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**

**Тема:**

«Развертывание и настройка кластера Hadoop»

Выполнил: Егоров А. С., группа: АДЭУ-211

Преподаватель: Босенко Т. М.

Москва

2024

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc183108889)

[ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 4](#_Toc183108890)

[ПРАКТИЧЕСКАЯ ЧАСТЬ 5](#_Toc183108891)

[Индивидуальное задание 8](#_Toc183108892)

[ЗАКЛЮЧЕНИЕ 16](#_Toc183108893)

# ВВЕДЕНИЕ

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).

# ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

Apache Hadoop — это пакет утилит, библиотек и фреймворков для построения систем работы с большими данными. Он предназначен для сбора, хранения и распределённой обработки сотен терабайт информации, которая поступает непрерывным потоком.

У Hadoop открытый исходный код, написанный на языке Java. Это значит, что пользователи могут работать с ним и модифицировать его бесплатно.

Hadoop работает по принципу MapReduce. Система состоит из кластеров — групп отдельных подсерверов, или узлов, которые используются как единый ресурс. Когда на кластер поступает обширная задача, Hadoop делит её на много мелких подзадач и выполняет каждую на своём узле. Это позволяет параллельно решать несколько задач и быстрее выдать конечный результат.

Hadoop используется для разнообразных задач, связанных с обработкой и анализом больших объёмов данных в ретейле, финансах, здравоохранении, науке.

# ПРАКТИЧЕСКАЯ ЧАСТЬ

1. Войдем под пользователем «hadoop»

Screenshot_1

Рисунок 1 – Вход под пользователем «hadoop»

1. Запустим Hadoop.

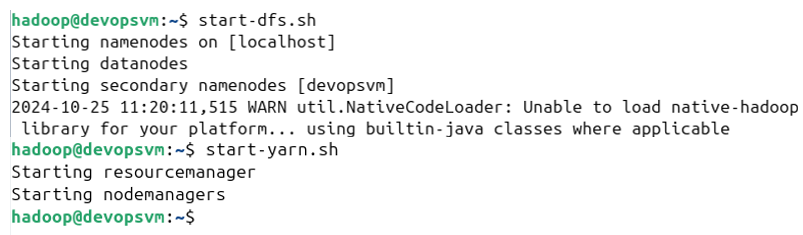
****

Рисунок 2 – Запуск Hadoop

1. Проверяем, что Hadoop успешно запустился.

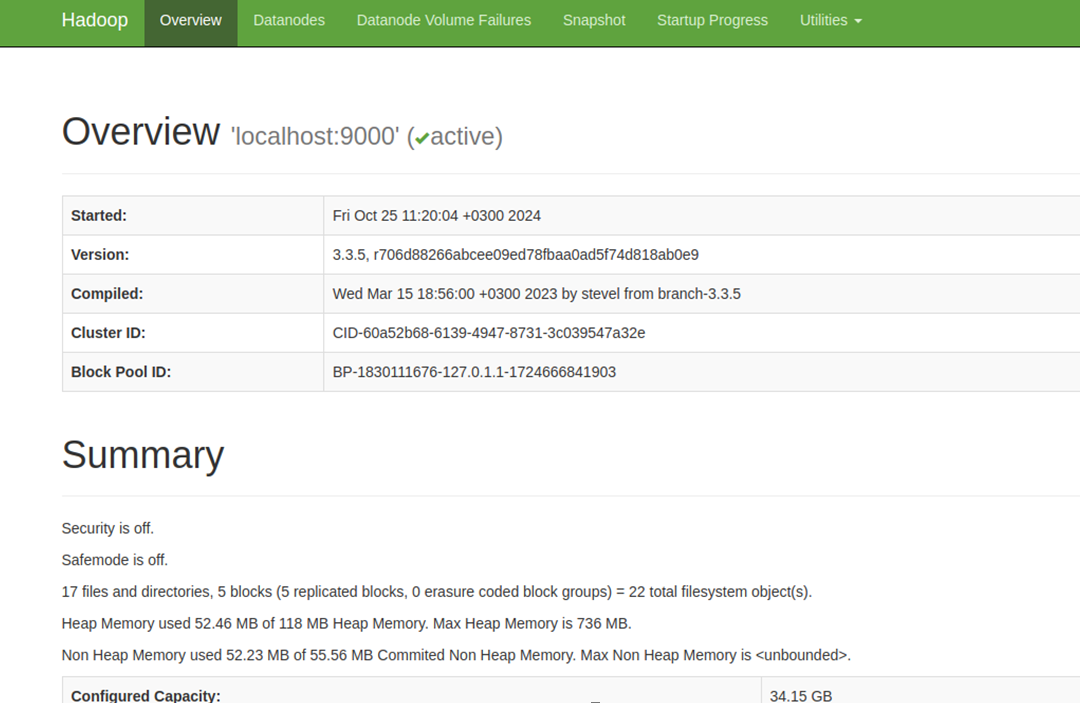


Рисунок 3 – Проверка запуска Hadoop

1. Просмотрим директорию.

