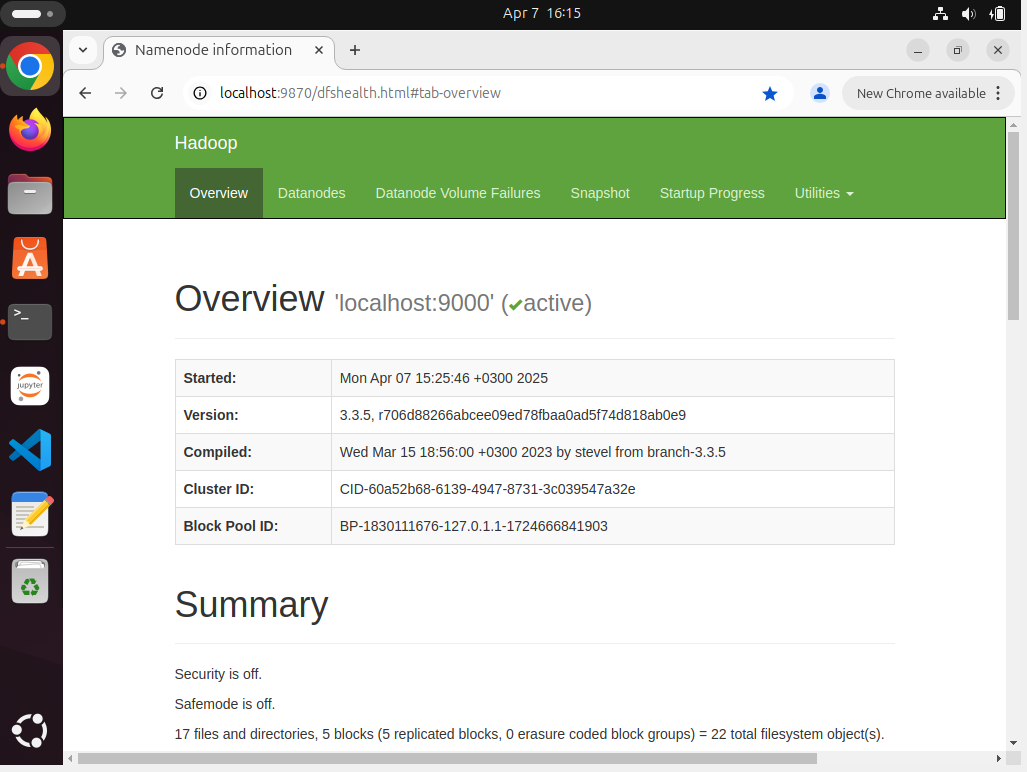
## Основная часть

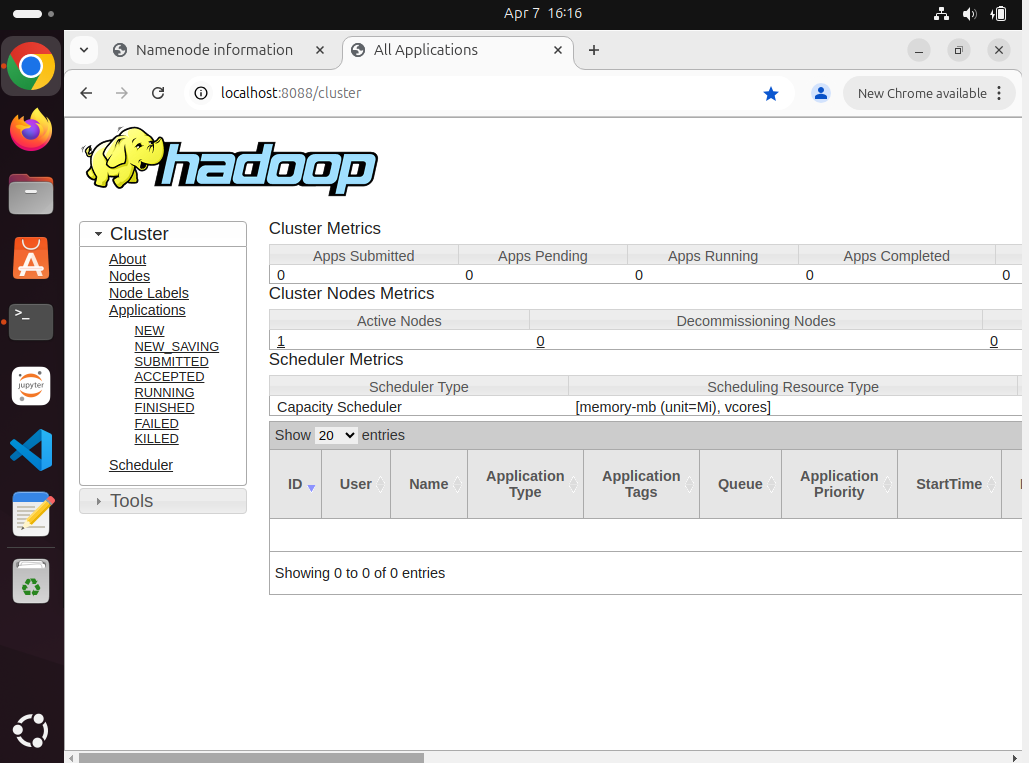
Задача 1. Запускаем yarm и dfs.





