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**»**

Выполнила: Сергеева А. И., группа: АДЭУ-211

Преподаватель: Босенко Т. М.

Москва

2024

**Цель работы:** 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).

**Ход работы**

**Вариант 12**

Объем данных постоянно растет, как и количество решений для хранения и обработки больших данных. Hadoop — это фреймворк с открытым исходным кодом, разработанный для хранения и обработки больших объемов данных. Он позволяет распределять данные и вычисления между множеством узлов в кластере, обеспечивая масштабируемость и устойчивость к сбоям. Apache Spark — это другой фреймворк с открытым исходным кодом, который используется для обработки больших данных, но в отличие от Hadoop, он ориентирован на быструю обработку данных в реальном времени. Spark может работать поверх Hadoop, используя Hadoop Distributed File System (HDFS) для хранения данных. Это позволяет Spark использовать все преимущества экосистемы Hadoop, что и будет продемонстрировано в данной лабораторной работе, данные будут храниться в HDFS, а обрабатываться с помощью Spark и Pandas.

Для начала нужно было переключиться на пользователя hadoop и запустить сценарии для запуска служб HDFS и YARN в кластере Hadoop (рисунок 1).

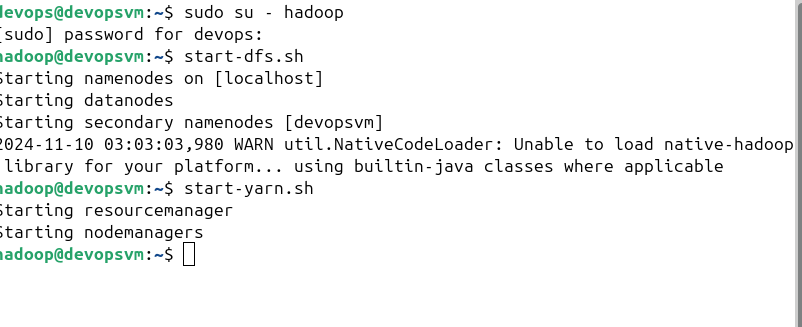


Рисунок 1 – Запуск сценариев для запуска служб HDFS и YARN в кластере Hadoop

В рамках варианта нужно было выгрузить данные с сайта московской биржи с данными по акциям за 2020 год для компании Сургутнефтегаз (SNGS) и сгенерировать для данных дополнительно миллион строк для демонстрации возможностей HDFS для хранения больших данных и Spark для их обработки (рисунки 2-3).