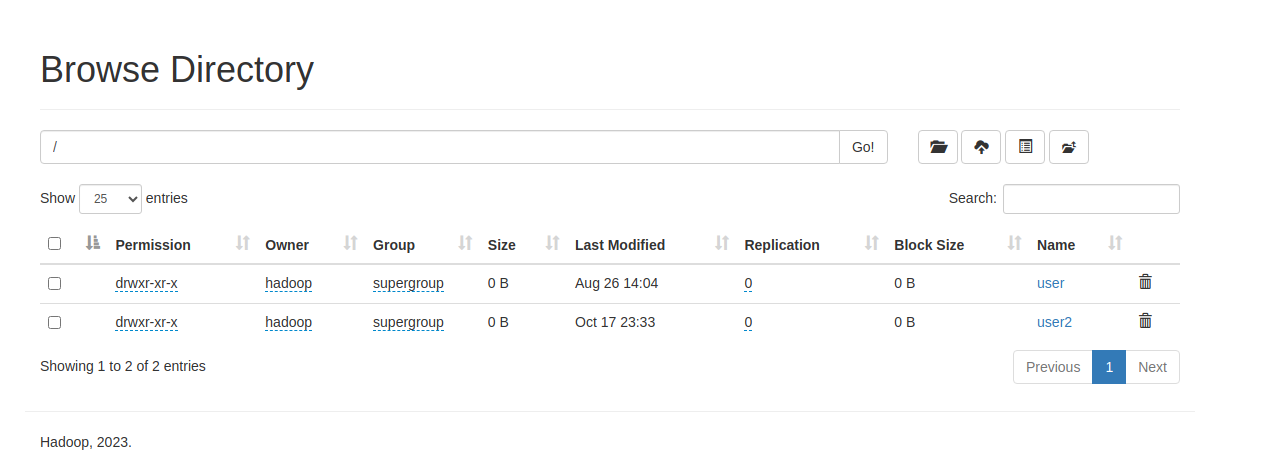


Рисунок 4 – Директория

1. Создадим директорию user4 в HDFS.



Рисунок 5 – Создание директории

1. Загрузим данные в HDFS через гитхаб.



Рисунок 6 – Загрузка данные в HDFS

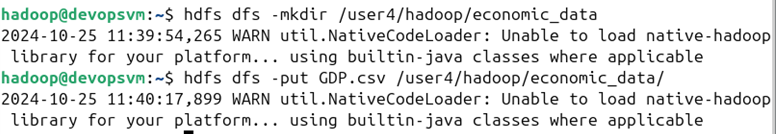


Рисунок 7 – Загрузка данные в HDFS (продолжение)

1. Предоставим другим пользователям доступ на изменения данных

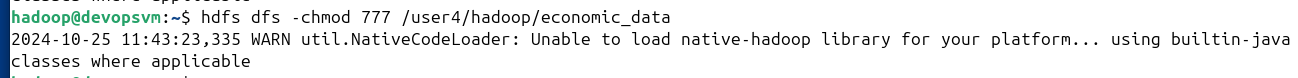


Рисунок 8 – Предоставление прав на изменение данных

## Индивидуальное задание

**Вариант 4.** Данные: Исторические данные по акциям Яндекса (YNDX) с сайта Московской биржи (moex.com)

Операции: Фильтрация данных за 2021 год, расчет средней цены закрытия, тренд анализа.

1. Надо загрузить таблицу в HDFS. Для Начала загружаем в гитхаб, чтобы узнать ссылку, с помощью которой будет загружаться уже в Hadoop.

