Historial de Versiones

Acá se van a detallar los diferentes cambios que se realicen al trabajo luego de cada entrega a fin de tener una mejor trazabilidad.

En el siguiente repositorio es el que contiene el versionado del trabajo:

`https://github.com/ulisescolina/UC-PYLP/tree/master/Integrador`

Versión	Cambios
1.0.0	Primer entrega.

Apache Hadoop: Una guía paso a paso

Instalación en un nodo simple

Ramirez Ulises
Universidad Nacional de Misiones
ulisesrcolina@gmail.com
18 de septiembre de 2021

Resumen

Este documento es una continuación de lo charlado en el documento introductorio al modelo MapReduce (MR) [1], aquí se procede a la descripción de cómo configurar un cluster Hadoop para el procesamiento paralelo y distribuido. En esta instancia la configuración se hará sobre un único nodo.

Algo imporntante a tener en cuenta es que los pasos seguidos para esta configuración pueden quedar obsoletos ante cambios en distintos paquetes de los cuales depende el presente tutorial. Se recomienda que ante la incapacidad de poder realizar una tarea se revisen los canales de distribución oficiales para el paquete en cuestión.

Palabras clave: *Apache Hadoop, Tutorial.*

1. Introducción

El comportamiento por defecto que tiene Hadoop le permite funcionar en entornos que solamente cuentan con un nodo, a esto se lo llama "modo local" (o también se lo denomina *standalone*), un nodo configurado en este modo permitirá la compilación/ejecución de los mappers y los reducers. A lo largo de las secciones siguientes se procederá con la configuración de un nodo Hadoop en modo local.

2. Pre-Requisitos

Los pre-requisitos para poder realizar el tutorial que se describe en este documento serán los siguientes:

1. **GNU/Linux:** Es necesario tener instalada alguna distribución GNU/Linux, este es el sistema de-facto para desarrollo y producción de Hadoop,

el mismo se probó en producción con clusters de más de 2000 nodos, y aunque se puede instalar el mismo en una máquina con Windows[1], este procedimiento no se cubre en el documento.

2. **Java:** Hadoop requiere que el sistema tenga instalada una versión de Java, la versión de java dependerá de la versión de Hadoop, para descubrir cual versión es la recomendada, se puede ver las versiones compatibles[2]. En este ejemplo no se está utilizando una versión oficial de Oracle, se está utilizando una versión libre, *OpenJDK 10.0.2*.
3. **ssh:** Esta herramienta para el acceso remoto a equipos de una red determinada debe estar instalada para que la versión pseudo-distribuida y distribuida de Hadoop pueda establecer comunicación inter-nodos (o inter-procesos para la versión pseudo-distribuida).

La instalación de Hadoop es sencilla de realizar, dependiendo de la conexión a internet y el grado de cumplimiento de los pre-requisitos se puede tener funcionando un "cluster" de un solo nodo en cuestión de minutos, primero vamos a recorrer los pasos necesarios para cumplimentar esto, un cluster de un solo nodo. Las versiones que posteriores a este documento detallarán el procedimiento para la instalación del software en un entorno con múltiples nodos.

3. Obtención del paquete

La primer tarea que tenemos que llevar a cabo es la obtención de Hadoop, para esto se seguirán las recomendaciones propuestas por los desarrolladores de descargar el archivo comprimido de la página oficial [3]. Por el momento se asume que el archivo descargado queda almacenado en:

$$\sim /Descargas/hadoop - 3{,}3{,}1.tar.gz$$

[1] https://cwiki.apache.org/confluence/display/HADOOP2/Hadoop2OnWindows

[2] `https://cwiki.apache.org/confluence/display/HADOOP/Hadoop+Java+Versions`

[3] `http://www.apache.org/dyn/closer.cgi/hadoop/common/`

NOTA:

Otra de las formas que se puede obtener una instalación de Hadoop, es a través de los repositorios oficiales de la distribución que se encuentre ejecutando en la máquina. Los repositorios compatibles con Ubuntu tendrán la posibilidad de agregar un PPA y obtener los paquetes necesarios para Hadoop.
Más información acerca del repositorio disponible para Ubuntu y derivados (Mint, Debian[a]) se pueden encontrar en el siguiente enlace:

`https://launchpad.net/~hadoop-ubuntu/+archive/ubuntu/stable`

Otras distribuciones (Fedora, RHEL, Manjaro, Gentoo, etc.) deberán buscar en sus repositorios oficiales en caso de querer seguir este acercamiento.

[a] Ubuntu es una derivación de Debian, pero hay veces en donde los paquetes de una distribución no tienen problemas funcionando en la otra, sin embargo, esto no se probó en el caso específico de Hadoop.

4. Configuración del entorno

Aquí se describen cuestiones que son necesarias con los pre-requisitos, pero que tienen que ver más con cuestiones de configuración del Sistema en general. Estas cuestiones son de especial importancia si se quiere establecer el entorno de acuerdo a las buenas prácticas para la implantación del cluster con múltiples nodos que actuen como *workers*, en este apartado de configuración de entorno tendría que procederse a la creación de usuarios específicos para el uso del cluster, la generación de claves asíncronas para la conexión con los demás nodos del cluster.