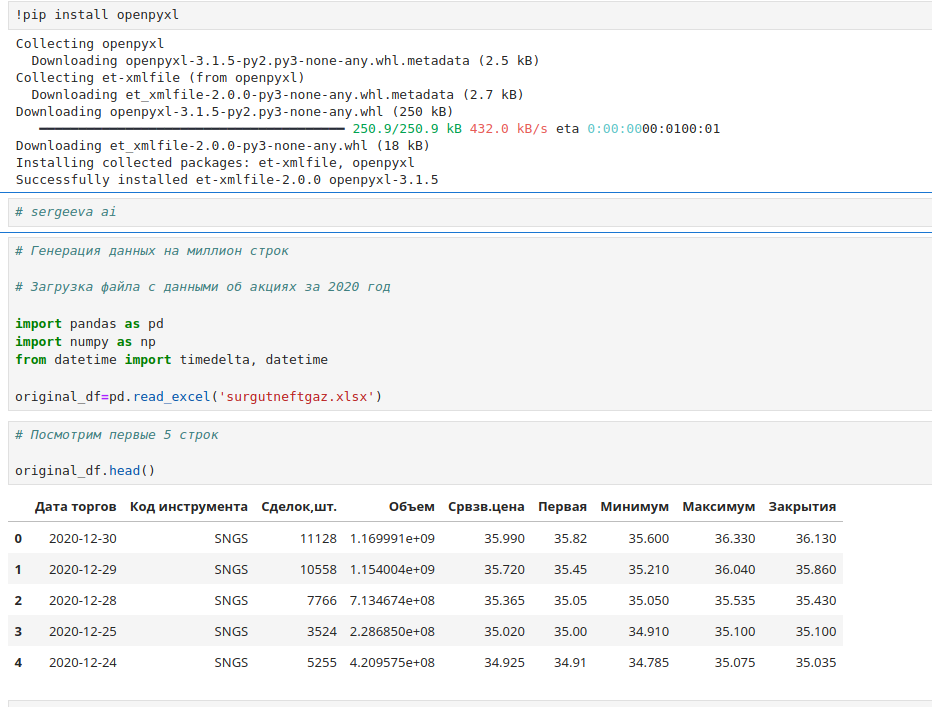


Рисунок 2 – Загрузка данных по акциям за 2020 год для SNGS

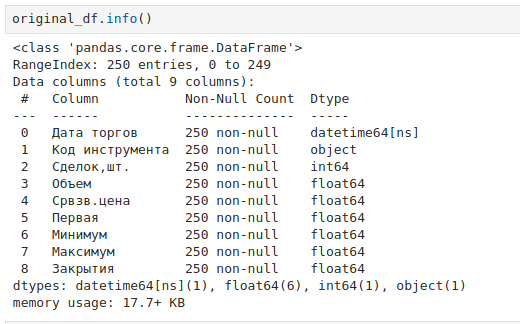


Рисунок 3 – Типы данных

Все типы данных исходного файла в порядке. Далее был осуществлен поиск дубликатов и проведена описательная статистика, дублей нет, всего 210 записей, минимальное количество сделок 3.524, максимальное 95.248, при этом среднее их число 20.262, с другими расчетами можно ознакомиться на рисунке 4.

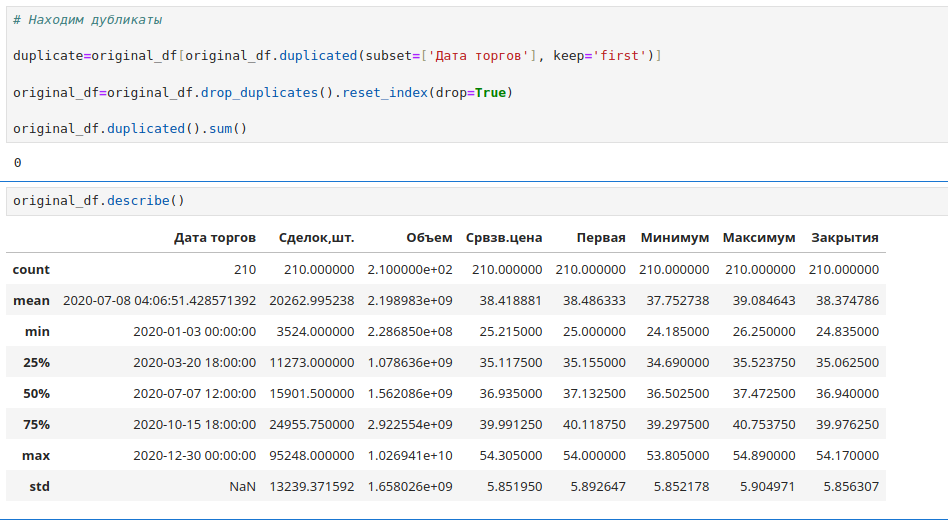


Рисунок 4 – Поиск дублей и описательная статистика

Далее была произведена генерация строк, код инструмента дублируется, т. к. смотрим по одной компании, задаем размеры для генераций в 1 миллион, при этом сделки генерируются в целые числа, дата в тип даты, а другие данные генерируются в числа с плавающей запятой (рисунки 5-6).

