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:

Распределенные системы

**Практическая работа №1**

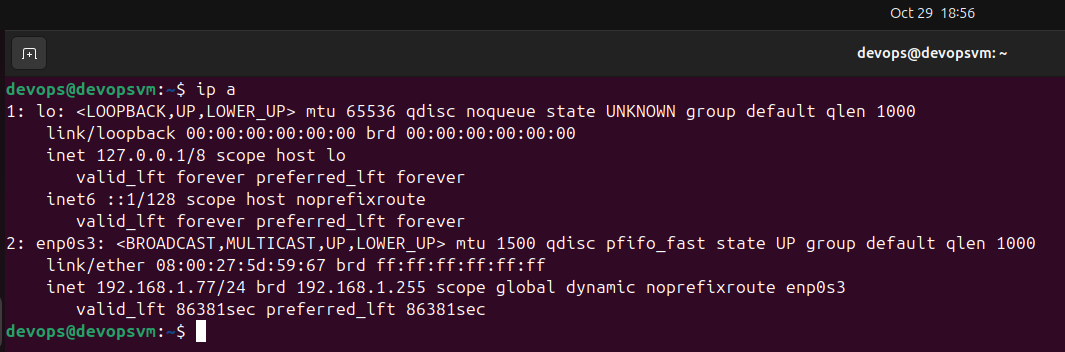
Выполнила: Сабитова А.Б., группа: АДЭУ-221

Преподаватель: Босенко Т. М.

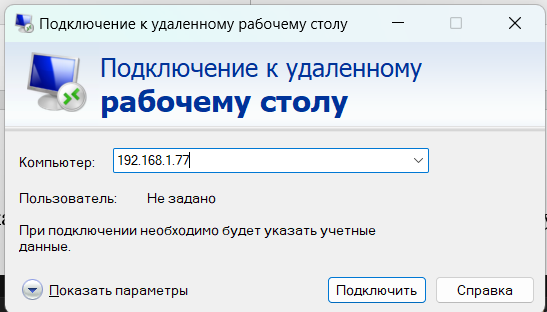
Москва

2024

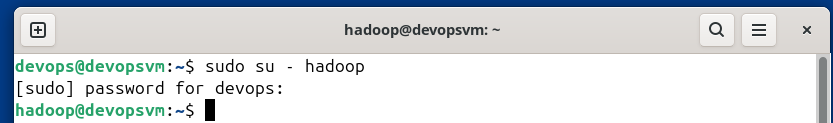
1. Запускаем виртуальную машину, и с помощью команды ip a узнаем ее ip-адрес



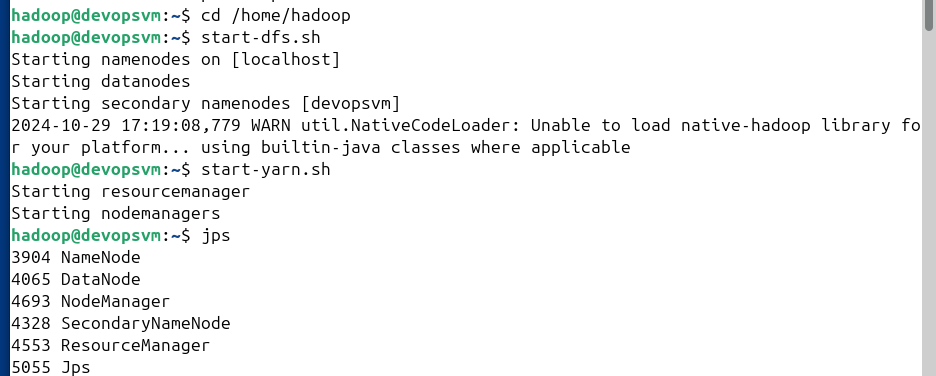
2. Подключаемся к удаленному рабочему столу:



