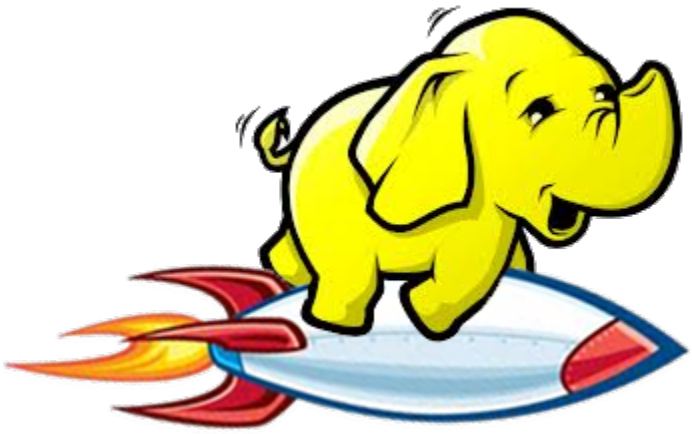
Развернуть Hadoop кластер из 5 машин с помощью Cloudera Manager



Задача: Развернуть работающий Hadoop-кластер с помощью бесплатного продукта Cloudera Manager.

Внимание! На каждую попытку дается 2 часа, этого должно хватить вам с запасом! Всего у вас 3 попытки. В результате вы сможете проверить правильность выполнения лабораторной работы с помощью автоматического скрипта автопроверки прямо в вашем личном кабинете.

Дедлайн: Четверг 12 марта 19:00
Жесткий дедлайн (минус 30%): Четверг 19 марта 19:00

Для того, чтобы начать работу войдите в свой личный кабинет с помощью вашей электронной почты @newprolab.ru

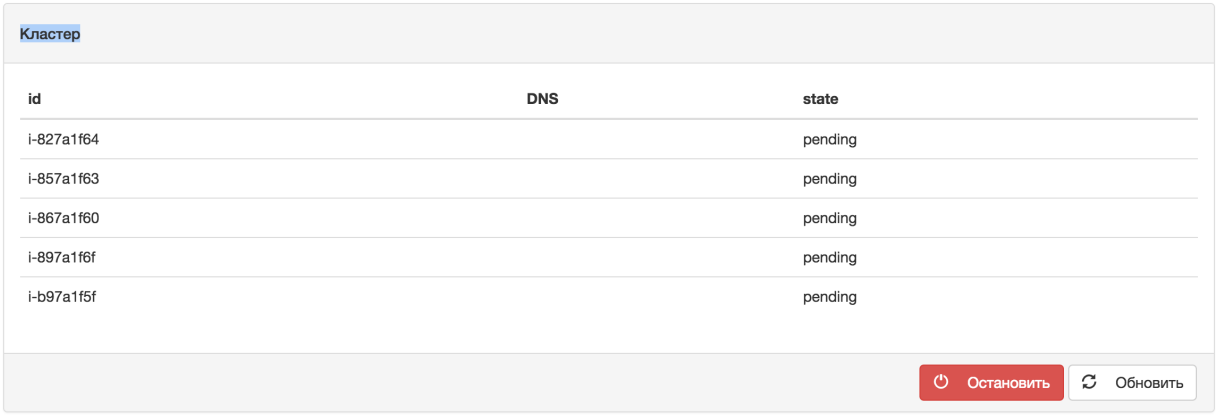
Если у вас нет ящика на @newprolab.ru, отправьте запрос на npl@digitaloctober.com

Получаем доступ к виртуальным машинам

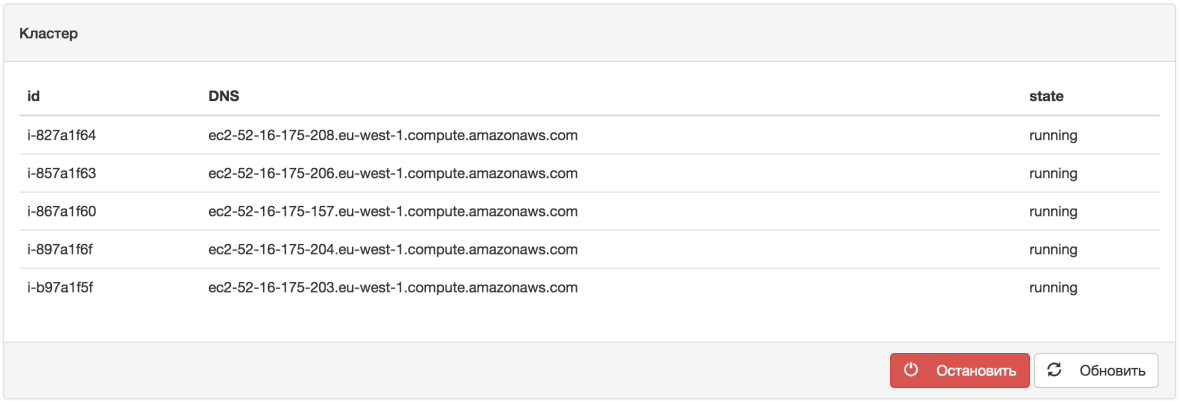
Зайдите в специальный кабинет для резервирования машин
http://bigdata.newprolab.com/amazon/

Зарезервируйте персональный пул виртуальных машин, нажав на кнопку “Запустить”.

Получите список машин для работы!



Подождите немного, потом нажмите кнопку “Обновить”, чтобы получить актуальную информацию о статусе запуска машин.

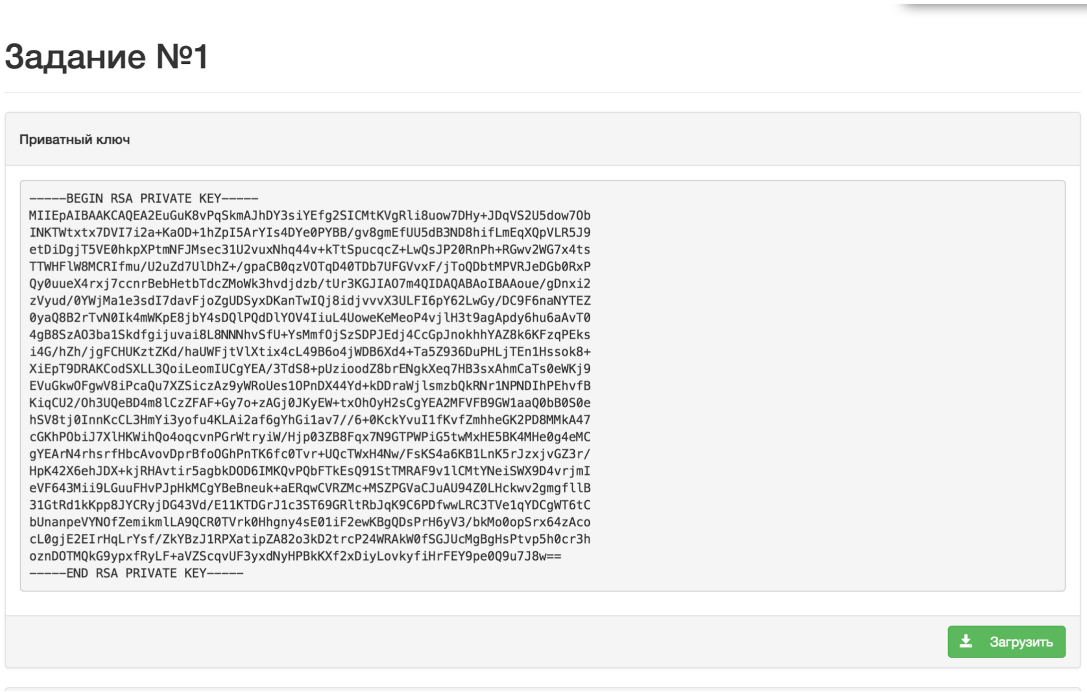


Для теста нам выделили следующие виртуальные машины:
ec2-52-16-84-121.eu-west-1.compute.amazonaws.com (здесь будет менеджер)
ec2-52-16-120-196.eu-west-1.compute.amazonaws.com
ec2-52-16-124-102.eu-west-1.compute.amazonaws.com

ec2-52-16-121-86.eu-west-1.compute.amazonaws.com
ec2-52-16-121-54.eu-west-1.compute.amazonaws.com

Подготовка приватного ключа

На этой же странице вам необходимо скачать приватный ключ. По умолчанию он сохранен в файл private.pem

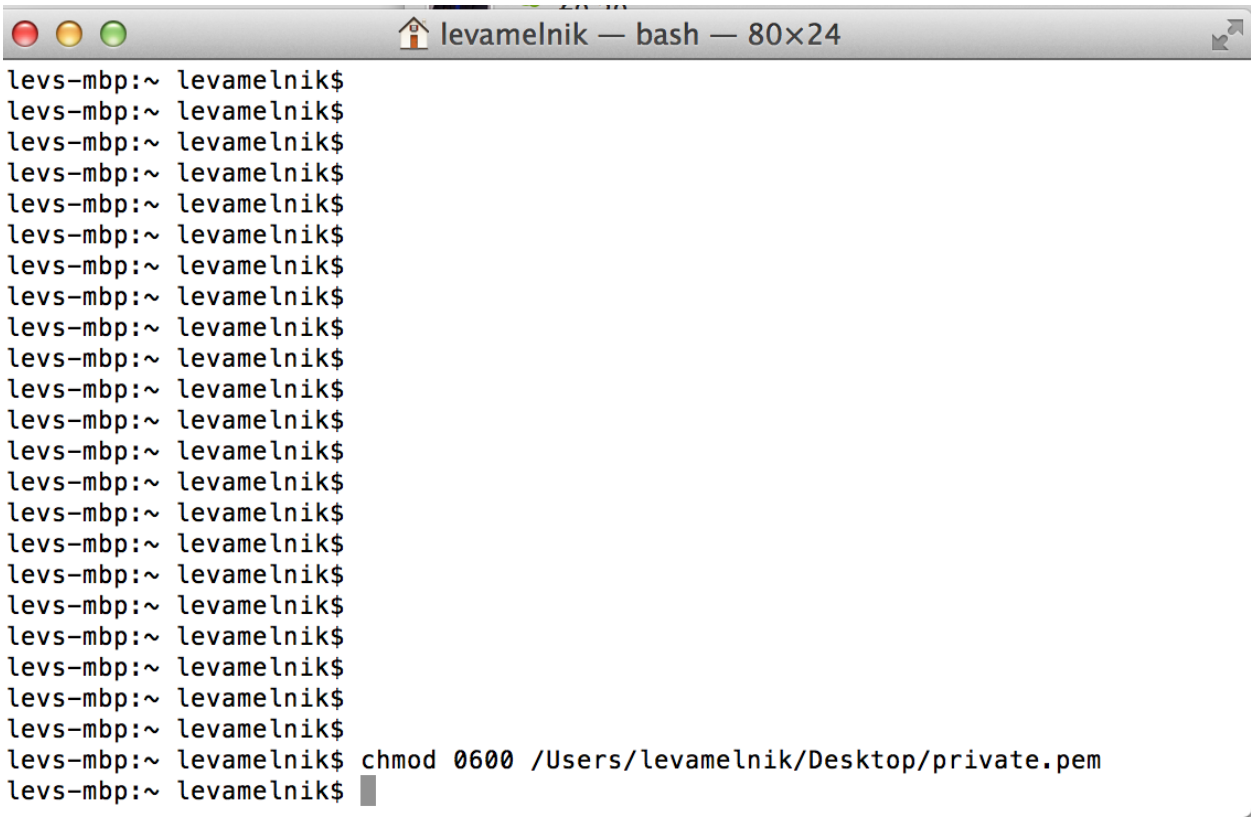


Выставляем права на доступ к ключу

Зайдите в консоль (мы будем использовать консоль в MacOS, но можно войти из любой операционной системы).

