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

ДИСЦИПЛИНА:

Инструменты для хранения и обработки больших данных

**Лабораторная работа 5.1**

**Развертывание и настройка кластера Hadoop**

Выполнил(а): Шведова С.С., группа: АДЭУ-211

Преподаватель: Босенко Т.М.

Москва

2024

**Цель:** ознакомление с процессом установки и настройки распределенных систем, таких как Apache(Arenadata) Hadoop. Изучить основные операции и функциональные возможности системы, что позволит понять принципы работы с данными и распределенными вычислениями. **Необходимое ПО:**

- Ubuntu 24.04 LTS (22.04, 20.04) или новее.

- Java 8 ил Java11 или новее.

- Apache Spark 3.4.3.

- Python 3.12+.

- pip (менеджер пакетов Python).

**Задачи:**

1. Установить Apache Hadoop;

2. Извлечь, обработать и проанализировать данные с целью выявления закономерностей, тенденций, и создание визуализаций на основе предоставленных данных.

**Вариант 13.** Данные: Исторические данные по акциям Мечела (MTLR) с сайта Московской биржи (moex.com) Операции: Фильтрация данных за последние 3 года, расчет медианной цены закрытия, группировка по кварталам.

Для начала надо войти пользователем hadoop (рисунок 1).



Рисунок 1. Вход пользователем hadoop.

Затем на рисунках 2-3 происходит запуск Hadoop.

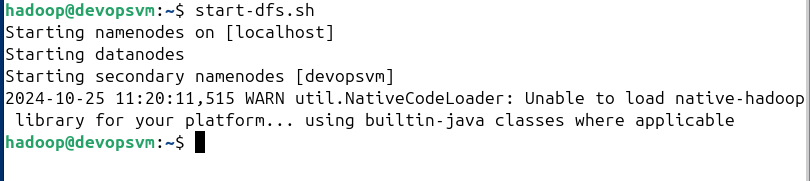


Рисунок 2. Запуск Hadoop

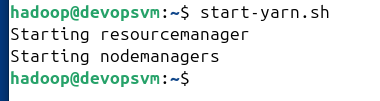


Рисунок 3. Запуск Hadoop

Как можно заметить на рисунке 4, Hadoop успешно запустился.