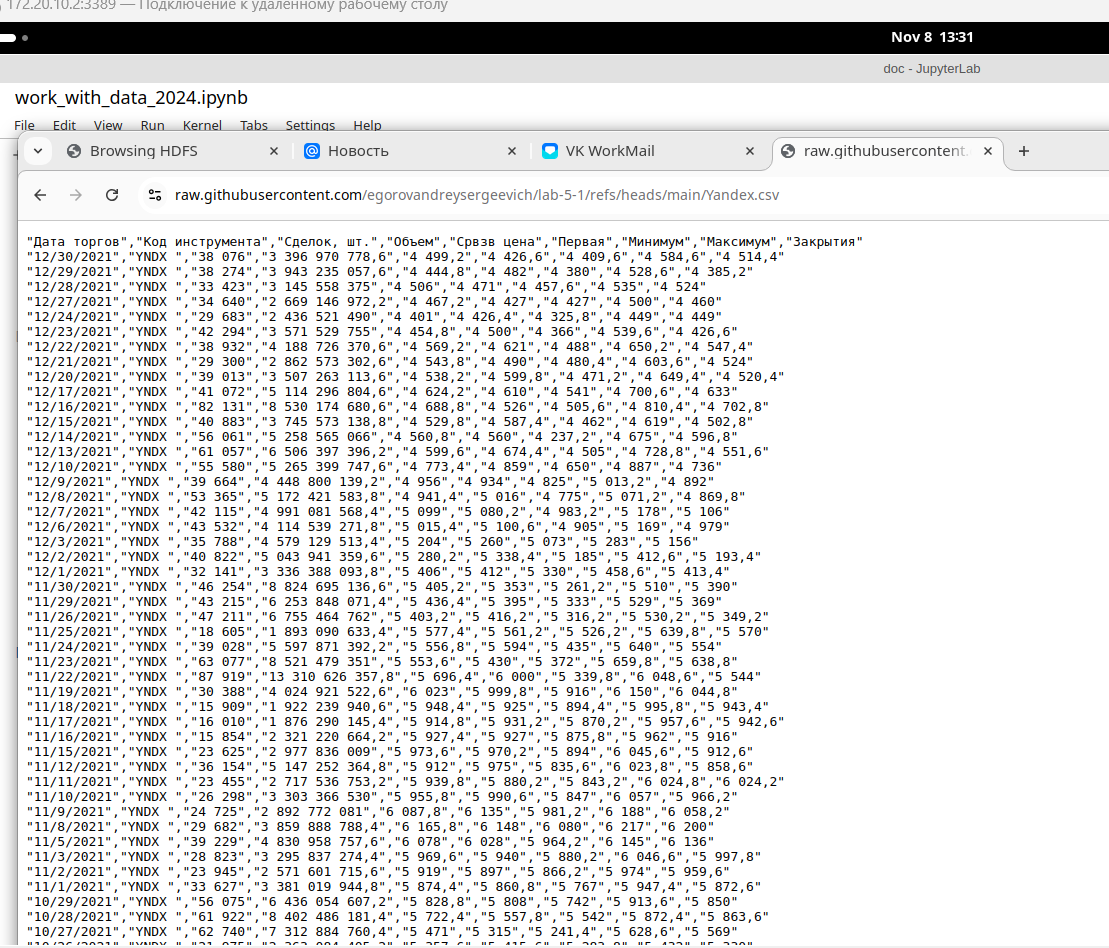
****

Рисунок 9 – Таблица, загруженная на Гитхаб

1. Загрузим таблицу в Hadoop с помощью ссылки из гитхаба.

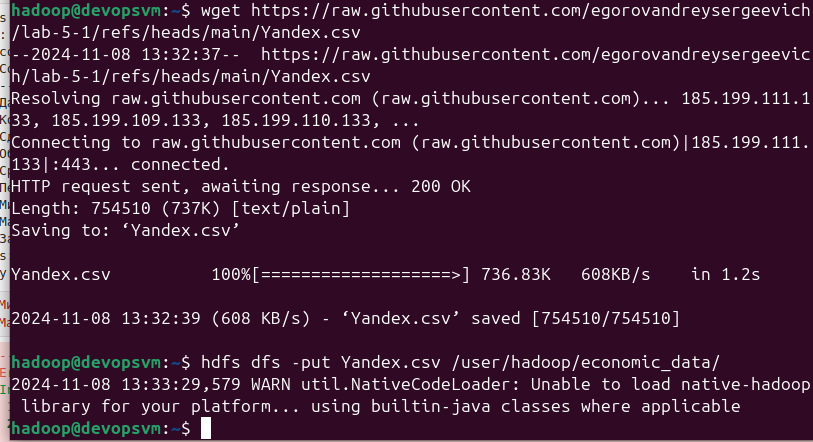
****

Рисунок 10 – Загрузка таблицы в Hadoop через Гитхаб

1. Проверяем, что все успешно загрузилось в Hadoop.

****

Рисунок 11 – Проверка успешности загрузки

1. Установим необходимые библиотеки для подключения к Hadoop.

****

Рисунок 12 – Установка необходимых библиотек

1. Прочитаем данные из таблицы и выведем первые 10 строк. Преобразуем в датафрейм.