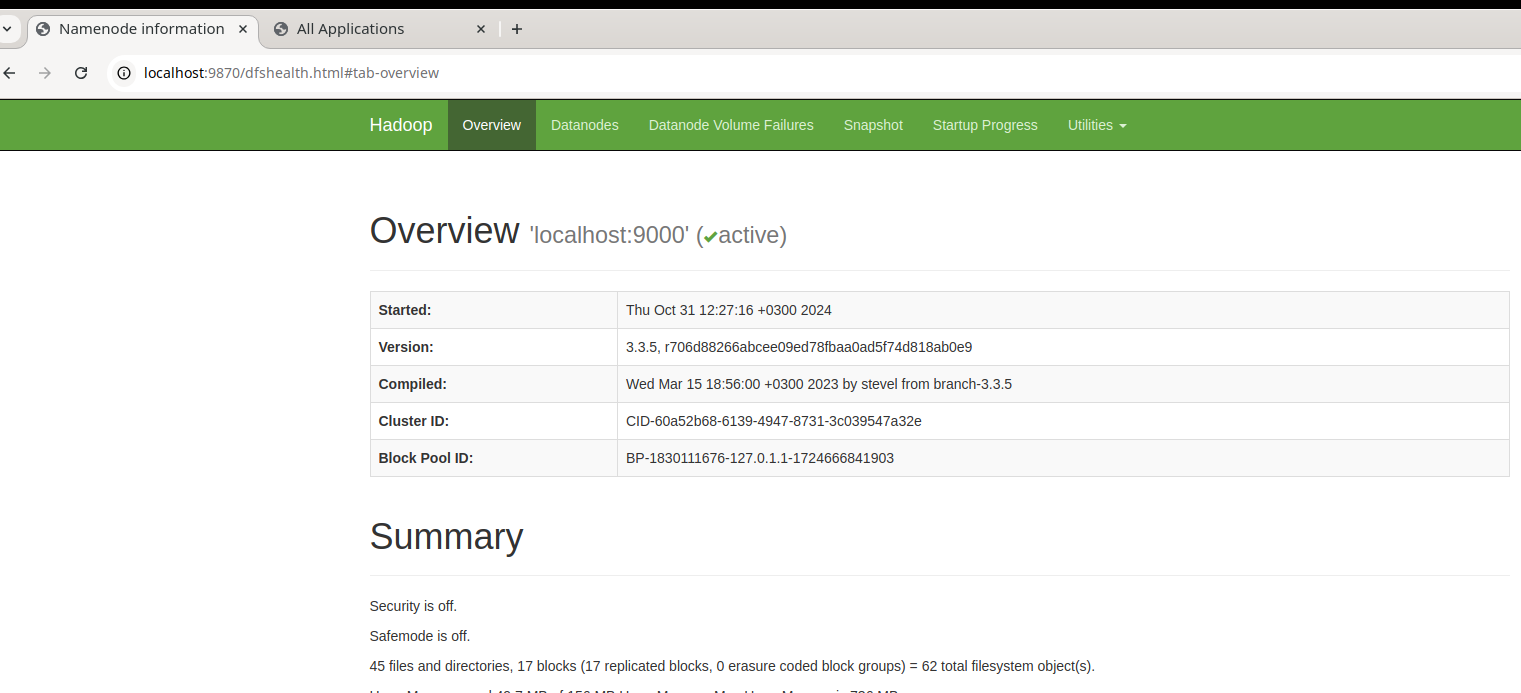
3. Переключаемся на пользователя hadoop

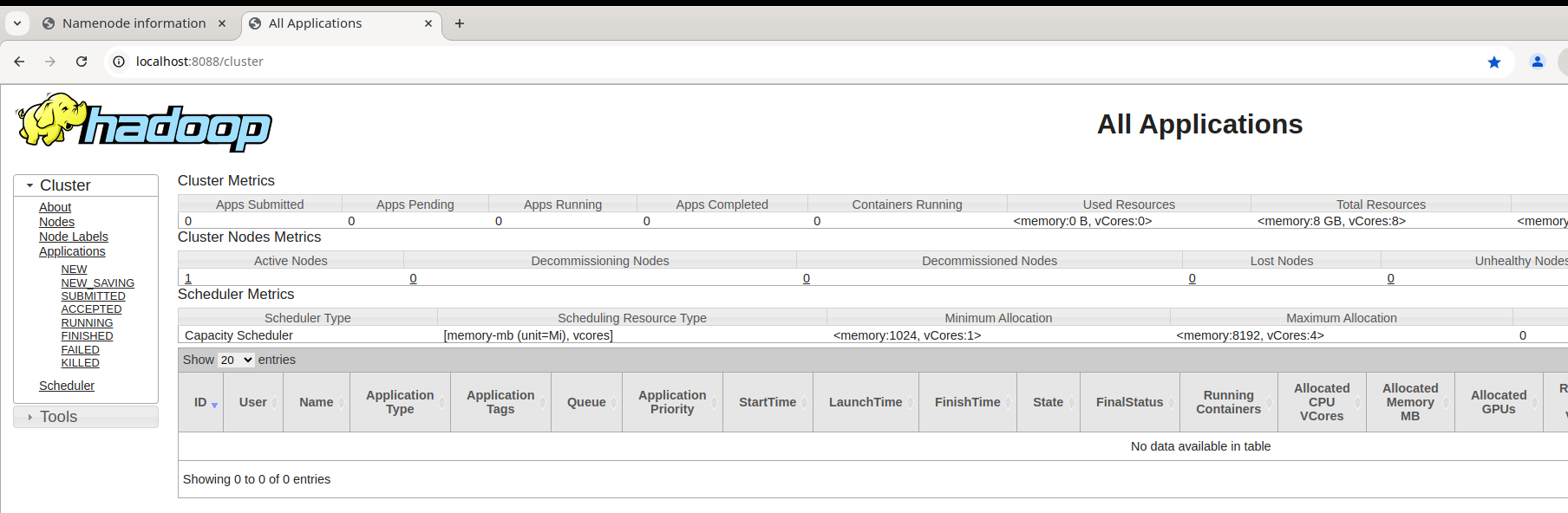


4. Запускаем HDFS и YARN, проверяем их работу

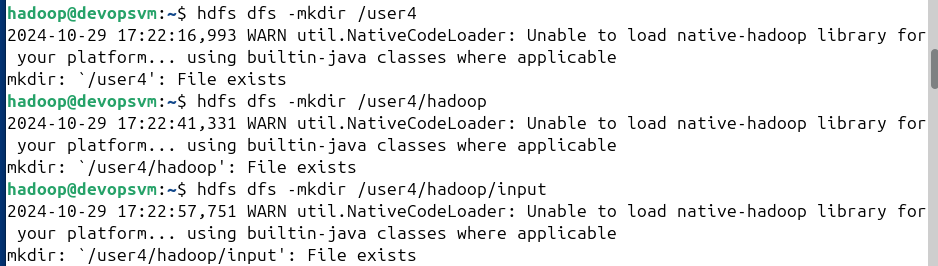


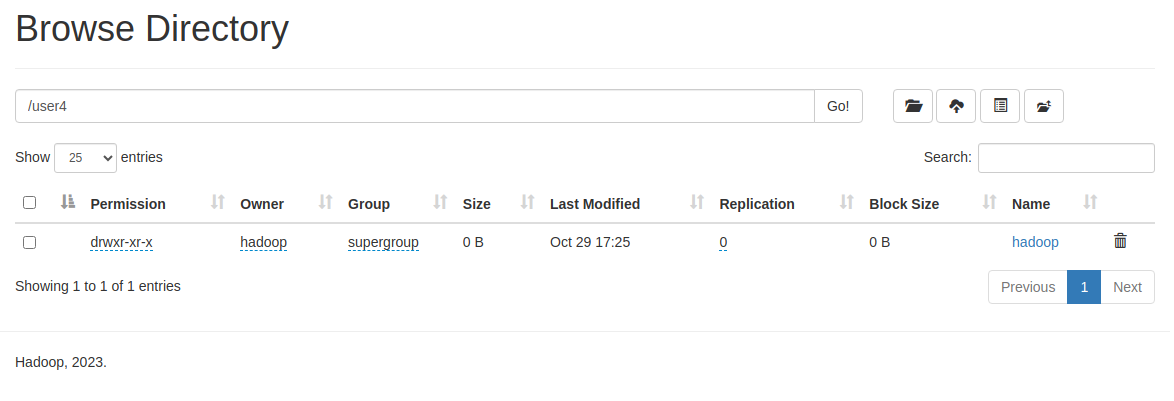
5. Запускаем их в браузере



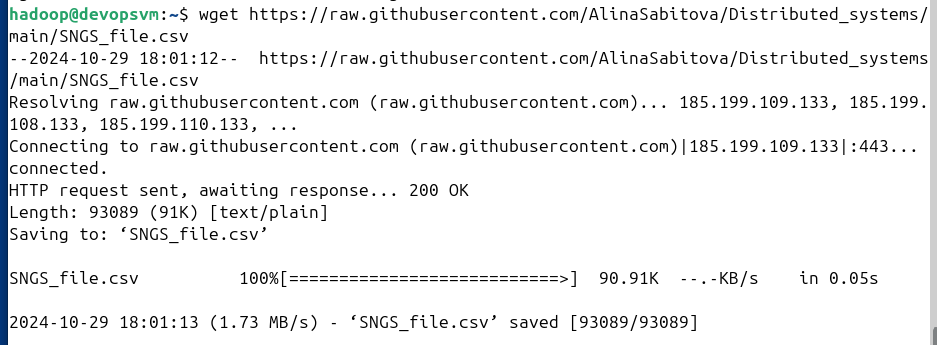


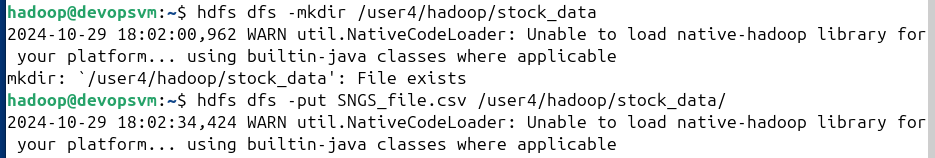
6. Создадим директорию в HDFS



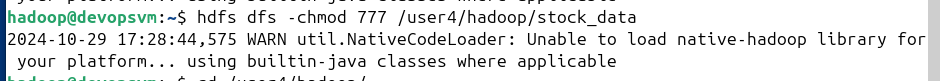


7. Загружаем данные в HDFS.

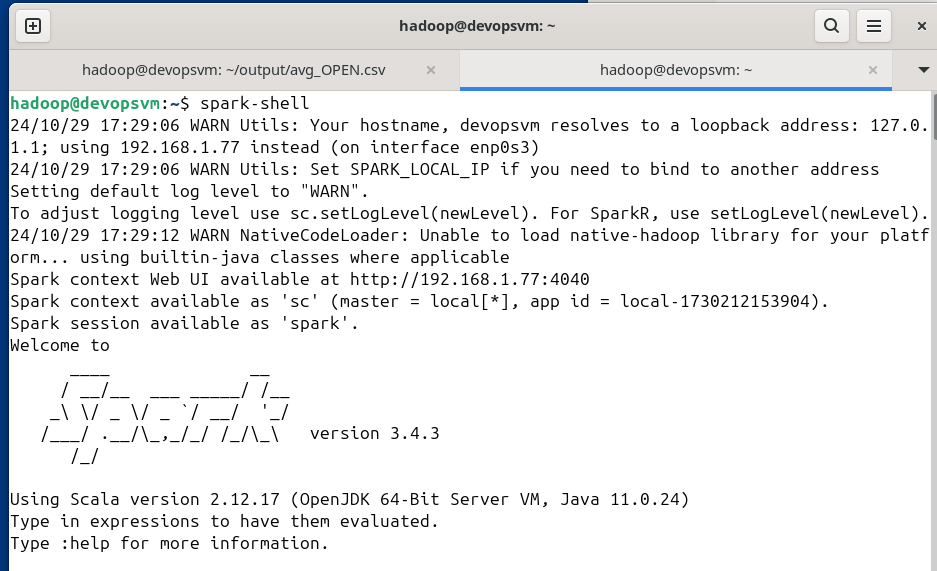




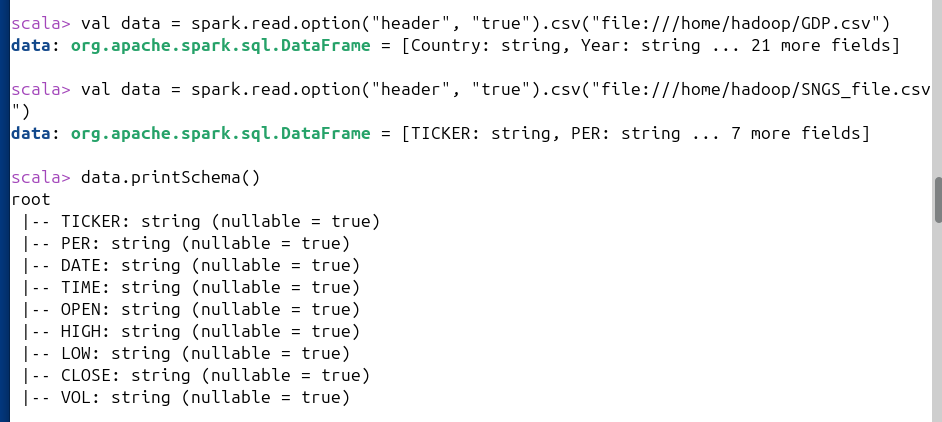
Предоставляем другим пользователям доступ на изменения данных:



