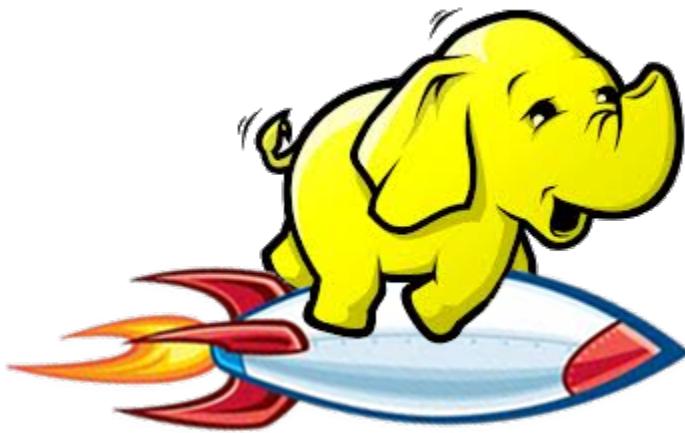


Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

# 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](mailto:@newprolab.ru)

Если у вас нет ящика на [@newprolab.ru](mailto:@newprolab.ru), отправьте запрос на [npl@digitaloctober.com](mailto: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”.

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

Кластер		
id	DNS	state
<button>⊕ Запустить</button>		<button>⟳ Обновить</button>

Получите список машин для работы!

Кластер		
id	DNS	state
i-827a1f64		pending
i-857a1f63		pending
i-867a1f60		pending
i-897a1f6f		pending
i-b97a1f5f		pending
<button>ⓧ Остановить</button>		<button>⟳ Обновить</button>

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

Кластер		
id	DNS	state
i-827a1f64	ec2-52-16-175-208.eu-west-1.compute.amazonaws.com	running
i-857a1f63	ec2-52-16-175-206.eu-west-1.compute.amazonaws.com	running
i-867a1f60	ec2-52-16-175-157.eu-west-1.compute.amazonaws.com	running
i-897a1f6f	ec2-52-16-175-204.eu-west-1.compute.amazonaws.com	running
i-b97a1f5f	ec2-52-16-175-203.eu-west-1.compute.amazonaws.com	running
<button>ⓧ Остановить</button>		<button>⟳ Обновить</button>

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

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

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

### Задание №1

Приватный ключ

```
-----BEGIN RSA PRIVATE KEY-----
MIIEPAIBAAQCAQEAE2euGuK8vPqSkmAjhDY3siYefg2S1CMtKVgRli8uow7DHy+JDqVS2U5dow70b
INKTWtx7DVt7i2a+Ka0d+1hZpI5ArYt40Ye0PYBB/gv8gmEfU5db3N08hiLfLmEqXqVLR5j9
etD1dgjT5VE0hkpzPtMfMsEc31U2vxNh4q4y+kTSpucqc2+LwQ5JP20Rpnh+RGwv2W7x4ts
TTWhf1w8MCRIfmU/U2uzd7UDhZ+/gpaC80q2V0Tp407Db7UFGVvxf/|ToQDbtMPVrJedG0@RxP
Qy8ueeX4rxj7ccnr8ebhetb7dcZMloK3hvjdzbz/Ur3KGIJA07m40IDQABoIAAoue/gbnx12
zVyud/YWijMale3sd17avFjozgUDSyxOKanTw10j8idjvvx3ULf16pY62Lw6gy/DC9F6naNYTEZ
0yaQ8B2rTvn01k4mWkpE8jbY4sDQLPqd1Y0V4IiuL4UoweKeMeoP4vj1H3t9agApdy6hu6aAvT0
4gb85zA03ba15kdfgiujuvabLBNNNNNv5fu+ysMmf0j5zSDPJEdj4CcpJnokhhYA8ZkGFkzqEkS
i4G/hZh/jgfCHUKz1Zkd/hauWFjtV1X1ix4cL49B6e4jWDB6xd4+Ta52936duPHLjTEn1Hssok8+
XiEp79RAKCodXL3Qo1leomIUCgYEA/3Tds8+puZlood2B8rENgkXeg7HB3sxahCatS@ewkj9
EVuGkwOfgw81pcu07XZ5iczaZ9yH0Rues10PnDX44Yd+DDraWjlsmzb0KRNr1NPNDInhPehvfb
KiqcU2/0h3Uqe0d4m8Lcz2FAF+Gy7o+zAGj0JKyEW+txoh0yh2sCgYEAE2MFVB9GWIaaQ0bb050e
hsv8tj01nkCcCL3HmY13yofu4KL12af6gYH61av7/6+0Kckyvui1FkvfzmhheGK2PBMmka47
cGKhP0b1j7XlHKWih0q4oqcvnPGrWtiryiw/Hjp032BBFqox7NGTPWP1G5twMxHE5BK4HHe0q4eMC
gYEARn4rhsrffhbcAvovDprBfoGhpNTk6fc0Tvr+H0cTwx4Nw/Fsk54a6KB1lnk5r1zxjvGZ3r/
Hpk42X6ehJDx-kjRHAvatir5aqbkD0D6IMQvPqBFTkEsQ91StTMRAF9v1lCMtNeiSwX9d4vrjmI
evF643Mii19LGuuFHVpJpHMCgYBeBneuk+aERqvCVR2Mc+MSZPGVaCJuAU94Z0LHckw2Zmgf1tB
31Gtrd1kkpp8JYCRYjDG43Vd/E11KTDrJ1c3ST69GRltRbjQKCGPDfkwLRC3TV1eqDcgWT6tC
bUnBoeVNoZemikmLLA90CR0TVrk0Hhgnv3/bkMo0op5rx64zAco
cL6gjEZErhqlrYsf/ZkyBz1RPXatipZ82o3kD2trcP24WRakw0fSGJUcmgBghsPtvph5h0cr3h
oznD0TMQkG9ypxFrylF+avZScqvuF3yxdNyHPKKXfx2DiyLovkyfiHrFEY9pe0Q9u73w==

-----END RSA PRIVATE KEY-----
```

### 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)

Программа “Специалист по большим данным”

Практическое задание: работа с Hadoop через Cloudera Manager

Кейс №1. Неделя 1.

```
levs-mbp:~ levamelnik$ ssh -i /Users/levamelnik/Desktop/private.pem ubuntu@ec2-52-16-84-121.eu-west-1.compute.amazonaws.com
Welcome to Ubuntu 12.04.5 LTS (GNU/Linux 3.2.0-75-virtual x86_64)

 * Documentation:  https://help.ubuntu.com/

 System information as of Wed Mar  4 21:18:36 UTC 2015

 System load:  0.0          Processes:      69
 Usage of /:   11.3% of 7.86GB  Users logged in:  0
 Memory usage: 1%           IP address for eth0: 172.31.19.28
 Swap usage:   0%

 Graph this data and manage this system at:
 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.nationalcablenetworks.ru
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ip-172-31-19-28:~$
```