Если вы работаете под Windows, то можете воспользоваться инструментом Putty
http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html

```
chmod 0600 private.pem
```

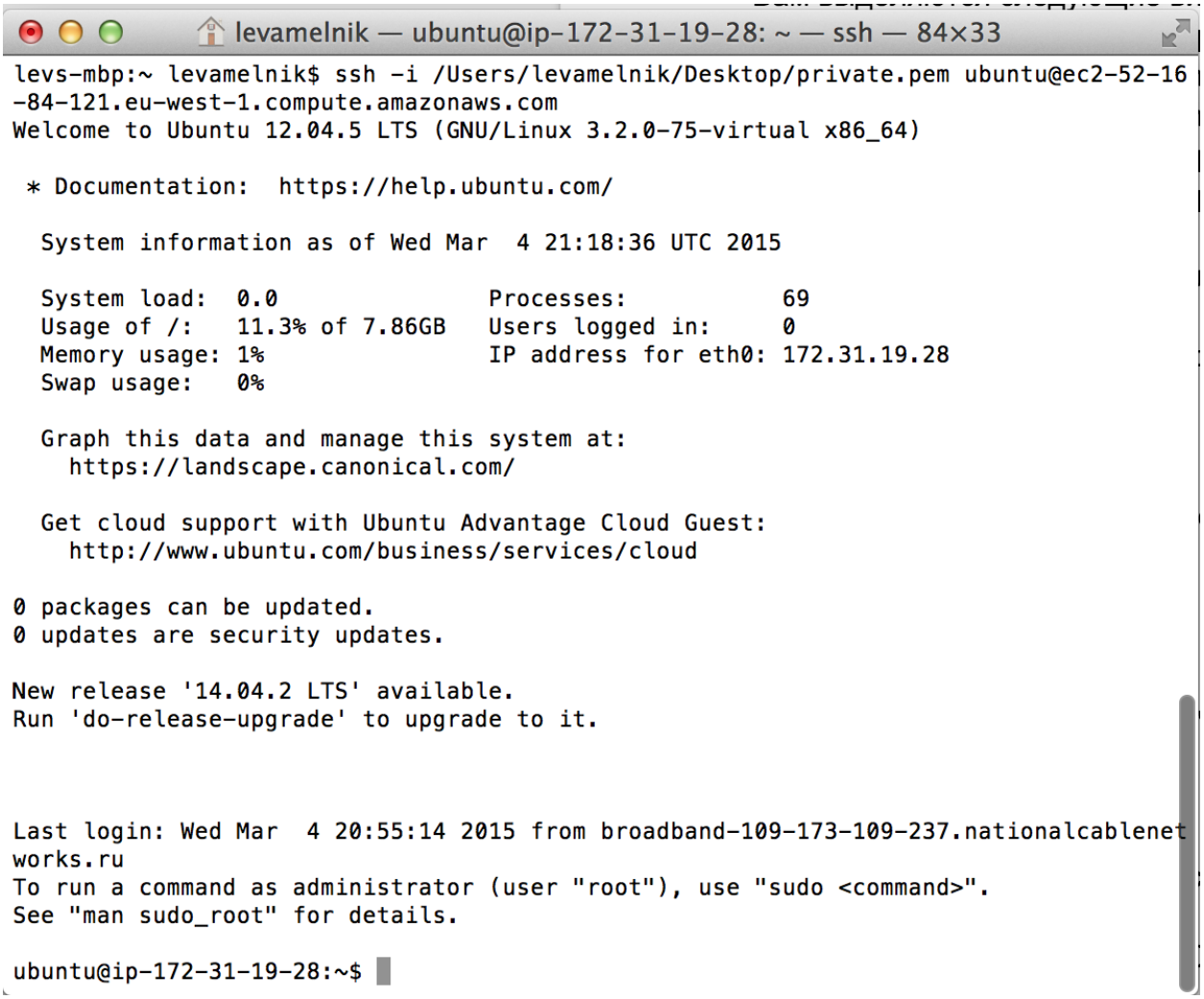


С помощью этой команды ключ становится доступным только текущему пользователю, в противном случае ssh будет считать его "слишком открытым" и "небезопасным".

Зайти на сервер №1 (сервер-менеджер)

Заходим на первый по списку сервер по SSH
ssh -i private.pem ubuntu@адрес_сервера

В нашем наборе первым идет сервер под номером (84-121)



В случае, если вы работаете с PuTTY, то вам придется переформатировать ключ в формат .ppk.
Можете почитать - http://support.cdh.ucla.edu/help/132-file-transfer-protocol-ftp/583-converting-your-private-key-

Инструкция по Putty

1. Загружаете приватный ключ с http://bigdata.newprolab.com/, как написано на странице 3.
2. Запускаете puttygen, Conversions -> Import key. Выбираете скачанный ключ. После этого - Save private key. Соглашаетесь сохранить без passphrase. Сохраняете в private.ppk.

3. Запускаете Putty, в Host Name вводите: ubuntu@HOST, где HOST - машина, выданная вам для кластера первой. Далее слева выбираете Connection -> SSH -> Auth, в Private Key указываете на private.ppk.
4. Возвращаетесь в Sessions влева, вводите в Save Session какое-нибудь имя, и Save. Это чтобы не вводить все второй раз. Потом - Open. Начнется соединение. В первый раз появится предупреждение, согласиться.

Скачиваем дистрибутив Cloudera Manager

wget http://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin

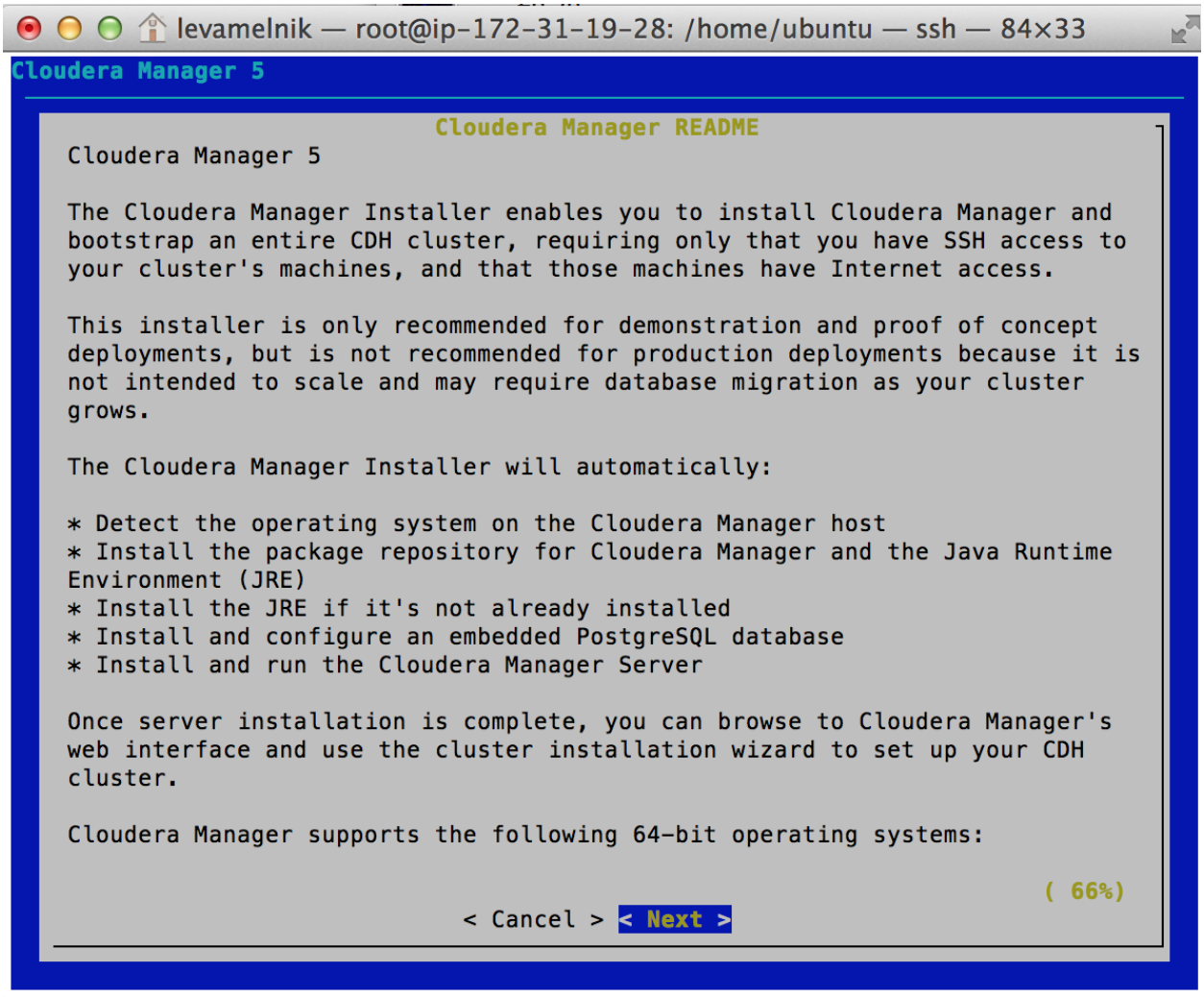
```
levamelnik — root@ip-172-31-19-28: /home/ubuntu — ssh — 84×33
    https://landscape.canonical.com/

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.

New release '14.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.



Last login: Wed Mar  4 20:55:14 2015 from broadband-109-173-109-237.nationalcablenet
works.ru
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ip-172-31-19-28:~$ wget http://archive.cloudera.com/cm5/installer/latest/clou
dera-manager-installer.bin
--2015-03-04 21:19:11--  http://archive.cloudera.com/cm5/installer/latest/cloudera-m
anager-installer.bin
Resolving archive.cloudera.com (archive.cloudera.com)... 54.230.196.188, 54.230.198.
12, 54.230.198.223, ...
Connecting to archive.cloudera.com (archive.cloudera.com)|54.230.196.188|:80... conn
ected.
HTTP request sent, awaiting response... 200 OK
Length: 514295 (502K) [application/octet-stream]
Saving to: `cloudera-manager-installer.bin'

100%[======================================>] 514,295     2.86M/s   in 0.2s

2015-03-04 21:19:11 (2.86 MB/s) - `cloudera-manager-installer.bin' saved [514295/514
```

