Máster en Big Data

Tecnologías de Almacenamiento

2. Hands-On: Despliegue de un clúster Hadoop - CDH

Àlex Balló Vergés

2018-2019

Índice

1. Introducción

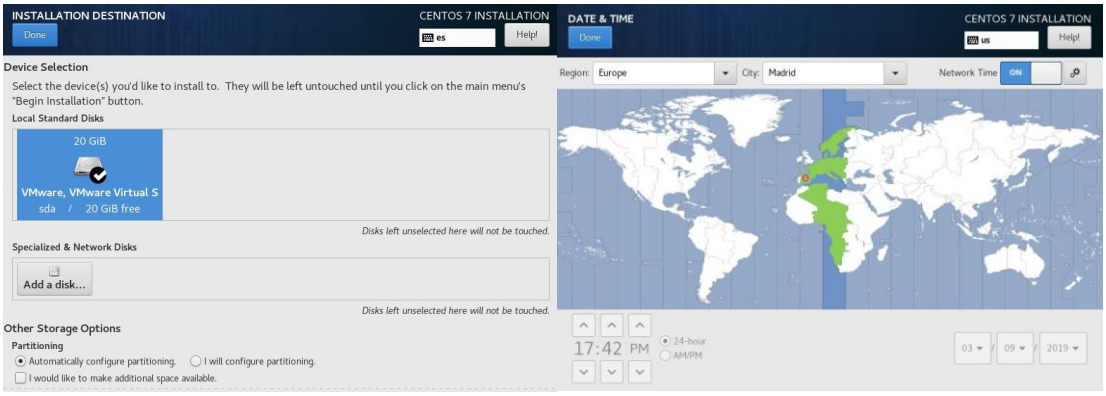
El objetivo de este Hands-On es desplegar un clúster Hadoop basado en una distribución Cloudera y contestar a las preguntas planteadas: 2.a), 5.a), 5.b), 5.c), 5.d)

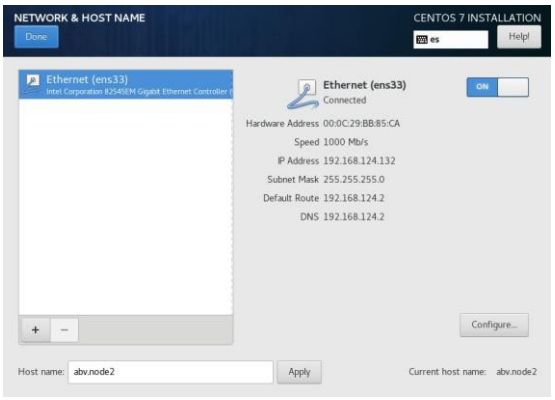
2. Despliegue del clúster

1- Instalar dos máquinas con CentOS

Utilizar el archivo CentOS-7-x86_64-Minimal-1708.iso proporcionado con este enunciado

Asignar los siguientes nombres a las máquinas:
<iniciales>.node1
<iniciales>.node2





Tips: recuerda configurar un hostname de la máquina durante la instalación y el mismo usuario de root para las dos máquinas

2- Preparar el SO para la instalación de hadoop, realiza la preparación en ambas máquinas

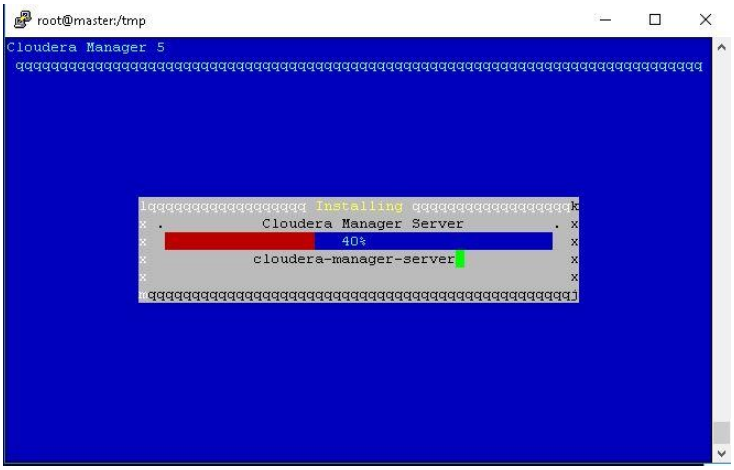
a. ¿Qué tres pasos, mínimos, son necesarios para considerar que tenemos el SO preparado?