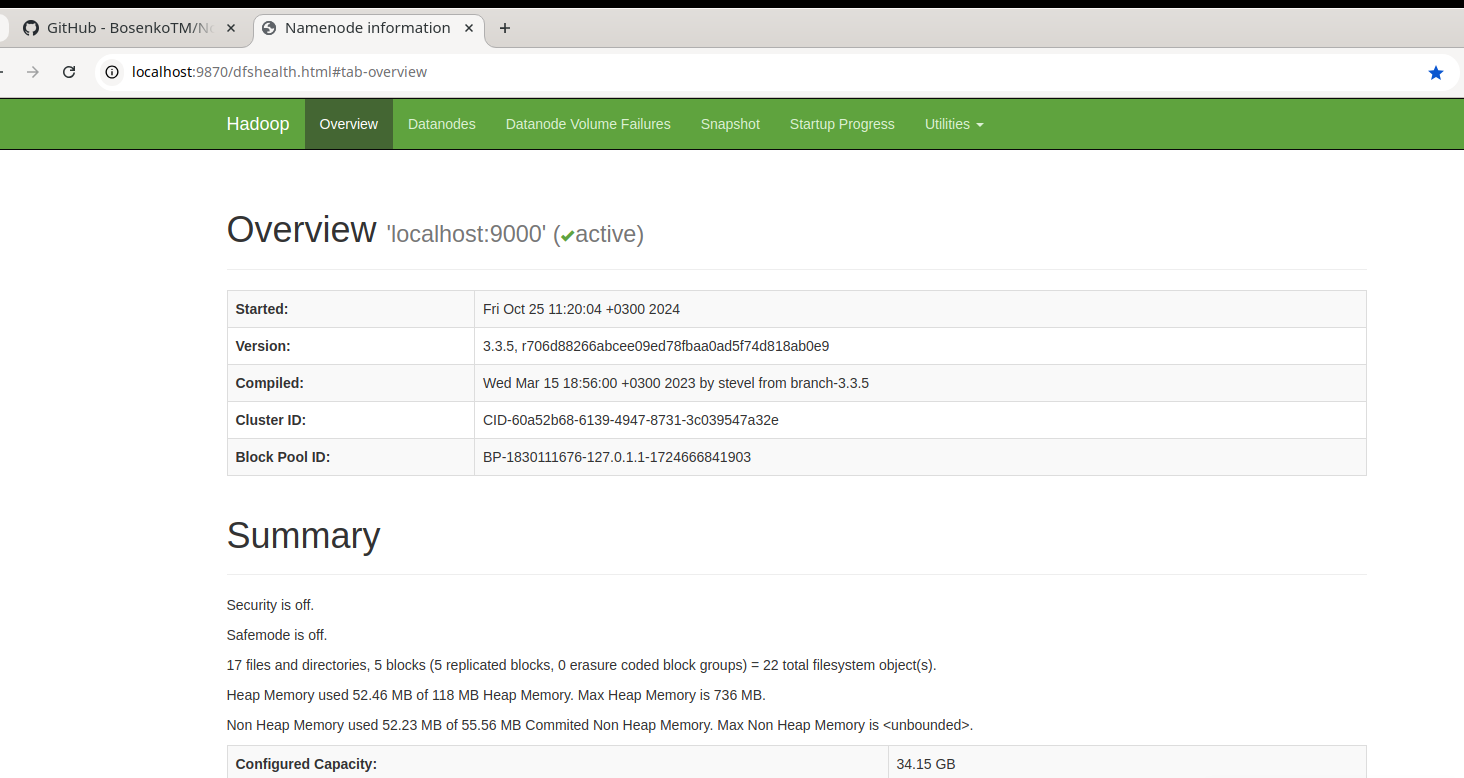


Рисунок 4. Hadoop успешно запустился

Директория продемонстрирована на рисунке 5.

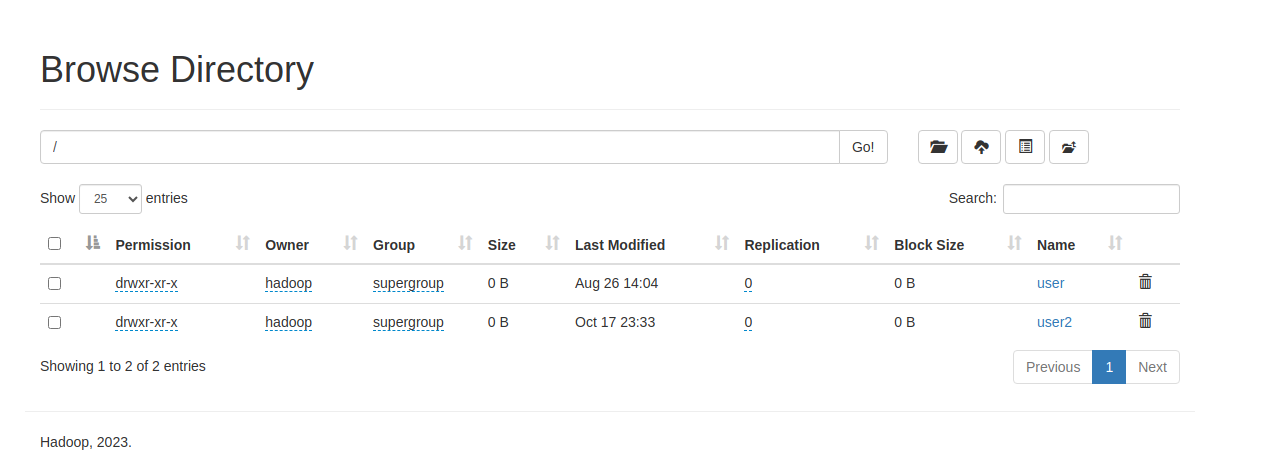


Рисунок 5. Директория

Кластеры Hadoop показаны на рисунке 6.

