### Introducción Al BIG DATA - CLOUDERA - Ecosistema HADOOP

#### **Big Data**

El Big Data surgió cuando Google estaba en el proceso de indexar toda la web. Se encontró con ficheros enormes que no cabían en ningún servidor. Partiendo de este problema se diseñó Hadoop. Cuando manejamos ficheros del entorno de 1 PByte ya hablamos de Big Data.

#### Hadoop

Hadoop es un framework o conjunto de herramientas distribuido, escalable, tolerante a fallos y de código abierto para almacenar, procesar y analizar Big Data.

#### HDFS

Hadoop Distributed File System. Es un sistema de ficheros, distribuido, escalable, tolerante a fallos, escrito en JAVA. Se sitúa por encima del sistema de ficheros nativo. Los archivos hdfs son Write Once, no permiten appends, acceso random ni escritura. Están pensados para contener Big Data. O sea datos que van a ser consultados y leídos.

### Instalación De MySQL

Normalmente los datos se han almacenado siempre en bases de datos relacionales o RDBMS. Y luego se han consultado estas bases de datos con SQL. Existen también bases de datos que no son relacionales. Son jerárquicas o de clave/valor. Y también existen bases de datos que no son SQL.

Nosotros como primera aproximación al problema del Big Data vamos a instalar y configurar una base de datos de las muchas que hay. He escogido MySQL porque es gratuita, accesible y es un estándar ampliamente utilizado.

Veremos que el Big Data es una evolución natural o si se quiere, una especialización de las RDBMS para el caso de grandes ficheros.

#### Instalación De MySQL En Linux

Aquí vamos a optar por el camino fácil. Como casi nadie tiene en su portátil Linux nativo sino que el sistema anfitrión siempre es Windows y Linux se utiliza en máquinas virtuales, pues os voy a seleccionar una máquina virtual Linux especialmente preparada para trabajar con MySQL.

Dicha máquina se puede descargar [**aquí**](https://home.mycloud.com/action/share/d38dd28b-e69b-48ec-b844-70b16943fdd4). Cuando pulséis descargar, aparecerá el fichero:

CentOS MySQL schemas.ova

Que es una máquina virtual de Oracle Virtual Box. Esta máquina tiene una particularidad y es que solo funciona bien con una determinada versión de Virtual Box, que se puede descargar [**aquí**](https://home.mycloud.com/action/share/6f942d48-0bb0-428f-89c7-97f9b67e861e). El resultado de la descarga es el fichero:

VirtualBox-6.1.42-155177-Win.exe

Que es el primer programa que deberemos instalar en nuestra máquina anfitriona. Una vez instalado. Deberemos instalar las extensiones de máquina virtual, que se pueden descargar [**aquí**](https://home.mycloud.com/action/share/721c1302-790a-4bff-9c20-dec19b0f0836). El resultado de la descarga es el fichero:

Oracle\_VM\_VirtualBox\_Extension\_Pack-6.1.42.vbox-extpack

Que deberá instalarse en segundo lugar en nuestra máquina anfitriona Windows.

Las extensiones de máquina virtual es una serie de utilidades para definir una carpeta compartida entre las máquinas virtual y anfitriona y para poder usar el corta, copia pega y arrastrar entre ambas máquinas. Siempre hay que instalar las extensiones en todas las máquinas virtuales.

Una vez instalado Oracle Virtual Box, lo arrancamos y debe aparecernos algo así:

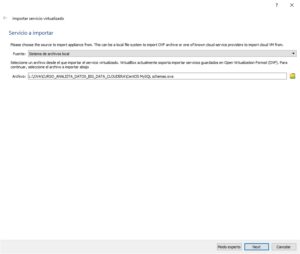


A vosotros aún no porque no tenéis la máquina instalada. A mi me sale abajo porque ya la he instalado. La manera de instalarla es la siguiente:

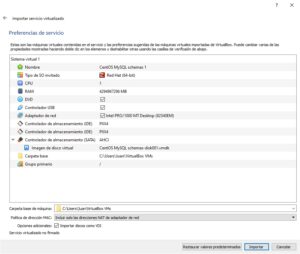
Pulsamos, Archivo, Importar servicio virtualizado y buscamos el archivo:

CentOS MySQL schemas.ova

Pulsamos siguiente:



Y luego pulsamos siguiente;



Y sin cambiar nada, pulsamos importar. Es posible que tarde unos minutos en hacer la importación, pero una vez terminada, ya no tendremos que hacer nada mas. No hay que modificar ningún parámetro ni configurar nada. Simplemente hacer la importación. Esto se hace una sola vez.