Configurar el archivo hosts para asociar los hostnames con las ips de cada nodo, deshabilitar el selinux y deshabilitar el Firewall.

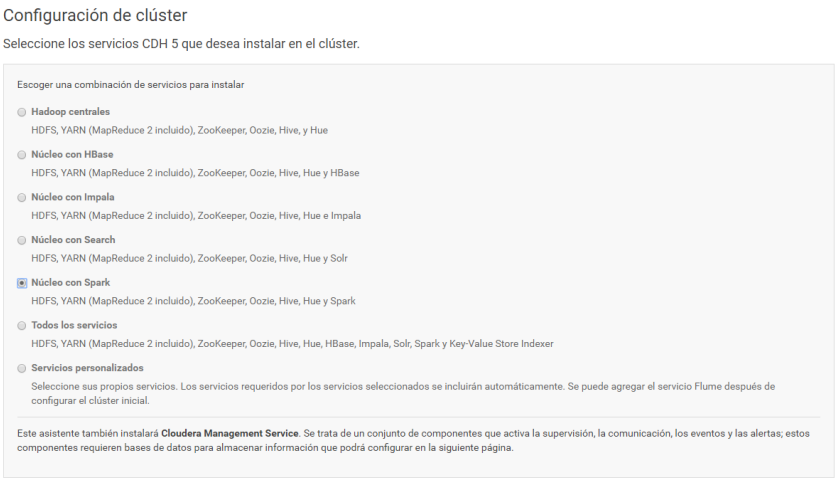
3- Descargar y ejecutar el instalador para Cloudera Manager

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

Tips: Recuerda proporcionar permisos de ejecución al instalador antes de ejecutarlo con chmod u+x Cloudera-manager-installer.bin



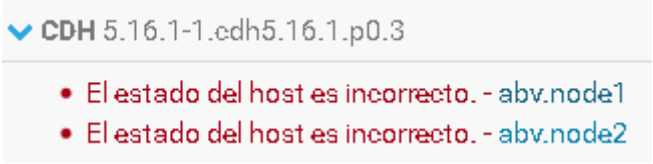
4- Realiza una instalación de Hadoop con Spark



Se ha hecho la instalación sin Spark por la falta de memoria en la máquina virtual.

(IMPORTANTE: No modifiques los roles de las máquinas, instala por defecto los que sugiere Cloudera Manager)

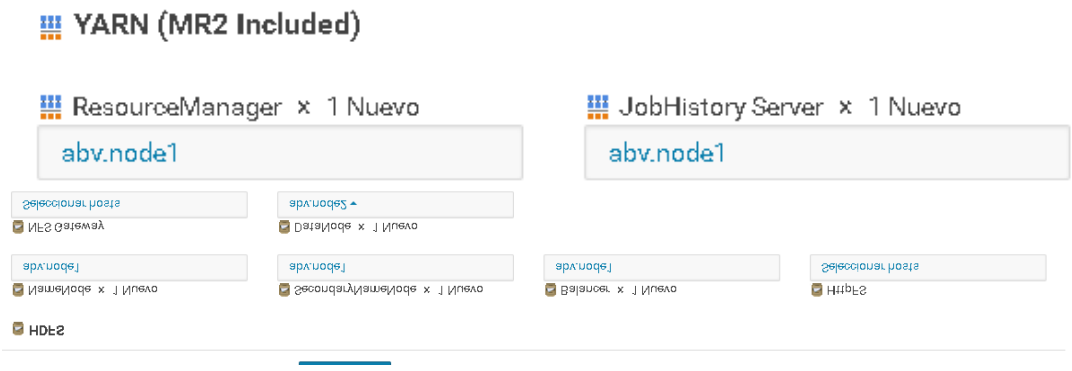
Después de fallar la primera instalación con el Yarn, aunque se reiniciará el servicio de Cloudera manager o se desinstalara el software, todo el rato aparecía un problema al distribuir los parcels en los hosts; hasta que se instalaron las máquinas virtuales de nuevo.



