

Instalación de Hadoop

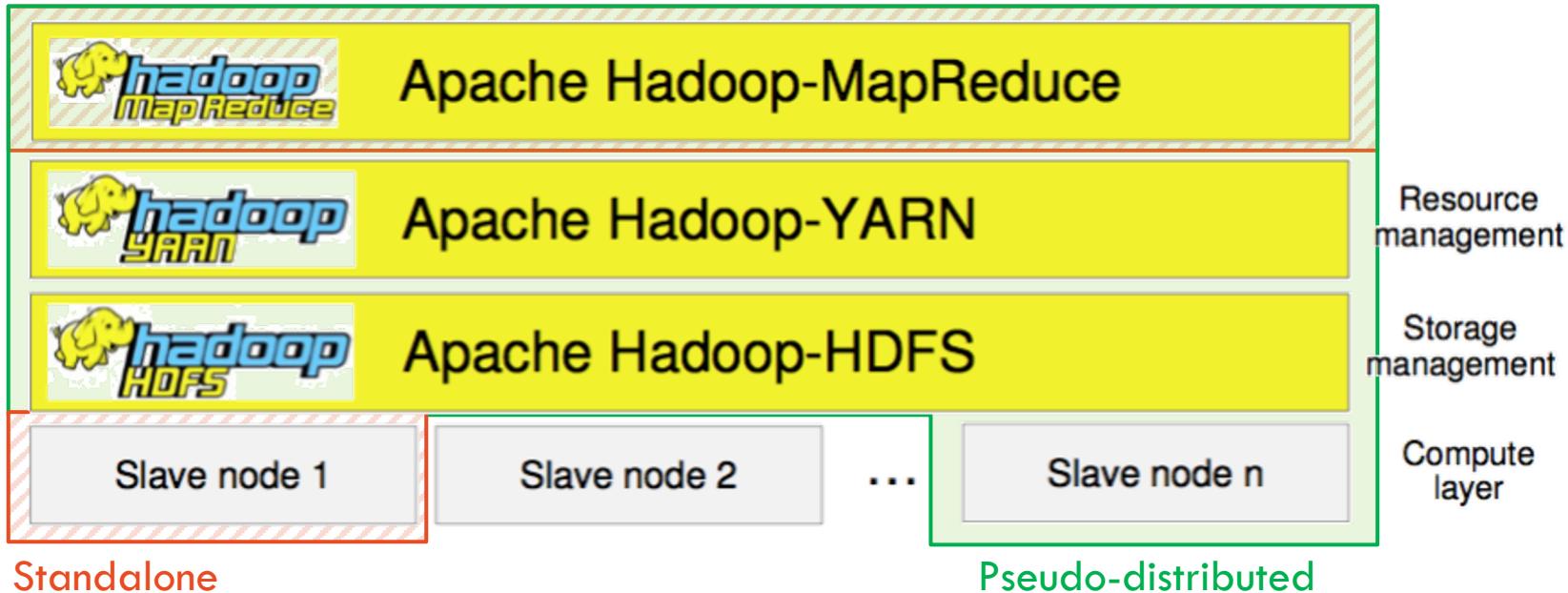
Fully distributed cluster

Instalación de Hadoop

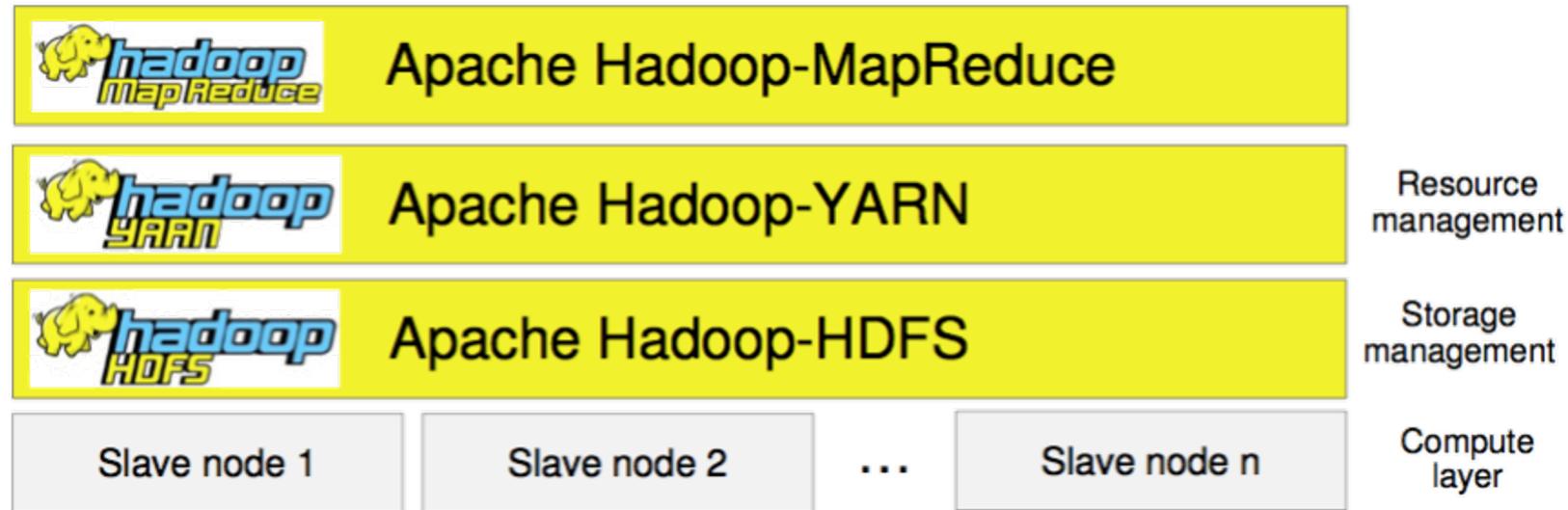
➤ Instalación *Fully Distributed*

- Los demonios de Hadoop se ejecutan en un cluster de máquinas
 - El trabajo de los demonios se puede repartir entre los nodos del cluster
- HDFS se emplea para distribuir datos entre todos los nodos
- A menos que se emplee un cluster pequeño (menos de 10 o 20 nodos), el NameNode y JobTracker deben ejecutarse en nodos dedicados
 - Para pequeños clusters pueden ejecutarse en el mismo nodo

Hasta ahora tenemos...

