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

Инструменты хранения и анализа больших данных

Смоляков Руслан Игоревич БД-241м

**Лабораторная работа 1.1. Введение в большие данные и их хранение. Инструменты обработки больших данных (Hadoop)**

**Вариант 23**

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**

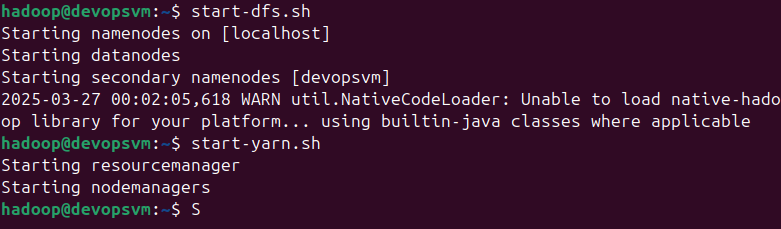
***Цель: изучить основные операции и функциональные возможности системы, что позволит понять принципы работы с данными и распределенными вычислениями.***

**Основная часть**

Переходим на нового пользователя:  

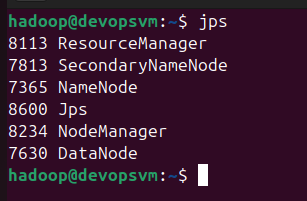

**Шаг 1. Запуск Hadoop.**  
start-dfs.sh

start-yarn.sh



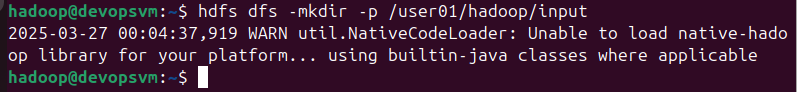
**Шаг 2. Проверка работы Hadoop.**

jps



**Шаг 3. Подготовка рабочего пространства**

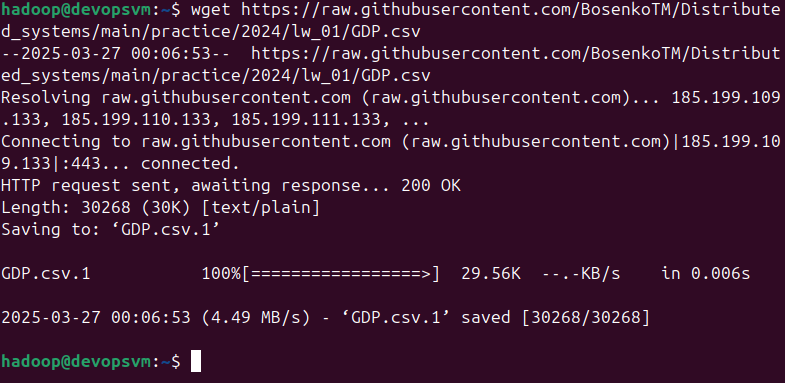
#Создание директории в HDFS   
hdfs dfs -mkdir -p /user01/hadoop/input



**Шаг 4. Загрузка и подготовка данных**

Скачивание файла с данными

wget [https://raw.githubusercontent.com/BosenkoTM/Distributed\_systems/main/practice/2024/l w\_01/GDP.csv](https://raw.githubusercontent.com/BosenkoTM/Distributed_systems/main/practice/2024/l%20w_01/GDP.csv)



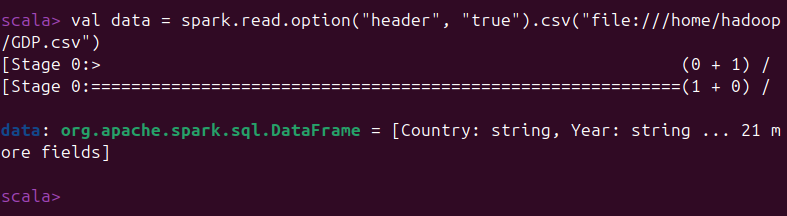
**Шаг 5. Обработка данных с помощью Spark**