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

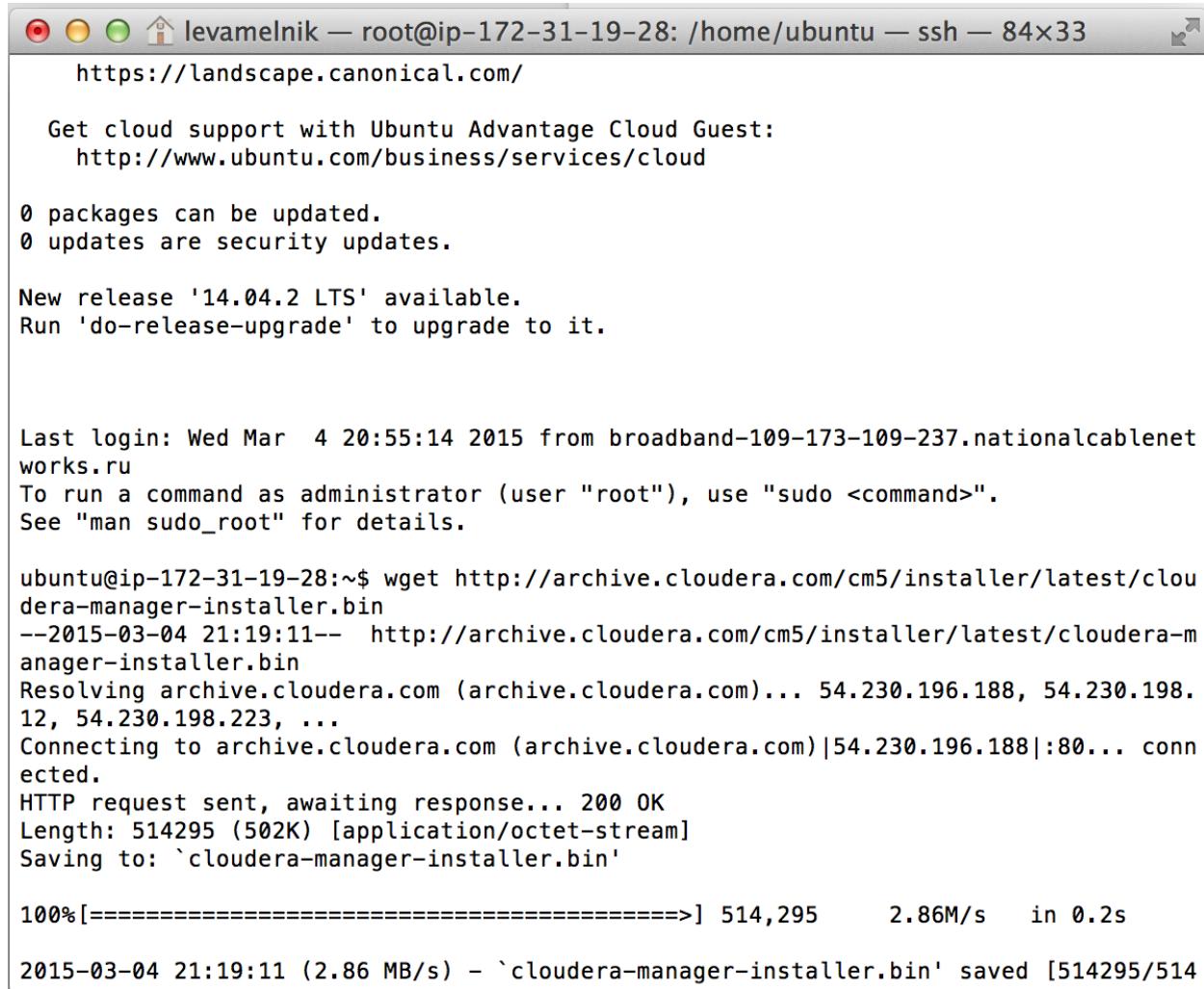
1. Загружаете приватный ключ с <http://bigdata.newprolab.com/>, как написано на странице 3.
2. Запускаете puttygen, Conversions -> Import key. Выбираете скачанный ключ. После этого - Save private key. Соглашаетесь сохранить без passphrase. Сохраняете в private.ppk.