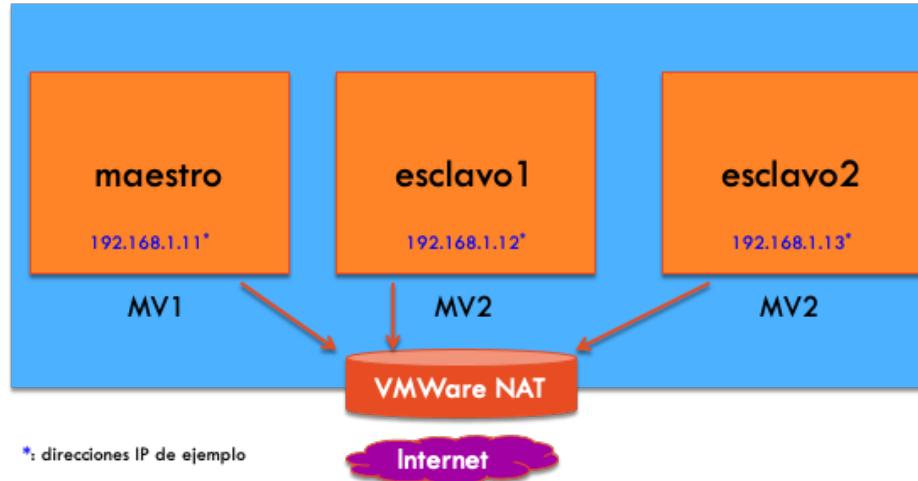


...y queremos...



¿Qué necesitamos?

- Tantos equipos como nodos vayamos a montar en el clúster
 - **En el mundo “físico”:** montaríamos un equipo físico por cada nodo
 - **En nuestro ejemplo de laboratorio:** instanciaríamos varias MV (una por nodo) dentro de nuestro equipo físico



Preparando nuestro clúster

- Hemos de asegurarnos de que **todos los nodos del clúster tienen una configuración compatible** que soporte la distribución de Hadoop
 - Podríamos instalar cada equipo por separado
 - Podríamos instalar un equipo y clonar los demás a partir de este
 - Con máquinas virtuales: copiar-pegar una MV, desplegar una imagen Vagrant...
 - Con contenedores: “docker run” de la imagen que hemos creado
 - Con equipos físicos: clonar la instalación con herramientas tipo [Clonezilla](#)

Pregunta: ¿es necesario que todos los nodos tengan exactamente la misma configuración?

Respuesta: ¡¡¡NO!!! Sólo compatible.

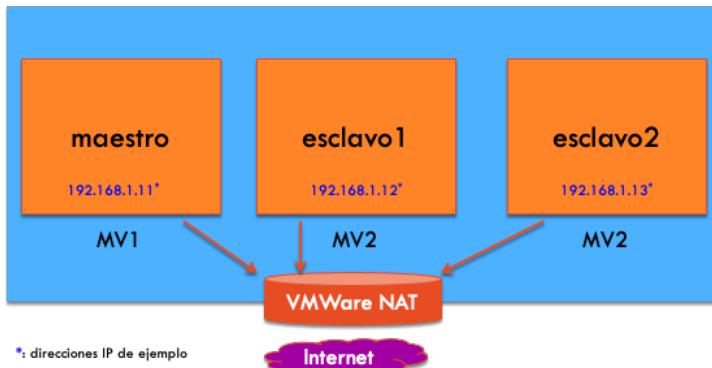


Preparando nuestro clúster

➤ Hemos de asegurarnos de que nuestros nodos...

1. Pueden "verse" unos a otros a través de la red Comunicaciones...

MVs en el mismo equipo físico:



Otra configuración:

Debemos configurar apropiadamente las:

- Direccionamiento de la subred que utilizemos
- Direcciones IP de cada nodo
- Máscaras de red
- Puerta de enlace (Gateway)

Preparando nuestro clúster

- Hemos de asegurarnos de que nuestros nodos...
 2. Se reconocen por su nombre (de red)...
 - La opción formal sería crear un servicio DNS para resolverlo
 - Podemos forzarlo editando ficheros de configuración:
 - Editamos el fichero /etc/sysconfig/network
 - NETWORKING=yes
 - HOSTNAME=maestro
 - Editamos el fichero /etc/hosts
 - 127.0.0.1 localhost
 - ::1 localhost
 - A.B.C.**11** maestro
 - A.B.C.**12** esclavo1
 - A.B.C.**13** esclavo2

Preparando nuestro clúster

➤ Hemos de asegurarnos de que nuestros nodos...

3. Están sincronizados con la misma fuente de tiempo...

- Necesitamos un servidor NTP con el que sincronizarse...
- E indicarlo en configuración del demonio NTP
 - Editamos el fichero /etc/ntp.conf
 - Añadimos línea: **server hora.uam.es**
- Para luego activar dicho demonio
 - Ejecutamos:
 - chkconfig ntpd on
 - service ntpd start

Pregunta: ¿por qué es VITAL que los nodos estén bien sincronizados?

Respuesta: Timeouts de inactividad, keepalives,...

Preparando nuestro clúster

- Hemos de asegurarnos de que nuestros nodos...
 - 4. Tenemos un espacio dedicado al HDFS en todos (o casi todos) los nodos
 - ¿Vamos a dedicar un disco en cada nodo (opción común)?
 - Dar formato al FS en el disco
 - Asegurarse de que el disco se monta en el arranque (fichero /etc/fstab)
 - ¿Sólo un espacio en uno de los discos?