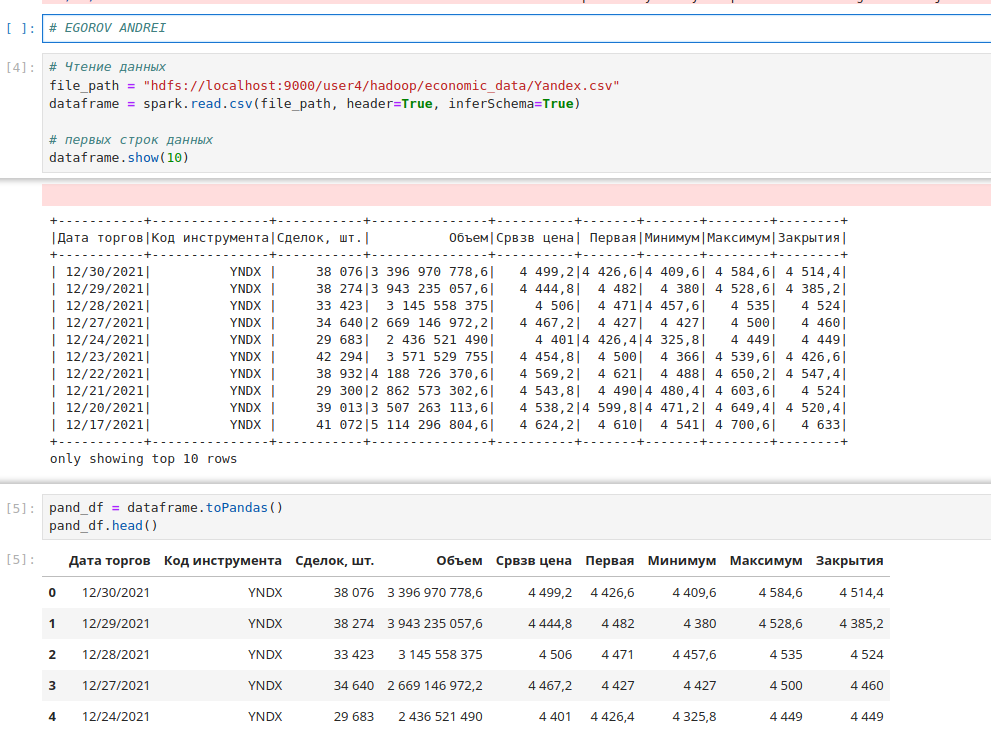
****

Рисунок 13 – Просмотр датафрейма

1. Отфильтруем данные по году (нужен 2021) и проверим их типы.

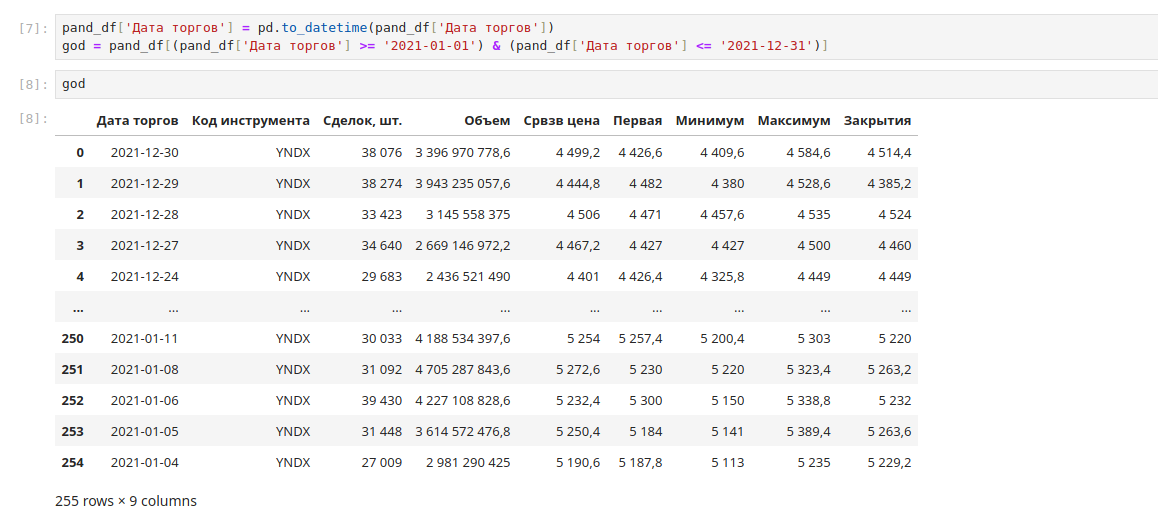
****

Рисунок 14 – Фильтр данных за 2021 год

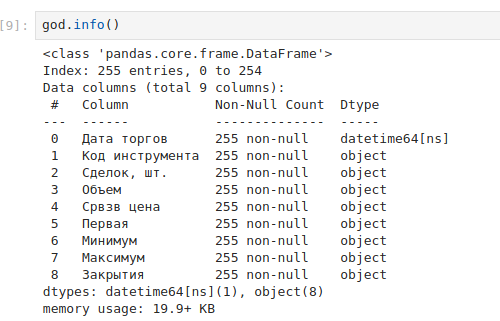
****

Рисунок 15 – Просмотр типов данных

1. На рисунки видно, что все данные типа object. Так как у нас числовые данных, изменим тип данных на float. Дата уже имеет нужный тип datetime64, а код инструмента оставим по умолчанию object.

