# Arenadata™ Hadoop

*Версия - v1.5.2*

## Инструкция по установке кластера

*2018 Arenadata™*

# Оглавление


**1 Загрузка дистрибутива кластера** **3**

**2 Подготовка к установке кластера** **8**
2.1 Системные требования . . . 8
2.2 Сбор информации . . . 12
2.3 Подготовка к установке Hadoop . . . 12

**3 Установка сервера Ambari 2.6.1** **17**
3.1 Настройка репозитория . . . 17
3.2 Установка сервера Ambari . . . 19
3.3 Настройка сервера Ambari . . . 20
3.4 Запуск сервера Ambari . . . 20

**4 Подготовка к установке основных компонентов ADH на кластер** **22**
4.1 Загрузка и извлечение архивов стека ADH . . . 22
4.2 Настройка локальных репозиториев . . . 23

**5 Запуск мастера установки** **24**
5.1 Изменение URL-адресов репозиториев . . . 24
5.2 Ввод имен узлов и SSH-ключа . . . 24
5.3 Выбор компонентов . . . 27
5.4 Назначение мастер-узлов . . . 27
5.5 Назначение Slave и Client узлов компонентов кластера . . . 28
5.6 Дополнительные настройки компонентов . . . 28
5.7 Установка, запуск и тестирование . . . 30

**6 Панель кластера** **33**

**7 Приложение 1. Ручная настройка подключения к базе данных** **34**

В инструкции приведены сведения необходимые для подготовки к установке кластера Arenadata Hadoop, к его настройке и запуску.

Инструкция может быть полезна администраторам, программистам, разработчикам и сотрудникам подразделений информационных технологий, осуществляющих внедрение кластера.

**Important:** Контактная информация службы поддержки – e-mail: info@arenadata.io

# Глава 1

# Загрузка дистрибутива кластера

Для загрузки дистрибутива кластера **Arenadata Hadoop** необходимо на сайте arenadata.io выбрать вкладку *“Скачать”*. При этом происходит переход на страницу *Магазин Программного Обеспечения Arenadata* (store.arenadata.io, Рис.1.3.), и открывается экранная форма заявки для загрузки продукта (Рис.1.1.).

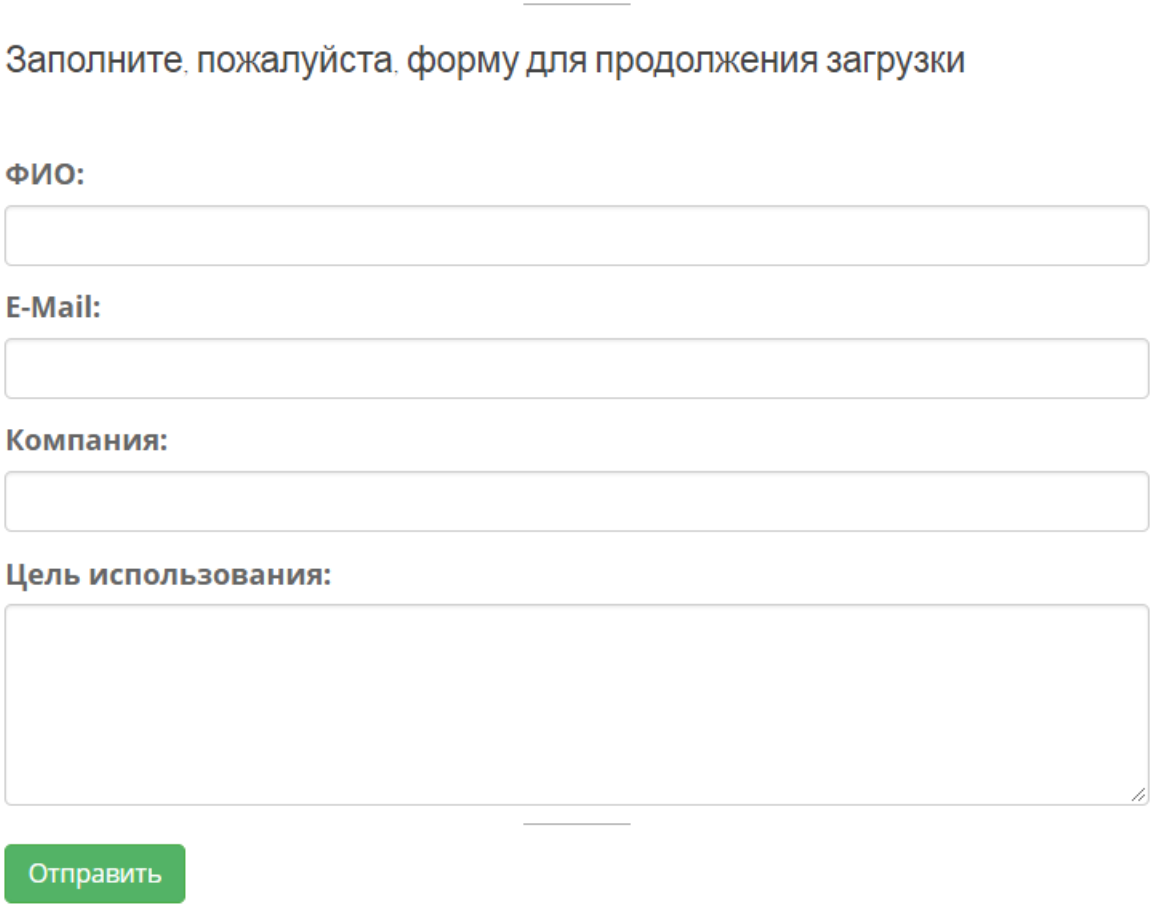



Рис.1.1.: Заявка для загрузки продукта

В экранной форме заявки следует заполнить следующие поля:

- *ФИО* – ФИО пользователя;

- *E-Mail* – адрес электронной почты пользователя;
- *Компания* – наименование компании пользователя;
- *Цель использования* – описание цели использования кластера.

Все поля являются обязательными для заполнения. После ввода данных необходимо нажать кнопку *Отправить*. При корректном заполнении полей выдается соответствующее сообщение (Рис.1.2.).




Рис.1.2.: Заявка принята

Окно сообщения следует закрыть. При этом происходит переход на страницу *Магазин Приложений* с перечнем продуктов **Arenadata** (Рис.1.3.).

При наведении курсора на продукт **ARENADATA HADOOP** предоставляется выбор действий:

- *Скачать* – переход на страницу загрузки дистрибутива кластера **ADH**;
- *Документация* – переход на страницу он-лайн документации кластера **ADH**;
- *Сайзинг оборудования* – переход на страницу расчета необходимого количества серверов для хранения данных в кластере **ADH**.

Для загрузки дистрибутива кластера следует выбрать пункт *Скачать*, при этом происходит переход на соответствующую страницу с кратким описанием продукта, ссылками на компоненты дистрибутива и информацией о релизе (Рис.1.4.).

В разделе *"Выберите версию продукта"* следует указать интересующую версию кластера **ADH**, а в поле *"Выберите ОС"* – используемую операционную систему. При этом в разделе *"Информация о релизе"* отображается дата выхода и тип релиза выбранной версии, ее номер, ссылки на лицензию использования и Release Notes.

В разделе *Ссылки на компоненты дистрибутива* находятся ссылки для загрузки компонентов выбранной версии кластера. После загрузки компонентов следует установить их, действуя в соответствии с последующими разделами настоящего документа.

На данной странице Магазина Приложений при открытии вкладки *"Подробнее о продукте"* появляется ссылка *"Сайзинг оборудования"* – ссылка на страницу калькулятора для расчета необходимого количества серверов заданной конфигурации для хранения и обработки указанного объема пользовательских данных в **HDFS** (Рис.1.5.).

На странице калькулятора есть возможность задать персональные параметры в соответствующих полях для расчета необходимого количества серверов заданной конфигурации для хранения и обработки указанного объема пользовательских данных в **HDFS**.