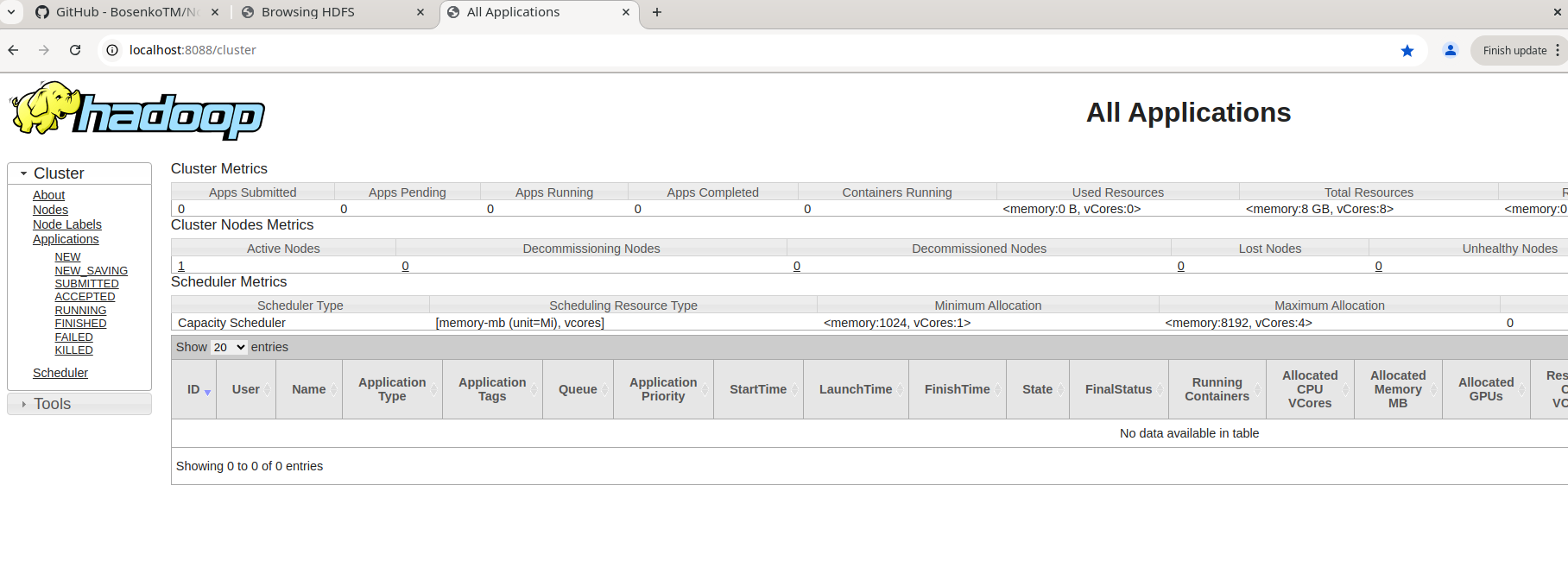


Рисунок 6. Кластеры Hadoop

Далее создается директория user4 в HDFS (рисунок 7).



Рисунок 7. Создание директории user4 в HDFS

Загрузка данных в HDFS через гитхаб показана на рисунках 8-9.

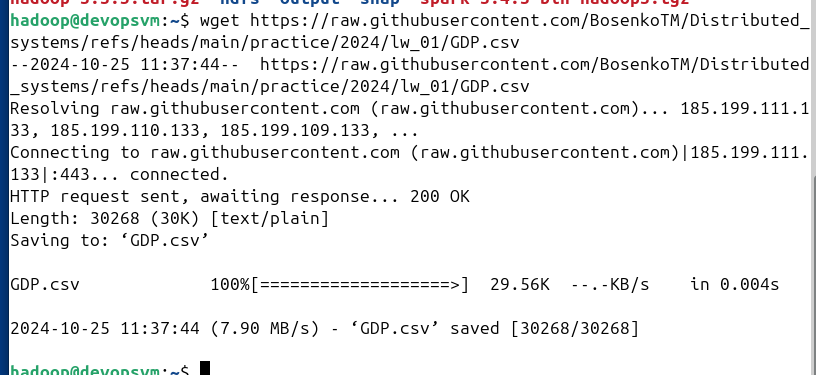


Рисунок 8. Загрузка данных в HDFS

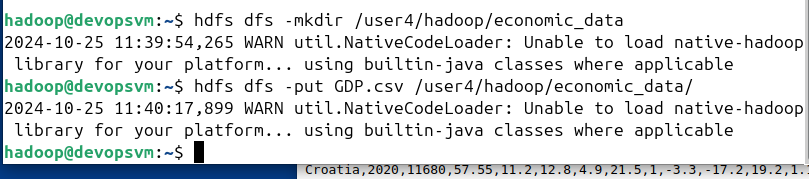


Рисунок 9. Загрузка данных в HDFS

Предоставление другим пользователям доступ на изменения данных на показано на рисунке 10.

