BIGDATA TEAM

# HW #01: HDFS proficiency




автор задания:

- Алексей Драль, aadral@bigdatateam.org
- Founder & Big Data Instructor @ BigData Team

BIGDATA TEAM

## 1. Описание задания и критериев оценивания

Все ответы на вопросы, полученную из системы информацию, сравнения и результаты исследований необходимо отобразить в файле домашней работы. За выполнение практических заданий учебного модуля по HDFS вы можете набрать 100%. Распределение баллов выглядит следующим образом:

- **1%** - задания уровня beginner
- **45%** - задания уровня intermediate
- **54%** - задания уровня advanced

Для простоты расчетов, соответствующие баллы (проценты) стоят около каждого упражнения. Сдача заданий производится посылкой YAML[1] файла, где напротив каждого упражнения будет написан ваш ответ. Шаблон находится по следующей ссылке:

- github:big-data-team/big-data-course/hdfs_quiz_template.yml

По умолчанию, в файле нужно указать команды, которые вы использовали для получения ответа на вопрос. Места, где нужно указать только число или текстовые комментарии, указаны явно. Ответы проверяются в автоматическом режиме, но в случае уточнений / вопросов их будут смотреть преподаватели и менторы курса, поэтому смело пишите в многострочном режиме комментарии для людей. Обратите внимание, что многострочные ответы указаны исключительно для примера, чтобы показать как ими пользоваться, а не с целью дать подсказку, где нужны однострочные или многострочные ответы.

Бонусы и штрафы:

- **100%** за плагиат в решениях (всем участникам процесса)
- **100%** за посылку решения после deadline
- **5%** за каждую дополнительную посылку в тестирующую систему (одна дополнительная посылка бесплатно)

Формула подсчета финальной оценки[2]:

$$max(0,\ 0.95^{max(0,\#\text{доп.посылок} - 1)} * (1 - \text{штраф.за.дедлайн.и.списывание})) * \text{оценка.по.тестам}$$

[1] Если вы не знакомы с форматом YAML файла - не стоит беспокоиться, это удобный JSON, который легко читают и люди и машины. Сдать задание можно легко, не зная всех тонкостей YAML. Этот формат имеет много дополнительных возможностей, если интересно изучить подробнее, то рекомендуем: документация PyYAML

[2] результат умножается на 10 (максимальная оценка) и округляется до первой цифры после точки

## 2. FAQ (часто задаваемые вопросы)

**Никогда раньше не использовал YAML, как проверить валидность файла?**

Проверить валидность файла можно с помощью онлайн-ресурсов:

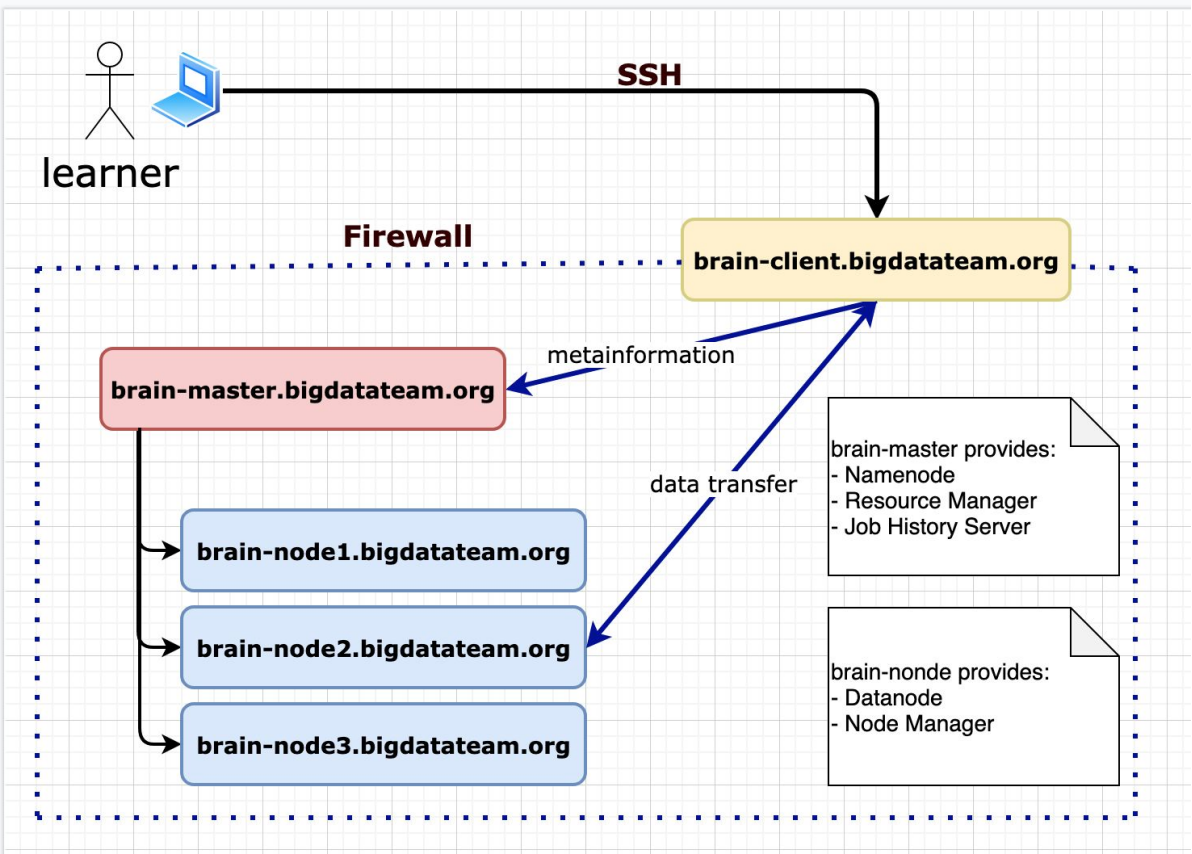
- https://yaml-online-parser.appspot.com/ (показывает результат парсинга)
- https://yamlchecker.com/ (есть встроенный syntax highlight)