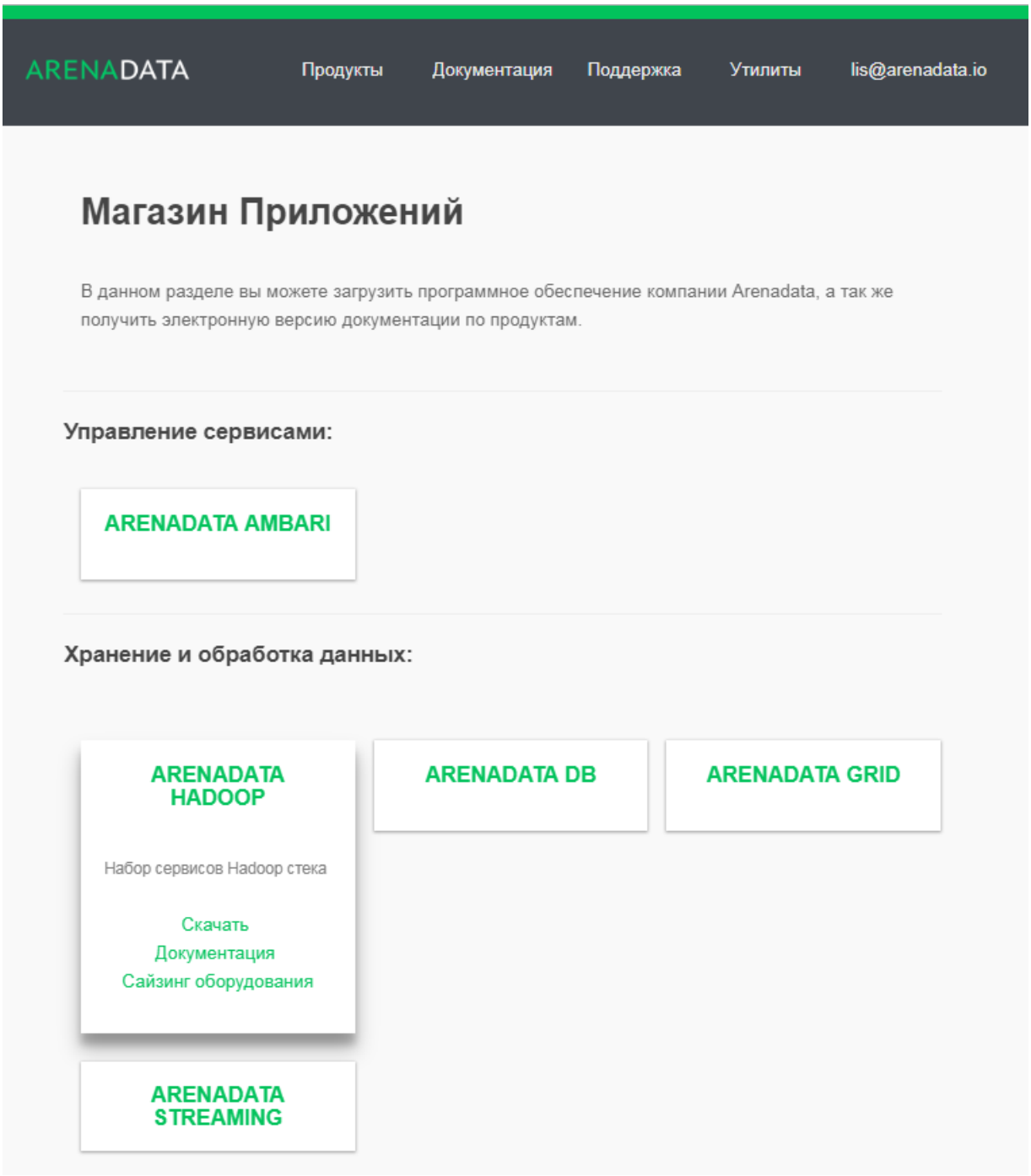



Рис.1.3.: Магазин Приложений

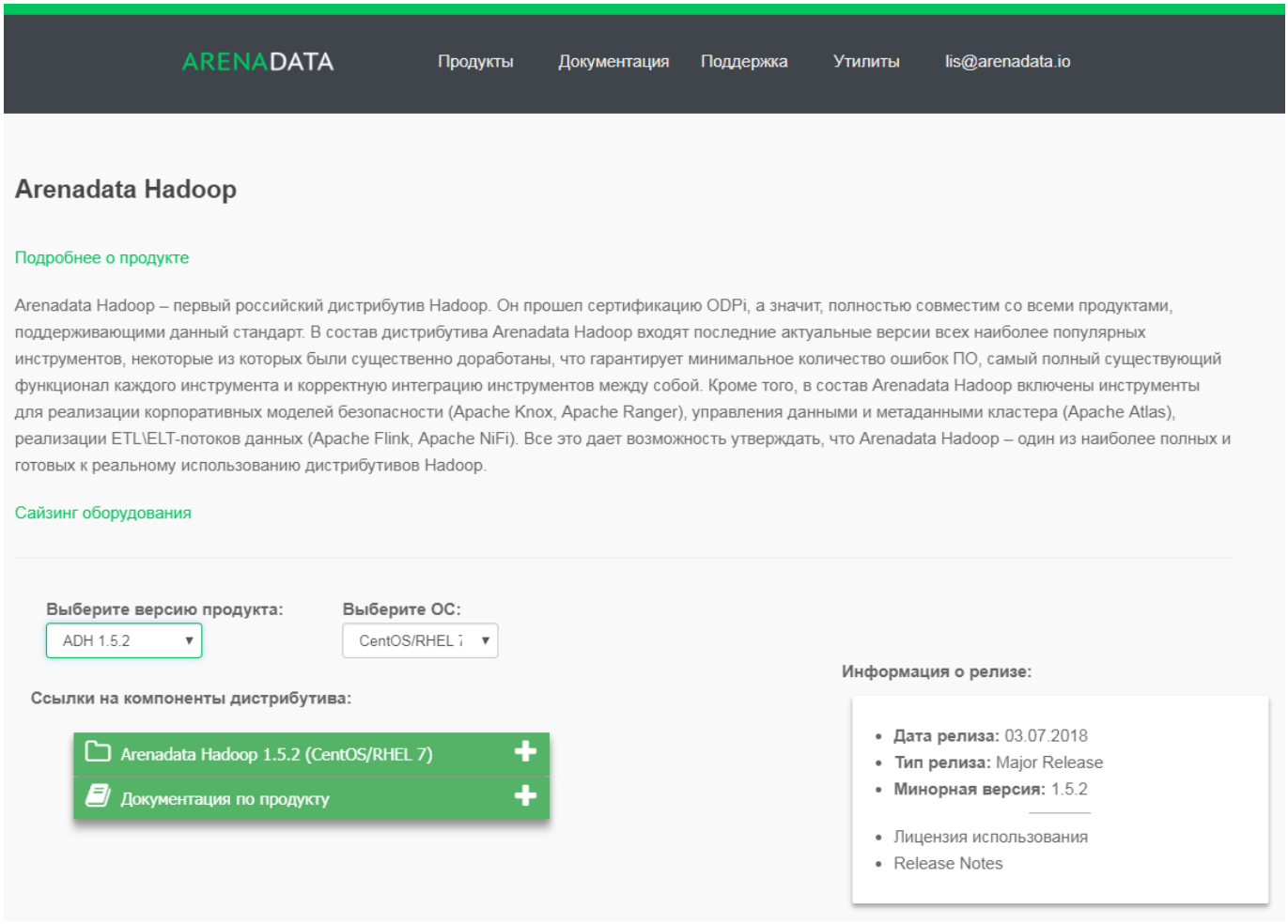



Рис.1.4.: Страница загрузки кластера Arenadata Hadoop

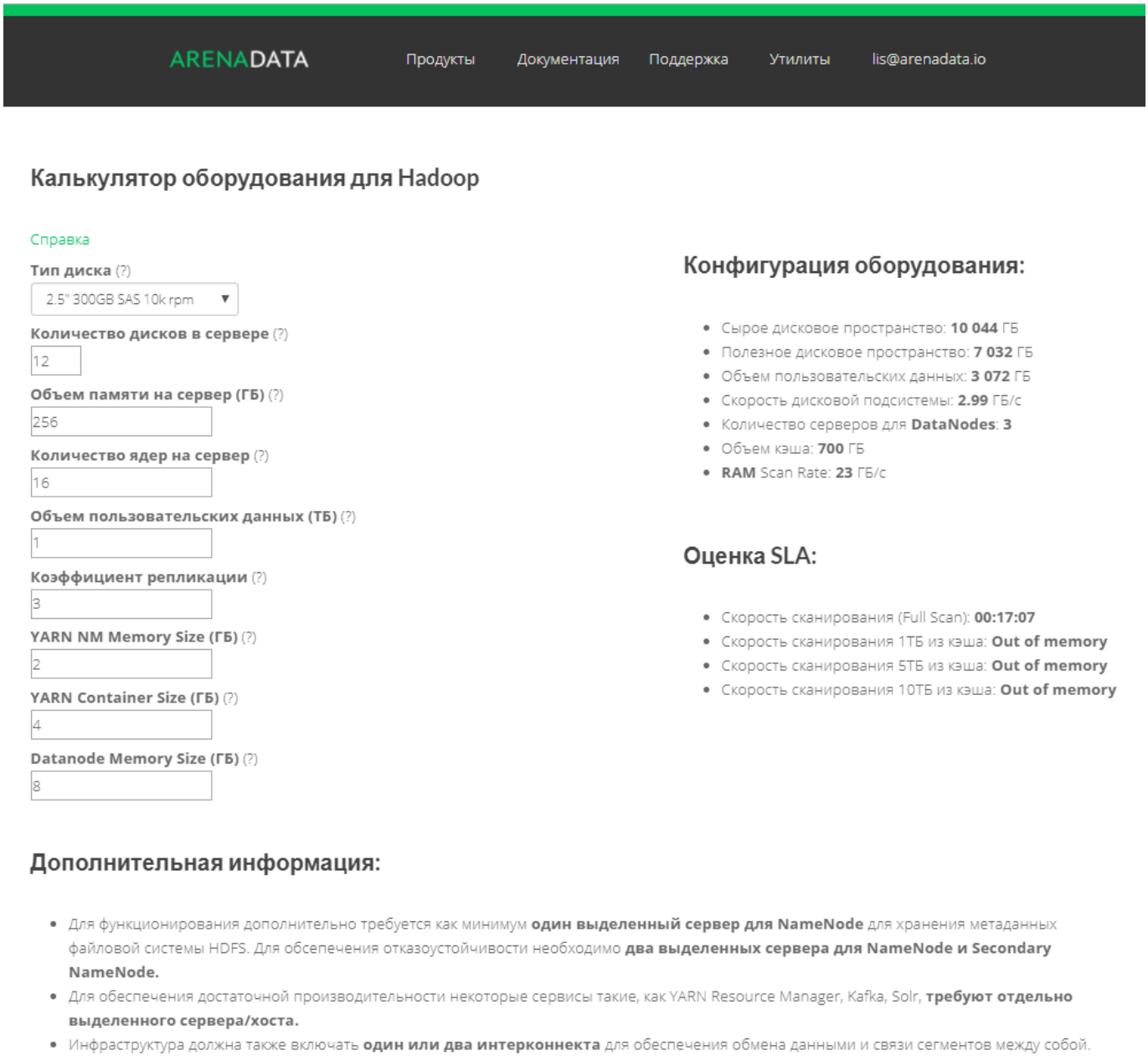



Рис.1.5.: Калькулятор оборудования для Hadoop

# Глава 2

# Подготовка к установке кластера

В разделе “Подготовка к установке кластера” приведена информация и даны материалы, которые необходимо подготовить для установки кластера **ADH** с помощью **Ambari**:

- Совместимость стека;
- Системные требования;
- Сбор информации;
- Подготовка к установке Hadoop.

**Ambari** обеспечивает комплексное решение для управления и мониторинга кластера **ADH**. Используя API-интерфейс **Ambari Web UI** и **REST**, можно развертывать, разрабатывать, управлять изменениями конфигурации и контролировать службы всех узлов кластера с центральной точки.

## 2.1 Системные требования

Для корректного запуска **Hadoop** необходимо соблюдение минимальных системных требований, представленных в таблице, а также описанных в разделах *Требования к памяти*, *Требования к объему пакета и количеству индексных дескрипторов* и *Проверка максимального количества открытых файловых дескрипторов*.

Таблица2.1.: Минимальные системные требования для установки кластера Arenadata Ambari

| Компонент | Требования |
|---|---|
| Платформа | • Intel x86_64<br>• IBM Power (Little-Endian) |

| Компонент | Требования |
|---|---|
| Операционная система | Поддерживаются следующие 64-разрядные ОС:<br>• Red Hat Enterprise Linux (RHEL) v7.x;<br>• CentOS v7.x;<br>• Suse Enterprise Linux 12sp3<br>Установщик использует множество пакетов из базовых репозиториев ОС. Если нет полного набора базовых репозиториев ОС, доступных для всех компьютеров во время установки, могут возникнуть проблемы.<br>В случае возникновения проблем с недоступными базовыми хранилищами ОС, необходимо обратиться к системному администратору для обеспечения проксирования или зеркалирования этих дополнительных репозиториев |
| Браузер | Мастер установки Ambari работает как веб-приложение на основе браузера. Поэтому необходимо наличие машины, способной использовать графический браузер для применения данного веб-инструмента.<br>Минимальные требуемые версии браузера:<br>• Internet Explorer 9.0 (устаревшая версия);<br>• Firefox 18;<br>• Google Chrome 26;<br>• Safari 5;<br>На платформе рекомендуется обновить браузер до последней, стабильной версии (кроме Internet Explorer 9.0) |
| Программное обеспечение | На всех узлах кластера необходимо установить следующие компоненты:<br>• YUM и rpm (RHEL / CentOS);<br>• Zypper и php_curl (SLES);<br>• Scp, curl, unzip, tar и wget;<br>• OpenSSL (v1.01, build 16 или новее);<br>• Python v2.7 |
| Java | Поддерживаются следующие среды выполнения Java:<br>• Oracle JDK 1.8 64-разрядный: минимум JDK 1.8.64;<br>Во время установки необходимо выбрать JDK 1.8, поскольку ADH 1.5 несовместим с JDK 1.7 |