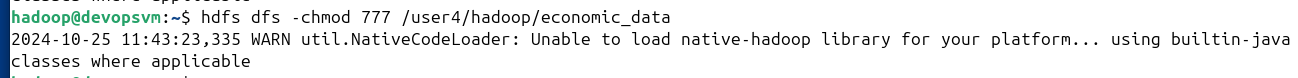


Рисунок 10. Предоставление другим пользователям доступ на изменение данных

**Индивидуальное задание**

Для начала надо загрузить исходный файл эксель с миллионом количеством строк весом 90 МБ, после перевода в CSV стало 67 МБ. На рисунке 11 видно, что файл успешно загружен в Hadoop.

C:\Users\Sonya\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Screenshot_1.png

Рисунок 11. Файл MLTR загружен в Hadoop

Затем нужно загрузить необходимые библиотеки и выполнить подключение (рисунок 12).

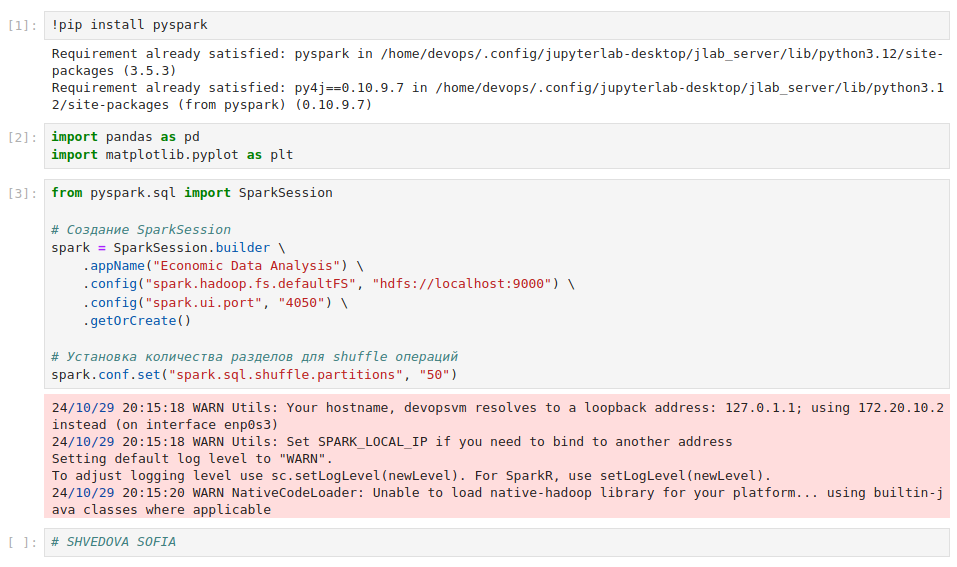


Рисунок 12. Подключение успешно выполнено

Далее нужно произвести чтение данных из HDFS и посмотреть строки данных (рисунок 13).

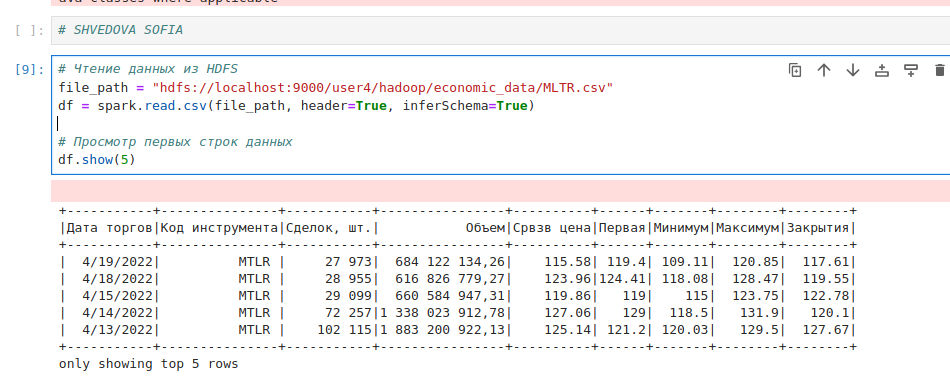


Рисунок 13. Чтение и просмотр первых строк данных

На рисунке 14 показан перевод в пандас и подсчет количества строк, чтобы убедиться в том, что все строки были загружены (их 1 миллион).