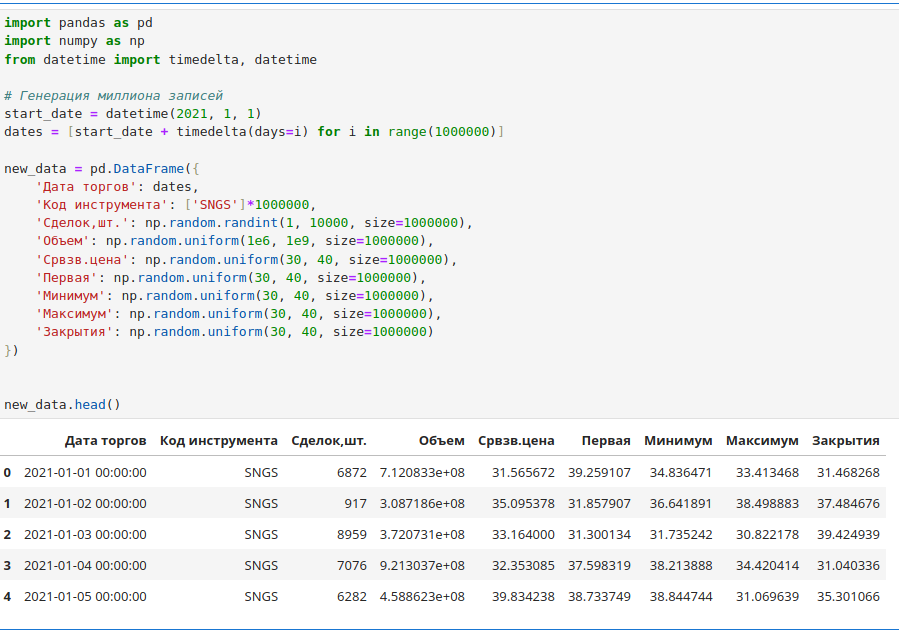


Рисунок 5 – Генерация миллиона строк

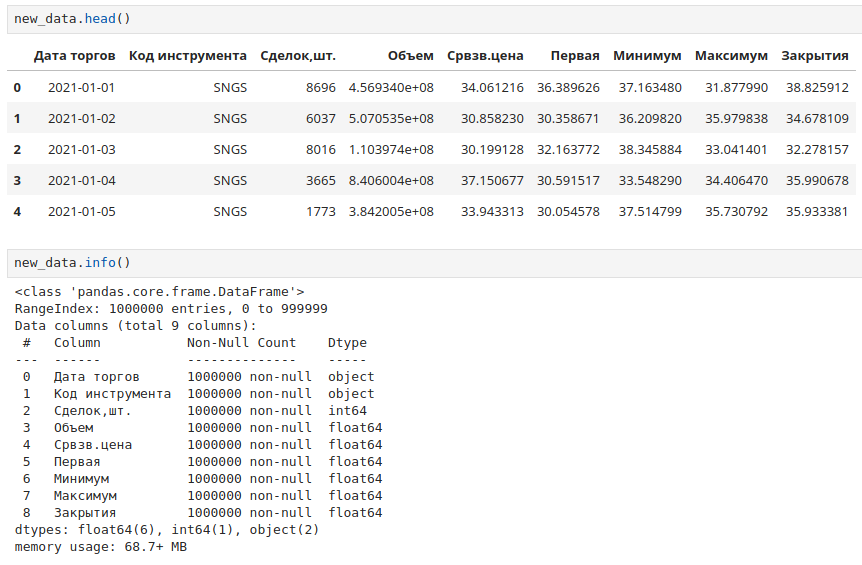


Рисунок 6 – Просмотр сгенерированных данных и информации о типах данных столбцов

Дата в сгенерированных данных имеет тип объекта, при этом каждая дата уникальна, т. к. не может за один день быть разная информация об акциях. Даты выходят за допустимый диапазон для типа даты в Pandas, ведь генерировался миллион уникальных записей по дате, поэтому их стоит оставить с типом данных объекта (есть данные, которые больше 11 апреля 2262 года). Датафреймы со сгенерированными и исходными данными были объединены в один, для этого в исходном датафрейме даты были преобразованы в тип объекта (рисунок 7).

