BIGDATA TEAM

#01: HDFS. Workshop.

1. Цель занятия.

Будем выделять 3 (содержательных) уровня владения HDFS:

0. Мимо проходил (passerby).
1. Новичок (beginner).
2. Продвинутый пользователь (intermediate).
3. Гуру (advanced).

Наша задача, начать прокачивать HDFS-скилы с уровня новичка до гуру и начнем мы это увлекательное занятие на семинаре. На заметку - только единицы в группе успевают выполнить все задания в рамках одного занятия. Поэтому не надо беспокоиться, если у Вы что-либо не успели. Всегда остается возможность продолжить погружение дома и иметь возможность спрашивать вопросы в Telegram-канале.

2. Определения.

Новичок:

- Умеет пробрасывать порт (port forwarding) для доступа к непубличным ресурсам
- Умеет пользоваться Web UI для того, чтобы просматривать HDFS.

Продвинутый пользователь умеет использовать “hdfs dfs” CLI для:

- Просмотра HDFS
- Перемещения файлов и папок в HDFS
- Перемещения файлов между локальной файловой системой (local FS) и HDFS

Гуру, должен уметь:

- Изменить реплику файла в HDFS с помощь “hdfs dfs” CLI
- Получить детальную информацию по файлам и блокам в HDFS с помощью “hdfs fsck” CLI
- Найти и прочитать содержимое реплики (replica’s data) на Datanode
- Найти и прочитать содержимое “edit.log” на Namenode
- Пользоваться curl для работы с HDFS через WebHDFS API

Для отслеживания прогресса в сравнении с остальными членами группы, мы будем пользоваться “Poll” в Telegram.

3. Задания.

3.1. Задания для beginner.

Задачи:

1. Пробросить порт (port forwarding) для доступа к HDFS Web UI[1]
2. 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?

3.2. Задания для 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. Вывести рекурсивно список всех файлов в /data/wiki.

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

2. 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. Ответьте на вопрос: какой фактор репликации используется для файлов и папок?
4. 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. 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. Создайте папку в домашней HDFS-папке Вашего пользователя, чтобы избежать конфликтов, на всякий случай используйте Ваш id (см. grades) в качестве префикса папки.
7. Создайте вложенную структуру из папок одним вызовом CLI.
8. Удалите созданные папки рекурсивно.
9. Что такое Trash в распределенной FS? Как сделать так, чтобы файлы удалялись сразу, минуя “Trash”?
10. Создайте пустой файл в HDFS.

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

11. Создайте небольшой произвольный файл (идеально - 15 строчек по 100 байт) и загрузите файл из локальной файловой системы (local FS) в HDFS.
12. Выведите HDFS-файл, его начало и конец (аналог консольных утилит - cat / head/ tail).
13. В чем разница между HDFS флагом “-tail” и локальной утилитой “tail”? Каким образом воспроизвести поведение “-tail” локально?
14. Сделайте копию файла в HDFS и переместите его на новую локацию (аналог консольных утилит - cp, mv)
15. Загрузите HDFS-файлы локально, а также объедините их в один файл при загрузке.

3.3. Задания для advanced.

Полезные флаги:

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

Задачи:

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

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:

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

4. Обратная связь

Обратная связь: http://rebrand.ly/mf2019q2_feedback_01_hdfs

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

Просьба потратить 1-2 минут Вашего времени, чтобы поделиться впечатлением, описать что было понятно, а что непонятно. Мы учитываем рекомендации и имеем возможность переформатируем учебную программу под Ваши запросы.