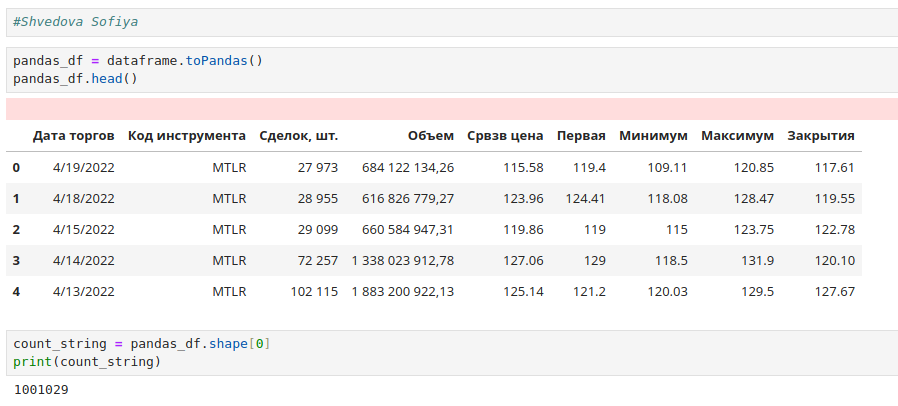


Рисунок 15. Перевод в пандас и подсчёт количества строк

Затем нужно провести фильтрацию данных за 3 года (будут взяты данные с 1 сентября 2021 года по 1 сентября 2024 года). Это показано на рисунке 16.

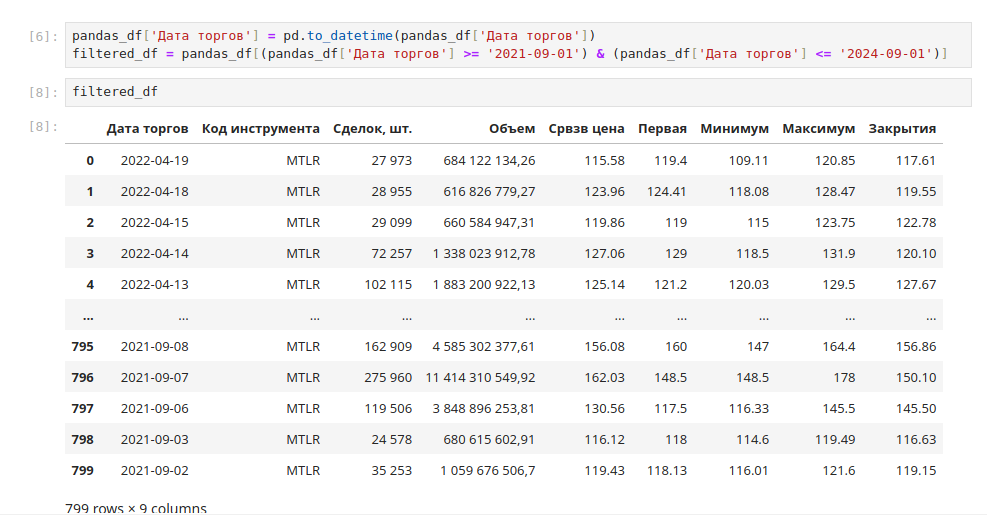


Рисунок 16. Фильтрация данных за 3 года

Далее нужно проверить типы данных. Как можно увидеть, все, кроме даты торгов и закрытия имеют тип object (рисунок 17).

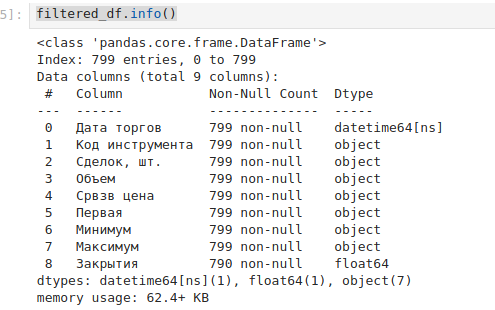


Рисунок 17. Типы данных

Перевод типа данных object в float показан на рисунке 18.