Preparando nuestro clúster

- Hemos de asegurarnos de que nuestros nodos...
 - 5. Configuración de accesibilidad de los puertos que utiliza cada servicio...
 - Consultar el listado de puertos usado en cada servicio
 - En entornos seguros: puede llegarse a desactivar el firewall
 - `chkconfig firewalld off`

Preparando nuestro clúster

- Hemos de asegurarnos de que nuestros nodos...
 - 6. Cumplen los requisitos adicionales que nos pida el gestor de cluster que vamos a utilizar...
- Ejemplo: Cloudera Manager o Ambari
 - Desactivar también “SELINUX”, unas extensiones de seguridad de Linux
 - Editamos el fichero /etc/selinux/config
 - Comentamos la línea: `SELINUX=enforcing`
 - Añadimos la línea: `SELINUX=disabled`
 - Editamos la configuración del sistema: /etc/sysctl.conf
 - Añadimos líneas:
`vm.swappiness = 0`
`net.ipv6.conf.all.disable_ipv6 = 1`
`net.ipv6.conf.default.disable_ipv6 = 1`

Recomendación de cloudera.
Evita que el SO haga swap

Desactivamos el uso del
protocolo de red IPv6

Preparando nuestro clúster

- Hemos de asegurarnos de que nuestros nodos...
 - 7. Que pueden intercambiar información sin autenticación por contraseña porque se han compartido las claves
 - Desde el nodo maestro del cluster (como usuario **root**), creamos y distribuimos una clave pública para interconectarnos
 - ssh-keygen
 - ssh-copy-id esclavo1
 - ssh-copy-id esclavo2

Preparando nuestro clúster

- Si hemos llegado hasta aquí tenemos N máquinas que se ven entre sí, están sincronizadas y “preparadas”...
- ...pero todavía no hemos desplegado nada...



Existen diversos proyectos para el **despliegue, gestión y monitorización** de un cluster Big Data:



Apache Ambari

cloudera manager



Despliegue del clúster hadoop (ejemplo Cloudera Manager)

Cloudera Manager

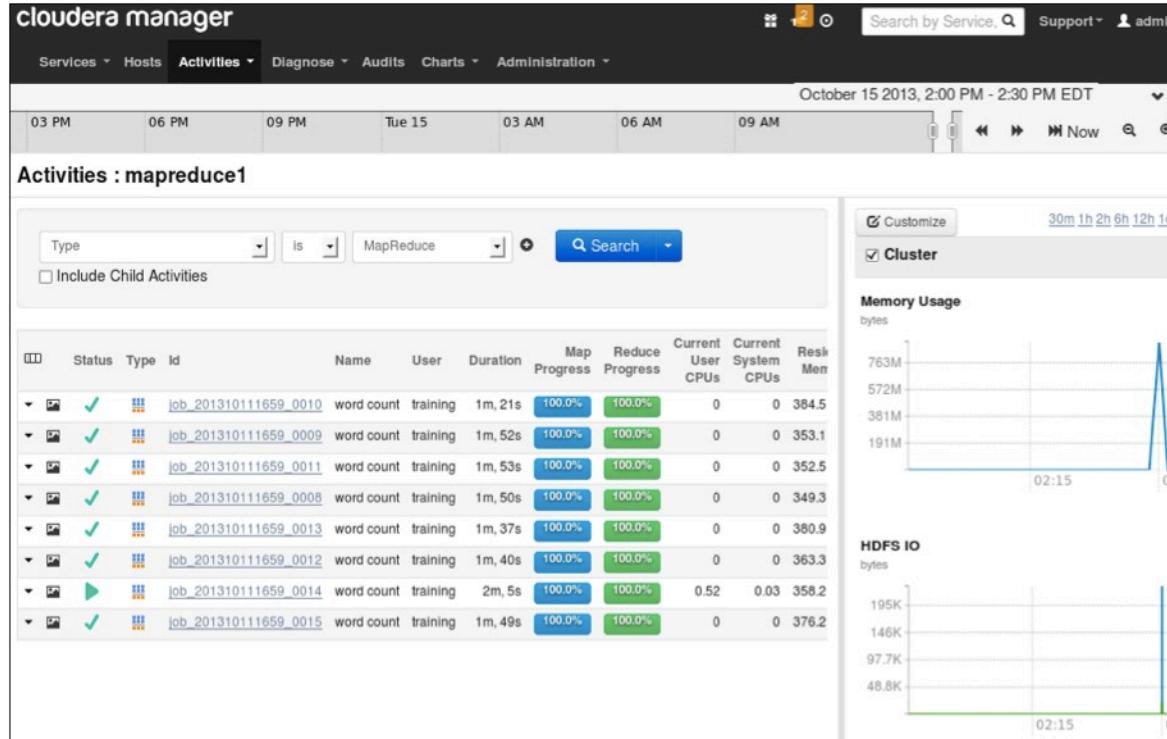
- ¿Qué es Cloudera Manager?
 - Aplicación diseñada para responder a las necesidades de usuarios profesionales de Hadoop
 - Instalar software de Hadoop en los nodos
 - Planificar y configurar servicios en un cluster
 - Monitorización de la actividad del cluster
 - Generación de informes de utilización del cluster
 - Gestionar usuarios y grupos con acceso al cluster

Cloudera Manager

- Despliegues automáticos
 - Instalación y configuración automática de servicios en los nodos del cluster
 - Permite la modificación de parámetros de configuración
 - Recomendaciones de mejores prácticas
 - Interfaz para parar/lanzar instancias de elementos del cluster
 - Punto único para obtener los ficheros de configuración que máquinas externas puedan necesitar para acceder al uso del cluster
 - Acceso y recopilación de logs de los diversos componentes del cluster

Cloudera Manager

➤ Monitorizando tareas en el cluster



Comienza la instalación del cluster BigData

Installation Phases

Phase 1: Install JDK

JDK required by the Server, Management Service, and CDH services is installed.

Cloudera Manager Installer

Installs supported versions of the Oracle JDK in /usr/java.

JDK

Install the same version of the supported versions of the Oracle JDK on each host and set the `JAVA_HOME` environment variable to the install directory.

Legend

- Interactive
- Command-Line

Phase 2: Set up DBs

Databases required by the Server, Management Service, and optional for some CDH services are installed, configured, and running.

Cloudera Manager Installer

Installs and configures embedded PostgreSQL packages and starts embedded database.

Embedded PostgreSQL Database

```
yum install cloudera-manager-server-db-2  
service cloudera-manager-server-db start  
Installs a PostgreSQL daemon on port  
7432 in  
/var/lib/cloudera-scm-server-db.
```

External Database

Install and start PostgreSQL, MySQL, or Oracle and create required databases.