Sin embargo, se omiten estas tareas para poder retomarlas en una versión futura del documento, primero nos ocuparemos de tener una versión funcional de Hadoop en un nodo.

5. Establecer el JAVA_HOME

Como se menciona en el documento que describe al software Hadoop[1], este es una herramienta construida en Java, y para su correcto funcionameiento se apoya sobre una instalación de la Java Virtual Machine (JVM), la forma de comunicar a Hadoop cuál es el directorio que contiene la instalación de la JVM es con la variable de entorno `JAVA_HOME`.

La variable de entorno del sistema que se denomina `JAVA_HOME` indica cual es el directorio en el cual se encuentra instalada la versión de Java que está utilizando la máquina. Para revisar si ya se tiene establecida la variable se puede realizar lo siguiente: "`echo $JAVA_HOME`".

Si el resultado de correr lo mencionado anteriormente es un directorio, el sistema tiene establecida la variable en cuestión, de lo contrario será necesario

que se haga manualmente.

JAVA_HOME

Para establecer una nueva variable de entorno en el sistema es suficiente con ejecutar

```
export JAVA_HOME="/usr/lib/jvm/java-10-openjdk/"
```

Listing 1: Exportar variable de entorno

El valor que debe obtener la variable de entorno va a depender de que versión de Java se tenga instalada además de si se tiene la versión que provee la empresa Oracle, o si se tiene la versión libre.

6. Descomprimir el paquete descargado

Hasta ahora se realizaron configuraciones referentes al entorno en el cual va a residir el nodo Hadoop, ahora se retoman las tareas relacionadas con el paquete descargado al inicio.

```
cd ~/Descargas/              # cambiar de directorio
tar -xzf hadoop-3.1.1.tar.gz # descompresion del archivo
mv hadoop-3.3.1.tar/hadoop-3.1.1/ ~/hadoop/
                 # se mueve todo el contenido a ~/hadoop/
```

Listing 2: Descompresión del paquete Hadoop

Lo listado anteriormente describe como realizar la descompresión de lo descargado en la sección 3, el primer paso es cambiar de directorio hacia donde se encuenra el paquete comprimido, esto se había asumido en la sección 3 que es `/Descargas/`, luego se descomprime con la herramienta `tar` (para ver más utilidades que proporciona la herramienta se puede ingresar `man tar` en un terminal), por último, se mueve el directorio y todo el contenido que resulta de la descompresión al directorio `/hadoop/`.

7. Prueba de funcionamiento

El directorio resultante de la descompresión contiene todo lo necesario para ejecutar Hadoop, en este documento nos vamos a enfocar en los contenidos de `bin` y los contenidos de `etc`.

El directorio bin contiene todos los puntos de acceso para la herramienta, se puede proceder a ejecutar lo siguiente para probar su funcionamiento:

```
cd ~/hadoop/
./bin/hadoop
```

Listing 3: Prueba de funcionamiento

Esto debería imprimir una pequeña ayuda acerca del paquete y es la prueba de que tenemos funcionando una instalación de Hadoop en el ordenador.

El directorio `etc`

En este directorio se tienen archivos de configuración o scripts para la modificación de valores por defecto utilizados dentro de Hadoop. Un ejemplo de esto y que ya se vió anteriormente es el establecimiento de la variable de entorno `JAVA_HOME`, si se revisa el archivo `hadoop-env.sh` se podrá encontrar parte de lo escrito con anterioridad para establecer la variable JAVA_HOME.

Una de las notas en el documento que es de interés es la siguiente:

```
Technically, the only required environment variable is
JAVA_HOME. All others are optional. However, the defaults
are probably not preferred. Many sites configure these
options outside of Hadoop, such as in /etc/profile.d
```

Esta da la pauta de que la configuración adecuada puede ser parte de un proceso que se tiene que adaptar a cada usuario, por lo que se tienen opciones por defecto, sin embargo, estas pueden no ser las adecuadas. Para el ejercicio que se realiza a lo largo del presente, se utilizan todos los valores por defecto por el hecho de que se esta configurando un nodo simple a modo de ejemplificación.

En versiones futuras del documento se planea continuar profundizando este asunto.

Referencias

[1] U. Ramirez, «Procesamiento Paralelo y Distribuido: Una respuesta a una creciente demanda de poder de procesamiento,» 2021. dirección: https://github.com/ulisescolina/UC-PYLP/blob/master/Integrador/integrador_pylp_ulises.pdf.

Siglas

MR MapReduce. 2

Glosario

cluster *Grupo* o también llamado *Granja de servidores*, es un término que se aplica a los sistemas distribuidos y hace referencia a un conjunto de máquinas interconectadas por una red de alta velocidad. 3