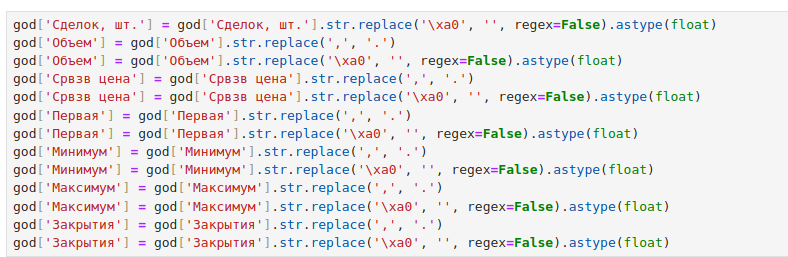
****

Рисунок 16 – Изменение типов данных

1. Проверим изменение типов данных.

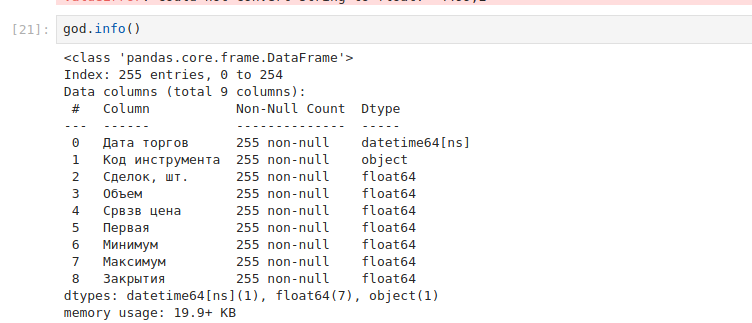
****

Рисунок 17 – Проверка изменений типов данных

1. Проведем статистический анализ с помощью функции describe.

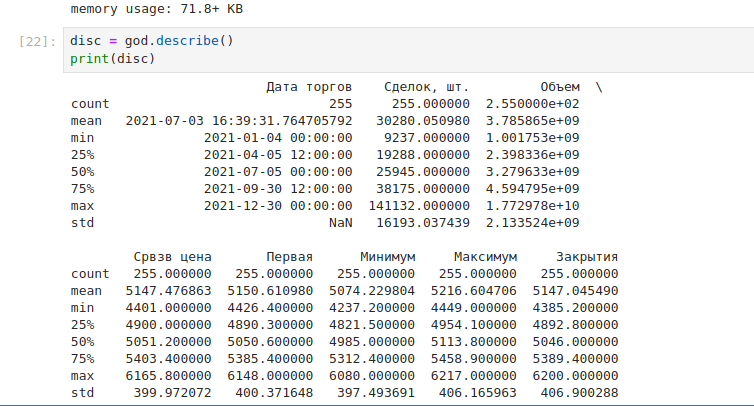
****

Рисунок 18 – Статистический анализ

1. Рассчитаем среднюю цену закрытия.

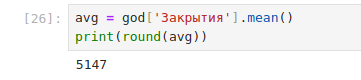
****

Рисунок 19 – Средняя цена закрытия

Итого средняя цена закрытия – 5147 за 2021 год. Минимальная за год была 4237, а максимальная 6217.

1. Изобразим на графике тренд закрытия.