Descargando el asistente

SÓLO EN LOS NODOS MAESTROS

- Descargamos la imagen del instalador de Cloudera
 - wget <http://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin>
 - chmod u+x cloudera-manager-installer.bin
- Ejecutamos el instalador del manager de Cloudera
 - ./cloudera-manager-installer.bin
 - Tras terminar la instalación, nos recomendará abrir un navegador
 - Si lo abrimos desde la MV del maestro: <http://localhost:7180>
 - Desde otra MV: <http://maestro:7180>

Estado actual de la instalación

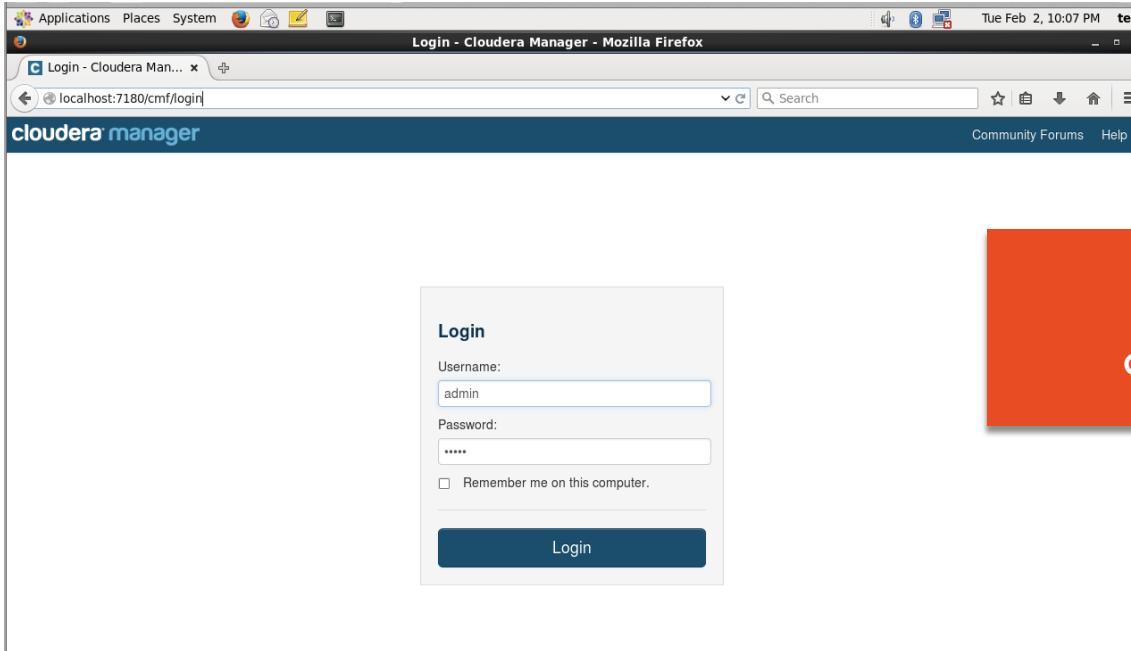


Comienza la instalación del cluster BigData

Installation Paths	A	B	C
Phase 3: Install Server Cloudera Manager Server installed and running on one host.	Cloudera Manager Installer Installs latest Cloudera Manager Server packages and Server. Requires Internet access and sudo access to Server host.	Package yum install cloudera-manager-server cloudera-manager-daemons vi /etc/cloudera-scm-server/db.properties service cloudera-manager-server start	Tarball tar xzf cloudera-manager*.tar.gz -C /opt/cloudera-manager service cloudera-manager-server start
Phase 4: Install Agents Cloudera Manager Agents installed and running on every host.	Cloudera Manager Installation Wizard Installs Cloudera Manager Agent package. Requires SSH credentials (password or key) for root or sudo-enabled user.	Package yum install cloudera-manager-agent cloudera-manager-daemons vi config.ini service cloudera-manager-agent start	Tarball vi config.ini service cloudera-manager-agent start
Phase 5: Install CDH and Managed Service SW CDH and managed service software installed on every host.	Cloudera Manager Installation Wizard Installs choice of CDH and managed service version and repo. Installs parcels or packages.	Parcel Remote or local repo or manual unpacking. API or UI.	Package yum install hadoop zookeeper hue oozie ...
Phase 6: Create, Configure, and Start CDH and Managed Services CDH and managed services configured and running.	Cloudera Manager Installation Wizard Creates, configures, and starts selected services, allows assignment of roles to hosts, and setting configuration properties. Auto-configures many options.	API POST /api/<version>/cm/deployment Best for scripting pre-configured deployments.	

Desplegando el cluster

- Interfaz gráfica a través de un navegador web
 - <http://maestro:7180>



Por defecto
admin(admin)

Desplegando el cluster

Welcome to Cloudera Manager - Cloudera Manager - Mozilla Firefox

localhost:7180/cmflicense/wizard?returnUrl=%2Fcmf%2Fexpress-wizard%2Fwelcome

End User License Terms and Conditions

Cloudera Standard License
Version 2015-08-06

END USER LICENSE TERMS AND CONDITIONS

THESE TERMS AND CONDITIONS (THESE "TERMS") APPLY TO YOUR USE OF THE PRODUCTS (AS DEFINED BELOW) PROVIDED PLEASE READ THESE TERMS CAREFULLY.

IF YOU ("YOU" OR "CUSTOMER") PLAN TO USE ANY OF THE PRODUCTS ON BEHALF OF A COMPANY OR OTHER ENTITY, YOU I EMPLOYEE OR AGENT OF SUCH COMPANY (OR OTHER ENTITY) AND YOU HAVE THE AUTHORITY TO ACCEPT ALL OF THE TEF ACCEPTED REQUEST (AS DEFINED BELOW) AND THESE TERMS (COLLECTIVELY, THE "AGREEMENT") ON BEHALF OF SUCH CO

BY USING ANY OF THE PRODUCTS, YOU ACKNOWLEDGE AND AGREE THAT:

(A) YOU HAVE READ ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT;
(B) YOU UNDERSTAND ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT;
(C) YOU AGREE TO BE LEGALLY BOUND BY ALL OF THE TERMS AND CONDITIONS SET FORTH IN THIS AGREEMENT

IF YOU DO NOT AGREE WITH ANY OF THE TERMS OR CONDITIONS OF THESE TERMS, YOU MAY NOT USE ANY PORTION OF THE PRODUCTS.

