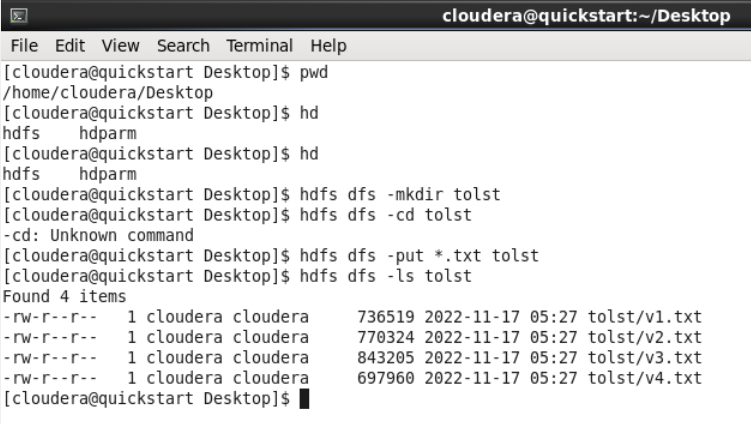
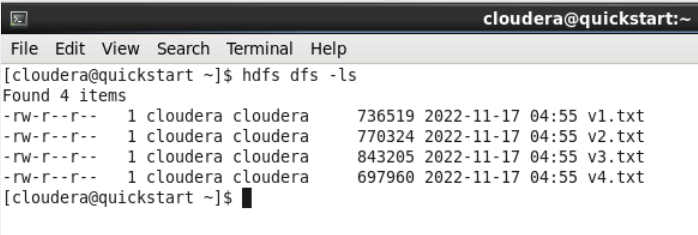
И выполняем следующие действия

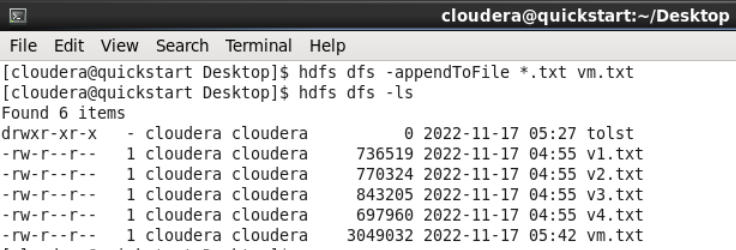
• Когда мы перетащили файлы с произведением Льва Толстого – мы перетащили их в файловую систему виртуальной машины, но не в HDFS, соответственно, в первую очередь нам нужно перенести их в папку нашего пользователя именно на HDFS.



• После того, как файлы окажутся на HDFS попробуйте выполнить команду, которая выводит содержимое папки. Особенно обратите внимание на права доступа к вашим файлам.

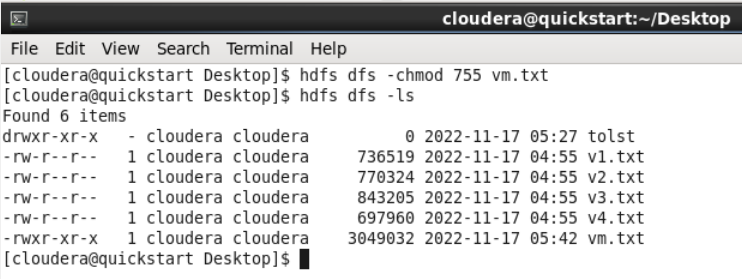


• Далее сожмите все 4 тома в 1 файл.