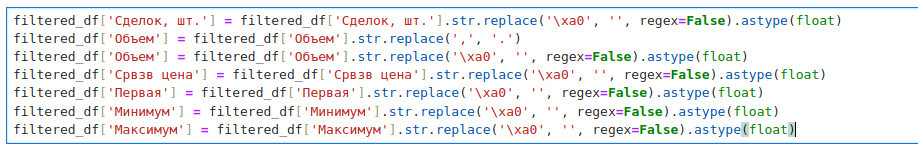


Рисунок 18. Перевод типа данных object в float

Как можно заметить на рисунке 19, тип данных успешно сменились на нужный, а также была проведена проверка структуры данных.

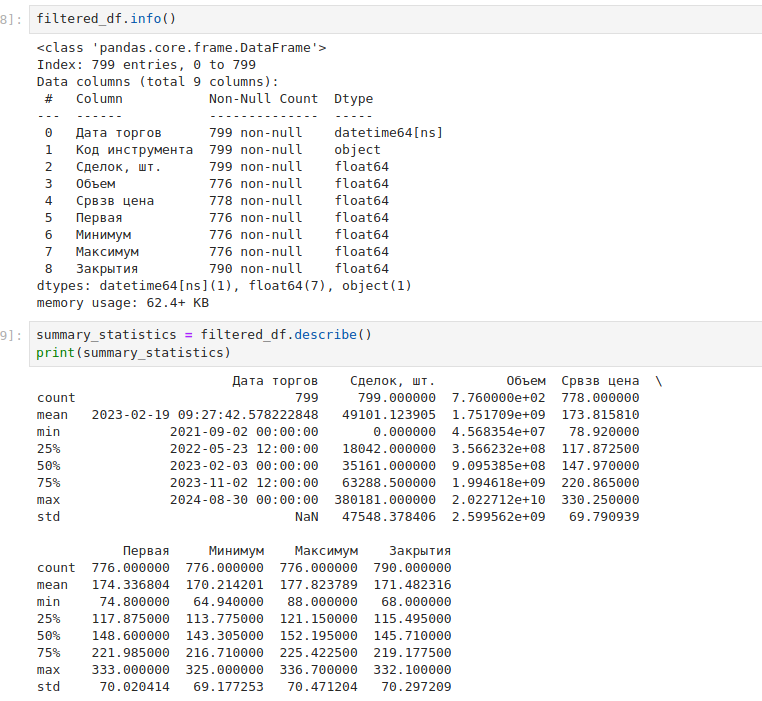


Рисунок 19. Тип данных сменился на float и проведена проверка структуры данных

Далее необходимо расчет медианной цены закрытия за последние 3 года. Это продемонстрировано на рисунке 20.

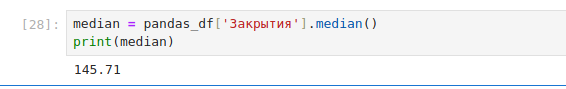


Рисунок 20. Расчет медианной цены закрытия за последние 3 года

Затем нужно было сгруппировать по кварталам. Для удобства восприятия были созданы столбцы «Год» и «Квартал». Также был произведен перевод «Дата торгов» в datetime. Для сделок выбрана сумма, чтобы посчитать общее количество сделок в определенный квартал, закрытие – среднее, максимум сделок – максимум, минимум соответственно минимум, а первая сделка – первое значение. Группировка продемонстрирована на рисунке 21.