Получаем права администратора

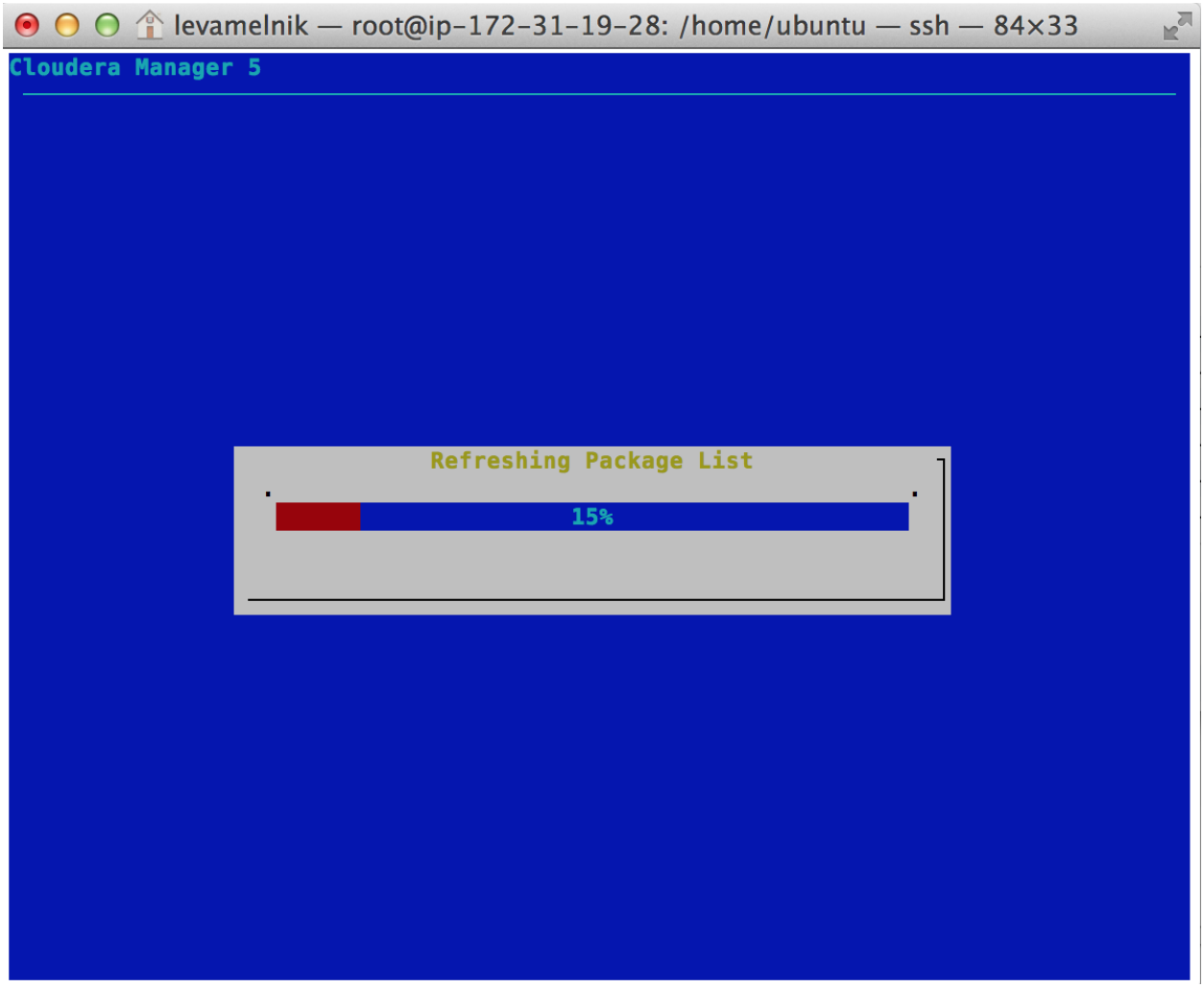
sudo su

Делаем дистрибутив исполняемым
chmod +x cloudera-manager-installer.bin

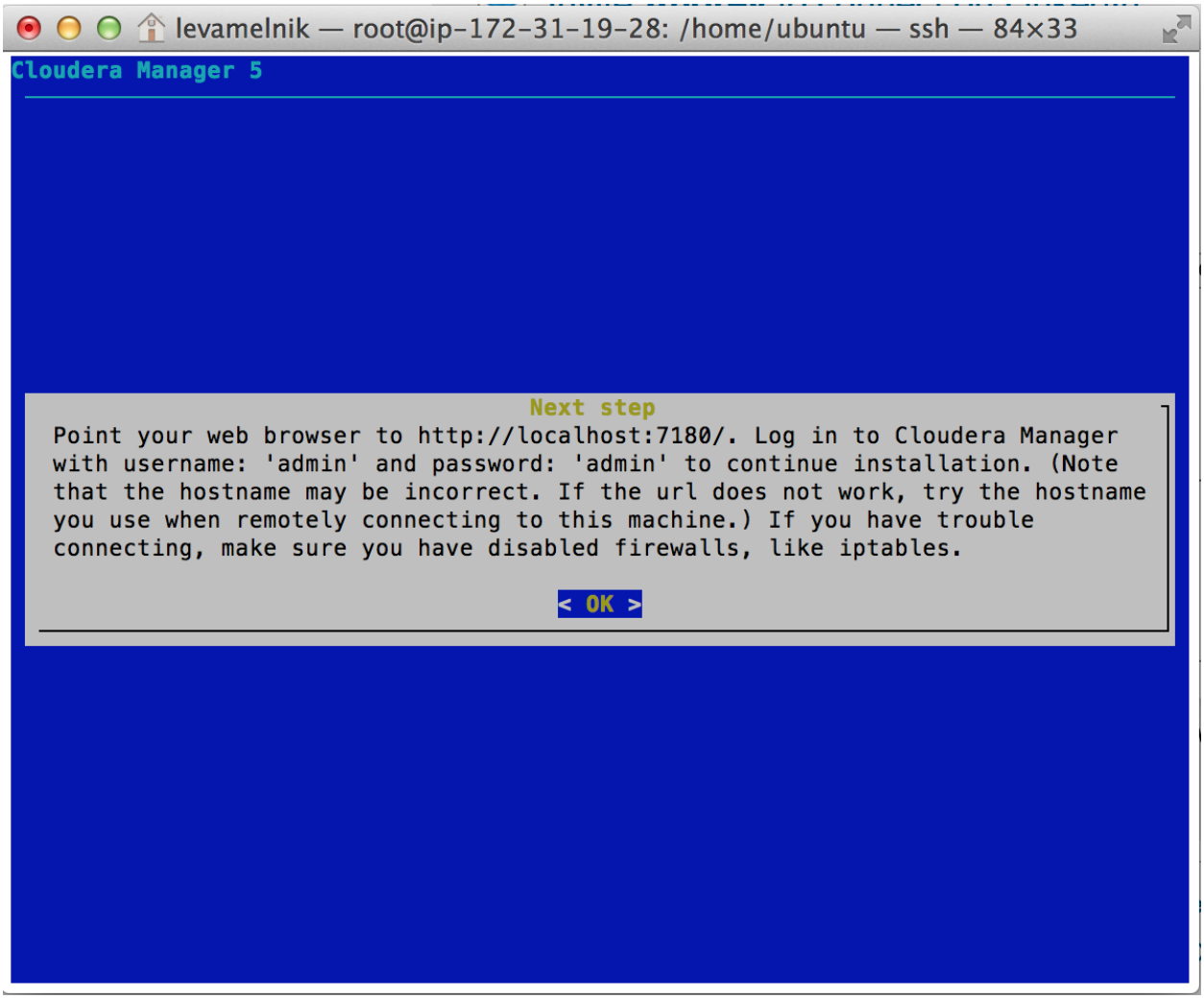
Запускаем дистрибутив
./cloudera-manager-installer.bin



На все вопросы отвечаем утвердительно, и ждём окончания установки.

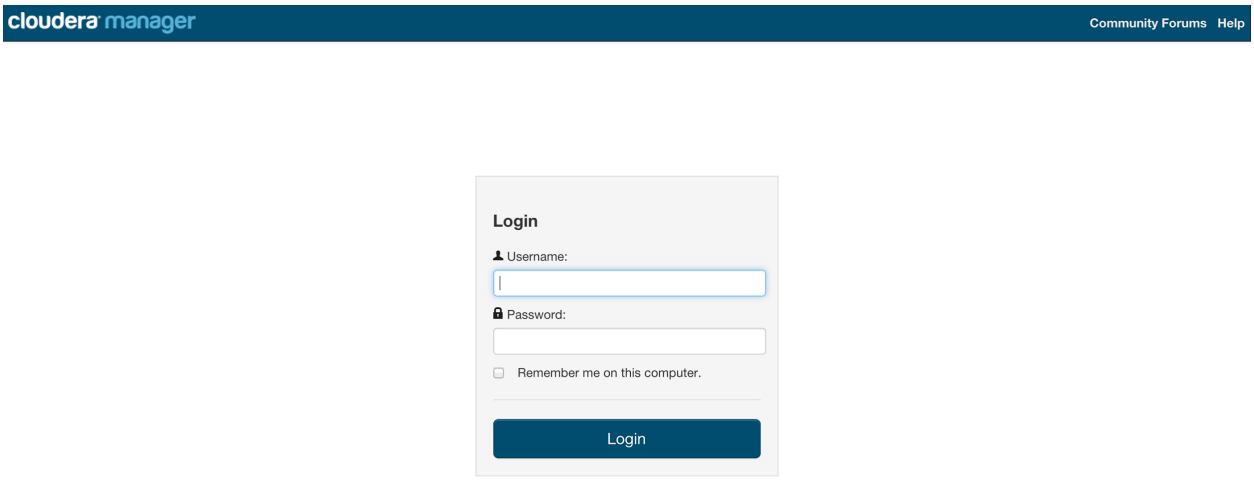


В финале будет сообщение вида "Зайдите на <domainname>:7180"



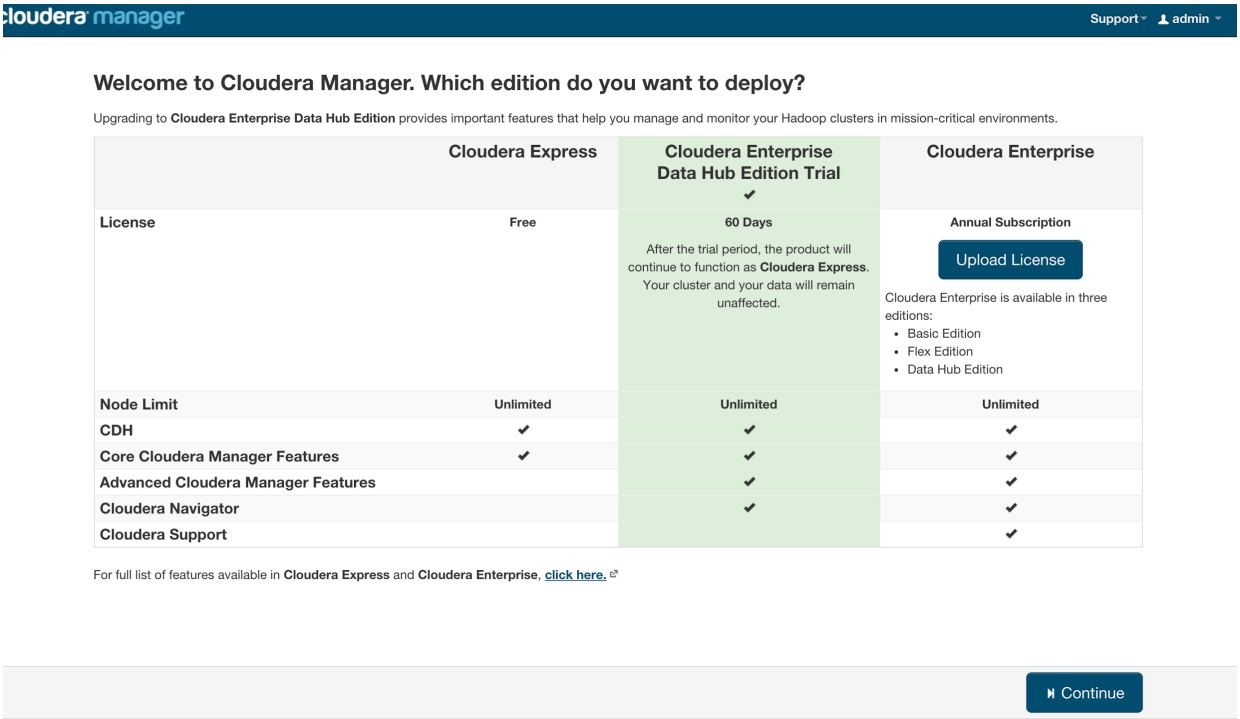
Все остальные шаги по установке делаются уже через браузер. В консоль вернёмся только потом для проверки.