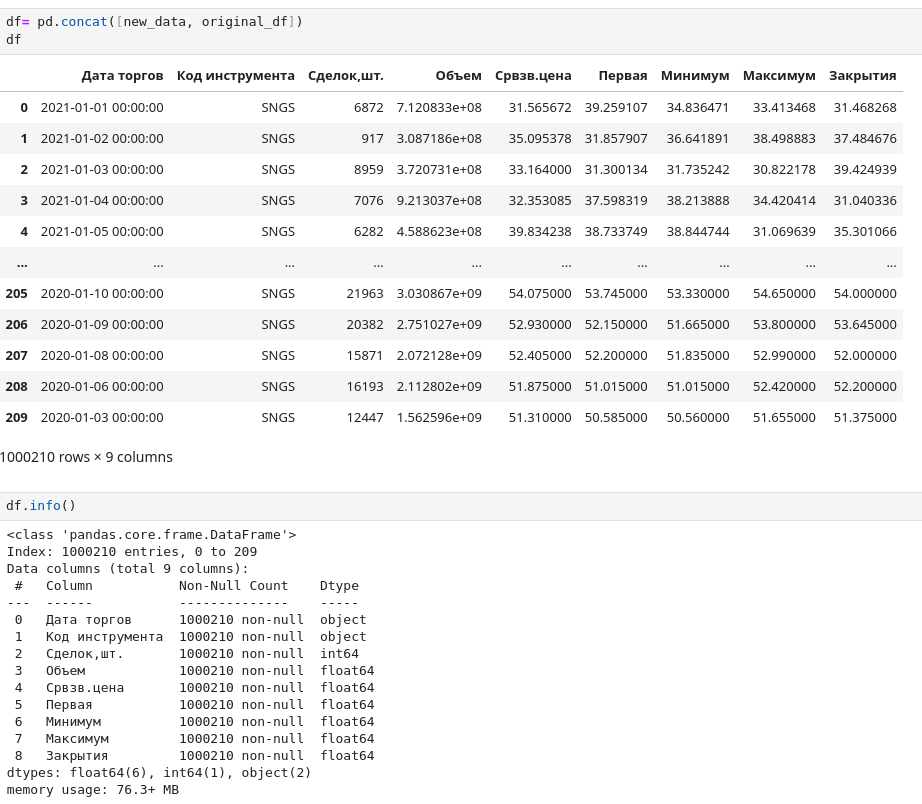


Рисунок 7 – Объединение таблиц и просмотр информации о столбцах

Далее файл был загружен в HDFS через Spark. Всего 1.000.210 строк. Была задана схема данных с указанием типов данных, выбран каталог для хранения economic\_data (рисунок 8).



Рисунок 8 – Загрузка файла с акциями в HDFS

Далее файл был выгружен из HDFS (рисунок 9).