Ahora arrancamos la máquina virtual. Desde:



Nos situamos en la máquina virtual señalada pulsamos el botón iniciar.

Pinchamos en el usuario cloudera y como password también pulsamos cloudera. Tenemos un CentOS Linux con los siguientes iconos en el desktop:



Tenemos el terminal y el programa MySQL Workbench, que son las dos maneras de acceder al RDBMS MySQL. Pues bien empezamos por el mas fácil, que es abrir un terminal y tecleamos:

mysql -ucloudera -pCl@7d3r4

Es decir entramos en MySQL con el usuario: cloudera y password: Cl@7d3r4



Una vez dentro de mysql, puedo ver si ya hay bases de datos cargadas para practicar:

mysql> show databases;

Y resulta que hay varias. otra ventaja de haber recurrido a esta máquina de Linux que ya viene con ejemplos de bases de datos precargados en el mysql. Vamos a utilizar una de las bases de datos que vienen:

mysql>use demobld;

Y ahora vamos a ver las tablas que tiene esa base de datos:

mysql>show tables;

Y podríamos ver los 10 primeros registros que tiene cualquiera de ellas:

mysql>SELECT \* FROM departments LIMIT 10;



Y así podríamos hacer cuantas consultas SQL quisiéramos sobre esa base de datos. Hay un excelente curso rápido de SQL compatible con MySQL y MaríaDB que se puede descargar [**aquí**](https://home.mycloud.com/action/share/e2d159ae-460f-4a37-8d2d-ce479504d512).

Yo recomiendo construir las consultas SQL en un fichero batch de varias consultas SQL que ejecutaremos en bloque desde la línea de comandos de mysql.

Los problemas de SQL no se suelen solucionar con una única instrucción, y el archivo de ejecución por lotes es el mejor método. Sería el equivalente a las macros de Access para quien conozca esa herramienta.

Veamos un ejemplo de uso de estos ficheros de script de consultas SQL. Para ello vamos a ver en primer lugar como nos podemos comunicar, como podemos transferir ficheros entre la máquina virtual y la máquina anfitriona (Windows).

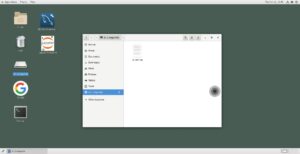
Bien, pues esta es otra de las ventajas de trabajar con una máquina virtual específica. Ya viene programado y configurado un sistema para hacer esta comunicación.

Lo único que debemos hacer es crear una carpeta en nuestra máquina anfitriona (Windows):

C:\Compartida

Así, respetando mayúsculas y minúsculas, tal y como está. Y dejamos en esa carpeta los ficheros que quiera compartir con la máquina CentOS. En mi caso voy a dejar un fichero de script llamado script1.sql

Una vez hecho esto vuelvo a la máquina anfitriona y doy doble click al icono que pone sf\_Compartida.

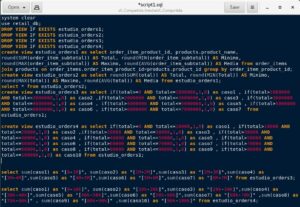


Y ahí podemos ver el archivo cuyo path completo es:

/media/sf\_Compartida/script1.sql

podemos editar este script que no es otra cosa que un fichero fuente con cualquier editor de texto. Por comodidad y sencillez recomiendo usar el que viene por defecto, en Applications, Accesories, Text editor.

Lo abrimos y vamos a buscar el fichero script1.sql para verlo:



Como veis es una serie de sentencias SQL que hacen algo complejo. No es para teclear en la línea de comando del terminal, porque son varias y seguro que nos equivocaríamos. Así que hacemos nuestro script así. Ahora vamos a ejecutarlo.

