# **Лабораторная работа №1. Установка виртуальной машины с** **Hadoop от Cloudera для работы с проектами**

# **экосистемы Hadoop**

Изучается процесс установки на собственном компьютере проекта верхнего уровня организации Apache Software Foundation - Hadoop и проектов его экосистемы для надежной, масштабируемой и распределенной обработки больших данных.

# **1. Установка VirtualBox**

Наиболее быстро познакомиться с распределенными системами Hadoop и hive можно при помощи средств виртуализации, например, VirtualBox.

В настоящей работе описан процесс установки виртуальной машины с использованием программного продукта *Oracle VM VirtualBox* (далее VirtualBox). Если кто-то предпочитает другую программу для виртуализации, то можно проделать то же самое и с вашим программным продуктом.

VirtualBox — программный продукт виртуализации для операционных систем Microsoft Windows, Linux, FreeBSD, macOS, Solaris/OpenSolaris, ReactOS, DOS и др.

Виртуализация - это предоставление набора вычислительных ресурсов или их логического объединения, абстрагированное от аппаратной реализации, и обеспечивающее при этом логическую изоляцию друг от друга вычислительных процессов, выполняемых на одном физическом ресурсе.

Примером использования виртуализации является возможность запуска нескольких операционных систем в облаке на одном компьютере: при том каждый из экземпляров таких гостевых операционных систем работает со своим набором логических ресурсов (процессорных, оперативной памяти, устройств хранения), предоставлением которых из общего пула, доступного на уровне оборудования, управляет хостовая операционная система — гипервизор. Также могут быть подвергнуты виртуализации сети передачи данных, сети хранения данных, платформенное и прикладное программное обеспечение: см. эмуляция.

Для установки ***Oracle VM VirtualBox*** для своей операционной системы скачиваем дистрибутив с сайта <https://www.virtualbox.org/wiki/Downloads> и устанавливаем его. Если возникают какие-то сложности, то по запросу в Google «Установка VirtualBox» можно найти десятки подробных инструкций (например: <https://faqpc.ru/kak-ustanovit-virtualbox/>), поэтому нет смысла описывать это здесь подробно.

**2**. **Установка виртуальной машины с** **Hadoop от Cloudera**

А) Переходим по ссылке <https://www.cloudera.com/downloads/quickstart_vms/5-12.html> выбираем в окне платформу VirtualBox (возможно к настоящему моменту, версия обновится и вас переадресует на более свежую версию образа, скачивайте его и выполняйте те же самые шаги)