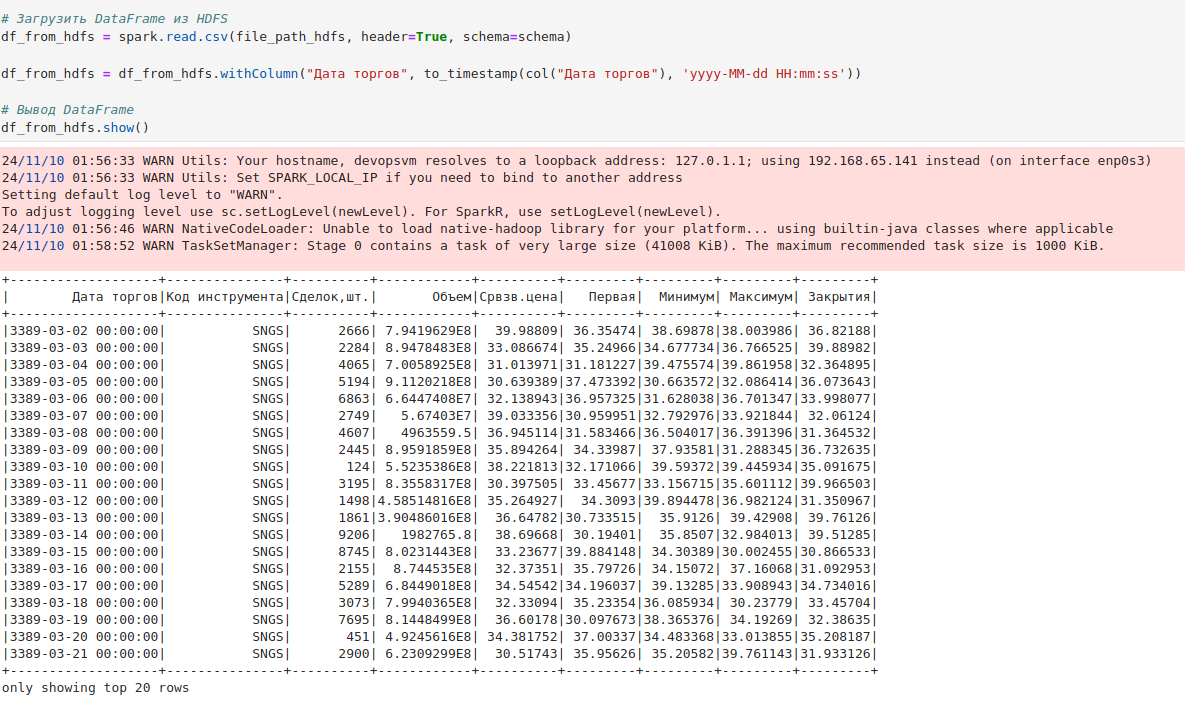


Рисунок 9 – Выгрузка файла с акциями из HDFS

Для анализа 2020 года данные были отфильтрованы и просмотрены типы данных для столбцов (рисунок 10).

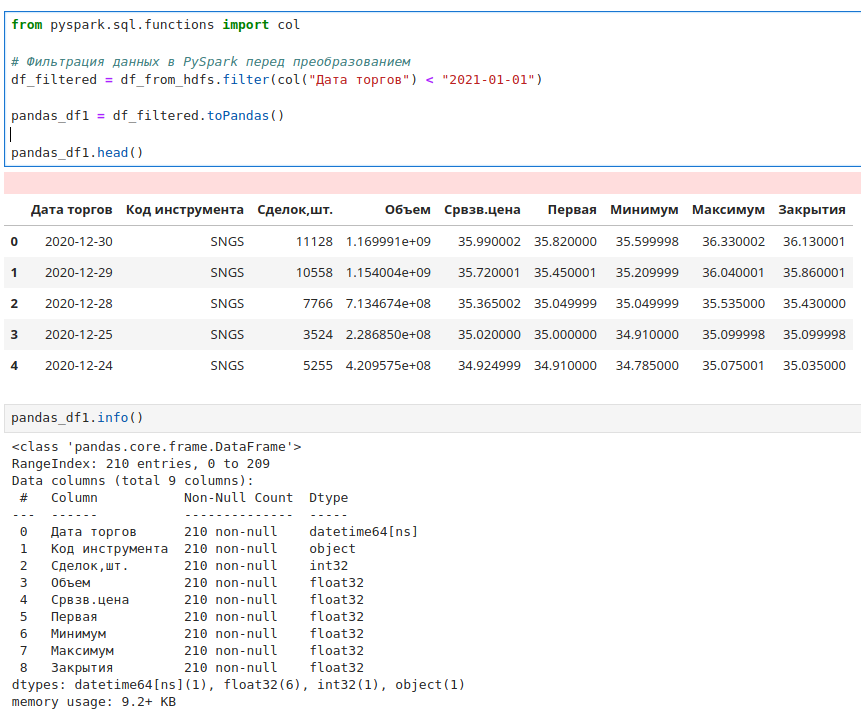


Рисунок 10 – Срез данных по акциям за 2020 год

Далее была проведена описательная статистика и выгружена библиотека для визуализации Altair (рисунок 11).