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

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 — 84x33
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/cloudera-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
```

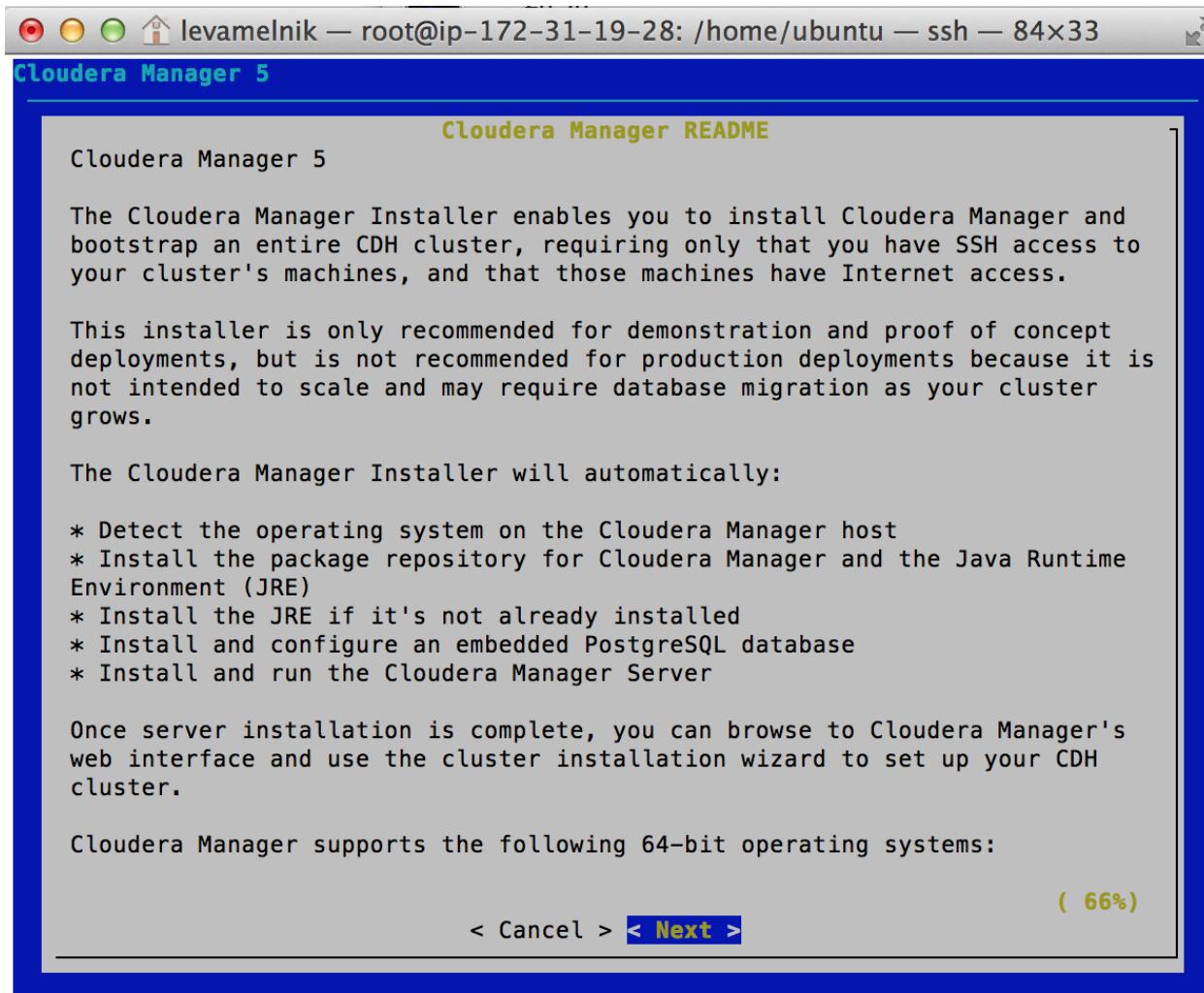
Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

### Делаем дистрибутив исполняемым