| Компонент | Требования |
|---|---|
| База данных | Ambari требует реляционную базу данных для хранения информации о конфигурации кластера и топологии. При установке ADH Stack с Hive или Oozie – для них также требуется реляционная база данных<br>• Ambari:<br>– PostgreSQL 8;<br>– PostgreSQL 9.1.13+, 9.3;<br>– MariaDB 5.5;<br>– Oracle 11g R2, 12c.<br>По умолчанию Ambari устанавливает инстанс PostgreSQL на хост сервера Ambari. При этом использование данного инстанса PostgreSQL, MySQL или Oracle необязательно. Дополнительная информация указана в пункте «Настройка сервера Ambari 2.6.1»<br>• Hive:<br>– PostgreSQL 8;<br>– PostgreSQL 9.1.13+, 9.3;<br>– MariaDB 5.5;<br>– Oracle 11g R2, 12c.<br>По умолчанию (на RHEL / CentOS), Ambari устанавливает инстанс MySQL на хосте Hive Metastore. Также можно использовать указанный инстанс PostgreSQL, MySQL или Oracle. Дополнительная информация указана в пункте «Настройка сервера Ambari 2.6.1»<br>• Oozie:<br>– PostgreSQL 8;<br>– PostgreSQL 9.1.13+, 9.3;<br>– MariaDB 5.5;<br>– Oracle 11g R2, 12c.<br>По умолчанию Ambari устанавливает инстанс Derby на хосте Oozie Server. При этом использование данного инстанса PostgreSQL, MySQL или Oracle необязательно. Использование Derby в качестве базы метаданных Oozie не рекомендуется для продуктивного использования. Дополнительная информация указана в пункте «Настройка сервера Ambari 2.6.1».<br>Инстанс Derby для рабочей среды не должен использоваться по умолчанию. Если планируется применение Derby для demo-версии, разработки или тестирования окружающей среды, перенос базы данных Oozie из Derby в новую базу данных доступен только в Community Edition<br>• Ranger:<br>– PostgreSQL 9.1.13+, 9.3;<br>– MariaDB 5.5;<br>– Oracle 11g R2, 12c.<br>Для Ranger необходим указанный инстанс PostgreSQL, MySQL или Oracle.<br>Для базы данных Ambari, если используется существующая база данных Oracle, необходимо убедиться, что инстанс Oracle работает на порте, отличном от 8080. Это необходимо, чтобы избежать конфликта с портом Ambari, установленном по умолчанию. Также следует обратиться к “Справочному руководству Ambari” для получения информации о возможных изменениях порта по умолчанию сервера Ambari.<br>Использование параметров базы данных Microsoft SQL Server или SQL Anywhere не поддерживается.<br>Если используется существующая база данных, развернутая на том же сервере, что и Ambari Server, необходимо, чтобы эта база работала на порте, отличном от 8080 |

### 2.1.1 Требования к памяти

Хост **Ambari** должен иметь как минимум *1 ГБ* оперативной памяти и *500 МБ* свободной. На любом узле можно проверить доступную память, выполнив команду:

```
free -m
```

В случае если планируется установить **Ambari Metrics Service** в кластер, для уточнения требований к ресурсам необходимо ознакомиться с разделом "Параметры настройки Ambari Metrics" в Справочном руководстве **Ambari**. В общем случае хосту, на котором планируется запускать **Ambari Metrics Collector**, необходимо иметь доступную память и дисковое пространство в зависимости от размера кластера.

Таблица2.2.: Требования к объему памяти в зависимости от размера кластера

| Количество хостов | Доступная память | Дисковое пространство |
|---|---|---|
| 1 | 1024 MB | 10 GB |
| 10 | 1024 MB | 20 GB |
| 50 | 2048 MB | 50 GB |
| 100 | 4096 MB | 100 GB |
| 300 | 4096 MB | 100 GB |
| 500 | 8096 MB | 200 GB |
| 1000 | 12288 MB | 200 GB |
| 2000 | 16384 MB | 500 GB |

Сведения в таблице предлагаются в качестве рекомендаций. Необходимо проверять каждый частный случай.

### 2.1.2 Требования к объему пакета и количеству индексных дескрипторов

Таблица2.3.: Приблизительные значения размера пакета и количества индексных дескрипторов

| | Размер пакета | Количество индексных дескрипторов |
|---|---|---|
| Ambari Server | 100MB | 5.000 |
| Ambari Agent | 8MB | 1.000 |
| Ambari Metrics Collector | 225MB | 4.000 |
| Ambari Metrics Monitor | 1MB | 100 |
| Ambari Metrics Hadoop Sink | 8MB | 100 |
| After Ambari Server Setup | Данные отсутствуют | 4.000 |
| After Ambari Server Start | Данные отсутствуют | 500 |
| After Ambari Agent Start | Данные отсутствуют | 200 |

### 2.1.3 Проверка максимального количества открытых файловых дескрипторов

Рекомендуемое максимальное количество открытых файловых дескрипторов – более *10000*. Для проверки текущего значения, заданного для максимального открытых файловых дескрипторов, необходимо выполнить следующие команды на каждом хосте:

```
ulimit -Sn
```

```
ulimit -Hn
```

Если значение максимального количества открытых файловых дескрипторов не превышает *10000*, то следует установить подходящее значение, выполнив команду:

```
ulimit -n 10000
```

## 2.2 Сбор информации

Перед развертыванием кластера **ADH** необходимо:

- Проверить полное доменное имя (FQDN) каждого хоста в кластере. Мастер установки Ambari поддерживает использование IP-адресов. Для проверки и установки FQDN необходимо воспользоваться командой:

  ```
  hostname -f
  ```

Развертывание всех компонентов **ADH** на одном хосте возможно, но не рекомендуется для продуктивного использования. Как правило, для минимального кластера настраивается три узла – один главный и два подчиненных.

- Определить список компонентов, которые необходимо настроить на каждом узле;
- Установить базовые каталоги, которые будут определены в качестве точек для хранения:
  - Узла NameNode;
  - Узла (узлов) DataNode (в этих каталогах не должно содержаться никаких других данных);
  - Узла Secondary NameNode;
  - Узла Oozie;
  - Узла YARN;
  - Узлов ZooKeeper, если устанавливается ZooKeeper;
  - Различных журналов, файлов *pid* и *db*, в зависимости от типа установки.

Необходимо использовать каталоги, которые предоставляют постоянные места хранения компонентов **ADH** и данных **Hadoop**.

**Important:** Не используйте */TMP* каталог для установки, так как файлы могут быть удалены в любое время

## 2.3 Подготовка к установке Hadoop

Для корректного развертывания **Arenadata Hadoop** необходимо выполнить следующие действия:

- Настройка беспарольного SSH доступа к хостам;
- Настройка учетных записей компонентов (опционально);
- Синхронизация часов на всех узлах кластера;
- Настройка DNS и NSCD;
- Настройка IPTables;
- Настройка SELinux, PackageKit и Umask.

### 2.3.1 Настройка беспарольного SSH доступа к хостам

Для **Ambari Server** необходимо, чтобы **Ambari Agents** были установлены на всех узлах кластера. Сервер **Ambari** связывается с агентами для выполнения установки и управления задачами кластера **Arenadata**.

Для автоматической установки Ambari-агентов на всех узлах кластера сервером **Ambari** необходимо для учетной записи, под которой производится установка (по умолчанию от *root*), настроить беспарольное SSH-соединение между хостом **Ambari Server** и всеми другими хостами в кластере.

Для настройки беспарольного **SSH** необходимо под выбранной учетной записью выполнить следующие инструкции (на примере *root*):

- Создать публичный и приватный ключи SSH на хосте сервера Ambari:

  ```
  ssh-keygen
  ```

- При запросе пароля для приватного ключа необходимо не задавая значения (поле для ввода пароля оставить пустым) нажать клавишу *Enter*;

- Скопировать публичный ключ SSH (*id_rsa.pub*) в учетную запись *root* на все узлы кластера:

  ```
  ssh-copy-id root@<remote.target.host>
  ```

  где *<remote.target.host>* – значение имени каждого хоста в кластере;

  В случае если во время первого подключения отображается предупреждающее сообщение, необходимо ответить *yes*:

  ```
  Are you sure you want to continue connecting (yes/no)?
  ```

- Убедиться, что с сервера Ambari выполняется подключение к каждому хосту в кластере с помощью SSH без пароля:

  ```
  ssh root@<remote.target.host>
  ```

- (Опционально) Сохранить копию приватного ключа SSH на компьютере, с которого планируется запуск веб-мастера установки Ambari Install Wizard.

Хост сервера **Ambari** использует аутентификацию публичного ключа **SSH** для удаленного доступа и установки Ambari-агента.

Так же Ambari-агенты возможно установить вручную на каждом узле кластера. В этом случае не будет необходимости создавать и распространять ключи **SSH**.

### 2.3.2 Настройка учетных записей пользователей

Для каждого сервиса **ADH** необходима учетная запись пользователя ОС. Мастер установки **Ambari** создает новые и сохраняет все существующие учетные записи пользователей и использует их при настройке сервисов **Hadoop**.

### 2.3.3 Синхронизация часов на всех узлах кластера

Часам на всех узлах кластера и на машине, запускающей браузер для доступа к веб-интерфейсу **Ambari**, необходимо иметь возможность синхронизации друг с другом. Для этого следует включить службу **NTP** и убедиться, что синхронизация происходит автоматически:

- RHEL/CentOS 7 & SUSE/SLES 12:

  ```
  systemctl is-enabled ntpd
  ```

Для настрйоки службы **NTP** на автоматический запуск при загрузке машины, необходимо выполнить следующую команду на каждом хосте:

- RHEL/CentOS 7 & SUSE/SLES 12:

  ```
  systemctl enable ntpd
  ```

  Для запуска службы **NTP**, необходимо выполнить команду на каждом хосте:

- RHEL/CentOS 7 & SUSE/SLES 12:

  ```
  systemctl start ntpd
  ```

### 2.3.4 Настройка DNS и NSCD

Все узлы в кластере должны быть настроены как для прямого, так и для обратного **DNS**.

В случае если не получается настроить **DNS**, необходимо отредактировать файл */etc/hosts* на каждом узле кластера так, чтобы он содержал IP-адрес и полное доменное имя (**FQDN**) каждого узла.

Следующие инструкции представлены в виде обзора и охватывают базовую настройку сети для общих узлов **Linux**. Разные версии и варианты **Linux** могут потребовать различные команды и процедуры.

**Hadoop** сильно зависит от **DNS** и выполняет многие DNS-запросы во время работы. Чтобы снизить нагрузку на инфраструктуру **DNS**, настоятельно рекомендуется использовать **Name Service Caching Daemon** (**NSCD**) на узлах кластера под управлением **Linux**. **NSCD** кэширует запросы хоста, пользователя и группы, а также обеспечивает лучшую производительность и снижает нагрузку на инфраструктуру **DNS**.