Посколько онлайн ловит не все проблемы, затем в обязательном порядке в Python:

```
>>> yaml.safe_load(open("/path/to/solution.yml"))
```

**Можно поподробнее про разницу local FS и HDFS?**

В первом приближении Hadoop кластер выглядит следующим образом (клиентский узел - brain-client, мастер-сервер (где работает NameNode) - brain-master, рабочие узлы кластера - brain-node1, brain-node2, …):



Под local FS понимается файловая система brain-client, с которого есть доступ к кластеру. В случае эмуляции через Docker-контейнер - это та файловая система, в которую вы попадаете при запуске терминала (посредством открытия сессии bash или через интерфейс jupyter ноутбука).

Что такое SSH tunneling и где посмотреть примеры проброса портов с подробным описанием?

- Порт (компьютерные сети)
- https://en.wikipedia.org/wiki/Port_forwarding
- SSH tunnel
- SSH Port Forwarding Example

Никогда не пользовался консолью Linux, где можно получить максимально быстрый ликбез?

Для хорошего погружения рекомендуем “The Unix Workbench” в формате:

- книги: https://seankross.com/the-unix-workbench/
- или онлайн курса на Coursera: https://www.coursera.org/learn/unix

## 3. Задания уровня beginner

Задачи:

1. Пробросить порт (port forwarding) для доступа к HDFS Web UI[3]
2. `[task ID: beginner.how_many_items_in_hdfs, score: 1%]` Воспользоваться Web UI для того, чтобы найти папку "`/backup_virtual`" в HDFS, а в ней логи сервиса - "`access_log`". Сколько подпапок в папке "`/backup_virtual/access_logs`" без учета рекурсии? (**в ответе ожидается одно число**).

## 4. 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. флаги "`-ls`" и "`-R`", чтобы:

1. `[task ID: intermediate.hdfs_list_recursively, score: 3%]` Вывести рекурсивно список всех файлов в `/data/wiki`.
2. `[task ID: intermediate.hdfs_list_recursively_human_readable, score: 3%]` 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. `[task ID: intermediate.hdfs_file_replication_factor, score: 1.5%]` Ответьте на вопрос: какой фактор репликации используется для файлов? В случае работы с Docker-контейнером, к ответу прибавьте 2. (**в ответе ожидается одно число**)
4. `[task ID: intermediate.hdfs_folder_replication_factor, score: 1.5%]` Ответьте на вопрос: какой фактор репликации используется для папок? (**в ответе ожидается одно число**)
5. `[task ID: intermediate.hdfs_describe_size, score: 3%]` 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.