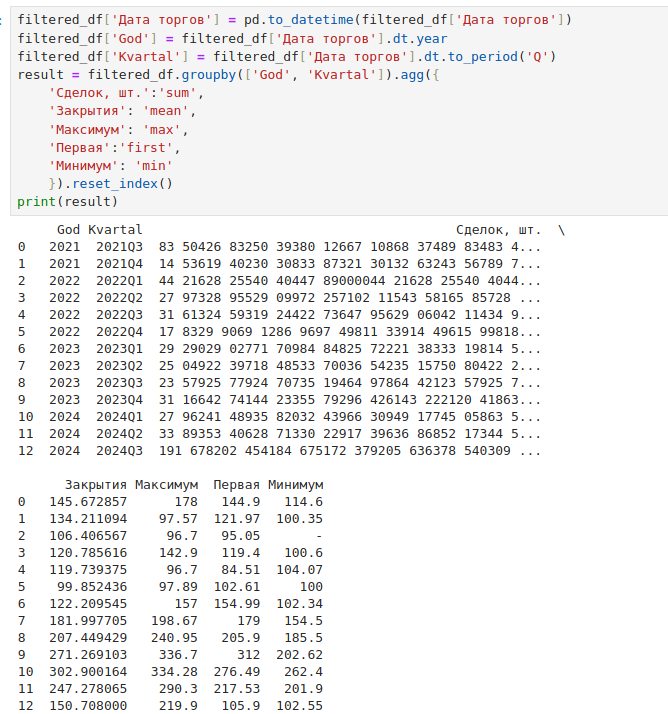


Рисунок 21. Группировка по кварталам

На рисунке 22 представлено в виде пандаса для удобства восприятия.

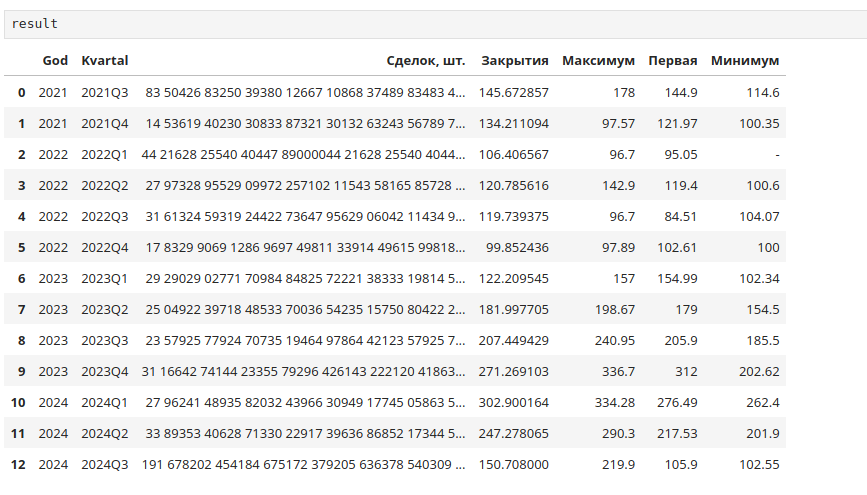


Рисунок 22. Группировка по кварталам в виде пандаса

Далее нужно провизуализировать полученные результаты. Была для этого использована библиотека matplotlib.pyplot, установлен формат для квартала, использованы функции для добавления заголовка и меток осей к графику, а также подписаны сами значения на каждый квартал закрытия (для удобства отображение было применено округление). Как можно заметить, в первый квартал 2024 года (январь-март) самая большое значение закрытия – 303, а самое маленькое – 4 квартал 2022 года (октябрь-декабрь), а именно 100. Компания Meltron AB производит осветительные приборы. Скорее всего, это связано с тем, что AB страдала от замедления бизнес-активности в предыдущие годы из-за пандемии COVID-19, с течением времени спрос на осветительные приборы значительно вырос. Также мог повлиять выпуск новой продукции. Сам график представлен на рисунке 23.