Заходим браузером по адресу http://<domainname>:7180 (в нашем примере сервер ec2-52-16-84-121.eu-west-1.compute.amazonaws.com:7180)



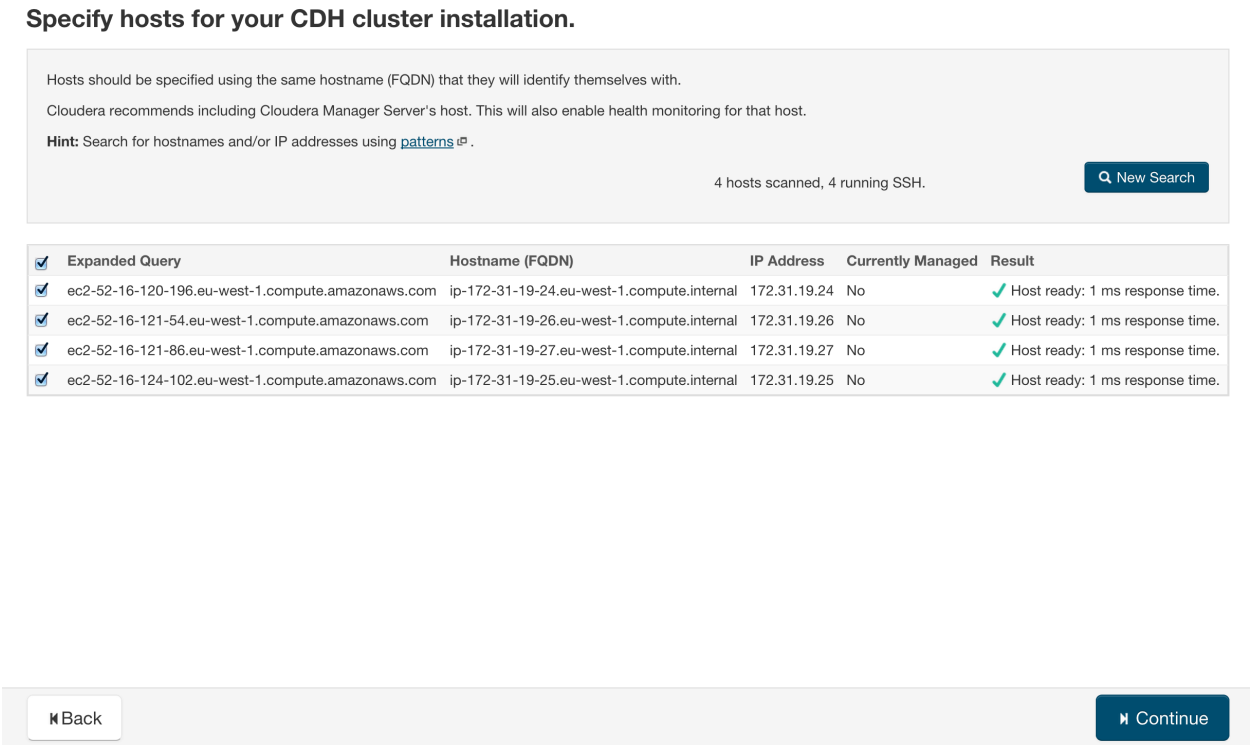
Логин/ пароль
admin/admin

Continue



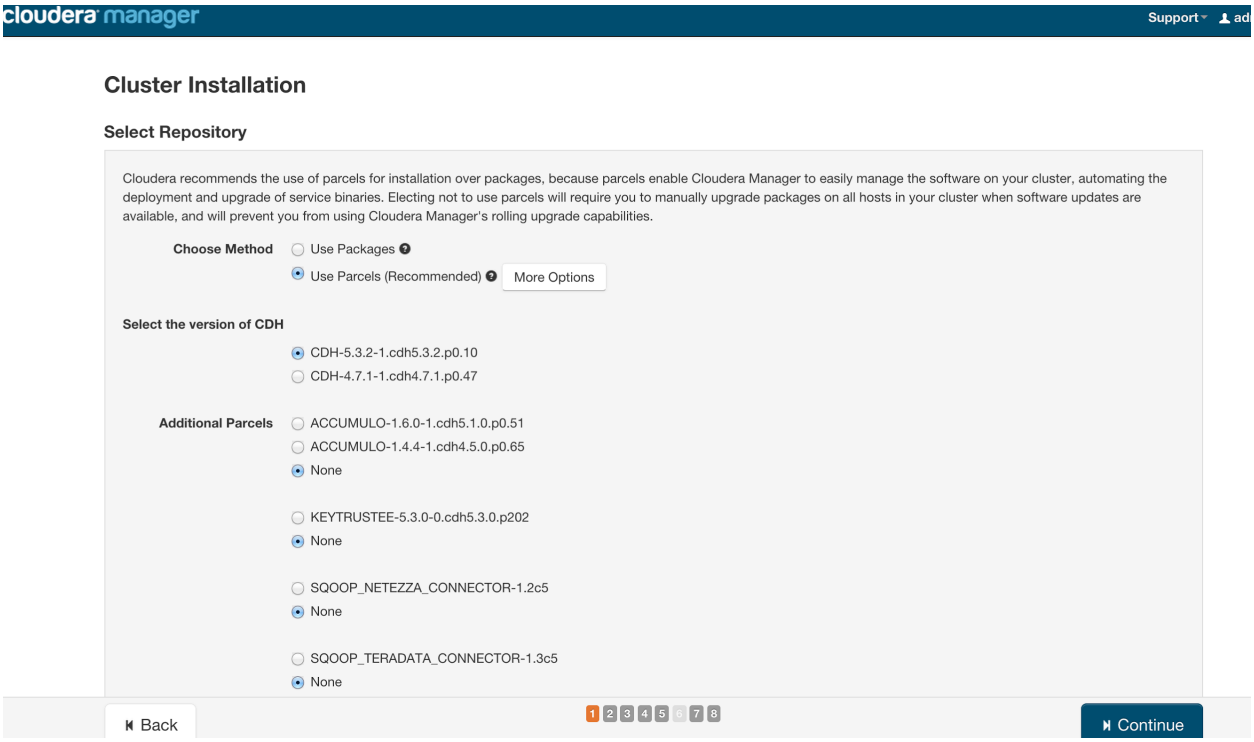
Continue

На следующем шаге необходимо вписать оставшиеся 4 домена

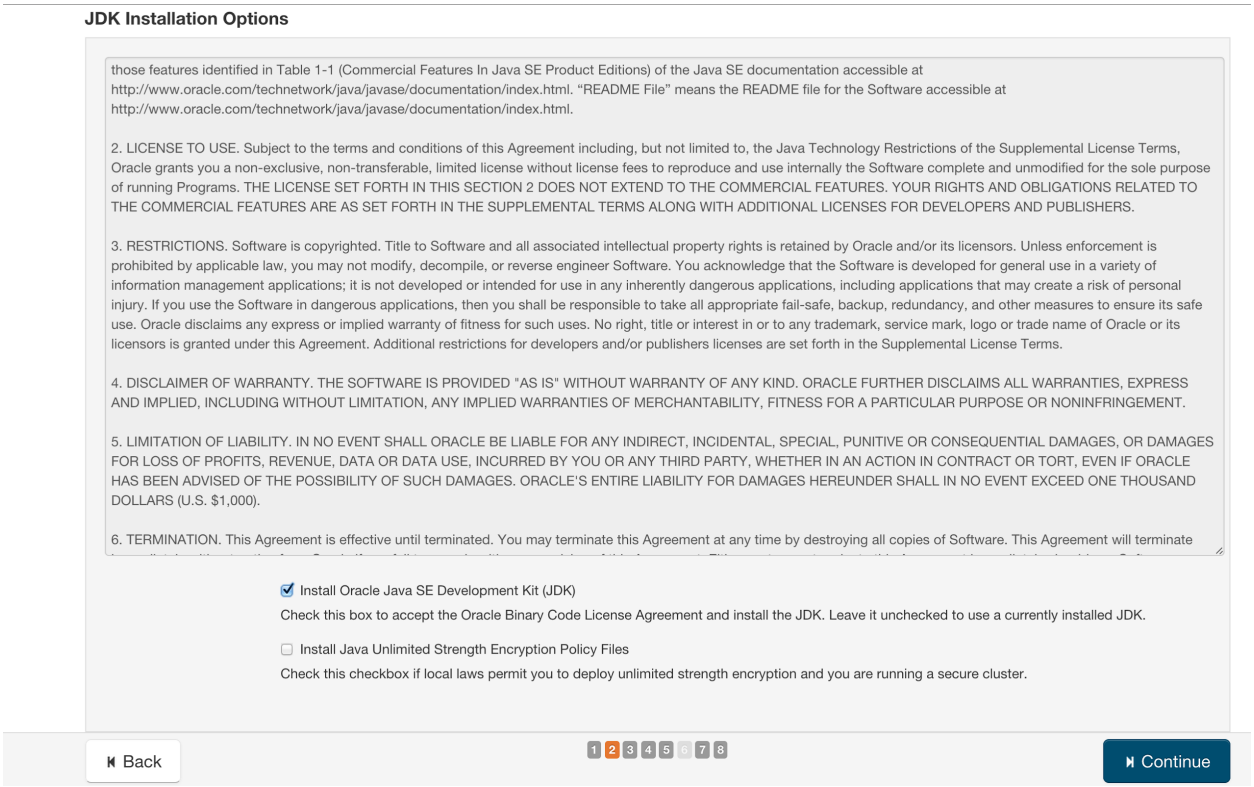


Continue

Cluster Installation / Select Repository - тоже Continue



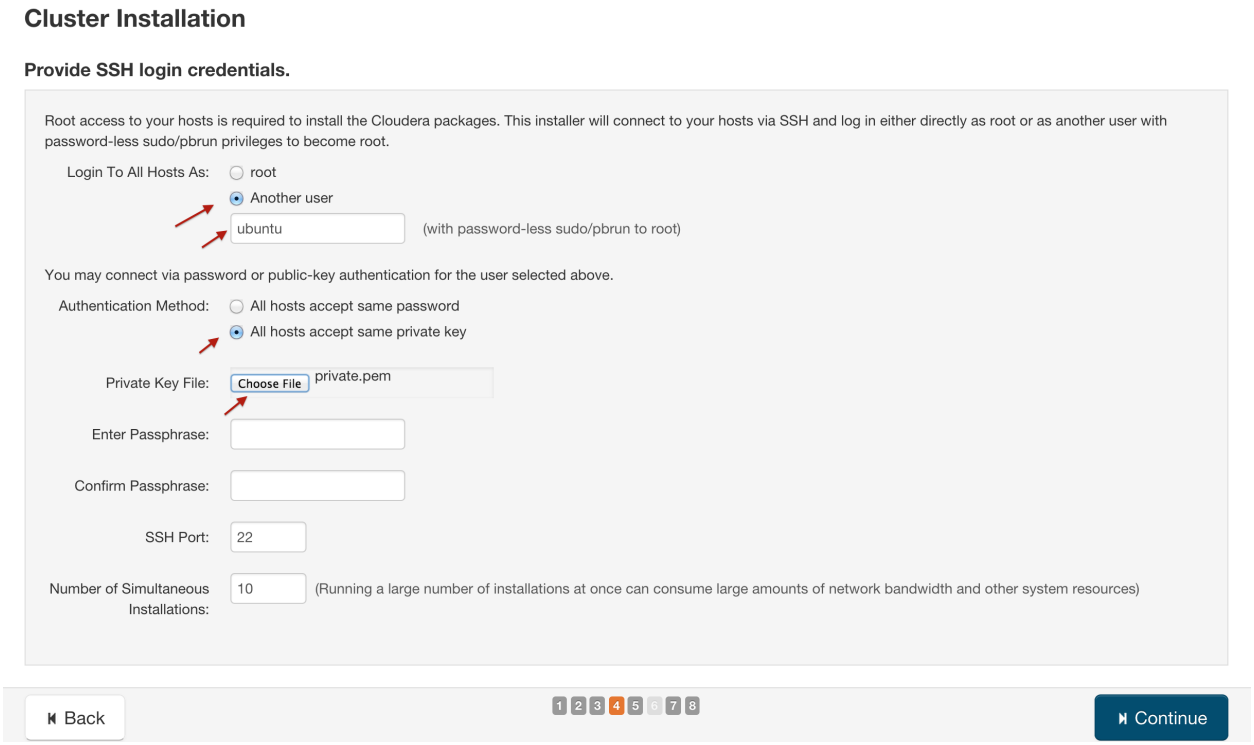
На следующем шаге необходимо поставить галочку напротив Oracle Java SE Development Kit



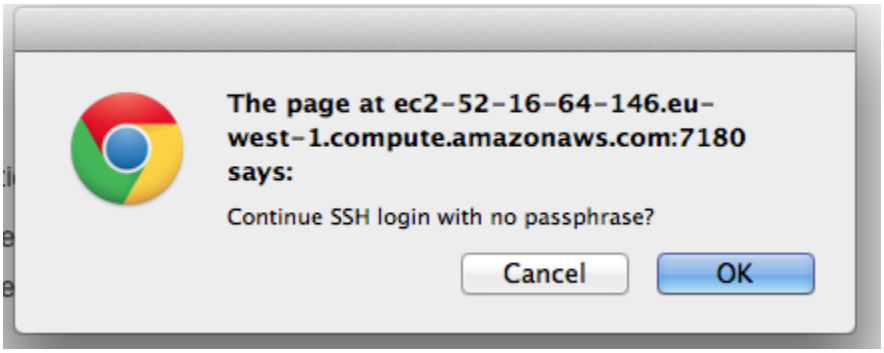