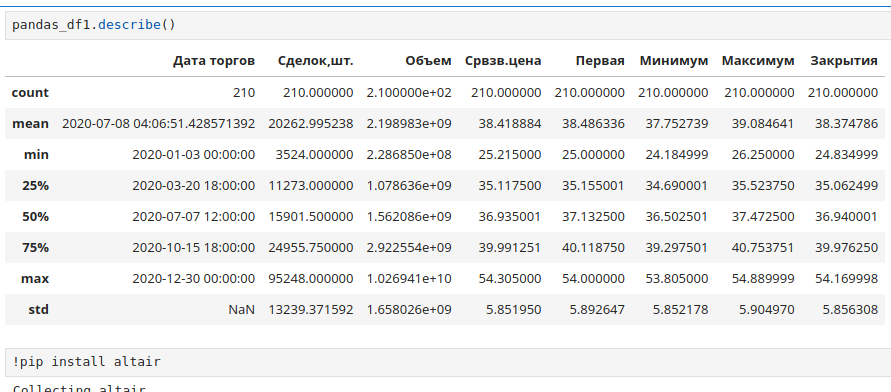


Рисунок 11 – Описательная статистика для выгруженных из HDFS данных

Цена закрытия является стандартным ориентиром, который используется инвесторами для отслеживания динамики акций с течением времени, поэтому был построен график для демонстрации динамики цены закрытия за год (рисунок 12).



Рисунок 12 – График динамики цены закрытия за год

Видно, что в период с февраля по апрель у акции наблюдается сильный спад, далее с небольшими колебаниями цена закрытия акции имеет тенденцию чуть снижаться.

Данный график был объединен с графиками по скользящим средним. Скользящие средние сглаживают ценовые колебания, что помогает определить тренды. Для анализа краткосрочных трендов я построила 10-дневную SMA, а для долгосрочных трендов 50-дневная (рисунок 13).