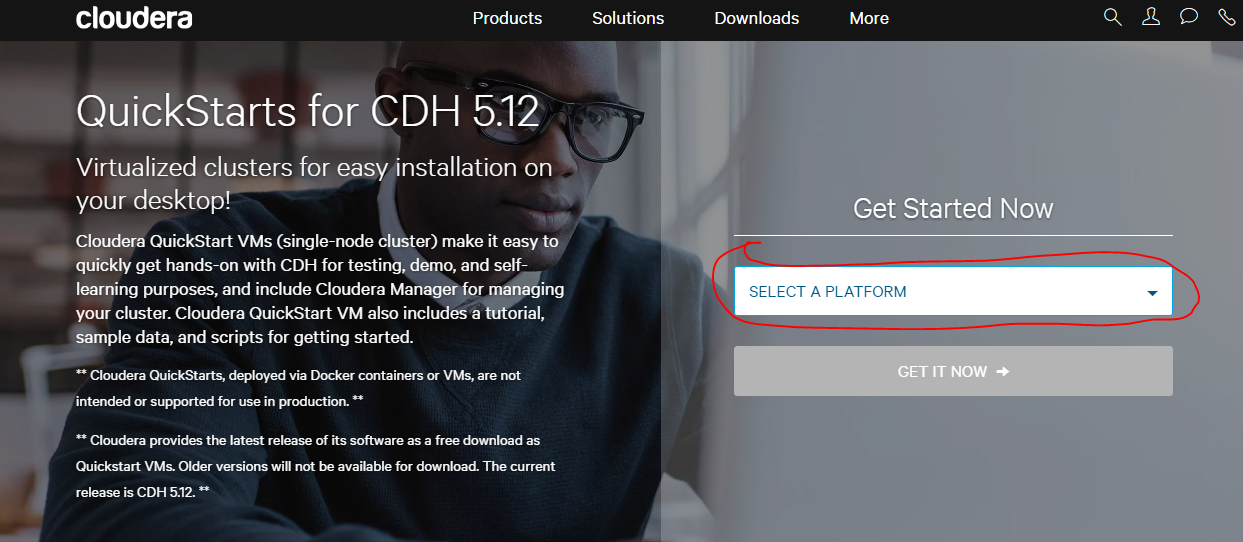


Рис. 1.

Б) Заполняем форму (писать можно, что угодно, но лучше что-то понятное)

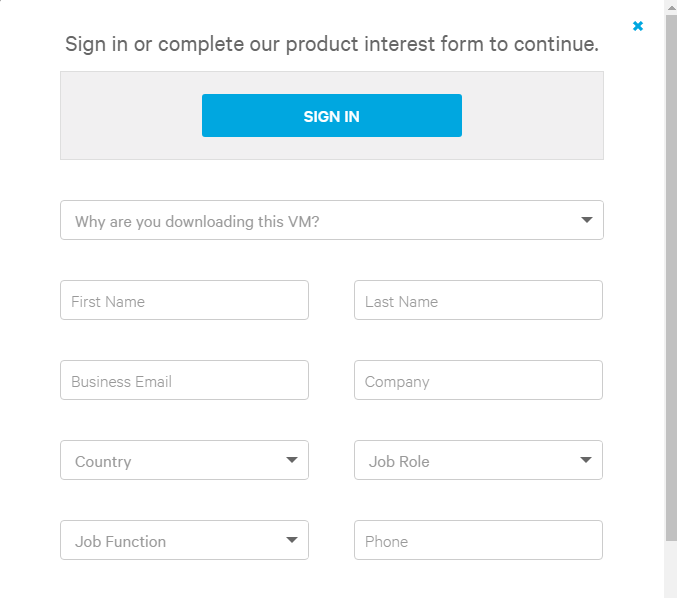


Рис.2.

В) Соглашаемся с лицензией

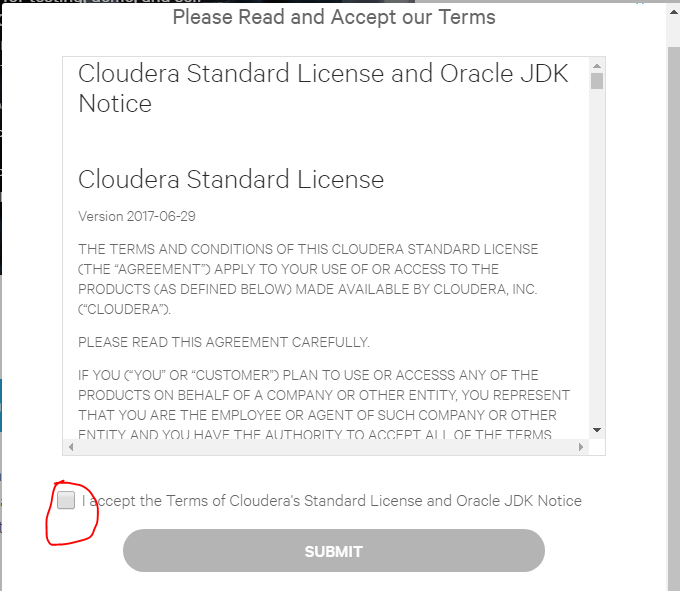


Рис. 3.

После нажатия кнопки SUBMIT (рис. 3) автоматически начнется скачивание, нужно дождаться его окончания.

Г) После окончания загрузки, необходимо разархивировать скачанный архив куда-нибудь и запомнить это место.