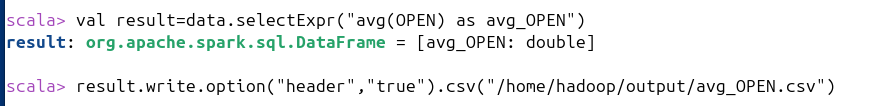
8. Запускаем Spark



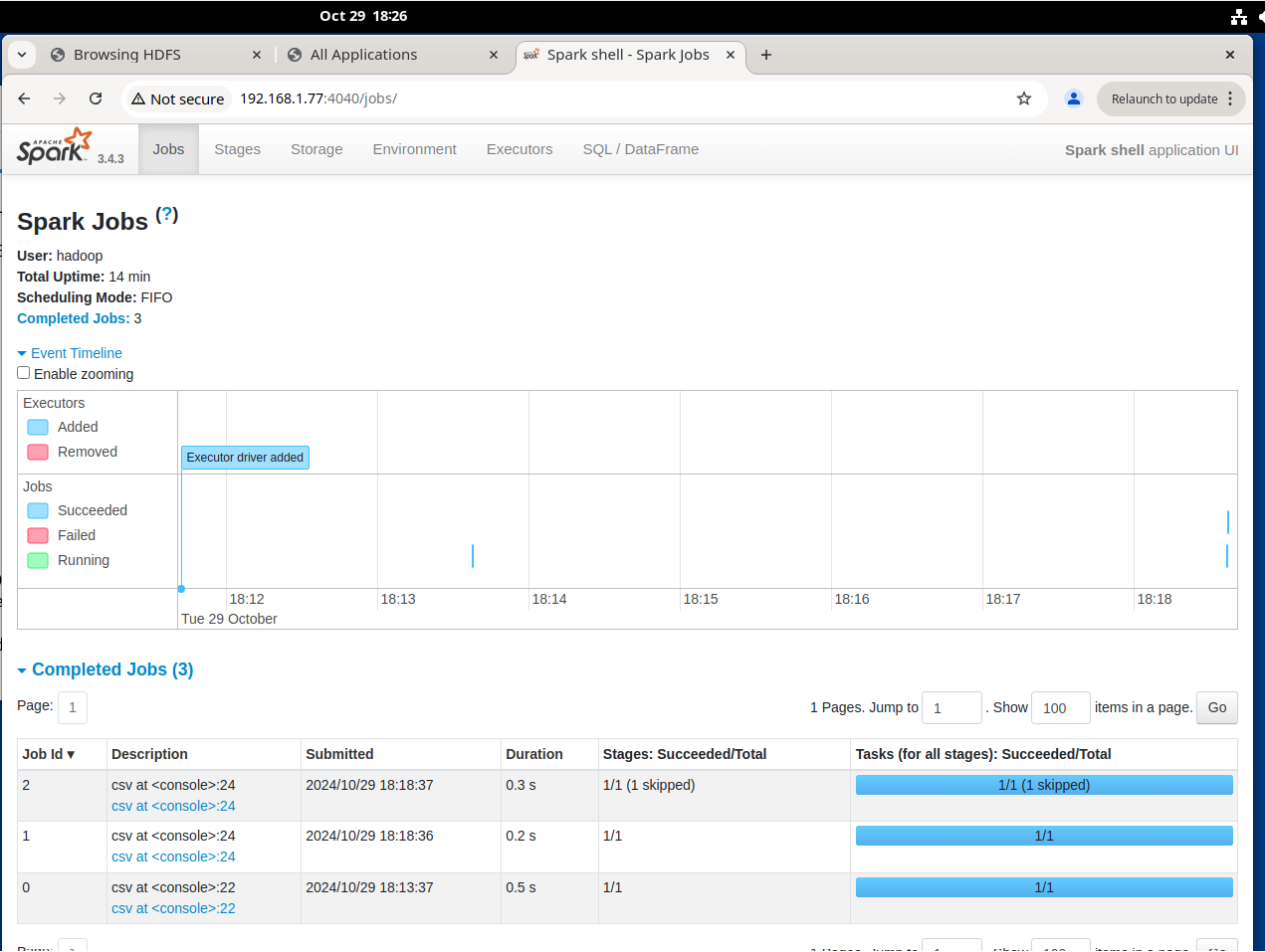
9. Загружаем данные и выводим схему DataFrame для проверки:



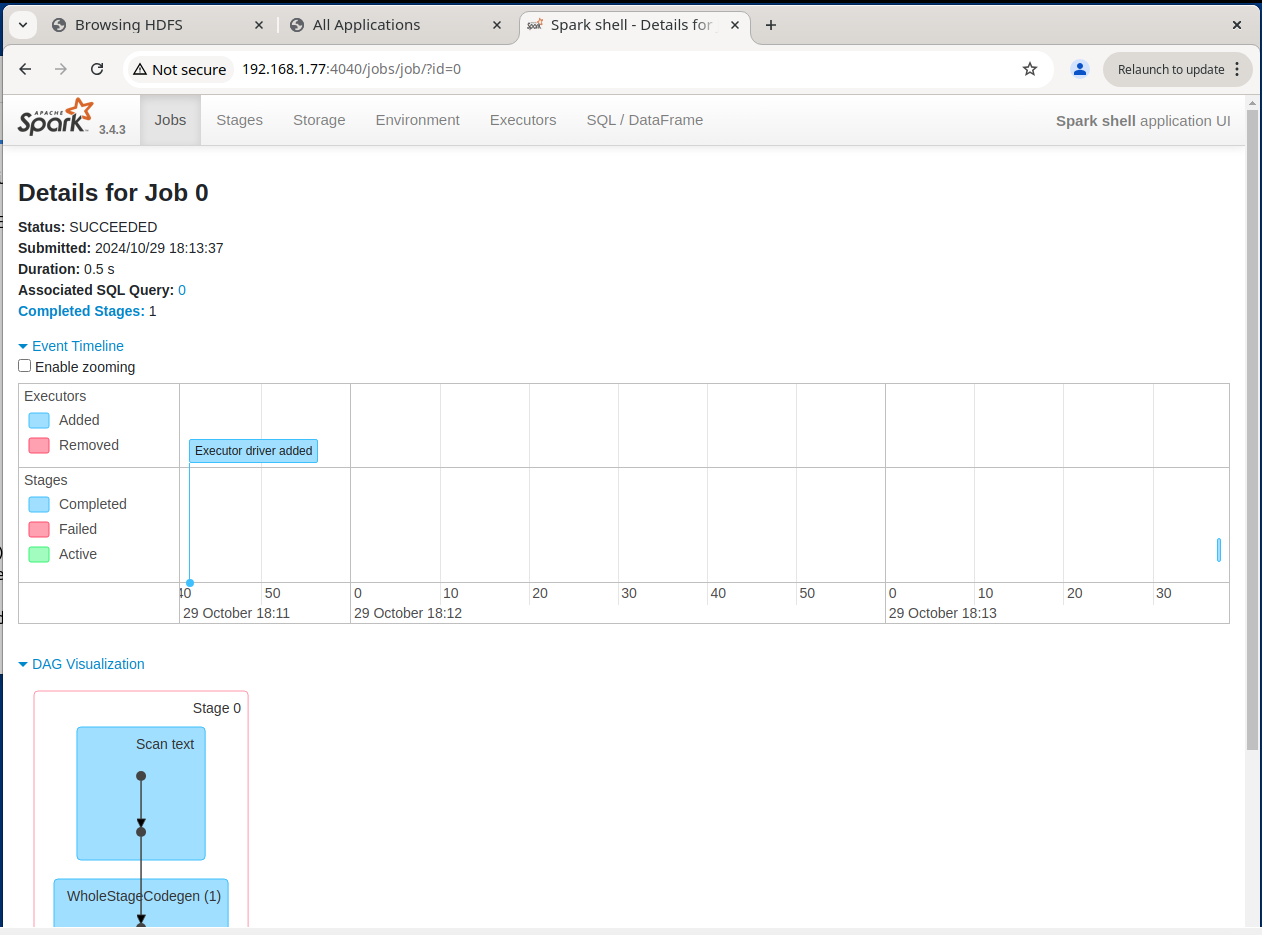
10. Вычислим среднее значение цены акций при открытии и сохраним результат в csv файл