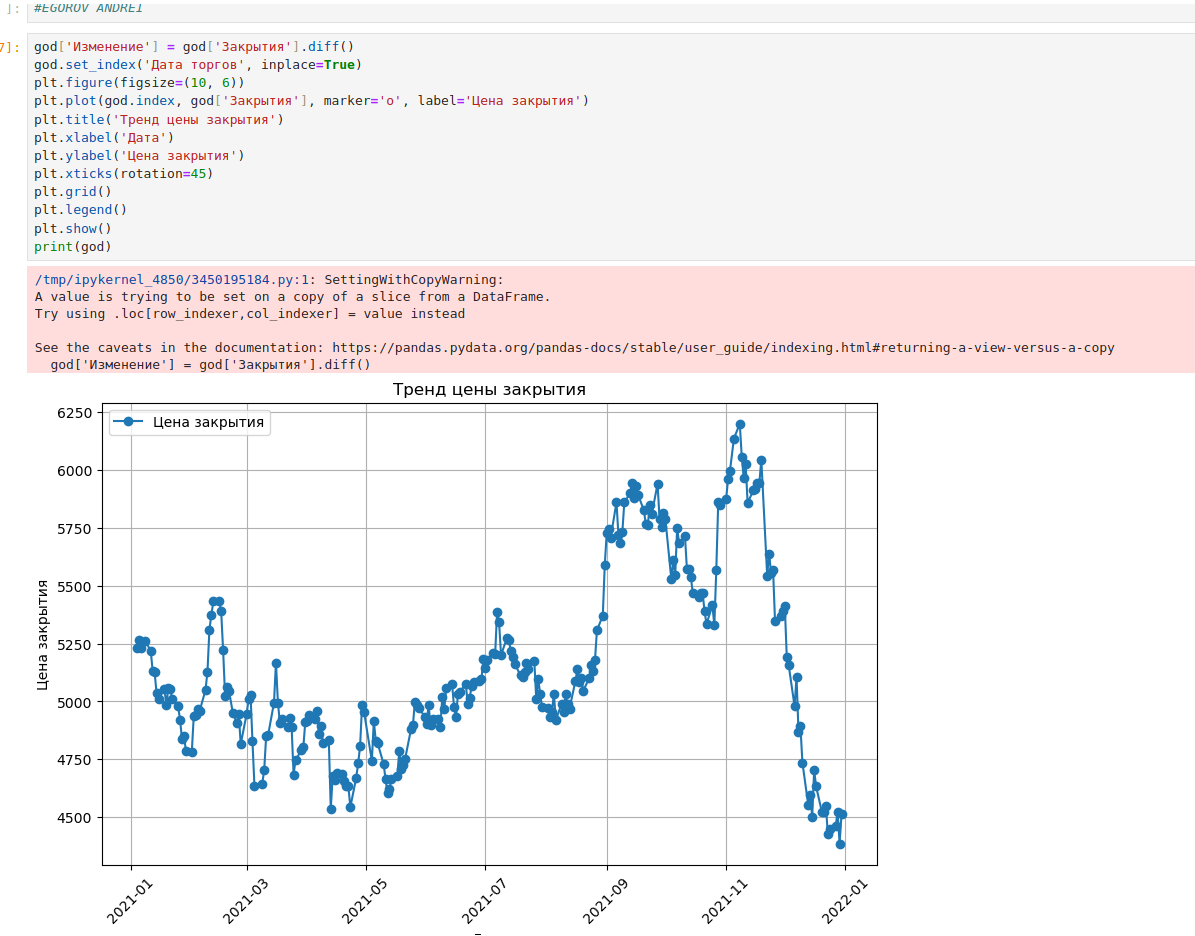
****

Рисунок 20 – График тренда закрытия

На графике видно, что цена закрытия постоянно изменяется, есть резкие скачки, особенно в конце года. Цены на акции российских компаний в конце 2021 года снижались из-за новостей-слухов о начале СВО, которая все же началась позже.

1. Сгруппируем данные по месяцам для построения дальнейшей визуализации.

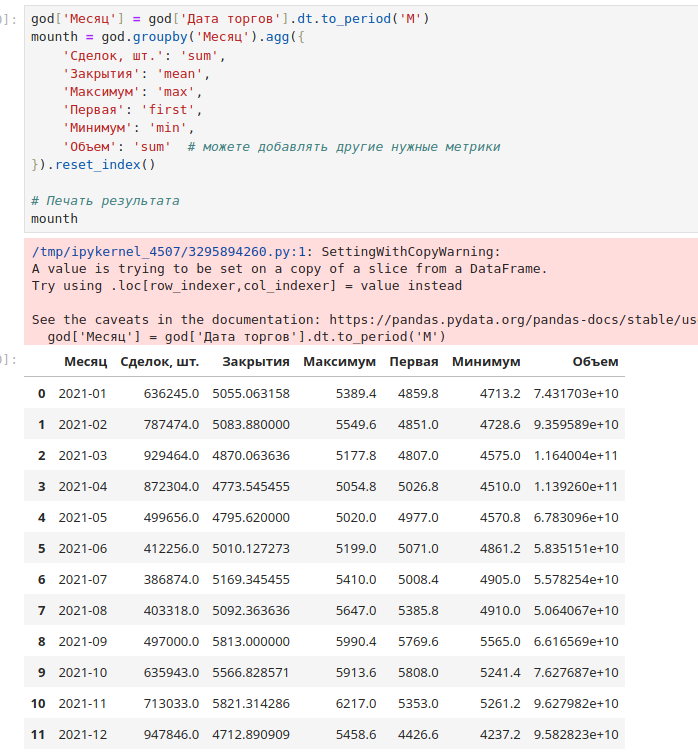
****

Рисунок 21 – Группировка данных по месяцам

1. Построим график распределения количества сделок по месяцам.