Д) Запускаем VirtualBox. Далее нажимаем Файл->Импорт конфигураций. Должно открыться такое окно (рис. 4):

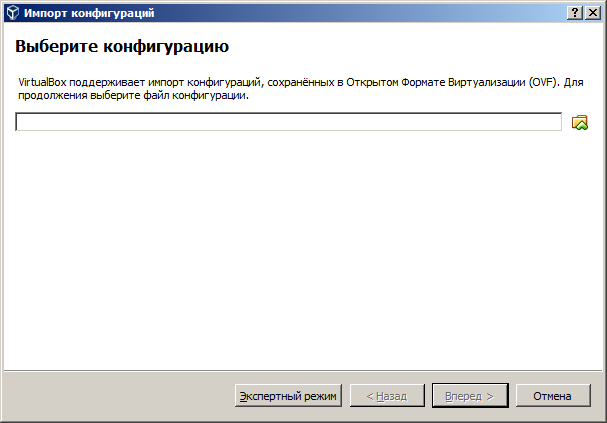


Рис. 4.

Далее нажимаем на значок папки (желтый с зеленой стрелкой) и выбираем папку в которую мы ранее разархивировали скачанный архив. В этой папке должен быть виден только один файл:

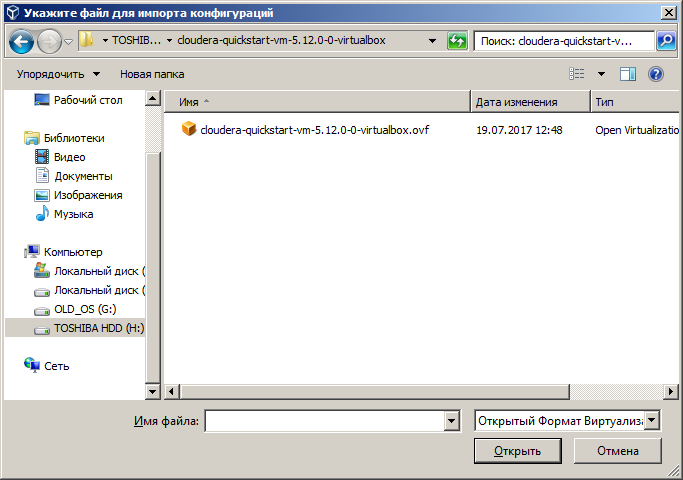


Рис.5.

Открываем этот файл. (Также это можно сделать непосредственно из графического интерфейса VirtualBox Файл->Импорт конфигураций)Нажимаем клавишу «Вперед»

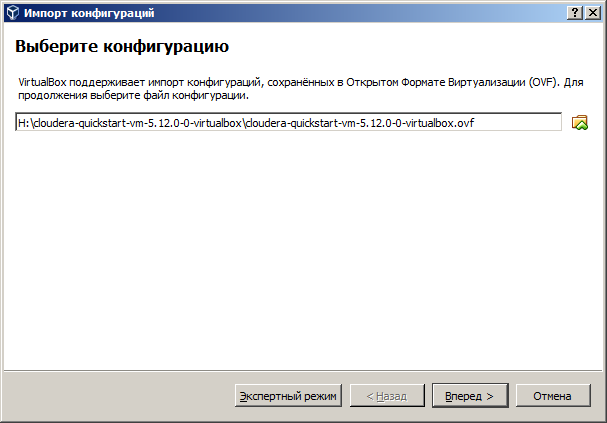


Рис.6.

Должно открыться окно (рис. 7):

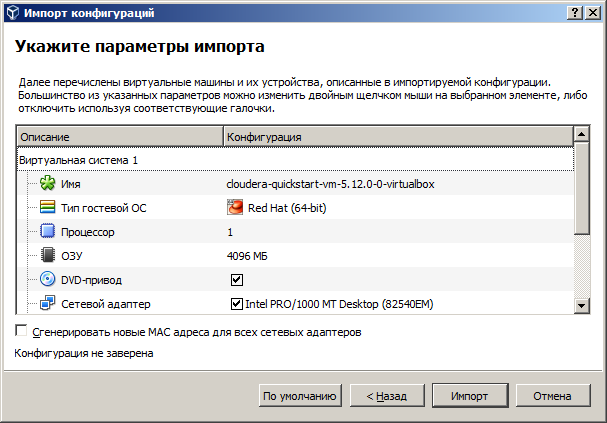


Рис.7.