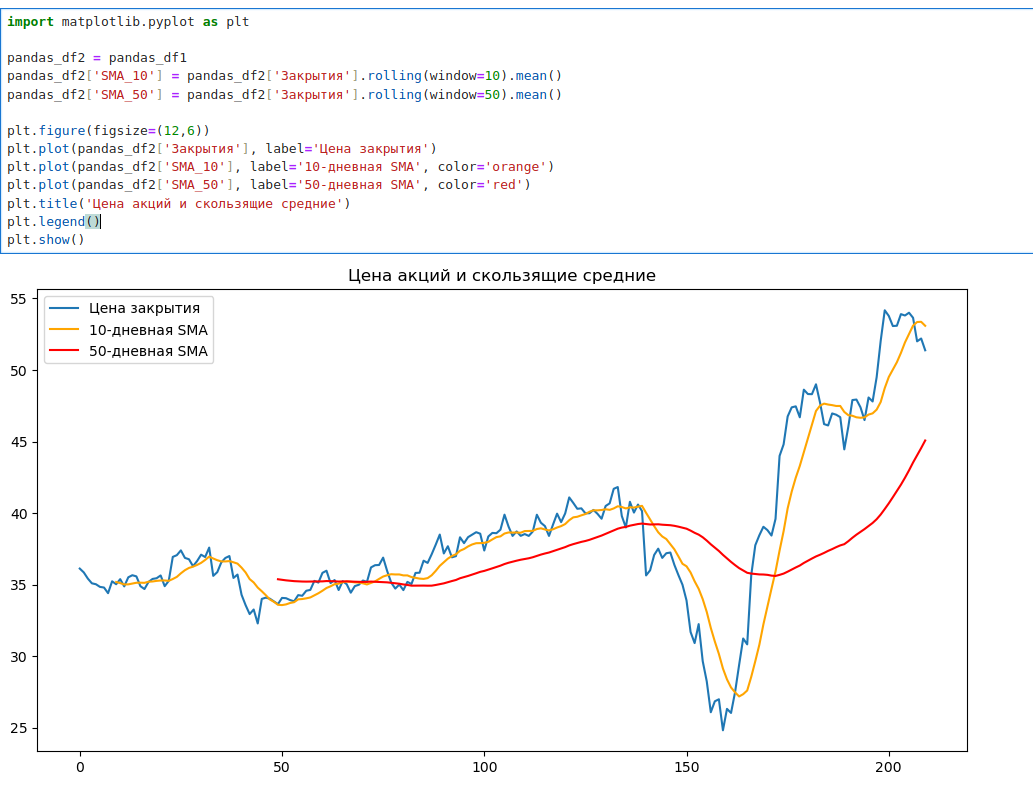


Рисунок 13 – График для цены акций и скользящих средних

Ближе к концу 2020 года наблюдаются нисходящие тренды в краткосрочной и долгосрочной перспективах, в другие периоды тренды восходящие, что говорит о спаде цен ближе к концу года.

Далее похожий график с динамикой на год был построен для объема сделок (рисунок 14).

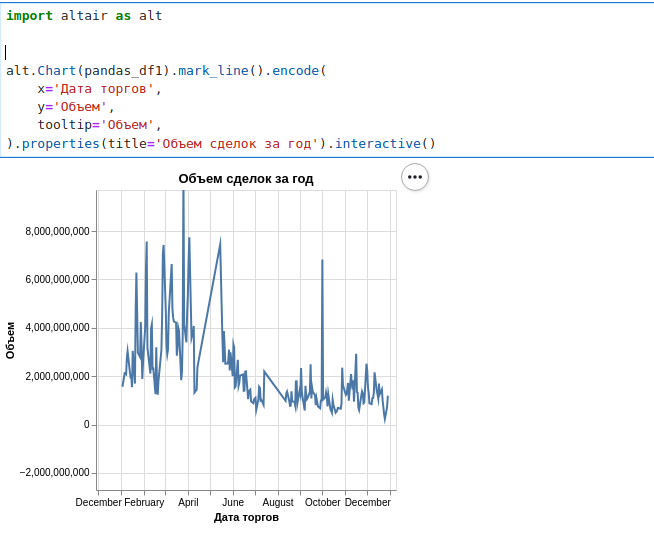


Рисунок 14 – График динамики объема сделок за 2020 год

Резкие изменения объема сделок наблюдаются в начале года и осенью, объем имеет постоянные колебания.

Была построена диаграмма рассеяния для объема сделок и цены закрытия (рисунок 15).

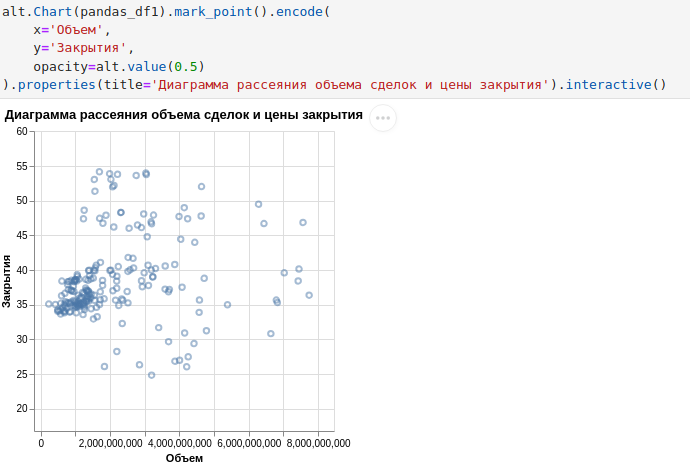


Рисунок 15 – Диаграмма рассеяния объема сделок и цены закрытия

Больше всего сделок в объеме до 2000 и с ценой до 40. При этом чем больше объем, тем меньше цена.

Далее был построен график волатильности за год (рисунок 16).