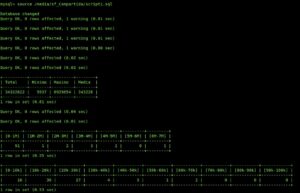
Abro un terminal y ejecuto las instrucciones:

mysql -ucloudera -pCl@7d3r4

de esta manera entramos en MySQL. Ahora ejecutamos el script:

mysql> source /media/sf\_Compartida/script1.sql

Y el resultado es:



Básicamente lo que hace este script es una distribución de frecuencias. os remito al curso de SQL para que aprendáis SQL, es la base del BigData.

Esto mismo se puede hacer sin utilizar el terminal. Para ello arrancamos el icono que pone "MySQL Workbench". Luego pulsamos el botón "ClouderaLocal".

Seleccionamos la base de datos retail\_db y abrimos el script1.sql y aparecerá algo así:



Si pulsamos el icono del rayo, se ejecutará el script y dará el mismo resultado que por el terminal.

A veces se pierde la conexión con MySQL. Si esto ocurre cambiar la conexión de red de la máquina virtual a NAT bridged y reiniciar la máquina.

Bueno, pues ya tenemos una idea de como instalar nuestra base de datos MySQL en Linux con ayuda de ésta máquina virtual. Tenemos un curso de SQL y hemos visto un ejemplo de como poder manejar una base de datos de ejemplo (hay varias) tanto desde el terminal como desde un entorno gráfico.

### Necesidad De BigData

El siguiente paso es responder a la pregunta ¿qué ocurre cuando manejamos ficheros enormes, del entorno o similares a 1 Petabyte?. Un Petabyte son 1024 teras, y un tera son 1024 gigas y un giga son 1024 megas. En 3 megas ya cabe una canción estándar.

Un fichero de 1PB no cabe en ningún disco conocido por grande que sea. Los discos mas grandes que se están haciendo ahora así que se puedan comprar en Amazon no pasan del los 20TB. Se impone pues que hay que trocear el fichero y almacenarlo en muchos discos distintos dentro de muchos pc distintos.

Para ello necesitamos un sistema de archivos distribuido. Esto será el HDFS. Y esa será la base del resto de utilidades del ecosistema Hadoop. Cloudera es una empresa que configura y da soporte a soluciones de BigData basadas en Hadoop.

Bueno, y ahora viene la gran pregunta: ¿Cómo nos instalamos un "cloudera"?

Pues es difícil, pero se puede hacer. Existe una máquina virtual que tiene una instalación de cloudera v. 6.3.2 con todas las utilidades y que se puede descargar [**aquí**](https://home.mycloud.com/action/share/e03ed6de-e515-414e-93db-8219652cfd0c).

Lamentablemente, la máquina anfitriona ha de tener al menos de 16 GB de RAM para que medio funcione cloudera, mucho mejor con 32. Si tu máquina anfitriona (Windows) tiene menos, no va a funcionar.

bien, una vez descargada la imagen de VMware nos habrá generado una carpeta con el nombre:

CDH\_6.3.2\_CentOS7

Esta carpeta, la dejamos en nuestro laboratorio, que lo ideal es que sea un disco independiente SSD, pero puede estar en cualquier disco.

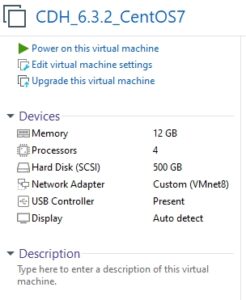
Ya explicamos en post anteriores, como instalar VMware, que se puede descargar [**aquí**](https://home.mycloud.com/action/share/6284d730-d4b0-4164-9091-6b04e317bec2). Y las tools par VMware, se pueden descargar [**aquí**](https://home.mycloud.com/action/share/bece7c19-a0a7-4a37-9531-777797035ddd).

En [**este post**](https://juanjosevivas.es/escaneo-de-redes-usando-metasploit/) explico como instalar la red dentro del VMware. La instalación de VMware así como las tools, entiendo que sabéis hacerla ya que si estáis leyendo esta web, es que sois usuarios avanzados.