По возможности необходимо установить наибольшее возможное значение ОЗУ, но не более ¾ всей доступной, тогда виртуальная машина будет работать быстрее.

Также рекомендуется сменить путь к виртуальному образу диска, так как он довольно объемный: 64 Гб, выберите тот диск, где у вас достаточно места. Для этого необходимо два раза щелкнуть по полю с путем к виртуальному образу диска (рис. 8), указать необходимую папку и дополнительно оставить название диска, то есть:

C:\Users\Kolychev\VirtualBox VMs\cloudera-quickstart-vm-5.12.0-0-virtualbox**\cloudera-quickstart-vm-5.12.0-0-virtualbox-disk1.vmdk**

То, что выделено жирным шрифтом, оставляется, остальное меняется на то, что вам удобно

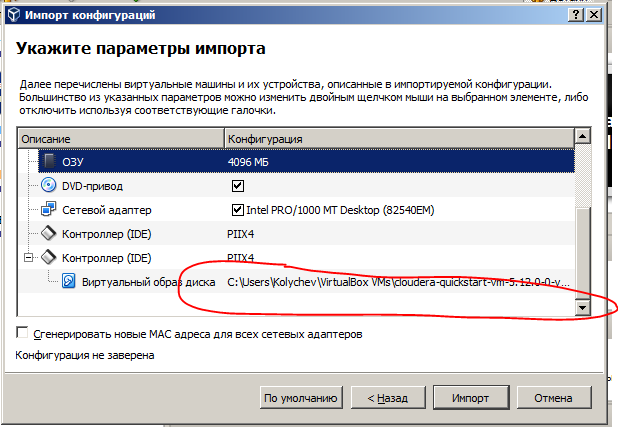


Рис.8.

После нажатия кнопки «Импорт» запустится импорт конфигурации (процесс займет около 10 минут):

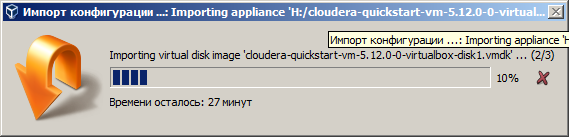


Рис.9.

После окончания виртуальная машина будет готова к использованию. Запускаться виртуальная машина будет довольно долго (5-10 минут), это абсолютно нормально, так как со стартом Linux запускаются все демоны Hadoop, Hive и других программных продуктов экосистемы, включенных в данный дистрибутив Hadoop.

Для удобства работы рекомендуется включить двунаправленный буфер, для того чтобы можно было копировать и вставлять из гостевой ОС в основную и наоборот: Устройства->Общий буфер обмена -> Двунаправленный

# **3**. **Работа с Hive**

После того как запуститься виртуальная машина и загрузится гостевая система, автоматически откроется браузер (Если этого не произойдет, запустите его самостоятельно). Для начала работы с Hive (рис. 10) необходимо нажать на закладку Hue, логин/пароль cloudera/cloudera

