Департамент образования города Москвы

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

ОТЧЕТ

по дисциплине «Инструменты для хранения и обработки больших данных»

Направление подготовки 38.03.05 – бизнес-информатика

Профиль подготовки «Аналитика данных и эффективное управление»

(очная форма обучения)

Выполнил:

Студент группы АДЭУ-211

Буц Игорь Денисович

Руководитель:

Босенко Тимур Муртазович

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.

**Вариант 2:**

Установка Apache Hadoop и выполнение задачи на копирование файлов в HDFS. Данные: Исторические данные по акциям Газпрома (GAZP) с сайта Московской биржи (moex.com)

**Решение**

1. Генерация примерных данных акция GAZ с помощью кода на Python

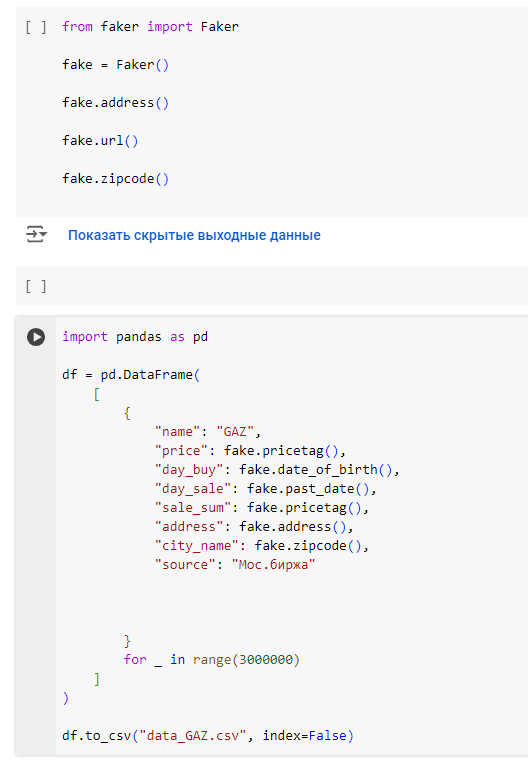


Рис. 1

1. Загрузка данных в Hadoop.
   1. Создание директории и пользователя

Рис. 2

* 1. Загрузка данных. Так как в Hadoop требуется загрузить файл более 128 мб, первым вариантом была загрузка из Git Hub. Но в Git Hub есть ограничения на размер файла – 25 мб.   
     Существует расширение для Git Hub: Git Hub LFS, который позволяет загружать большие файлы. Далее приведен пример использования расширения:

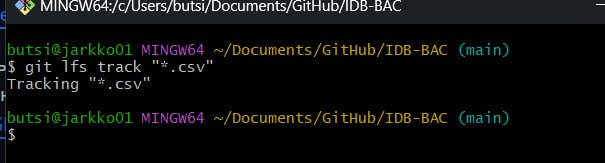


Рис. 3

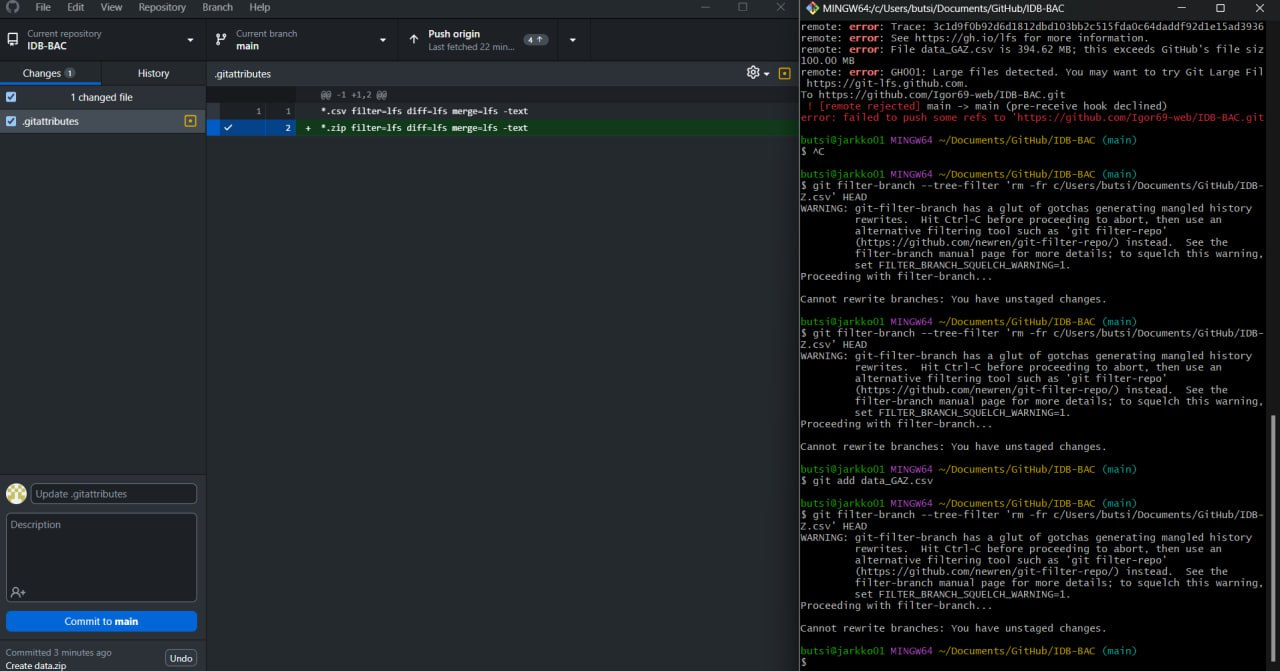
На рисунке 4 (Рис. 4) показан способ загрузки большого файла в репозиторий Git Hub, но можно заметить, что существует ограничение в 100 мб.

Рис. 4

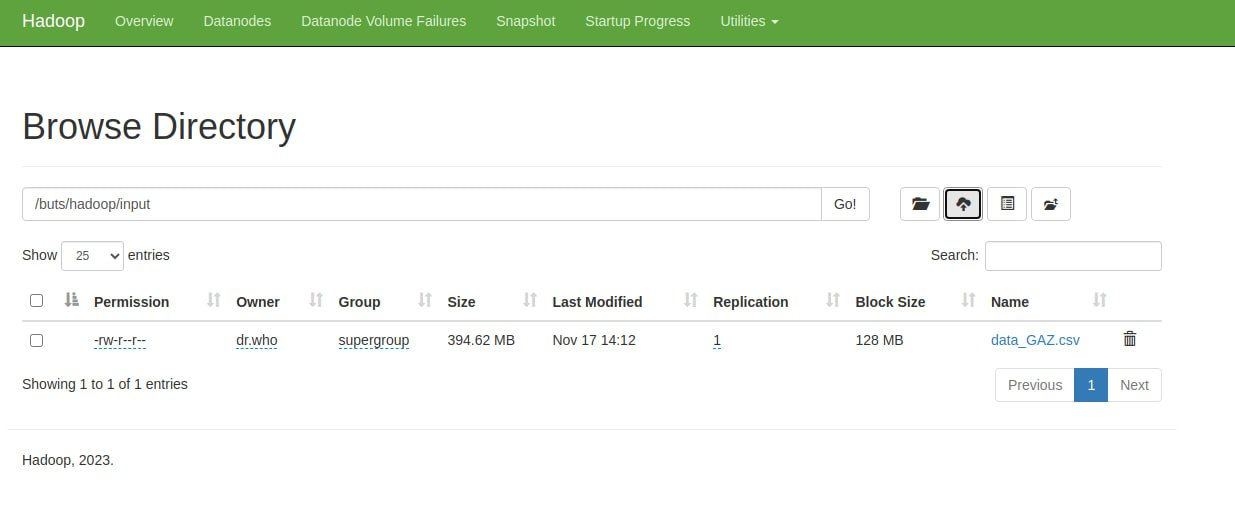
* 1. Загрузка вручную

Рис. 5