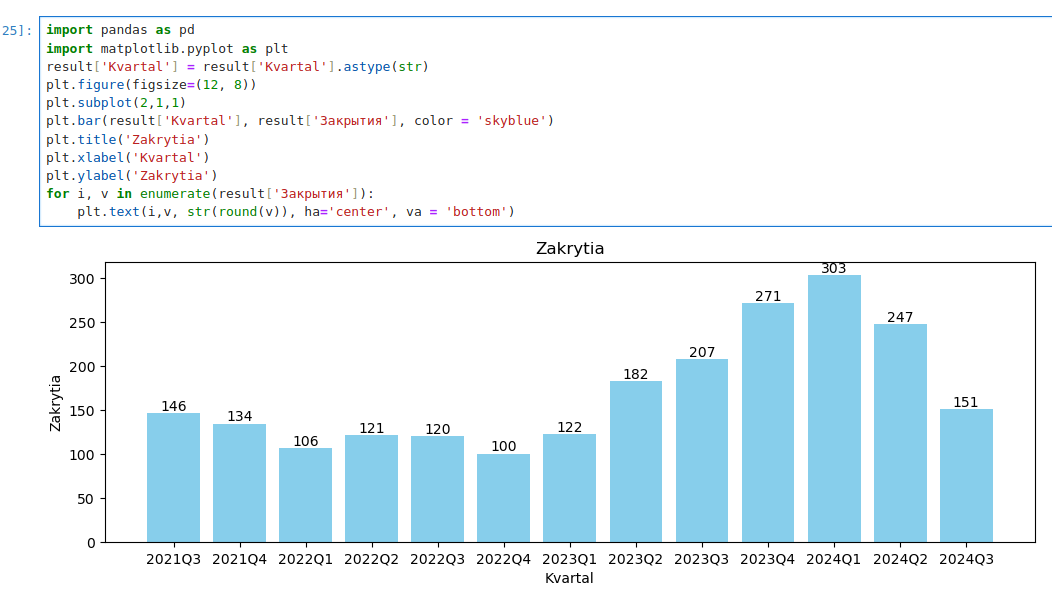


Рисунок 23. Цена закрытия по кварталам

График сделок по кварталам продемонстрирован на рисунке 24, что демонстрирует, что в 2023 году в 4 квартале (октябрь-декабрь) резко был всплеск количества сделок, что довольно схож с предыдущим графиком, так как этот квартал был второй по цене закрытия. Самое низкое – 4 квартал 2022 года, как и в предыдущем анализе.

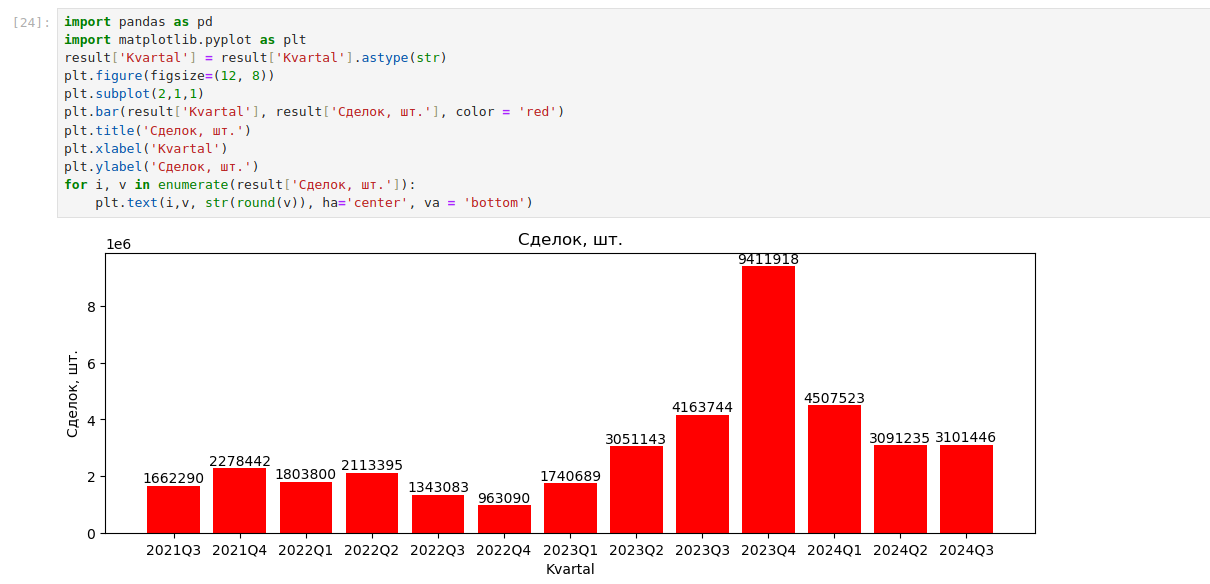


Рисунок 24. График сделок по кварталам

**Вывод:**

Hadoop - это мощный инструмент для решения проблем с обработкой и анализом больших наборов данных. Он обеспечивает масштабируемость, отказоустойчивость, экономичность и доступность для широкого круга приложений. Была проведена установка Apache Hadoop, а также извлечены, обработаны и проанализированы данные, выявлены закономерности, тенденции, и созданы визуализации на основе предоставленных данных, а именно исторические данные компании MLTR.