Практика HDFS

1. Практика посвящена работе с Hadoop, и нам, по классике, хотелось бы установить себе Hadoop локально. Однако устанавливать Hadoop локально в свою ОС не самое лучшее решение(хотя вполне рабочее), т.к. основным компонентом Hadoop является HDFS – т.е. целая файловая система и соседство двух ФС на одном рабочем ноутбуке может привести к множеству проблем. Поэтому мы рекомендуем 2 решения – установка экосистемы Hadoop на виртуальную машину или установка через Docker. Безусловно есть возможность познакомиться с экосистемой Hadoop воспользовавшись услугами облачных провайдеров, но у данного решения тоже есть свои издержки, в первую очередь - финансовые.

a. В данном примере мы покажем как устанавливать экосистему Hadoop с помощью виртуальной машины. Поэтому нам в первую очередь нужна сама виртуальная машина, самое популярное решение на рынке – это VirtualBox. Скачайте VirtualBox с оф. сайта <https://www.virtualbox.org/wiki/Downloads> и установите на ваш ПК.

b. Далее нам необходимо скачать официальную сборку Cloudera(один из крупнейших поставщиков решений в экосистеме Hadoop) - <https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.zip>

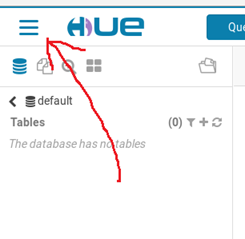
c. После скачивания и установки VirtualBox вам необходимо импортировать сборку Cloudera в VirtualBox и выполнить запуск. Перед запуском желательно выделить немного больше ресурсов для виртуальной машины. В идеале – 8 гб оперативки и 2 ядра процессора. После запуска виртуальной машины у вас должно получится следующее



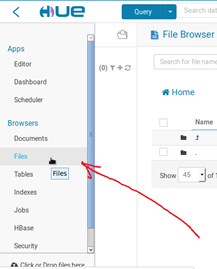
d. Если вы видите данную картину, значит установка экосистемы Hadoop в виртуальной машине прошла успешно. Поздравляем! Теперь вам доступны для использования HDFS, Hive, Spark и прочее. В первую очередь рекомендуем ознакомиться с HUE – программным обеспечением, которое дает возможность взаимодействовать с экосистемой Hadoop с помощью пользовательского интерфейса.

 e. Для авторизации в HUE используйте следующее сочетание логина и пароля по умолчанию:  cloudera/cloudera.

 f. После авторизации в HUE с помощью иконки меню в левом верхнем углу

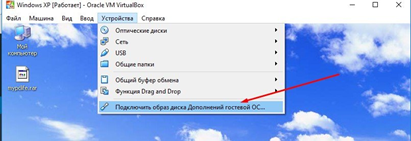


Выберите раздел Files

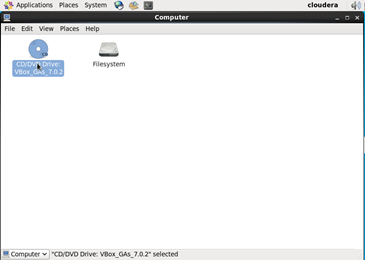


Поздравляем – вы только что попали в файловую систему HDFS, а конкретно в home-папку вашего пользователя. Пока в ней пусто, но это ненадолго.

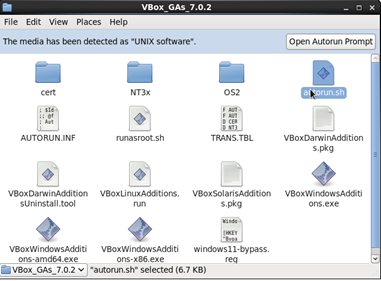
g. Теперь нам необходимо активировать возможность обмениваться файлами между вашей локальной ОС и виртуальной гостевой ОС. Для этого активируйте подключение образа диска гостевой ОС.