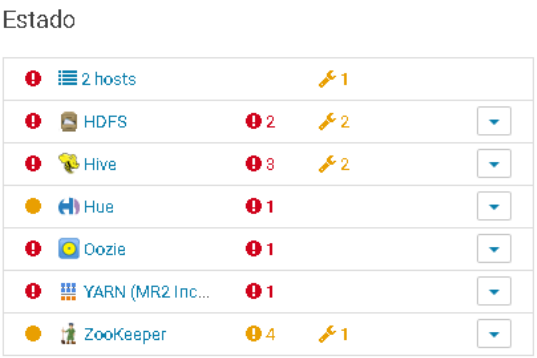
5- Una vez finalizada la instalación, realiza los siguientes apartados:

a. ¿Si tuvieses que elegir uno de los nodos, cual dirías que juega el rol de máster y cual el de esclavo? (adjunta una captura de pantalla de los roles por host que justifique tu explicación)

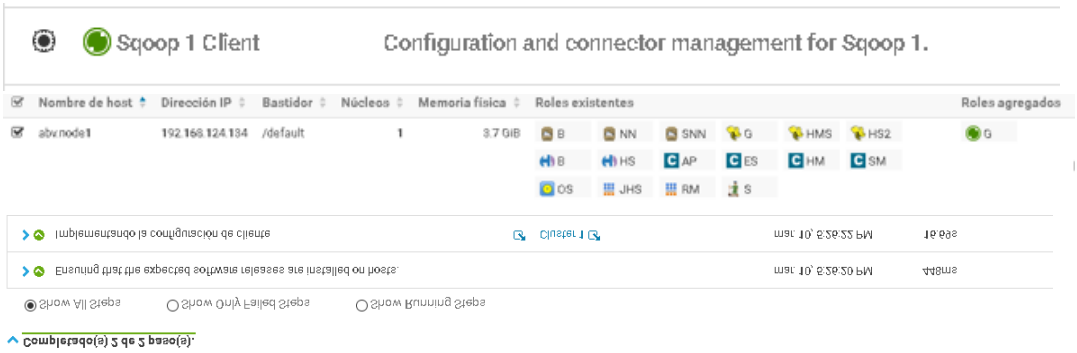
El rol de máster seria el nodo al que hemos instalado Cloudera ya que el otro solo tendría la función de procesar y almacenar la información.



Cluster instalado:



b. Instala el servicio de Sqoop para el clúster (captura de pantalla de los pasos realizados)



c. Crea una nueva cola de ejecución en YARN, **llamada** test_LaSalle con los siguientes parámetros:
   i. Mínimo de recursos: 1 núcleo / 1 GB de memoria
   ii. Máximo de recursos: 2 núcleos / 2GB
   iii. Máximas aplicaciones en ejecución: 1

Búsqueda			Mínimo de recursos		Máximo de recursos		Máximo de aplicaciones en ejecución
Nombre	Peso	%	CPU	Memoria	CPU	Memoria	
root			1 vcores	1 GiB	100 %	50 %	1
root.default	1	50.0%	1 vcores	1 GiB	100 %	50 %	1
root.users	1	50.0%	1 vcores	1 GiB	100 %	50 %	1

Como la màquina virtual tiene 4gb de memoria se ha añadido el 50% de recursos de la memoria y el 100% de los núcleos de CPU porqué solo tenemos uno por cada máquina.

```
[root@abv ~]# nproc --all
1
```

d. ¿Fíjate que existe un error en el servicio de HDFS llamado underreplication? ¿Qué quiere decir? Realiza los cambios necesarios en la configuración de HDFS para solucionarlo
   i. ¿Después de solucionarlo, seguramente la alerta no haya desaparecido, a causa de qué?