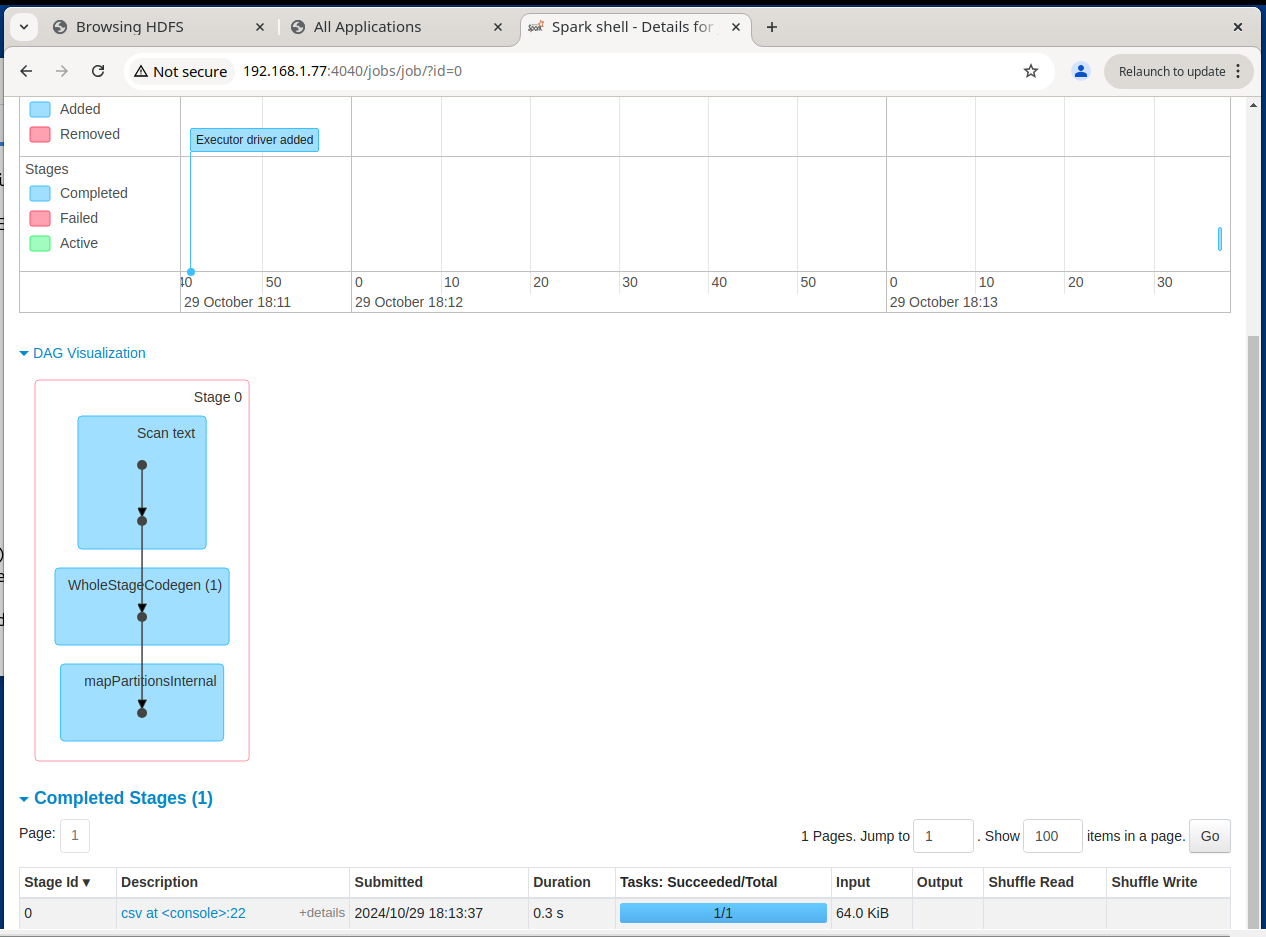


11. Посмотрим отображение этих процессов в браузере

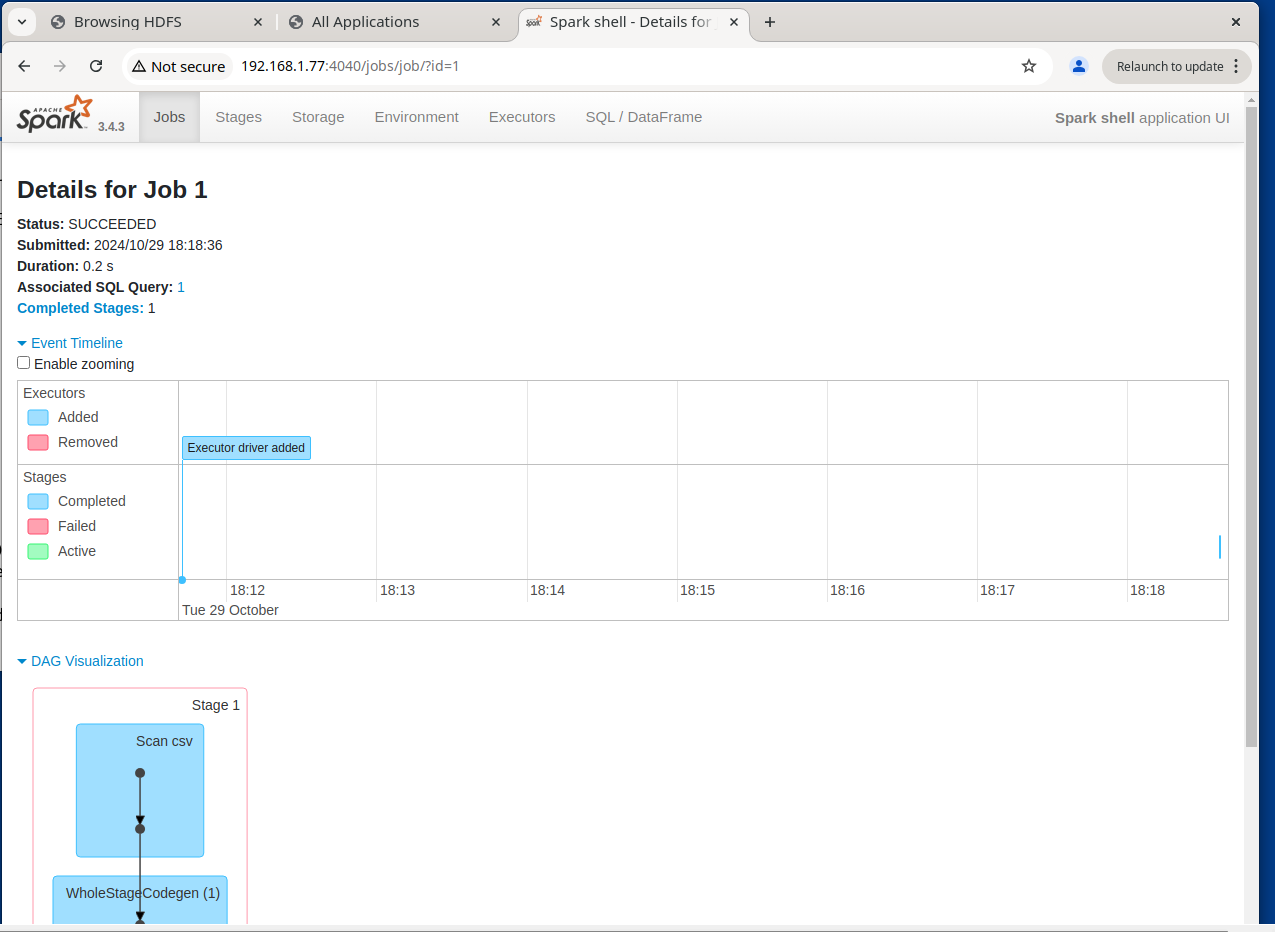


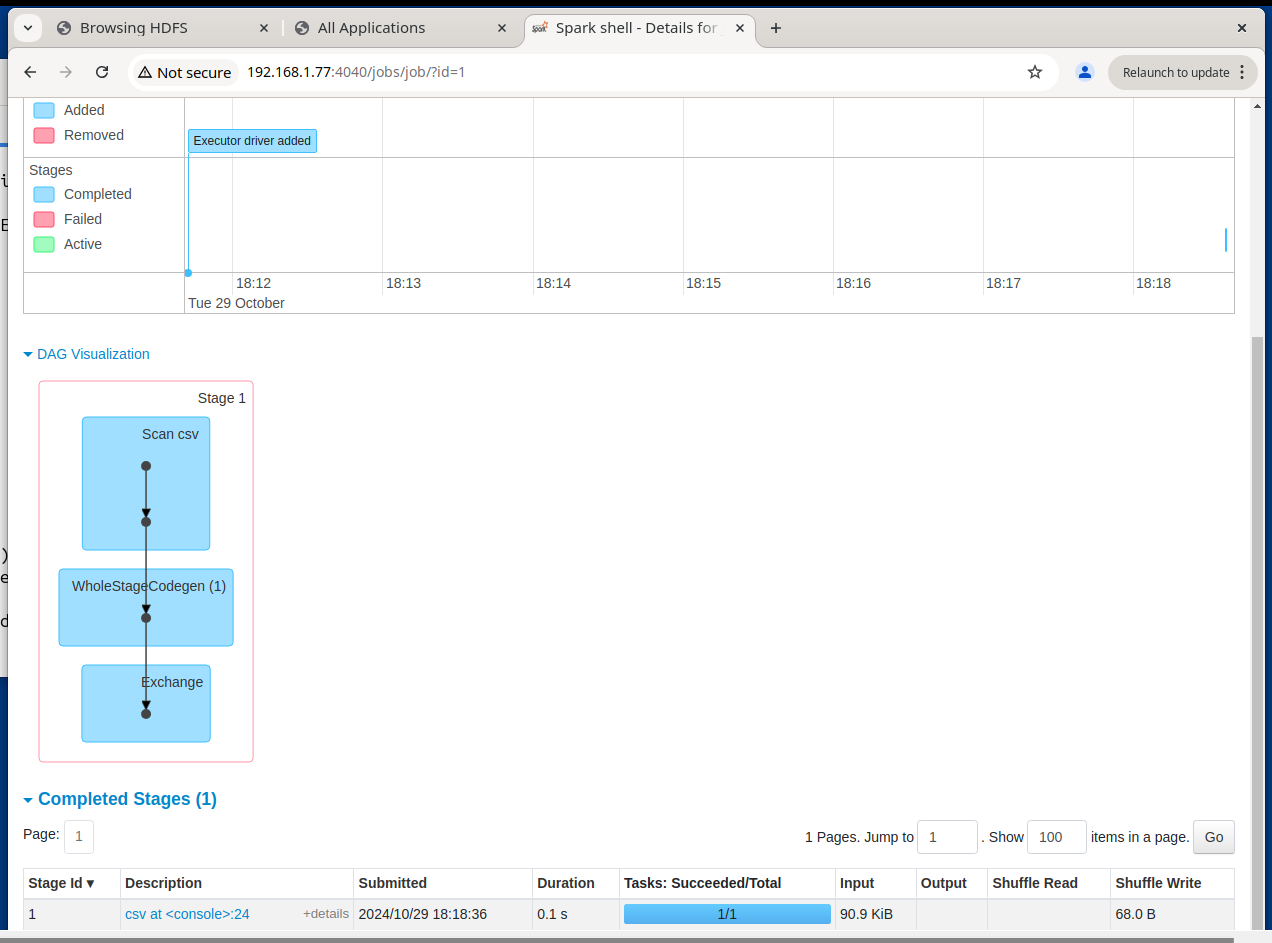
Job 0:



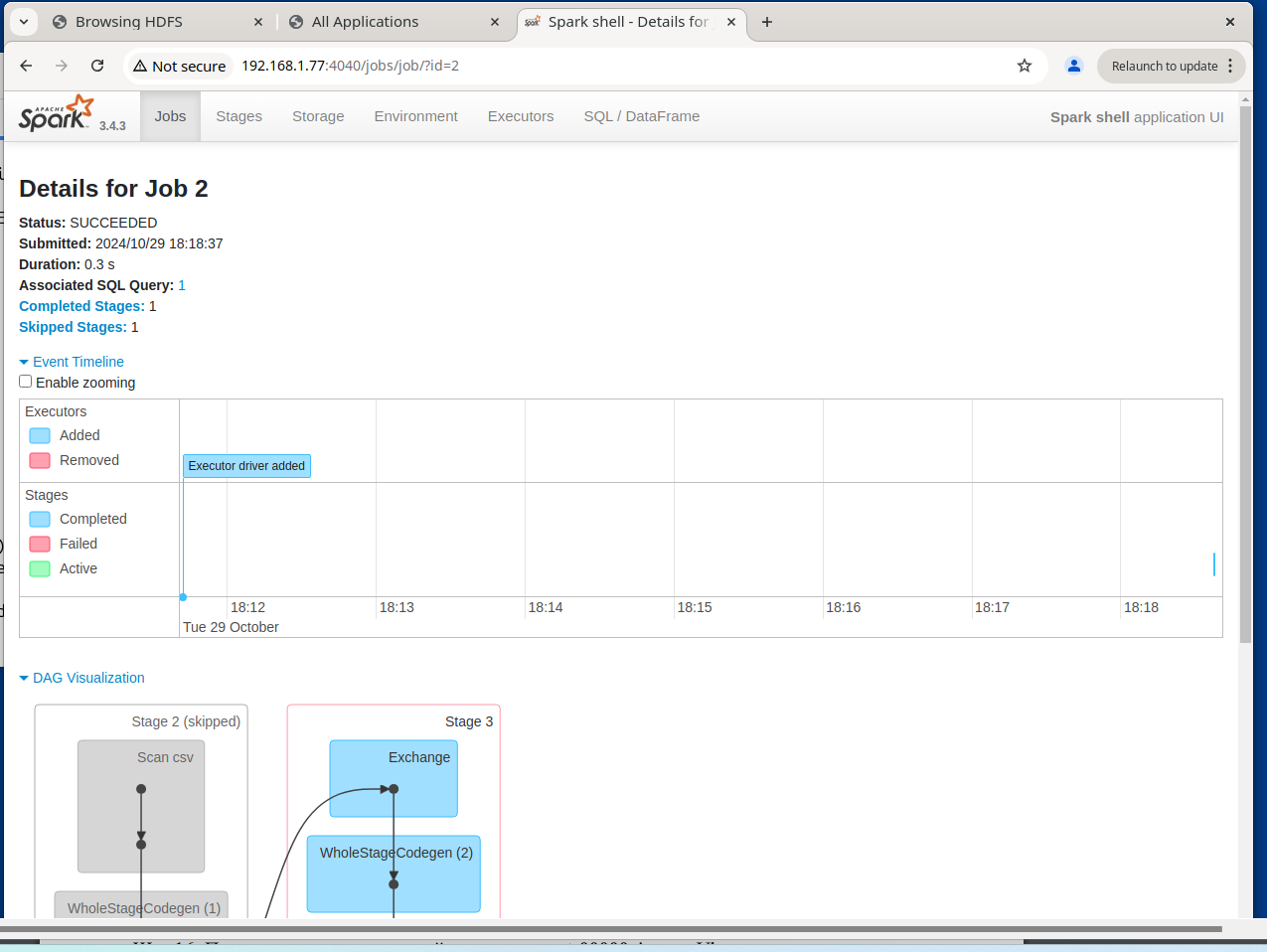


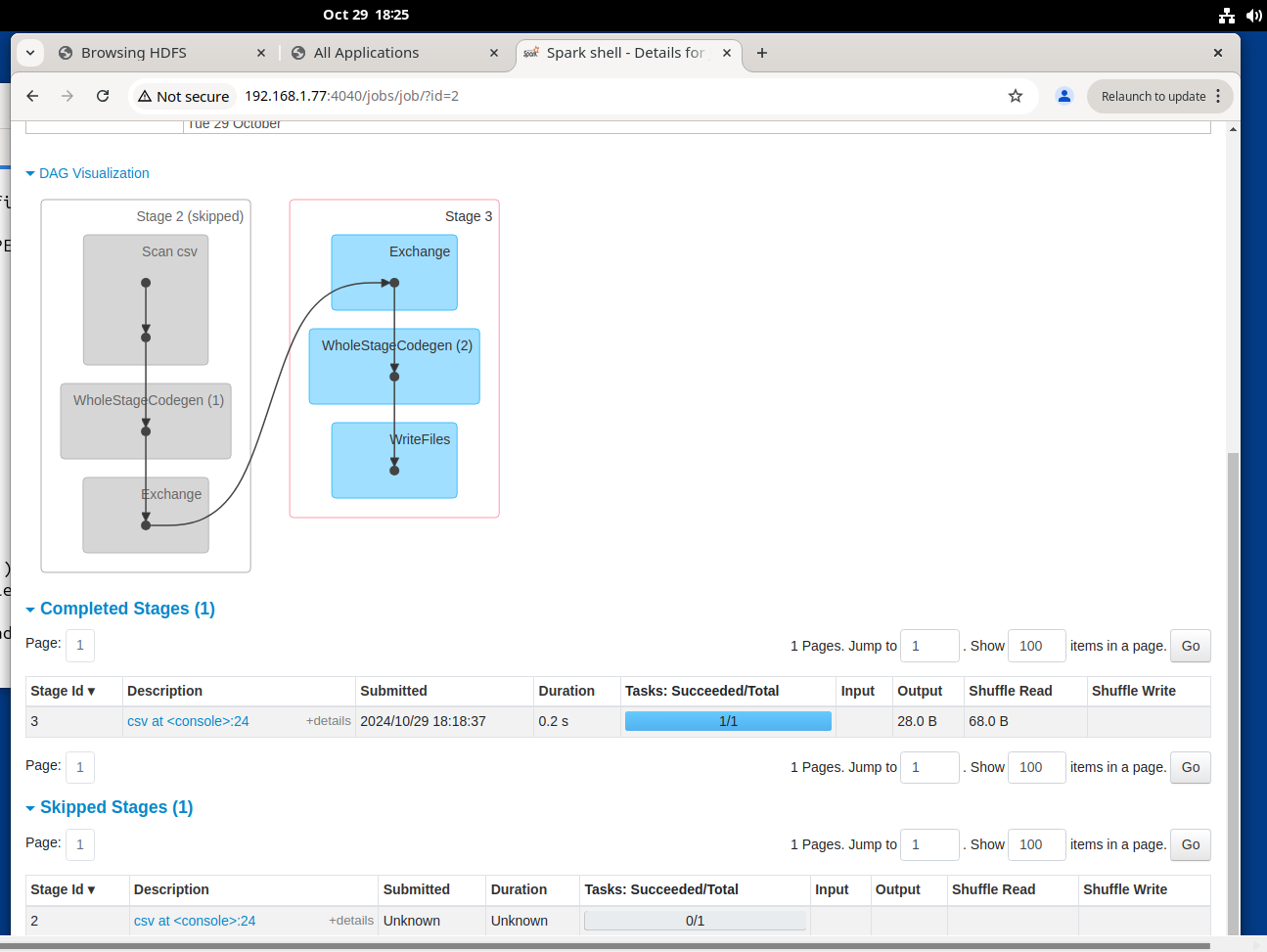
Job 1:



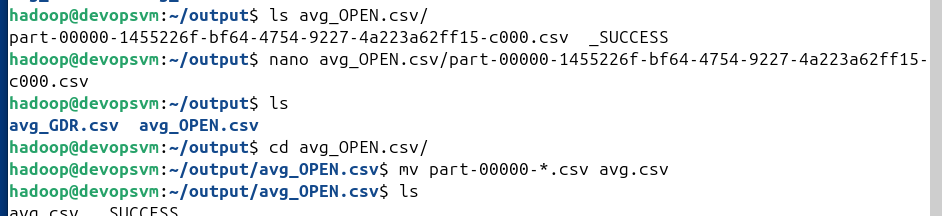


Job 2:

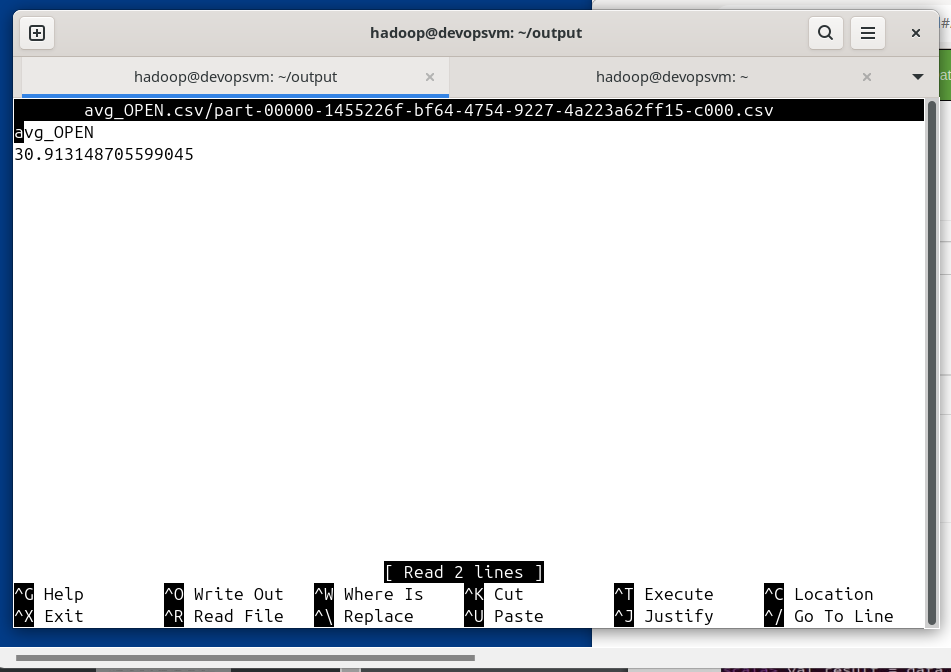




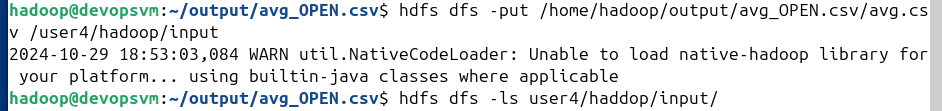
12. Переименуем полученный результат part-00000-\*.csv в Ubuntu avg\_OPEN.csv