spark-shell



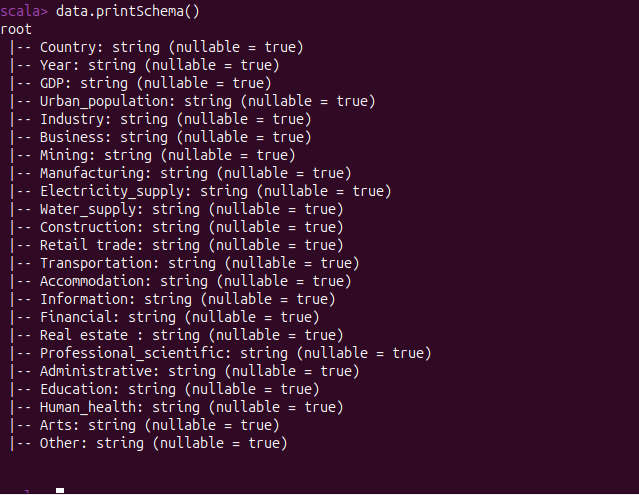
// Загрузка данных из HDFS с правильным URI

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



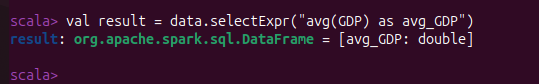
// Проверка схемы данных

data.printSchema()



// Вычисление среднего значения GDP

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

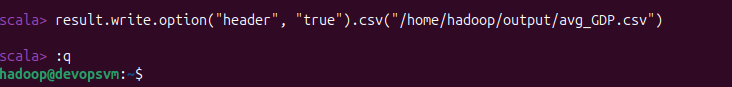


// Сохранение результата в CSV файл

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

// Выход из Spark Shell

:q



**Шаг 6. Работа с результатами**

**Переходим в директорию с результатами**

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

**Переименование файла с результатами**

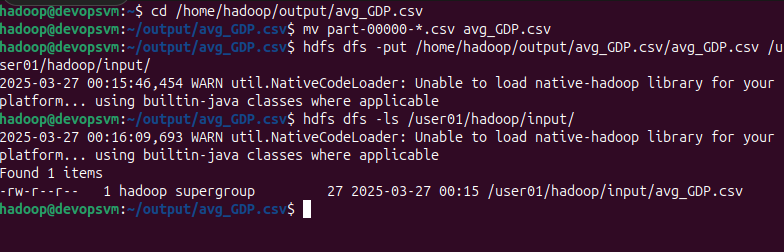
mv part-00000-\*.csv avg\_GDP.csv

**Загрузка результатов в HDFS**

hdfs dfs -put /home/hadoop/output/avg\_GDP.csv/avg\_GDP.csv /user01/hadoop/input/

**Проверка загрузки**

hdfs dfs -ls /user01/hadoop/input/



**Шаг 7. Завершение работы с Hadoop**

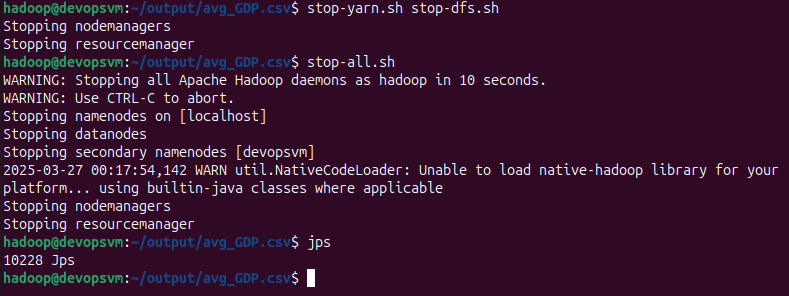
stop-yarn.sh stop-dfs.sh

**Для полной остановки всех Hadoop-демонов:**

stop-all.sh

**Проверка остановки всех процессов:**

jps



**Задание для самостоятельной работы**

1. Загрузите данные по акциям другой компании.

2. Выполните аналогичный анализ для новых данных.

3. Сравните результаты анализа двух компаний.

4. Напишите Spark-приложение, которое находит дни с максимальным объемом торгов для обеих компаний.

**Постановка задачи**

Проанализировать экономические данные, содержащиеся в вашем файле, который

находится в файловой системе Hadoop (HDFS). Задача заключается в извлечении, обработке, и

анализе данных с целью выявления закономерностей, тенденций, и создания визуализаций на

основе предоставленных данных.

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.