```
chmod +x cloudera-manager-installer.bin
```

### Запускаем дистрибутив

```
./cloudera-manager-installer.bin
```

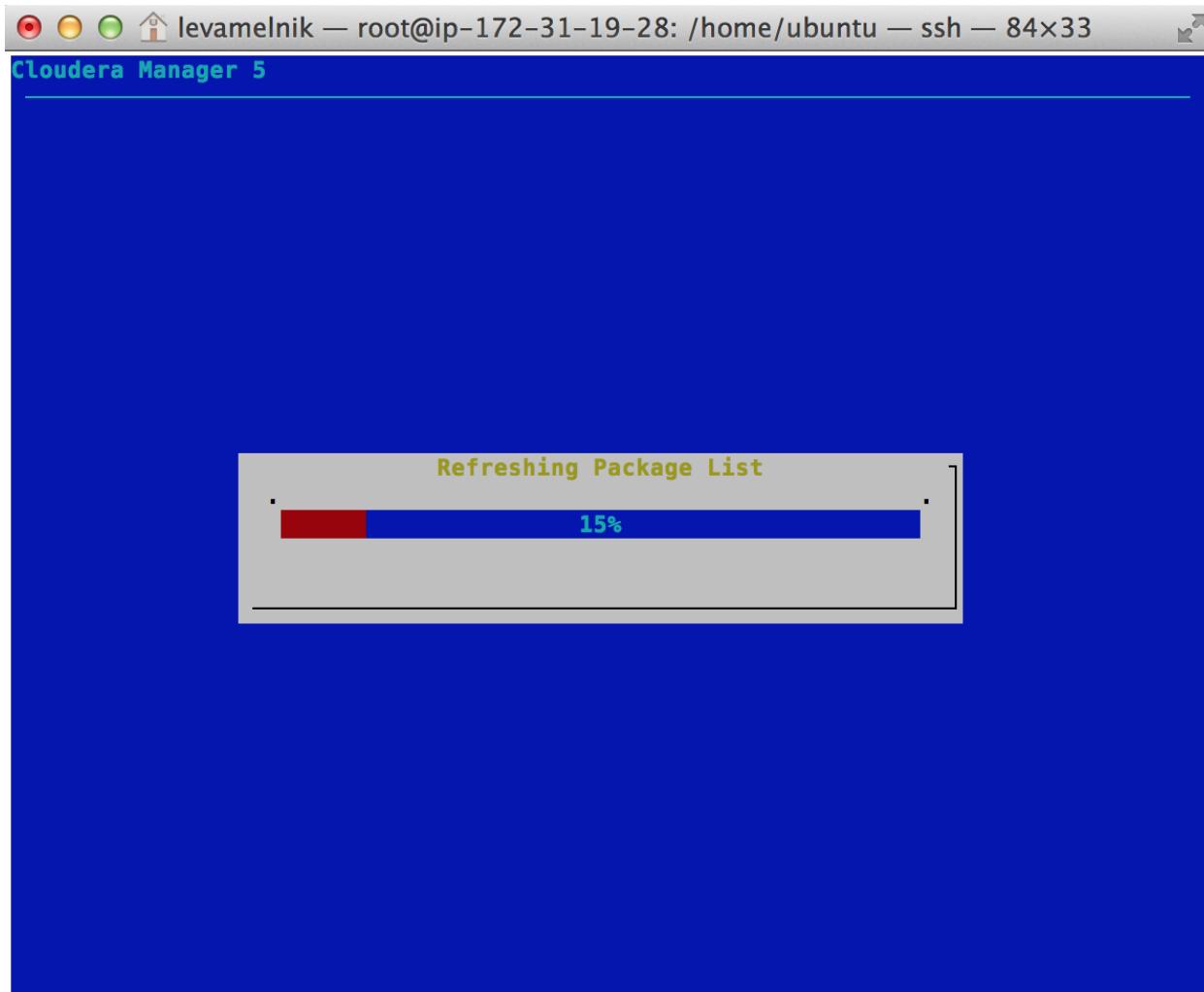


На все вопросы отвечаю утвердительно, и ждём окончания установки.

Программа “Специалист по большим данным”

Практическое задание: работа с Hadoop через Cloudera Manager

Кейс №1. Неделя 1.

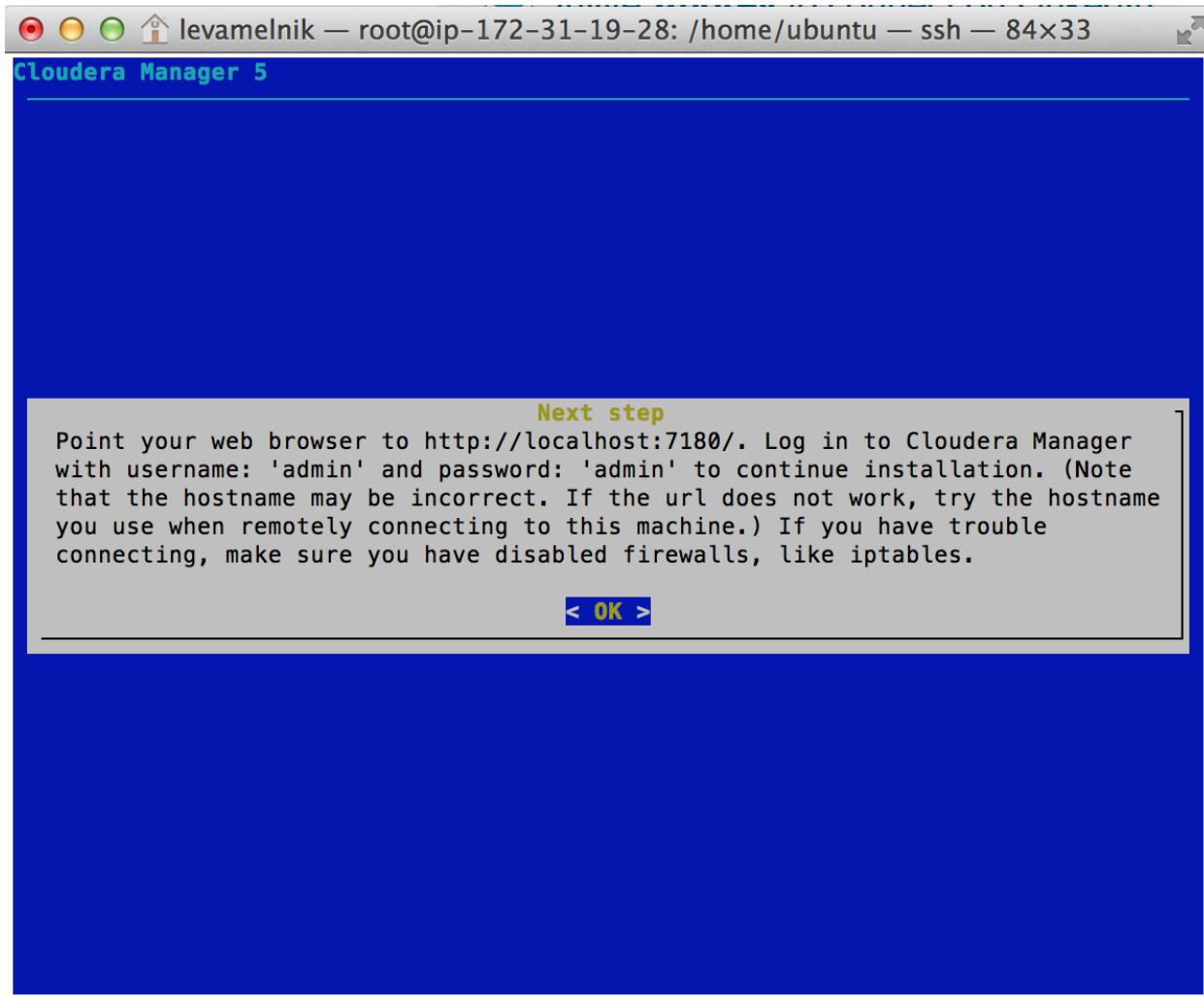


В финале будет сообщение вида "Зайдите на <domainname>:7180"

Программа “Специалист по большим данным”

Практическое задание: работа с Hadoop через Cloudera Manager

Кейс №1. Неделя 1.



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)**

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.



## Логин/ пароль

admin/admin

Continue



### Welcome to Cloudera Manager. Which edition do you want to deploy?

Upgrading to Cloudera Enterprise Data Hub Edition provides important features that help you manage and monitor your Hadoop clusters in mission-critical environments.

	Cloudera Express	Cloudera Enterprise Data Hub Edition Trial	Cloudera Enterprise
License	Free	60 Days  After the trial period, the product will continue to function as <b>Cloudera Express</b> . Your cluster and your data will remain unaffected.	Annual Subscription  <a href="#">Upload License</a>
Node Limit	Unlimited	Unlimited	Unlimited
CDH	✓	✓	✓
Core Cloudera Manager Features	✓	✓	✓
Advanced Cloudera Manager Features		✓	✓
Cloudera Navigator		✓	✓
Cloudera Support			✓

For full list of features available in Cloudera Express and Cloudera Enterprise, [click here](#).

Continue

Continue

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

## На следующем шаге необходимо вписать оставшиеся 4 домена

Specify hosts for your CDH cluster installation.

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.  
Cloudera recommends including Cloudera Manager Server's host. This will also enable health monitoring for that host.  
Hint: Search for hostnames and/or IP addresses using [patterns](#).

4 hosts scanned, 4 running SSH.

[New Search](#)

Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
<input checked="" type="checkbox"/> ec2-52-16-120-196.eu-west-1.compute.amazonaws.com	ip-172-31-19-24.eu-west-1.compute.internal	172.31.19.24	No	Host ready: 1 ms response time.
<input checked="" type="checkbox"/> ec2-52-16-121-54.eu-west-1.compute.amazonaws.com	ip-172-31-19-26.eu-west-1.compute.internal	172.31.19.26	No	Host ready: 1 ms response time.
<input checked="" type="checkbox"/> ec2-52-16-121-86.eu-west-1.compute.amazonaws.com	ip-172-31-19-27.eu-west-1.compute.internal	172.31.19.27	No	Host ready: 1 ms response time.
<input checked="" type="checkbox"/> ec2-52-16-124-102.eu-west-1.compute.amazonaws.com	ip-172-31-19-25.eu-west-1.compute.internal	172.31.19.25	No	Host ready: 1 ms response time.

[Back](#)

[Continue](#)

Continue

**Cluster Installation / Select Repository - тоже Continue**

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

The screenshot shows the 'Cluster Installation' process in Cloudera Manager. The current step is 'Select Repository'. A note at the top states: 'Cloudera recommends the use of parcels for installation over packages, because parcels enable Cloudera Manager to easily manage the software on your cluster, automating the deployment and upgrade of service binaries. Electing not to use parcels will require you to manually upgrade packages on all hosts in your cluster when software updates are available, and will prevent you from using Cloudera Manager's rolling upgrade capabilities.' Below this, under 'Choose Method', 'Use Parcels (Recommended)' is selected. Under 'Select the version of CDH', 'CDH-5.3.2-1.cdh5.3.2.p0.10' is selected. Under 'Additional Parcels', 'None' is selected for all options: ACCUMULO, KEYTRUSTEE, SQOOP\_NETEZZA\_CONNECTOR, and SQOOP\_TERADATA\_CONNECTOR.

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

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

JDK Installation Options

those features identified in Table 1-1 (Commercial Features In Java SE Product Editions) of the Java SE documentation accessible at <http://www.oracle.com/technetwork/java/javase/documentation/index.html>. “README File” means the README file for the Software accessible at <http://www.oracle.com/technetwork/java/javase/documentation/index.html>.

2. LICENSE TO USE. Subject to the terms and conditions of this Agreement including, but not limited to, the Java Technology Restrictions of the Supplemental License Terms, Oracle grants you a non-exclusive, non-transferable, limited license without license fees to reproduce and use internally the Software complete and unmodified for the sole purpose of running Programs. THE LICENSE SET FORTH IN THIS SECTION 2 DOES NOT EXTEND TO THE COMMERCIAL FEATURES. YOUR RIGHTS AND OBLIGATIONS RELATED TO THE COMMERCIAL FEATURES ARE AS SET FORTH IN THE SUPPLEMENTAL TERMS ALONG WITH ADDITIONAL LICENSES FOR DEVELOPERS AND PUBLISHERS.

3. RESTRICTIONS. Software is copyrighted. Title to Software and all associated intellectual property rights is retained by Oracle and/or its licensors. Unless enforcement is prohibited by applicable law, you may not modify, decompile, or reverse engineer Software. You acknowledge that the Software is developed for general use in a variety of information management applications; it is not developed or intended for use in any inherently dangerous applications, including applications that may create a risk of personal injury. If you use the Software in dangerous applications, then you shall be responsible to take all appropriate fail-safe, backup, redundancy, and other measures to ensure its safe use. Oracle disclaims any express or implied warranty of fitness for such uses. No right, title or interest in or to any trademark, service mark, logo or trade name of Oracle or its licensors is granted under this Agreement. Additional restrictions for developers and/or publishers licenses are set forth in the Supplemental License Terms.