### 2.3.5 Редактирование файла хоста

Для редактирования файла хоста необходимо используя текстовый редактор открыть файл *hosts* на каждом узле кластера:

```
vi/etc/hosts
```

И добавить строку для каждого хоста, состоящую из IP-адреса и **FQDN**, например:

```
1.2.3.4 <fully.qualified.domain.name>
127.0.0.1 localhost.localdomain localhost
::1 localhost6.localdomain6 localhost6
```

### 2.3.6 Проверка имени хоста

Чтобы убедиться, что имя хоста установлено, необходимо выполнить следующую команду:

```
hostname -f
```

В результате именю хоста возвращается значение *fully.qualified.domain.name*. Для задания имени на каждом узле кластера необходимо использовать команду:

```
hostname <имя узла>
```

### 2.3.7 Редактирование файла конфигурации сети

С помощью текстового редактора необходимо открыть файл конфигурации сети на каждом узле и установить ему требуемую конфигурацию сети для каждого узла:

`vi /etc/sysconfig/network`

Следует изменить свойство *HOSTNAME*, чтобы задать полное доменное имя:

```
NETWORKING=yes
HOSTNAME=<fully.qualified.domain.name>
```

### 2.3.8 Настройка IPTables

Для взаимодействия во время установки **Ambari** с развернутыми узлами необходимо, чтобы определенные компоненты были открыты и доступны. Самый простой способ сделать это – временно отключить **IPTables**:

- RHEL/CentOS 7:

```
systemctl disable firewalld
systemctl stop firewalld
```

- SUSE/SLES 12:

```
rcSuSEfirewall2 stop
chkconfig SuSEfirewall2 off
```

По завершению установки следует перезапустить **IPTables**.

В случае если протоколы безопасности предотвращают отключение **IPTables**, можно продолжить работу с включенными **IPTables** при условии, что все необходимые порты открыты и доступны.

В процессе установки Ambari-сервера **Ambari** проверяет, работают ли **IPTables**. В случае если **IPTables** запущены, выдается предупреждение, напоминающее о необходимости проверить, что требуемые порты открыты и доступны.

В мастере установки на шаге подтверждения узла также выдается предупреждение для каждого хоста с запущенными **IPTables**.

### 2.3.9 Настройка SELinux, PackageKit и Umask

Работа с **SELinux** не поддерживается, поэтому необходимо его отключить. Для этого следует выполнить команду:

- RHEL/CentOS 7:

```
:command:`setenforce 0`
```

И установить значение */etc/selinux/config*.

На установочном узле, где запущен RHEL/CentOS с установленным **PackageKit**, с помощью текстового редактора необходимо открыть */etc/yum/pluginconf.d/refresh-packagekit.conf* и выполнить следующее изменение:

- RHEL/CentOS 7:

```
:command:`enabled=0`
```

**UMASK** устанавливает разрешения по умолчанию или базовые разрешения при создании нового файла или папки в системе **Linux**. Большинство дистрибутивов **Linux** устанавливают **UMASK** значение по умолчанию *022*. Это разрешение допускает чтение, запись и выполнение *755* для новых файлов или папок. Значение **UMASK** равное *027* допускает чтение, запись и выполнение *750* для новых файлов или папок.

**Ambari** поддерживает оба значения **UMASK**. Например, чтобы установить значение **UMASK** равное *022*, необходимо запустить команду в корне на всех хостах:

```
vi /etc/profile
```

Затем добавить следующую строку:

```
umask 022
```

# Глава 3

# Установка сервера Ambari 2.6.1

Установка сервера **Ambari 2.6.1** проходит в несколько этапов:

- Настройка репозитория;
- Установка сервера Ambari;
- Настройка сервера Ambari;
- Запуск сервера Ambari.

## 3.1 Настройка репозитория

Установка сервера **Ambari** выполняется с помощью пакетного менеджера из репозитория, содержащего соответствующий пакет. При этом допускается использование как удаленного репозитория, доступного через сеть Интернет, так и размещенного в локальной сети кластера (например, если из соображений безопасности доступ к сети Интернет ограничен).

**Important:** Репозиторий должен находиться на доступном со всех узлов кластера хосте

### 3.1.1 Настройка удаленного репозитория

Настройка удаленного репозитория не отличается от настройки любого дополнительного репозитория. Для добавления репозитория необходимо выполнить от имени *root* команду:

- RHEL/CentOS 7:

```
yum-config-manager --add-repo <URL репозитория Ambari>/ambari.repo
```

- SUSE/SLES 12:

```
zypper addrepo <URL репозитория Ambari>/ambari.repo
```

### 3.1.2 Настройка локального репозитория

Для настройки нового сервера репозитория необходим веб-сервер *httpd*. Следует убедиться, что сервер *httpd* запускается на хосте, который служит в качестве репозитория:

- RHEL/CentOS 7:

```
systemctl status httpd
```

- SUSE/SLES 12:

```
systemctl status apache2.service
```

В случае если сервер не запущен, необходимо его установить и запустить:

- RHEL/CentOS 7:

```
systemctl start httpd
```

- SUSE/SLES 12:

```
systemctl start apache2.service
```

### Создание промежуточного каталога

Для извлечения архивов для стеков **Ambari** и **ADH** рекомендуется использовать промежуточный каталог.

Каждый архив представляет собой архивный репозиторий и имеет скрипт *setup_ repo.sh*, создающий ссылку из корня документа *httpd server/var/www/html* в каталог, из которого извлекается архив. Необходимо, чтобы промежуточный каталог и все верхнеуровневые каталоги были читаемыми и доступными пользователю, выполняющему процесс *httpd* (**apache**), а лучше сделать их доступными для всех пользователей кластера:

```
mkdir /staging
chmod a+rx /staging
```

**Important:** Не используйте каталог */TMP* в качестве промежуточного, так как файлы могут быть удалены в любое время

### Загрузка и распаковка архива Ambari

**Arenadata Ambari** поставляется как архив репозитория, который необходимо извлечь на сервер репозитория.

На узел, который используется в качестве репозитория, необходимо загрузить архив **Ambari 2.6.1** в ранее созданный промежуточный каталог, или в каталог:

- RHEL/CentOS 7:

```
wget https://storage.googleapis.com/arenadata-repo/ambari/2.6.1/repos/ambari.repo -O /etc/yum.repos.d/
```

- SUSE/SLES 12:

```
wget https://storage.googleapis.com/arenadata-repo/ambari/2.6.1/repos/ambari-sles.repo - O /etc/zypp/
↪repos.d/
```

- PPC64le RHEL7:

```
wget https://storage.googleapis.com/arenadata-repo/ambari/2.6.1/repos/ambari-ppc64le.repo - O /etc/
↪zypp/repos.d/
```

Необходимо убедиться, что все родительские каталоги до промежуточного имеют доступ *"r + x"* для всех пользователей, поскольку данный каталог будет использоваться для создания локального репозитория.

После загрузки **Ambari 2.6.1** необходимо извлечь архив в промежуточный каталог. Например:

```
tar -xvf /staging/AMBARI-2.6.1.tar -C /staging/
```

#### Настройка локального репозитория

Для настройки локального репозитория необходимо на хосте, используемом в качестве репозитория, выполнить скрипт *setup_ repo.sh*, входящий в состав архива **Ambari**:

```
/staging/AMBARI-2.6.1/setup_repo.sh
```

В скрипте предполагается, что в корневом каталоге репозитория веб-сервер устанавливает */var /www /html* и создает ссылку *ambari-<версия>*, указывающую на извлеченный архив.

Необходимо убедиться, что репозиторий доступен на веб-сервере:

```
curl http://localhost/AMBARI-2.6.1/repodata/repomd.xml
```

Скрипт также создает определенный репозиторий **Ambari** и помещает его в файл:

- RHEL/CentOS 7:

```
/etc/yum.repos.d/ambari.repo
```

- SUSE/SLES 12:

```
/etc/zypp/repos.d/ambari.repo
```

Данный файл должен быть доступен на хосте администратора, где будет установлен сервер **Ambari**.

**Important:** Репозиторий Ambari должен быть доступен для всех узлов кластера

Необходимо проверить наличие доступа к следующему URL-адресу с хоста администратора и с узлов кластера:

```
http://<yum.repo.host.fqdn>/AMBARI-2.6.1
```

## 3.2 Установка сервера Ambari

Сервер **Ambari** устанавливается из RPM-пакета по команде:

- RHEL/CentOS 7:

```
yum install ambari-server
```

- SUSE/SLES 12:

```
zypper install ambari-server
```

Данная команда устанавливает сервер **Ambari**, являющийся сервером веб-приложений, на порт *8080*. Также устанавливает инстанс сервера **PostgreSQL** на порт *5432*.

## 3.3 Настройка сервера Ambari

Сервер **Ambari** необходимо настроить для корректной работы.

В случае если инстанс **PostgreSQL** настроен на порт по умолчанию, следует выполнить следующую команду:

```
ambari-server setup
```

В процессе настройки необходимо указать или принять по умолчанию параметры:

- *Учетная запись пользователя* – для запуска Ambari-сервера можно выбрать любую учетную запись (необязательно выполнять вход от *root*). В случае если пользователя не существует, он создается автоматически;
- *Java JDK* – для загрузки Oracle JDK 1.8 необходимо ввести значение *1* и принять лицензию Oracle JDK для загрузки файлов из Oracle. При этом установка JDK выполняется автоматически;
- *База данных* – выбор базы данных:

  ```
  Enter advanced database configuration
  ```

  В командной строке необходимо ответить *n* или *y*:

  – *n* – для использования с Ambari стандартной встроенной базы данных PostgreSQL. По умолчанию для базы данных PostgreSQL устанавливается имя “ambari” и логин / пароль принимают значения *ambari / bigdata*.

  – *y* – при необходимости использования с Ambari уже существующей базы данных PostgreSQL, MySQL или Oracle вместо предлагаемой по умолчанию. Далее для выбранной базы данных необходимо указать параметры подключения (см. Приложение 1.).

## 3.4 Запуск сервера Ambari

После установки сервера **Ambari** запуск его осуществляется по команде:

```
ambari-server start
```

Для проверки статуса сервера необходимо использовать команду:

```
ambari-server status
```

Для остановки сервера необходимо использовать команду:

```
ambari-server stop
```

Сервер **Ambari** доступен на порту *8080*. По умолчанию для него установлена следующая учетная запись:

User: *admin*

Password: *admin*

**Important:** Рекомендуется сменить пароль после первого входа в систему

Для входа в веб-интерфейс **Ambari** необходимо в адресной строке браузера указать адрес сервера:

```
http://<адрес сервера>:8080
```

При этом запрашивается логин и пароль. После авторизации открывается веб-интерфейс **Ambari** (Рис.3.1.).