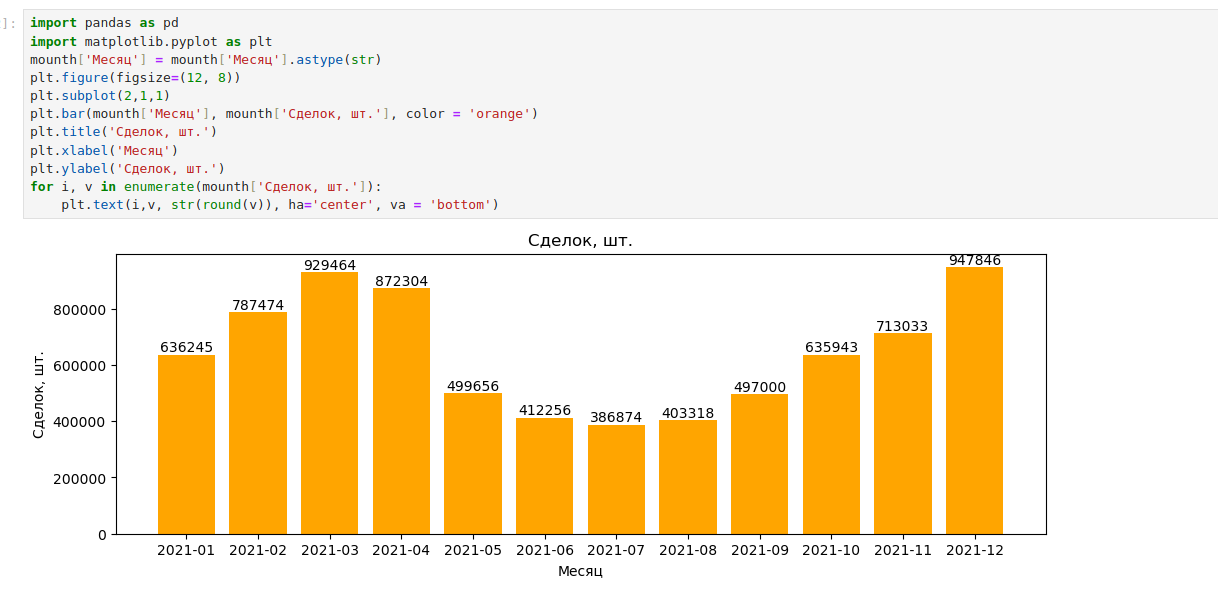
****

Рисунок 22 – График количества сделок по месяцам

На графике видно, что наибольшее количества сделок – февраль, март, апрель и декабрь. 16 февраля был опубликован отчет Яндекса, после чего начались активные сделки с акциями. В декабре многие активно продавали акции российских компаний из-за скорого начала СВО, санкций и снижения российского рынка.

1. Построим график максимальной цены акции в каждом месяце.