**Вариант 23**

Трансформация данных:

Вычисление статистических параметров и месячная агрегация

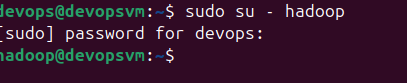
Данные для анализа:

Данные по акциям X5 Retail Group (FIVE): Finam (https://www.finam.ru/profile/mo ex-akcii/x5-retailgroup/export/), MOEX (https://www.moex.com/ ru/issue.aspx?board=TQBR&code=FIVE)

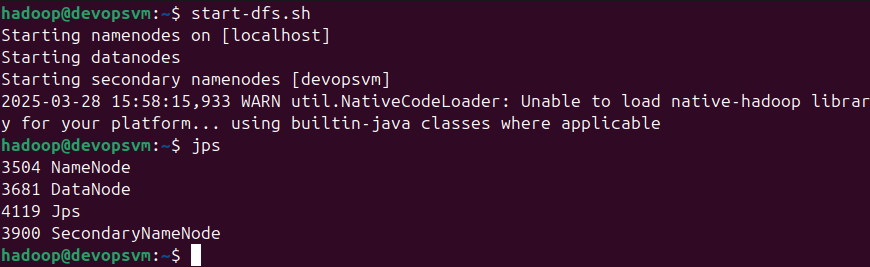
Операции анализа:

Фильтрация данных за последние 2 года, расчет средней цены закрытия, группировка по месяцам

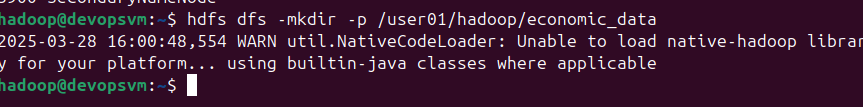
1. Переходим на пользователя «hadoop»



Подключаемся к hadoop и проверяем:



1. Создаем директорию «economic\_data» в HDFS

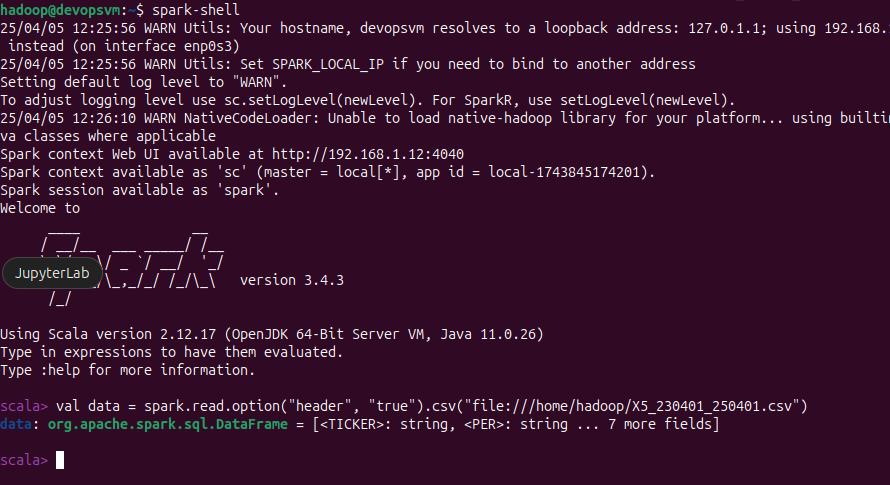


1. Загружаем данные по акциям X5 Retail Group и обрабатываем с помощью spark

Требуемые данные – X5\_230401\_250401.csv

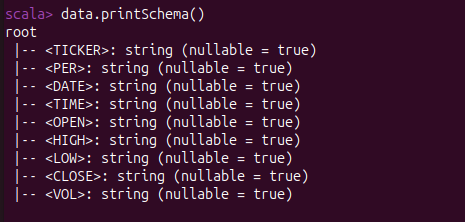


Загрузка данных из HDFS с правильным URI:



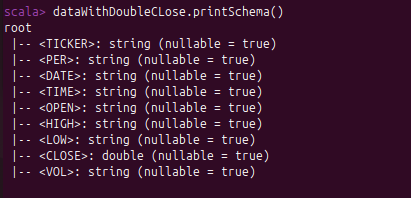
Проверка схемы данных

data.printSchema()



Видно, что цена закрытия «CLOSE» представлена в строчном формате, требуется изменить его на формат double, для этого выполним следующие команды:  