Yes, I accept the End User License Terms and Conditions.
If your download and use of Cloudera Manager are on behalf of a company that has an existing agreement with Cloudera for the use of the software, your action does not modify that existing agreement.

Back Continue

Read localhost

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 Cloudera Express . Your cluster and your data will remain unaffected.	Annual Subscription Upload License Select License File Upload
Node Limit	Unlimited	Unlimited	Unlimited
CDH	✓	✓	✓
Core Cloudera Manager Features	✓	✓	✓

Back Continue

Elegimos Cloudera Express

Deplegando el cluster

- Indicamos la lista de nodos que componen nuestro cluster
- Se verifican

The screenshot shows two browser windows side-by-side. Both windows have the URL `localhost:7180/cmf/express-wizard/hosts` and the title "cloudera manager".

Left Window (Smaller View):

Title Bar: cloudera manager

Content: Specify hosts for your CDH cluster installation.

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

Table:

Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
<input checked="" type="checkbox"/>	hadoop-master	10.11.12.100	No	✓ Host ready: 2 ms response time.
<input checked="" type="checkbox"/>	hadoop-slave1	10.11.12.101	No	✓ Host ready: 0 ms response time.

Buttons: Back, Continue

Right Window (Larger View):

Title Bar: Specify hosts for you... localhost:7180/cmf/express-wizard/hosts cloudera manager

Content: Specify hosts for your CDH cluster installation.

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

Table:

Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
<input checked="" type="checkbox"/>	hadoop-master	10.11.12.100	No	✓ Host ready: 2 ms response time.
<input checked="" type="checkbox"/>	hadoop-slave1	10.11.12.101	No	✓ Host ready: 0 ms response time.

Buttons: New Search, Back, Continue

Desplegando el cluster

- Se instalan los servicios de CM en todos los nodos

Seleccionar NO instalar JDK (ya lo tenemos en nuestras máquinas)

Cluster Installation - ... ×

localhost:7180/cmf/express-wizard/wizard#step=installStep

cloudera manager

Cluster Installation

Installation in progress.

0 of 2 host(s) completed successfully. Abort Installation

Hostname	IP Address	Progress	Status
hadoop-master	10.11.12.100	<div style="width: 50%;">50%</div>	Detecting Cloud
hadoop-slave1	10.11.12.101	<div style="width: 10%;">10%</div>	Copying installat

1 2 3 4 5 6 7 8

Back

Continuar

cloudera manager

Asistencia técnica ▾

Instalación de clúster

La instalación se ha realizado correctamente.

5 de 5 host(s) completados correctamente.

Nombre de host	Dirección IP	Progreso	Estado
nodo-principal	150.244.64.82	<div style="width: 100%;">100%</div>	La instalación se ha realizado correctamente. Detalles
nodo1	150.244.64.81	<div style="width: 100%;">100%</div>	La instalación se ha realizado correctamente. Detalles
nodo2	150.244.64.80	<div style="width: 100%;">100%</div>	La instalación se ha realizado correctamente. Detalles
nodo3	150.244.64.78	<div style="width: 100%;">100%</div>	La instalación se ha realizado correctamente. Detalles
nodo4	150.244.64.79	<div style="width: 100%;">100%</div>	La instalación se ha realizado correctamente. Detalles

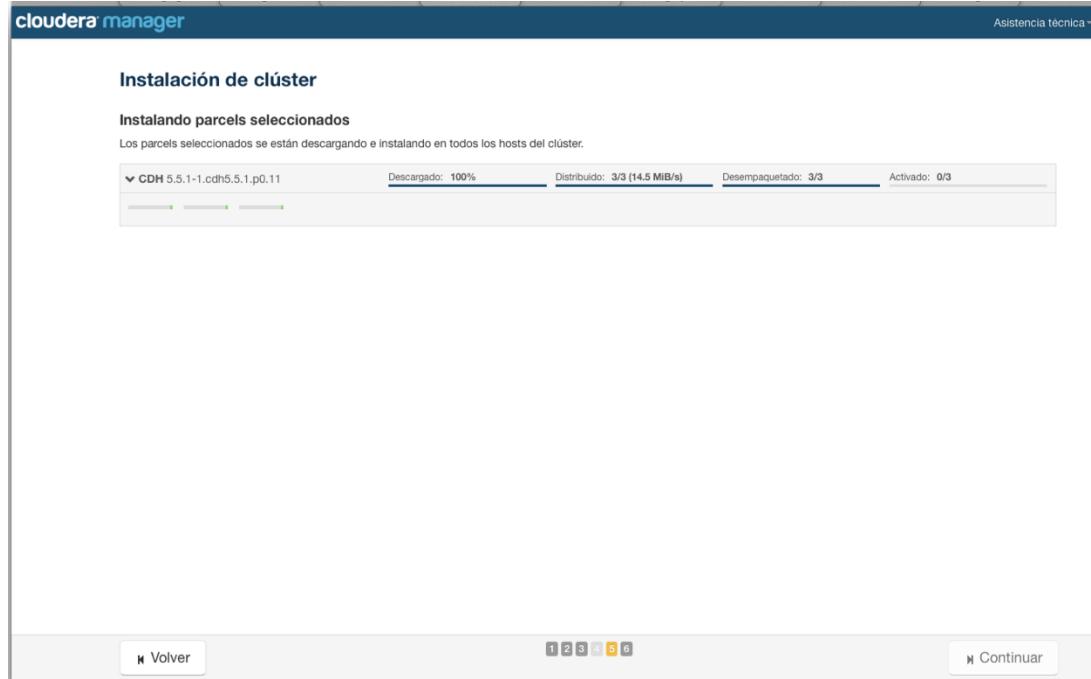
1 2 3 4 5 6 7 8

Volver

Continuar

Despliegue del cluster

- Se descargan y despliegan los parcels
 - Distribuciones preconfiguradas



Despliegue del cluster

- Se nos da a elegir los componentes de software a instalar
- Dentro de los componentes existentes en el parcel (CDH) instalado

Configuración de clúster

Seleccione los servicios CDH 5 que desea instalar en el clúster.