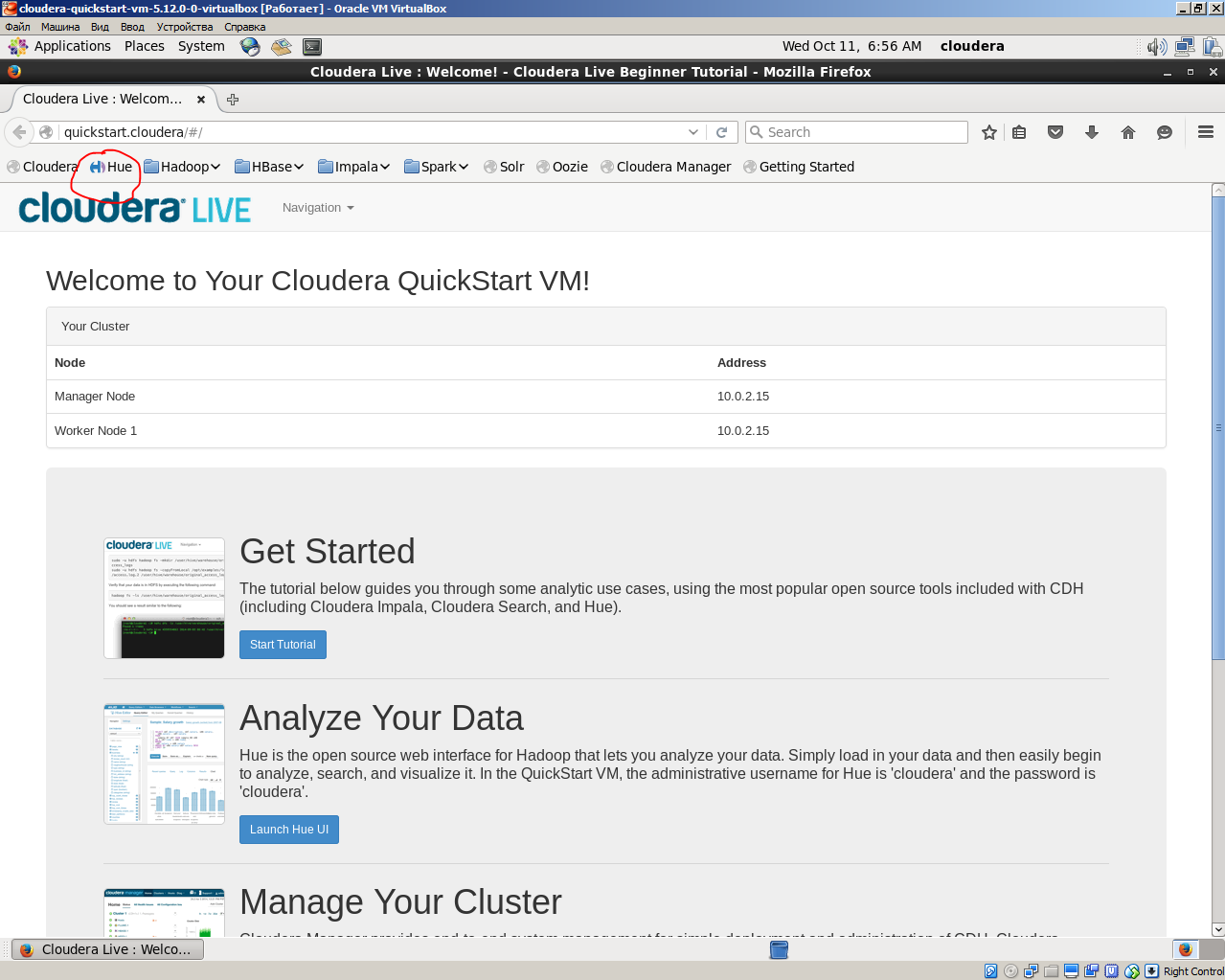


Рис.10.

Далее необходимо выбрать (рис. 11) Query Editor -> Hive и Вы у цели.