Arrancaremos VMware:

Menú, File, Open y nos vamos a la carpeta donde está la máquina cloudera que acabamos de descargar. Entramos en la carpeta y abrimos el fichero .vmx

Y tendremos algo así:



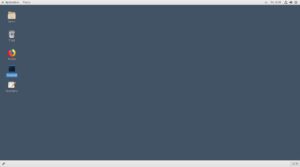
Arrancamos la máquina:

Los datos básicos de esta instalación cloudera en CentOS son:

Cloudera QuickStart VM 6.3.2  
=======================================  
CentOS 7 + GNOME Based  
Java Eclipse & Scala Eclipse IDE Included  
MySql With 'retail\_db' Installed  
Minimum System Requirement - 2/4 Cores + 16GB RAM

---------------------------  
CentOS GUI Login 'Base User' Password - BaseUser@123  
'root' Password - BaseUser@123  
---------------------------  
sudo user - osboxes  
sudo password - BaseUser@123  
---------------------------  
MySql user - root  
MySql password - bigdata  
---------------------------  
Cloudera Manager user - admin  
Cloudera Manager password - admin

Os recomiendo sacar el teclado en pantalla para teclear la clave de entrada al CentOS. Una vez dentro veremos algo así:



Leer bien lo que acabo de escribir para entrar en esta máquina virtual, ahí está toda la información. Bien, una vez dentro vamos a configurar la red para que esto funcione.

En primer lugar voy a suponer que habéis instalado la red VMnet8 tal y como explico en el enlace de mas arriba. No tiene ningún misterio, se hace en 5 minutos.

Aseguraros de que el adaptador de red está conectado a esa red que acabamos de definir en el VMware.

Ahora abrimos un terminal y vemos nuestra dirección IP

ifconfig

Y tomamos nota de nuestra IP



Si habéis hecho bien la configuración de red, tiene que haberos dado una IP que empiece por 10.10.1, imaginemos que esa IP es 10.10.1.3, cada uno pondrá la que le haya suministrado sus sistema, esta IP me la he inventado a efectos de ilustrar la instalación.

Ahora editamos el siguiente archivo:

sudo gedit /etc/hosts

Y añado al final del archivo la línea:

10.10.1.3 quickstar-bigdata

El fichero, al final debe mostrar algo así:

127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4  
::1 localhost localhost.localdomain localhost6 localhost6.localdomain6  
10.10.1.3 quickstart-bigdata

Salimos grabando

A continuación editamos el siguiente archivo:

sudo gedit /etc/cloudera-scm-agent/config.ini

Y cambiamos la línea:

# Hostname of the CM server.  
server\_host=quickstart-bigdata

Tiene que quedar como acabo de escribir.

Salimos grabando y a reinicializamos los servicios:

sudo systemctl restart cloudera-scm-agent

sudo systemctl restart cloudera-scm-server

sudo tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log

Ahora vamos a comprobar que todo funciona: Es posible que haya que reiniciar la máquina virtual.

Arrancamos el Firefox:

pulsamos en cloudera manager:

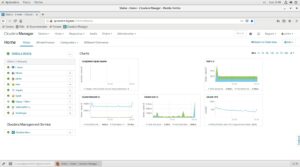
usuario: admin

password: admin

Ahora vamos a hacer un restart de los servicios. Entramos en cloudera management services, luego en instancias. Marcamos Host Monitor y Server Monitor, luego actions, luego restart y luego close.

Luego Cloudera Manager, HDFS, carpeta azul, botón restart, Stale Services, Finish.

Si todo va bien, todos los servicios y sistemas de nuestro clúster de BigData debería estar todo en verde:



### Instalación De MySQL Workbench En CLOUDERA

Sabemos que en nuestro clúster de CLOUDERA de un solo nodo instalado en la máquina CentOS7 que estuvimos analizando en el post anterior, tenemos instalada una base de datos MySQL cuyos parámetros eran:

MySql user – root MySql password – bigdata

Pero no tenemos instalado el MySQL Workbench y eso vamos a hacer ahora. Entramos en la máquina CentOS7 y abrimos un terminal. En Centos y Red Hat 7 existe un bug que impide la ejecución de MySQL Workbench en ciertas versiones. La instalación aparentemente se realiza sin problemas pero al ejecutarlo aparecen errores del tipo:

GtkDialog mapped without a transient parent. This is discouraged.  
Process 3730 (mysql-workbench-bin) of user 0 killed by SIGSEGV – dumping core  
\*\*\* Segmentation fault

Este error ocurre con las últimas versiones de MySQL Workbench, concretamente con 6.3.9 y 6.3.10.