Так как в исходной схеме изменять формат нельзя, создаем новую схему с измененным форматом

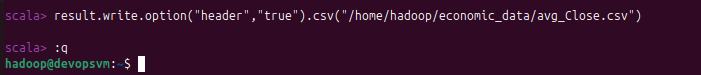
Проверяем новую схему:  


Видим, что формат изменился на «double»

Расчет средней цены закрытия:



Сохранение результата в CSV файл и выход из spark:



1. Работа с результатами

Переходим в директорию с результатами

cd /home/hadoop/ economic\_data /avg\_Close.csv

Переименование файла с результатами

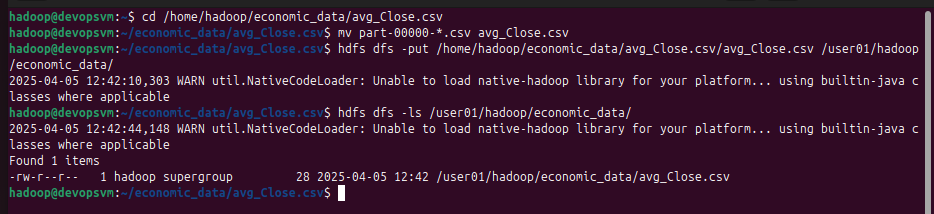
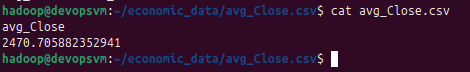
mv part-00000-\*.csv avg\_Close.csv

Загрузка результатов в HDFS

hdfs dfs -put /home/hadoop/economic\_data/avg\_Close.csv/avg\_Close.csv /user01/hadoop/economic\_data/

Проверка загрузки

hdfs dfs -ls /user01/hadoop/economic\_data/

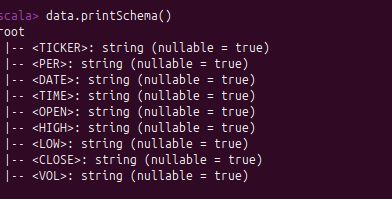
Смотрим вычисленное значение:  


Выполнили основное задание по вариантам, приступаю к следующему:

1. Исследование и очистка данных.

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

- Убедиться, что все данные корректны, и преобразовать необходимые столбцы в числовые форматы, если они изначально представлены в виде строк.

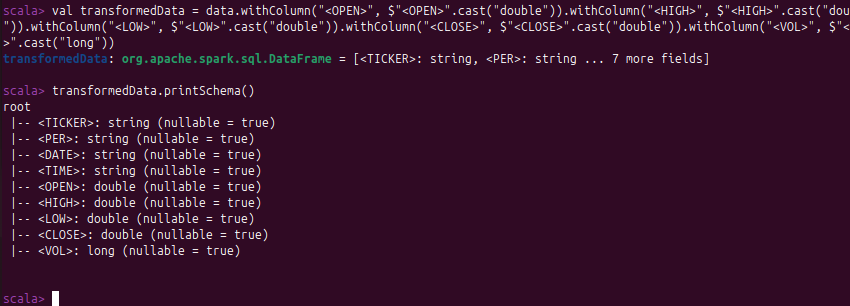


В моем случае это будут:  
**<OPEN>, <HIGH>, <LOW>, <CLOSE>**

* Эти поля представляют цены открытия, максимума, минимума и закрытия акций. (Формат «double»)

**<VOL>**

* Это объем торгов акцией (количество акций, проданных за определенный период). (Формат «long»

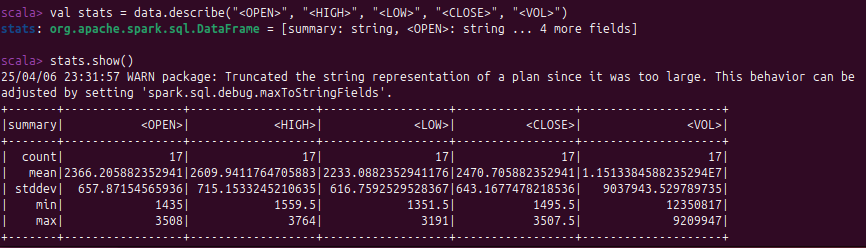
Меняем и проверяем:  


Делаем те же действия, что и раньше, но в этот раз заменим уже все типы данных переменны, которые должны быть представлены в числовом формате

2. Анализ данных.