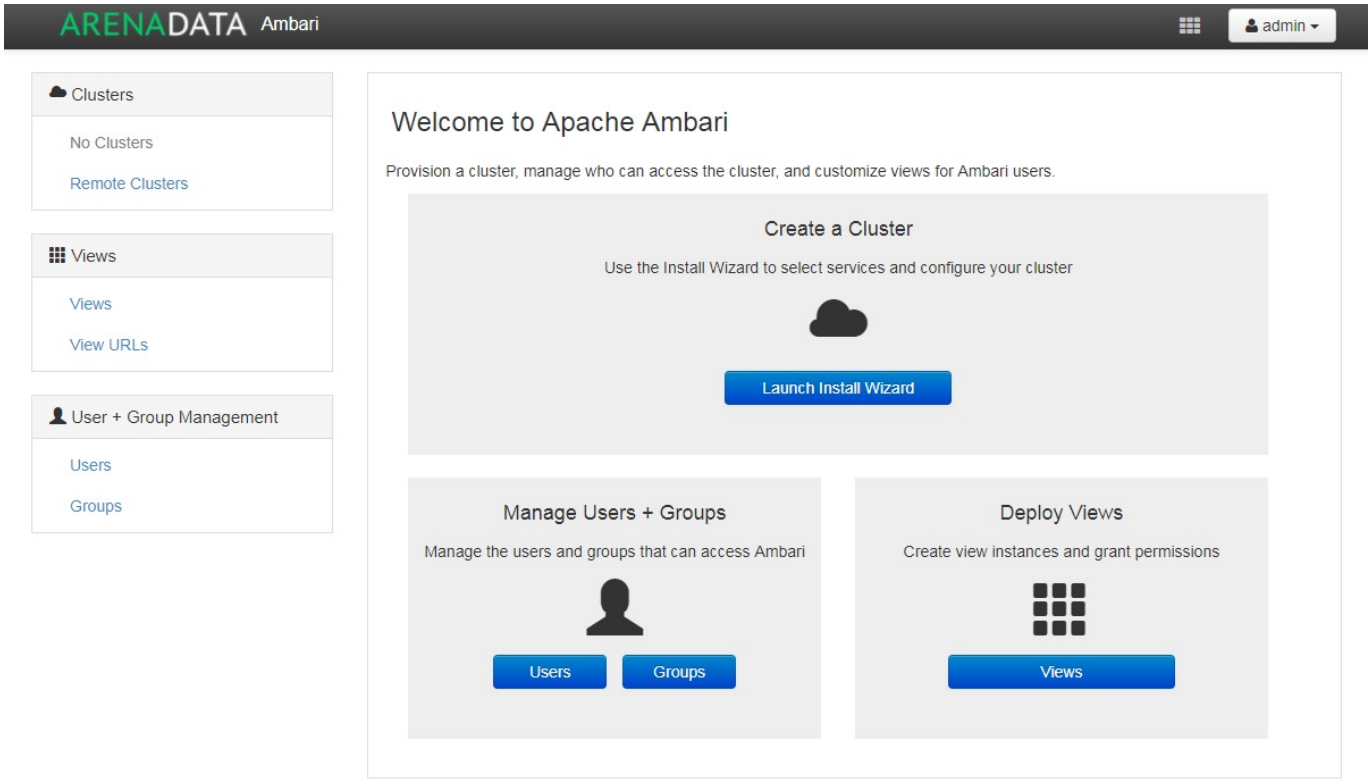



Рис.3.1.: Веб-интерфейс Ambari до настройки кластера

# Глава 4

# Подготовка к установке основных компонентов ADH на кластер

Основные компоненты **ADH** устанавливаются из репозиториев, которые определяются при первичной настройке кластера. Как и в случае репозитория **Ambari**, допускается использование удаленных и локальных репозиториев.

Удаленные репозитории уже заданы в **Ambari** как предлагаемые по умолчанию, для их настройки и использования не требуется дополнительных действий.

Для настройки локальных репозиториев необходимо выполнить действия, аналогичные настройке локального репозитория **Ambari**:

- Загрузить и извлечь архивы стека ADH;
- Настроить локальные репозитории.

## 4.1 Загрузка и извлечение архивов стека ADH

Архивы стека **ADH** необходимо установить на машине, где размещен репозиторий. В случае если для сервера репозитория используется выделенная машина, то архивы стека **ADH** следует установить на хосте администратора, использованном для установки сервера **Ambari**.

Необходимо загрузить и распаковать следующие архивы в выделенном для них месте (при этом следует избегать использования каталога */tmp*):

- *ADH-1.5.2* – RPM-пакеты для сервисов Hadoop, таких как HDFS, YARN, Hbase, Hive, Zookeeper;
- *ADH-UTILS-1.5.2* – дополнительные сервисы и библиотеки, используемые для мониторинга и оповещения серверов кластера.

В случае если архивы загружены в каталог */tmp*, то для их распаковки в каталоге, например, */staging* необходимо выполнить следующую команду:

```
tar -xvf /tmp/{stack}.tar -C /staging/
```

Для использования локальных репозиториев **ADH** и **ADH UTILS** необходимо выполнить настройки, описанные в пункте *Настройка локальных репозиториев*.

## 4.2 Настройка локальных репозиториев

Стек **ADH** поставляется в виде архива репозитория, который необходимо развернуть на сервере репозитория так, чтобы при этом он был доступен серверу **Ambari** и всем узлам кластера.

Каждый репозиторий стека содержит скрипт *setup_repo.sh*, для которого необходимо выполнение следующих требований:

- Сервер репозитория доступен всем узлам кластера;
- Корень сервера репозитория находится в */var/www/html/*.

Скрипт каждого стека создает символическую ссылку в документе сервера репозитория, указывающую на местоположение извлеченного архива стека, и создает файл с местоположением репозитория в каталоге:

- RHEL/CentOS 7:

```
/etc/yum.repos.d/
```

- SUSE/SLES 12:

```
/etc/zypp/repos.d/
```

Для каждого стека необходимо запустить скрипт установки локального репозитория:

`/staging/{stack}/setup_repo.sh`

По завершению установки скрипт выводит URL-адрес репозитория. Данный URL потребуется при установке кластера **ADH** с использованием сервера **Ambari**.

В случае если сервер репозитория установлен не на хосте администратора (где установлен сервер **Ambari**), необходимо скопировать созданные файлы определения местоположения репозитория из папки репозитрия на хост администратора, где установлен сервер **Ambari**.

Затем необходимо проверить правильность настройки репозитория, выполнив две команды от узла администратора:

- RHEL/CentOS 7:

```
yum clean all
yum repolist
```

- SUSE/SLES 12:

```
zypper clean -a
zypper repos
```

При корректной настройке выдается список репозиториев стека.

# Глава 5

# Запуск мастера установки

Для создания кластера необходимо после входа в **Ambari** запустить мастер установки, нажав в главной экранной форме кнопку *Launch Install Wizard*. Мастер установки проводит по шагам, необходимым для создания нового кластера **ADH**. Некоторые этапы установки требуют особого внимания:

- Изменение URL-адресов репозиториев;
- Ввод имен узлов и SSH-ключа или ручная установка Ambari-агентов;
- Выбор компонентов;
- Назначение мастер-узлов для компонентов;
- Назначение Slave и Client узлов;
- Дополнительные настройки компонентов.

## 5.1 Изменение URL-адресов репозиториев

Для того чтобы открыть список репозиториев, необходимо в блоке "Stacks" установить флаг в поле *ADH 1.5* и раскрыть блок "Advanced Repository Options", при этом **Ambari** предлагает указать URL-адреса репозиториев (Рис.5.1.).

В полях "Base URL" необходимо указать URL-адреса репозиториев, которые были получены при запуске скрипта *setup_ repo.sh*. Данные URL-адреса всегда можно уточнить в файлах <имя репозитория>.repo*.

Для установки из публичного репозитория **Arenadata** необходимо выбрать соответствующий пункт меню "Use Public Repository" (Рис.5.2.).

Репозитории можно обновить после развертывания кластера через **Ambari UI** (*"Admin → Repositories"*).

## 5.2 Ввод имен узлов и SSH-ключа

В разделе "Install Options" следует указать полные доменные имена (**FQDN**) для узлов, которые будут содержать кластер. Можно задавать диапазоны имен с помощью квадратных скобок, например, *host [01-10].domain* описывает *10* хостов. В случае если применяется **EC2**, необходимо использовать имена внутренних частных DNS-узлов.