Проверяем запущенные ресурсы.

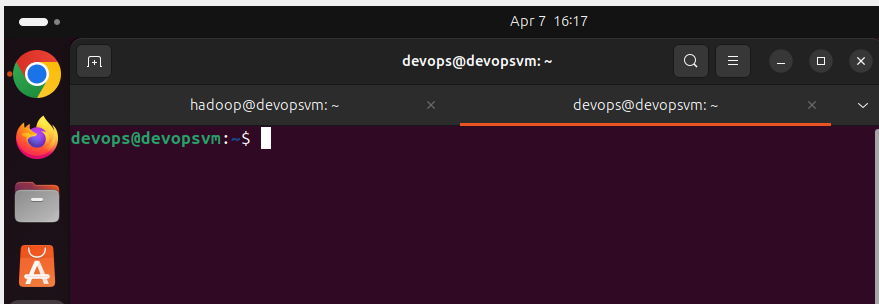




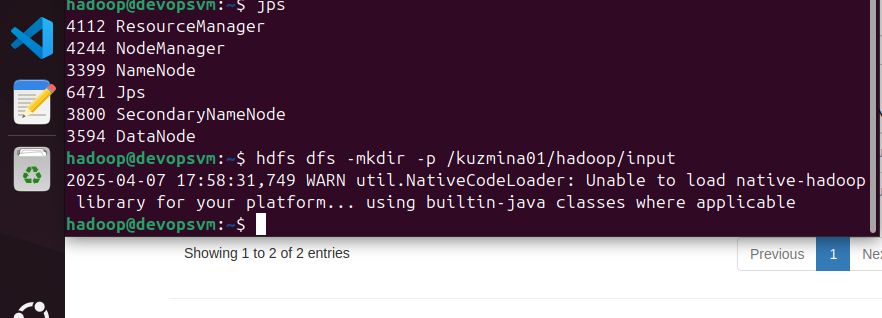
Проверяем что поднялась файловая система и yarn.

**Задача 2:** Получим данные из источника, трансформируем и посмотрим результат.

Открываем два окна для администрирования и для работы.



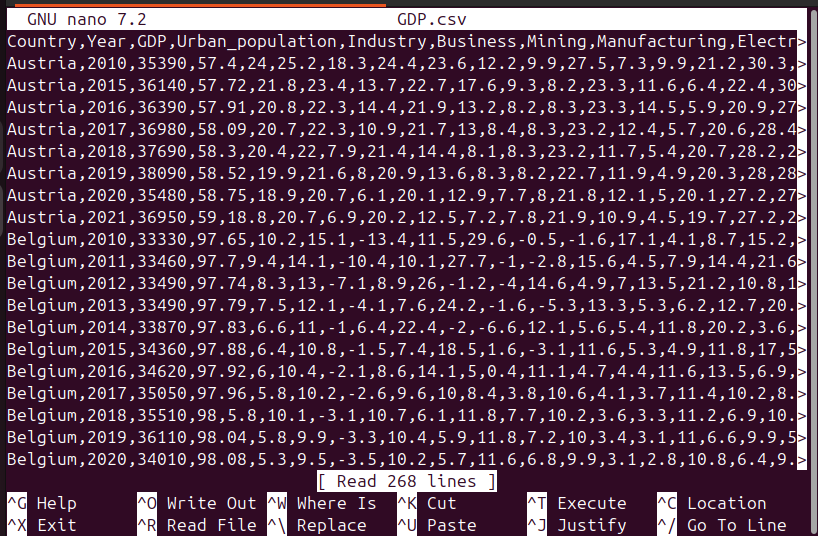
Создаем своего пользователя.