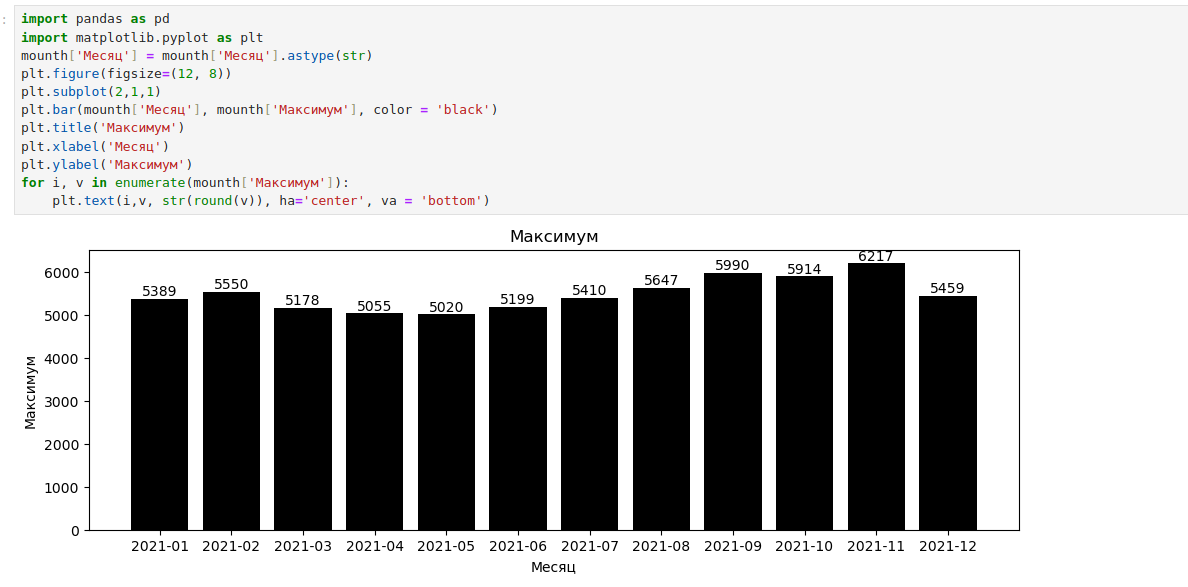
****

Рисунок 23 – График максимальной цены по месяцам

На графике видно, что произошло резкое падение максимальной цены акции на 758 рублей в декабре

1. Построим график минимальной цены акции по месяцам.

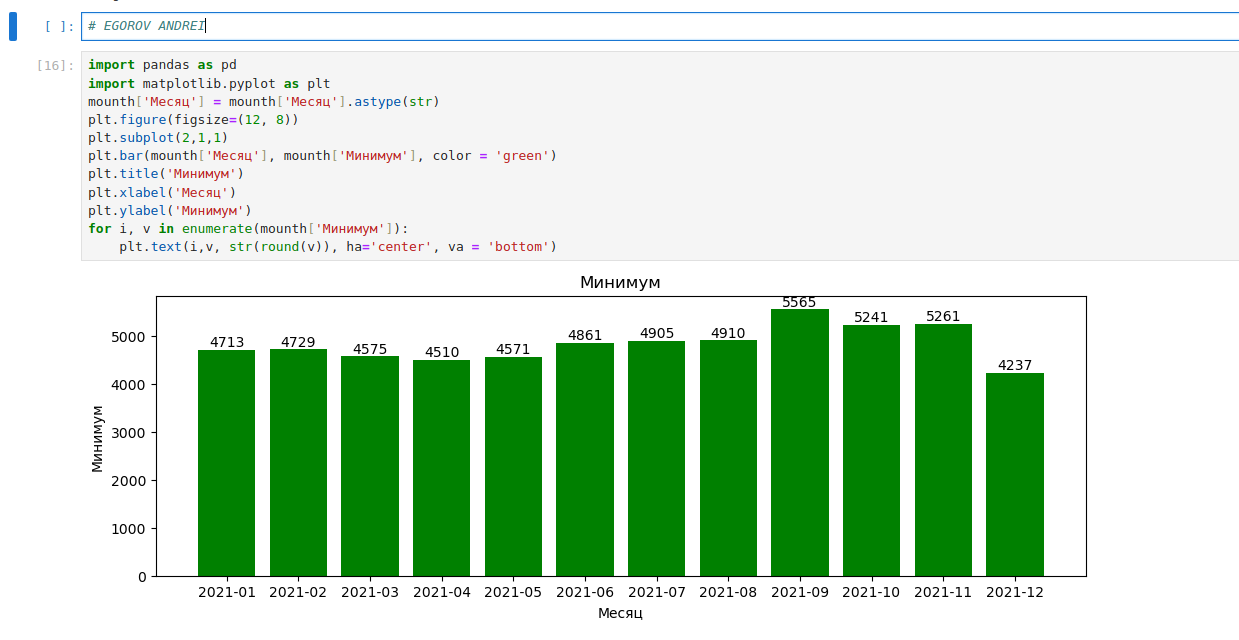
****

Рисунок 24 – График минимальной цены по месяцам

На графике видно, что вслед за падением максимальной цены в дкабре, произошло падение и минимальной.

# ЗАКЛЮЧЕНИЕ

Hadoop представляет собой мощный инструмент для эффективной обработки и анализа огромных объемов данных. Он способен обеспечить масштабируемость, устойчивость к сбоям, экономическую эффективность и доступность для различных приложений. В рамках работы была установлена система Apache Hadoop, в ходе которой данные были извлечены, обработаны и проанализированы, что позволило выявить тренды и проанализировать цену акций Яндекса. На основе проанализированных данных также были созданы визуализации.

В 2021 году цена акций Яндекса колебалась в зависимости от различных факторов, таких как финансовые результаты компании, изменения в отрасли интернет-технологий, макроэкономическая обстановка и другие внешние факторы. Несмотря на нестабильность рынка, акции Яндекса в целом показали положительную динамику в 2021 году, только в конце года появился резкий тренд на снижение акций.