- Escoger una combinación de servicios para instalar
- Hadoop centrales**
HDFS, YARN (MapReduce 2 incluido), ZooKeeper, Oozie, Hive, Hue y Sqoop
 - Núcleo con HBase**
HDFS, YARN (MapReduce 2 incluido), ZooKeeper, Oozie, Hive, Hue, Sqoop y HBase
 - Núcleo con Impala**
HDFS, YARN (MapReduce 2 incluido), ZooKeeper, Oozie, Hive, Hue, Sqoop e Impala
 - Núcleo con Search**
HDFS, YARN (MapReduce 2 incluido), ZooKeeper, Oozie, Hive, Sqoop y Solr
 - Núcleo con Spark**
HDFS, YARN (MapReduce 2 incluido), ZooKeeper, Oozie, Hive, Hue, Sqoop y Spark
 - Todos los servicios**
HDFS, YARN (MapReduce 2 incluido), ZooKeeper, Oozie, Hive, Hue, Sqoop, HBase, Impala, Solr, Spark y Key-Value Store Indexer
 - Servicios personalizados**
Seleccione sus propios servicios. Los servicios requeridos por los servicios seleccionados se incluirán automáticamente. Se puede agregar el servicio Flume después de configurar el clúster inicial.

Este asistente también instalará **Cloudera Management Service**. Se trata de un conjunto de componentes que activa la supervisión, la comunicación, los eventos y las alertas; estos componentes requieren bases de datos para almacenar información que podrá configurar en la siguiente página.

« Volver

1 2 3 4 5 6

» Continuar

Despliegue del cluster

Configuración de clúster

⌘ Primera ejecución Comando

Estado: **Running** Hora de inicio: feb. 6, 3:10:31 AM

Abortar

Detalles

Completed 4 of 7 step(s).

Todo Solo erróneos Running Only

Step	Contexto	Hora de inicio	Duración	Acciones
➤ ✓ Implementar la configuración de cliente Successfully deployed all client configurations.	Cluster 1	feb. 6, 3:10:31 AM	15.5s	
➤ ✓ Realizando la Primera ejecución para Cloudera Management Service, ZooKeeper Se han completado 2 pasos correctamente.		feb. 6, 3:10:46 AM	26.12s	
➤ ✓ Iniciando servicio HDFS Se han completado 1 pasos correctamente.		feb. 6, 3:11:12 AM	38.36s	
➤ ✓ Iniciando servicio YARN (MR2 Included) Se han completado 1 pasos correctamente.		feb. 6, 3:11:51 AM	64.43s	
➤ ⚡ Iniciando servicio Hive Solo se han completado los pasos 0/1. Esperando el resto.		feb. 6, 3:12:55 AM		
➤ Iniciando servicio Oozie				
➤ Iniciando servicio Hue				

« Volver

1 2 3 4 5 6

Continuar »

Despliegue del cluster

cloudera manager

Asistencia técnica ▾ admin ▾

Configuración de clúster

¡Enhorabuena!

Los servicios están instalados, configurados y ejecutándose en su clúster.

1 2 3 4 5 6

Volver Finalizar



Estado actual de la instalación



Utilizando el clúster

Nuestro cluster operativo

cloudera manager Clústeres Hosts Diagnóstico Auditorias Gráficos Administración Buscar (Hotkey: /) Asistencia técnica admin

Inicio

Estado Todos los problemas de estado 6 Configuración 4 Todos los comandos recientes Agregar clúster

30 minutos anterior 6 de febrero de 2016, 3:22 CET

Pruebe Cloudera Enterprise Data Hub Edition durante 60 días

Cluster 1 (CDH 5.5.1, Remesas)

Hosts	1	1
HDFS	1	1
Hive		
Hue		
Oozie		
YARN (MR2 Incl...)	1	
ZooKeeper		1

Cloudera Management Service

Cloudera...	2	2
-------------	---	---

Gráficos

CPU de clúster

percent
0% 50% 100%
03 AM 03:15
Cluster 1, Uso de la CPU del host en Hosts 2%

IO de disco del clúster

bytes / second
0 9.5M/s 19.1M/s
03 AM 03:15
Total de Bytes... 152K/s Total de Byte... 25.5K/s

IO de red del clúster

bytes / second
0 4.8M/s 9.5M/s 14.3M/s
03 AM 03:15
Total de Bytes... 8.5K/s Total de Bytes... 3.5K/s

E/S de HDFS

bytes / second
0 4.8M/s 9.5M/s
03 AM 03:15
Total de Bytes... 2.8b/s Total de Bytes le... 1b/s

30m 1h 2h 6h 12h 1d 7d 30d

Hue

- Hadoop User Experience
- Agrega en un **único interfaz** los componentes más comunes de hadoop
- Acceso: **http://<nombre-nodo>:8888**



No tiene por qué ser el nodo principal. Se elige al instalar.



Hue

The screenshot shows the Hue web interface with several menu items highlighted by red boxes:

- Query Editors** (highlighted)
- Workflows** (highlighted)
- Explorador de archivos** (highlighted)
- Job Browser** (highlighted)

The main content area displays the "Asistente de configuración" (Configuration Wizard) for Hadoop. Step 1: Comprobar configuración is completed successfully.

Comprobando la configuración actual

Archivos de configuración ubicados en `/var/run/cloudera-scm-agent/process/38-hue-HUE_SERVER`

Todo correcto. Comprobación de configuración aprobada.

Volver **Siguiente**

Hue y el logotipo de Hue son marcas comerciales de Cloudera, Inc.