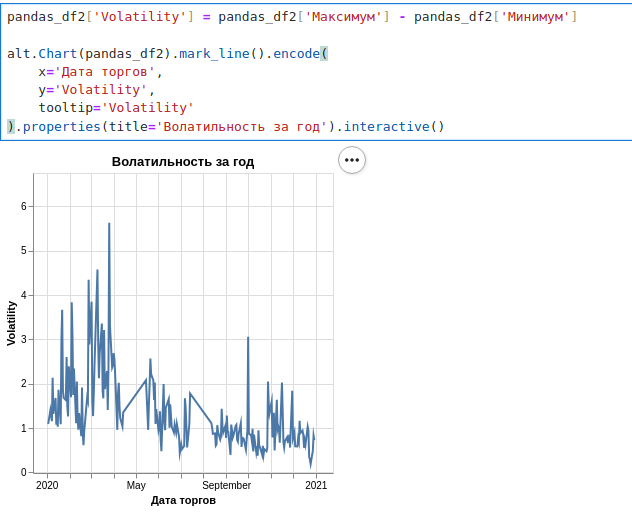


Рисунок 16 – Волатильность за год

Волатильность скачкообразная, при этом в середине года она более стабильная, цены из-за такой волатильности подвержены резким скачкам и вкладываться в такие акции рисковано.

Далее была построена корреляционная матрица для поиска зависимостей и рассчитано стандартное отклонение для первой цены акции (рисунок 17).

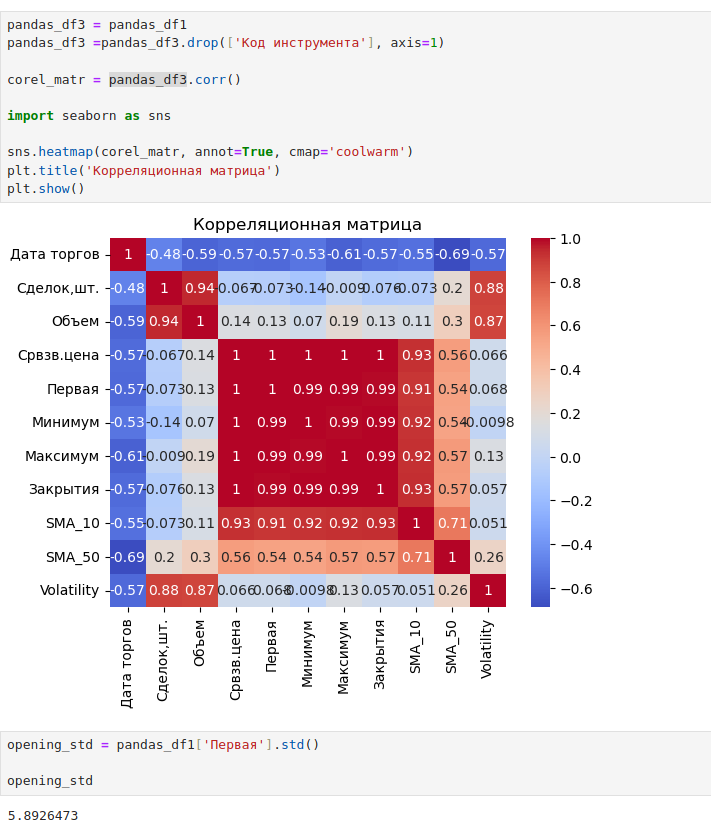


Рисунок 17 – Корреляционная матрица

С ростом даты торгов уменьшаются цены акции, с ростом сделок растет их объем и уменьшается цена акций.

**Выводы**

1. Использование HDFS (Hadoop Distributed File System) позволяет эффективно хранить и управлять большими объемами данных благодаря распределенной архитектуре. Это особенно актуально для работы с данными акций, которые могут образовывать большие массивы информации за длительные промежутки времени.
2. Spark предоставляет мощные инструменты для параллельной обработки данных, что позволяет существенно ускорить анализ и обработку больших массивов информации. Это особенно заметно при выполнении операции фильтрации данных, когда брался только 2020 год, а также при получении и загрузке данных в HDFS.
3. Анализ данных по акциям за 2020 год показал спад цены акций к концу года, волатильность скачкообразность цен и риски при вкладывании денег в них. Также была изучена зависимость объема продаж с ценой акции и рассмотрены долгосрочные и краткосрочные тренды изменения цены акций.