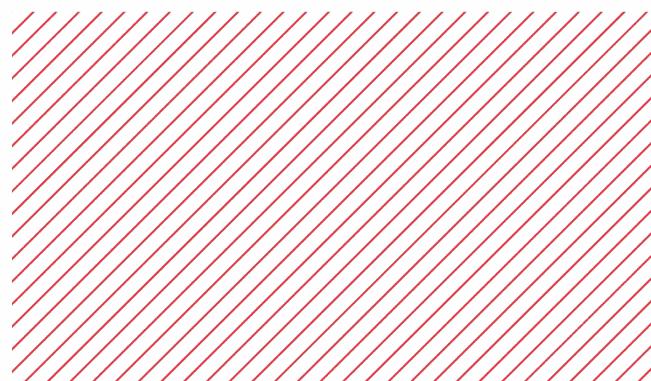


академия  
больших  
данных



# HW01: HDFS



# Описание работы и критерии оценивания

---

Для доступа на кластер вы получите EC2-ключ emr.pem, который вы будете использовать для подключения через ssh. Сохраните данный ключ в папке на вашей локальной машине и в данной папке через терминал (командную строку) пропишите следующую команду:

```
ssh -i emr.pem hadoop@ec2-3-249-21-2.eu-west-1.compute.amazonaws.com
```

В случае ошибки с правами нужно назначить chmod 400 emr.pem для ключа

Если все сделано правильно, то вы должны увидеть надпись EMR в вашем терминале (командной строке). После этого вы можете писать команды в консоли для взаимодействия с HDFS.

Максимально возможное количество баллов за работу: 75 баллов

Результаты в виде списка команд отправлять: [muzalevsky.ds@gmail.com](mailto:muzalevsky.ds@gmail.com)

Бонусы и штрафы:

- **100%** за плагиат
- **30%** за посылку решения в течение недели после deadline
- **10%** за повторные сдачи

# Задания уровня “Beginner”

---

Веб интерфейс HDFS (WebUI HDFS) доступен в EMR через порт 50070. Для того, чтобы пробросить порт с вашей локальной машины и открыть веб интерфейс в вашем браузере через `http://localhost:port_name`, используется следующая команда:

```
ssh -i emr.pem -N -L port_name:master_node_public_dns_name:50070 hadoop@ master_node_public_dns_name
```

Вместо `port_name` вы можете использовать любой открытый локальный порт.

Задачи:

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?

# Задания уровня “Intermediate”

---

Все следующие задачи используют консольную утилиту “`hdfs dfs`”. Чтобы получить документацию / подсказку по HDFS-утилите или флагу, можно набрать:

- `hdfs dfs -usage`
- `hdfs dfs -help`
- `hdfs dfs -usage ls`
- `hdfs dfs -help ls`

См. флаг “`-ls`”, чтобы:

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

См. 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.

# Задания уровня “Intermediate”

---

См. флаги “-mkdir” и “-touchz”

1. [4 балла] Создайте папку в корневой HDFS-папке Вашего пользователя
2. [4 балла] Создайте в созданной папке новую вложенную папку.
3. [4 балла] Что такое Trash в распределенной FS? Как сделать так, чтобы файлы удалялись сразу, минуя “Trash”?
4. [4 балла] Создайте пустой файл в подпапке из пункта 2.
5. [3 балла] Удалите созданный файл.
6. [3 балла] Удалите созданные папки.

См. 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
2. [4 балла] Выведите содержимое HDFS-файла на экран.
3. [4 балла] Выведите содержимое нескольких последних строчек HDFS-файла на экран.
4. [4 балла] Выведите содержимое нескольких первых строчек HDFS-файла на экран.
5. [4 балла] Переместите копию файла в HDFS на новую локацию.

# Задания уровня “Advanced”

---

Полезные флаги:

- Для “hdfs dfs”, см. “-setrep -w”
- hdfs fsck /path -files - blocks -locations

Задачи:

2. [6 баллов] Изменить replication factor для файла. Как долго занимает время на увеличение / уменьшение числа реплик для файла?
3. [6 баллов] Найдите информацию по файлу, блокам и их расположениям с помощью “hdfs fsck”
4. [6 баллов] Получите информацию по любому блоку из п.2 с помощью “hdfs fsck -blockId”.  
Обратите внимание на Generation Stamp (GS number).