Probando el cluster

- Ejecutar aplicación de ejemplo para asegurarnos de que el clúster está correctamente instalado
- Pista:
 - /opt/cloudera/parcels/CDH/
 - Directorio *bin*
 - *hadoop*
 - Directorio *jars*
 - *hadoop-examples.jar wordcount ...*
 - *hadoop-streaming ...*

Instalación de componentes

- Cloudera Manager se encarga de aislarnos de todos los detalles de instalación de cada uno de los servicios/demonios que componen el cluster
- /opt/cloudera/parcels/CDH/lib/hadoop/etc/hadoop/**hdfs-site.xml**

```
<?xml version='1.0' encoding='UTF-8'?>
<!--Autogenerated by Cloudera Manager-->
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:///hadoop-hdfs/dfs/nn</value>
  </property>
  <property>
    <name>dfs.namenode.servicerpc-address</name>
    <value>cluster1bigdata.ii.uam.es:8022</value>
  </property>
  <property>
    <name>dfs.https.address</name>
    <value>cluster1bigdata.ii.uam.es:50470</value>
  </property>
  <property>
    <name>dfs.https.port</name>
    <value>50470</value>
  </property>
  <property>
    <name>dfs.namenode.http-address</name>
    <value>cluster1bigdata.ii.uam.es:50070</value>
  </property>
  <property>
    <name>dfs.replication</name>
    <value>3</value>
  </property>
  <property>
    <name>dfs.blocksize</name>
    <value>134217728</value>
  </property>
  <property>
    <name>dfs.client.use.datanode.hostname</name>
    <value>false</value>
  </property>
```

Lecciones aprendidas (de las dos instalaciones)

- Hadoop puede ejecutarse de tres modos
 - Local
 - Pseudo-distributed
 - Distributed
- Ficheros de configuración
 - Pares claves – valor
 - /etc/hadoop
- Los demonios de hadoop generan logs que podemos consultar en caso de que las cosas empiecen a fallar
 - /var/log

Información adicional

➤ Más información en:

- Guía completa de instalación

http://www.cloudera.com/documentation/enterprise/5-3-x/topics/cm_ig_intro_to_cm_install.html

- Configuración de puertos (si se activa un firewall)

http://www.cloudera.com/documentation/archive/manager/4-x/4-5-4/Configuring-Ports-for-Cloudera-Manager-Enterprise-Edition/cmeecp_topic_4.html

- Resumen gráfico:

http://www.cloudera.com/documentation/enterprise/5-3-x/topics/cm_ig_intro_to_cm_install.html

- Quick Start Guide:

http://www.cloudera.com/documentation/enterprise/latest/topics/cm_qs_quick_start.html

Despliegue en Cloud (ejemplo AWS)

Ejemplo: crear clúster y lanzar tarea (Amazon EMR)

- Amazon Elastic MapReduce (EMR)
 - Cloud **público** de Amazon
 - Caso particular Amazon EC2
 - Creación y gestión de clústers para cómputo Big Data
 - Instalaciones predefinidas
 - Elección del HW apropiado entre configuraciones existentes
 - Aplicación intensiva en memoria, CPU, ...
 - Elección de componentes HW
 - Versiones de Hadoop, Spark, ... soportadas

Ejemplo: crear clúster y lanzar tarea (Amazon EMR)

Create Cluster - Advanced Options

[Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release ✓

- Hadoop
- Flume
- Pig
- Zookeeper
- Hue
- Spark

Edit software

Entire cluster

{"classifi

Add step

EMR Release - 5.x

- emr-5.8.0
- emr-5.7.0
- emr-5.6.0
- emr-5.5.0
- emr-5.4.0
- emr-5.3.1
- emr-5.3.0
- emr-5.2.2
- emr-5.2.1
- emr-5.2.0
- emr-5.1.0
- emr-5.0.3
- emr-5.0.0

EMR Release - 4.x

- emr-4.9.2
- emr-4.9.1
- emr-4.8.4
- emr-4.8.3
- emr-4.8.2
- emr-4.8.0
- emr-4.7.2
- emr-4.7.1
- emr-4.7.0
- emr-4.6.0
- emr-4.5.0
- emr-4.4.0

Ejemplo: crear clúster y lanzar tarea (Amazon EMR)

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release emr-5.8.0

- | | | |
|--|--|--|
| <input checked="" type="checkbox"/> Hadoop 2.7.3 | <input checked="" type="checkbox"/> Zeppelin 0.7.2 | <input type="checkbox"/> Tez 0.8.4 |
| <input type="checkbox"/> Flink 1.3.1 | <input checked="" type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.3.1 |
| <input type="checkbox"/> Pig 0.16.0 | <input type="checkbox"/> Hive 2.3.0 | <input type="checkbox"/> Presto 0.170 |
| <input type="checkbox"/> ZooKeeper 3.4.10 | <input type="checkbox"/> Sqoop 1.4.6 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 3.12.0 | <input type="checkbox"/> Phoenix 4.11.0 | <input type="checkbox"/> Oozie 4.3.0 |
| <input checked="" type="checkbox"/> Spark 2.2.0 | <input type="checkbox"/> HCatalog 2.3.0 | |

Edit software settings (optional) i

Enter configuration

Load JSON from S3

```
[{"classification": "spark", "properties": {"maximizeResourceAllocation": "true"}]}
```

Add steps (optional) i

Step type

Auto-terminate cluster after the last step is completed

[Cancel](#)

[Next](#)

Ejemplo: crear clúster y lanzar tarea (Amazon EMR)

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration ⓘ

Instance types

Instance type	vCPU	Memory (GB)	Storage (GiB)
c3.xlarge	4	7.5	80 SSD
c3.2xlarge	8	15	160 SSD
c3.4xlarge	16	30	320 SSD
c3.8xlarge	32	60	640 SSD
c4.large	2	3.8	EBS only
c4.xlarge	4	7.5	EBS only
c4.2xlarge	8	15	EBS only
c4.4xlarge	16	30	EBS only
c4.8xlarge	36	60	EBS only
d2.xlarge	8	30.5	6144 SSD
d2.2xlarge	16	61	12288 SSD

option
and
m bid price: \$ 0.05 ⓘ
and
m bid price: \$ 0.05 ⓘ
and
m bid price: \$ 0.05 ⓘ

[Cancel](#) [Save](#)

Ejemplo: crear clúster y lanzar tarea (Amazon EMR)

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration

If you need more than 20 EC2 instances, [see this topic.](#)

Instance group configuration

Uniform instance groups

Specify a single instance type and purchasing option for each node type.