4. DISCLAIMER OF WARRANTY. THE SOFTWARE IS PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND. ORACLE FURTHER DISCLAIMS ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT.

5. LIMITATION OF LIABILITY. IN NO EVENT SHALL ORACLE BE LIABLE FOR ANY INDIRECT, INCIDENTAL, SPECIAL, PUNITIVE OR CONSEQUENTIAL DAMAGES, OR DAMAGES FOR LOSS OF PROFITS, REVENUE, DATA OR DATA USE, INCURRED BY YOU OR ANY THIRD PARTY, WHETHER IN AN ACTION IN CONTRACT OR TORT, EVEN IF ORACLE HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. ORACLE’S ENTIRE LIABILITY FOR DAMAGES HEREUNDER SHALL IN NO EVENT EXCEED ONE THOUSAND DOLLARS (U.S. \$1,000).

6. TERMINATION. This Agreement is effective until terminated. You may terminate this Agreement at any time by destroying all copies of Software. This Agreement will terminate

Install Oracle Java SE Development Kit (JDK)  
Check this box to accept the Oracle Binary Code License Agreement and install the JDK. Leave it unchecked to use a currently installed JDK.

Install Java Unlimited Strength Encryption Policy Files  
Check this checkbox if local laws permit you to deploy unlimited strength encryption and you are running a secure cluster.

◀ Back

1 2 3 4 5 6 7 8

▶ Continue

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**

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

## Cluster Installation

### Provide SSH login credentials.

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login To All Hosts As:  root  Another user  
 ubuntu (with password-less sudo/pbrun to root)

You may connect via password or public-key authentication for the user selected above.

Authentication Method:  All hosts accept same password  All hosts accept same private key

Private Key File:  private.pem

Enter Passphrase:

Confirm Passphrase:

SSH Port: 22

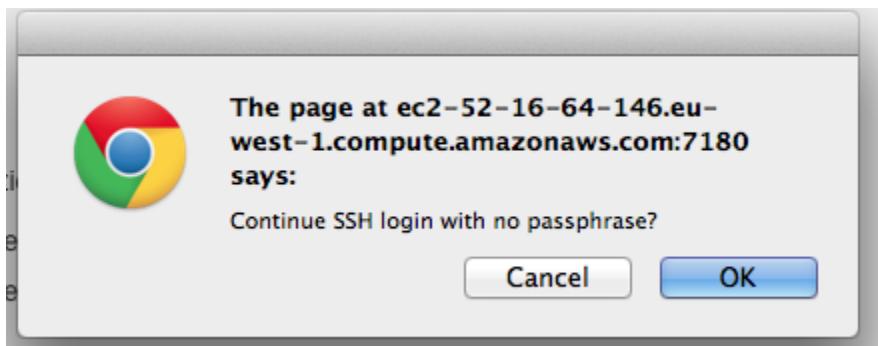
Number of Simultaneous Installations: 10 (Running a large number of installations at once can consume large amounts of network bandwidth and other system resources)

Back

1 2 3 4 5 6 7 8

Continue

Да, это ок двигаться вперед без passphrase!



Ждем окончания инсталляции

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

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

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

## Cluster Installation - и снова ждем

cloudera manager

Support

### Cluster Installation

Installation in progress.

0 of 4 host(s) completed successfully. [Abort Installation](#)

Hostname	IP Address	Progress	Status	<a href="#">Details</a>
ip-172-31-19-24.eu-west-1.compute.internal	172.31.19.24	<div style="width: 50%;"></div>	Refreshing package metadata...	<a href="#">Details</a>
ip-172-31-19-25.eu-west-1.compute.internal	172.31.19.25	<div style="width: 50%;"></div>	Refreshing package metadata...	<a href="#">Details</a>
ip-172-31-19-26.eu-west-1.compute.internal	172.31.19.26	<div style="width: 25%;"></div>	Executing installation script...	<a href="#">Details</a>
ip-172-31-19-27.eu-west-1.compute.internal	172.31.19.27	<div style="width: 50%;"></div>	Refreshing package metadata...	<a href="#">Details</a>

## Ура! Next!

cloudera manager

Support adm

### Cluster Installation

Installation completed successfully.

4 of 4 host(s) completed successfully.

Hostname	IP Address	Progress	Status	<a href="#">Details</a>
ip-172-31-19-24.eu-west-1.compute.internal	172.31.19.24	<div style="width: 100%;"></div>	✓ Installation completed successfully.	<a href="#">Details</a>
ip-172-31-19-25.eu-west-1.compute.internal	172.31.19.25	<div style="width: 100%;"></div>	✓ Installation completed successfully.	<a href="#">Details</a>
ip-172-31-19-26.eu-west-1.compute.internal	172.31.19.26	<div style="width: 100%;"></div>	✓ Installation completed successfully.	<a href="#">Details</a>
ip-172-31-19-27.eu-west-1.compute.internal	172.31.19.27	<div style="width: 100%;"></div>	✓ Installation completed successfully.	<a href="#">Details</a>

Back

1 2 3 4 5 6 7 8

Continue

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

## Finish

### Cluster setup - выбираем Custom services HDFS и YARN

HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and HBase

**Core with Impala**  
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and Impala

**Core with Search**  
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and Solr

**Core with Spark**  
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and Spark

**All Services**  
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, HBase, Impala, Solr, Spark, and Key-Value Store Indexer

**Custom Services**  
Choose your own services. Services required by chosen services will automatically be included. Flume can be added after your initial cluster has been set up.

Service Type	Description
<input type="checkbox"/> HBase	Apache HBase provides random, real-time, read/write access to large data sets (requires HDFS and ZooKeeper).
<input checked="" type="checkbox"/> HDFS	Apache Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute hosts throughout a cluster to enable reliable, extremely rapid computations.
