Pasos instalación HDFS con Hadoop

En la parte de preparación del nodo:

- Se configura una IP fija. (netplan)
- Se cambia el nombre del nodo (hostname)
- Se actualiza el archivo /etc/hosts con los nombres e IP de los nodos que componen el clúster.
- Se instala JAVA-8 y se configura la variable de entorno JAVA_HOME para que apunte a esta versión.
- Se instala openssh server, se crea un par de llaves público-privada y se intercambian entre todos los nodos del clúster.
- Preparamos el montaje del segundo disco duro en la carpeta "discogrande"

En la parte de configuración de Hadoop:

Descargamos la última versión y la descomprimimos en una ruta amigable.
Se crea una variable de entorno llamada HADOOP_HOME con la ruta a la carpeta descomprimida.
Actualizamos el PATH para incluir las carpetas /bin y /sbin de Hadoop.
Configuramos el archivo core-site.xml con los datos del nodo "maestro" y los copiamos al resto de nodos.

- fs.defaultFS
  El nombre del sistema de archivos predeterminado. Un URL cuyo esquema y autoridad determinan la implementación del sistema de archivos. El esquema del uri determina la propiedad config (fs. SCHEME.impl) nombrando la clase de implementación del sistema de archivos. La autoridad del uri se utiliza para determinar el host, el puerto, etc. de un sistema de archivos.Configuramos el archivo hdfs-site.xml en el namenode.
- dfs.namenode.name.dir
  Determina en qué parte del sistema de archivos local el nodo de nombre DFS debe almacenar el nombre table(fsimage). Si se trata de una lista de directorios delimitada por comas, la tabla de nombres se replica en todos los directorios, por redundancia.
- dfs.replication
  Replicación de bloques predeterminada. El número real de replicaciones se puede especificar cuándo se crea el archivo. El valor predeterminado se utiliza si no se especifica la replicación en el tiempo de creación.Configuramos el archivo hdfs-site.xml en un datanode y lo copiamos al resto de datanodes.
- dfs.datanode.data.dir
  Determina en qué parte del sistema de archivos local un nodo de datos DFS debe almacenar sus bloques. Si se trata de una lista de directorios delimitada por comas, los datos se almacenarán en todos los directorios con nombre, normalmente en diferentes dispositivos. Los directorios deben etiquetarse con los tipos de almacenamiento correspondientes ([SSD]/[DISCO]/[ARCHIVO]/[RAM_DISK]/[NVDIMM]) para las directivas de almacenamiento HDFS. El tipo de almacenamiento predeterminado será DISK si el directorio no tiene un tipo de almacenamiento etiquetado explícitamente. Los directorios que no existan se crearán si el permiso del sistema de archivos local lo permite.

Formateamos el namenode: hdfs namenode -format

Configuración avanzada de Hadoop:

- seconds.dfs.heartbeat.interval
  Determina el intervalo de latidos del nodo de datos en segundos. Puede usar el siguiente sufijo (no distingue entre mayúsculas y minúsculas): ms(millis), s(sec), m(min), h(hour), d(day) para especificar la hora (como 2s, 2m, 1h, etc.). O proporcione el número completo en segundos (por ejemplo, 30 por 30 segundos). Si no se especifica ninguna unidad de tiempo, se asume

- dfs.namenode.heartbeat.recheck-interval
  Este tiempo decide el intervalo para comprobar si hay nodos de datos caducados. Con este valor y dfs.heartbeat.interval, también se calcula el intervalo para decidir si el nodo de datos está obsoleto o no. La unidad de esta configuración es el milisegundo.

- dfs.hosts
  Asigna un nombre a un archivo que contiene una lista de hosts a los que se les permite conectarse al nodo name. Se debe especificar el nombre completo de la ruta de acceso del archivo. Si el valor está vacío, se permiten todos los hosts.
- dfs.hosts.exclude
  Asigna un nombre a un archivo que contiene una lista de hosts a los que no se les permite conectarse al nodo name. Se debe especificar el nombre completo de la ruta de acceso del archivo. Si el valor está vacío, no se excluye ningún host.

- dfs.blocksize
  El tamaño de bloque predeterminado para los archivos nuevos, en bytes. Puede utilizar el siguiente sufijo (que no distingue entre mayúsculas y minúsculas): k(kilo), m(mega), g(giga), t(tera), p(peta), e(exa) para especificar el tamaño (por ejemplo, 128k, 512m, 1g, etc.), o proporcionar el tamaño completo en bytes (por ejemplo, 134217728 para 128 MB).

Arrancamos el namenode, los datanodes y confirmamos que están vivos mediante la interfaz web. hdfs –daemon start namenode, hdfs dfs –daemon start datanode.
Si hemos añadido los datanode en el fichero workes podemos utilizar start-dfs.sh

Comandos

- Hdfs dfs - …, donde ... son muy parecidos a lo de Linux con ext4
- hdfs getconf -confKey
  Obtiene una clave específica de la configuración hdfs dfsadmin –reconfig

  Inicia la reconfiguración u obtiene el estado de una reconfiguración en curso, u obtiene una lista de propiedades reconfigurables. El segundo parámetro especifica el tipo de nodo. El tercer parámetro especifica la dirección del host. Para el inicio o el estado, datanode admite livenodes como tercer parámetro, que iniciará o recuperará la reconfiguración en todos los datanodes activos.

- hdfs dfsadmin –report
  Informa sobre la información básica del sistema de archivos y las estadísticas, El uso de dfs puede ser diferente del uso de "du", ya que mide el espacio bruto utilizado por la replicación, las sumas de comprobación, las instantáneas, etc. en todos los DN. Se pueden usar indicadores opcionales para filtrar la lista de DataNodes mostrados. Los filtros se basan en el estado de la DN (por ejemplo, activo, muerto, desmantelamiento) o en la naturaleza de la DN (por ejemplo, nodos lentos, nodos con mayor latencia que sus pares).