*Это объединение с локальной директории в облако.  
Как объединить то, что находится в облаке сразу там в облаке – выяснить не удалось.*

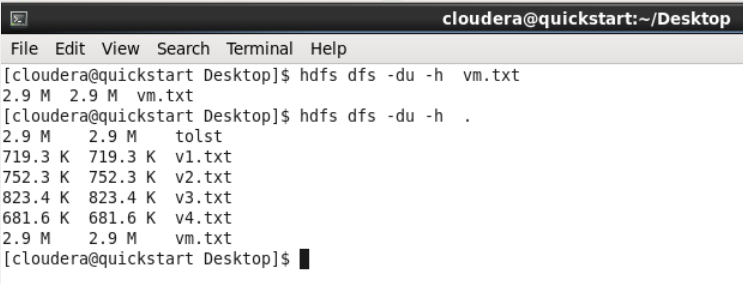
• Теперь давайте изменим права доступа к нашему файлу. Чтобы с нашим файлом могли взаимодействовать коллеги, установите режим доступа, который дает полный доступ для владельца файла, а для сторонних пользователей возможность читать и выполнять.



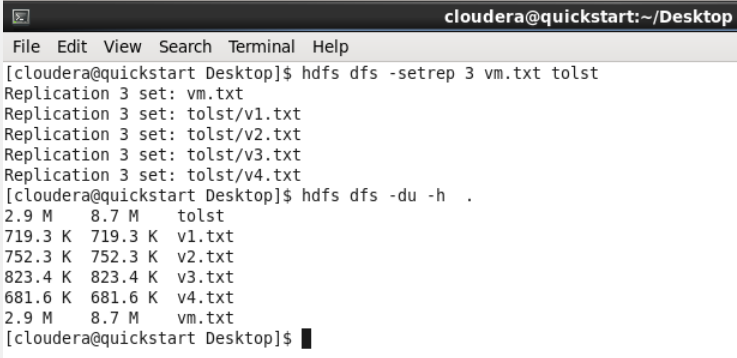
• Попробуйте заново использовать команду для вывода содержимого папки и обратите внимание как изменились права доступа к файлу.

+

• Теперь попробуем вывести на экран информацию о том, сколько места на диске занимает наш файл. Желательно, чтобы размер файла был удобночитаемым.



• На экране вы можете заметить 2 числа. Первое число – это фактический размер файла, а второе – это занимаемое файлом место на диске с учетом репликации. По умолчанию в данной версии HDFS эти числа будут одинаковы – это означает, что никакой репликации нет – нас это не очень устраивает, мы хотели бы, чтобы у наших файлов существовали резервные копии, поэтому напишите команду, которая изменит фактор репликации на 2.

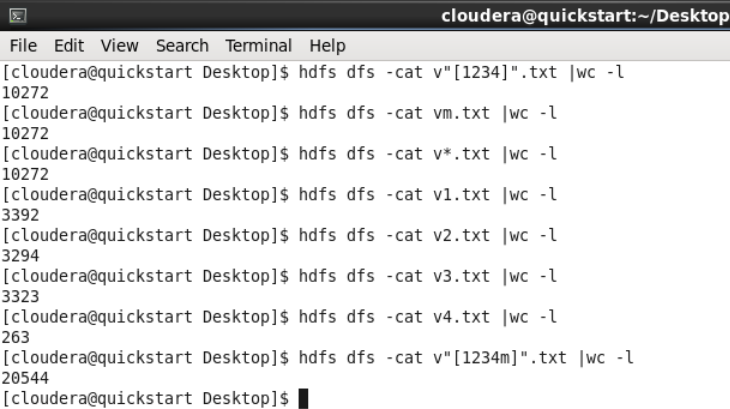


*Сделал на 3 , наверное, не критично.*

• Повторите команду, которая выводит информацию о том, какое место на диске занимает файл и убедитесь, что изменения произошли.

+

• Напишите команду, которая подсчитывает количество строк в вашем файле



• В качестве результатов вашей работы, запишите ваши команды и вывод этих команд в отдельный файл и выложите его на github.

Дополнительные ссылки:

• Видеоинструкция по установке cloudera на VirtualBox: https://www.youtube.com/watch?v=8rsNXdvFqBc

• Docker-образ экосистемы Hadoop: https://github.com/big-data-europe/docker-hadoop-spark-workbench