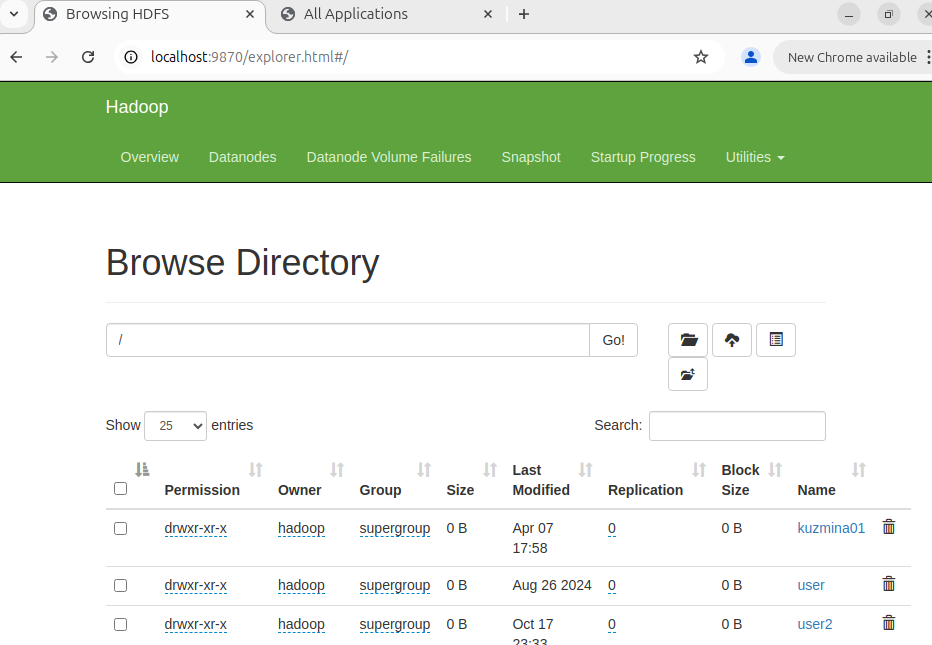
Качаем файлы.



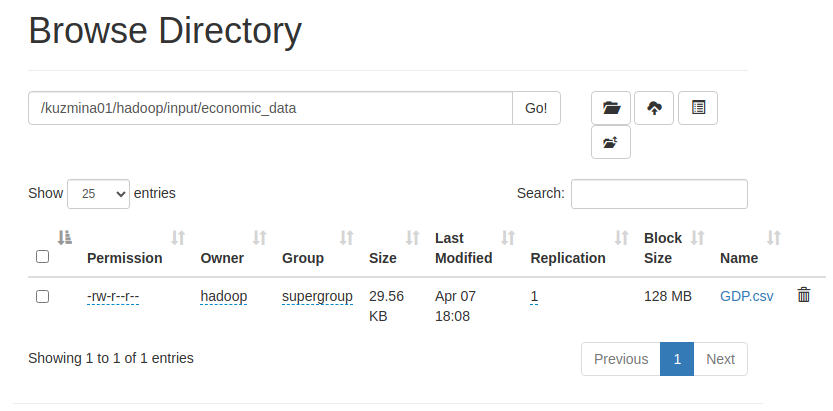
Смотрим данные.



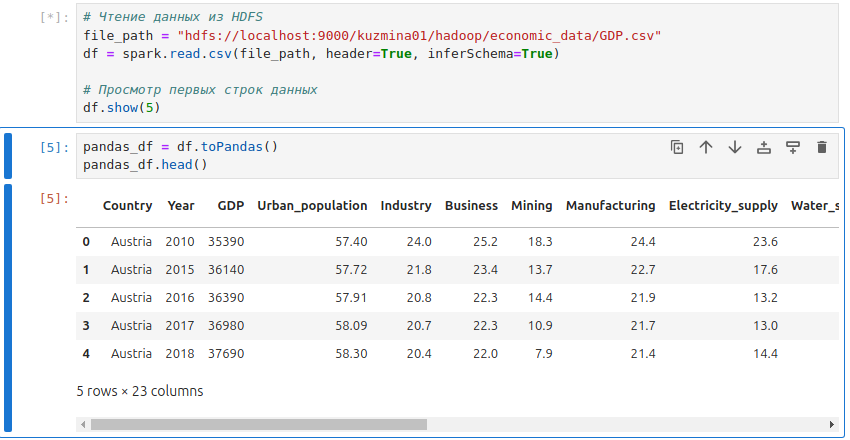
Создали каталог.



Задача 3: Перегоняем в каталог данные.



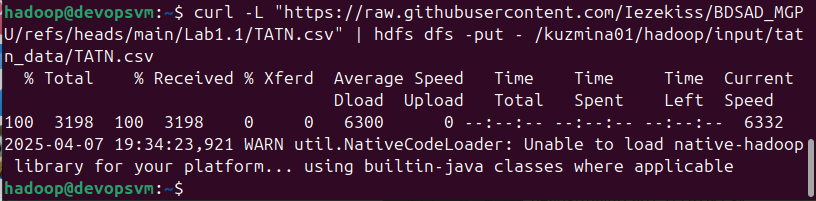
Проверяем визуализацию данных.

