IES de Teis
Avda. de Galicia, 101
36216 – Vigo

Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es

Práctica Clúster HDFS

Objetivo.

- Demostrar capacidad para crear una red de equipos interconectados.

Fecha de entrega: 15 de marzo de 2024 hasta las 23:00h (hora del servidor)

No es necesario realizar capturas de pantalla de la preparación de la máquina virtual.

Elabora y envía un documento PDF con tu nombre completo y DNI en la portada. Usa una nueva hoja por cada apartado, copia el enunciado y contesta. Las capturas de pantalla deben ser de "pantalla completa", no solo del detalle.

Preparación.

Necesitaremos 4 equipos virtualizados con las siguientes características:

- Un disco de 50GB
- Un segundo disco de 100GB
- Un mínimo de 2 procesadores asignados.
- Sistema operativo: Ubuntu 22.04
- Usuario: hadoop
- Contraseña: BigData.,
- El disco de 100GB debe estar montado de manera permanente en la siguiente ruta "/home/hadoop/discogrande"
- Necesitaremos que algún nodo tenga interfaz gráfica para poder interactuar con HDFS a través de un navegador web.
- Cada equipo debe tener como nombre el tuyo más un número, por ejemplo, javi1, javi2, javi3 y javi4.
- Cada equipo podrá establecer una conexión ssh a cualquier otro usando su IP y por su nombre.
- Las conexiones ssh se autentificarán con llaves públicas de cada nodo.
- Todos los equipos tendrán salida a internet.
- Debes instalar en todos los nodos JAVA 8. (sudo apt install openjdk-8-jdk)
- Los nodos tendrán una IP fija manual que seguirá este patrón: IP acabada en 101 para el nodo1, IP acabada en 102 para el nodo2,…

IES de Teis
Avda. de Galicia, 101
36216 – Vigo

Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es

Enunciado.

1.- Realiza una conexión ssh de tu nodo1 a tu nodo4 usando el nombre del nodo y sin necesidad de introducir contraseña. Captura pantalla donde se vea que usas el nombre del nodo en lugar de su IP y que no has tenido que introducir contraseña.

2.- Comprueba en la web https://hadoop.apache.org/releases la ruta del paquete hadoop-3.3.6 para tu plataforma, descárgalo con wget y descomprímelo de manera que su contenido quede en la ruta "/home/hadoop/hadoop". Repite esta operación en todos los nodos. Realiza una captura de pantalla del comando "ls -la /home/hadoop/hadoop".

3.- Edita el archivo "/etc/environment" para añadir lo siguiente:

- Una línea con: JAVA_HOME:"/usr/lib/jvm/java-8-openjdk-amd64"
- Una línea con: HADOOP_HOME:"/home/hadoop/hadoop"
- Añadir al final del PATH las rutas de bin y sbin de hadoop (ojo con los dos puntos y las comillas): /home/hadoop/hadoop/bin:/home/hadoop/hadoop/sbin

Realiza una captura del comando "cat /etc/environment". Explica con tus palabras para que sirve lo que acabas de hacer en este archivo.

4.- Modifica la configuración del archivo "home/hadoop/hadoop/etc/hadoop/core-site.xml" para indicar que la propiedad "fs.defaultFS" tiene un valor de hdfs://nodo1:9000. Donde nodo1 será uno de tus nodos que has elegido para la función de NameNode. Haz lo mismo en todos los nodos. Realiza la captura de pantalla del comando "cat /home/hadoop/hadoop/etc/core-site.xml" en cualquiera de los nodos.

¿Qué estamos indicando en esa configuración? ¿Por qué todos los nodos comparten la la misma configuración en el archivo core-site.xml?

5.- En el nodo que realizará las funciones de NameNode modifica el archivo *"/home/hadoop/hadoop/etc/hadoop/hdfs-site.xml"* para incluir las siguientes propiedades:

IES de Teis
Avda. de Galicia, 101
36216 – Vigo
Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es

- Nombre: *"dfs.namenode.name.dir"* con valor "/home/hadoop/discogrande/namenode"
- Nombre: *"dfs.replication"* con valor "2"

Realiza una captura de pantalla del comando "cat /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml".