<input type="checkbox"/> Hive	Hive is a data warehouse system that offers a SQL-like language called HiveQL.
<input type="checkbox"/> Hue	Hue is a graphical user interface to work with Cloudera's Distribution Including Apache Hadoop (requires HDFS, MapReduce, and Hive).
<input type="checkbox"/> Impala	Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires Hive service and shares Hive Metastore with Hive.
<input type="checkbox"/> Isilon	EMC Isilon is a distributed filesystem.
<input type="checkbox"/> Key-Value Store Indexer	Key-Value Store Indexer listens for changes in data inside tables contained in HBase and indexes them using Solr.
<input type="checkbox"/> MapReduce	Apache Hadoop MapReduce supports distributed computing on large data sets across your cluster (requires HDFS). <b>YARN (MapReduce 2 Included) is recommended instead. MapReduce is included for backward compatibility.</b>
<input type="checkbox"/> Oozie	Oozie is a workflow coordination service to manage data processing jobs on your cluster.
<input type="checkbox"/> Solr	Solr is a distributed service for indexing and searching data stored in HDFS.
<input type="checkbox"/> Spark	Apache Spark is an open source cluster computing system. This service runs Spark as an application on YARN.
<input type="checkbox"/> Sqoop 2	Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. The version supported by Cloudera Manager is <b>Sqoop 2</b> .
<input checked="" type="checkbox"/> YARN (MR2 Included)	Apache Hadoop MapReduce 2.0 (MRv2), or YARN, is a data computation framework that supports MapReduce applications (requires HDFS).

**Back** 1 2 3 4 5 6 **Continue**

**Continue (просто полезная информация)**

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

## Cluster Setup

### Customize Role Assignments

You can customize the role assignments for your new cluster here, but if assignments are made incorrectly, such as assigning too many roles to a single host, this can impact the performance of your services. Cloudera does not recommend altering assignments unless you have specific requirements, such as having pre-selected a specific host for a specific role.

You can also view the role assignments by host. [View By Host](#)

#### HDFS

NameNode x 1 New	SecondaryNameNode x 1 New	Balancer x 1 New	HttpFS
ip-172-31-19-24.eu-west-1.compute.int...	ip-172-31-19-24.eu-west-1.compute.int...	ip-172-31-19-24.eu-west-1.compute.int...	Select hosts

#### NFS Gateway

DataNode x 3 New
Select hosts

ip-172-31-19-[25-27].eu-west-1.compute.int...

#### Cloudera Management Service

Service Monitor x 1 New	Activity Monitor	Host Monitor x 1 New	Reports Manager x 1 New
ip-172-31-19-24.eu-west-1.compute.int...	Select a host	ip-172-31-19-24.eu-west-1.compute.int...	ip-172-31-19-24.eu-west-1.compute.int...
Event Server x 1 New	Alert Publisher x 1 New		
ip-172-31-19-24.eu-west-1.compute.int...	ip-172-31-19-24.eu-west-1.compute.int...		

#### YARN (MR2 Included)

Back

1 2 3 4 5 6

Continue

## Review Changes - Continue

### Progress - ждем зеленых галок везде

cloudera manager

Support ▾ admin

#### Cluster Setup

##### Progress

Command	Context	Status	Started at	Ended at
First Run		In Progress	Mar 4, 2015 9:59:19 PM UTC	

##### Command Progress

Completed 1 of 8 steps.

Checking if the name directories of the NameNode are empty. Formatting HDFS only if empty. Successfully formatted NameNode. <a href="#">Details ↗</a>
Starting HDFS Service <a href="#">Details ↗</a>
Creating HDFS /tmp directory
Creating MR2 job history directory
Creating NodeManager remote application log directory
Starting YARN (MR2 Included) Service
Starting Cloudera Management Service Service
Deploying Client Configuration

Back

1 2 3 4 5 6

Continue

## 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. <a href="#">Details ↗</a>
✓ Starting HDFS Service Successfully started HDFS service <a href="#">Details ↗</a>
✓ Creating HDFS /tmp directory Successfully created HDFS directory /tmp. <a href="#">Details ↗</a>
✓ Creating MR2 job history directory Successfully created HDFS directory. <a href="#">Details ↗</a>
✓ Creating NodeManager remote application log directory Successfully created HDFS directory.

 Back   Continue

## Finish

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

The screenshot shows the 'Cluster Setup' page of the Cloudera Manager interface. At the top, a message says 'Congratulations!' followed by 'The services are installed, configured, and running on your cluster.' Below this is a progress bar with steps 1 through 6, where step 6 is highlighted in orange. At the bottom are 'Back' and 'Finish' buttons.

### ЧТО-ТО ПОЛУЧИЛОСЬ

The screenshot shows the 'Home' page of the Cloudera Manager interface. It displays the status of 'Cluster 1' (CDH 5.3.2, Parcels) with sections for Hosts, HDFS, and YARN. Below this is the 'Cloudera Management Service' section. To the right, there are four charts: 'Cluster CPU' (Host CPU Usage Across Hosts: 19.5%), 'Cluster Disk IO' (NO DATA), 'Cluster Network IO' (NO DATA), and 'HDFS IO' (NO DATA). The top navigation bar includes 'Home', 'Clusters', 'Hosts', 'Diagnostics', 'Audits', 'Charts', 'Backup', 'Administration', and a search bar.

**ВХОДИМ В КОНСОЛЬ**

Программа “Специалист по большим данным”

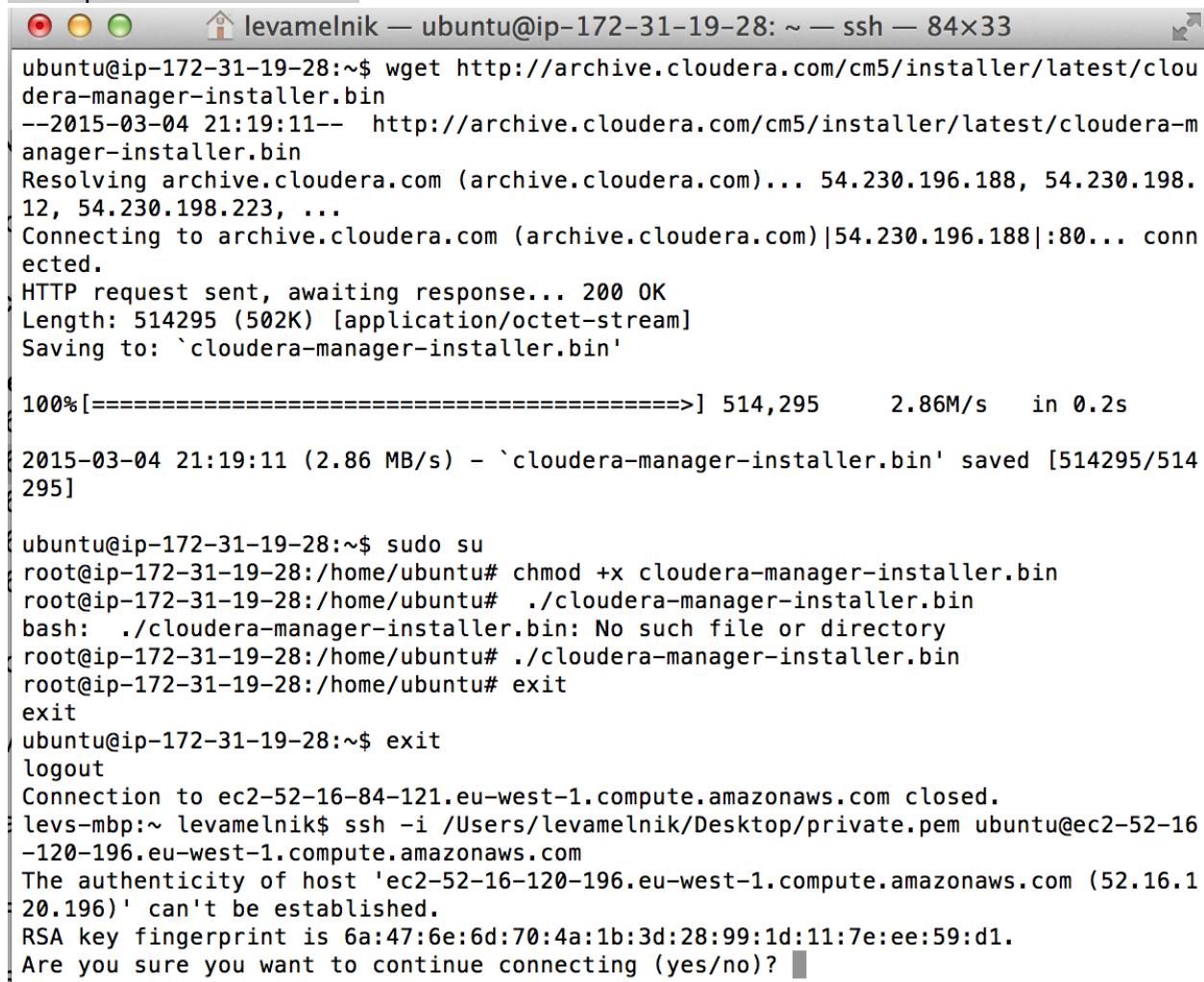
Практическое задание: работа с Hadoop через Cloudera Manager

Кейс №1. Неделя 1.

*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
```



