BIGDATA TEAM

HW #01: HDFS proficiency

Deadline: 15.06.2019, 23:59

0. Описание задания и критериев оценивания.

Все ответы на вопросы, полученную из системы информацию, сравнения и результаты внесения изменений необходимо отобразить в файле домашней работы. Для оцениваемых заданий написан балл, который будет получен за выполнение, максимум 100 баллов. Шаблон находится по следующей ссылке:

- https://rebrand.ly/mf2019q2_hw_hdfs_template

1. Задания уровня beginner.

Задачи:

1. Пробросить порт (port forwarding) для доступа к HDFS Web UI[1]
2. [1 балл] Воспользоваться Web UI для того, чтобы найти папку “/data” в HDFS, а в ней логи какого-то сервиса (см. “access_log”). Сколько папок и файлов в указанной папке с логами?

[1] См. User Guides - http://rebrand.ly/mf2019q2_user_guides

2. 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. [3 балла] Вывести рекурсивно список всех файлов в /data/wiki.
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] Ответьте на вопрос: какой фактор репликации используется для файлов и папок?
4. [3 балла] Ответьте на вопрос: полученный выводе размер файла - это актуальный размер файла или же объем пространства, занимаемый с учетом всех реплик этого файла?

См. флаг “-du“

5. [3 балла] Получите размер пространства, занимаемый всеми файлами (с учетом рекурсии) внутри /data/wiki (т.е. на выходе ожидается одно число / одна строка)

См. флаги “-mkdir” и “-touchz“

6. [3 балла] Создайте папку в домашней HDFS-папке Вашего пользователя, чтобы избежать конфликтов, на всякий случай используйте Ваш id (см. grades) в качестве префикса папки.
7. [3 балла] Создайте вложенную структуру из папок одним вызовом CLI.
8. [3 балла] Удалите созданные папки рекурсивно.
9. [3 балла] Что такое Trash в распределенной FS? Как сделать так, чтобы файлы удалялись сразу, минуя “Trash”?
10. [3 балла] Создайте пустой файл в HDFS.

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

11. [3 балла] Создайте небольшой произвольный файл (идеально - 15 строчек по 100 байт) и загрузите файл из локальной файловой системы (local FS) в HDFS.
12. [3 балла] Выведите HDFS-файл, его начало и конец (аналог консольных утилит - cat / head/ tail).

13. [3 балла] В чем разница между HDFS флагом “-tail” и локальной утилитой “tail”? Каким образом воспроизвести поведение “-tail” локально?
14. [3 балла] Сделайте копию файла в HDFS и переместите его на новую локацию (аналог консольных утилит - cp, mv)
15. [3 балла] Загрузите HDFS-файлы локально, а также объедините их в один файл при загрузке.

3. Задания уровня advanced.

Полезные флаги:

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

Задачи:

1. [6 баллов] Изменить replication factor для файла. Как долго занимает время на увеличение / уменьшение числа реплик для файла?
2. [6 баллов] Найдите информацию по файлу и блокам с помощью “hdfs fsck” CLI
3. [6 баллов] Получите информацию по любому блоку из п.2 с помощью "hdfs fsck -blockId”. Обратите внимание на Generation Stamp (GS number).
4. [6 баллов] Воспользуйтесь пользователем hdfsuser[2], чтобы найти физические реплики на Datanode’ах и исследовать файловую структуру Namenode (e.g. edits.log)

Extras:

- Сравните поведение (локальной) консольной утилиты “find" и распределенной (HDFS) утилиты find (hdfs dfs -find).

4. Задания по WebHDFS.

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

Цель - научиться делать запросы к Namenode (NN).

Пример запроса на чтение файла с помощью curl:

[2] Для всех слушателей курсы мы сделали беспарольный доступ с помощью команды “sudo -i -u hdfsuser”

>> curl -i "http://virtual-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. [6 баллов] Получить данные файла размером в 100B.
2. [6 баллов] Научиться пользоваться опцией “follow redirects” с помощью curl (см. “man curl”).
3. [6 баллов] Получить детализированную информацию по файлу (см. file status)
4. [6 баллов] Изменить параметр репликации файла с помощью curl
5. [6 баллов] Дозаписать данные в файл (append). Подсказка - обратите внимание, что это запрос типа “POST”.

5. Сроки сдачи и правила оформления задания.

Deadline: 15.06.2019, 23:59

Оформление задания:

- Код задания (Short name): **HW1:HDFS**.
- Выполненное ДЗ сохраните в файл MF2019Q2_<фамилия>_HW#.xlsx, к примеру -- MF2019Q2_Ivanov_HW1.xlsx.
- Присылайте выполненное задание на почту bigdata_mf2019q2@bigdatateam.org с темой письма “Short name. ФИО.”. Например: “HW1:HDFS. Иванов Иван Иванович.”
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по сссылке: http://rebrand.ly/mf2019q2_feedback_hw01. Это позволит нам 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 телеграм-канале: http://rebrand.ly/mf2019q2_telegram_join
- На почту: bigdata_mf2019q2@bigdatateam.org

6. Дорешка.

Решения после получения фидбека на решение ДЗ можно улучшить. Разрешается одна досылка в течение 1й недели после окончания дедлайна за ДЗ. Соответственно, фидбек за дорешенные ДЗ вы получите в течение 24 часов после окончания deadline следующего ДЗ.

Дорешивать неработающие задания - нельзя. Это позволит исправить дисбаланс между

присланными **НЕработающими** заданиями **ДО** deadline

VS

присланными **работающими** заданиями **ПОСЛЕ** deadline

Всем удачи!