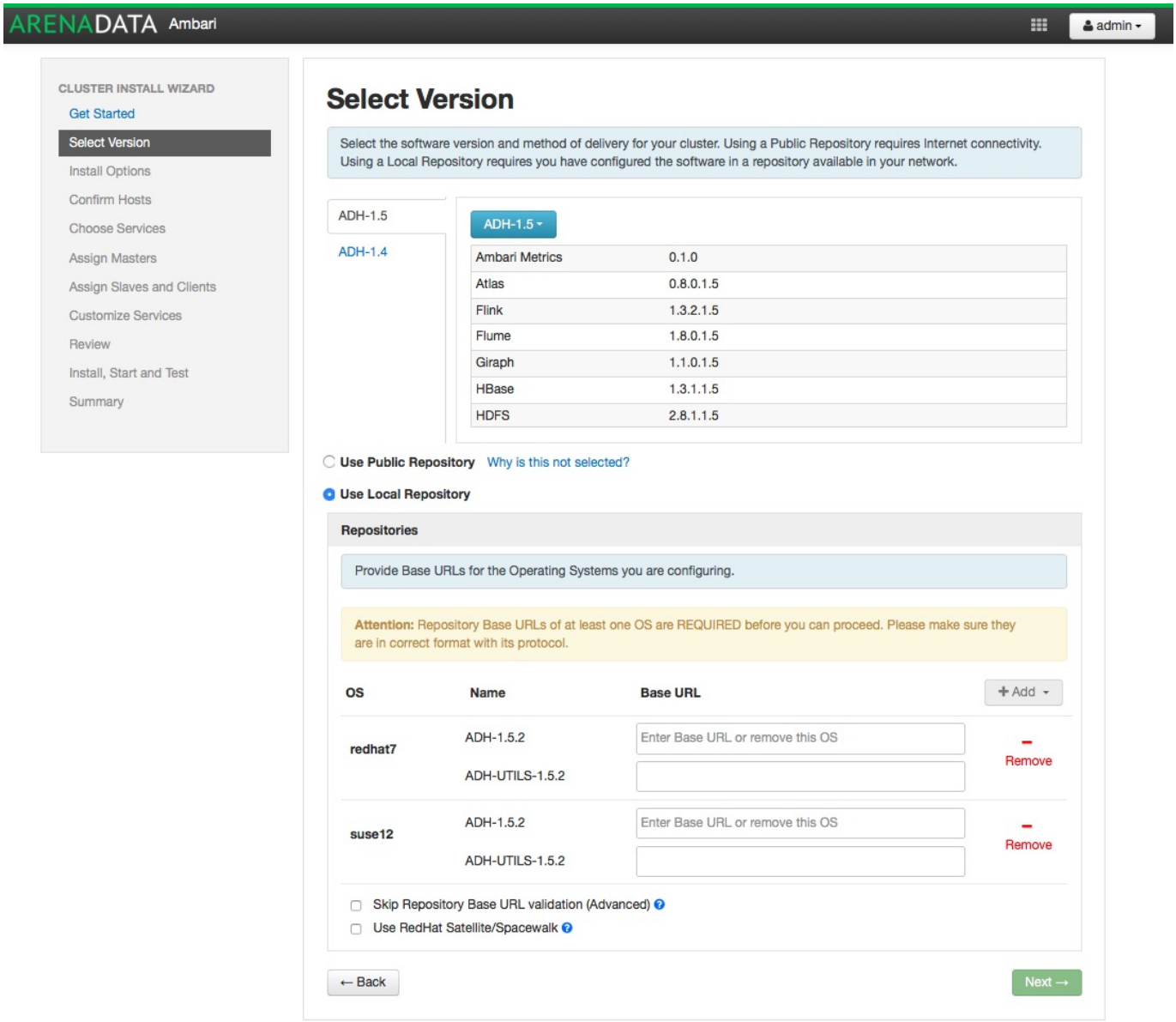



Рис.5.1.: Выбор стека

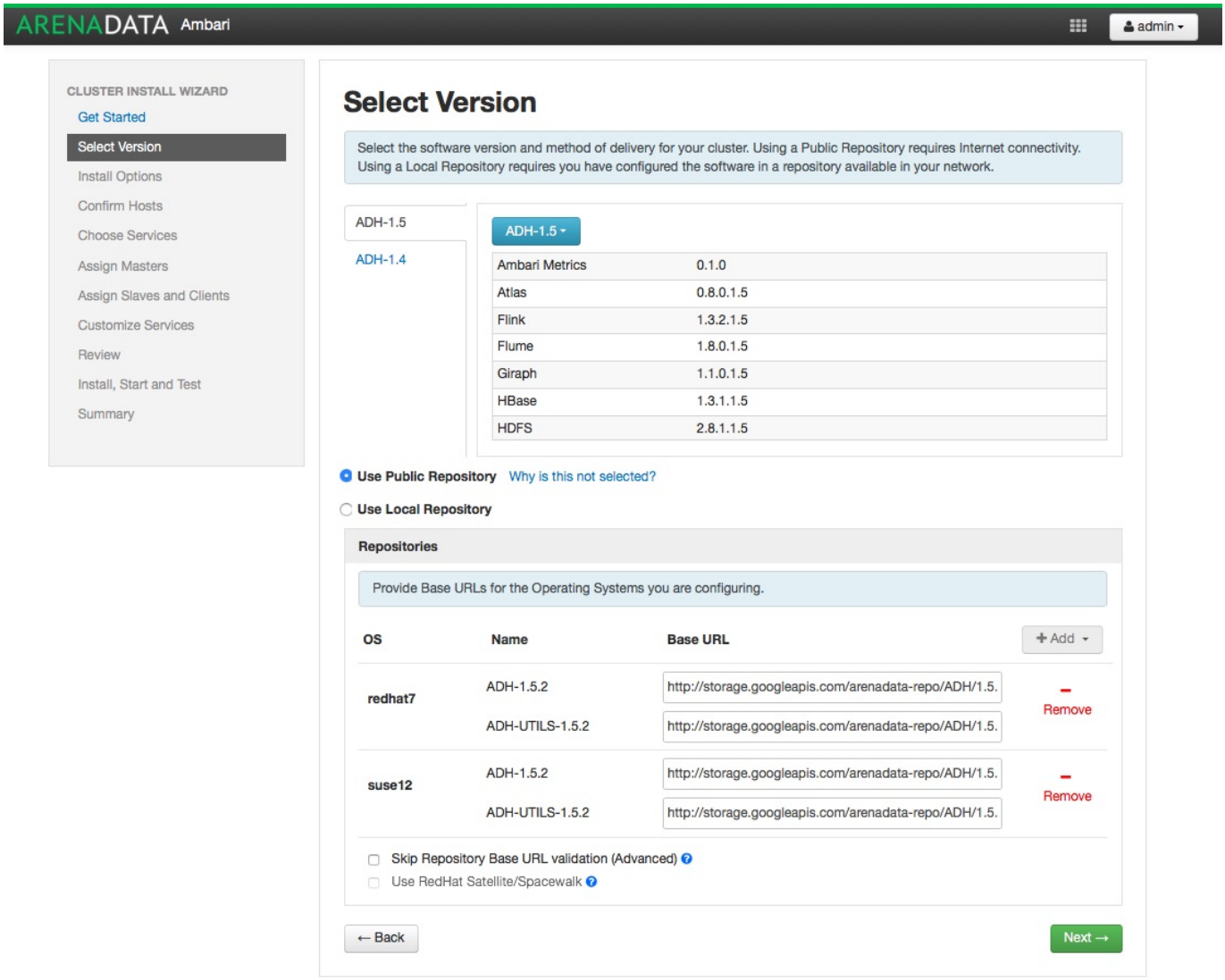



Рис.5.2.: Установка из публичного репозитория

Для автоматической регистрации Ambari-агентов на узлах кластера необходимо ввести закрытый ключ, который использовался для настройки беспарольного SSH для кластера. Можно передать сам файл *id_rsa* или скопировать и вставить его содержимое в экранную форму.

### 5.2.1 Ручная установка Ambari-агентов

В случае если нет возможности предоставить закрытый ключ беспарольного **SSH**, следует произвести установку Ambari-агентов вручную. Для этого на каждом узле кластера необходимо выполнить следующие шаги:

- Установить репозиторий Ambari, скопировав файл *ambari.repo* с сервера репозитория;
- Установить Ambari-агент, выполнив команду:
- RHEL/CentOS 7:

```
yum install ambari-agent
```

- SUSE/SLES 12:

```
zypper install ambari-agent
```

- Изменить конфигурацию Ambari-агента */etc/ambari-agent/conf/ambari-agent.ini* для определения его на сервере Ambari:
  - [server]
  - hostname={ambari.server.hostname}
  - url_port=8440
  - secured_url_port=8441
- Запустить Ambari-агент, выполнив команду:

  ```
  ambari-agent start
  ```

  Ambari-агент зарегистрируется на сервере при его запуске.

## 5.3 Выбор компонентов

На начальном этапе установки необходимо выбрать компоненты **ADH**, которые следует установить. При этом для любой инсталляции следует установить **HDFS** и **Zookeeper**, остальные компоненты возможно установить позднее (Рис.5.3.).

В случае если выбирается компонент **Ambari Metrics**, то для контроля кластера можно использовать **Ambari**. Если данный компонент не выбирается, выдается предупреждение, которое можно игнорировать в случае, если кластер планируется контролировать с помощью других инструментов. При этом **Ambari Metrics** можно будет добавить в кластер позднее.

## 5.4 Назначение мастер-узлов

Необходимо назначить мастер-узлы компонентов кластера (Рис.5.4.).

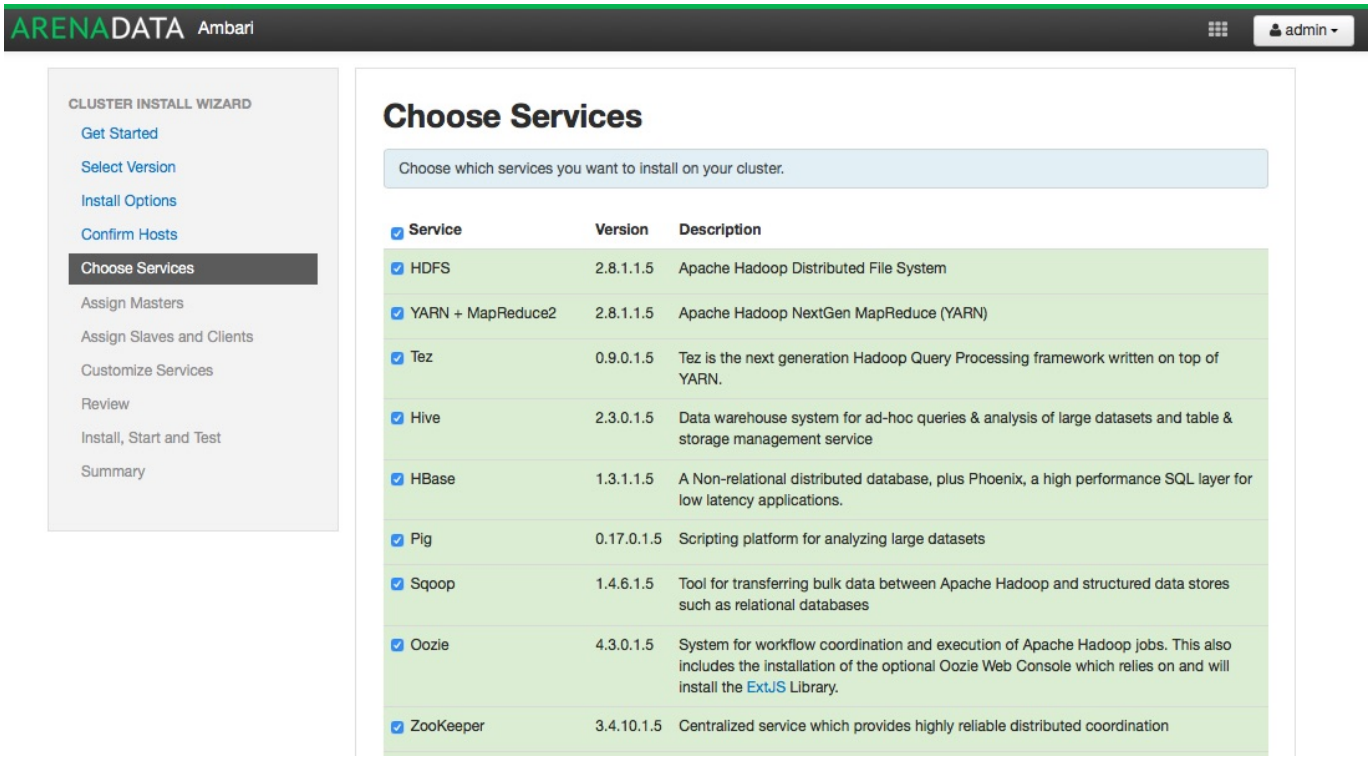



Рис.5.3.: Выбор компонентов

**Important:** Если Hive Metastore использует новую базу данных *PostgreSQL*, компонент HIVE METASTORE не должен находиться на хосте AMBARI

Данное ограничение объясняется тем, что оба компонента будут пытаться использовать порт *5432*. В случае абсолютной необходимости совместного размещения указанных компонентов на одном и том же хосте, предварительно следует переконфигурировать базу данных **PostgreSQL** на порт, отличный от *5432*, и выбрать опцию “Existing PostgreSQL Database” для конфигурации **Hive Metastore**.

## 5.5 Назначение Slave и Client узлов компонентов кластера

Необходимо назначить **Slave** и **Client** узлы, на которых будут разворачиваться соответствующие компоненты кластера (Рис.5.5.).

## 5.6 Дополнительные настройки компонентов

На экранной форме отображаются параметры конфигурации отдельных компонентов, автоматически сгенерированных установщиком **Ambari** на основе параметров кластера. Параметры каждого компонента можно менять по своему усмотрению в зависимости от планируемого использования того или иного компонента кластера.

В случае если для какого-либо обязательного параметра установщик не может предложить значение по умолчанию, перед продолжением установки данные параметры необходимо указать вручную (на Рис.5.6.