[3] См. User Guides

См. флаг “-du“

6. [task ID: intermediate.hdfs_cumulative_size, score: 3%] Приведите команду для получения размера пространства, занимаемого всеми файлами (с учетом рекурсии, но без учета фактора репликации) внутри “/data/wiki”. На выходе ожидается одна строка с указанием команды.

См. флаги “-mkdir” и “-touchz“

7. [task ID: intermediate.hdfs_create_folder, score: 3%] Создайте папку в домашней HDFS-папке Вашего пользователя, чтобы избежать конфликтов, на всякий случай используйте Ваш id (см. grades) в качестве префикса папки.
8. [task ID: intermediate.hdfs_create_nested_folder, score: 3%] Создайте вложенную структуру из папок одним вызовом CLI. Символы ‘;’ и ‘&’ в команде запрещены. Решить задачу нужно не объединением нескольких команд в одну строку, а вызовом одной команды.
9. [task ID: intermediate.hdfs_remove_nested_folders, score: 3%] Удалите созданные папки рекурсивно.
10. [task ID: intermediate.hdfs_trash_behavior, score: 3%] Что такое Trash в распределенной FS (ответ текстом)? Как сделать так, чтобы файлы удалялись сразу, минуя “Trash” (указать команду)?
11. [task ID: intermediate.hdfs_create_empty_file, score: 3%] Создайте пустой файл в HDFS.

См. флаги “-put”, “-cat”, “-tail”, “-cp”, “-get”, “-getmerge”

12. [task ID: intermediate.hdfs_create_small_file, score: 3%] Создайте небольшой произвольный файл (идеально - 15 строчек по 100 байт) и загрузите файл из локальной файловой системы (local FS)[4] в HDFS.
13. [task ID: intermediate.hdfs_output_file, score: 1%] Выведите содержимое HDFS-файла на экран.
14. [task ID: intermediate.hdfs_output_file_end, score: 1%] Выведите конец HDFS-файла на экран.
15. [task ID: intermediate.hdfs_output_file_start, score: 1%] Выведите содержимое нескольких первых строчек HDFS-файла на экран.
16. [task ID: intermediate.hdfs_tail_vs_unix_tail, score: 3%] Разберитесь в чем разница между HDFS флагом "-tail" и локальной утилитой "tail". С помощью какой команды (флага) можно воспроизвести поведение HDFS "-tail" локально?
17. [task ID: intermediate.hdfs_copy_file, score: 1.5%] Сделайте копию файла в HDFS.

[4] См. FAQ

18. `[task ID: intermediate.hdfs_move_file, score: 1.5%]` Переместите копию файла в HDFS на новую локацию.
19. `[task ID: intermediate.hdfs_download_and_concatenate, score: 3%]` Загрузите HDFS-файлы локально[5], объединив их в один файл во время загрузки одним вызовом CLI.

## 5. Задания уровня advanced

**Задачи на консольную утилиту "hdfs dfs"**

Полезные флаги:

- Для "`hdfs dfs`", см. "`-setrep -w`"
- `hdfs fsck /path -files - blocks -locations`

Задачи:

1. `[task ID: advanced.hdfs_set_file_replication, score: 6%]` Изменить replication factor для файла (команда). Как долго занимает время на увеличение / уменьшение числа реплик для файла (текст/обсуждение в чатах)?
2. `[task ID: advanced.hdfs_get_files_and_block, score: 6%]` Найдите информацию по файлу, блокам и их расположениям с помощью "`hdfs fsck`" CLI
3. `[task ID: advanced.hdfs_get_block_information, score: 6%]` Получите информацию по любому блоку из п.2 с помощью "`hdfs fsck -blockId`". Обратите внимание на Generation Stamp (GS number).
4. `[task ID: advanced.hdfs_dfs_architecture, score: 6%]` Выберите произвольный файл в HDFS. Узнайте из каких блоков он состоит. Воспользуйтесь пользователем hdfsuser[6], чтобы найти физические реплики этого блока на Datanode'ах. Также изучите структуру и содержимое snapshot (fsimage) и транзакций сервиса Namenode. Слепок файловой структуры Namenode (e.g. edits.log) предоставлен в HDFS по адресу "/data/namenode_example". Скопируйте все введенные команды в терминале, это будет являться ответом на задачу.

