ДЗ #1

Задания уровня "Beginner"

1. Пробросить порт (port forwarding) для доступа к HDFS Web UI
2. [3 балла] Воспользоваться Web UI для того, чтобы найти папку "/data" в HDFS. Сколько подпапок в указанной папке /data?

Жмём "Utilities" -> "Browse the file system" -> набираем "/data"
Видим единственную подпапку "texts"
Ответ: одна

Задания уровня "Intermediate"

Флаг "-ls"

1. [3 балла] Вывести список всех файлов в /data/texts

```
$hdfs dfs -ls /data/texts
Found 1 items
-rw-r--r--   1 hadoop hadoop        714 2020-09-18 20:50
/data/texts/twain.txt
```

2. [3 балла] См. п.1 + вывести размер файлов в "human readable" формате (т.е. не в байтах, а например в МБ, когда размер файла измеряется от 1 до 1024 МБ).

```
$hdfs dfs -ls -h /data/texts
Found 1 items
-rw-r--r--   1 hadoop hadoop        714 2020-09-18 20:50
/data/texts/twain.txt
```

3. [3 балла] Команда "hdfs dfs -ls" выводит актуальный размер файла (actual) или же объем пространства, занимаемый с учетом всех реплик этого файла (total)? В ответе ожидается одно слово: actual или total.

Ответ: actual

Флаг "-du"

1. [3 балла] Приведите команду для получения размера пространства, занимаемого всеми файлами внутри "/data/texts". На выходе ожидается одна строка с указанием команды.

`hdfs dfs -du -s /data/texts`

Флаги "-mkdir" и "-touchz"

1. [4 балла] Создайте папку в корневой HDFS-папке Вашего пользователя

`hdfs dfs -mkdir a.kopnin`

2. [4 балла] Создайте в созданной папке новую вложенную папку.

`hdfs dfs -mkdir a.kopnin/subdirectory`

3. [4 балла] Что такое Trash в распределенной FS? Как сделать так, чтобы файлы удалялись сразу, минуя "Trash"?

Ответ: это аналог "корзины", т.е. по умолчанию при удалении файлы попадают в директорию `.Trash` текущего пользователя, а не удаляются перманентно. Чтобы удалить файл минуя "Trash", следует указать флаг `-skipTrash` для команды `hdfs dfs -rm`

4. [4 балла] Создайте пустой файл в подпапке из пункта 2.

`hdfs dfs -touchz a.kopnin/subdirectory/empty.txt`

5. [3 балла] Удалите созданный файл.

В задании не сказано, что файл нужно удалить перманентно, поэтому флаг `-skipTrash` указывать не будем.

`hdfs dfs -rm a.kopnin/subdirectory/empty.txt`

6. [3 балла] Удалите созданные папки.

`hdfs dfs -rm -r a.kopnin`

Флаги "-put", "-cat", "-tail", "-distcp"

1. [4 балла] Используя команду "-distcp" скопируйте рассказ О'Генри "Дары Волхвов" henry.txt из s3://texts-bucket/henry.txt в новую папку на HDFS

```
hdfs dfs -mkdir a.kopnin
hadoop distcp s3://texts-bucket/henry.txt a.kopnin/henry.txt
```

2. [4 балла] Выведите содержимое HDFS-файла на экран.

`hdfs dfs -cat a.kopnin/henry.txt`

3. [4 балла] Выведите содержимое нескольких последних строчек HDFS-файла на экран.

`hdfs dfs -tail a.kopnin/henry.txt`

4. [4 балла] Выведите содержимое нескольких первых строчек HDFS-файла на экран.

`hdfs dfs -cat a.kopnin/henry.txt | head`

5. [4 балла] Переместите копию файла в HDFS на новую локацию.

```
hdfs dfs -mkdir a.kopnin/new_dir
hdfs dfs -mv a.kopnin/henry.txt a.kopnin/new_dir/henry.txt
```

Задания уровня "Advanced"

2. [6 баллов] Изменить replication factor для файла. Как долго занимает время на увеличение / уменьшение числа реплик для файла?

```
hdfs dfs -setrep -w 2 a.kopnin/new_dir/henry.txt
hdfs dfs -setrep -w 1 a.kopnin/new_dir/henry.txt
```

Увеличение числа реплик с 1 до 2 заняло 8 секунд, тогда как обратное уменьшение до одной - 17 секунд.

3. [6 баллов] Найдите информацию по файлу, блокам и их расположениям с помощью "hdfs fsck"

`hdfs fsck a.kopnin/new_dir/henry.txt -files -blocks -locations`

4. [6 баллов] Получите информацию по любому блоку из п.2 с помощью "hdfs fsck -blockId". Обратите внимание на Generation Stamp (GS number).

`hdfs fsck -blockId blk_1073745505`