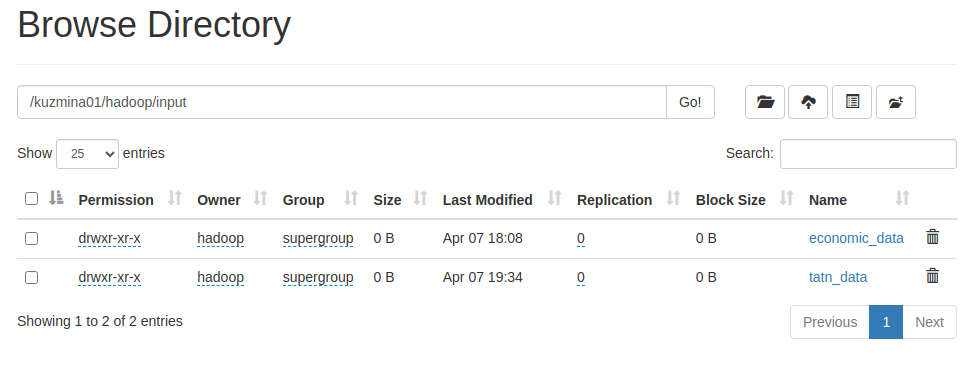


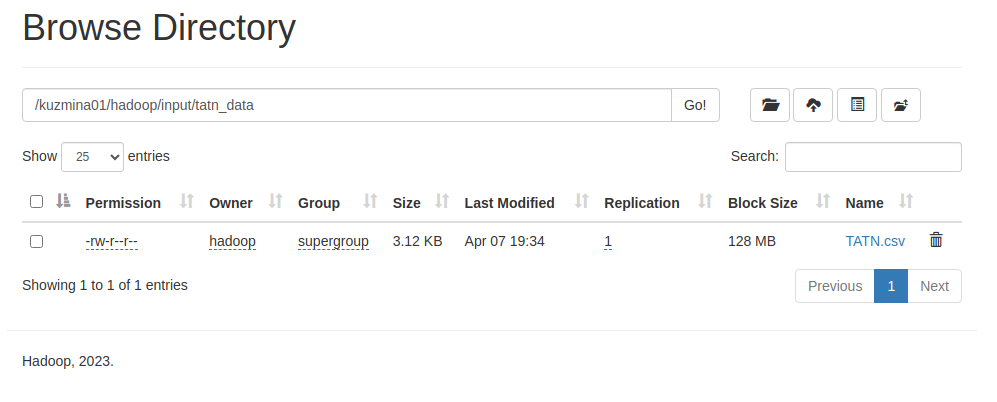
Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Индивидуальное задание **Вариант 11**  
  
Добавляем свои данные с гитхаба.



Проверяем, загрузились ли.  




Запускаем в блокноте свои данные.







Строим графики.

Изображение выглядит как линия, График, Шрифт, диаграмма

Автоматически созданное описание

# 

# Заключение

**Вывод**:

В ходе выполнения лабораторной работы были изучены основные принципы работы с распределенной файловой системой Hadoop (HDFS) и инструментами обработки больших данных. Цель работы — освоение базовых операций управления данными и распределенными вычислениями — достигнута. Реализованы следующие задачи:

1. **Подключение к HDFS и загрузка данных**:

- Успешно настроено окружение Hadoop, запущены HDFS и YARN.

- Данные загружены в HDFS, созданы необходимые каталоги и обеспечен корректный доступ к файлам.

2. **Исследование и очистка данных**:

- Проведена проверка структуры данных, преобразование типов столбцов и обработка некорректных значений.

- Устранены пропуски и аномалии, что обеспечило готовность данных для анализа.

3. **Анализ данных**:

- Выполнен базовый статистический анализ (средние значения, минимумы, максимумы).

- Построены временные ряды, выявлены тенденции в динамике экономических показателей.

4. **Визуализация**:

- Созданы графики временных рядов и диаграммы для наглядного представления результатов.

5. **Сохранение результатов**:

- Результаты анализа экспортированы в форматы CSV и изображений.

- Обработанные данные сохранены в HDFS для дальнейшего использования.

6. **Автоматизация**:

- Разработан Jupyter Notebook, автоматизирующий процесс загрузки, анализа и визуализации данных.

Работа с Hadoop и HDFS продемонстрировала их эффективность для хранения и обработки больших объемов данных. Полученные навыки критически важны для задач бизнес-аналитики, где требуется масштабируемость и скорость обработки данных.