**Задание 1.**

1. Загрузите gender.csv в JupyterHub.

(через кнопку)

1. Скопируйте файл в домашнюю директорию своей учетной записи в hdfs.

hdfs dfs -put /home/std03/gender\_train.csv /user/std03

1. Переименуйте, используя свое Фамилию Имя – ivanov\_ivan.csv.

hdfs dfs -mv gender\_train.csv ivanov.csv

1. Создайте директорию со своими ФИ ivanov\_ivan в своей папке в HDFS, скопируйте туда файл.

hdfs dfs -mkdir ivanov

hdfs dfs -put ivanov\_ivan.csv ivanov

1. Отобразите содержимое файла и его размер.

hdfs dfs –cat ivanov\_ivan.csv

hdfs dfs -du ivanov\_ivan.csv

1. Выведите размер родительской папки и всех вложенных объектов с указанием гб/мб/кб

hdfs dfs -du –h /user

hdfs dfs -du -s –h

1. Создайте дубликат директории ivanov\_ivan и ее содержимого с другим названием.

hdfs dfs –cp input int

1. Удалите дубликат.

hdfs dfs -rm –r int

1. Загрузите файл ivanov\_ivan.csv в локальную директорию JupyterHub.

hdfs dfs –get ivanov\_ivan.csv /home/std03

**Задание 2.**

Напишите 10 отличий HDFS от обычной файловой системы Linux/Windows/Mac.

<https://stackoverflow.com/questions/17932374/what-is-the-difference-between-hdfs-and-ntfs-and-fat32>

1) горизонтальная масштабируемость за счёт дешевого железа

2) отказоустойчивость

репликация данных ( по умолчанию 3 уровня)

перезапуск задач

3) распределенные вычисления на кластере из множества машин

скрывает сложности распределенных систем

освобождает разработчиков от проблем системного уровня

передача данных

доставка кода к данным и др.

последовательный доступ к данным

4) поддержка принципа типа "один раз записать данные и много раз считывать в дальнейшем"

5) подходит для хранения большого объёма данных

6) Работает поверх облачных файловых систем

7) В отличие от локальных файловых систем, в HDFS (Hadoop Distributed File System) невозможно изменение (модификация) файла. Файлы в HDFS могут быть записаны лишь однажды, причем одновременно запись в файл осуществляет только один процесс.