¿Qué indicamos con la primera propiedad? ¿Qué indicamos con la segunda propiedad? ¿Por qué no replicamos el contenido de este archivo en el resto de los nodos?

6.- En cada uno de los nodos que realizarán las funciones de DataNode modifica el archivo "/home/hadoop/hadoop/etc/hadoop/hdfs-site.xml" para que incluya la propiedad de "dfs.datanode.data.dir" con el valor "/home/hadoop/discogrande/datanode". Captura la pantalla del comando "cat /home/hadoop/hadoop/etc/hadoop/hdfs-site"

Explica con tus palabras la posible razón por la que el mismo archivo "hdfs-site.xml" es distinto en NameNodes y en DataNodes.

Justifica con tus palabras si consideras que sería bueno (o no) fusionar ambas versiones de "hdfs-site.xml" en una sola en la que aparezcan todas las propiedades.

7.- Formatea el sistema de ficheros HDFS en el NameNode. Captura la pantalla del comando que has usado y su salida.

Indica qué comando has usado. Indica desde qué nodo has ejecutado el comando. En este momento en el que aún no hay ningún namenode ni datanode encendido, ¿qué dirías que está haciendo este formateo?

8.- Arranca únicamente el namenode desde su nodo. Captura pantalla en la que se vea el comando que usas y su salida.

9.- Desde el namenode realiza una captura de pantalla de la salida del comando "hdfs dfsadmin -report".

Interpreta y explica lo que puedas de la salida del comando.

10.- Arranca los datanodes de uno en uno desde cada nodo. Captura pantalla en la que se vea el comando que usas y su salida en alguno de los nodos.

11.- Desde el namenode realiza una captura de pantalla de la salida del comando "hdfs dfsadmin -report".Si tienes una interfaz gráfica también puedes capturar la web en la ip del namenode:9870, apartado datanodes.

Interpreta y explica lo que puedas de la salida del comando.

12.- Realiza una captura de la salida del comando "hdfs dfs -df -h /" ejecutado en cualquiera de los nodos. Interpreta con tus palabras el resultado.

13.- Apaga un datanode y confirma que se muestra como caído. Realiza una captura de pantalla de la salida del comando donde se pueda ver esta información. Con los parámetros por defecto, Hadoop dará por caído un nodo cuando no tenga conexión durante los últimos 10 minutos y medio.

Justifica con tus palabras qué valor de tiempo especificarías en segundos como tope para dar un nodo por caído en un clúster de 100 máquinas dentro de un mismo CPD.

Justifica con tus palabras qué valor de tiempo especificarías en segundos como tope para dar un nodo por caído en un clúster de miles de máquinas repartidas en CPDs de distintos países.

14.- Cambia en el archivo "hdfs-site.xml" las siguientes propiedades:

- Nombre: "dfs.heartbeat.interval", Valor: 1
- Nombre: "dfs.namenode.heartbeat.recheck-interval", Valor: 500

Al reiniciar los nodos, el tiempo de detección de nodos caídos será de 11 segundos. Realiza una captura de pantalla donde se vea un nodo caído con pocos segundos de retraso.

15.- Copia el archivo que has descargado previamente (hadoop-3.3.6.tar.gz) a HDFS con el siguiente comando:

"hdfs dfs -copyFromLocal /home/hadoop/hadoop-3.3.6.tar.gz /".

Captura la pantalla con el comando "ls" aplicado a hdfs para listar los documentos que están en la raíz de HDFS.

16.- Desde la interfaz web del namenode en el puerto 9000 averigua en qué nodos está replicado el archivo que has subido. Realiza una captura donde se pueda ver esta información. Indica con tus palabras qué debería pasar si uno de los nodos que mantiene una réplica cae.