Теперь перейдите в Computer и убедитесь, что у вас появился образ диска



Переходите внутрь и выполните autorun.sh



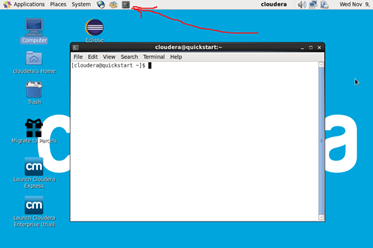
После выполнения autorun.sh – перезапустите вашу виртуальную машину.

h. Теперь вам доступна функция drag&drop и вы можете с помощью курсора мыши перетаскивать файлы из вашей локальной ОС в вашу виртуальную ОС. Переходим на сайт <https://all-the-books.ru/authors/tolstoy-lev-nikolaevich/> и скачиваем все тома произведения «Война и мир» в формате txt.

i. Далее перетаскиваем эти файлы в любое место на вашей виртуальной машине.

j. Теперь мы приступаем к изучению базовых команд работы с файловой системой HDFS. Для взаимодействия с HDFS мы будем использовать терминал. Ознакомиться со списком базовых команд hdfs можно в официальной документации: <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html#du>

k. Запускаем терминал



И выполняем следующие действия

• Когда мы перетащили файлы с произведением Льва Толстого – мы перетащили их в файловую систему виртуальной машины, но не в HDFS, соответственно, в первую очередь нам нужно перенести их в папку нашего пользователя именно на HDFS.

#Создаю каталог с именем /hadoop

bin/hadoop dfs -mkdir / hadoop

• После того, как файлы окажутся на HDFS попробуйте выполнить команду, которая

выводит содержимое папки. Особенно обратите внимание на права доступа к вашим файлам.

hadoop fs -ls  / user/training/hadoop

• Далее сожмите все 4 тома в 1 файл

Так как:

**hadoop archive -archiveName NAME -p <parent path> <src>\* <dest>**

**-NAME: имя сжатого файла, которое вы можете взять самостоятельно;**

**-parent dir: родительский каталог, в котором находится сжатый файл;**

**-src: имя файла для сжатия;**

**-hdfs dst: путь для хранения сжатых файлов**

**ex:**

**hadoop archive -archiveName hadoop.har -p /user 1.txt 2.txt /des**

**# Сжать файлы 1.txt и 2.txt в каталоге / user в hdfs в файл с именем hadoop.har и сохранить их в каталоге / des в hdfs**

**# Если 1.txt, 2.txt не записаны, все каталоги и файлы в каталоге / user сжимаются в файл с именем hadoop.har и сохраняются в каталоге / des в hdfs**

**# Чтобы отобразить содержимое har, вы можете использовать следующую команду: hadoop fs -ls /des/hadoop.har**

**# Показать файлы, сжатые с помощью har, вы можете использовать следующую команду: hadoop fs -ls -R har: ///des/hadoop.har**

Но лучше второй подход:

hadoop fs -cat [dir]/\* | hadoop fs -put - [destination file]

Однако здесь загружаеть призодится все данные из HDFS на узел, с которого выполняются команды (локальный). Только потом они загружаются в HDFS. Это не слишком эффективно, если много данных

• Теперь давайте изменим права доступа к нашему файлу. Чтобы с нашим файлом могли взаимодействовать коллеги, установите режим доступа, который дает полный доступ для владельца файла, а для сторонних пользователей возможность читать и выполнять.

**-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...**

**-R: Рекурсивная модификация**

**# Пользователь команды должен быть владельцем или суперпользователем файла**

• Попробуйте заново использовать команду для вывода содержимого папки и обратите внимание как изменились права доступа к файлу.

hadoop fs -ls /user/training/hadoop

• Теперь попробуем вывести на экран информацию о том, сколько места на диске занимает наш файл. Желательно, чтобы размер файла был удобночитаемым.

hadoop fs -df hdfs:/

• На экране вы можете заметить 2 числа. Первое число – это фактический размер файла, а второе – это занимаемое файлом место на диске с учетом репликации. По умолчанию в данной версии HDFS эти числа будут одинаковы – это означает, что никакой репликации нет – нас это не очень устраивает, мы хотели бы, чтобы у наших файлов существовали резервные копии, поэтому напишите команду, которая изменит фактор репликации на 2.

hadoop fs -cat [dir]/\* > tmp\_local\_file

hadoop fs -copyFromLocal tmp\_local\_file

hadoop fs -cp /user/training/\*.txt /user/training/hadoop

• Повторите команду, которая выводит информацию о том, какое место на диске занимает файл и убедитесь, что изменения произошли.

hdfs dfs -count [-q] [-h] <paths>

• Напишите команду, которая подсчитывает количество строк в вашем файле

hadoop fs -cat hadoop/purchases.txt

• В качестве результатов вашей работы, запишите ваши команды и вывод этих команд в отдельный файл и выложите его на github.

Дополнительные ссылки:

• Видеоинструкция по установке cloudera на VirtualBox: <https://www.youtube.com/watch?v=8rsNXdvFqBc>

• Docker-образ экосистемы Hadoop: <https://github.com/big-data-europe/docker-hadoop-spark-workbench>