- Провести базовый статистический анализ данных:

С помощью метода «describe» вычислим средние значения, минимум и максимум. Показать результаты с помощью метода «show».

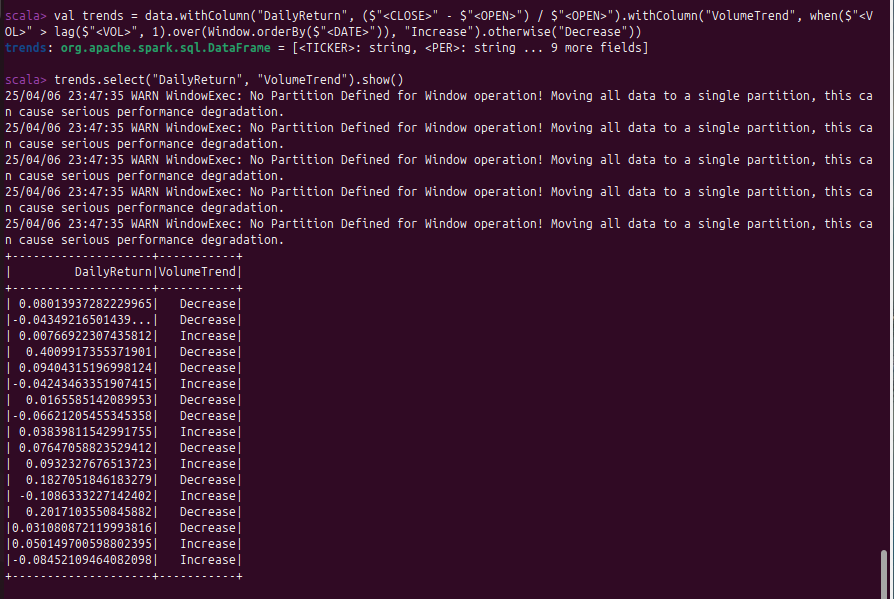


- Проанализировать и выявить тенденции.

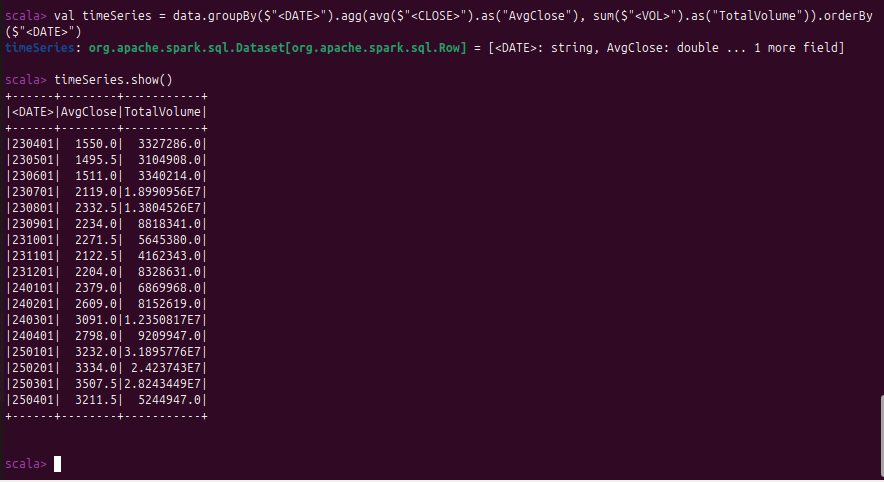
Чтобы проанализировать тенденции, можно рассчитать изменение цен (**<CLOSE>** - **<OPEN>**) и другие показатели. Например, вычислим дневную доходность "DailyReturn" и тренд изменения объема торгов "VolumeTrend":

Дневная доходность **DailyReturn** - рассчитывается как **(CLOSE - OPEN) / OPEN.**

**VolumeTrend** - Определяет, увеличивается или уменьшается объем торгов по сравнению с предыдущим днем.



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

Группируем данные по дате и вычисляет среднюю цену закрытия (AvgClose) и общий объем торгов (TotalVolume):  


3. Визуализация данных.

- Построить графики (например, графики временных рядов).