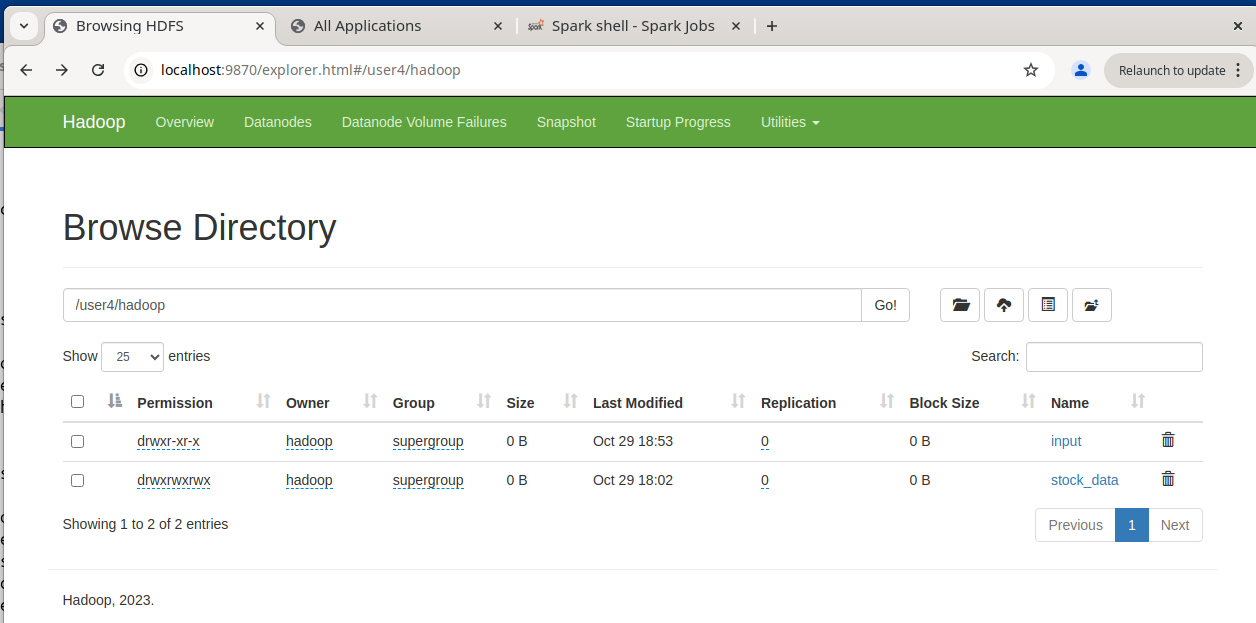
Посмотрим результат:

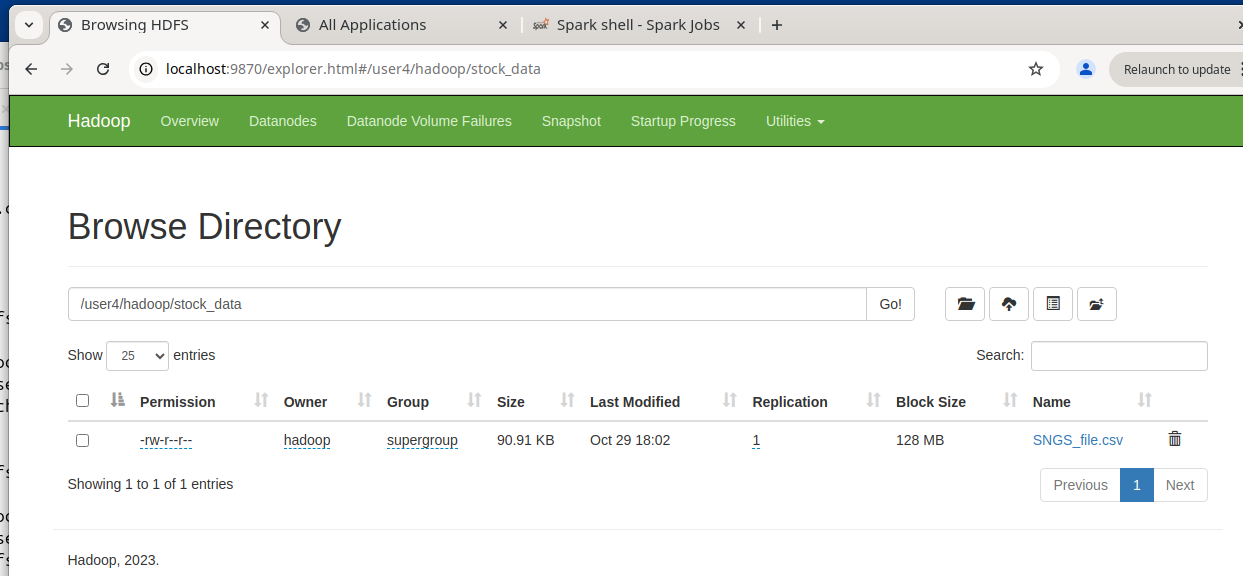


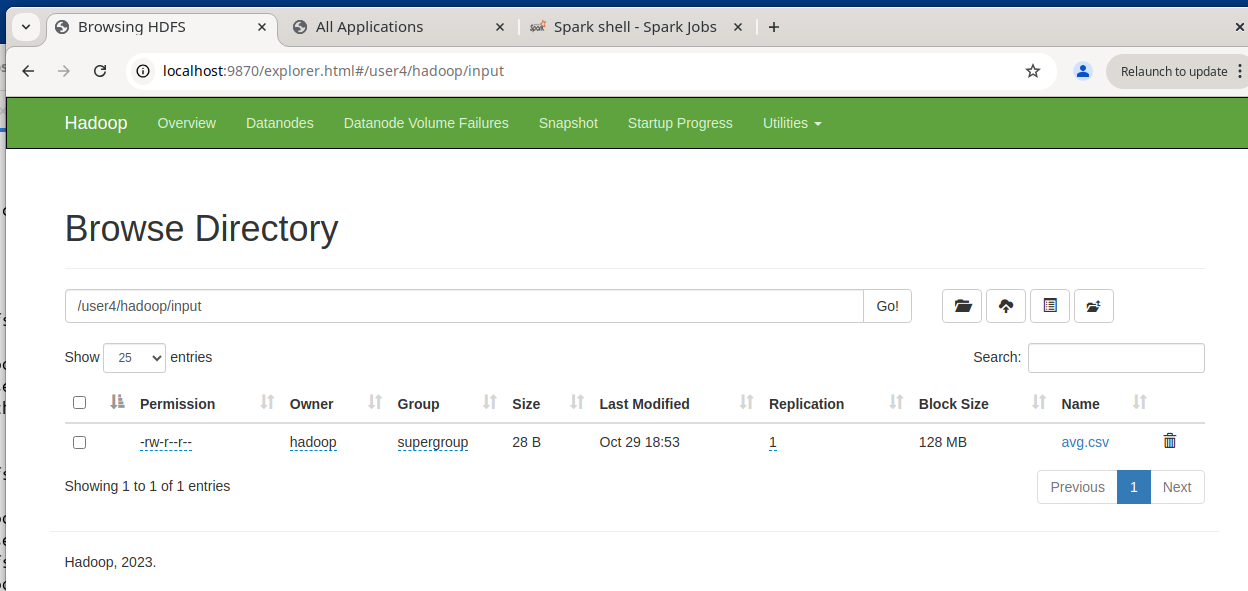
13. Перенесем данные в HDFS



Проверим, что файл загрузился



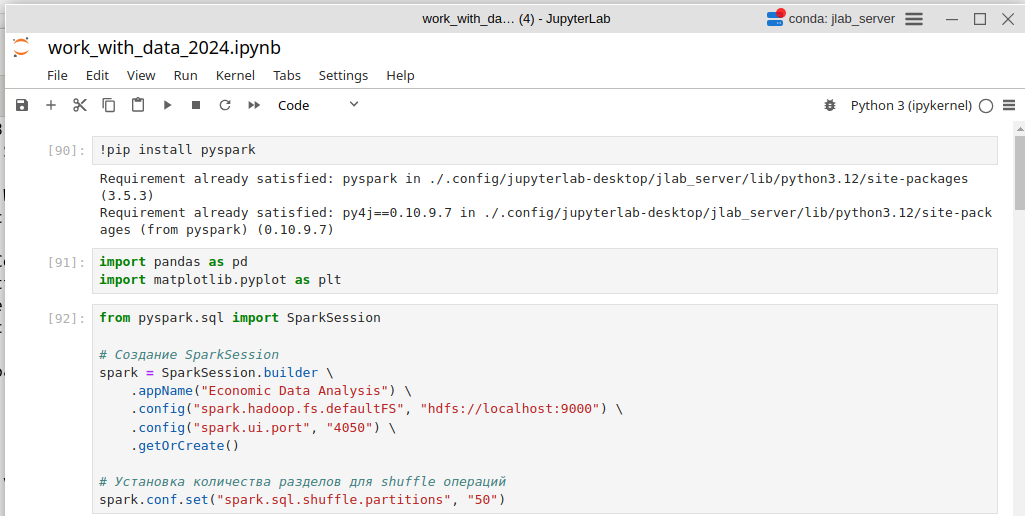




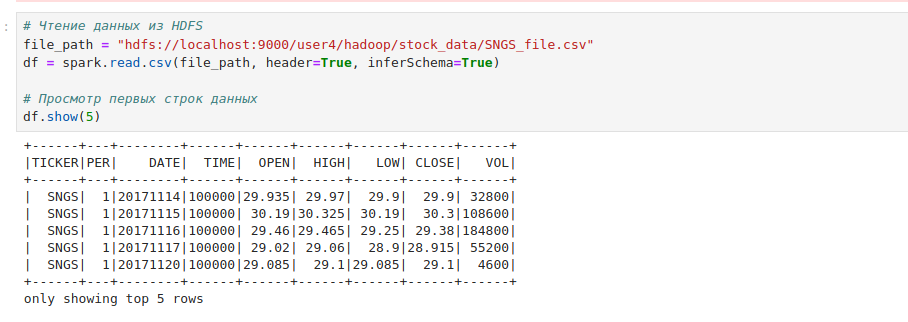
14. Переходим в Jupiter Notebook и выполняем следующие операции:

* фильтрация данных за 2020 год
* расчет стандартного отклонения цены открытия
* тренд анализа.

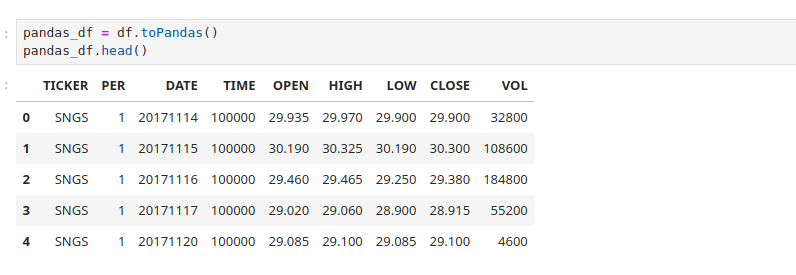
Устанавливаем pyspark, библиотеки pandas и matplotlib, создаем сессию



15. Выполняем чтение данных из HDFS

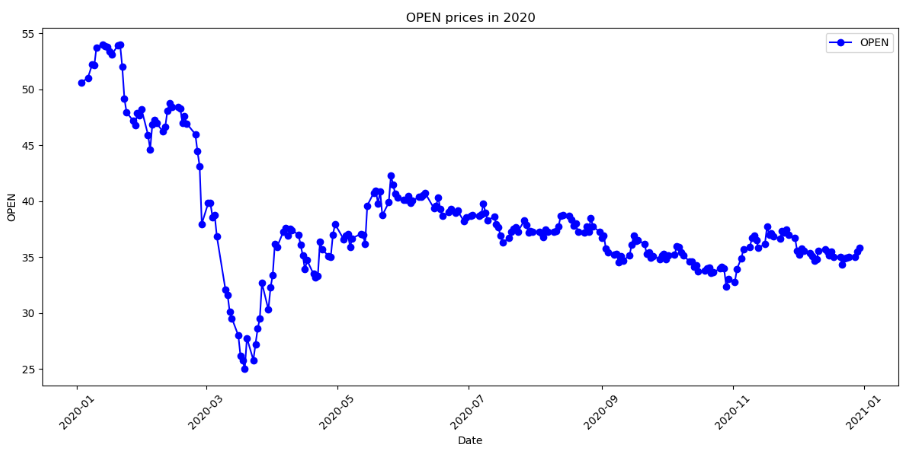


16. Выводим первые 5 строчек получившегося датафрейма

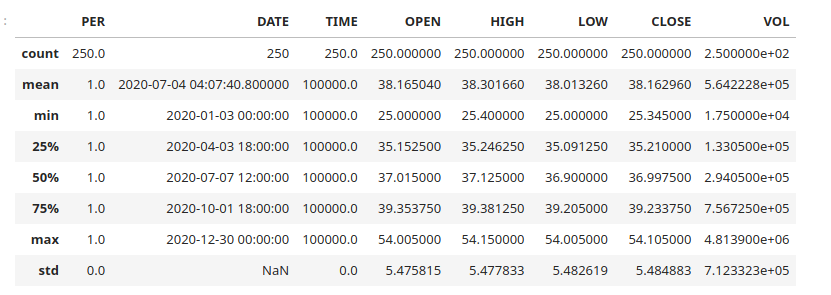


17. Выполняем фильтрацию данных за 2020 год, и визуализируем изменения цены акций при открытии



****

А также выведем статистику по всем показателям за 2020 год:

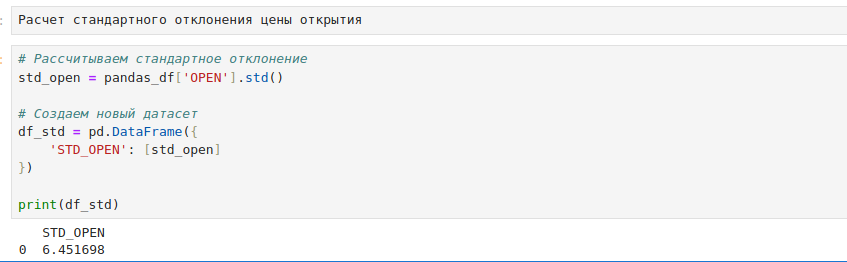


Преобразовываем данные обратно в pyspark dataframe и сохраняем их в HDFS

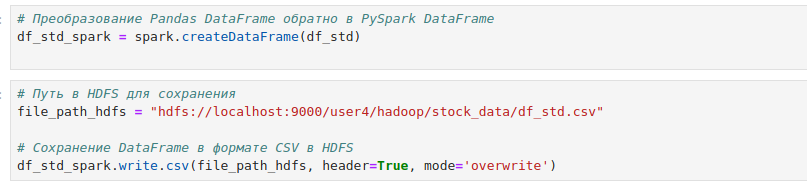


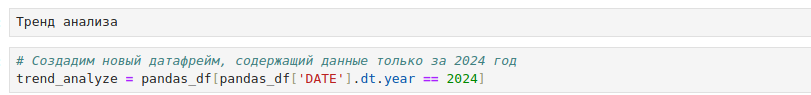


18. Рассчитаем стандартное отклонение цены открытия и запишем это значение в новый датафрейм



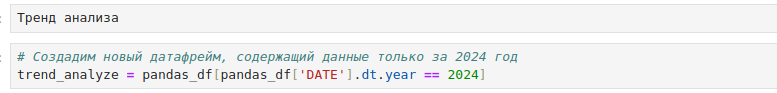
Преобразовываем данные обратно в pyspark dataframe и сохраняем их в HDFS