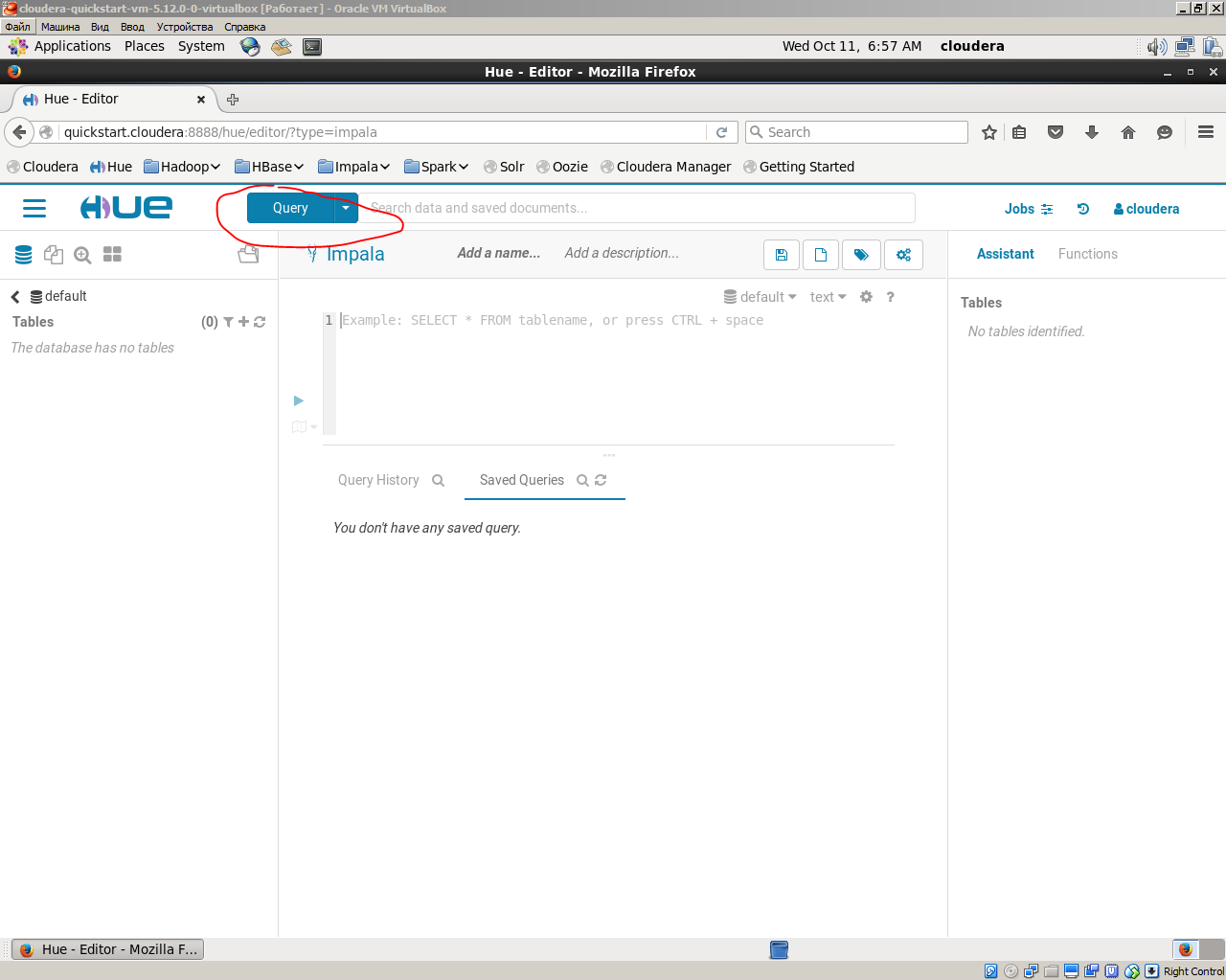


Рис.11.

Для того чтобы перенести какой-нибудь файл в хранилище Hive для начала работы с ним, необходимо создать директорию в **Hive** командой:

hadoop fs -mkdir /user/hive/warehouse/<имя директории>/

Для копирования файла из локальной файловой системы в хранилище Hive используется команда:

hadoop fs -copyFromLocal <путь к файлу в локальной файловой системе> <путь к каталогу созданному в Hive>

***Пример:***

hadoop fs -copyFromLocal /home/cloudera/Desktop/price /user/hive/warehouse/price/

Здесь файл расположенный на рабочем столе скопируется в каталог price в хранилище Hive.