Cluster Installation / Enable Single User Mode - Continue

Hint! Галку «enable single mode» не ставить!

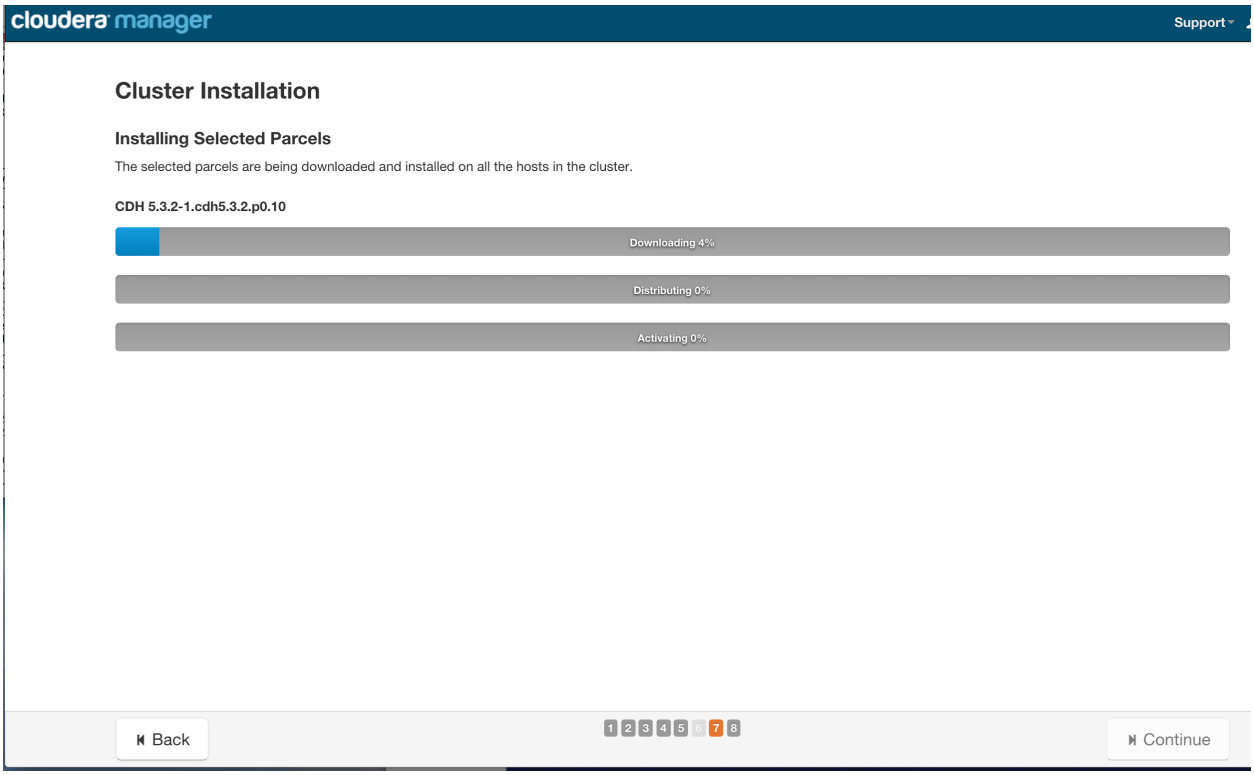
Выбираем правильные параметры согласно рисунку ниже, подгружаем сюда наш файл private.pem



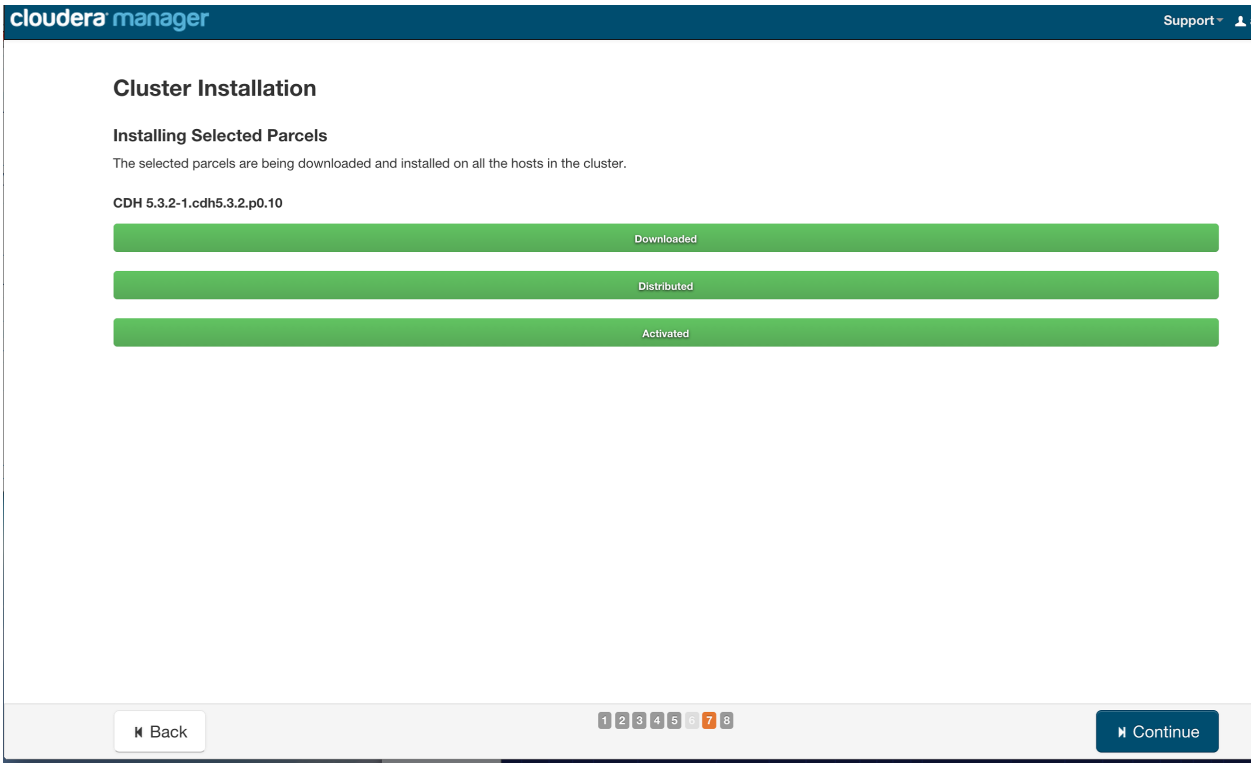
Да, это ок двигаться вперед без passphrase!



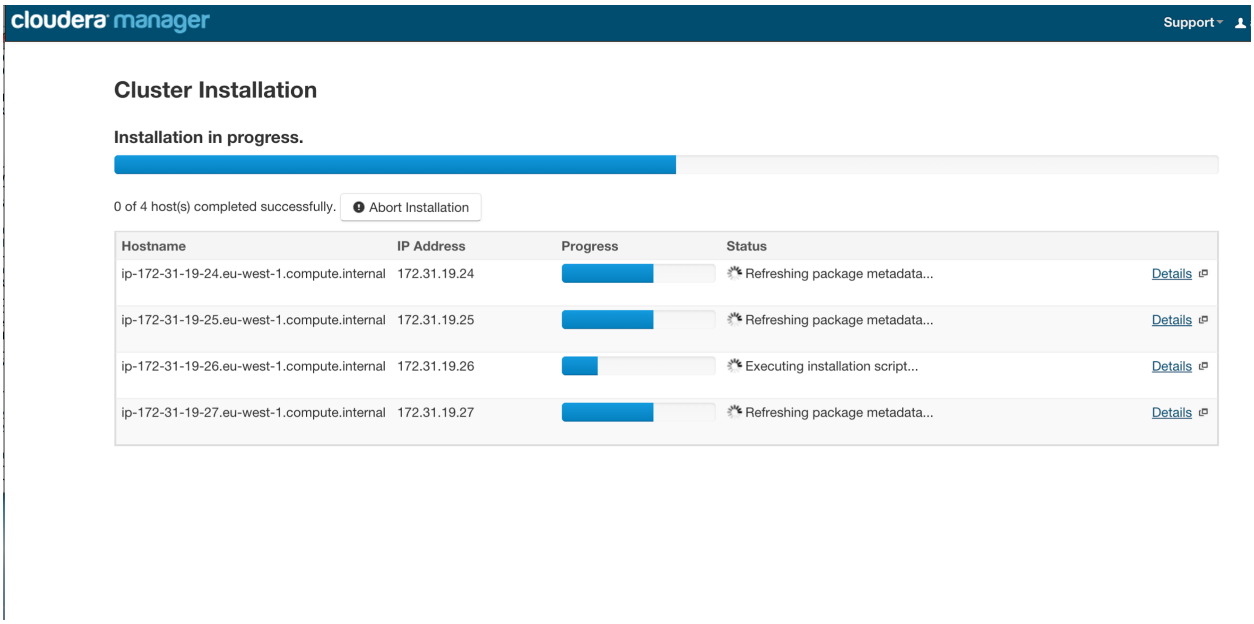
Ждем окончания инсталляции


cloudera manager
Support
Cluster Installation
Installing Selected Parcels
The selected parcels are being downloaded and installed on all the hosts in the cluster.
CDH 5.3.2-1.cdh5.3.2.p0.10
Downloading 4%
Distributing 0%
Activating 0%
Back
1 2 3 4 5 6 7 8
Continue