Instance fleets

Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#)

Network

vpc-dd6b4ab5 (172.31.0.0/16) (default)

[Create a VPC](#) 

EC2 Subnet

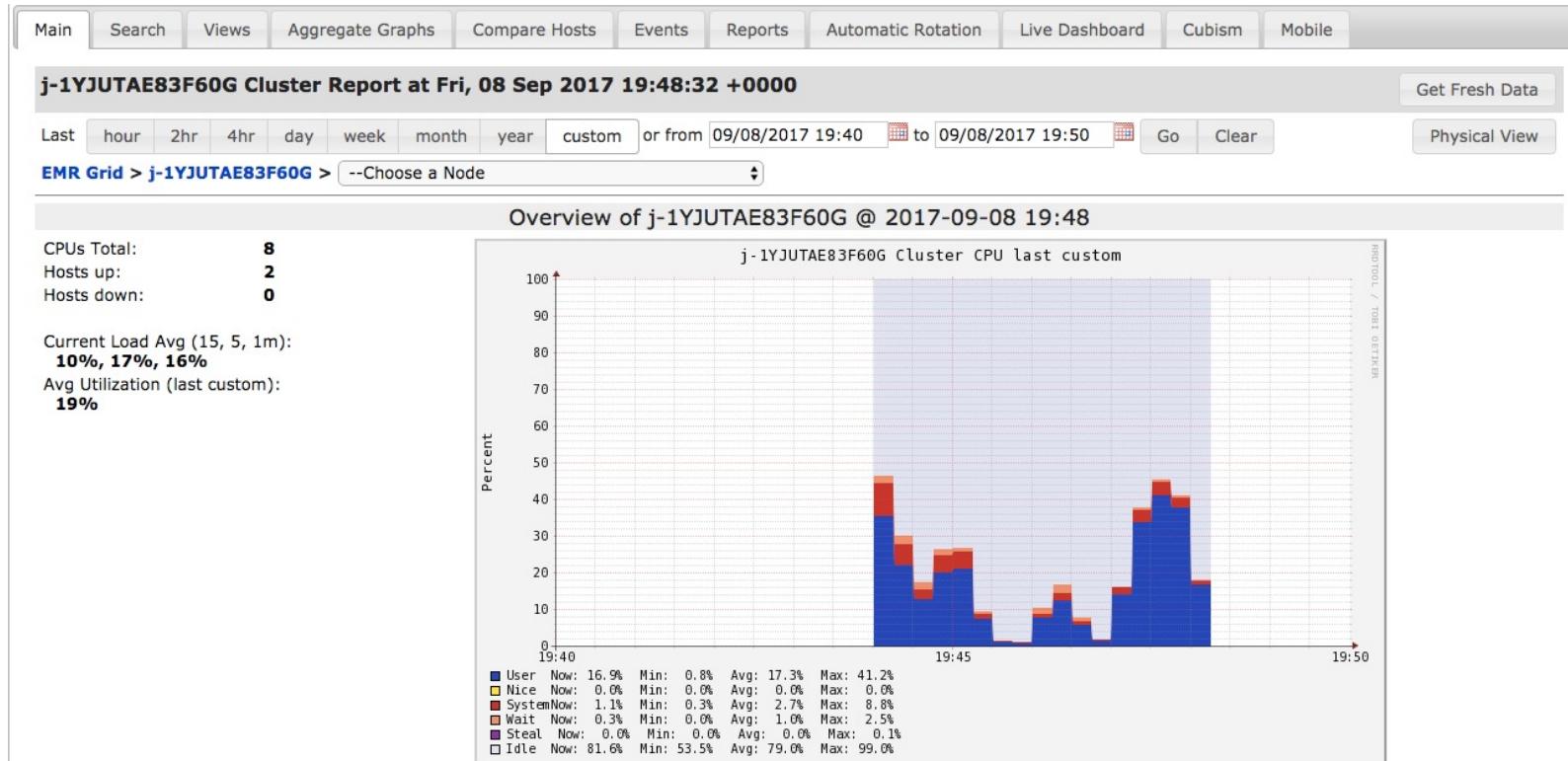
subnet-33fed449 | Default in eu-central-1b

Root device EBS volume size

10 GiB 

Node type	Instance type	Instance count	Purchasing option
Master	m3.xlarge 	1 Instances	<input type="radio"/> On-demand <input checked="" type="radio"/> Spot Maximum bid price: \$ 0.05 
Master Instance Group	8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none 		
Core	m3.xlarge 	1 Instances	<input type="radio"/> On-demand <input checked="" type="radio"/> Spot Maximum bid price: \$ 0.05 
Core Instance Group	8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none 		

Ejemplo: crear clúster y lanzar tarea (Amazon EMR)



Ejemplo: crear clúster y lanzar tarea (Amazon EMR)

The screenshot shows the Hue Job Browser interface. At the top, there are search fields for 'Nombre de usuario' and 'Texto', and a filter section for job status ('Succeeded', 'Running', 'Failed', 'Killed') and time ('in the last 7 days'). Below this is a table listing a single job entry:

Logs	ID	Name	Application Type	Status	Usuario	Maps	Reduces	Queue	Priority	Duration	Submitted	Action
	1504899669944_0001	Zeppelin	SPARK	RUNNING	zeppelin	10%	10%	default	N/A	2m:32s	09/08/17 12:47:14	Kill

At the bottom, it says 'Showing 1 to 1 of 1 entries' and has navigation buttons for 'Previous', '1', and 'Next'.

Ejemplo: crear clúster y lanzar tarea (Amazon EMR)

- Al crear un clúster en EMR
 - Se eligen los componentes SW
 - Se elige la configuración HW
 - Se indican configuraciones adicionales
 - Pinchar el botón “create”, entonces EMR:
 - Reserva y crea contenedores virtuales para componer el clúster
 - Arranca esos contenedores virtuales
 - Realiza la instalación/configuración de todos el SW
 - Más aún: redimensionar clúster, configuraciones de seguridad...

¿Preguntas?