The screenshot shows a terminal window titled "levamelnik — ubuntu@ip-172-31-19-28: ~ — ssh — 84x33". The window displays the following command and its output:

```
ubuntu@ip-172-31-19-28:~$ wget http://archive.cloudera.com/cm5/installer/latest/cloudera-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
295]

ubuntu@ip-172-31-19-28:~$ sudo su
root@ip-172-31-19-28:/home/ubuntu# chmod +x cloudera-manager-installer.bin
root@ip-172-31-19-28:/home/ubuntu# ./cloudera-manager-installer.bin
bash: ./cloudera-manager-installer.bin: No such file or directory
root@ip-172-31-19-28:/home/ubuntu# ./cloudera-manager-installer.bin
root@ip-172-31-19-28:/home/ubuntu# exit
exit
ubuntu@ip-172-31-19-28:~$ exit
logout
Connection to ec2-52-16-84-121.eu-west-1.compute.amazonaws.com closed.
levs-mbp:~ levamelnik$ ssh -i /Users/levamelnik/Desktop/private.pem ubuntu@ec2-52-16-120-196.eu-west-1.compute.amazonaws.com
The authenticity of host 'ec2-52-16-120-196.eu-west-1.compute.amazonaws.com (52.16.1
20.196)' can't be established.
RSA key fingerprint is 6a:47:6e:6d:70:4a:1b:3d:28:99:1d:11:7e:ee:59:d1.
Are you sure you want to continue connecting (yes/no)?
```