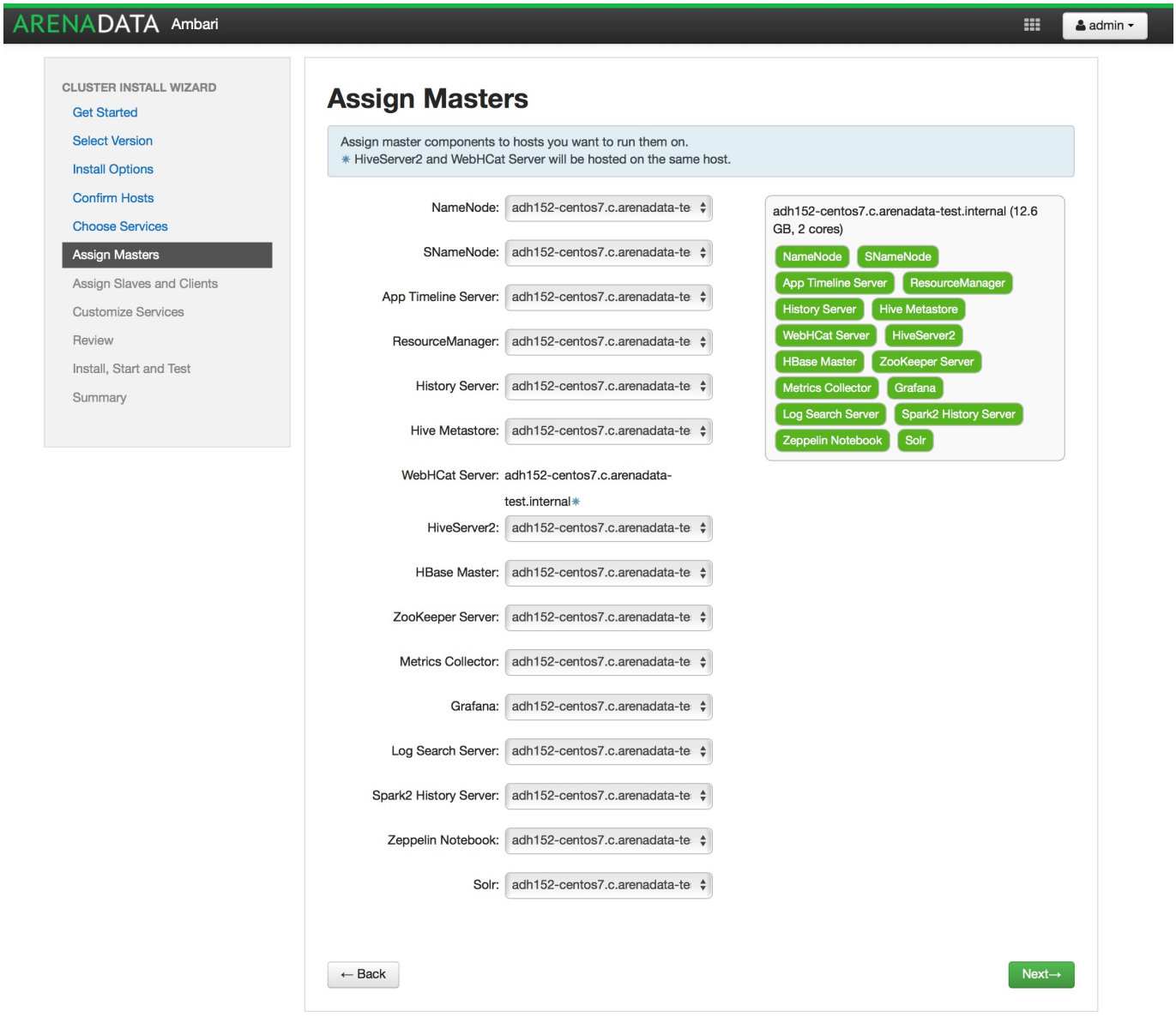



Рис.5.4.: Назначение мастер-узлов

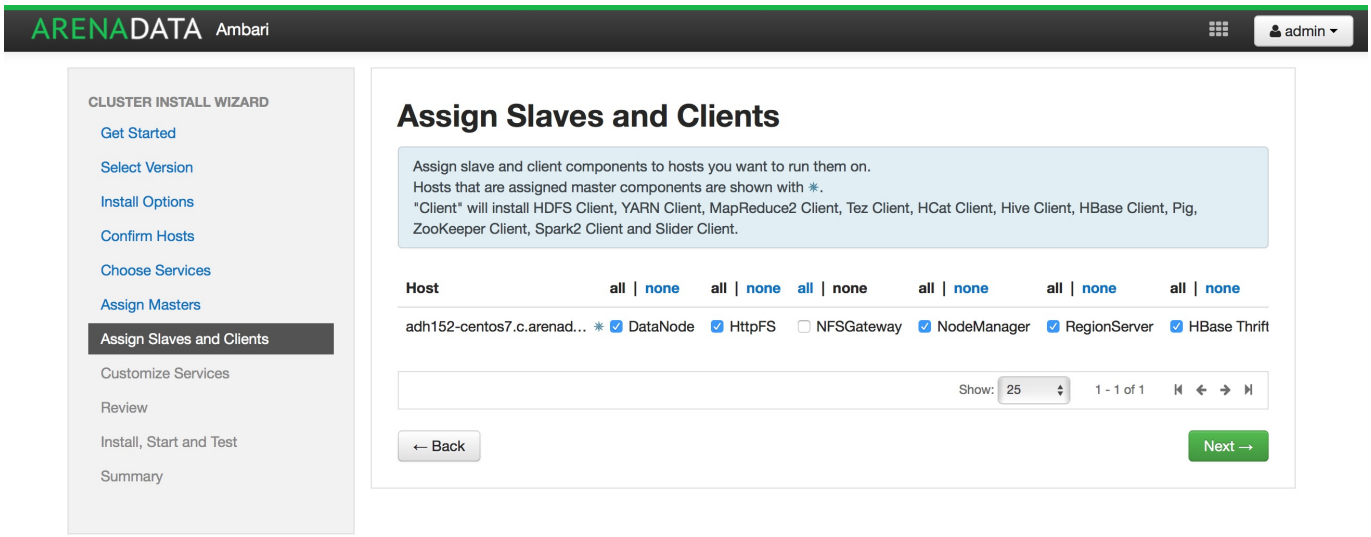



Рис.5.5.: Назначение Slave и Client узлов

приведен пример, когда для компонентов *Hive*, *Oozie*, *Ambari Metrics*, *Knox* необходимо указать пароли для внутренних баз данных).

**Important:** Каталоги для размещения данных HDFS (параметр “DataNode Directories” сервиса HDFS) не должны содержать никаких других данных, в том числе данных других компонентов. Это связано с тем, что при старте DataNode указанные каталоги очищаются, и может произойти потеря данных

**Important:** В случае если компонент DataNode сервиса HDFS устанавливается менее, чем на трех узлах кластера, необходимо задать соответствующее значение параметра *DFS Replication Factor* указанного компонента

## 5.7 Установка, запуск и тестирование

На экранной форме отображается ход развертывания кластера на каждом узле (Рис.5.7.).

Каждый компонент, который разворачивается вместе с хостом, устанавливается, запускается и проходит простой тест для проверки работоспособности.

При этом есть возможность просмотра подробной информации о завершенных и ожидающих задачах для каждого хоста (Рис.5.8.). Для этого необходимо нажать ссылку в столбце “Message” (см. Рис.5.7.).

По завершению установки компонентов появляется сообщение *Successfully installed and started the services*, в котором необходимо нажать кнопку *Next*.

Для окончания установки необходимо на странице “Summary” проверить список завершенных задач и нажать кнопку *Complete*. При этом открывается панель инструментов кластера.