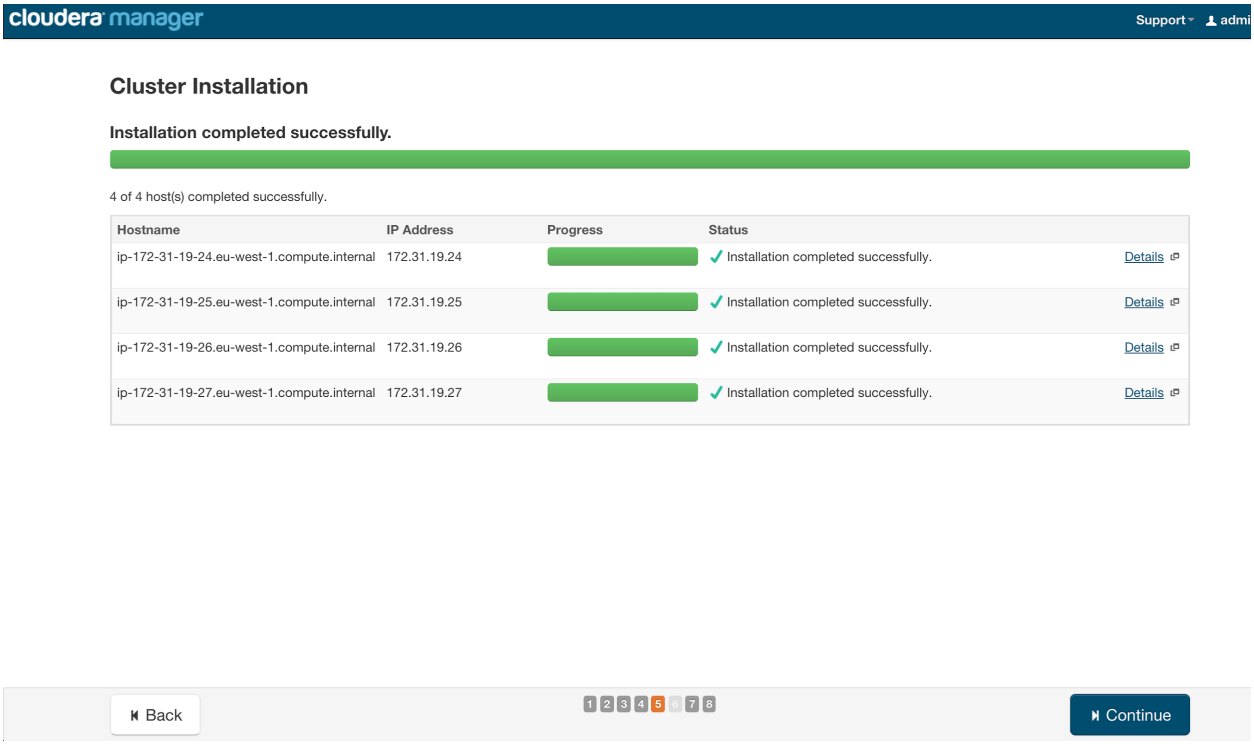
Победа!


cloudera manager
Support
Cluster Installation
Installing Selected Parcels
The selected parcels are being downloaded and installed on all the hosts in the cluster.
CDH 5.3.2-1.cdh5.3.2.p0.10
Downloaded
Distributed
Activated
Back
1 2 3 4 5 6 7 8
Continue

Cluster Installation - и снова ждем

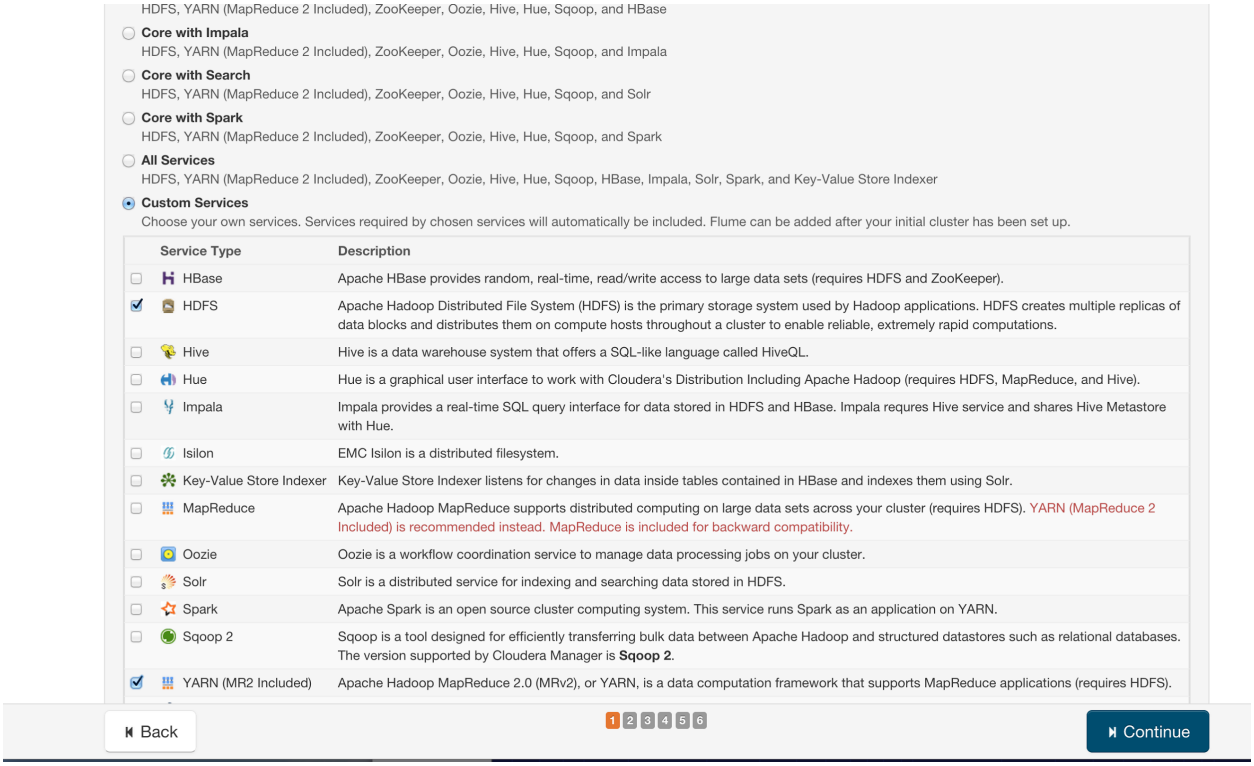


Ура! Next!