**Задачи на работу с сервисом WebHDFS**

См. документацию по адресу https://hadoop.apache.org/docs/r1.0.4/webhdfs.html

Цель - научиться делать запросы к Namenode (NN) и Datanode'ам (DN) с помощью curl.

[5] См. FAQ: на edge-ноду (client), с которой есть доступ к кластеру и где вы запускаете команду "hdfs dfs" в консоли.

[6] Для всех слушателей курсы мы сделали беспарольный доступ с помощью команды "`sudo -i -u hdfsuser`"

Пример запроса на чтение файла с помощью curl:

```
>> curl -i
"http://brain-master:50070/webhdfs/v1/data/access_logs/big_log/access.log.2015-12-10?op=OPEN"
```

Найдите по какому адресу (Location) на какую Datanode нужно обращаться для чтения данных из реплики.

Задачи:

1. `[task ID: advanced.webhdfs_read_100B, score: 6%]` Прочитайте 100B из произвольного файла в HDFS с помощью WebHDFS.
2. `[task ID: advanced.webhdfs_curl_follow_redirects, score: 6%]` Научитесь пользоваться опцией “follow redirects” с помощью `curl` (см. “`man curl`”).
3. `[task ID: advanced.webhdfs_get_file_detailed_information, score: 6%]` Получите детализированную информацию по файлу (см. `file status`)
4. `[task ID: advanced.webhdfs_change_file_replication, score: 6%]` Измените параметр репликации файла с помощью `curl`
5. `[task ID: advanced.webhdfs_append_to_file, score: 6%]` Дозапишите данные в файл (append). Подсказка - обратите внимание, что это запрос типа “POST”.

## 6. Правила оформления задания

Оформление задания:

- Код задания (Short name): **HW1:HDFS(Quiz)**.
- Выполненное ДЗ сохраните в файл **`MADEBD2021Q1`**`_<Surname>_<Name>_HW#.yml`, к примеру - **`MADEBD2021Q1`**`_Dral_Alexey_HW1.yml`.
- Для того, чтобы сдать задание необходимо:
  - Зарегистрироваться и залогиниться в сервисе Everest
  - Перейти на страницу приложения: BDT-grader-MADE-BD
  - Выбрать вкладку Submit Job (если отображается иная).
  - Выбрать в качестве “Task” значение: **HW1:HDFS(Quiz)** [7]
  - Загрузить в качестве “Task solution” файл с решением
  - В качестве Sender ID указать тот, который был выслан по почте
- Если Вы видите надпись "You are not allowed to run this application" во вкладке Submit Job в Everest, то на данный момент сдача закрыта (нет доступных для сдачи домашних заданий, по техническим причинам или другое). Попробуйте,

[7] Сервисный ID: hdfs.quiz

пожалуйста, еще раз через некоторое время. Если Вы еще ни разу не сдавали, у коллег сдача работает, но Вы видите такое сообщение, сообщите нам об этом.

- Ситуации:
  * система оценивания показывает оценку (Grade) < 0, а отчет (Grading report) не помогает решить проблему (пример помощи: в случае неправильно указанного Sender ID система вернет -2 и информацию о том, что его нужно поправить);
  * показывает 0 и в отчете (Grading report) не указано, какие тесты не пройдены.
  Если Вы столкнулись с какой-то из них присылайте ссылку на выполненное задание (Job) на почту с темой письма “Short name. ФИО.”. Например: “**HW1:HDFS(Quiz)**. Иванов Иван Иванович.”
  Пример ссылки: https://everest.distcomp.org/jobs/67893456230000abc0123def
  **Внимание:** Если до дедлайна остается меньше суток, и Вы знаете (сами проверили или коллеги сообщили), что сдача решений сломана, обязательно сдайте свое решение и напишите письмо, как написано выше, чтобы мы видели, какое решение Вы имели до дедлайна и смогли его оценить.
- Перед отправкой задания, оставьте, пожалуйста, отзыв о нём по ссылке: http://rebrand.ly/mailbd2021q1_feedback_hw. Это позволит скорректировать учебную нагрузку по следующим заданиям (в зависимости от того, сколько часов уходит на решение ДЗ), а также ответить на интересующие вопросы.

Любые вопросы / комментарии / предложения можно писать в Discord-канал курса или на почту bigdata_made2021q1@bigdatateam.org

Всем удачи!