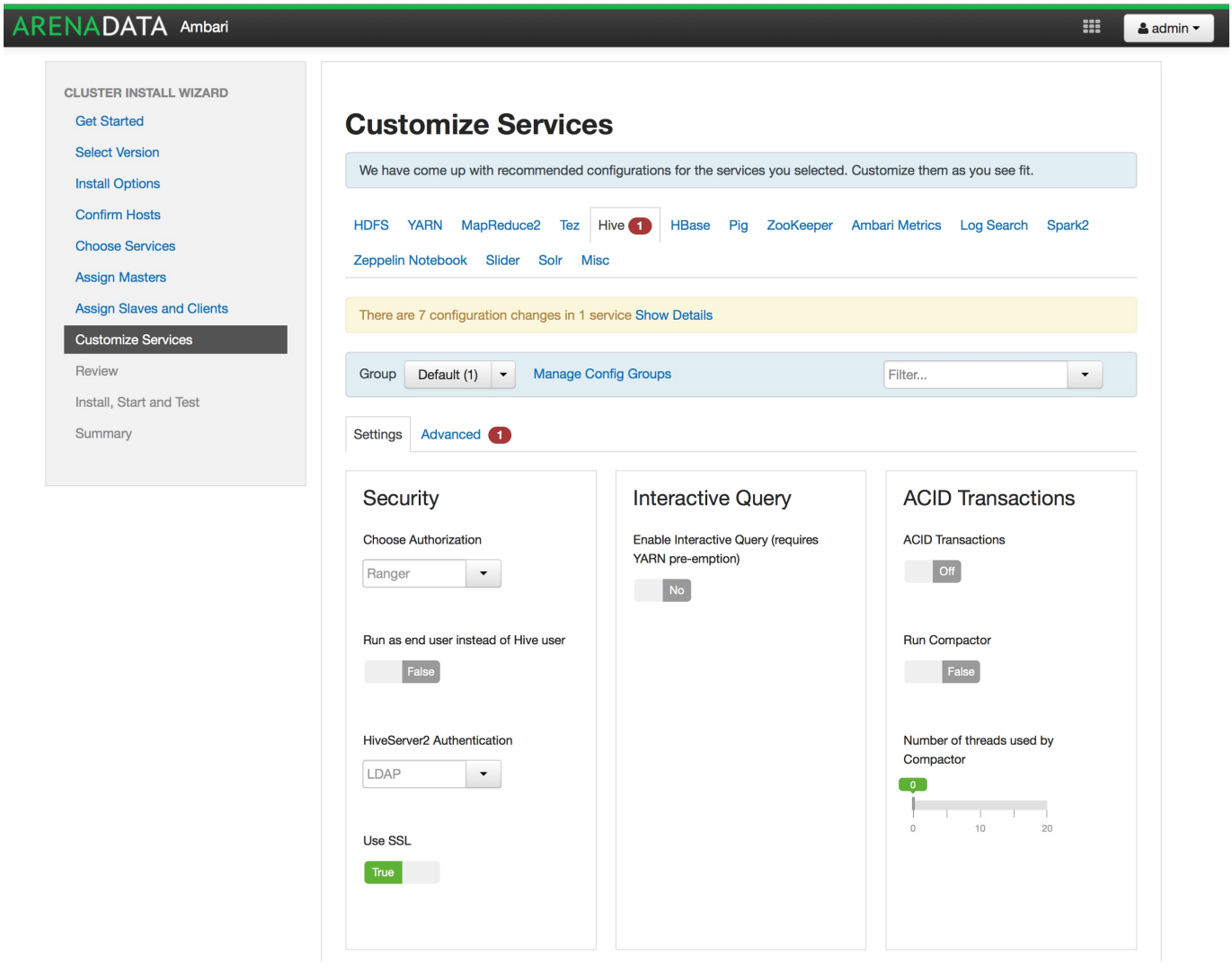



Рис.5.6.: Дополнительные настройки компонентов

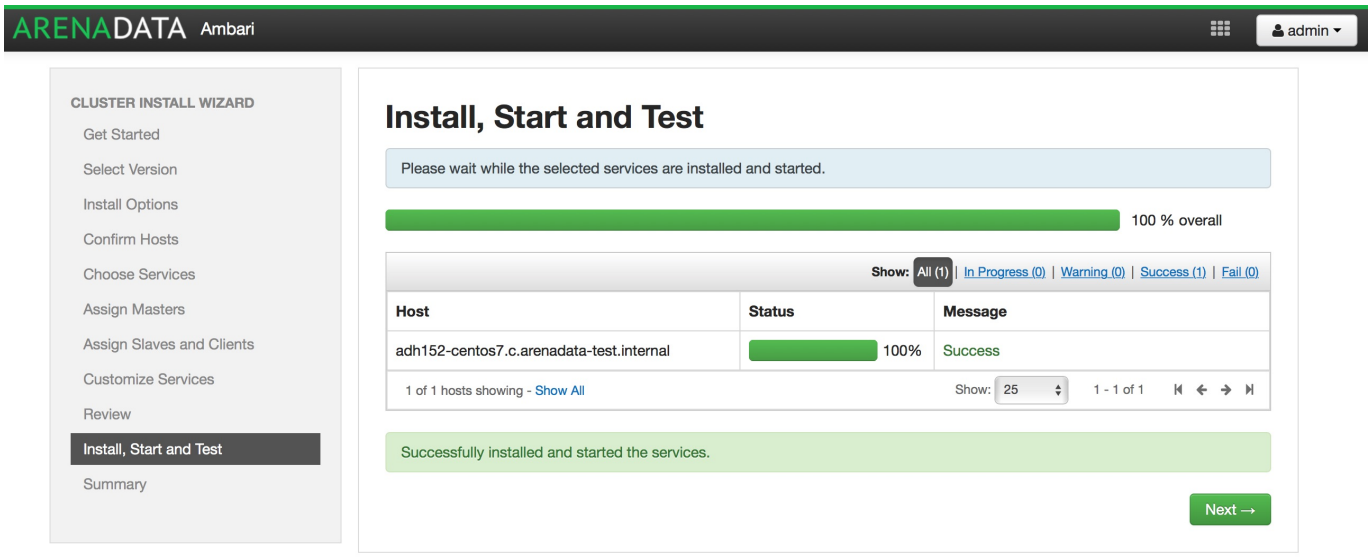



Рис.5.7.: Ход развертывания кластера

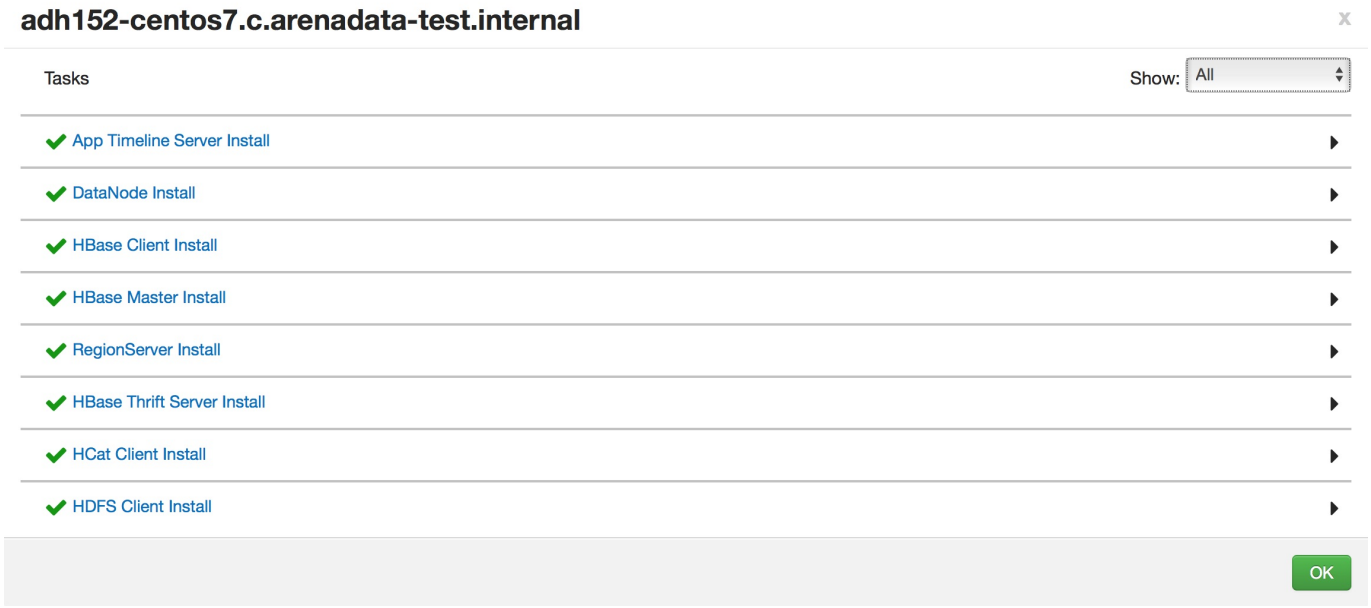



Рис.5.8.: Информация о задачах хоста

# Глава 6

# Панель кластера

Панель инструментов является центральным местом и отображает развернутые сервисы и их статус. Здесь можно добавлять новые сервисы или хосты, останавливать и запускать сервисы и компоненты, анализировать показатели мониторинга и выполнять конкретные действия сервиса (Рис.6.1.).

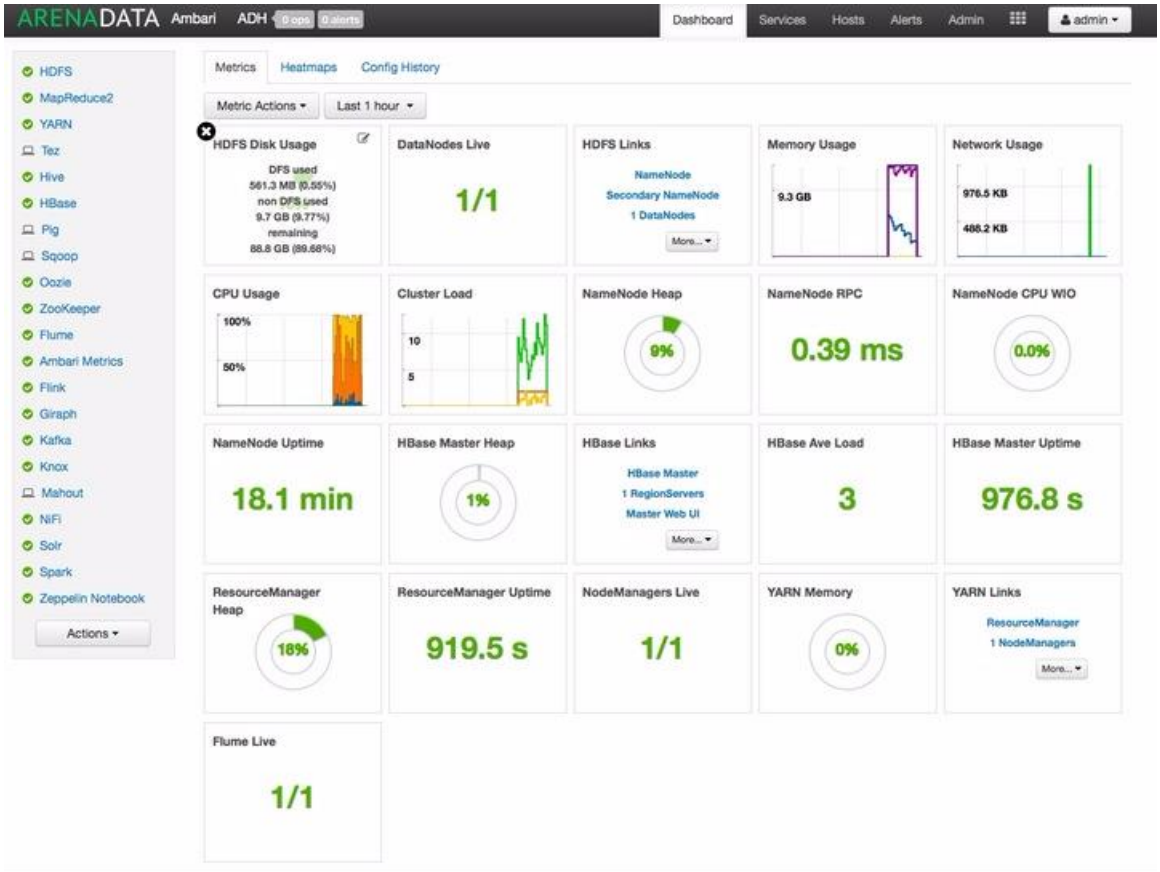



Рис.6.1.: Панель кластера

# Глава 7

# Приложение 1. Ручная настройка подключения к базе данных

Если в процессе настройки сервера **Ambari** необходимо отличное от используемого по умолчанию подключение к базе данных следует в командной строке нажать клавишу *y*:

```
Enter advanced database configuration
```

Если инстанс **PostgreSQL** настроен на порт, отличный от предлагаемого по умолчанию, для настройки **Ambari** необходимо выполнить следующие шаги:

- Открыть в текстовом редакторе конфигурационный файл PostgreSQL */var/lib/pgsql/data/pg_hba.conf*. Чтобы позволить пользователю *ambari* подключиться к базе данных, необходимо в конце файла добавить следующие строки:

```
local all ambari md5
host all ambari 0.0.0.0/0 md5
host all ambari ::/0 md5
```

- Чтобы подключить порт, выбранный не по умолчанию, следует открыть файл */etc/sysconfig/pgsql/postgresql* и добавить в него строку с номером необходимого порта. Например, чтобы подключить порт *10432* следует указать:

```
PGPORT=10432
```

- Перезапустить базу данных PostgreSQL:

```
service postgresql restart
```

- Подключиться к базе данных под *postgres* (супер-пользователь) и выполнить следующие настройки:

```
psql -U postgres -p 10432;
postgres=# CREATE DATABASE ambari;
postgres=# CREATE USER ambari WITH ENCRYPTED PASSWORD 'bigdata';
postgres=# \c ambari;
ambari=# CREATE SCHEMA ambari AUTHORIZATION ambari;
ambari=# ALTER SCHEMA ambari OWNER TO ambari;
ambari=# ALTER ROLE ambari SET search_path to 'ambari','public';
ambari=# \q
```

- Выполнить команду установки Ambari:

```
ambari-server setup --database=postgres --databasehost=localhost--databaseport=10432 --
↪databasename=ambari --databaseusername=ambari--databasepassword=bigdata
```

- Чтобы убедиться, что *postgres* подключен к хосту *databasehost*, необходимо использовать следующую команду:

```
netstat -anp | egrep <port>
```

- Выполнить файл *Ambari-DDL-Postgres-CREATE.sql* в PostgreSQL для завершения настройки:

```
psql -f /var/lib/ambari-server/resources/Ambari-DDL-Postgres-CREATE.sql -U ambari -p 10432 -d ambari
```

- При запросе пароля необходимо ввести значение *bigdata*.