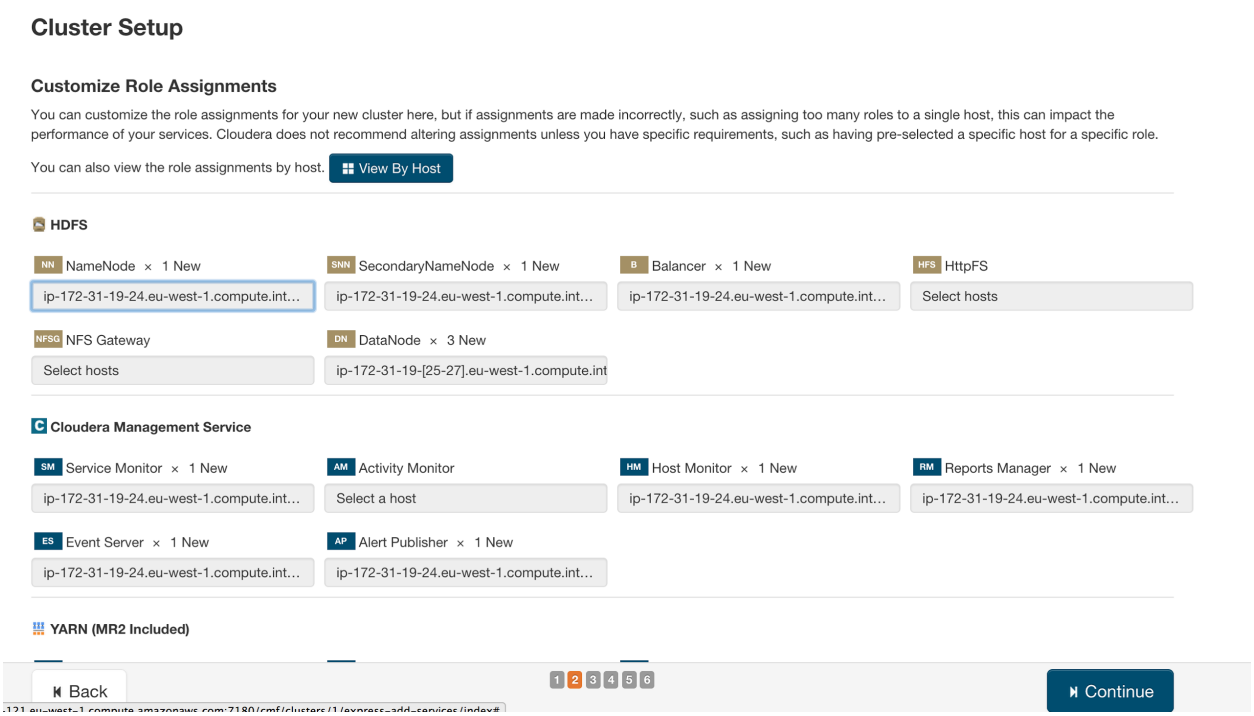


Finish

Cluster setup - выбираем Custom services HDFS и YARN

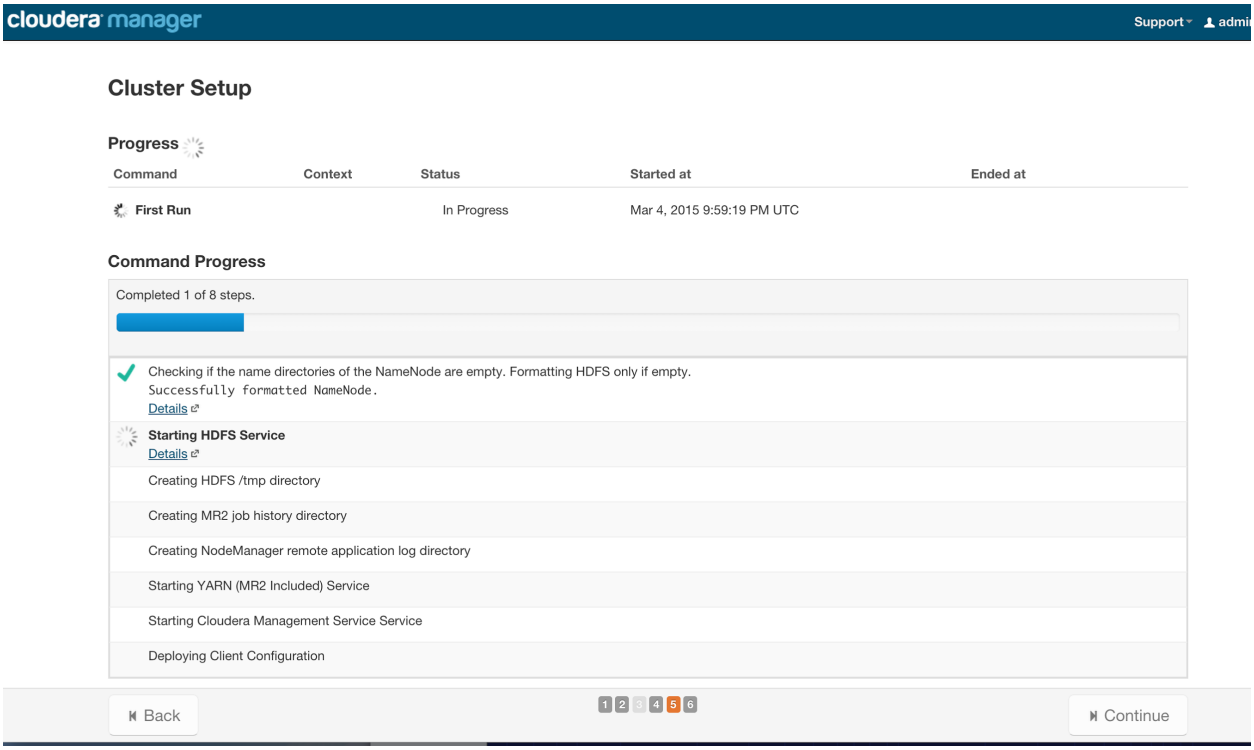


Continue (просто полезная информация)

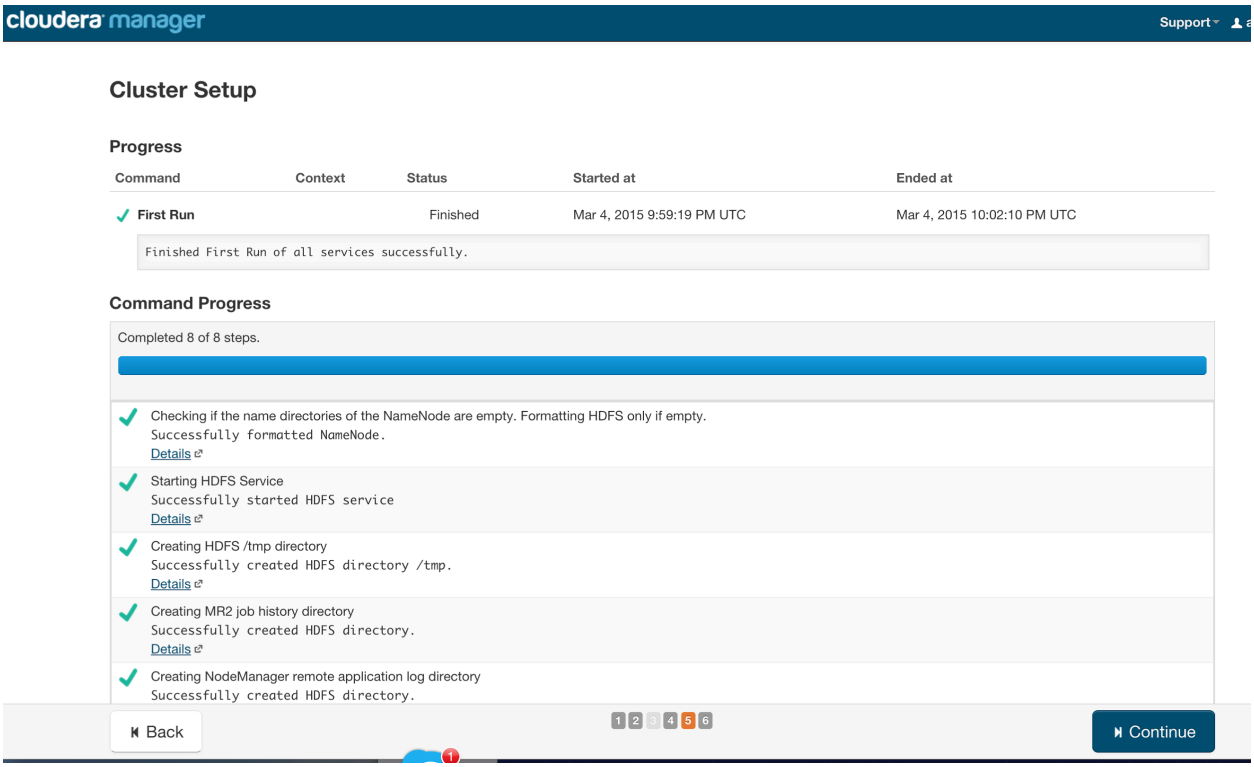


Review Changes - Continue

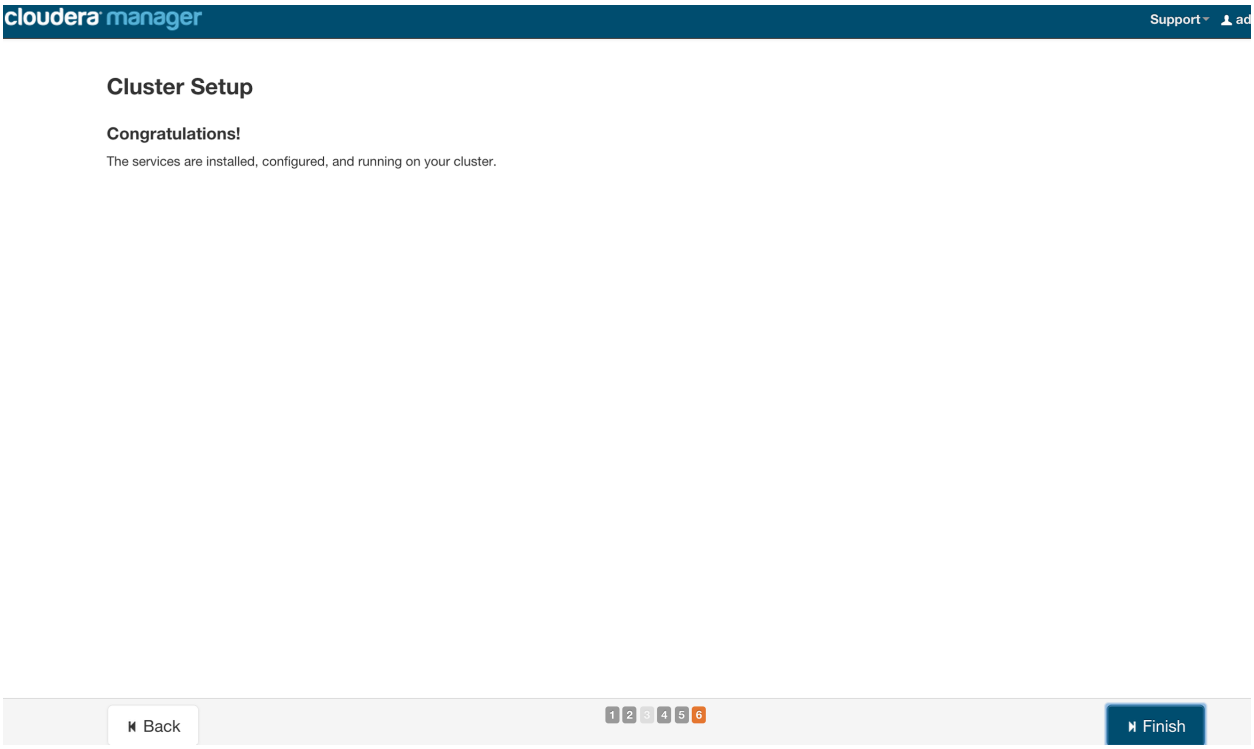
Progress - ждем зеленых галок везде



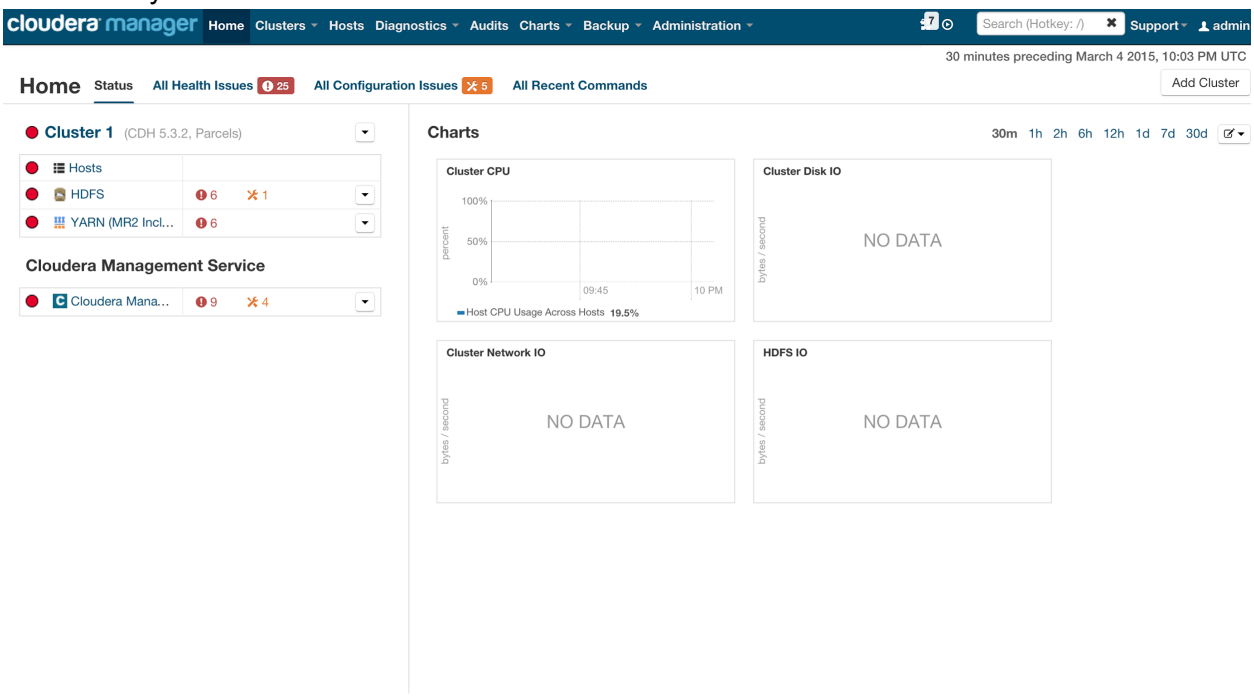
Зеленые галки! Continue


cloudera manager
Support
Cluster Setup
Progress
Command
Context
Status
Started at
Ended at
First Run
Finished
Mar 4, 2015 9:59:19 PM UTC
Mar 4, 2015 10:02:10 PM UTC
Finished First Run of all services successfully.
Command Progress
Completed 8 of 8 steps.
Checking if the name directories of the NameNode are empty. Formatting HDFS only if empty.
Successfully formatted NameNode.
Details
Starting HDFS Service
Successfully started HDFS service
Details
Creating HDFS /tmp directory
Successfully created HDFS directory /tmp.
Details
Creating MR2 job history directory
Successfully created HDFS directory.
Details
Creating NodeManager remote application log directory
Successfully created HDFS directory.
Back
1 2 3 4 5 6
Continue