La solución pasa por **instalar la versión 6.3.8**. Esta versión sí funciona en Centos/RHEL 7 pero hace falta instalar previamente unas dependencias para que se ejecute correctamente y el repositorio **EPEL**. Estas librerías son **libzip y tinyxml**. Veamos entonces cual sería el proceso:

Abrimos un terminal y tecleamos los siguientes comandos:

sudo su

yum-config-manager –disable cloudera-manager

Este primer comando elimina por defecto el ir primero al repositorio de Cloudera, que además ni siquiera está disponible. Da un error 404.

sudo yum install epel-release -y

sudo yum install libzip tinyxml

wget http://repo.mysql.com/yum/mysql-tools-community/el/7/x86\_64/mysql-workbench-community-6.3.8-1.el7.x86\_64.rpm

yum install rpm mysql-workbench-community-6.3.8-1.el7.x86\_64.rpm -y

Y si todo ha ido bien, ya debe estar instalado nuestro MySQL Workbench y debe haber aparecido un icono en Applications/Programming

### 

Arrastraremos el icono al desktop de nuestra máquina cloudera. una vez allí pulsaremos doble click. Saldrá la siguiente pantalla:



Pulsamos el botón de conexión situado en la esquina superior izquierda. Debería salir una ventana ya del MySQL Workbench:



Como podemos ver ya hay varias bases de datos instaladas en el sistema, entre ellas retail\_db. Vamos a recordar esta base de datos que nos será útil mas adelante.

### Instalación De HUE En CLOUDERA

Unos de los servicios mas importantes de CLOUDERA es el HUE, que es un entorno gráfico de usuario para Hadoop. Mas en concreto para HDFS – Hive – Impala. Recordemos que HDFS es la capa de almacenamiento, un sistema de ficheros distribuido que se monta sobre el sistema de archivos subyacentes que básicamente es el de Linux. Hive e Impala son motores SQL para hacer consultas sobre datos que están en HDFS. HUE viene a ser algo parecido a lo que es MySQL Workbench para la base de datos MySQL.

Bien, pues procedemos a la instalación de HUE.

Arrancamos FireFox.

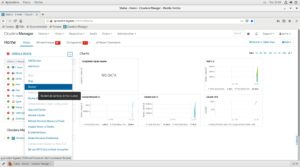
Y pulsamos el botón de Cloudera Manager. Ingresamos el usuario/contraseña que es admin/admin:



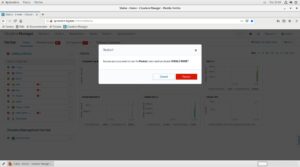
El clúster debería presentar todos los servicios en verde (o casi todos), tal como esto:



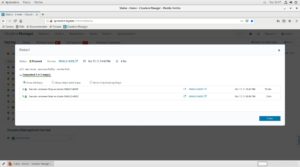
Pero a veces puede ocurrir que muchos servicios se presenten en rojo, tal como la siguiente imagen. En ese caso, en el desplegable que hay al lado de “single-node”, seleccionamos “restart”:



Y comienza el proceso de reinicialización de todos los servicios:



Pulsamos en restart:



Y pulsamos close. Esto debería haber reinicializado todos los servicios. Si alguno no estuviera reiniciado, podemos hacer un reinicio solo de ese servicio desplegando el menú correspondiente situado a la derecha del servicio.

Unos servicios tienen dependencias de otros. Es probable que si por ejemplo, HDFS está reiniciándose, muchos otros servicios como Hive o Impala estén pausados ya que dependen de que HDFS esté presente, que al fin y al cabo es el sistema de archivos.

Una vez que todo esté en verde, desplegaremos el menú situado a la derecha de “SINGLE-NODE” y seleccionaremos “add service”

Bueno, y con esto finalizamos este post, que tenía por objetivo introducirnos en el mundo del BigData e instalar un sistema en miniatura pero absolutamente real de BigData en nuestro PC.

