BIGDATA TEAM

Как работать с HDFS, используя docker-образ HYSH

1. Установка Docker

Установить Docker можно как на Linux системы, так и на Mac OS и даже Windows. Если у Вас есть выбор, на какую систему поставить Docker, то наш совет - смотреть в сторону Linux (основная платформа для Docker, инструкции далее также основаны на опыте работы с именно системой Ubuntu этого семейства) и не смотреть в сторону Windows (немалое количество недоступных функций и вопросов совместимости).

В Интернете можно найти множество инструкций по установке, но официальной, как правило, достаточно:

- https://docs.docker.com/engine/install/

Для перехода к инструкции для конкретной системы необходимо нажать на ее название в таблице поддерживаемых платформ. Во время ожидания установки можно ознакомиться со следующим разделом

2. Что такое docker-образ HYSH

Для обучения мы разработали специализированный docker-*образ* (проводя аналогию с установкой операционной системы с диска, образ - это установочный флэш-накопитель/диск), который позволяет пользователю на своей системе быстро развернуть *контейнер* (если говорить, что образ - это флэш-накопитель/диск (носитель информации), то контейнер - это запущенная/установленная с носителя система) с системой с установленными **H**DFS, **Y**ARN, **S**park, **H**ive. Внутри контейнера, конечно, нет name/master node и data/worker node как самостоятельных систем, но концепции эти реализованы в виде виртуальных узлов по одному каждого типа. Аналогом консоли клиентского узла в кластере здесь играет система, в которую Вы попадаете, запуская контейнер.

3. Запуск HYSH-контейнера для изучения HDFS

Следующая *однострочная bash*-команда (в инструкции строки переносятся) запустит контейнер и подключит терминал контейнера к терминалу Вашей системы:

docker run -it --rm -v $(pwd)/hysh_work:/home/jovyan/work -p 50070:50070 --name hysh_cont bigdatateam/hysh-data:py3-wikitwitter bash

Предупреждение: если Ваша система не поддерживает мультиоконный режим по умолчанию (например, версия Ubuntu без графического интерфейса), то Вы потеряете доступ к своей основной системе, пока не закроете контейнер; избежать эту проблему в Linux системах можно с помощью оконного менеджера *tmux*.

Разберем команду:

- *docker run* - команда docker для запуска контейнера;
- *-it* - аргумент, который позволяет подключить терминал Вашей системы к терминалу контейнера;
- *--rm* - удаляет контейнер после его отключения/выхода из его терминала, например, командой *exit* или сочетанием клавиш *Ctrl+c* . **Предупреждение:** при удалении контейнера крайне вероятно (есть некоторые исключения с большими сложностями) не сможете восстановить файлы, если они не сохранены в правильном месте, а также историю команд;
- *$(pwd)* - это команда *bash* (популярное терминальное приложение в Linux-системах), которая возвращает абсолютный путь до текущей директории. С большой вероятностью Ваше терминальное приложение также поддерживает эту команду.
  Для Windows: https://stackoverflow.com/a/41489151
- *-v $(pwd)/hysh_work:/home/jovyan/work* - командует монтировать (если отсутствует, то предварительно создать) директорию hysh_work в текущей директории Вашей системы к контейнеру как директорию */home/jovyan/work* . Все файлы, которые Вы сохраните в директории */home/jovyan/work*, будут доступны на Вашей системе в директории *$(pwd)/hysh_work* , остальные файлы в контейнере могут быть потеряны (см. **Предупреждение** для *--rm);*
- *-p 50070:50070* позволяет в браузере Вашей системы посмотреть HDFS UI по адресу http://localhost:50070/ . Если Ваша система доступна из Интернета и незащищена Firewall-ом, то более безопасным способом будет поменять стандартный порт на нестандартный, например, так: *-p 51170:50070*, но это изменение надо постоянно иметь в виду, потому что инструкции написаны под стандартный порт, а проверяющая система ожидает ссылки с использованием стандартного порта. Также, если Вы запускаете docker на удаленной системе, то Вам понадобится сделать проброс портов (Port Forwarding);
- *--name hysh_cont* устанавливает выбранное имя контейнера вместо случайно сгенерированного;
- *bigdatateam/hysh-data:py3-wikitwitter* - название docker-образа;
- *bash* - команда, которую должен запустить контейнер, в данном случае контейнер запустит терминальное приложение *bash*.

4. Полезная литература

Для более глубокого погружения и повышения навыков взаимодействия с перечисленными системами рекомендуем обратиться к следующим ресурсам:

- Официальная инструкция для начинающих пользователей Docker
- Одна из инструкций для начинающих пользователей Bash-терминала
- Официальная инструкция для начинающих пользователей tmux

5. FAQ

Я не могу найти директорию/файлы в директории *hysh_work* не обновились. Написал "*-v hysh_work:/home/jovyan/work*" (*иначе говоря, указал только имя директории*).
Значением до двоеточия для *-v* может быть как *абсолютный путь* (начинающийся с /, такой путь выдает *$(pwd)*) так и название именованного виртуального диска для контейнера. Соответственно, ошибка не выводится, потому что это считается валидным именем, но при этом это другая сущность.