Finish



Что-то получилось

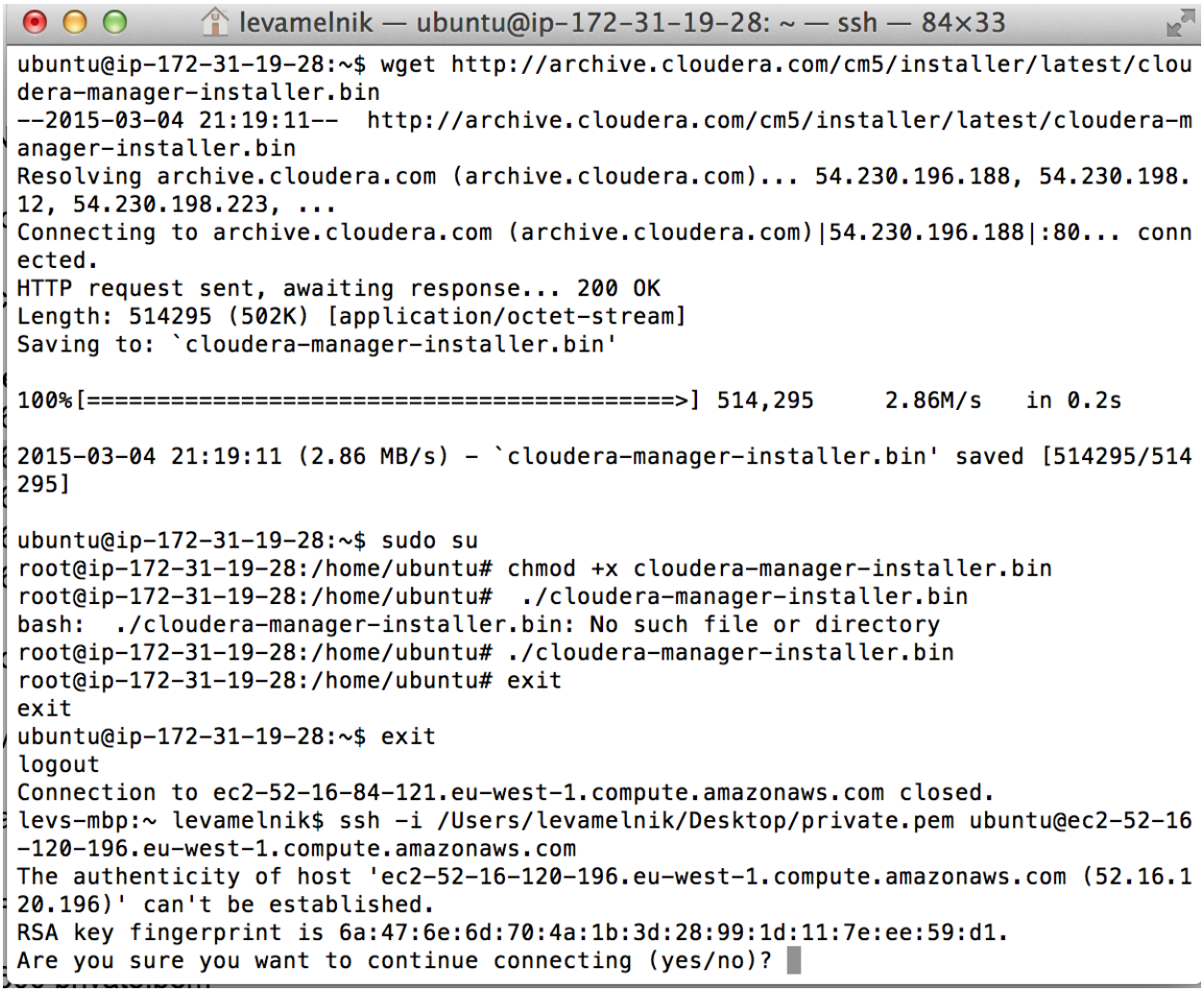


Входим в консоль

Hint! Если вы еще не вышли из предыдущего сервера, то вам понадобится один или несколько раз написать команду Exit

Подключаемся по SSH ко второму по порядку серверу

ssh -i /Users/levamelnik/Desktop/private.pem ubuntu@ec2-52-16-120-196.eu-west-1.compute.amazonaws.com



Yes

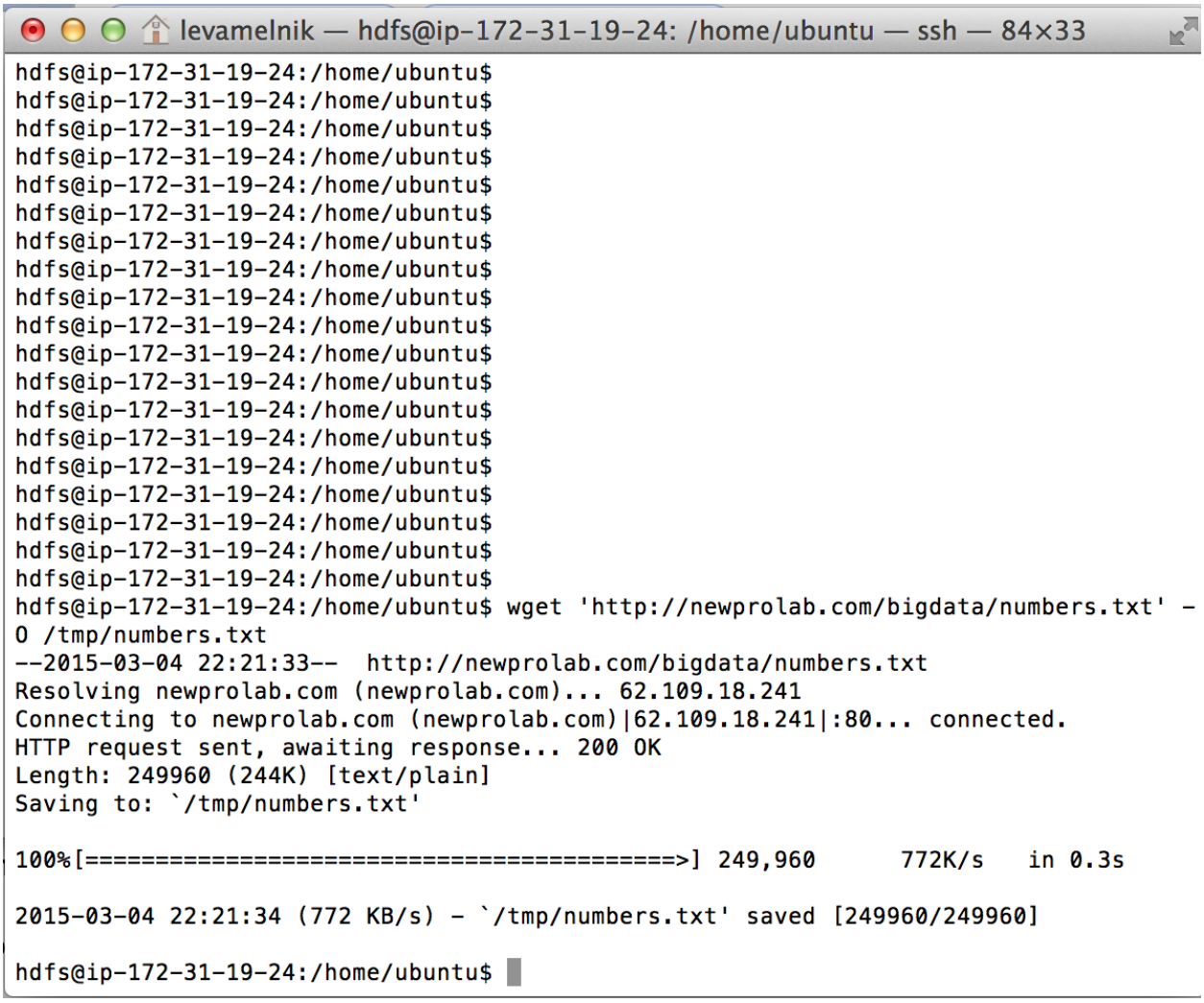
sudo su hdfs

Победа!!!!

Теперь попробуем выполнить простое задание.

Загружаем данные с сервера NPL

```
wget 'http://newprolab.com/bigdata/numbers.txt.lzma' -O /tmp/numbers.txt.lzma
lzma -d /tmp/numbers.txt.lzma
```



hadoop fs -mkdir -p /users/numbers

Потом

hadoop fs -put /tmp/numbers.txt /users/numbers/

Отдаем в работу

hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -input /users/numbers -output /users/numbers/result -mapper "cat" -reducer "uniq -c"

```
levamelnik — hdfs@ip-172-31-19-24: /home/ubuntu — ssh — 84×33
Length: 249960 (244K) [text/plain]
Saving to: `/tmp/numbers.txt'

100%[======================================>] 249,960      772K/s   in 0.3s

2015-03-04 22:21:34 (772 KB/s) - `/tmp/numbers.txt' saved [249960/249960]

hdfs@ip-172-31-19-24:/home/ubuntu$ hadoop fs -mkdir -p /users/numbers
hdfs@ip-172-31-19-24:/home/ubuntu$ hadoop fs -put /tmp/numbers.txt /users/numbers/
hdfs@ip-172-31-19-24:/home/ubuntu$
hdfs@ip-172-31-19-24:/home/ubuntu$
hdfs@ip-172-31-19-24:/home/ubuntu$
hdfs@ip-172-31-19-24:/home/ubuntu$
hdfs@ip-172-31-19-24:/home/ubuntu$
hdfs@ip-172-31-19-24:/home/ubuntu$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-m
apreduce/hadoop-streaming.jar -input /users/numbers -output /users/numbers/result -m
apper "cat" -reducer "uniq -c"
packageJobJar: [] [/opt/cloudera/parcels/CDH-5.3.2-1.cdh5.3.2.p0.10/jars/hadoop-stre
aming-2.5.0-cdh5.3.2.jar] /tmp/streamjob3553239476098183853.jar tmpDir=null
15/03/04 22:26:25 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-19
-24.eu-west-1.compute.internal/172.31.19.24:8032
15/03/04 22:26:26 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-19
-24.eu-west-1.compute.internal/172.31.19.24:8032
15/03/04 22:26:26 INFO mapred.FileInputFormat: Total input paths to process : 1
15/03/04 22:26:26 INFO mapreduce.JobSubmitter: number of splits:2
15/03/04 22:26:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_142550
6457029_0001
15/03/04 22:26:27 INFO impl.YarnClientImpl: Submitted application application_142550
6457029_0001
15/03/04 22:26:27 INFO mapreduce.Job: The url to track the job: http://ip-172-31-19-
24.eu-west-1.compute.internal:8088/proxy/application_1425506457029_0001/
15/03/04 22:26:27 INFO mapreduce.Job: Running job: job_1425506457029_0001
```

Смотрим результат

```
hadoop fs -cat /users/numbers/result/* | sort | tail
```

В результате должно получиться

```
49995 9
49996 7
49996 8
49997 5
49997 6
49998 3
49998 4
49999 1
49999 2
```

50000 0

Если вы совершили ошибку в последнем задании, то можете стереть данные из HDFS и начать все сначала при помощи команды

```
hadoop fs -rm -r -f /users/ && hadoop fs -mkdir -p /users/numbers
```

Проверьтре правильность выполнения задания на
http://bigdata.newprolab.com/amazon/