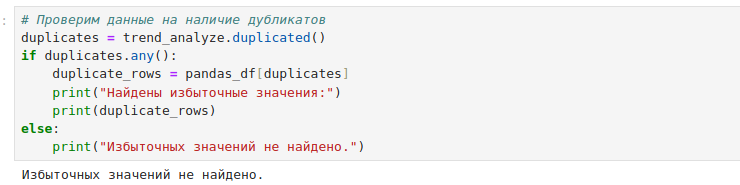


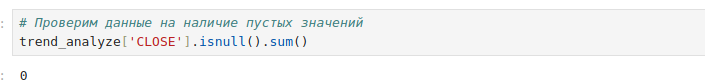
19. Выполним тренд анализа

Создадим новый датафрейм с данными только за этот год

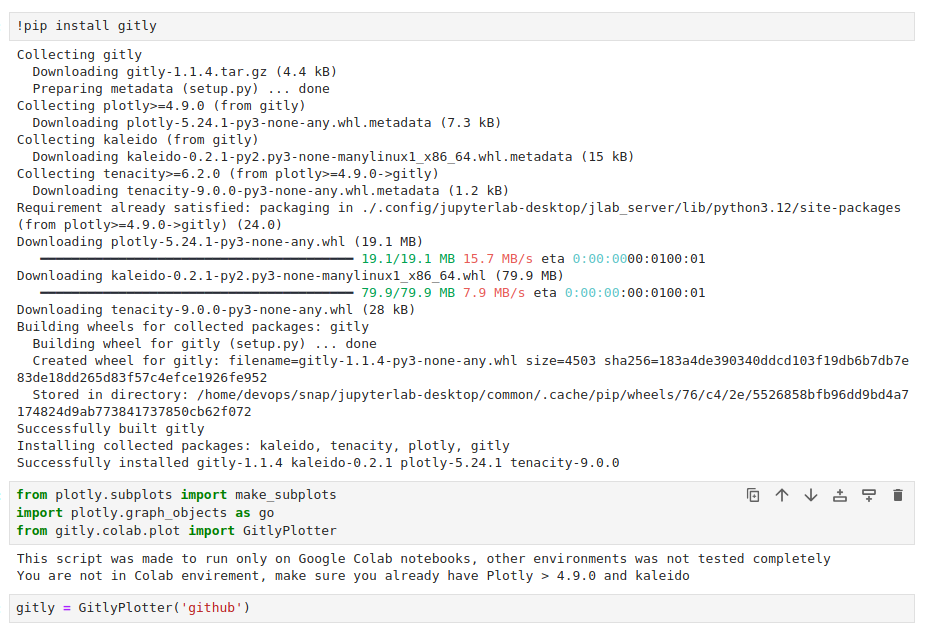


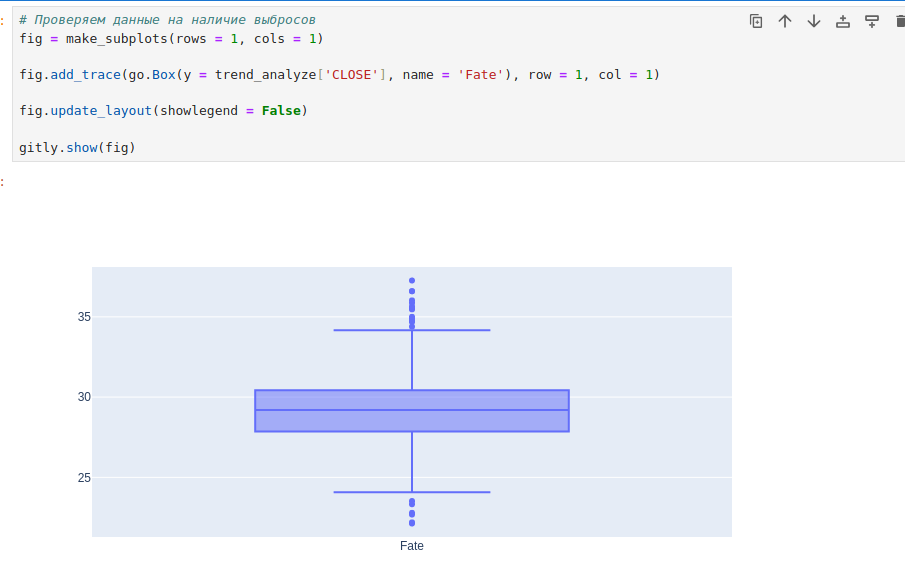
Проверим данные на наличие дубликатов, выбросов и пустых значений





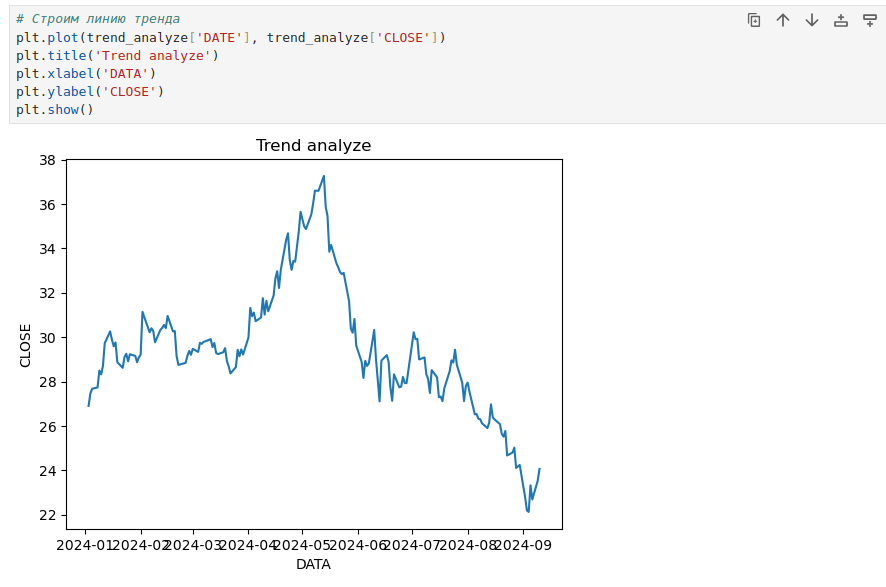
Для проверки на наличие выбросов установим библиотеку gitly



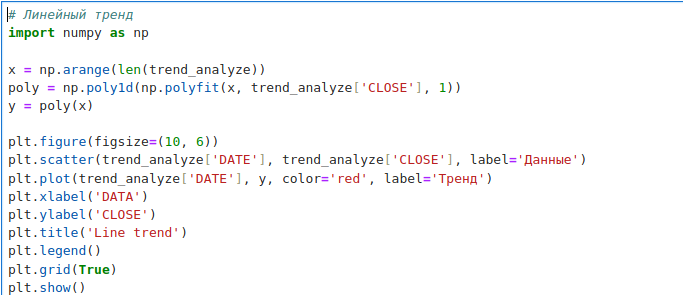


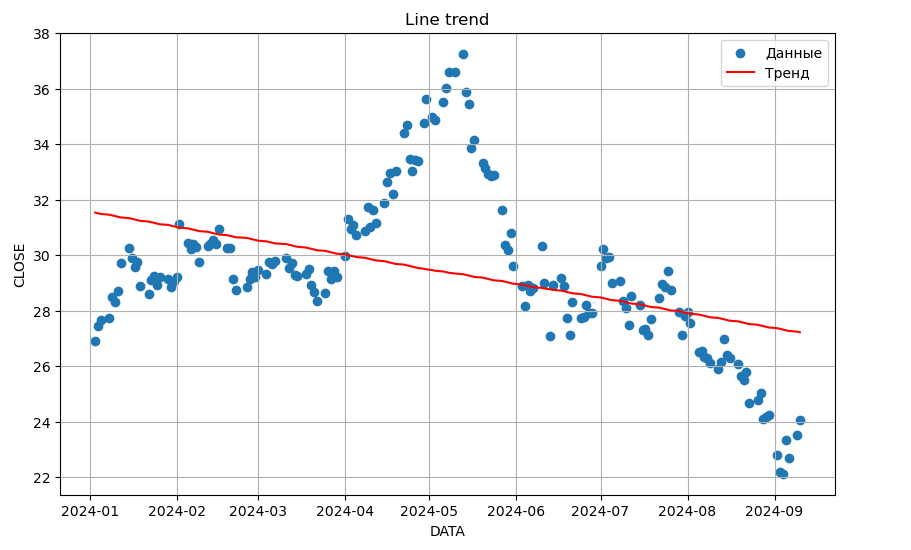
Выбросы отсутствуют. Значений, отклоняющихся от среднего немного, и отклонение не такое значительное. Это просто колебание цены акций.

Далее построим линию тренда, проанализируем цены акции при закрытии за 2024 год

****

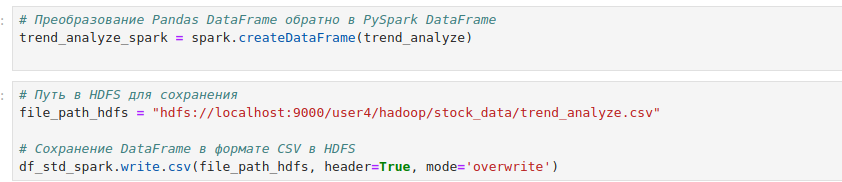
Построим диаграмму рассеивания с линией тренда



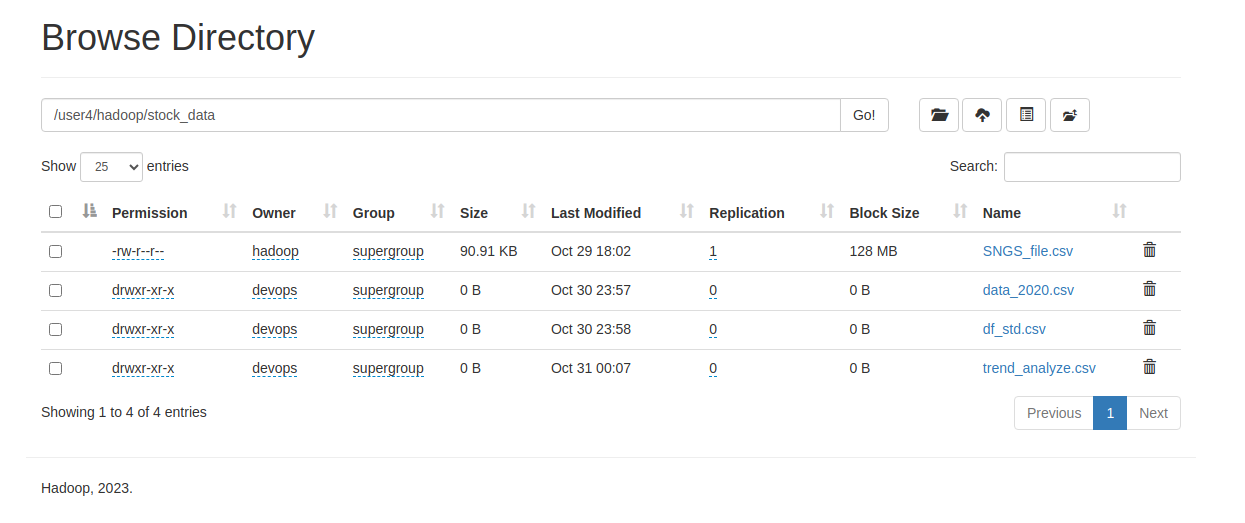


По данной диаграмме можно сделать вывод, что общая тенденция – это снижение цены акций при закрытии торгов в течении года. В мае был достаточно резкий подъем, но затем цена снова пошла на снижение. В сентябре цены начали повышаться, но возможно эта тенденция не сохранится, и тенденция снова пойдет на снижение.

Преобразовываем данные обратно в pyspark dataframe и сохраняем их в HDFS



20. Проверим, сохранились ли данные в HDFS



21. Остановим все процессы