Yes

```
sudo su hdfs
```

Программа “Специалист по большим данным”

Практическое задание: работа с Hadoop через Cloudera Manager

Кейс №1. Неделя 1.

```
levs-mbp:~ levamelnik$ ssh -i /Users/levamelnik/Desktop/private.pem ubuntu@ec2-52-16-120-196.eu-west-1.compute.amazonaws.com
The authenticity of host 'ec2-52-16-120-196.eu-west-1.compute.amazonaws.com (52.16.1.20.196)' can't be established.
RSA key fingerprint is 6a:47:6e:6d:70:4a:1b:3d:28:99:1d:11:7e:ee:59:d1.
Are you sure you want to continue connecting (yes/no)? Yes
Warning: Permanently added 'ec2-52-16-120-196.eu-west-1.compute.amazonaws.com,52.16.120.196' (RSA) to the list of known hosts.
Welcome to Ubuntu 12.04.5 LTS (GNU/Linux 3.2.0-75-virtual x86_64)

 * Documentation:  https://help.ubuntu.com/

 System information as of Wed Mar  4 22:13:32 UTC 2015

 System load:  0.14          Processes:           84
 Usage of /:   81.7% of 7.86GB  Users logged in:     0
 Memory usage: 33%          IP address for eth0: 172.31.19.24
 Swap usage:   0%

 Graph this data and manage this system at:
 https://landscape.canonical.com/

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

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

Last login: Wed Mar  4 21:39:35 2015 from ip-172-31-19-28.eu-west-1.compute.internal
ubuntu@ip-172-31-19-24:~$ sudo su hdfs
hdfs@ip-172-31-19-24:/home/ubuntu$
```

**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
```

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

```
levamelnik — hdfs@ip-172-31-19-24: /home/ubuntu — ssh — 84x33
hdfs@ip-172-31-19-24: /home/ubuntu$ 
hdfs@ip-172-31-19-24: /home/ubuntu$ wget 'http://newprolab.com/bigdata/numbers.txt' -O /tmp/numbers.txt
--2015-03-04 22:21:33-- http://newprolab.com/bigdata/numbers.txt
Resolving newprolab.com (newprolab.com)... 62.109.18.241
Connecting to newprolab.com (newprolab.com)|62.109.18.241|:80... connected.
HTTP request sent, awaiting response... 200 OK
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
```

ПОТОМ

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

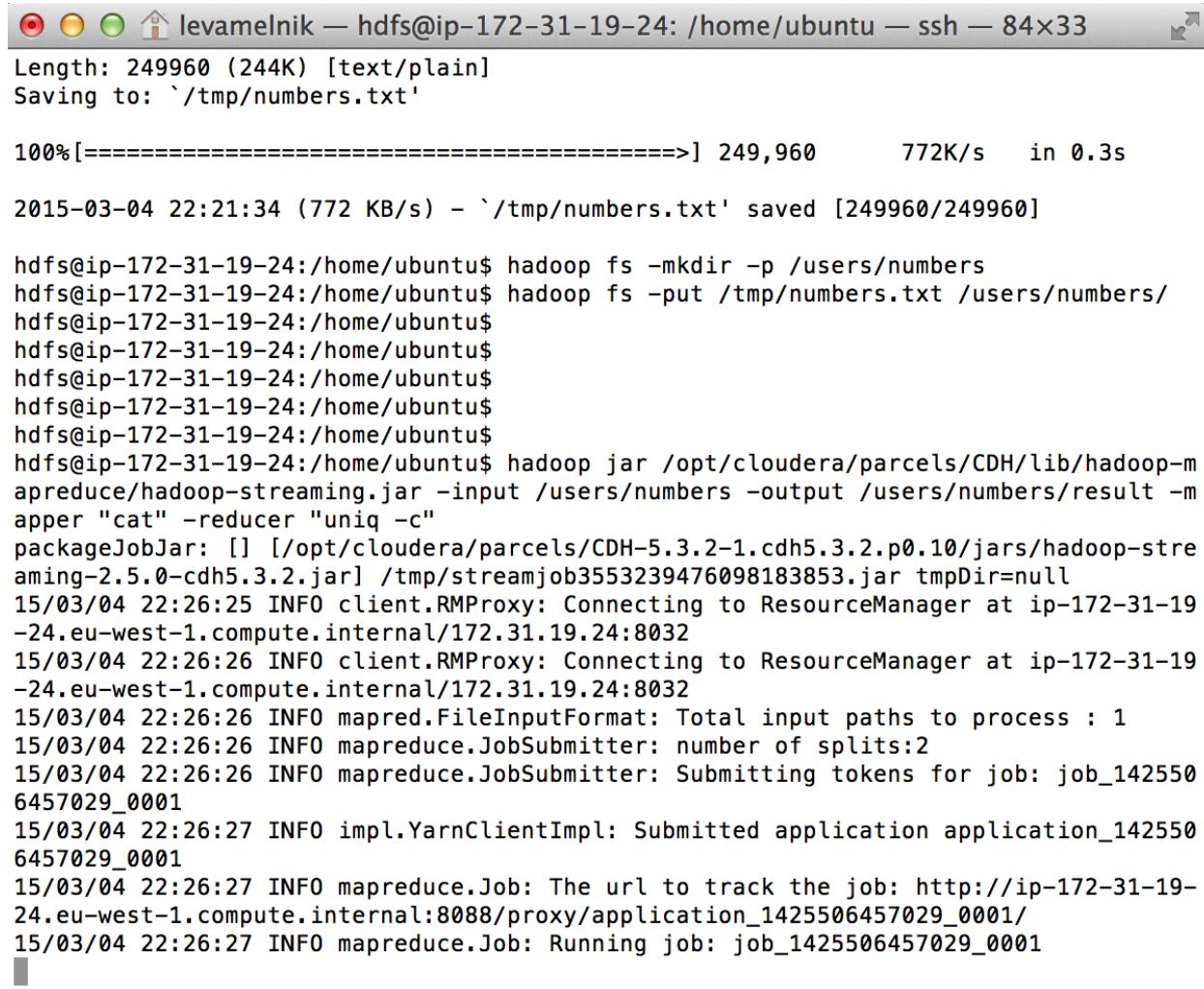
### Отдаем в работу

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

Программа “Специалист по большим данным”

Практическое задание: работа с Hadoop через Cloudera Manager

Кейс №1. Неделя 1.



levamelnik — hdfs@ip-172-31-19-24: /home/ubuntu — ssh — 84x33

```
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-mapreduce/hadoop-streaming.jar -input /users/numbers -output /users/numbers/result -mapper "cat" -reducer "uniq -c"
packageJobJar: [] [/opt/cloudera/parcels/CDH-5.3.2-1.cdh5.3.2.p0.10/jars/hadoop-streaming-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_1425506457029_0001
15/03/04 22:26:27 INFO impl.YarnClientImpl: Submitted application application_1425506457029_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
```

Программа “Специалист по большим данным”  
Практическое задание: работа с Hadoop через Cloudera Manager  
Кейс №1. Неделя 1.

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
```

Проверьте правильность выполнения задания на  
<http://bigdata.newprolab.com/amazon/>